Proteome-wide mapping of sequence-function relationships using mistranslation

Kyle N. Hess

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

Stanley Fields

James Bruce

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Proteome-wide mapping of sequence-function relationships using mistranslation

Kyle N. Hess

Chair of the Supervisory Committee:
Judit Villén
Department of Genome Sciences

Amino acid substitutions fuel molecular innovations across the tree of life, yet they also underpin the majority of molecular, cellular, and organismal dysfunction. Delineating between such disparate mutational outcomes is critical in fulfilling the promise of precision medicine and harnessing the power of proteomics for understanding and engineering protein biology. In this dissertation, I present experiments establishing modular and high-throughput proteomic methods to characterize the effects of amino acid substitutions on protein structure and function *en masse*. Specifically, I showcase Miro (Chapter 2), a proteomics platform that expands mutational scans from single proteins to entire proteomes. I helped establish Miro in *Saccharomyces cerevisiae* and, once established, I applied this technology to systematically probe the effects of non-canonical amino acid (ncAA) substitutions on protein thermal stability (Chapter 3). Specifically, I first developed a high-throughput thermal stability assay inspired by Thermal Proteome Profiling and Proteome Integral Solubility Alteration. Using this streamlined method, I then coupled it with eight mistranslated proteomes and quantified the effects of ~9000 ncAA

substitutions on the stability of >700 proteins. I computationally mapped substitutions back to protein structure to reveal a significant role of local sequence contexts in shaping the impact of a ncAA substitution. I also expanded my analysis to generate protein-specific mutational sensitivity maps, which uncovered clusters of deleterious mutations close in both sequence and three-dimensional space. Many of these clusters also overlapped with regions of known function, highlighting how positional ncAA sensitivity can illuminate functional protein regions. I then coupled this high-throughput stability assay with small molecules to map ATP binding sites across the yeast proteome (Chapter 4). Lastly, I used TPP to identify substrates of the SARS-CoV-2 protease, NSP5, in human cell lines (Chapter 5), by looking for changes in protein stability that arise due to the expression of different NSP5 constructs (wildtype, catalytically dead, or GFP only).

TABLE OF CONTENTS

LIST OF FIGURES

DEDICATION

In loving memory of my dear friend, Vincent A. Fernandez (1991-2021).
Full of unbridled curiosity and unconditional love. I miss you Vinnie. I always will.

And in loving memory of my uncle, Kurt W. Hess (1944-2020).
A fearless scientist and passionate environmentalist steadfast in his fight for change.

ACKNOWLEDGEMENTS

I have been fortunate in my life to be surrounded by incredibly generous and kind individuals who have offered so much of their time, wisdom, mentorship, love, and support. These individuals have been a core part of my scientific journey and development; I am indebted to them for the scientist and person I have become.

First and foremost, I want to thank the scientific mentors and colleagues for whom, without their support, this journey would not have been possible. Nikolas Nikolaidis, who was the first to put a pipette in my hand and allow me to perform experiments in his lab at California State University, Fullerton. Thank you for helping me establish a love and appreciation for molecular biology and evolution, which engendered much of my fascination with the natural world.

Thank you Judit Villén and Ricard Rodriguez-Mias, who have been my role models and have become a core part of my scientific identity. Judit, thank you for fostering my scientific and personal growth, and supporting my pursuits outside of the lab in science education and outreach. Thank you for giving me the freedom to take my project in a myriad of directions. Your scientific guidance and mentorship has been instrumental in my path towards becoming an experimental and computational scientist, and in achieving my long-term career goals. Ricard, thank you for being by my side at the bench over the past six years. I will always cherish the late nights in lab harvesting yeast, the invaluable life advice, and our frequent visits to the whiteboard to talk through ideas. Our friendship has been the highlight of my graduate school experience and I look forward to seeing it continue to grow. And to the both of you: thank you for entrusting me with what has turned out to be both an intellectually challenging and incredibly rewarding thesis project that helped me grow into the scientist I am today.

I would also like to thank the individuals who are a part of my broader scientific community. Stanley Fields, who helped shape my approach to genomic and proteomic technology development. Christine Quietsch and Debbie Nickerson, both of whom helped me establish a sense of identity and belonging, respectively, within the Department of Genome Sciences. Fiona McAllister and Niclas Olsson, who let me test the waters of industry for a summer. Rob Lawrence, whose conversations were pivotal when choosing the MCB PhD program at UW. Stephanie Zimmerman, a colleague and mentor who instilled in me a sense of courage and optimism when developing novel technologies. Ben Wiggins, Jolie Carlisle, and Michael Cargill, all of whom shaped my identity as a science educator. And my committee members, Stanley Fields, Kelley Harris, Dustin Maly, and James Bruce, whose thoughtful feedback and questions during committee meetings helped guide my project in fruitful directions.

I want to thank all the individuals I have had the fortune of meeting in the Molecular and Cellular Biology program and Department of Genome Sciences. Never in my wildest dreams would I have imagined finding such an amazing group of creative, kind-hearted, and humble scientists, administrators, and staff, who have built a world-class, supportive, and scientifically-rigorous interdisciplinary training environment second to none. I would also like to thank the MCB and GS class of 2016 and all the friends I have made within both of these communities.

I also want to thank the vibrant Villén lab community. First, I would like to thank the cohort of 2016 graduate students in the lab, Ian Smith, Bianca Ruiz, and Anthony Barente, all of whom I had the chance to grow with for the last six years, and who have become family for me and Julia. Thank you Sam Entwisle, Miguel Martin-Perez, and Ariadna Llovet-Soto, for the positive impact you all had on my early development as a scientist, and for helping me feel at home in the lab. Thank you Mario Leutert, for being a wonderful friend and mentor, and for filling the lab

with joy. And thank you to the wonderful community of friends and colleagues that have joined the lab over the years, Alex Hogrebe, Sophie Moggridge, Matt Berg, and Alexis Chang. I am grateful I had the chance to conduct exciting research in the presence of such amazing people.

I also want to thank my uncle Kurt Hess and dear friend Vincent Fernandez. Kurt encouraged me to take education more seriously after high school and completely changed my life's trajectory through conversations during my senior year. Vincent Fernandez was a consistent presence during graduate school, whose curiosity and optimism were a constant source of comfort and revitalization. I will miss Kurt and Vinnie dearly.

Thank you to my loving family. Thank you Loretta, Mike, and Jenny, for being a source of unwavering support for me and Julia. Thank you to my Mom and Dad, Scott and Connie, for always supporting me; whether it was theater, soccer, or pursuing a degree in biology, I always knew I had your full love and support. Thank you Mom for instilling in me a sense of wonder, curiosity, and gusto for life, and always encouraging me to find my higher purpose. And thank you Jason and Jessica for the many adventures over the years. Thank you Jason for being my first goalkeeper coach and helping me learn how to persevere through hardships in the process.

And above all else, thank you to my spouse and life partner, Julia. You have always been a grounding force in my life. Thank you for your unconditional love and support, and for your intrepid spirit throughout this six year journey. Your kind heart, sense of humor, and radiant positivity helped me persevere through graduate school, and helped me remember the important things in life. I love you with all my heart, and I am so excited to be returning home with you and our little bean, Maggie. And of course, thank you Sweet Cat, for changing our lives forever after breaking into our apartment on that fateful afternoon in November 2017.

Chapter 1. Sequence-function relationships

1.1 A revolution in protein biochemistry

Protein biochemistry is in the midst of a revolution. Structures for entire proteomes can be predicted from amino acid sequence with the accuracy and resolution of crystal structures (Jumper et al. 2021; Tunyasuvunakool et al. 2021; Leman et al. 2022; Baek et al. 2021); missense variants can be accurately classified *in silico* as benign or disease-causing for many disease genes (Frazer et al. 2021), and may soon aid clinical diagnostics; millions of mutant proteins can be assayed in parallel to inform sequence-function relationships (Fowler et al. 2010) or to map the clinical effects of genetic variants (Findlay et al. 2018); cell signaling dynamics can be quantified proteome-wide (Leutert et al. 2019; Needham et al. 2021; Humphrey, Azimifar, and Mann 2015); the functions of post-translational modifications can be accurately predicted (Ochoa et al. 2020); and synthetic molecular machines can be designed from scratch (Courbet et al. 2022). Protein biochemistry is in the midst of a revolution, and is transforming the world as we know it.

Technological developments in genomics and proteomics lie at the heart of this revolution. In genomics, next-generation DNA sequencing (Shendure and Ji 2008) has lowered the cost of sequencing by five orders of magnitude, enabling high-throughput genetic screens (Y. Zhou et al. 2014), deep mutational scanning (Fowler and Fields 2014), multiplexed genetic engineering (Wang et al. 2009), and high-throughput variant characterization (Fowler et al. 2010; Starita et al. 2017). In proteomics, high-resolution mass spectrometers, many of which now fit on laboratory benchtops, routinely acquire 10's to 100's of scans per second and quantify 100's to 1000's of peptides and proteins per minute (Bekker-Jensen et al. 2020). These advances in

instrumentation have coincided with advances in sample multiplexing (Ong et al. 2002; J. Li, Van Vranken, Pontano Vaites, et al. 2020), software (Kong et al. 2017; Schweppe et al. 2020), and acquisition strategies (Ludwig et al. 2018; McAlister et al. 2014; Meier et al. 2018), with impacts in drug development (Savitski et al. 2014; Perrin et al. 2020; Savitski et al. 2018), post-translational modification (PTM) characterization (Smith et al. 2021; Potel et al. 2021; Zecha et al. 2018, 2022; Bludau et al. 2021; J. X. Huang et al. 2019), and biomarker discovery (Geyer et al. 2016, 2021). The convergence of genomics and proteomics technologies has transformed protein biochemistry and is ushering in the age of high-throughput protein science.

1.2 The sequence-function paradigm

Central to high-throughput protein science is the protein sequence-function paradigm. The sequence-function paradigm posits that protein sequences encode all the molecular information needed to form protein structure and function. Much of the support for this relationship was established throughout the 20th century, which cemented our understanding of some of the most fundamental cellular processes in biology. For example, the structural elucidation of aminoacyl-tRNA synthetases (aaRSs) enabled a mechanistic understanding of amino acid discrimination and translational fidelity (Perona and Hadd 2012). Solving the structures of this heterogenous protein family revealed general principles underlying amino acid discrimination, spanning mechanisms of steric hindrance (Arnez, Dock-Bregeon, and Moras 1999), molecular recognition of amino acid-specific chemical moieties in aaRS active sites (Cavarelli et al. 1998; Delagoutte, Moras, and Cavarelli 2000), to the “double-sieve” editing of mischarged tRNA products when the active site is more promiscuous (Fersht 1977; Fersht and Dingwall 1979; Fukai et al. 2000). As a result of this basic understanding between aaRS structure and function, scientists have gone on to engineer orthogonal translation systems for applications in protein biochemistry and synthetic biology (Zimmerman et al. 2018; Berg et al. 2019; Cervettini et al.

2020).

However, relying solely on structural information to infer protein function is, in and of itself, limited for several reasons. First, structure-and-function is not a one-to-one relationship; a variety of structural solutions can exist for any given protein function (Galperin, Walker, and Koonin 1998; Doolittle 1994). Most of the members of the serine protease family, for example, share the serine-histidine-aspartate catalytic triad essential to their activities. Yet, different family members have converged on this function during evolution, arriving at this endpoint with diverse structural arrangements (Rawlings and Barrett 1993). Examples of convergent evolution can also be found in other protein families, such as for some aaRSs (Perona and Hadd 2012) and nucleotide exchange factors (Sondermann et al. 2001).

Second, protein structures are essential scaffolds for the evolution of new protein functions. As a result, any given protein structure can give rise to a variety of novel functions. The family of Hsp70 molecular chaperones, for example, is a highly-conserved protein family deeply rooted in the tree of life whose structure and essential functions have remained (mostly) conserved, but whose auxiliary functions have diversified as the family expanded (Faust et al. 2020; McCallister et al. 2015). In some instances, proteins can acquire new functions without undergoing much sequence change at all. As a fascinating example, crystallin proteins in the eyes of avian species are closely related to several metabolic enzymes (Wistow and Piatigorsky 1987; Piatigorsky and Wistow 1991), with some crystallins being identical to their metabolic counterpart (G. J. Wistow, Mulders, and de Jong 1987). These structural and functional conundrums complicate the structure-function paradigm. Identical functions can arise from different protein structures, and proteins with the same structure can evolve dissimilar functions.

It was not until the second half of the 20th century that protein sequence emerged as an

essential component to the structure-function paradigm. Through several fundamental experiments with ribonuclease, Christian Anfinsen showed that a protein's sequence was enough to recreate a protein's structure and function (C. B. Anfinsen 1973; C. B. Anfinsen et al. 1961, 1954), revealing their biological encoding in protein sequence. We are now seeing the ramifications of this fundamental discovery come to fruition in the form of accurate structure prediction from sequence alone (Jumper et al. 2021; Leman et al. 2022; Baek et al. 2021; Jiang et al. 2016; Radivojac et al. 2013). Anfinsen's discovery also played a monumental role in bridging the gap between protein biochemistry and genomics. Genetic variation within coding sequences now had a *mechanistic* basis for being "seen" by natural selection. Genetic variation that altered amino acid sequence could be understood in the context of how those changes affect the "native conformation" of a protein, its function, and, as a result, how those changes impact organismal adaptation (Christian B. Anfinsen 1959). Unsurprisingly, comparative genomics has discovered signals of protein structure and function embedded in the evolutionary history of proteins (Socolich et al. 2005; Halabi et al. 2009; Süel et al. 2003; Cong et al. 2019). The inclusion of sequence in the structure-function paradigm paved the way for high-throughput protein science in the age of genomics.

1.3 Approaches to map sequence-function relationships

Since Anfinsen's fundamental discoveries more than 60 years ago, two technological waves of experimental approaches have emerged that empower scientists to map protein sequence-function relationships. An important commonality between these waves was the development of methods for (1) generating genetic diversity; and (2) sequencing and measuring that diversity.

The first wave of technologies was the result of harmonizing site-directed mutagenesis,

polymerase chain reaction (PCR), and Sanger sequencing. These technologies, which were awarded Nobel prizes for their transformative nature, enabled scientists to introduce amino acid substitutions in proteins with surgical precision. Novel mutagenesis applications soon emerged, such as alanine scanning (Cunningham and Wells 1989), saturation mutagenesis (Zheng, Baumann, and Reymond 2004), and amber codon suppression (Noren et al. 1989; Cornish, Mendel, and Schultz 1995), which enabled scientists to probe protein sequence-function relationships, direct the evolution of novel or enhanced protein functions, and explore the site-specific effects of non-natural amino acid substitutions for the first time.

The second wave of technologies has been driven by next generation sequencing and an expansion of methods to generate mutational libraries. Advances in high-throughput DNA sequencing and synthesis have resulted in methods like deep mutational scanning (Fowler and Fields 2014), which measures the function of thousands of mutant proteins per experiment (Fowler et al. 2010; Faure et al. 2022). These methods allow the effects of missense variants to be inferred by quantifying a variant's frequency within a mutational library through experimental time (i.e. pre-, post-selections), much like tracking allele frequencies through evolutionary time. Several creative approaches to generate genetic diversity *in vitro* and *in vivo* have also emerged *(Wang et al. 2009; Hess et al. 2016; Ravikumar et al. 2018; Esvelt, Carlson, and Liu 2011; Hanna et al. 2021; Sharon et al. 2018; Cirino, Mayer, and Umeno 2003; Plesa et al. 2018)*, which, in-turn, have opened the door to new screening technologies that move beyond cell growth and fitness, such as the use of microscopy (Hasle et al. 2020), cell sorting (Matreyek et al. 2018), yeast two-hybrids (Starita et al. 2015), display technologies (Garrett et al. 2021).

Despite the huge success of these technologies in the last decade, many avenues still exist for improving mutagenesis and screening technologies. For one, these technologies typically come at the cost of proteome coverage (albeit, by design). Most workflows screen libraries composed

of a single protein or protein domain (Melamed et al. 2013; Starita et al. 2015; Findlay et al. 2018), decreasing overall throughput and limiting the types of generalizations that can be made to other protein functions or families. Second, many of the tried-and-true approaches infer the effect of missense variants by linking the variant to cell or organismal survival. While linking variant effects with fitness is relatively straightforward for proteins with essential functions, non-essential proteins and functions require sophisticated selections that are, at times, limited in their ease-of-portability across genes and experimental systems. Modular methods (i.e. methods that can be coupled with different proteins with minimal change in experimental design) have increased overall portability of these functional selections. For example, Variant Abundance by Massively Parallel-sequencing (VAMP-seq) (Matreyek et al. 2018) is compatible with any protein that remains soluble when fused with GFP at its N- or C-terminus. However, even for a modular method like VAMP-seq, certain proteins are more compatible to the required manipulations and modifications than others, and some deleterious variants may be masked by confounding effects arising from tagging a protein with GFP.

More recently, complementary mutagenesis approaches have emerged that characterize sequence variants across entire genomes (Després et al. 2020; Sharon et al. 2018; Hanna et al. 2021). While these methods improve scale, they are still limited to linking variant effects with cellular or organismal fitness. Additionally, some of the more recent technologies rely on promiscuous base editors, which can introduce a spectrum of substitutions at any given site. This promiscuity presents an additional challenge of inferring, with accuracy, the most likely genetic change at a locus that brought about the observed cellular-level phenotypic change.

1.4 Mapping sequence-function relationships with mass spectrometry (MS)

Mass spectrometry-based proteomics has the potential to complement many of the genomics-based workflows presented above by supporting both the scale of variants tested (from single proteins to multiple proteins) and the modularity of functional selections (classic biochemical assays with minimal change in experimental design needed from protein-to-protein) to map sequence-function relationships. First, mass spectrometers have advanced significantly in the last decade, with state-of-the-art instruments routinely measuring hundreds to thousands of proteins and protein variants (in the form of PTMs) within a sample in a single experiment (Schweppe et al. 2020; Humphrey, Azimifar, and Mann 2015; Leutert et al. 2019). There is every reason to assume current instruments are fast and sensitive enough to quantify the tens of thousands of variants contained within a mutational library, just like they do for PTMs.

Second, several mass spectrometry-based methods already probe protein biophysics, protein structure, and protein function at the scale of entire proteomes (Savitski et al. 2014; Feng, De Franceschi, Kahraman, Soste, Melnik, Boersema, De Laureto, et al. 2014; Andre Mateus, Savitski, and Piazza 2021). Unsurprisingly, these methods have recently been used to characterize the effects of post-translational modifications on protein structure and function in high-throughput (Smith et al. 2021; Potel et al. 2020; Zecha et al. 2018, 2022; J. X. Huang et al. 2019; Wu et al. 2020), which has illustrated the protein-specific and site-specific insights that can be gained from multi-dimensional biochemical assays.

With that being said, mass spectrometry-based proteomics does have several intrinsic constraints that make high-throughput variant characterization in *conventional* mutational

libraries challenging (i.e. by conventional, I mean a library of single or double mutants for a single protein). The predominant challenge is rooted in traditional “bottom-up” proteomic workflows, which rely on enzymatic digestion of protein samples followed by detection of peptides by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS).

While these methods are well-suited for samples with a diversity of proteins (i.e. an organismal proteome), there are two major bottlenecks when applying these methods to large collections of a mutant protein. First, in order to characterize a mutant protein, the peptide spanning the amino acid substitution needs to be detected. However, not all peptides are observable to the same degree. Some peptides are too short or too long; others ionize poorly or are of low intensity (due a protein or mutant protein being low abundance, for example). This detectability problem means that the coverage of a mutational library will be dependent on which stretches of amino acid sequence within a protein are observable, placing an upfront constraint on library size and design. Furthermore, peptide coverage is also protein dependent, making choice of protein an additional constraint when using conventional proteomic workflows.

Setting aside the challenge associated with peptide detection, the second major limitation with applying current proteomics methods to conventional mutational libraries is that, as the library size increases (e.g. from >100 protein variants to >1000), there is a significant *decrease* in the signal-to-noise ratio for the peptides of interest. The peptides containing any given amino acid substitution, which are found in only one variant within the library, rapidly become the *least* abundant peptides, with background peptides that are shared across *all* members in a library outcompeting the peptide of interest. Exciting alternative approaches, such as “peptide barcoding” (Egloff et al. 2019; Matsuzaki et al. 2021), have emerged as a solution to this challenge. However, these approaches assume barcodes do not alter the underlying protein phenotype being tested, which, from our own observations, is an assumption that still needs

thorough testing and will likely vary from protein-to-protein. Additionally, these approaches are not scalable to the entire proteome. Thus, additional technologies are needed that can map sequence-function relationships at the scale of proteomes.

1.5 Organization of this thesis

In this dissertation, I describe the development of a novel mass spectrometry-based method to map sequence-function relationships for entire proteomes. This technology was invented by Dr. Ricard Rodriguez-Mias, and some of the experiments I showcase in Chapter 2 represent my efforts to help establish the technology in yeast, which was done in collaboration with Ricard and several members in the Villén, Fields, and Noble labs. These early experiments were integral to the applications of this technology I present in Chapter 3 and Chapter 4. Lastly, in Chapter 5, I describe an extension of some of the methods presented here to identify endogenous substrates of the SARS-CoV-2 protease NSP5, which was done in collaboration with Dr. Ian Smith and Dr. Mario Leutert. I finish in Chapter 6 with a broad overview of the state of the field and exciting technological developments on the horizon.

1.5.1 A proteomics platform to characterize substitutions proteome-wide

The effects of amino acid substitutions on protein functions can be explored at depth for single proteins, but can we expand these approaches to probe substitutions across many proteins simultaneously? In Chapter 2, I described a method that leverages proteome-wide mistranslation and biochemical selections with a mass spectrometry-based readout to characterize the effects of amino acid substitutions across entire proteomes.

1.5.2 Towards an atlas of substitution sensitivity in yeast

With the method from Chapter 2 in hand, the effects of amino acid substitutions across entire proteomes can be probed in high-throughput. However, what kinds of information can be gleaned from integrating substitution sensitivity across different mistranslated proteomes? In Chapter 3, I discuss two developments in this area. First, I describe a high-throughput thermal stability assay compatible with mistranslated proteomes that can determine the site-specific effects of amino acid substitutions on protein thermal stability. This assay was a key technological development that enabled scaling these functional selections to many mistranslated proteomes, replicates, and conditions. Second, I apply this method to eight mistranslated proteomes and identify patterns of residue sensitivity that uncover sequence-function relationships across the yeast proteome.

1.5.3 Unmasking residue functions using secondary selections

The majority of amino acid substitutions I quantified in Chapter 3 did not significantly alter protein thermal stability. Are most residues dispensable to structure and function? In Chapter 4, I present data that suggest residue sensitivity to ncAA substitutions depends on molecular contexts, and that adding additional selections helps unmask these sensitivities. I describe experiments in Chapter 4 that illustrate how these methods can be coupled with small molecules to identify small molecule binding sites, highlighting future opportunities to expand the scope of secondary selections to reveal residue sensitivity.

1.5.4 A novel approach to identify protease substrates proteome-wide

Lastly, I showcase an application of these stability assays in combination with protein turnover to identify substrates of proteases proteome-wide. This project was done in collaboration with

fellow lab members and highlights some of our lab's efforts to use proteomic technologies to better understand SARS-CoV-2 biology.

Chapter 2. Mapping sequence-function relationships across proteomes

This Chapter is based on the following preprint:

Proteome-wide identification of amino acid substitutions deleterious for protein function

Ricard A. Rodriguez-Mias[#,1,*], Kyle N. Hess[#,1,2], Bianca Y. Ruiz[1], Ian R. Smith[1], Anthony S. Barente[1], Stephanie M. Zimmerman[1], Yang Y. Lu[1], William S. Noble[1], Stanley Fields[1], Judit Villén[1*]

[1]Department of Genome Sciences, University of Washington

[2]Graduate Program in Molecular and Cellular Biology, University of Washington

[#]equal author contribution

*corresponding authors

DOI: https://doi.org/10.1101/2022.04.06.487405

Author Contributions: This project was a joint collaborative effort between Ricard Rodriguez-Mias (RRM) and Kyle Hess (KH). RRM invented Miro, which has been patented with co-inventors Judit Villén and Stanley Fields. RRM made Figure 2.1. RRM generated, collected, and analyzed the data in Figure 2.3. KH generated, collected, and analyzed the data presented in Figure 2.4. KH generated and collected data for Figure 2.2, which was analyzed by both KH and RRM. RRM and KH contributed equally to manuscript writing and figure generation.

2.1 Summary

DNA sequencing has led to the discovery of millions of mutations that change the encoded protein sequences, but the impact of nearly all of these mutations on protein function is unknown. We addressed this scarcity of functional data by developing Miro, a proteomic technology that uses mistranslation to introduce amino acid substitutions and biochemical assays to quantify functional differences of thousands of protein variants by mass spectrometry. We apply this technology to the proteome of yeast to reveal amino acid substitutions that impact protein structure, ligand binding, protein-protein interactions, protein post-translational modifications, and protein thermal stability. Adapting Miro to human cells will provide a means to efficiently accelerate our mechanistic interpretation of genomic mutations to predict disease risk.

2.2 Introduction

Changes to protein sequences as a consequence of mutations in DNA can alter protein structure and function, impact cellular and organismal physiology, and result in disease (Shendure and Akey 2015). Thus, deciphering which amino acid substitutions have functional consequences is a major focus of studies in biology, protein engineering, and medical genetics. The predominant approach to determine the functional effect of a change to a protein sequence has been mutagenesis followed by measurement of the resulting mutant protein's function. However, developments in genomics have accelerated the discovery of mutations such that tens of millions of missense variants in the human genome need functional interpretation (Karczewski et al. 2020; Lek et al. 2016; 1000 Genomes Project Consortium et al. 2015). Traditional mutagenesis approaches are severely underpowered for this task, and even recent high-throughput implementations, such as deep mutational scanning (Starita et al. 2017; Fowler et al. 2010), assess the effect of mutations in only one protein per experiment. Considering the ~20,000 human genes and the vast proteome encoded by these genes, such mutational

experiments would require an inordinate amount of time to characterize the missense variants discovered to date.

To overcome this major bottleneck in genome analysis, we developed Miro, an approach to introduce amino acid substitutions into proteins and assess their consequences on protein function at a proteome-wide scale. First, Miro leverages the permissive nature of aminoacyl tRNA synthetases, which can incorporate closely related non-canonical amino acids (ncAAs) during protein translation in living systems ranging from bacteria to human cells (Budisa 2006; Richmond 1962; Cowie et al. 1959). Cells grown in the presence of ncAAs generate statistical proteomes composed of thousands of protein quasispecies, each defined as a collection of protein variants sharing most of their sequence, with the exception of the randomly introduced ncAAs. Although similar to their cognate counterparts, ncAAs can introduce significant changes to hydrophobicity, pKa, and secondary structure and thereby serve as a useful tool to probe for positional sensitivity in protein sequences. Second, Miro applies a selection to statistical proteomes to classify protein variants according to some biochemical property, such as the ability to fold correctly into a soluble form or to interact with another protein. Third, Miro uses mass spectrometry (MS) to quantify protein variants and assess differences in the biochemical property for proteins with the ncAA vs. the native amino acid at each position, thereby determining the functional tolerance of the substitution. Because of the global nature of ncAA substitutions, the many biochemical properties that can be assessed, and the capacity of mass spectrometry to quantify peptides, Miro has the potential to identify amino acid substitutions that are likely to be deleterious at an unprecedented scale.

2.3 Results

2.3.1 Development of the Miro technology

The overall scheme for Miro is shown in Figure 2.1 for a representative ncAA and biochemical assay. A proline tRNA can be mischarged with a proline analogue, resulting in the random incorporation of the analogue at proline codons and the generation of a statistical proteome (Figure 2.1a). The statistical proteome is then subjected to a biochemical assay, e.g. a protein affinity purification that separates mistranslated variants of a protein based on their ability to bind to a tagged protein (Figure 2.1b). Following digestion of the proteins into peptides, mass spectrometry quantifies the abundance of each peptide with a ncAA substitution relative to its native version, before and after affinity purification (Figure 2.1c). Depletion of a peptide with a ncAA substitution in the purified fraction indicates that the substitution impaired the protein-protein interaction, while enrichment of a substituted peptide indicates that the substitution enhanced the interaction.

Miro does not rely on modifications to the DNA or to the cellular translation machinery and thus should be generalizable to *in vitro* translation systems and to many organisms and cell types, including bacteria, yeast, and mammalian cells. Because the biochemical assays are carried out on protein lysates, the toxicity of ncAAs towards the cells or translation system should have minimal consequences on the ability to interpret the functional effects of substitutions.

2.3.2 Incorporation of ncAAs and properties of peptide containing ncAAs

To determine the feasibility of Miro, we first sought to identify ncAAs that both incorporate in the proteome and slow down cell growth, an indication that the ncAA may alter protein function. Additionally, ncAAs with fractional incorporation are ideal, so that individual protein variants contain a small number of substitutions. We screened a panel of 26 ncAAs (Supplementary Table 1) for toxicity and incorporation in *S. cerevisiae* by growing cultures in the presence of a ncAA at one of eight different concentrations and assessing incorporation by mass spectrometry for one or two of these concentrations (Methods). In total, 13 of these ncAAs inhibited yeast

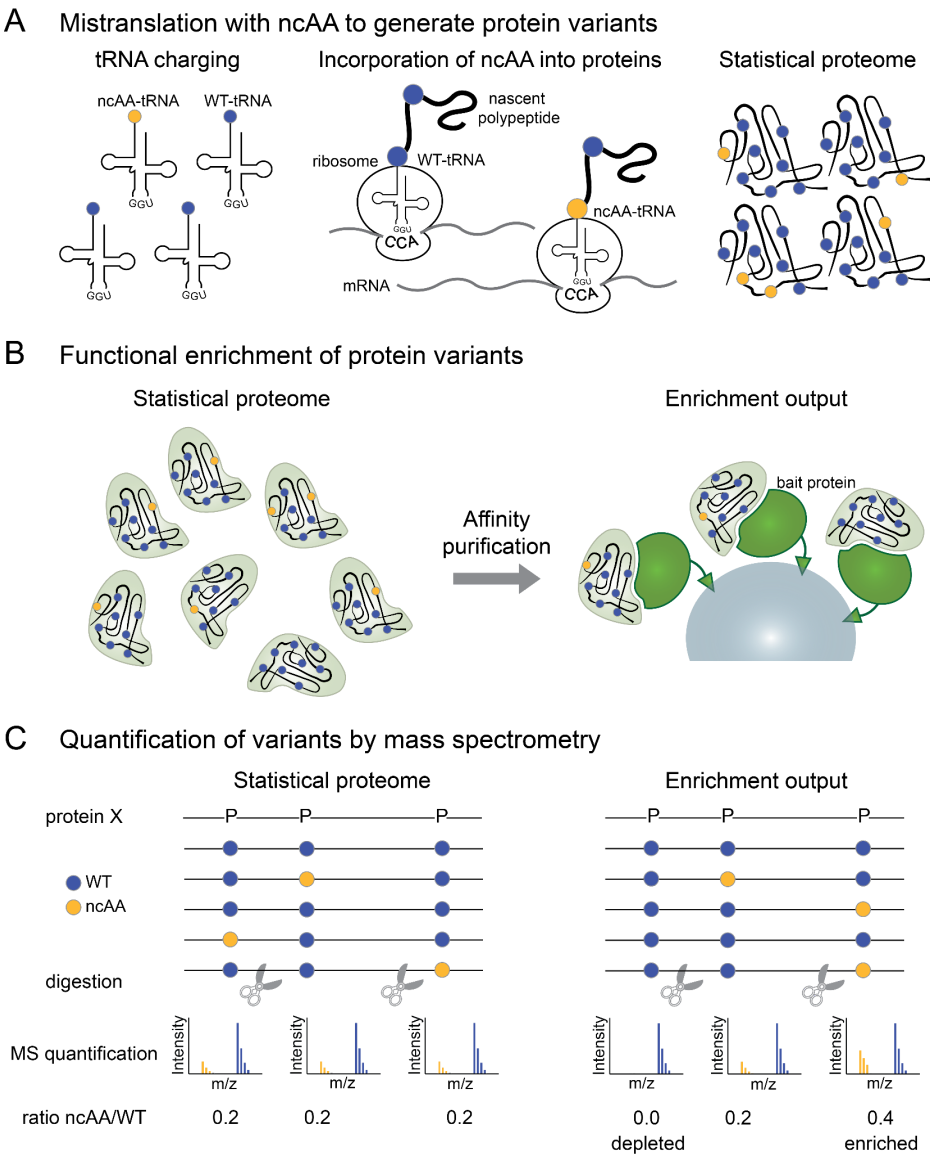


Figure 2.1. Overview of the Miro method.

A) Mistranslation with non-canonical amino acids (ncAAs) to generate protein variants. Non-canonical amino acids are recognized by tRNA synthetases, charged onto their cognate tRNAs, and incorporated randomly into proteins at cognate positions, alongside wild type amino acids, to produce a statistical proteome. B) Functional enrichment of protein variants. The statistical proteome is subjected to a biochemical selection assay to assess the impact of ncAA substitutions on protein function. One example is an affinity purification assay in which protein variants with impaired binding to their partner protein will be depleted from the enrichment output. C) Quantification of variants by mass spectrometry. Mass spectrometry is used to determine the relative abundance of ncAA-containing peptides compared to their wild type cognate versions to determine the incorporation of ncAA at a given position. A comparison of ncAA incorporation before and after the functional enrichment can identify amino acid positions that are functionally important (e.g., in this case for protein-protein interaction). Blue circles represent wild type amino acid proline and yellow circles represent a ncAA analog of proline.

growth by at least 25% (Figure 2.2a, Supplementary Figure 2.1). Assessing incorporation, 12 ncAAs showed greater than 2% incorporation in the yeast proteome (Figure 2.2b). Aromatic amino acid analogues with single fluorine substitutions had the highest incorporation, likely due to the conservative nature of the chemical change. We also found several analogues that were toxic without appreciable incorporation, suggesting toxicity mechanisms unrelated to mistranslation. Overall, this screen yielded a set of eight analogues that incorporate at one of seven cognate amino acid positions and are toxic to cells. This set of ncAAs is usable for functional selections in the Miro protocol.

Next, we assessed the measurability of peptides containing a ncAA substitution by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) by comparing their elution profile and MS/MS fragmentation spectra to those for their matching wild type sequences in LC-MS/MS runs of yeast mistranslated proteomes. Most ncAAs had minimal effects on peptide chromatographic retention time (Figure 2.2c), and the observed effects were generally in line with the chemical modification introduced by the ncAA. For example, replacement of hydrogen atoms with fluorine tended to increase peptide hydrophobicity and lead to delayed elution (Figure 2.2c analogues F2, W1, Y1), while the addition of hydroxyl groups had the opposite effect (Figure 2.2c analogue W2). Furthermore, replacing proline with the 4-membered ring analogue azetidine-2-carboxylic acid (azetidine) advanced elution, due to both decreased hydrophobicity as a result of ring contraction and to a higher propensity towards the cis configuration of the peptide bond (Figure 2.2c analogue P1, Supplementary Figure 2.2a-b) (Kern, Schutkowski, and Drakenberg 1997; Tsai, Overberger, and Zand 1990). The retention time shift of this substitution was dependent on peptide length, relative sequence position, and neighboring amino acid (Supplementary Figure 2.2c-e).

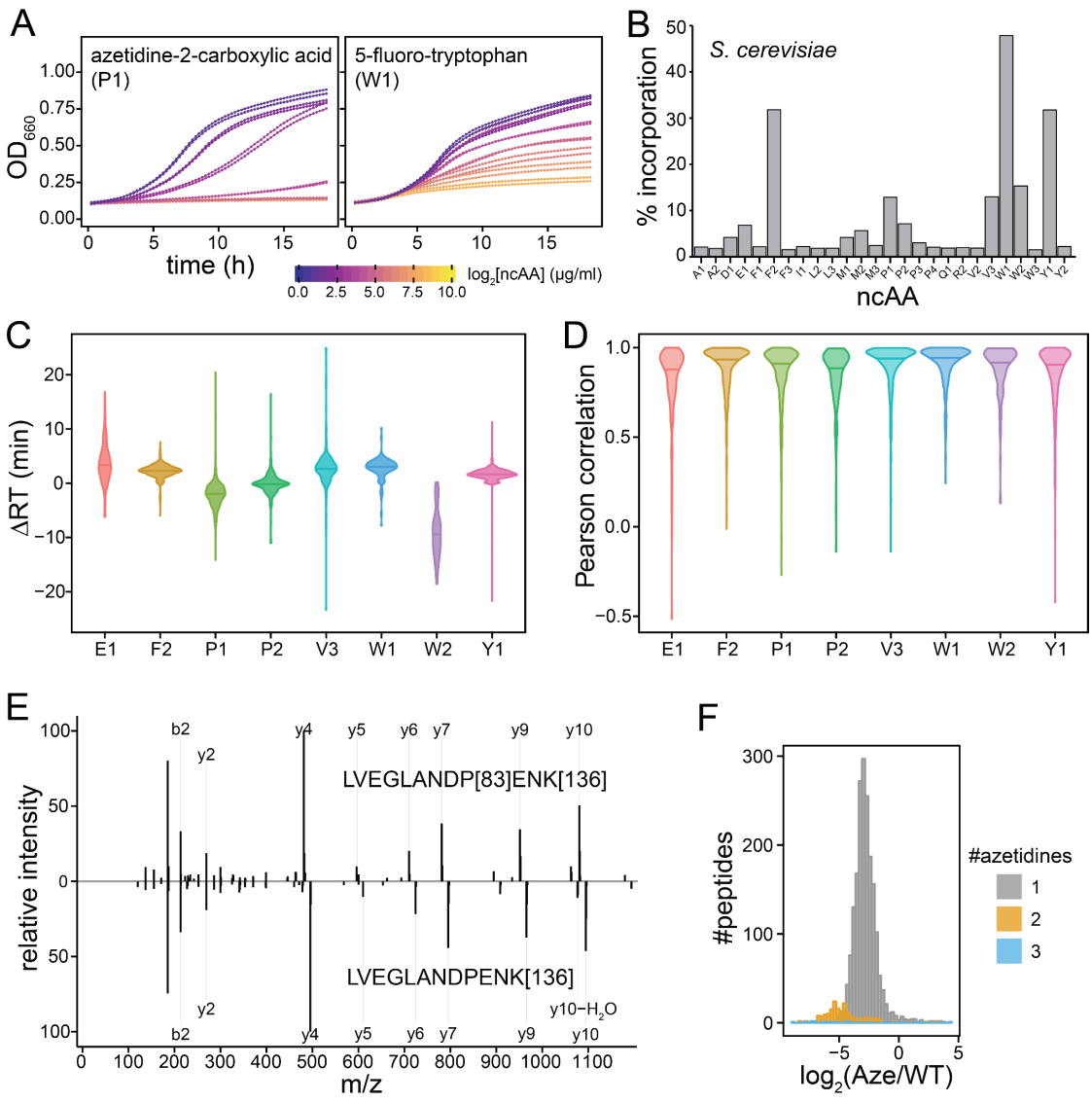


Figure 2.2. Toxicity and incorporation of ncAA and physicochemical properties of misincorporated peptides.

A) S. cerevisiae growth curves in the presence of increasing amounts of the ncAAs azetidine (P1) and 5-fluoro-tryptophan (W1) to assess toxicity of non-canonical amino acids (full data in Supplementary Figure 2.1). B) Incorporation of ncAA on the S. cerevisiae proteome. C) Violin plot showing retention time differences between peptides containing ncAAs E1, F2, P1, P2, V3, or W1 and their wild type counterparts. D) Violin plot showing Pearson correlations between MS/MS spectra for peptide precursors containing ncAAs E1, F2, P1, P2, V3, or W1 and their wild type counterparts. E) Mirror MS/MS plot for a charge 2+ precursor of peptide LVEGLANDPENK from Rps12. The spectrum at the top corresponds to the peptide containing azetidine-2-carboxylic acid in place of proline and the bottom spectrum corresponds to the wild type sequence. F) Distribution of P1 incorporation ratios at proline sites on the S. cerevisiae proteome. Each data point corresponds to a measured peptide and is colored according to the number of azetidine residues.

Peptides containing ncAAs can be identified and the precise position of the substitution localized within the peptide sequence using a similar analysis strategy to that for identifying post-translational modifications, i.e. by allowing a defined mass shift from the protein sequence database on the precursor and some of the fragment ions. This strategy allows us to assess ncAA substitutions at a positional resolution of a single amino acid. MS/MS spectra for wild type peptides and peptides containing a single ncAA substitution were similar, as shown by the high spectra correlations (Figure 2.2d) and direct spectra comparison for a representative proline-containing peptide (Figure 2.2e). These similarities indicate that peptide pairs can be reliably quantified and compared at both the full MS and the MS/MS levels.

Finally, in order to accurately quantify ncAA incorporation, we implemented a pulsed-labeling strategy in yeast whereby isotopically-heavy lysine was added to growing cultures at the same time as the ncAA of interest. This strategy mitigates experimental variability due to the chronic toxicity of the ncAA and improves estimates of incorporation ratios across the proteome, specifically by labeling proteins that were synthesized after yeast were exposed to the ncAA. To test this approach, we quantified azetidine incorporation in the yeast proteome. Azetidine incorporation was quantified for 2,112 unique peptides containing one or more proline residues across 789 proteins. Observed incorporation ratios were relatively uniform, with a median incorporation of 10.5% and interquartile range between 5.9 and 18.4%, suggesting that incorporation in yeast is stochastic in nature and, as expected, that the lowest incorporation corresponds to peptides with multiple azetidine substitutions (Figure 2.2f).

Together, we show that eight ncAAs, when individually added to growth media, were toxic and incorporated stochastically into the yeast proteome by mistranslation. Generally, peptide pairs with and without the ncAA showed similar elution and fragmentation in LC-MS/MS analysis. These results demonstrate the feasibility of the mistranslation approach with ncAAs to assess

the functional impact of amino acid substitutions at the scale of the whole proteome. Additionally, because mass spectrometry measurements on Miro are carried out at the peptide level, the impact of substitutions can be assessed at the site level, with the peptides containing a ncAA collectively reporting for all the protein variants containing the ncAA at that site, whereas its wild type peptide counterpart reports for all the variants that do not have the ncAA at that site.

2.3.3 Effects of substitutions on protein function

We next decided to explore the impact on aspects of protein biology of azetidine mistranslation, given its reported proteotoxic effects (Fowden and Richmond 1963), efficient incorporation into the yeast proteome (Figure 2.2b), and ability to alter protein secondary and tertiary structure (Tsai, Overberger, and Zand 1990; Baum et al. 1975). We applied Miro with azetidine to four functional readouts: i) the ability of an *in vitro* translated protein to fold and interact; ii) the ability of yeast ribosomal proteins to assemble into the ribosome; iii) the post-translational modification of yeast proteins; and iv) the thermal stability of the yeast proteome.

2.3.3.1 Effects of azetidine substitutions on protein structure and interactions

For our initial assay, we applied azetidine to a mammalian *in vitro* translation system synthesizing a protein construct composed of three binding modules: a glutathione S-transferase (GST) domain, a human influenza hemagglutinin (HA) tag, and a polyhistidine (His) tag (Figure 2.3a). We reasoned that this model system would allow us to maximize proline site coverage and assess overall protein folding and interactions in an affinity purification assay, while largely eliminating potential bias due to degradation or toxicity.

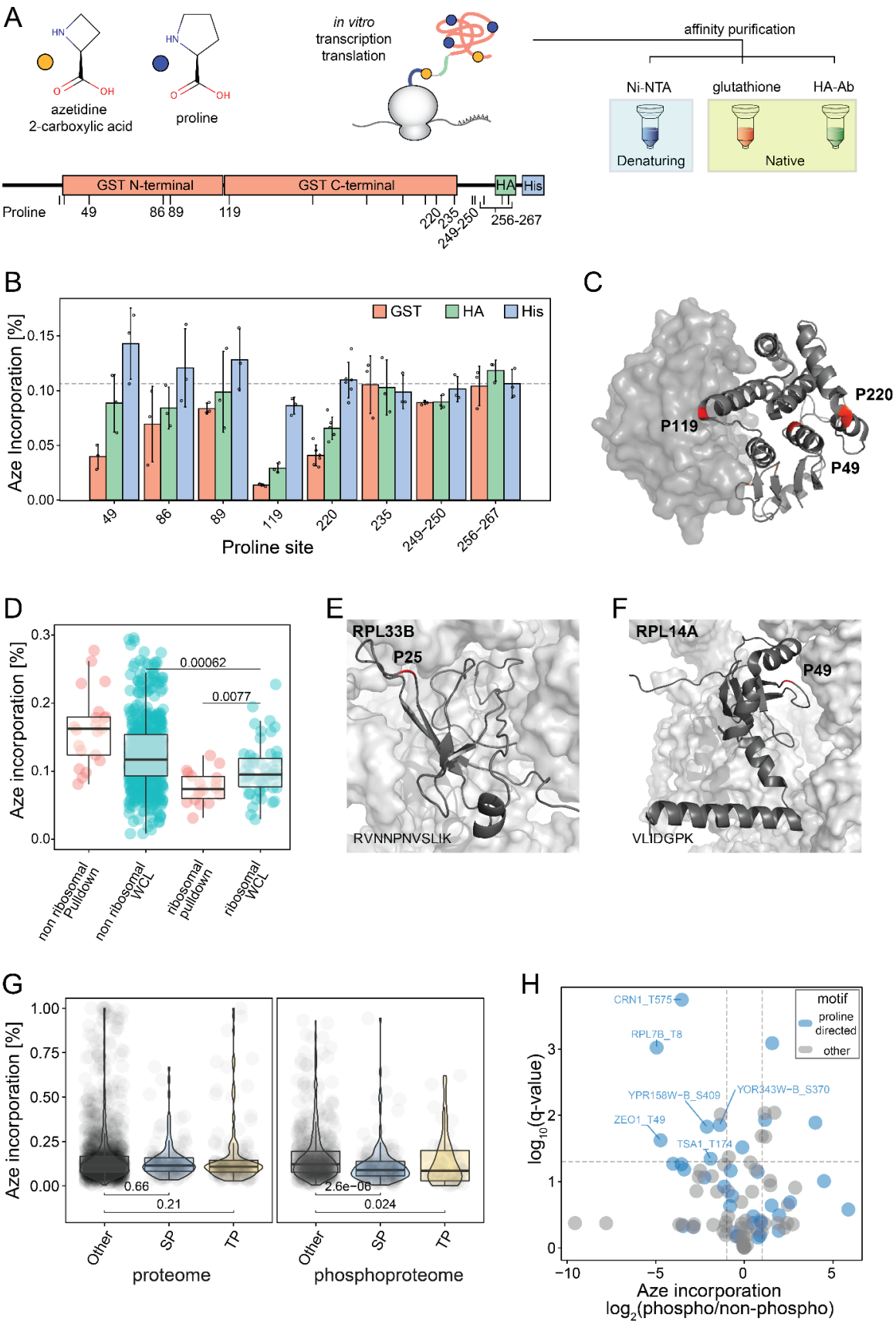

A
azetidine
2-carboxylic acid
proline
in vitro
transcription
translation
affinity purification
Ni-NTA
glutathione
HA-Ab
Denaturing
Native
GST N-terminal
GST C-terminal
HA
His
Proline
49
86 89
119
220
235
249-250
256-267
B
GST
HA
His
Aze Incorporation [%]
0.15
0.10
0.05
0.00
49
86
89
119
220
235
249-250
256-267
Proline site
C
P220
P119
P49
D
Aze incorporation [%]
0.3
0.2
0.1
0.0
0.00062
0.0077
non ribosomal Pulldown
non ribosomal WCL
ribosomal pulldown
ribosomal WCL
E
RPL33B
P25
RVNNPNVSLIK
F
RPL14A
P49
VLIDGPK
G
Aze incorporation [%]
1.00
0.75
0.50
0.25
0.00
0.66
0.21
2.6e−06
0.024
Other
SP
TP
proteome
phosphoproteome
H
motif
proline directed
other
CRN1_T575
RPL7B_T8
YPR158W−B_S409
YOR343W−B_S370
ZEO1_T49
TSA1_T174
log10(q-value)
3
2
1
0
−10
−5
0
5
Aze incorporation
log2(phospho/non-phospho)

Figure 2.3. Functional impact of azetidine-2-carboxylic acid incorporation.

A) Orthogonal affinity purifications of an HA- and His-tagged glutathione-S-transferase construct that has been mistranslated with azetidine (P1) in a cell-free mammalian in vitro transcription and translation system. B) Relative incorporation of azetidine at proline sites for the three purifications indicated in panel A. C) Surface and cartoon representation of a GST dimer (PDB:1Y6E) showing proline sites with the highest depletion of azetidine in red. D) Boxplot showing incorporation of azetidine at proline sites for the whole yeast cell lysate (WCL) and for the Rpl16B pulldown eluate (PD). Median incorporation rates across replicates per site are shown, and peptides are grouped by fraction and protein of origin (ribosomal or non-ribosomal). The total number of quantified sites per group was: non-ribosomal_PD n=29, non-ribosomal_WCL n=563, ribosomal_PD n=20, ribosomal_WCL n=55 and p-values for group comparisons are computed using a two-sided Wilcoxon test. E, F) Structural context of two proline sites where azetidine incorporation appears as significantly depleted upon ribosome purification: Rpl33B Pro25 (E) and Rpl14A Pro49 (F). The ribosome is shown as surface representation, ribosomal proteins of interest as cartoons (PDB: 4V6I), and relevant proline sites in red. G) Boxplot showing azetidine incorporation into proline positions for the yeast proteome and phosphoproteome. Median azetidine incorporation across replicates is displayed, and peptides are grouped according to the presence of SP, TP or other sequence motifs. Total number of sites quantified for proteome are: (SP motif: n=118, TP motif: n=134, Other: n =1426) and phosphoproteome: (SP motif: n=141, TP motif: n=41, Other: n=488) respectively; and p-values for group comparisons are computed using a two-sided Wilcoxon test. H) Volcano plot displaying azetidine incorporation at proline sites in phosphopeptides relative to their non-phosphorylated counterparts. A t-test was applied to azetidine incorporation rates across replicates for phosphorylated peptides and their non-phosphorylated counterparts (n=40), and p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method.

In vitro translation was carried out in the presence of an azetidine excess so as to compete with the proline present in the reaction mixture. We carried out three distinct affinity purifications: a denaturing purification against the histidine tag to assess azetidine incorporation at proline positions throughout the protein, and native affinity purifications against GST and against the HA tag to assess the impact of proline-to-azetidine replacement on the ability of the protein to fold into a native structure (both native purifications) and to interact with its ligand glutathione (only the GST affinity purification) (Figure 2.3a). Using targeted mass spectrometry, we obtained quantitative information for 11 of the 17 proline sites in the sequence, and calculated intensity ratios between the azetidine-containing peptides and their corresponding wild type forms across all purifications (Figure 2.3b).

Azetidine incorporation ratios observed in the denaturing purification were uniform across proline positions at ~10% (Figure 2.3b, Supplementary Dataset 1), validating that ncAA misincorporation in an *in vitro* translation system is largely unbiased. Native purifications, on the other hand, featured several proline positions that were significantly depleted of azetidine, particularly positions located within the GST domains (Figure 2.3b). This depletion is indicative of azetidine-induced misfolding, which can lead to insolubility prior to an affinity purification. Even though the two native purifications shared similar trends, azetidine depletion generally seemed more pronounced in the GST affinity purification (Figure 2.3b), with a few site-level differences where azetidine depletion was even more pronounced in the GST purification (Figure 2.3b).

Thus, the substantial depletion of azetidine at proline 119 in our construct for both purifications is likely the result of its location at the start of the α-helix 4 in the GST C-terminal subdomain. The α-helices 4 and 5 participate in a hydrophobic lock and key arrangement with the other GST monomer (Sayed, Wallace, and Dirr 2000), driving the dimerization and crucial to the overall fold

(Figure 2.3c). Conversely, the more extreme depletion of azetidine at proline 49 in the GST purification is likely due to the fact this residue plays an important role in scaffolding the glutathione-binding interface (Figure 2.3c). Indeed, proline 49 in our construct is located at the N-terminus of α-helix 1 opposite to β-sheet 1 and the conserved catalytic tyrosine residue and mutations at these sites can severely affect glutathione binding (Manoharan et al. 1992). Prolines 86 and 89 are also located in the glutathione binding interface; they were, however, not very sensitive to azetidine substitution. Interestingly, the least affected position was proline 89, perhaps due to the fact that its peptide bond adopts a cis conformation.

2.3.3.3 Effects of azetidine substitutions on protein interactions

We next asked whether Miro could identify amino acid positions that are important for the assembly of proteins into a large complex, the ribosome. The yeast cytosolic ribosome is composed of 79 different proteins that assemble with four ribosomal RNAs into a large and a small subunit. We grew *S. cerevisiae* in the presence of azetidine to obtain 5-10% incorporation across the proteome, lysed the cells in native conditions, and affinity-purified the ribosome using an epitope-tagged version of the Rpl16B protein. We analyzed protein mixtures in the whole cell lysate and the ribosome-purified samples to quantify azetidine at each proline position.

We identified over 1200 proteins across two or more replicates. As expected, the Rpl16B pulldown eluate was highly enriched in ribosomal proteins, which contributed almost half of the signal for these samples. This eluate also captured a large fraction of ribosome-associated proteins, as evidenced by Gene Ontology enrichment analysis (Supplementary Figure 2.3).

In the whole cell lysate sample, we observed lower azetidine incorporation in ribosomal proteins compared to non-ribosomal proteins (p-value=0.0024) (Figure 2.3d). Azetidine in ribosomal

proteins was further reduced in the purified ribosome fraction (p-value=0.035) (Figure 2.3d), suggesting that in most cases azetidine disrupts ribosomal protein interactions. We identified several proline sites on proteins of the 60S ribosomal subunit with a two-fold depletion of azetidine after pulldown: Rpl33B Pro25 (BH-adjusted p-value=0.014), Rpl14A Pro49 (BH-adjusted p-value=0.021) and Rpl4A P329 (BH-adjusted p-value=0.014). Rpl33B Pro25 is located on the loop connecting strands β1 and β2, which extensively contacts the 25S rRNA (Figure 2.3e). Similarly, Rpl14A Pro49 is located in the turn that connects the β3 and β4 strands, which interact with both 25S rRNA and Rpl20A (Figure 2.3f). Finally, Rpl4A P329 mediates a helix-turn-helix motif at the C-terminal extension region, which has been shown to facilitate ribosomal recruitment (Stelter et al. 2015).

2.3.3.4 Effects of azetidine substitutions on protein post-translational modifications

Amino acid substitutions in a substrate protein can affect the ability of modifying enzymes to recognize and catalyze a post-translational modification. We sought to determine whether Miro could be used to identify positions in which a ncAA substitution altered the phosphorylation levels of the protein. In addition to its prominent role in protein structures by forming β-turns, proline is an important residue for recognition by kinases in the CMGC group. These kinases, which include cyclin-dependent kinases (CDKs) and mitogen-activated protein kinases (MAPKs), require a proline at the +1 position of the phospho-acceptor serine or threonine. Therefore, we reasoned that proline substitutions may impact the phosphorylation of SP and TP short linear motifs. To test the effect of such substitutions, we grew *S. cerevisiae* in the presence of azetidine to obtain 5-10% incorporation at proline positions across the proteome.

We measured the content of azetidine at proline sites for whole cell lysate peptides and immobilized metal affinity chromatography (IMAC)-enriched phosphopeptides. Globally, we found that phosphopeptides with a Pro at position +1 to Ser or Thr showed significant depletion of azetidine (Figure 2.3g). Prolines in other sequence positions relative to a phosphorylated amino acid were not affected by azetidine substitution (Figure 2.3g).

For 91 proline-containing peptides, we were able to quantify azetidine incorporation in both their phosphorylated and non-phosphorylated forms (Figure 2.3h) across multiple replicates. Similar to the global trends, the most significantly depleted azetidine incorporations occurred at phosphorylation sites on [S|T]P proline-directed motifs. These instances map to unstructured and readily accessible protein regions, which suggests that azetidine incorporation alters the cis/trans peptide bond conformation balance and likely affects the recognition by kinases or phosphatases (X. Z. Zhou et al. 2000).

2.3.3.5 Effects of azetidine substitutions on protein stability

Lastly, we assessed whether Miro could measure the site-specific effects of ncAA substitutions across an entire proteome. As a proof of concept, we measured the impact of proline-to-azetidine substitutions on protein thermal stability in the *S. cerevisiae* proteome. To do this, we generated a statistical proteome with 5-10% azetidine incorporation in proline positions and implemented thermal proteome profiling (TPP) (Savitski et al. 2014), a proteomics assay that measures protein thermal stability of thousands of proteins by mass spectrometry. We adapted TPP to produce and compare site-specific thermal denaturation curves for protein variants with and without ncAA substitutions (Figure 2.4a).

Across two replicates, we identified 1,506 azetidine-containing peptides mapping to 700 yeast proteins, with 660 of these peptides detected and quantified in duplicate in 276 yeast proteins (Supplementary Figure 2.6a). After stringent filtering (see Methods), we measured changes in stability for 465 unique azetidine substitutions across 203 proteins (Figure 2.4b). On average, replacing proline with azetidine had a mild destabilizing effect on protein thermal stability (Figure 2.4c, mean ΔAUC = -0.5 ± 2.9, p-value = 1e-4, two-sided t-test). In total, 43 azetidine substitutions significantly altered melting behavior (permuted non-parametric analysis of response curves (Childs et al. 2019), Benjamini-Hochberg corrected p-value < 0.01), most of which were destabilizing (30 destabilizing *vs.* 13 stabilizing) (Figure 2.4c).

We next analyzed several site-specific structural and evolutionary features for each proline residue where we detected an azetidine substitution in order to understand why azetidine incorporation was destabilizing (D), stabilizing (S), or non-significant (NS). We found that proline residues across all major structural elements (coils, helices, strands, turns) and disordered regions were sensitive to azetidine, although substitutions in β-sheets were generally more destabilizing compared to coiled regions and turns (Supplementary Figure 2.4c). We also found that destabilizing substitutions occurred at proline residues with lower solvent accessible surface area (Figure 2.4d, Wilcoxon rank-sum test, p-value = 0.023), and these same sites were predicted to be more sensitive to natural amino acid substitutions (predictions of ΔΔG from mutfunc database (Wagih et al. 2018); Wilcoxon rank-sum test, p-value = 0.0038, Figure 2.4e), suggesting that these proline residues may be intolerant to most amino acid changes. Lastly, an unexpected finding was that azetidine sensitivity was not correlated with the evolutionary conservation of proline residues (Supplementary Figure 2.4c). Taken together, these data suggest that the effect of azetidine may depend less on broad structural features shared across the proteome, and more on local structural and functional contexts specific to each protein.

Next, we asked whether we could identify functional amino acid residues by mapping the effects of azetidine substitutions back to protein structure. We focused on the essential glycolytic enzyme phosphoglycerate kinase-1 (Pgk1), due to its high coverage in our dataset (9 of the 17 possible proline-to-azetidine substitutions) and range of azetidine effects (mean ΔAUC = -1.64 ± 1.40; Figure 2.4f). Most substitutions had an effect on Pgk1's stability (6 out of 9 azetidine substitutions significantly destabilizing Pgk1; Figure 2.4f). Mapping these positions back to Pgk1's structure reveals slight differences among Pgk1's domains. Three of the most destabilizing substitutions (azetidine incorporation at Pro45, Pro79, or Pro407) fall in and around the 3-phosphoglycerate-binding domain (Figure 2.4g). Conversely, the nucleotide-binding domain contains a mix of destabilizing and neutral substitutions, with two destabilizing substitutions falling directly in Pgk1's nucleotide-binding pocket (Pro337, Pro338) (Figure 2.4h). Pro205, which has been functionally implicated in Pgk1's catalytic cycle(McHarg et al. 1999), was unexpectedly tolerant to azetidine (Figure 2.4i). Interestingly, this proline adopts a cis conformation in several Pgk1 crystal structures from different species (Figure 2.4i). Additionally, replacing Pro205 with histidine or phenylalanine, which locks the trans-conformation, has been shown to decrease Pgk1 stability (McHarg et al. 1999). This result suggests that Pro205 may be tolerant to proline-like analogues that maintain the cis conformation, such as azetidine, and more generally that other cis-conforming proline residues across the proteome may tolerate these types of substitutions (Supplementary Figure 2.4c).

We also show that Miro can help pinpoint structural elements or regions within proteins that play an important role in protein stability. For example, we found a cluster of destabilizing proline-to-azetidine substitutions within the glycolytic enzyme phosphoglycerate mutase-1 (Gpm1) that all fell within a stretch of four consecutive proline residues (Supplementary Figure 2.5a). These four residues are all in trans conformation, forming a polyproline-II helix (Supplementary Figure 2.5a). Azetidine incorporation into polyproline peptides has been shown

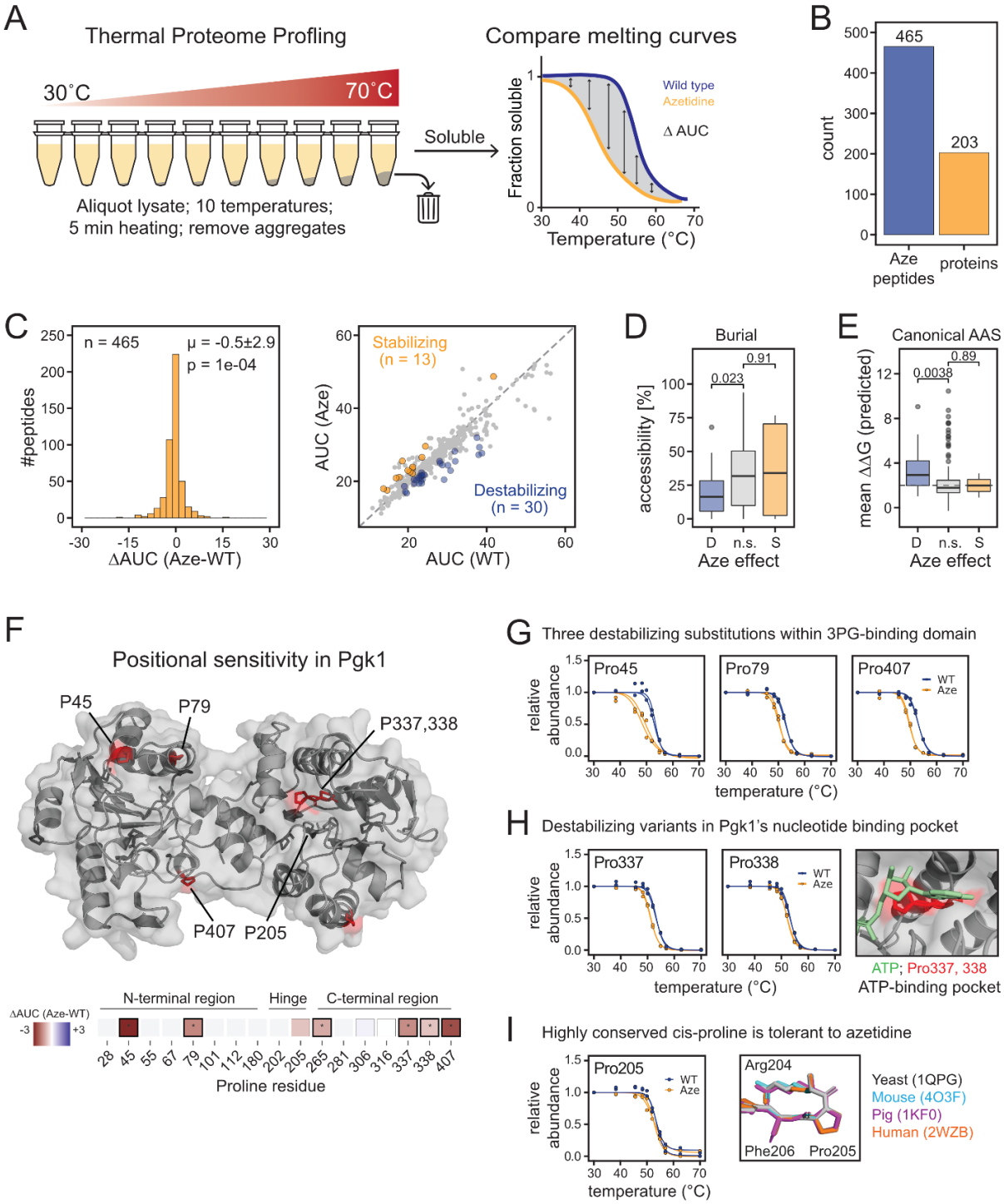

A
Thermal Proteome Profling
30˚C
70˚C
Soluble
Aliquot lysate; 10 temperatures;
5 min heating; remove aggregates
Compare melting curves
Fraction soluble
Temperature (°C)
Wild type
Azetidine
Δ AUC
B
count
465
203
Aze peptides
proteins
C
n = 465
μ = -0.5±2.9
p = 1e-04
#peptides
ΔAUC (Aze-WT)
Stabilizing (n = 13)
Destabilizing (n = 30)
AUC (Aze)
AUC (WT)
D
Burial
accessibility [%]
0.023
0.91
D
n.s.
S
Aze effect
E
Canonical AAS
mean ΔΔG (predicted)
0.0038
0.89
Aze effect
F
Positional sensitivity in Pgk1
P45
P79
P337,338
P407
P205
ΔAUC (Aze-WT)
-3
+3
N-terminal region
Hinge
C-terminal region
28 45 55 67 79 101 112 180 202 205 265 281 306 316 337 338 407
Proline residue
G
Three destabilizing substitutions within 3PG-binding domain
relative abundance
Pro45
Pro79
Pro407
WT
Aze
temperature (°C)
H
Destabilizing variants in Pgk1's nucleotide binding pocket
Pro337
Pro338
ATP; Pro337, 338
ATP-binding pocket
I
Highly conserved cis-proline is tolerant to azetidine
Pro205
Arg204
Phe206
Yeast (1QPG)
Mouse (4O3F)
Pig (1KF0)
Human (2WZB)

Figure 2.4. Impact of azetidine substitutions on protein thermal stability.

A) Thermal proteome profiling workflow implemented with mistranslated yeast lysates. B) Total number of unique proline-to-azetidine substitutions and total number of proteins with at least one quantified azetidine substitution. C) Left: Distribution of melting curve differences between peptides containing azetidine and matching wild type peptides. Right: Scatterplot showing the relationship between areas under the melting curve (AUC) and the effect of azetidine on protein stability. Substitutions that significantly alter thermal stability (BH-adjusted p-value < 0.01) are colored in yellow (stabilizing) or blue (destabilizing). D) Relationship between the stability effects of proline-to-azetidine substitutions and proline surface accessibility (Destabilizing: n=25; Non-significant: n=364; Stabilizing: n=9). Listed p-values are computed from Wilcoxon rank-sum test. E) Relationship between the stability effects of proline-to-azetidine substitutions and the predicted ΔΔG of natural substitutions at the same prolines (Destabilizing: n=17; Non-significant: n=135; Stabilizing: n=2). Listed p-values are computed from Wilcoxon rank-sum test. F) Structural representation of Pgk1 showing the stability effects of proline-to-azetidine substitutions. Proline residues with an asterisk are statistically significant (BH-adjusted p-value < 0.01). G-I) Melting curves for Pgk1 residues of interest from panel F classified as destabilizing subsitutions in the 3-phosphoglycerate binding domain (G), destabilizing substitutions in the ATP binding pocket (H), and conserved cis-proline (I).

to influence the cis/trans conformation of the entire helix(Tsai, Overberger, and Zand 1990). While this stretch of prolines has not been previously implicated in Gpm1 stability or function, our data suggest this conformationally-rigid polyproline region is important for Gpm1 stability, and any transition towards a polyproline-I helix (all cis) may be detrimental.

Taken together, these data illustrate how patterns of stability-altering ncAA substitutions within a protein reveal residues and regions important for structure and function. Increasing the depth of substitutions covered in these assays and mistranslating with a variety of ncAAs will bring us closer to interpretable positional sensitivity maps for each protein in the proteome.

2.4 Discussion

Miro is a technology that harnesses mistranslation to produce protein variants *en masse* and mass spectrometry readouts to functionally annotate the consequences of amino acid substitutions on protein function. Using mistranslation with the ncAA azetidine-2-carboxylic acid as a proof of concept, we demonstrate that Miro can identify residues important for protein folding, interaction, and stability.

Mistranslation with ncAAs does not require genetic manipulation. We report a set of ncAAs that incorporate into proteins and can be used in Miro to identify positions that are highly sensitive to substitution. Some amino acids allow for the incorporation of a variety of ncAAs with diverse chemical changes, which should help to interpret the functional consequences of substitutions. For instance, experiments exploring multiple proline analogues with different cis/trans propensities can illuminate protein positions that are sensitive to the conformational context versus general sensitivity to any amino acid substitution.

Miro should be extensible to natural amino acid substitutions by altering the cellular translation machinery to either enhance naturally-occurring mistranslation (Schwartz and Pan 2017) or to produce tailored substitutions, for example, by the use of engineered tRNAs (Berg et al. 2019; Zimmerman et al. 2018). Additionally, the same high throughput mass spectrometry assays developed and implemented in Miro could be used to characterize and functionally annotate variant libraries generated for deep mutational scanning experiments (Fowler et al. 2010).

Similar to other discovery-type mass spectrometry proteomics approaches, Miro does not achieve complete sequence coverage due to sample complexity and the wide range of peptide abundances. In the case of Miro, this limitation is further amplified by the low incorporation of ncAAs necessary to facilitate data interpretation. However, sample enrichment strategies, targeted mass spectrometry acquisition approaches, and future instrumentation improvements should help in the detection of substitution sites and alleviate the coverage issue.

Miro is cost effective, scalable to most types of amino acids, and versatile with respect to the model system in which it can be employed. In this work, we have shown application in yeast and *in vitro* translation reactions; however, we have also screened a subset of these amino acid analogs in *E. coli* and observed similar rates of toxicity and proteome-wide mistranslation. We expect the method to also be adaptable to cultured mammalian cells.

Lastly, Miro is versatile in terms of the protein biochemical properties and functions that can be probed. We have assessed the effects of azetidine substitutions on protein folding, interactions, modifications, and thermal stability, but Miro can readily be extended to assay protein aggregation, enzymatic activity, small molecule binding, and subcellular localization. Further development of Miro with an expanded suite of non-canonical amino acids and the use of

engineered tRNAs to carry out natural substitutions will enable the generation of proteome-wide mutational sensitivity maps for human proteins. As companions to genome sequencing efforts, these maps should allow us to interpret the clinical significance of millions of mutations in the human genome. Furthermore, from the standpoint of protein science, Miro will underpin efforts towards detailed sequence-function relationships, enabling the rational design of proteins, including those with enhanced properties for pharmaceutical and biotechnological applications.

2.5 Methods

2.5.1 Amino acid analogue toxicity screen and incorporation tests.

S. cerevisiae strain BY4741 was grown in duplicate overnight in minimal media supplemented with +Ura, +Leu, +Met, +His, 2% glucose at 30^{o}C, diluted into fresh media to a final OD of 0.1, and grown for two additional doublings. At OD of 0.4, cultures were diluted 1:1 in a 96-well plate containing fresh minimal media and a ncAA of interest at one of eight concentrations. The ncAA concentrations used were 2-fold serial dilutions from a starting concentration of 1000 µg/ml (F1, L3, M2, M3), 500 µg/ml (A1, A2, F2, F3, I1, L2, M1, P2, P3, P4, Q1, V2, V3, W2, W3), or 250 µg/ml (D1, E1, P1, R2, W1, Y1, Y2). Optical density of cultures was monitored for 18 h in a temperature-controlled plate reader (BioTek) measuring OD660 every 15 min.

To assess analogue toxicity, we measured the area under the growth curve (AUC) and established relative toxicity to a control growth curve present in each plate. To do this, we first took the average OD_{660} readings from each time point collected for growth curves. We then scaled growth curves between different plates by subtracting the very first OD reading for each growth curve. We then calculated AUC by summing up all the OD readings for each individual condition (i.e. analog, concentration). To assess the effect of analogue treatment on relative

growth during this time window, we then divided the AUC values for each condition by the AUC for the control, untreated condition.

Collected growth curves were then used to select ncAA concentrations for incorporation analysis. For ncAAs that showed dose-dependent changes in cell growth, we qualitatively selected an analogue concentration around the IC_{50}. For ncAAs that were non-toxic, we chose either 500 µg/mL or 1000 µg/mL. In order to assess incorporation, a single replicate of yeast was grown in 20 ml of minimal media (+Ura, +Leu, +Met, 2% glucose) at 30ºC to OD 0.2 at which point both heavy lysine and a ncAA were added to the culture. One exception were cultures exposed to P1, where we assessed incorporation in synthetic complete media (2% glucose) to guide downstream experiments with this analog. Cultures were harvested at OD 1.0 by centrifuging cultures at 2,850 x g and 4°C for 10 min. Cell pellets were resuspended in ice-cold sterile water, pelleted at 10,000 x g and 4°C for 10 min, snap-frozen in liquid nitrogen, and stored at -80°C until cell lysis and sample preparation. One biological replicate was collected for each analog tested. For each analog, we estimate percent incorporation by dividing the number of ncAA-containing peptide-spectrum matches (PSMs) identified in a run and normalizing by the total number of PSMs that could have containing the ncAA substitution (i.e. number of ncAA-containing PSMs plus number of wild-type PSMs containing the cognate amino acid).

Yeast strain BY4741 was grown in lysine drop out synthetic complete media (SC -Lys) at 30ºC to OD_{600} 0.2, at which point U-^{13}C,^{15}N-lysine (heavy lysine, K8) proline analogue azetidine-2-carboxylic acid was added at 90-100 µg/ml. Cells were harvested at OD_{600} 1.0 by centrifuging cultures at 2,850 x g and 4°C and decanting growth media. Three replicate pellets were resuspended in ice-cold sterile water and pelleted again at 10,600 x g at 4°C. Cell pellets were snap-frozen in liquid nitrogen and stored at -80°C until cell lysis.

2.5.2 Azetidine-2-carboxylic acid incorporation into in vitro translated protein and pulldown.

Mammalian in vitro translation system (Thermo Fisher Scientific IVT Kit #88330) was used to synthesize GST-HA-His protein as described by the manufacturer. Briefly, HeLa cell lysate, reaction mix and accessory proteins were supplemented with pT7CFE1-CGST-HA-His vector, 1mM L-lysine, 1mM L-arginine, protease inhibitors (complete mini EDTA-free, Roche), and 16 mM azetidine-2-carboxylic acid. Reaction proceeded for 8h at 30°C.

In vitro translated protein was purified over 20 µL of anti-HA agarose beads (Sigma-Aldrich), glutathione agarose beads (GoldBio) , or 50 µL of Co2+-NTA beads (GoldBio). Binding was done in 50 mM Tris buffer pH 8.2, 120 mM NaCl, and elution with 100 mM glycine pH 2.5 and 20 mM glutathione, respectively. Co2+-NTA purification was carried out in 50 mM Tris, pH 8.2, 120 mM NaCl, 8M Urea buffer, and eluted with 300 mM imidazole. Purified protein was reduced with 5 mM DTT for 30 min at 55°C, alkylated with 15 mM iodoacetamide for 30 min at RT, and quenched with 15 mM DTT for 15 min at RT. LysC digestion was carried out at a 1:50 enzyme/protein ratio for 2 h at 37°C and pH 8.9. Digestion was quenched with 1% TFA and peptides were desalted over C18 stage tips prior to LC-MS/MS.

2.5.3 Cell lysis, protein reduction, alkylation, and digestion

Frozen cell pellets were thawed on ice and resuspended in lysis buffer (50 mM Tris pH 8.2, 75 mM NaCl, 8 M urea, 50mM β-glycerophosphate, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, 50mM NaF, protease inhibitor (complete mini Roche, 1 tablet/10 ml). Cells were

lysed at 4°C by repeated vigorous agitation with 0.5mm zirconia/silica beads in a bead beater (Biospec) using 4 cycles of 60 s with 75 s of rest in between cycles. Lysates were cleared by centrifugation at 211 x g for 2 min to remove beads and at 10,600 x g for 8 min to remove cell debris, both at 4°C. Protein concentration was determined by BCA assay (Pierce, Thermo Fisher Scientific). Proteins were reduced with 5 mM DTT at 55°C for 30 min, alkylated with 15 mM iodoacetamide for 30 min at RT, and quenched with 15 mM DTT for 15 min at RT. Protein extracts were digested with lysyl-endopeptidase, LysC (Wako) in 50 mM Tris pH 8.9 overnight at RT and 1:100 or 1:50 enzyme to protein ratio. All samples were acidified to pH 2 by addition of trifluoroacetic acid (TFA) to 0.5% final concentration to inactivate the digestive enzyme.

2.5.4 Peptide desalting

Peptides were desalted by reversed-phase solid phase extraction over Sep-Pak tC18 cartridges (Waters) in a vacuum manifold as previously described (Swaney et al. 2013) or stage tips (Rappsilber, Ishihama, and Mann 2003; Rappsilber, Mann, and Ishihama 2007) using Empore C18 material (3M), depending on the peptide scale. Sep-Pak tC18 cartridges were equilibrated with sequential additions of 100% acetonitrile (ACN), 70% ACN with 0.25% acetic acid (AA), 40% ACN with 0.5% AA, and 0.1% TFA. Peptide samples were loaded onto the column and washed with 0.1% TFA and 0.5% AA. Peptide samples were eluted by sequential additions of 750 µl of 40% ACN with 0.5% AA, and 750 µl of 70% ACN with 0.25% AA. For stage tips, peptides were applied to conditioned Empore C18 material, washed with 40 µl 0.1% TFA, and eluted with 40 µl 70% ACN with 0.25% acetic acid (AA). All peptide samples were lyophilized prior mass spectrometry analysis, or phosphopeptide enrichment.

2.5.5 Phosphopeptide enrichment

Phosphopeptide enrichment was performed by immobilized metal affinity chromatography (IMAC) as previously described (Swaney et al. 2013). Briefly, aliquots of 1 mg peptide were resuspended in 80% ACN, 0.1% TFA and incubated with IMAC beads at a ratio of 1 µl 5% slurry (in 80% ACN, 0.1% TFA) per 10 µg peptides. Phosphopeptide-bound IMAC beads were washed with 80% ACN, 0.1% TFA. Phosphopeptides were eluted with 70% ACN, 1% NH_4OH, 29% water. Phosphopeptide eluents were stage-tip desalted and lyophilized as described above.

2.5.6 Pulldown of yeast ribosomal proteins

The strain expressing tagged YNL069C from the MORF collection (Gelperin et al. 2005) was grown in biological triplicate using lysine drop out synthetic complete media (SC -Lys) overnight at 30°C and raffinose as carbon source. Cells were inoculated at OD 0.1 into SC media containing K8, 90 µg/mL azetidine-2-carboxylic acid concentration and 1% galactose to induce Rpl16B expression. Cells were harvested at OD 0.8 by centrifuging cultures at 2,850 x g at 4°C and decanting growth media. Cell pellets were resuspended in ice-cold sterile water and pelleted again at 10,600 x g at 4°C. Cell pellets were snap-frozen in liquid nitrogen and stored at -80°C until cell lysis.

For each biological replicate, cells were resuspended at 4°C in native lysis buffer consisting of 50 mM Tris, 100 mM NaCl, pH 7.5, 0.05% Tween 20 and protease inhibitors (complete mini 1 tablet/10 mL, Roche). Cell suspensions were lysed by four cycles of bead beating (1 min beating, 1.5 min rest). Lysates were centrifuged at 211g for 2 min to remove beads and at 10,600 x g for 8 min to remove cell debris. Supernatants were applied to 60 µL of conditioned IgG Sepharose 6 Fast Flow beads (GE Healthcare) and incubated at 4°C for 3 h. Beads were washed with 50 mM Tris, 100 mM NaCl, pH 7.5, and the bound proteins were eluted with 100 mM glycine pH 2.5. Proteins in the eluate were precipitated with 20% TCA, and their pellet

washed with cold 10% TCA solution, followed by cold acetone, and drying. Protein pellet was resuspended with a buffer containing 50 mM ammonium bicarbonate and 10% ACN, and digested for 4 h with 2 µg of lysyl-endopeptidase (LysC). Digestion was quenched with 0.5% TFA and LysC peptides were desalted on styrene divinylbenzene stage tips (3M) and lyophilized.

2.5.7 Thermal proteome profiling in yeast cell lysates

Two replicates of pelleted yeast (BY4741) containing azetidine mistranslation were resuspended in 650 µl of native lysis buffer (50 mM HEPES pH 7.5, 75 mM NaCl, 2 mM MgCl2, protease inhibitors) and lysed by 4 cycles of bead beating (1 minute each with 1 min rest on ice). Lysates were centrifuged for 10 min at 21,000 x g to remove cell debris. Supernatant was collected and transferred to 2 mL tubes. These lysates (2.25 mg/ml concentration) were then aliquoted into PCR tubes on ice. PCR tubes were incubated on a thermal cycler in two phases: first, a 5-min incubation at 30oC; second, a 5-min incubation at 10 different temperatures (30˚C, 38.3˚C, 45.6˚C, 48.3˚C, 50.0˚C, 52.0˚C, 54.7˚C, 57.0˚C, 62.6˚C, 70.0˚C) for 5 min. After temperature treatment, lysates were incubated at room temperature for 5 min. All samples were then centrifuged at 17,000 x g 4˚C for 1 h. After centrifugation, 75 µl from each temperature-treated sample was mixed 1:1 with denaturing buffer (9M urea, 10 mM DTT, 50 mM HEPES pH 8.9, 75 mM NaCl) and incubated at 55˚C for 30 minutes. 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 30 min at RT. Protein concentration was measured on an additional aliquot treated at 30˚C using the BCA assay.

For each temperature, 75 µg of reduced and alkylated protein lysate was digested with LysC at a 1:50 enzyme:substrate ratio, shaking overnight at RT. Digestion was stopped by addition of

10% TFA to a final concentration of 1.5% TFA. Precipitate was removed by centrifugation, peptides were cleaned up by solid phase extraction on a µHLB Oasis Plate (Waters), and eluates dried down by vacuum centrifugation. 25 µg of dried peptides were resuspended in 100 mM HEPES buffer pH 8.5, 30% acetonitrile, and were labeled with 100 µg of TMT10plex isobaric label reagent (Thermo Fisher Scientific) for 1h at room temperature. The reaction was quenched by addition of 5% hydroxylamine to a final concentration of 0.5% and 30 min incubation. TMT channels corresponding to the different temperatures were pooled together prior to acidification to pH 3 with hydrochloric acid. Acidified peptides were desalted using Sep-Pak tC18 columns (Waters). To increase coverage in Thermal Proteome Profiling experiments, peptides were stage tip fractionated using a high pH reversed phase stepwise elution approach(Lawrence et al. 2015). Briefly, peptides were acidified with 1% TFA, and loaded over conditioned Empore SDB-RPS C18 material (3M), washed and eluted into 4 fractions using a buffer of 20 mM ammonium hydroxide with stepping 5%, 10%, 20% and 80% acetonitrile. Peptide fractions were acidified with 10% FA and dried down by vacuum centrifugation. Samples were analyzed on an Orbitrap Lumos Tribrid mass spectrometer using an SPS-MS3 TMT method (McAlister et al. 2014).

2.5.8 Mass spectrometry

Lyophilized peptide, phosphopeptide, and TMT-labeled peptide samples were resuspended in 3% ACN, 4% formic acid and subjected to liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS).

Peptide samples were loaded into 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 8 to 30% ACN in 0.15% FA over 70 min at 350 nl/min delivered by an Easy1000 nLC system (Thermo Fisher Scientific).

Peptides were online analyzed on a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Mass spectra were collected using a data dependent acquisition method. For each cycle a full MS scan (300-1500 m/z, resolution 70,000, AGC target 3e6) was followed by ten MS/MS scans (isolation width 2.0 Da, 26% normalized collision energy, resolution 17,500, AGC target 5e4) on the top 20 most intense precursor peaks.

TMT labeled peptides 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. Peptides were separated by a 120 min gradient of 8-35% ACN in 0.15% formic acid (gradient was optimized for the high pH reverse phase fractions) delivered at 350 nl/min by a nanoACQUITY UPLC (Waters) and online analyzed on a Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific). Mass spectra were collected using a data dependent SPS-MS3 acquisition method (McAlister et al. 2014), using 5-sec cycles of one full MS scan on the Orbitrap mass analyzer (500-1200 m/z, resolution 60,000, AGC target 5e5), followed by MS/MS scans on the most intense precursor peaks using CID fragmentation and acquisition in the linear ion trap (isolation width of 0.5 Da, normalized collision energy 30, rapid, AGC target 1e4), 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 2.5 Da, normalized collision energy 55, resolution of 50,000, 5e4 AGC, max injection time 86 ms).

Analysis of azetidine incorporation on GST-HA-His protein in vitro translated construct was carried out using both data dependent and targeted acquisition approaches on a Q-Exactive instrument. Peptides were eluted over a similar analytical LC set up as above using a gradient elution of 12 to 40% ACN in 0.15% FA over 60 min at 350 nl/min delivered by an Easy1000 nLC system (Thermo Fisher Scientific). DDA methods were acquired as described above, whereas the targeted PRM method consisted of a full MS scan followed by up to 20 targeted MS/MS scans as defined by a time-scheduled inclusion list that included wild type and azetidine 2 carboxylic acid substituted peptides. MS/MS scan was carried out at 35k resolution, 5e5 AGC target, 100 ms maximum injection time, 2 m/z isolation window, 27% normalized collision energy, centroid mode.

2.5.9 Mass spectrometry data analysis

Raw files were converted to mzXML and mzML formats using ReAdW and MSconvert, and MS/MS spectra were searched against a target/decoy protein sequence database using Comet (version 2015.01, version 2019.01.02) (Eng, Jahan, and Hoopmann 2013) to identify peptides or MaxQuant (version 1.6.5.0) (Cox et al. 2009). *Saccharomyces cerevisiae* (orf_trans_all.fasta downloaded from the Saccharomyces Genome Database in 2016), or human (Uniprot UP000005640 downloaded 2016/09) protein sequence databases were used and tagged bait proteins were included in the database for the affinity purification experiment samples.

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 1.0005 Da 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) and N-terminal acetylation (+42.0106 Da). Variable modifications were also used to search for the incorporation of non-canonical amino acids. For instance, variable modification of -14.0156 Da on proline residues reported for the substitution of proline with azetidine-2-carboxylic acid. Database searches of phosphopeptide samples included variable phosphorylation modification on serine, threonine and tyrosine (+79.9663 Da). Dynamic SILAC samples were searched with light lysine (K0) and heavy (K8, +8.0142 Da) variable modifications in binary mode. TMT-labeled samples were searched with constant modification (+229.1629 Da) on lysines and peptide N-termini. Search results were filtered with Percolator (Percolator version 3.01) (Käll et al. 2007) 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 (Nesvizhskii et al. 2003).

For phosphopeptide samples, phosphosite localization was performed using an in-house implementation of Ascore (Beausoleil et al. 2006) using a fragment mass tolerance of 0.4 Da. Phosphosites with Ascore ≥13 were considered localized (> 95% confidence for localization).

Analysis of targeted acquisition data for in vitro translated GST-HA-His pulldown samples was carried out using Skyline (version 3.6.1). Signal extraction was performed on +2, +3, +4, and +5 precursors and +1, +2 b and y fragment ions. Peptide identifications and chromatographic peak boundaries were refined manually, and precursor MS1 intensities were used to calculate azetidine-2-carboxylic acid incorporation. Peptides containing multiple prolines for which azetidine-2-carboxylic acid substitutions could not be resolved chromatographically (n=5), average incorporation rates were calculated.

2.5.10 Selection of peptides for melting curve fitting

For fitting melting curves, we first applied several stringent filtering criteria. First, we only considered peptides that were fully cleaved, in order to reduce confounding effects due to digestion biases across the temperature range. Second, we only considered peptides containing a heavy lysine, constraining our analysis to the statistical proteome synthesized after exposure to the ncAA. Lastly, we only considered PSMs where at least 5 of the top 10 most intense fragment ions in the MS2 belonged to the assigned peptide, reducing ratio compression from co-isolated precursors. 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.

2.5.11 Peptide-level melting curve normalization

To account for differences in the amount of material labeled in each channel, we applied a normalization approach similar to a previous approach (Miettinen et al. 2018). Briefly, we selected a set of proteins with relative fold-changes between 0.5 and 1.5 across the entire temperature range, and with a minimum of 3 unique, heavy-labeled, non-redundant, and non-azetidine 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.

2.5.12 Significance tests for the effect of azetidine substitutions on protein thermal stability

To identify azetidine substitutions that significantly alter protein thermal stability, we compared the melting curves of azetidine-containing peptides with their matching wild type peptides using non-parametric analysis of response curves (NPARC) (Childs et al. 2019). In order to estimate the null distribution for our dataset, we took a modified approach to the original implementation of NPARC. We first calculated F-statistics for each peptide pair (degrees of freedom: df1 = 34, df2 = 3). Specifically, peptides that spanned the same ncAA substitution were first consolidated into a site-level quantification by taking the median fold-change across channels. Then, instead of relying on the theoretical F-distribution to calculate a p-value, we generated a permuted dataset of null F-statistics and used these null values to calculate empirical p-values. Specifically, for each observed peptide pair, we generated 500 permuted melting curve comparisons by sampling (without replacement) the relative fold-changes at each temperature, across both replicates and peptide fold-changes. We then fit null and alternative melting curves in R, calculated F-statistics, and combined the 500 permutations across all 465 peptide pairs to generate a dataset of 232,500 permuted curves, each with an associated F-statistic. Any permuted models that failed to converge were first removed before calculating empirical p-values, leaving us with a final dataset of 230,665 null F-statistics (F_{null}). We then calculated empirical p-values for each observed F-statistic ($F_{observed}$) using the following formula:

$$p = \frac{1 + (number\ of\ instances\ where\ F_{null} > F_{observed})}{1 + total\ number\ of\ F_{null}}$$

All p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method in R.

2.5.13 Bioinformatics

Gene Ontology Term enrichment for ribosomal pulldown was carried out using YeastMine server (https://yeastmine.yeastgenome.org/). Ontology terms for Biological function , cellular component and molecular function were analyzed for proteins identified in the ribosomal pulldown fraction against a background list of proteins generated from whole cell lysate samples. p-values were corrected using the Holm-Bonferroni method and a 0.05 significance threshold was used.

We assessed the structural and evolutionary context for all proline residues where we detected an azetidine substitution using a variety of computational tools. For all analyses, we only used peptides containing a single azetidine substitution.

For secondary structure analysis, we downloaded PDB files of predicted protein structures from AlphaFold (Jumper et al. 2021) (downloaded 2021-09-21). We assigned secondary structure elements to each residue using DSSP with the Bio3d package in R (Grant et al. 2006). The 8-state SSE assignments were then consolidated to 4-state SSE (Helix, Sheet, Turn, Coil) by collectively calling H, G, I states as Helix; E and B states as Sheet; S and T as Turn; and all others as Coil. We calculated relative solvent accessibility by dividing the solvent accessible surface area (SASA) output from DSSP (Kabsch and Sander 1983) by a list of maximum SASA per residue (Tien et al. 2013). Prediction scores were extracted from each PDB file, and residues with prediction scores less than 70 were excluded from our analysis. Proline residues with omega bond angles between -30˚ and 30˚ were considered to reside in cis. Disordered predictions for each residue in the yeast proteome were downloaded from

http://bioinfadmin.cs.ucl.ac.uk/disodb/. A disordered region was defined as a segment of 30 or more consecutive residues in a protein with a prediction of being disordered.

In order to compare the effects of natural amino acid substitutions with ncAA substitutions, we downloaded the entire *in silico* mutagenesis dataset (both the homology-based and experimental models) from the mutfunc database (Wagih et al. 2018). We then calculated a mean $\Delta\Delta G$ for each proline residue and used those values as representative of the overall mutational sensitivity of a proline residue.

For our evolutionary analysis, we submitted FASTA sequences obtained from Uniprot (accessed 11 April 2021) to the BLAST server from the command line. Sequences were submitted to BLASTp using default parameters with `refseq_protein` as the search database and `Eukarya[ORGN]` as the entrez query. After searches were complete, we removed any hits with sequence identity less than 30% and E-score > 0.05. We then generated a multiple sequence alignment for each protein in R using the Bio3d and msa package, using the filtered hits as input sequences and default parameters for protein sequence alignments. We then calculated positional entropy within each sequence alignment using the Bio3d package (Grant et al. 2006) and the Shannon entropy algorithm. Raw H22 entropy scores were used in our final analysis. We only analyzed positions with more than 100 sequences in the alignment and less than 30% gaps.

2.5.14 Data availability

The mass spectrometry data have been deposited to the ProteomeXchange Consortium (https://www.ebi.ac.uk/pride/archive/) via the PRIDE partner repository with the dataset identifiers: PXD031230, PXD031233, PXD031255.

2.5.15 Acknowledgements

We thank members of the Villén laboratory for enriching scientific discussions. This work is primarily supported by a Medical Research Program grant from the W.M. Keck Foundation (to J.V., S.F., and W.S.N.). This work is supported in part by the University of Washington's Proteomics Resource (UWPR95794). The work and personnel involved in the project are additionally supported by NIH grants R35GM119536 (J.V.), R01AG056359 (J.V.), and RM1HG010461 (JV. and S.F.); and Human Frontiers Science Program grant RGP0034/2018 (J.V.). K.N.H and I.R.S. were supported by NIH training grant T32HG000035. A.S.B. was supported by NIH training grant T32LM012419. B.Y.R. was supported by NSF Graduate Research Fellowship DGE-1256082. S.M.Z. was a Washington Research Foundation fellow of the Life Sciences Research Foundation.

Chapter 3: Identifying residues important for stability, structure, and function

This Chapter is based on unpublished observations

3.1 Summary

Amino acid substitutions fuel evolutionary innovation and drive organismal dysfunction. Delineating between such disparate mutational outcomes is critical for basic science and in fulfilling the promise of precision medicine and synthetic biology. Towards this end, we have developed Miro, a proteome-wide method that assays the functional effects of amino acid substitutions by the steps of mistranslation with a non-canonical amino acid, a selection that separates mistranslated proteins by their function, and mass spectrometry to identify the protein variants in each functional class. Here, we couple Miro with a high-throughput thermal protein stability assay that subjects mistranslated proteins to stepwise increases in temperature followed by measurement of the area under the melting curve. Using eight different non-canonical amino acids in turn, we assess changes in thermal stability for 8,919 unique substitutions in 715 proteins across the yeast proteome. This approach allows us to delineate the features of amino acid sites whose substitution correlates with decreased thermal stability. Further, we find that alternative ncAA substitutions at the same site can have differing effects, reflecting structural constraints at these sites. Finally, we identify clusters of residues sensitive to substitution that often correspond to protein regions with discrete functional activities. These results indicate that the use of Miro with multiple types of substitution provides insight into protein structure and function.

3.2 Introduction

Sequence-function relationships are ubiquitous across biological molecules and systems. Understanding their guiding principles has advanced structural prediction (Jumper et al. 2021), directed evolution (Wittmann, Yue, and Arnold 2021; Hsu et al. 2022), human genetics (Frazer et al. 2021), and synthetic biology (P.-S. Huang, Boyken, and Baker 2016; Kuhlman et al. 2003). The challenge now lies in assigning distinct structural or functional roles to individual amino acid

residues within proteins, which would empower our predictions regarding the effects of millions of missense variants (Karczewski et al. 2020) and post-translational modifications (Ochoa et al. 2020) on protein biology. High-throughput functional genomics and proteomics methods offer an opportunity to bridge this gap between variant identification and variant annotation.

Recent advances in DNA sequencing have enabled the characterization of protein variants in high-throughput (Fowler and Fields 2014). These methods have, in-turn, advanced our understanding of protein sequence-function relationships (Fowler et al. 2010; Song et al. 2020), in addition to advancing our ability to determine the effects of clinical variation within disease genes (Findlay et al. 2018). However, these approaches come at the cost of proteome coverage and sophisticated functional selections, many of which are limited in their ease-of-portability across genes and experimental systems. To combat the limitation of scale, complementary mutagenesis approaches have emerged that characterize sequence variants across entire genomes (Després et al. 2020; Sharon et al. 2018; Hanna et al. 2020). While these tackle the limitation of scale, they are, thus far, constrained to variants affecting cellular or organismal fitness, limiting their applications to essential genes and functions.

We recently developed Miro (Rodriguez-Mias et al. 2022), a proteome-wide method that characterizes the effects of amino acid substitutions on protein structure and function *in vitro* or *in vivo*. Miro first generates collections of protein variants via proteome-wide mistranslation with a non-canonical amino acid. Then, these mistranslated proteomes are coupled to biochemical selections with a mass spectrometry-based readout to quantify the effect of an amino acid substitution on a biochemical property of interest. Our initial implementation of Miro showcased multiple biochemical selections applied to a single yeast proteome mistranslated with azetidine-2-carboxylic acid (Rodriguez-Mias et al. 2022). Here, we apply a single biochemical selection to yeast proteomes generated by mistranslation with one of eight different ncAAs.

First, we benchmark a high-throughput stability assay inspired by proteome integral solubility alteration (Gaetani et al. 2019). We then apply this assay to each of the eight mistranslated proteomes, gaining insight into structural and functional contexts underlying ncAA sensitivity. We then use this method to identify functional protein regions in a subset of yeast proteins. Taken together, these data illustrate the ability of Miro to map sequence-functional relationships proteome-wide.

3.3 Results

A high-throughput method to quantify the effects of amino acid substitutions on protein thermal stability

To measure protein thermal stability in high-throughput, we developed a modified version of the proteome integral solubility alteration (PISA) assay (Gaetani et al. 2019), shown in Figure 3.1. The original implementation of PISA involves: (1) incubating lysates or cells with or without a small molecule of interest; (2) thermal denaturation by splitting samples across PCR tubes and incubating at one of several temperatures along a temperature gradient; (3) pooling the samples into a single tube after temperature treatment and discarding aggregates via centrifugation; (4) measuring changes in protein stability between conditions by measuring the difference in the area under the melting curve between treated and untreated samples. Our approach extends the PISA assay to measure differences in stability between protein variants through a simple modification: the introduction of a 30˚C-treated control channel. This channel provides a measurement of protein abundance before thermal denaturation, which is used to normalize measurements of protein abundance *after* temperature treatment to enable an estimate of relative stability (R_s), a measurement of how much protein has aggregated within a temperature window. These R_s measurements are akin to melting temperatures, enabling a reliable comparison of stability between different proteins and variants within the same sample.

We benchmarked modified PISA against a yeast proteome with azetidine-2-carboxylic acid substitutions proteome-wide. Cell pellets from five biological replicates were lysed in a non-denaturing buffer and heated to 30˚C or to temperatures between 46˚C and 56˚C. In total, we quantified R_s for 13,583 peptides and 1,516 yeast proteins. We observed partial-to-full aggregation for 1,078 yeast proteins (R_s < 95%), with 438 yeast proteins remaining stable across these temperatures (R_s > 95%) (Figure 3.1B). Estimates of R_s at the peptide-level accurately reflected protein-level R_s (R^2 = 0.88; Figure 3.1C), and estimates of R_s were highly reproducible (Pearson's R^2 peptide = 0.80; protein = 0.92; pep-to-prot = 0.84) (Figure 3.1D).

Next, we assessed modified PISA's ability to identify azetidine-sensitive residues across the proteome. To do this, we first statistically compared the R_s value of ncAA-containing peptides (i.e. "ncAA") with the R_s value of the unmodified counterpart peptides (i.e. "wildtype" or "WT"). Using this comparison, we categorized azetidine substitutions as non-significant (n.s.), destabilizing, or stabilizing. Next, we asked whether these annotations correspond to known changes in stability measured in a prior azetidine-mistranslated proteome (Rodriguez-Mias et al. 2022). We focused on a set of 164 overlapping azetidine substitutions for which the melting temperature in TPP was less than 56˚C (i.e. the max temperature used here) and the wildtype peptide R_s was less than 100%. Across these 164 overlapping sites, PISA categorized 21 substitutions as destabilizing, 7 as stabilizing, and the remaining 136 as non-significant (paired t-test, BH-adjusted p-value < 0.05, Figure 3.1E). PISA-destabilizing and PISA-stabilizing substitutions were significantly more likely to be destabilizing or stabilizing, respectively, in TPP (destabilizing, median ΔAUC = -1.84, wilcoxon rank-sum test, p-value = 9.5×10^{-07}; stabilizing, median ΔAUC = 1.88, wilcoxon rank-sum test, p-value = 0.008) compared to non-significant substitutions (Figure 3.1E). Taken together, these data confirm that peptide-level R_s values reflect and can measure the effects of amino acid substitutions on protein thermal stability.

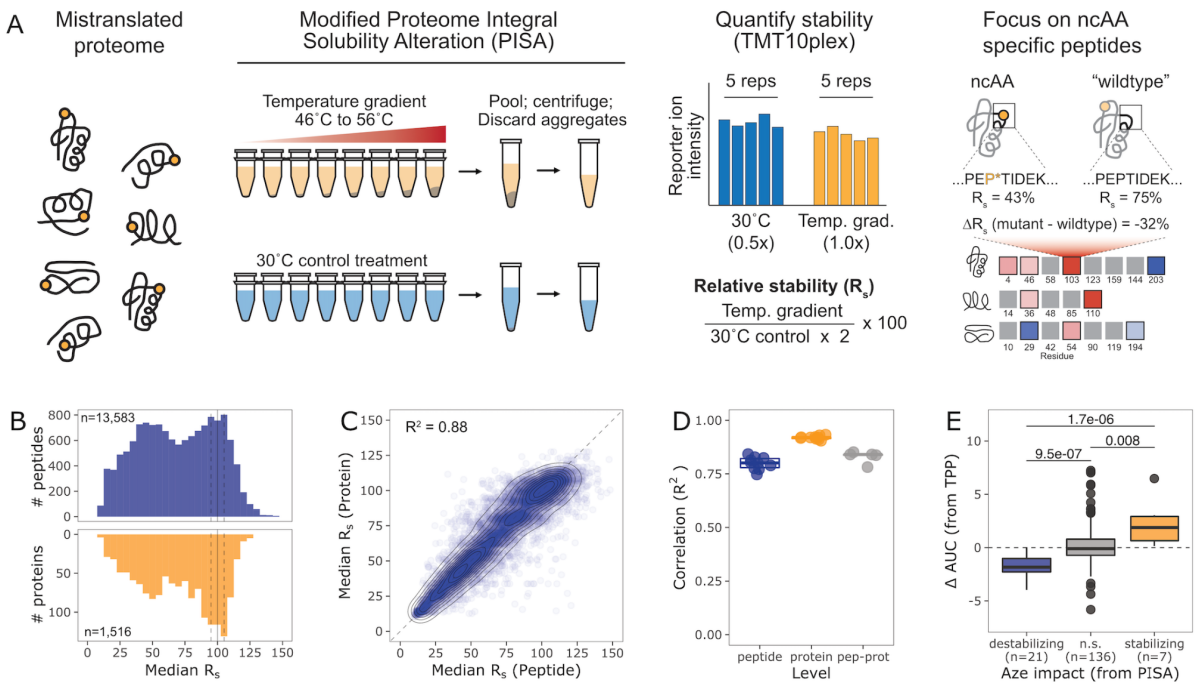


Figure 3.1. A high-throughput thermal stability assay to measure the effects of amino acid substitutions.

(A) Workflow to characterize the effects of amino acid substitutions on protein thermal stability. (B) Distribution of median relative stability (R_s) values for peptides and proteins measured across five biological replicates of an azetidine-mistranslated yeast proteome treated between 46˚C and 56˚C. (C) Correlation between median peptide R_s and median protein R_s. (D) Correlation of R_s values for all pairwise comparisons between five replicates for peptides, proteins, and correlation between peptides and their corresponding proteins (pep-prot). (E) Site-specific effects of azetidine substitutions measured in modified PISA and in a prior dataset measuring the effect of azetidine substitutions on protein stability with thermal proteome profiling (data from (Rodriguez-Mias et al. 2022)).

Towards an atlas of non-canonical amino acid sensitivity in yeast

We next sought to leverage the increased throughput and simplicity enabled by modified PISA to probe the effects of different ncAAs on yeast protein thermal stability. We focused our efforts on eight ncAAs that incorporate in place of one of seven cognate amino acids, which we selected to explore a range of chemical changes across the yeast proteome (Figure 3.2A).

To generate mistranslated proteomes for each ncAA, we first established relationships between ncAA concentration and proteome-wide mistranslation in yeast. Our goal was to identify ncAA concentrations that generate proteomes in which most newly-synthesized proteins would contain, on average, no more than two ncAA incorporation events per molecule synthesized (Supplemental Figure 3.1). To do this, we took a three step approach. First, we screened seven of these eight analogs (the exception being azetidine, due to prior information regarding toxicity) for dose-dependent toxicity in yeast growing on synthetic complete media (Supplemental Figure 3.2A) and determined an IC50 for each ncAA (Supplemental Figure 3.2B). Second, for each analog, we selected 3-5 concentrations around the IC50 (Table 1), grew cultures of yeast at these concentrations, and assessed the relationship between ncAA concentration and (a) number of ncAA-containing peptides (Supplemental Figure 3.3A) and (b) mistranslation levels using mass spectrometry (Supplemental Figure 3.3B-D). Lastly, we used these dose-dependent relationships to guide which ncAA concentrations to use to generate desired levels of mistranslation (Table 1). Using this approach, we established dose-dependent relationships between analog concentrations and proteome-wide mistranslation for eight different ncAAs (Supplemental Figure 3.3D). We generated five biological replicates of each minimally-mistranslated proteome and then coupled them with modified PISA to quantify the site-specific effects of ncAA substitutions.

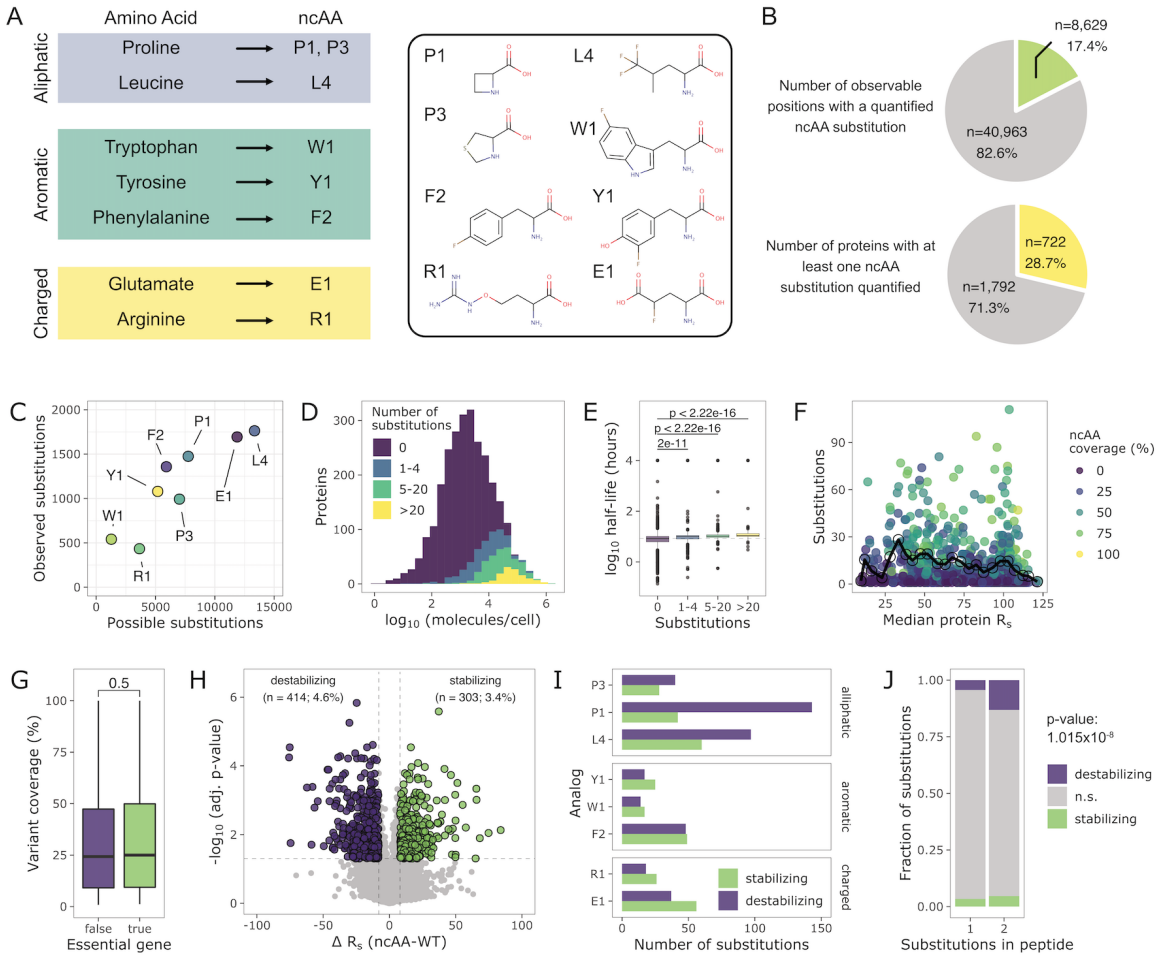


Figure 3.2. Global features of the mistranslated meltome in yeast.

(A) Eight mistranslated proteomes and the corresponding structures for each mistranslating non-canonical amino acid. (B) Top: Number of positions across the entire detectable yeast proteome where a ncAA substitution was detected; Bottom: Total number of proteins in the detected yeast proteome with at least one ncAA substitution quantified. (C) Number of unique ncAA substitutions observed across eight different mistranslated proteomes as a function of the total number of possible substitutions based on detected wildtype peptides. (D) Distribution of protein abundances in yeast, with different colors representing proteins grouped based on the numbers of ncAA variants quantified. Abundance estimates from (Kulak et al. 2014). (E) Differences in protein half-life based on the number of variants quantified for different proteins. Half-life measurements from (Christiano et al. 2014). (F) The number of quantified ncAA substitutions and overall variant coverage as a function of protein relative stability. (G) Variant coverage between essential and non-essential genes in yeast. (H) Site-level effects of ncAA substitutions on protein thermal stability in yeast (paired t-test in Limma; n=8,919). (I) Site-specific effects of ncAA substitutions for each mistranslated proteome. (J) Fraction of substitutions that alter thermal stability based on the number of substitutions contained within the peptide.

Across these eight mistranslated proteomes, we quantified the effects of 8,919 ncAA substitutions mapping back to at least one of 8,629 unique positions across 722 yeast proteins (Figure 3.2B). In total, these sites and proteins represent 28.7% of the quantified yeast proteome and 17.4% of all possible positions based on the set of observed wildtype peptides (Figure 3.2B). In general, the coverage of ncAA substitutions varied across each sample and each protein. We captured more ncAA substitutions for analogs that mistranslated at frequent amino acids (Figure 3.2C), more substitutions in abundant proteins (Figure 3.2D), and more substitutions in proteins with a slower half-life (Figure 3.2E). Interestingly, there was no significant difference in the number of quantified variants based on protein thermal stability (Figure 3.2F), and no significant bias in variant coverage between essential and non-essential yeast proteins (Figure 3.2G). The lack of bias in variant coverage between essential and non-essential genes reflects the indiscriminate nature of mistranslation and suggests the predominant constraint on variant coverage is peptide detectability (although there may also be underlying biological constraints, i.e., altered turnover of mutant proteins). Equal variant coverage between essential and non-essential genes also highlights a major advantage of Miro, its ability to measure variant effects in both essential and non-essential proteins without the need to couple effects with indirect readouts (e.g., cell fitness, fluorescence intensity).

Next, we assessed the site-specific effects of ncAA substitutions on protein thermal stability. Out of the 8,919 quantified substitutions, 717 (~8%) significantly altered thermal stability (Figure 3.2H). These were slightly biased towards destabilizing substitutions (414 destabilizing vs. 303 stabilizing) (Figure 3.2H), which was more pronounced for certain ncAAs, such as azetidine (P1) and trifluoro-leucine (L4) (Figure 3.2I). These more pronounced effects suggest that the observed global effect of a ncAA on protein stability is related to the severity of the chemical change introduced by the analog. Along similar lines, we found that variants peptides for which we quantified the effect of a double mistranslation event (i.e. 2 ncAA substitutions) were

significantly more likely to have altered stability compared to those with single mistranslation events (Fisher exact test, p-value = 1.015×10^{-8}, Figure 3.2J), underscoring the compounding effects of additional substitutions on protein stability. However, we were surprised to find that most substitutions (~92%) did not alter thermal stability. This “tolerance” to ncAA substitutions supports earlier findings with azetidine (Rodriguez-Mias et al. 2022) and suggests more broadly that the site-specific effect of a ncAA substitution is strongly tied to local structural and functional contexts for each site within each protein.

Structural contexts partly explain ncAA sensitivity

Next, we systematically explored the structural contexts surrounding stability-altering substitutions to better understand factors contributing to ncAA sensitivity. To do this, we mapped ncAA substitutions back to their positions within predicted protein structures from AlphaFold (Jumper et al. 2021), and extracted site-level molecular features (Figure 3.3A). We also included *in silico* predictions of $\Delta\Delta G$ and SIFT scores for all 19 natural amino acid substitutions at these positions from the mutfunc database (Wagih et al. 2018), as well as predicted residue disorder.

Our analysis identified several features correlating with ncAA sensitivity that match prior observations with natural amino acid substitutions (Figure 3.3B). For example, sites where ncAA substitutions had a destabilizing effect were significantly less solvent exposed compared to sites with non-significant substitutions (wilcoxon rank-sum test, p-value = 6.4×10^{-8}; Figure 3.3C). This same pattern has been observed with natural amino acid substitutions, with buried amino acid residues generally more sensitive to substitutions compared to surface-exposed residues (Iqbal et al. 2020; Matreyek et al. 2018). We also observed correlations with thermodynamic and evolutionary properties of residues. For example, destabilizing substitutions occurred at residues that had a significantly higher average $\Delta\Delta G$ from *in silico* predictions of natural amino

acid substitutions (wilcoxon rank-sum test, p-value = $8.2x10^{-7}$; Figure 3.3D). Interestingly, stabilizing substitutions only weakly correlated with structural features (Figure 3.3B), suggesting a more complicated mechanistic interpretation is required for stabilizing substitutions.

A more analog-centric approach yielded additional insights related to ncAA sensitivity (Figure 3.3B). For example, destabilizing fluoroglutamate (E1) and canavanine (R1) substitutions were more likely to occur at buried glutamic acid or arginine residues, respectively (E1, wilcoxon rank-sum test p-value =$3.3x10^{-4}$; R1, wilcoxon rank-sum test p-value = 0.0054, Figure 3.3E). Fluoroglutamate is particularly interesting because the pKa of the side-chain carboxylic acid is indistinguishable from glutamate. However, in the context of more hydrophobic environments, such as at the core of protein structures, pKa values of glutamate residues can increase by several orders of magnitude (Isom et al. 2010) and are likely more sensitive to fluoroglutamate. Interpreted in this light, these data illustrate how ncAA sensitivity can yield information about residue burial in lieu of structural information.

As an example, we highlight fluoroglutamate (E1) and canavanine (R1) substitutions within glycerol-3-phosphate phosphatase (Gpp1) (Figure 3.3F). Incorporation of E1 or R1 at either Glu178 or Arg157 significantly destabilized Gpp1. Both of these substitutions occur at highly-connected buried residues at the core of the protein (Figure 3.3F-I). These effects are potentially driven by the ncAA disrupting these interactions. However, solvent accessibility is not the only predictor for ncAA sensitivity, as seen by the lack of sensitivity at Glu79 to E1, which is also buried (Figure 3.3F). Taken together, the multiple protein facets correlating with residue sensitivity illustrates some of the complexities uncovering broad generalizations underpinning ncAA sensitivity, but also the ability to use ncAA sensitivity to classify amino acids into subtypes.

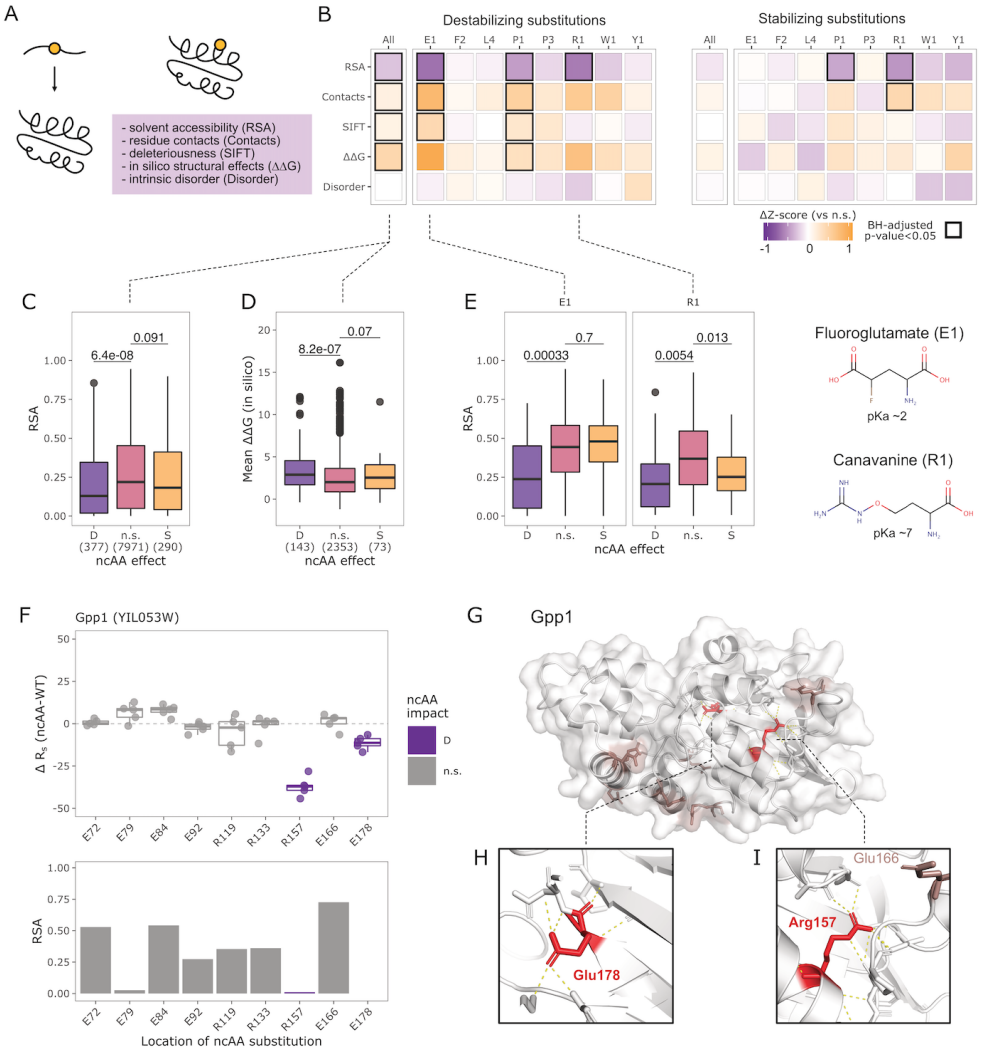


Figure 3.3. Structural contexts partly explain ncAA effects.

(A) Each unique position across the yeast proteome where a ncAA substitution was quantified was mapped back to predicted protein structures from AlphaFold and site-level characteristics were extracted. (B) Global and analog-specific correlations between ncAA effect and structural and evolutionary features. Left side: correlates with destabilizing substitutions; Right side: correlates with stabilizing substitutions. Fill color represents the scaled difference between the destabilizing or stabilizing substitutions and the non-significant substitutions. Max differences were set to -1 and 1 for visualization purposes. Boxes with a bold outline are features with significant correlation with ncAA effect (BH-adjusted p-values < 0.05). All statistical tests were performed on the unscaled values. (C-D) Correlation between ncAA effect and relative solvent accessibility (C) and average ΔΔG (D). Values for panel (D) were obtained from in silico models generated using FoldX and stored in the mutfunc database. (E) Analog-centric correlations between fluoroglutamate (E1) and canavanine (R1) with solvent accessibility. (F) Example effects of E1 and R1 substitutions detected in Gpp1 (top panel) and their correlation with solvent accessibility (bottom panel). (G) Predicted structure of Gpp1. (H-I) Residue interactions surrounding Glu178 (H) and Arg157 (I). Yellow lines represent polar interactions.

Interestingly, there were limited associations between ncAA sensitivity and residue disorder. While this lack of correlation suggests that regions of intrinsic disorder are more tolerant to amino acid substitutions, we also observed a systematic depletion of substitutions in intrinsically disordered regions, suggesting the relationship between ncAA sensitivity and disorder is not clear cut from these data.

Residues at interaction interfaces are sensitive to ncAAs

Next, we took a deeper look at stability-altering substitutions occurring at interaction interfaces. We focused on interaction interfaces for two reasons: (1) thermal stability assays, such as PISA and TPP, are acutely sensitive to interactions (Tan et al. 2018; Becher et al. 2018; Dai et al. 2018); and (2) TPP and PISA can be used to identify PTMs at interfaces regulating interactions based on changes in protein stability (Smith et al. 2021; J. X. Huang et al. 2019). Given the types of chemical changes introduced by ncAAs, we reasoned that we could identify interaction residues using a similar approach.

To explore the effects of ncAA substitutions at interface residues, we integrated known interactions from the INSIDER database (M. J. Meyer et al. 2018) and from the mutfunc database (Wagih et al. 2018) that were based on crystalized interactions documented in the PDB. Across our dataset, we quantified the effect of ncAA substitutions at 2,119 positions in 115 proteins found within the INSIDER or mutfunc database (excluding major macromolecular complexes like the ribosome or proteasome). Of these, 458 positions occurred at interaction interfaces, while 1,661 ncAA substitutions occurred outside known interaction interfaces (Figure 3.4A). We confirmed that interface residues were significantly more surface exposed compared to residues outside known interfaces (Figure 3.4B). However, we saw no significant difference in the sensitivity of these residues to ncAA substitutions (Figure 3.4C). Out of 458 substitutions at

interfaces, 36 (7.9%) significantly altered protein thermal stability (25 destabilizing, 11 stabilizing), which matches what we observed for residues outside of known interfaces, with 122 out of 1,661 substitutions (7.3%) altering thermal stability (72 destabilizing, 50 stabilizing).

For a subset of substitutions occurring at interaction interfaces, we were able to compare changes in ΔΔG caused by all 19 natural substitutions modeled in the context of an interaction (inter) compared to as a monomer (mono) from the mutfunc database. Interestingly, destabilizing ncAA substitutions at interaction interfaces had a significantly higher predicted ΔΔG in the context of the interaction compared to the monomer (paired wilcoxon rank-sum test, p-value = 0.0017, n = 20, Figure 3.4C), and also had a significantly higher predicted ΔΔG in the interaction context when compared to sites with non-significant substitutions quantified in our assay (wilcoxon rank-sum test, p-value = 0.0009, Figure 3.4C).

For some proteins in our dataset, we captured multiple stability-altering substitutions across an interaction interface. For example, in yeast pyruvate kinase (Pyk1), we captured two stability-altering substitutions at the dimer interface (Figure 3.4E). Incorporation of azetidine (P1) at Pro274 destabilized Pyk1 (ΔR_s= -14.9, BH-adjusted p-value = 0.0019, Figure 3.4E), whereas incorporation of trifluoroleucine (L4) at Leu277 stabilized Pyk1 (ΔR_s = +9.3, BH-adjusted p-value = 0.018, Figure 3.4E). Neither of these residues have been experimentally shown to be involved in Pyk1 dimer formation, but both are in direct contact with the other subunit (Figure 3. 4E). Interestingly, one of these positions is highly conserved across eukaryotes, but has diverged within yeast (Figure 3.4F). Comparing the structures of yeast and human Pyk1 dimers reveals a conservation of residue contacts between each subunit, despite their evolutionary divergence (Figure 3.4F). The sensitivity of these residues to ncAA substitutions illustrates how experimental information can pinpoint interacting residues where site-level evolutionary metrics alone may conceal their importance.

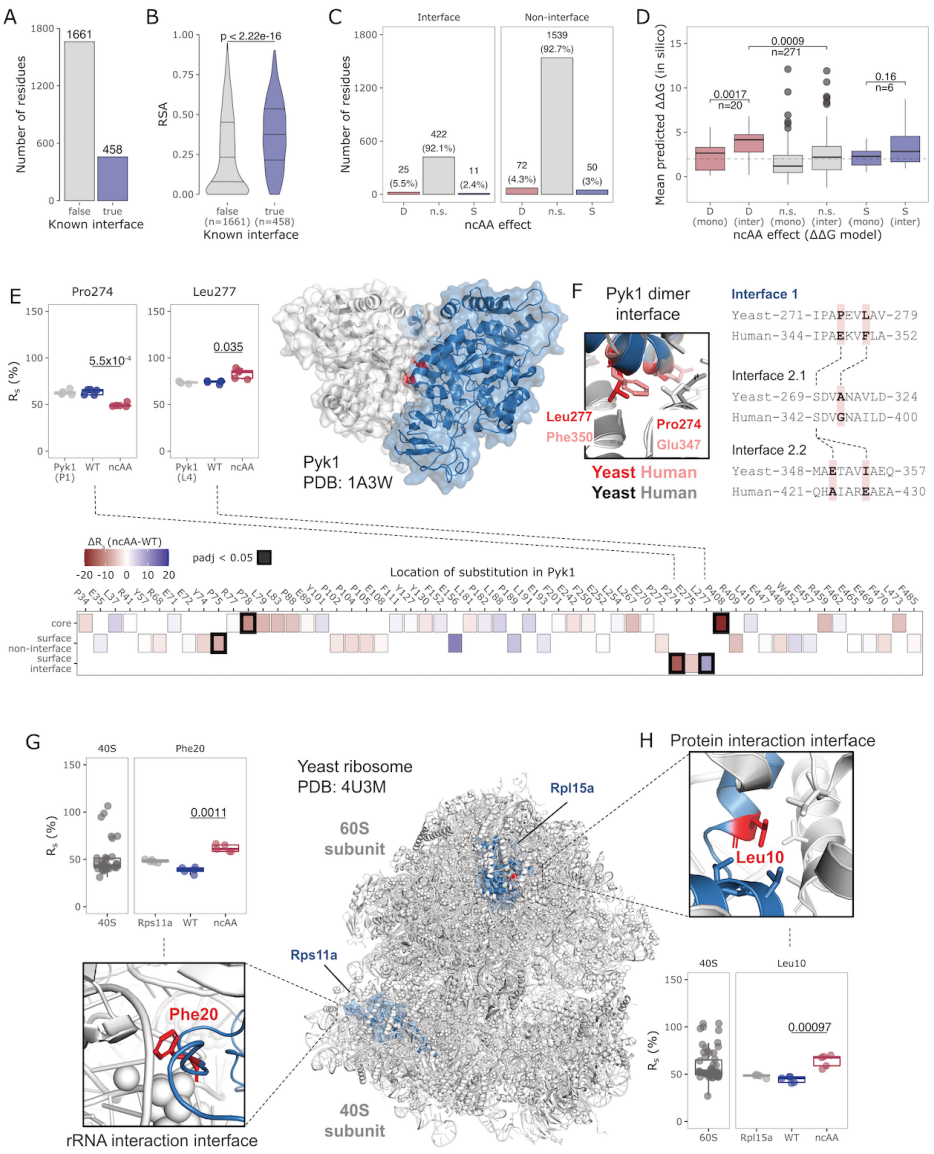


Figure 3.4. Residues at interaction interfaces are sensitive to ncAA substitutions.

(A) Total number interface residues and non-interface residues with a quantified ncAA substitution in this study. (B) Interface residues and relative solvent accessibility. (C) Effect of ncAA substitutions at interface vs non-interface residues. (D) Modeled effects of natural substitutions on stability of protein-protein interactions. (E) ncAA sensitivity map for pyruvate kinase (Pyk1; YAL038W). The two highlighted substitutions (Pro274; Leu277) are residues at the Pyk1 dimer interface. The ncAA sensitivity map highlights the effect of different ncAA substitutions occurring at different locations with Pyk1. Sites with a thick black border significant (BH-adjusted p-value < 0.05). (F) Structural and evolutionary comparison of interface residues between yeast Pyk1 and human Pyk1 where ncAA substitutions had an effect on protein stability. (G) Effect of F2 incorporation at Phe20 in Rps11a. (H) Effect of L4 incorporation at Leu10 in Rpl15a.

We also identified stability-altering substitutions within protein complexes that may impact protein-protein and protein-RNA interactions. For example, in the ribosome, the incorporation of fluorophenylalanine (F2) at Phe20 in Rps11A increased stability (ΔR_s= +14.9, BH-adjusted p-value = 0.0019, Figure 3.4G). This site sits at the interface between Rps11a and ribosome RNA (Figure 3.4G). In Rpl15a, incorporation of trifluoroluecine at Leu10 also increased protein stability (ΔR_s= +14.9, BH-adjusted p-value = 0.00097, Figure 3.4H). This site sits at an interaction interface within the ribosome (Figure 3.4H). While the observed stabilizing effects of these substitutions may reflect a direct biophysical change in stability, another possibility is that these substitutions reduce the co-aggregation propensity of these variants with the rest of their respective ribosomal subunit, either due to decreased occupancy of that variant within ribosomal molecules or altered subunit binding affinity. Taken together, these data showcase how surface-exposed residues that are sensitive to ncAA substitutions can pinpoint positions involved in biomolecule interactions or protein function.

Residue sensitivity is specific to ncAA identity

As the above analyses illustrate, different structural contexts play an important role in shaping the effect of ncAA substitutions. We next asked a complementary question: do different ncAA substitutions occurring in the *same* structural context elicit different effects on protein thermal stability? To answer this question, we focused on two ncAAs, azetidine (P1) and thioproline (P3), that introduce unique chemical changes at proline residues; P1 perturbs cis-trans isomerization (Kern, Schutkowski, and Drakenberg 1997) and hydrophobicity (loss of methylene group), while P3 alters hydrophobicity and sterics (Figure 3.5A).

In total, we quantified changes in thermal stability for 1,371 azetidine substitutions and 919 thioproline substitutions, totaling 2,290 unique ncAA substitutions occurring at 1,655 proline

residues (Figure 3.5B). Of these, we quantified the effects of both substitutions at 644 proline residues (Figure 3.5B). In general, proline-to-azetidine substitutions were more likely to significantly alter protein thermal stability compared to thioproline substitutions, in agreement with the degree of chemical change introduced by each substitution (Figure 3.5C).

Focusing on the 644 overlapping proline residues yielded additional insights regarding the factors contributing to ncAA-specific sensitivity. The vast majority of overlapping sites were not sensitive to either substitution (539 that were non-significant in both). However, out of the 105 positions where ncAAs did alter thermal stability, the sites were mutually exclusive, with 69 sites exclusively sensitive to azetidine and 28 sites exclusively sensitive to thioproline (Figure 3.5D, Figure 3.5E). Only 4 proline residues showed the same sensitivity towards azetidine and thioproline, with 4 residues showing opposing sensitivities (Figure 3.5D). This minimal overlap in ncAA effect suggests that, at least in the context of proline, different ncAA substitutions behave like different cognate amino acid substitutions. This minimal overlap also indicates that even structurally-similar analogs can yield insight into differential residue sensitivity.

A deeper dive into these overlapping residues and the factors correlating with differential sensitivity reveals several molecular features. For example, residues uniquely sensitive to azetidine were enriched in alpha-helices compared to thioproline-sensitive and non-significant residues (Figure 3.5F). Residues uniquely sensitive to azetidine were also less likely to be in cis-conformation, albeit this is a correlation without statistical significance.

To illustrate substitution-specific sensitivities within a protein, we highlight three different proline residues within Rpl12a (Pro30, Pro34, and Pro39) where we quantified the effects of azetidine and thioproline (Figure 3.5G). Across these three residues, each ncAA had a unique effect. For example, incorporation of azetidine or thioproline at Pro30 had no effect on Rpl12a stability.

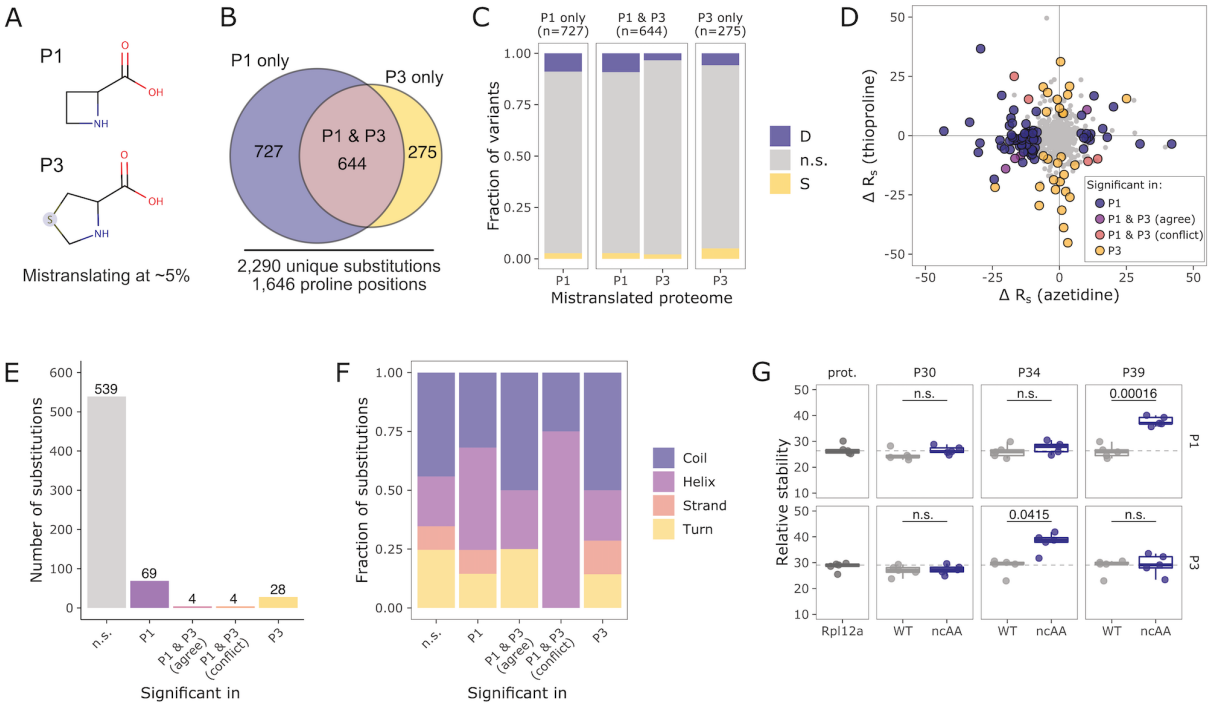


Figure 3.5. Residue sensitivity is specific to the ncAA substitution.

(A) Structures of azetidine (P1) and thioproline (P3). (B) Overlap of proline sites where we quantified the effects of P1 or P3 on protein thermal stability. (C) Fraction of substitutions with a significant effect on protein thermal stability, and whether the sensitivity of these proline residues was quantified P1 only, P3 only, or in both mistranslated proteomes. (D) Correlation of the effects of P1 or P3 mistranslation at the overlapping 644 sites shared between each mistranslated proteome. (E) Annotated effects of the overlapping substitutions as sensitive in P1, both P1 and P3 and same directional change, both P1 and P3 but different directional change, or just in P3. (F) Relationship between residue sensitivity and secondary structure assignment of proline residue. (G) Example of proline residue sensitivity across P1 and P3 in the ribosomal protein Rpl12a.

However, incorporation of thioproline at Pro34 significantly increased Rpl12a stability, while azetidine had no effect (Figure 3.5G). Conversely, incorporation of azetidine at Pro39 significantly increased Rpl12a stability, while thioproline had no effect (Figure 3.5G). Interestingly, a neighboring phosphorylated serine (pSer38) plays an important regulatory role in translation (Imami et al. 2018) and increases protein thermal stability (Smith et al. 2021). These two findings for pSer38, in combination with azetidine sensitivity of Pro39 and the known molecular consequence of azetidine substitutions (effects on size/trans isomerization), suggests cis/trans isomerization may play a mechanistic role in the regulatory function of Ser39 in Rpl12a function. Taken together, these data highlight how using multiple ncAAs to mistranslate at the same residues can yield additional insights related to residue function and residue sensitivity.

Clusters of sensitive residues in protein structures

Lastly, given the structural and functional correlations with ncAA sensitivity, we asked whether there were any clusters of sensitive amino acid residues in protein structures across the proteome, and whether these also correlate with protein function. To do this, we developed a statistical pipeline that “scans” proteins for “thermally-sensitive regions” (TSRs), which we defined as clusters of amino acids within 5 Å of each other that have a significantly higher absolute ΔR_s than what is expected by chance. Residues that fell into these clusters were then labeled as TSR residues, whereas other residues were defined as non-TSR residues. We also labeled residues as “not tested” if fewer than two ncAA substitutions were quantified.

We focused our analysis on yeast proteins where at least three positions were significantly affected by ncAA substitutions. Applying this filter trimmed the dataset down to 77 yeast proteins and 2,629 unique positions where we quantified the effects of at least one ncAA substitution (note, we excluded all double mutants from this analysis). For any residues with multiple ncAA

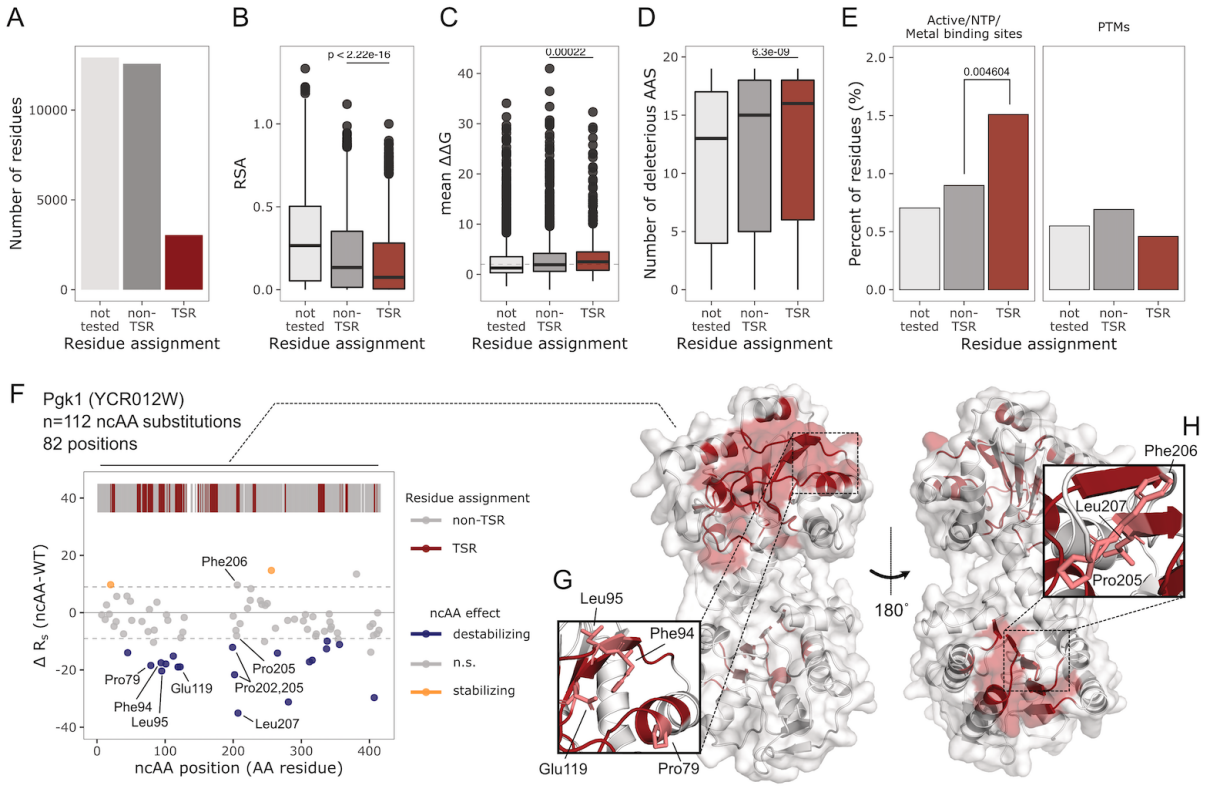


Figure 3.6. Spatial clusters of sensitive residues correlate with structure and function.

(A) Number of residues across 43 proteins that are assigned to thermally-sensitive regions (TSRs), non-TSRs, or that were not compatible with our statistical pipeline (not tested). Correlation between residue assignment and (B) relative solvent accessibility (RSA) (C) predicted changes in stability by natural amino acid substitutions (D) number of deleterious substitutions according to SIFT, and (E) significant enrichments of functional residues within TSRs. Function annotations downloaded from Uniprot (F) Substitution sensitivity map for Pgk1. Each point represents the effect of a ncAA substitution at a specific position within Pgk1. The color represents the effect of that ncAA on Pgk1 stability. The lines at the top of the plot represent residue assignment, i.e., TSR or non-TSR. Residues without assignments did not pass our filters for statistical testing. (G) cluster of destabilizing substitutions in the N-terminal domain of Pgk1. (H) Cluster of destabilizing substitutions in the “hinge” region of Pgk1.

substitutions mapping to those sites, we took the substitution and ΔR_s values that had the most statistically-significant effect on protein stability. We then applied a final filtering criterion of requiring proteins to have multiple residues in TSRs and non-TSRs. After this filter, we applied our pipeline to 28,542 residues, 15,619 of which passed filters for statistical testing (minimum of two residues where we quantified an ncAA substitution). Out of these, a total of 3,043 were assigned to TSRs (BH-adjusted p-value < 0.05, permutation test for ΔR_s of cluster) (Figure 3.6A).

We next asked whether TSR residues were reflective of structural or functional features of proteins. Our analysis revealed significant correlations between TSR residues and solvent accessibility (Figure 3.6B, p-value < 2.22×10^{-16}, Wilcoxon-rank sum test), predicted $\Delta\Delta G$ of all 19 natural substitutions (FIgure 3.6C, p-value = 0.00022, Wilcoxon-rank sum test), and the overall predicted deleteriousness of natural substitutions (Figure 3.6D, p-value = 6.3×10^{-9}, Wilcoxon-rank sum test). These data support the notion that residues sensitive to ncAA substitutions are part of larger networks of important protein regions. We also found a modest but significant enrichment of TSR residues known to be important for certain types of protein functions. For example, there was an enrichment of residues important for small molecule binding, active sites, metal binding, and binding to nucleotide triphosphates (Figure 3.6E, p-value = 0.004604, Fisher's Exact Test). Conversely, there was no difference in PTMs located at TSR residues versus non-TSR residues. These data suggest that clusters of sensitive residues may also pinpoint regions of function within proteins.

An example of this type of clustering can be found in the glycolytic enzyme, phosphoglycerate kinase 1 (Pgk1) (Figure 3.6F). We quantified the effect of 112 ncAA substitutions (92 single and 20 double substitutions) at 82 different positions in Pgk1, 27 of which significantly altered Pgk1 stability (~24%). While the substitution sensitivity map yields insights into individual positions

with particular sensitivity to ncAAs (Figure 3.6F), integrating ncAA sensitivity and TSR residues reveals functional regions within Pgk1. For example, several TSR residues reside in the “hinge” region of Pgk1, which is important for structural motions during catalysis (Figure 3.6H). Additionally, these TSRs may shed light on regions of proteins that may have novel structural or functional roles. For example, two surface-exposed clusters in the N-terminal domain are sensitive to several ncAAs. There are no known “functions” for these regions.

3.4 Discussion

Sequence-function relationships frame our understanding of genome and proteome function. Methods that probe these relationships at scale remain limited, hampering our mechanistic interpretation of genetic variation and post-translational modifications, with implications for clinical genetics, molecular evolution, and protein engineering. Here, we applied a high-throughput thermal stability assay on eight different mistranslated proteomes, revealing correlations between stability-altering ncAA substitutions and protein structure and function for hundreds of proteins in yeast.

By coupling the same functional selection with multiple mistranslated proteomes, we could begin to make generalizations about the basis for ncAA sensitivity. First and foremost, these data support the fundamental idea that proteins are sensitive to amino acid substitutions, regardless of whether those substitutions are of natural origin (cognate) or unnatural (ncAAs). We quantified hundreds of amino acid residues in yeast proteins that were inhospitable to amino acid change, even for the most “modest” of substitutions, such as fluorotryptophan (W1) (Figure 3.2).

Second, the effect of a ncAA substitution (as stabilizing vs destabilizing) is, in and of itself, informative to the molecular phenotype underlying the substitution's effect. We observed strong correlations between site-specific sensitivity of residues to ncAAs and structural features, such as solvent accessibility, predicted $\Delta\Delta G$, and residue centrality within protein structures, all of which are paralleled by observations with cognate substitutions (Figure 3.3). Importantly, these correlations were strongest for destabilizing substitutions, indicating that decreased protein thermal stability is a universal feature of amino acid substitutions, regardless of their natural or unnatural origins. However, we found several examples of stabilizing substitutions, many of which did not correlate with protein structural features. Many of these stabilizing substitutions may pinpoint residues with strong functional importance (Figure 3.4).

Third, different ncAA substitutions at the same residue can elicit different effects on stability, very much like natural substitutions (Figure 3.5). In the context of this study, this result suggests most of the residues that were tolerant to ncAA substitutions may, in-fact, be sensitive to other substitutions or molecular contexts that we did not explore in this study. Furthermore, this results suggests we could use Miro with multiple, structurally-similar ncAAs that replace the same residue to probe targeted questions about residue function, e.g. what are the effects of withdrawing increasingly more electron density from tyrosine residues with more fluorination? Future studies that generate mistranslated proteomes with different substitutions occurring at the same residues will yield additional molecular insights that have the potential to generalize across proteins.

Fourth, integrating positional sensitivity across different mistranslated proteomes is informative for elucidating protein function. As we show with Pgk1, despite quantifying ncAA sensitivity across separate mistranslated proteomes and experiments, mapping their effects back to protein sequence and structure reveals clustering in 3-dimensional space. Furthermore, these

clusters illuminate residues known to be functional, such as the residues within the hinge region of Pgk1. Spatial clustering of sensitive residues within a protein lines up with prior observations of mutational hotspots and PTM hotspots in regulatory regions of proteins (K.-L. Huang et al. 2021; Beltrao et al. 2012).

The mistranslated proteomes we quantified here represent a small fraction of the total possible substitution space. Part of this limited coverage is driven by constraints on solubility; the modified PISA workflow we developed and implemented across mistranslated proteomes was performed on native lysates. Porting these methods and systems into mammalian cells, which are more amenable to in-cell thermal stability assays, may increase overall coverage. Advances in mass spectrometry instrumentation and sample acquisition will also improve variant coverage.

Lastly, we observed that only a small fraction of quantified substitutions impacted protein thermal stability. While one explanation is that most residues are tolerant to ncAAs, an alternative view is that residues may have sensitivities under specific contexts that are missing in a lysate, for example, being in the presence of a small molecule or metabolite, or binding partner. We envision these “tolerant” mistranslated proteomes are actually a benefit for Miro, and open up applications coupling modified PISA with additional spike-ins or modifications. For example, PISA and thermal proteome profiling have already been shown to be sensitive to protein-protein, protein-RNA, and protein-small molecule interactions. We envision applying these same modifications to large collections of protein variants to reveal context-specific residue sensitivities, for example, residues involved in small molecule binding.

Mapping sequence-function relationships for all the proteins encoded within genomes remains a monumental task, yet all the more important with the vast wealth of genomic and proteomic

information accumulating in the last decade. Porting Miro into mammalian systems will begin closing the gap on contextualizing the effects of substitutions at these sites in human proteins, with potential impacts in clinical genetics. We envision the tools and methods presented here to be highly modular and generalizable to a variety of systems and proteins to map protein sequence-function relationships at scale.

3.5 Methods

3.5.1 Yeast strains

All experiments were done with the *Saccharomyces cerevisiae* haploid strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0), a direct descendant from FY2, which is itself a direct descendant of S288C. The only exception was for the R1 mistranslated proteome, which were done in the *Saccharomyces cerevisiae* haploid strain YMD1914, an arginine auxotroph.

3.5.2 Establishing dose-dependent toxicity and mistranslation in yeast

Cultures of BY4741 were grown overnight at 30˚C in synthetic complete media (SCM) containing 6.7 g/L yeast nitrogen base, 2 g/L of synthetic complete mix minus lysine, 2% glucose. Cultures of YMD1914 were grown overnight in the same conditions as above, but with the addition of 150 mg/L of proline and 75 mg/L of isotopically-heavy arginine (Arg6). These cultures were used to seed fresh 60 mL cultures in matching fresh media to an OD of 0.1, and grown for two additional doublings. When OD of the cultures reached 0.4, cultures were diluted 1:1 in a 96-well plate containing 100 µl of fresh matching media containing a ncAA of interest at

one of eight concentrations (Table 1). Optical density of cultures was monitored for 18 hours in a temperature-controlled plate reader (BioTek) measuring OD600 every 15 minutes.

Toxicity of each ncAA was assessed by comparing the area under the growth curve (AUC), determined by summing the average OD measurements across the 18-24 hour time window for each analog and concentration and normalizing relative to the set of control growth curves in each plate. We established dose-response relationships by comparing relative AUC for each analog and qualitatively selecting 3-5 concentrations around the IC50 for downstream incorporation analysis by mass spec.

To assess incorporation, we grew single replicates of yeast in 20 mL of synthetic complete media at 30˚C. We specifically used an overnight culture for each strain grown in the same conditions mentioned above and used these cultures to seed 20 mL fresh strain-specific median (see paragraphs above) and grew at 30˚C to a starting OD of 0.3, at which point cultures of yeast were diluted 1:1 with fresh media containing heavy lysine and an analog of interest. Cultures were harvested at OD 1.0 by centrifuging cultures at 8,000 x g and 4°C for 10 min. Cell pellets were resuspended in ice-cold sterile water, pelleted at 21,000 x g and 4°C for 10 min, snap-frozen in liquid nitrogen, and stored at -80°C until cell lysis and sample preparation. One biological replicate was collected for each analog and concentration tested. For each analog, we estimated site-level incorporation by dividing the MS1 signal intensity associated with a ncAA-containing peptide by the MS1 signal intensity of the corresponding wildtype peptide. To model the relationship between analog concentration and proteome-wide mistranslation, we fit a linear model that incorporates analog concentrations and the median observed site-level incorporation frequency. This model was then used to assess concentrations for generating the mistranslated proteomes for downstream selection with modified PISA (Table 1).

Table 3.1. Concentrations of ncAAs tested for toxicity and mistranslation

Non-canonical amino acid	Toxicity screen (mg/L)	Mistranslation screen (mg/L)	Final [ncAA]
4-Fluoroglutamate	5000 2500 1250 625 312.5 156.25 78.125 39.0625	650 325 162.5 81.25 40.625	300 mg/L
Fluorophenylalanine	10000 5000 2500 1250 625 312.5 156.25 78.125	160 80 40 20 10	16 mg/L
Trifluoroleucine	5000 2500 1250 625 312.5 156.25 78.125 39.0625	2500 1250 625 312.5 156.25	80 mg/L
Azetidine-2-carboxylic acid	N/A	500	100 mg/L
Thioproline	10000 5000 2500 1250 625 312.5 156.25 78.125	10000 5000 2500	7.5 g/L
Canavanine	600 300 150 75	180 120 80 54	75 mg/L

	37.5 18.75 9.375 4.6875	36	
5-fluorotryptophan	5000 2500 1250 625 312.5 156.25 78.125 39.0625	1250 416.67 138.89 46.3 15.43	420 mg/L
Fluorotyrosine	5000 2500 1250 625 312.5 156.25 78.125 39.0625	650 216.67 72.22 24.07 8.02	100 mg/L

3.5.3 Generating mistranslated proteomes

Cultures of BY4741 were grown overnight at 30˚C in SCM containing 6.7 g/L yeast nitrogen base, 2 g/L of synthetic complete mix minus lysine, 2% glucose. Cultures of YMD1914 were grown overnight at 30˚C in SCM containing 6.7 g/L yeast nitrogen base, 2 g/L of synthetic complete mix minus lysine, minus arginine, 75 mg/L isotopically-heavy arginine (Arg6), 150 mg/L proline, 2% glucose. These cultures were used to seed six fresh 60 mL cultures of the same media composition for each strain at OD_{600} 0.025, which were grown at 30˚C. When cultures reached OD_{600} 0.15, isotopically-heavy lysine was supplemented at 0.872 mM final concentration, along with canavanine at the concentration listed below (Table 1).

All cultures were harvested at OD_{600} ~ 1 by centrifugation in 50 mL Falcon tubes at 8,000 × g, 4˚C, for 10 min. Supernatant was decanted and yeast pellets were washed by resuspension in 1 mL ice-cold sterile water and centrifugation in 2 mL screw cap tubes at 21,000 × g, 4˚C, for 10 min. Yeast pellets were washed one more time by resuspension in 1 mL ice-cold sterile water, split into two separate 2 mL screw cap tubes (500 µL each) at 21,000 × g, 4˚C, for 10 min; supernatant was decanted and pellets were snap-frozen in liquid nitrogen and stored at −80˚C.

3.5.4 Thermal denaturation of mistranslated proteomes

Frozen yeast cell pellets were resuspended in 400 µL of non-denaturing lysis buffer (25 mM HEPES pH 7.5, 75 mM NaCl) containing 0.25x protease inhibitors (Pierce) on ice. Cells were lysed by bead beating with 0.5 mm 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 × g for 1 min to remove the beads and then at 21,000 × g for 10 min at 4°C to remove cell debris. To bring all protein extracts to the same concentration, a BCA was performed and an additional amount of lysis buffer was added to bring protein extracts to 2.5 mg/mL.

For each biological replicate, cell extracts were aliquoted into two strips of PCR tubes (1×8 for the temperature gradient and 1×8 for the 30°C) dispensing 20 µL of protein extract per tube. All samples were initially equilibrated to 30°C for 5 min. Temperature gradient samples were subjected to 46.0°C, 47.2°C, 48.7°C, 50.3°C, 51.7°C, 53.2°C, 54.8°C, and 56.0°C, one tube to each temperature, for 5 min. In parallel, controls were subjected to an additional 30.0°C temperature treatment for 5 min. All samples were cooled down to room temperature for 5 min. For each replicate, temperature gradient samples were all pooled into one tube and 30°C controls were pooled into a separate tube prior to centrifugation at 17,100 × g for 60 min at 4°C. The soluble protein fraction for the temperature gradient and 30.0˚C controls were combined 1:1

with denaturing lysis buffer (25 mM HEPES pH 8.9, 75 mM NaCl, 9 M Urea). The 30.0˚C controls were then diluted one more time, this time by combining 1:1 with equal mixture of non-denaturing lysis buffer and denaturing lysis buffer (with buffer composition around 25 mM HEPES pH 8.2, 75 mM NaCl, 4.5 M Urea). Protein concentration was measured with a BCA assay.

3.5.5 Proteomics sample preparation and desalting

Protein samples were reduced with 5 mM dithiothreitol (DTT) for 30 min at 55°C, followed by alkylation with 15 mM iodoacetamide for 30 min at room temperature in the dark. The alkylation reaction was quenched with 5 mM DTT at room temperature for 15 min. The pH was adjusted to 8.5 with 1 M HEPES pH 8.9 and 100 µg of sample was digested overnight with Lysyl endopeptidase (LysC; Wako Chemicals) at a 1:50 enzyme:substrate ratio. Digestion was quenched with 10% trifluoroacetic acid (TFA) to a final concentration of 1.5% and pH ~2-3.

Peptide samples were desalted by solid-phase extraction over 2 mg Oasis HLB 96-well µElution Plates using vacuum pressure. Packing material was conditioned with 200 µL methanol, 3 × 200 µL 100% acetonitrile, 200 µL 75% acetonitrile, 0.5% acetic acid, 200 µL 50% acetonitrile, 0.5% acetic acid, 3 × 200 µL of 0.1% TFA. Peptides were then loaded under low vacuum pressure, washed with 3 × 200 µL 0.1% TFA and 200 µl 0.5% acetic acid. Peptides were eluted into PCR tubes with 50 µl of 50% acetonitrile, 0.5% acetic acid and 50µL 75% acetonitrile, 0.5% acetic acid, and aliquoted into 15 µg aliquots for downstream labeling. All samples were lyophilized by vacuum centrifugation and stored at −80°C before labeling with tandem mass tags (TMTs).

3.5.6 Peptide labeling and desalting

Peptide aliquots (20 µg) were resuspended in 20 µl of 100 mM HEPES pH 8.2, 30% acetonitrile, and labeled with 5 µl of TMT10plex solution (equivalent to 50 µg of TMT10plex). Reactions were

incubated at room temperature for 1 hour followed by a 15 min quench with 2.5 µl of 5% hydroxylamine solution followed at room temperature. Samples were then pooled and acidified to pH 2-3 using a final concentration of 1% TFA. Excess acetonitrile was removed by brief vacuum centrifugation followed by desalting over a 50 mg Sep-Pak tC_{18} cartridge (Waters). Packing material was washed with 1 mL methanol, 3 × 1 mL 100% acetonitrile, 1 mL 75% acetonitrile, 0.5% acetic acid, 1 mL 50% acetonitrile, 0.5% acetic acid, and equilibrated with 3 × 1 mL 0.1% TFA. Peptides were then loaded by gravity, washed with 3 × 1 mL 0.1% TFA and 1 mL 0.5% acetic acid. Peptides were eluted with 750 µL of 50% acetonitrile, 0.5% acetic acid and 750 µL 75% acetonitrile, 0.5% acetic acid. Samples were vortexed, dried by vacuum centrifugation, and stored at −80°C before offline peptide fractionation.

3.5.7 Peptide fractionation

Peptides were fractionated by pentafluorophenyl reversed-phase fractionation (Grassetti, Hards, and Gerber 2017) using a Waters XSelect HSS PFP 2.5 µm 2.1 x 150 mm column. Approximately 100 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.

3.5.8 Mass spectrometry data acquisition

3.5.8.1 Data-dependent acquisition of dose-dependent mistranslated yeast proteomes

Lyophilized peptides were resuspended in 5% ACN, 5% formic acid and, using either an EASY-nLC (1000 and 1200 series; ThermoFisher Scientific), loaded on a 100 µm x 3 cm trap column packed with 3 µm C18 beads (Dr Maisch), and separated on a 50˚C-heated 35 cm analytical column packed with 1.9 µm C18 beads (Dr. Maisch) using a 90-minute gradient of 80% acetonitrile, 0.1% formic acid. Mass spectra were collected on either a Q-Exactive mass spectrometer (Thermo Fisher Scientific) or Orbitrap Eclipse mass spectrometer (ThermoFisher Scientific) using a data dependent acquisition method. For each cycle a full MS scan (300-1500 m/z, resolution 70,000, AGC target 3e6) was followed by MS/MS scans (isolation width 2.0 Da, 26% normalized collision energy, resolution 17,500, AGC target 5e4) on the top 20 most intense precursor peaks.

3.5.8.2 Data-dependent acquisition of temperature-treated mistranslated proteomes

Lyophilized TMT-labeled peptide samples were resuspended in 5% ACN, 5% formic acid and subjected to liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Fractions of TMT-labeled samples (0.5 µg - 1 µg) were collected on an Orbitrap Eclipse mass spectrometer (ThermoFisher Scientific) coupled to an EASY-nLC 1200 LC (ThermoFisher Scientific). Peptides were loaded on a 100 µm x 3 cm trap column packed with 3 µm C18 beads (Dr. Maisch) and separated on a 50˚C-heated 35 cm analytical column packed with 1.9 µm C18 beads (Dr. Maisch) using a 90-minute gradient of 80% acetonitrile, 0.1% formic acid. Peptides were analyzed online using data-dependent acquisition using an MS3-based method with real time search and 3-second cycle time. First, MS1 data were collected using the Orbitrap (120,000 resolution; maximum injection time 50 ms; AGC 4e5, mass range 400-1600 m/z, charge states 2-6, dynamic exclusion 30 seconds). MS2 scans of the most intense precursors were performed in the ion trap with CID fragmentation (isolation window 0.8 Da, rapid, NCE 35%, maximum injection time 50 ms; AGC 1e4). An online real-time search algorithm

(Schweppe et al. 2020) was used to search spectra against a yeast FASTA database (static modifications: TMT10plex on N-terminus and lysines; variable modifications: mass shift associated with ncAA, heavy lysine, methionine oxidation) with maximum 1 missed cleavage and 3 variable modifications. MS2 scans that passed custom RTS thresholds (Charge 2 precursors: Xcorr > 0.75; Charge 3+ precursors: Xcorr > 1.25; all charge states: PPM error < 20, dCn > 0.1) were sent for MS3 quantification. The top 10 matching fragment ions were isolated using synchronous precursor selection (McAlister et al. 2014) followed by HCD fragmentation and collected for an MS3 scan in the Orbitrap (resolution of 60,000, NCE of 55%, maximum injection time of 120 ms, isolation window 0.8 Da, and AGC of 1e5). Dynamic exclusion was enabled to exclude fragmented precursors from repeated MS/MS selection for 60 sec.

3.5.9 Mass spectrometry data processing

Raw files were converted to mzML formats using msconvert, and MS/MS spectra were searched against a target/decoy protein sequence database using Comet (v2019.01.02)(Eng, Jahan, and Hoopmann 2013). *Saccharomyces cerevisiae* (orf_trans_all.fasta downloaded from the Saccharomyces Genome Database in 2014). 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 0.6 Da tolerance with 0.4 Da offset for MS/MS acquired on a linear ion trap mass analyzer. For all data collected using real time search, ion trap mass tolerances were set to match the RTS version, i.e., in this case 0.6 Da tolerance with 0.4 Da offset. 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 (Table 2). For instance, variable

modification of -14.0156 Da on proline residues reported for the substitution of proline with azetidine-2-carboxylic acid. Dynamic SILAC samples were searched with light lysine (K0) and heavy (K8, +8.0142 Da) variable modifications in binary mode. TMT-labeled samples were searched with constant modification (+229.1629 Da) on lysines and peptide N-termini. Search results were filtered with Percolator (Percolator version 3.01) (Käll et al. 2007) 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.

Table 3.2. Mass shifts searched in offline comet and real time search for in each ncAA

Non-canonical amino acid	Compound I.D.	Variable modification mass shift
4-Fluoro-DL-glutamate	E1	+17.99057813134
Fluorophenylalanine	F2	+17.99057813134
Trifluoroleucine	L4	+53.97173439402
Azetidine-2-carboxylic acid	P1	-14.01565006452
Thioproline	P3	+17.95642111038
Canavanine	R1	-4.04086445552
5-fluorotryptophan	W1	+17.99057813134
Fluorotyrosine	Y1	+17.99057813134

3.5.10 Quantifying peptide- and protein-level relative stability

For each mistranslated proteome, all scan-level TMT reporter ion intensities were first corrected for isotopic difference, followed by correcting for differences in sample loading by summing reporter ion intensities across each of the ten TMT channels and applying correction factors so that summed intensities were the same across all channels. Importantly, we normalized the 30˚C and temperature gradient channels separately, since the proteomic composition was

known to be different between the 30˚C samples and the temperature gradient-treated samples. We also removed a subset of scans containing reporter ions with reproducibly-high variance that was completely agnostic to the peptide or protein identifications and quantifications. Specifically, we applied a peptide and protein agnostic percentile cutoff for each mistranslated proteome, removing scans with CVs for either the 30˚C or the temperature gradient samples in the top 1%. Lastly, to account for potential systematic differences in the distribution of relative stability, a correction factor was applied to each mistranslated proteome such that the dominant peak of non-melting proteins all aligned at an R_s value equal to 100%. The relative stability for each peptide was quantified using the equation below. Lastly, protein relative stability was determined as the median R_s value for all peptides quantified for that protein, with a requirement of at least two unique peptides quantified per protein.

$$R_s = \frac{Reporter\ ion\ intensity_{temperature\ gradient}}{Reporter\ ion\ intensity_{30^{\circ}C\ control} \times 2} \times 100$$

3.5.11 Identifying stability-altering substitutions

To identify substitutions that significantly alter protein thermal stability, we compared the R_s values of ncAA-containing peptides against wildtype peptides using a paired t-test in R and limma. Specifically, we first required that all comparisons were between fully tryptic peptides containing a heavy lysine. Second, to consolidate from peptide-level to site-level quantifications, we required that all versions of peptide mapping to the same site be seen in the ncAA and wildtype form (i.e. if one of the versions is a methionine oxidized peptide, then the oxidized version must have been quantified for both the ncAA and wildtype peptide). Then, after ensuring overlap between ncAA and wildtype peptides, we collapsed peptides into site-level quantifications by taking the median R_s value for each replicate. We then compared the ncAA R_s

value with the wildtype R_s value using a paired t-test in R with the limma package, accounting for mean-variance relationships. All p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg correction. Statistical testing, p-values, and p-value corrections were done separately for each mistranslated to control for sample-specific effects on protein stability. Lastly, to determine an effect size cutoff, we required that ΔR_s of ncAA substitutions be greater than the median standard deviation ΔR_s across mistranslated proteomes between wildtype peptides and their corresponding protein-level R_s values.

3.5.12 Bioinformatics

We assessed the structural and evolutionary context for all residues where we detected a ncAA substitution using a variety of computational tools. For all analyses, we only used peptides containing a single ncAA substitution.

For structural analysis, we downloaded PDB files of predicted protein structures from AlphaFold (Jumper et al. 2021) (downloaded 2021-09-21). We extract solvent accessibility for each residue using DSSP (Kabsch and Sander 1983) with the Bio3d package in R (Grant et al. 2006). We calculated relative solvent accessibility by dividing the solvent accessible surface area (SASA) output from DSSP (Kabsch and Sander 1983) by a list of maximum SASA per residue (Tien et al. 2013). Prediction scores were extracted from each PDB file, and residues with prediction scores less than 70 were excluded from our analysis. Residues with omega bond angles between -30˚ and 30˚ were considered to reside in cis. Disordered predictions for each residue in the yeast proteome were downloaded from http://bioinfadmin.cs.ucl.ac.uk/disodb/. A disordered region was defined as a segment of 30 or more consecutive residues in a protein with a prediction of being disordered.

In order to compare the effects of natural amino acid substitutions with ncAA substitutions, we downloaded the entire *in silico* mutagenesis dataset (both the homology-based and

experimental models) and all SIFT scores for the yeast proteome from the mutfunc database (Wagih et al. 2018). We then calculated a mean ΔΔG and summed the number of “deleterious” natural substitutions (SIFT score < 0.05) for each residue and used those values as representative of the overall mutational sensitivity of a residue.

To identify thermally-sensitive regions, we developed a statistical pipeline that took the following approach: (1) first, we established a null distribution of ΔR_s values for each protein by generating 10000 random permutations of the observed ΔR_s for each replicate, shuffling the randomizing position ΔR_s. Importantly, we took the absolute value of ΔR_s at this point before proceeding further. We then calculated the median permuted ΔR_s for each position and each permutation. (2) Next, we performed an *in silico* “scan” for each protein by moving along primary amino acid sequence, one residue at a time, and extracting all neighboring residues within 5 Å (which included the residue of interest) where we quantified the effect of a ncAA substitution. We defined this set of neighboring residues as a “cluster”. We then calculated a “cluster”-specific ΔR_s for both the observed dataset and the 10000 permutations by taking the median absolute ΔR_s value for all observed residues surrounding that amino acid position. (3) We calculated a cluster-specific p-value based on this empirical distribution, which was defined as:

$$p = \frac{[Number\ of\ instances\ abs(\Delta Rs_{observed}) < abs(\Delta Rs_{permuted})] + 1}{10000 + 1}$$

All p-values within each protein were corrected for multiple-hypothesis testing using the Benjamini-Hochberg method in R. An important note - *before* p-value correction, we removed any clusters that were composed of one or fewer ncAA substitutions (i.e. we required a minimum of two observed ΔR_s).

Lastly, clusters of residues that had a significantly higher absolute ΔR_s than what was observed by chance (BH-corrected empirical p-value < 0.05) were all classified as “TSR” residues. Any residues that did not fall into these clusters but were quantified by at least two ncAA substitutions were classified as “non-TSR” residues. All remaining residues were classified as “not tested”.

3.5.13 Data availability

All raw data used to generate the analyses presented in this paper will be available on ProteomeXChange upon submission for publication.

3.5.14 Code availability

All R scripts to reproduce the figures and analysis generated here will be available on github upon submission for publication of this manuscript.

Chapter 4. Identifying residues important for small molecule binding

This Chapter is based on unpublished observations

4.1 Summary

Metabolites orchestrate a diverse array of molecular processes, from metabolic flux and transcription, to cell signaling. They exert their effects by binding and altering the activity of proteins. Despite the essential roles of metabolites in cell biology, their wide-spread interactions with proteins, and the residues mediating these interactions, remain enigmatic and biased towards a small subset of the proteome. Here, we present a novel approach to map protein-metabolite interactions across proteomes at single amino acid resolution. This approach involves first generating collections of mutant proteins using mistranslation with non-canonical amino acids. Lysates of these proteomes are then incubated with a small molecule of interest, followed by application of a high-throughput thermal stability assay to identify residues that lead to dose-dependent changes in protein stability. We optimize this method to probe protein-ATP interactions in yeast lysates and show that small molecule binding events can be detected in a proteome containing proteome-wide fluoroglutamate mistranslation. We use the effects of fluoroglutamate substitutions to identify glutamate residues that mediate direct and allosteric binding to ATP. We envision this approach being generalizable to other biological systems and workflows, enabling targeted biological questions, such as elucidating binding interfaces, or catalytic residues, to be explored proteome-wide and at single amino acid resolution.

4.2 Introduction

Protein functions are dynamically tuned by their interactions with metabolites. Through a variety of mechanisms, metabolites bind proteins and regulate metabolic flux (Gerhart and Pardee 1962), alter a protein's biomolecular interactions with DNA (Gilbert and Muller-Hill 1966) and other proteins (Piazza et al. 2018), and modulate transcriptional (Sellick and Reece 2003), enzymatic (Gerhart and Pardee 1962), and signaling activity (X. Li et al. 2010). These regulatory interactions inform the cell about nutrient supply, bioenergetic status, and metabolic

homeostasis, enabling real-time decisions regarding growth, proliferation, and survival. Despite their fundamental importance to cell and organismal biology, the amino acid residues mediating these interactions remain poorly understood.

Protein-metabolite interactions have traditionally been studied using a combination of biophysical and biochemical approaches, such as circular dichroism (Greenfield and Fasman 1969), nuclear magnetic resonance (Shuker et al. 1996), and differential scanning calorimetry (Jelesarov and Bosshard 1999). While tried-and-true, these methods are unable to scale to entire proteomes. More recently, mass spectrometry-based methods have emerged that capture protein-metabolite interactions by exploiting changes in protein structure upon metabolite binding. These methods have identified protein-metabolite interactions by measuring changes in protease susceptibility (Feng, De Franceschi, Kahraman, Soste, Melnik, Boersema, de Laureto, et al. 2014; Piazza et al. 2018), oxidation (Tran, Adhikari, and Fitzgerald 2014), and thermal stability (Savitski et al. 2014; Reinhard et al. 2015; Huber et al. 2015). These same principles have also been used to identify interactions between proteins in the same complex (Tan et al. 2018; Becher et al. 2018; Piazza et al. 2018). However, these methods have yet to pinpoint individual residues important for a protein-metabolite interaction.

We recently developed Miro, a proteomics method that maps sequence-function relationships at single amino acid resolution proteome-wide (Rodriguez-Mias et al. 2022). The method works by generating proteomes containing proteome-wide mistranslation with non-canonical amino acids (ncAAs) and then quantifying the effects of these ncAA substitutions using classic biochemical assays with a mass spectrometry-based readout. We recently coupled Miro with a high-throughput stability assay, quantifying the effects of close to 9000 amino acid substitutions in >700 yeast proteins. One surprising observation was that the majority of ncAA substitutions (>90%) did not alter protein thermal stability. However, the ability to generate a large collection

of protein variants that are only mildly altered in their stability opens the door to the possibility of a variety of secondary functional selections that can help “sensitize” or “unmask” the effect of a ncAA substitution under a specific condition.

Here, we showcase a novel extension of Miro that capitalizes on this idea of secondary selections to identify protein-small molecule interactions proteome-wide. Specifically, we couple a fluoroglutamate-mistranslated proteome with or without ATP, and detect binding events by measuring ATP-dependent changes in protein thermal stability. We first highlight the importance of temperature window selection when coupling our high-throughput stability assay with small molecules. We then go on to identify ATP-binding proteins in mistranslated proteomes by their ATP-induced stabilization. Lastly, we couple these assays to different concentrations of ATP to identify individual residues with concentration-dependent sensitivities. Some of these residues are known binding sites, but several are residues that exist outside ATP-binding pockets, highlighting the unique ability of Miro to potentially identify residues important for allosteric interactions. The method presented here will broaden our understanding of sequence and structural factors governing protein-metabolite interactions.

4.3 Results

Identifying ATP-sensitive yeast proteins with modified PISA

We recently developed a modified version of the proteome integral solubility assay (PISA) (Gaetani et al. 2019) that extends this stability assay to measure relative stability (R_s), a protein parameter akin to melting temperatures measured by Thermal Proteome Profiling (Savitski et al. 2014). This approach, which we will refer to here as modified PISA, can be used to assess changes in protein stability across conditions and proteins. Here, we sought to extend modified PISA to measure metabolite-induced changes in protein thermal stability.

We first set out to optimize modified PISA for detecting small molecule binding events in a lysate. We focused on narrow temperature windows for thermal denaturation to increase the sensitivity for metabolite-induced changes in protein stability (J. Li, Van Vranken, Paulo, et al. 2020) and benchmarked our approach using ATP, given its regulatory and catalytic roles, and widespread interactions with the yeast proteome. For thermal denaturation, we used four narrow temperature ranges (T1 through T5) between 48˚C and 58˚C (listed in Figure 4.1A, left panel, and Methods). Wildtype yeast lysates were incubated with either 1x PBS or 1x PBS + 10 mM ATP for 10 minutes before heating to either 30˚C or temperatures between 48˚C and 58˚C. After thermal denaturation, lysates were pooled, aggregates were discarded by centrifugation, and soluble fractions were processed for downstream analysis by mass spectrometry.

As expected, the global distribution protein R_s correlated with each temperature window, with a systematic shift towards lower R_s values as temperatures increased (Figure 4.1A, left panel). We also noticed a mild global stabilizing effect by 10 mM ATP (Figure 4.1A, right panel). A more targeted protein analysis revealed systematic stabilization of known nucleotide (ADP/ATP, GDP/GTP)-binding proteins by 10 mM ATP (Figure 4.1B). However, the degree of ATP-dependent stabilization varied across the five temperature windows (Figure 4.1C). For example, Actin (Act1) was completely stabilized by 10 mM ATP and resistant to thermal denaturation across all five temperature windows (Figure 4.1C, left panel). Conversely, ATP had no effect on aspartyl-tRNA synthetase (AspRS), as evident by the lack of precipitation within any of the temperature windows tested (Figure 4.1C, middle panel). Lastly, Hsp82 represents an example of many ATP-sensitive proteins for which the choice of temperature window (and likely the choice of ATP concentration) played an important role in detecting ATP-dependent stabilization (Figure 4.1C, right panel). At lower temperature windows (T1, T2), the observed difference in protein stability (ΔR_s) was minimal. However, at higher temperature windows

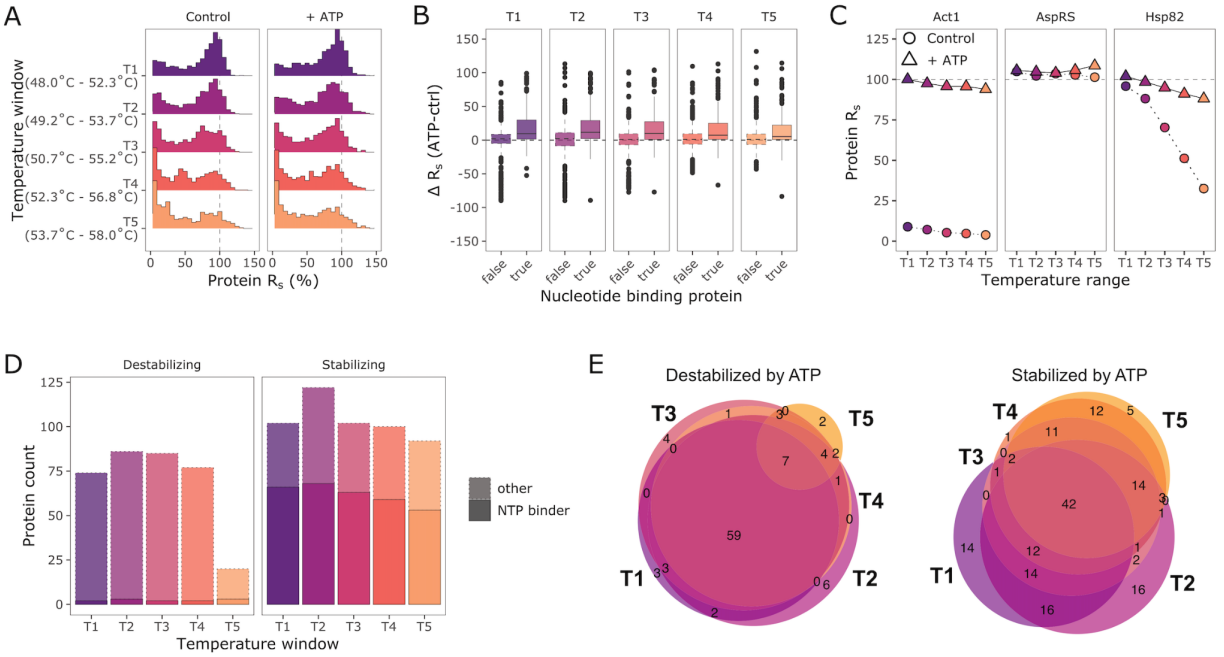


Figure 4.1. A modified high-throughput stability assay to probe the effects of metabolites on protein thermal stability.

(A) Distribution of protein relative stability (R_s) across five different temperature windows applied on yeast lysates with or without 10 mM ATP (n = 887 yeast proteins in each distribution). (B) Difference in protein R_s in the presence or absence of 10 mM ATP across five temperature windows. (C) Temperature window- and ATP-dependent protein R_s measurements for three example proteins that are known ATP-binding proteins. (D) Number of yeast proteins that are stabilized or destabilized in the presence of 10 mM ATP. Proteins were called as stabilized or destabilized if the absolute value of their measured ΔR_s was greater than 25. (E) Euler plot showing the overlap of proteins with altered stability across the five temperature windows tested here.

(T3, T4, T5), 10 mM had a substantial effect on Hsp82 thermal stability. These proteins highlight the importance of temperature windows in allowing the detection of small molecule-induced changes in protein stability.

Each temperature window tested here captured different subsets of the yeast proteome altered by 10 mM ATP. Because we acquired only one replicate across temperature windows and treatments, we applied a stringent qualitative cutoff to identify proteins that were “destabilized” or “stabilized” by ATP. Proteins with a $\Delta R_s < -25$ in the presence of 10 mM ATP were considered destabilized, and proteins with a $\Delta R_s > 25$ were considered stabilized. Proteins that were destabilized by 10 mM ATP (Figure 4.1D) were enriched for ribosomal proteins, whereas proteins stabilized by 10 mM ATP were enriched for nucleotide-binding proteins (Figure 4.1D). Each temperature window captured unique subsets of the yeast proteome stabilized by ATP (Figure 4.1E; right euler plot). This finding is in direct contrast to destabilized proteins, which were shared across the temperature windows (Figure 4.1E; left euler plot). Taken together, these data illustrate the importance of temperature selection in PISA when quantifying metabolite-induced changes in protein stability.

Identifying ATP-sensitive proteins in a mistranslated proteome

Next, we asked whether we could detect ATP-binding proteins in the context of a mistranslated proteome. To do this, we generated yeast proteomes containing proteome-wide fluoroglutamate substitutions at glutamate residues and coupled these proteomes with different concentrations of ATP (0 mM, 2 mM, or 10 mM) (Figure 4.2A). We selected fluoroglutamate due to (1) its ease of incorporation across the proteome; (2) its relatively mild effect on protein stability; and (3) the involvement in several yeast proteins of glutamate residues in interactions with ATP. Specifically, we generated four biological replicates of yeast exposed to fluoroglutamate and incubated

lysates of these samples with ATP (0 mM, 2 mM, or 10 mM) for 10 minutes, followed by heating at one of several temperatures between 51.5˚C and 55.5˚C. In addition to the temperature-treated samples, we treated the same lysate at 30˚C as a control with or without 10 mM ATP (i.e. the highest concentration tested here) to assess for indirect effects of ATP on protein stability.

After stringent filtering, we quantified ATP-dependent changes in stability for 694 yeast proteins (Figure 4.2A). We assessed whether the introduction of ATP resulted in a global dose-dependent shift in protein stability. At the 2 mM treatment, 50 proteins had a significant change in stability (7.2% of the quantified proteome) (Figure 4.2B), with a strong bias towards stabilizing effects compared to destabilizing (43 stabilized vs 7 destabilized) (Figure 4.2C). In the presence of 10 mM ATP, 215 proteins had a significant change in stability (31% of proteome) (Figure 4.2B), with the same bias towards stabilizing effects, although there was a larger fraction of protein that were destabilized by ATP at this concentration (129 stabilized vs 86 destabilized) (Figure 4.2C). In our control treatment (10 mM ATP in a 30˚C treated lysate), we identified 86 proteins with a significant change in “stability”. These proteins were strongly enriched for ribosomal proteins, which suggests that there is a significant pool of ribosomes that are insoluble at lower temperatures, as observed previously (Sridharan et al. 2019). Additionally, two proteins showed a significant decrease in stability in the control sample, one of which was Ndk1, a protein involved in transferring gamma phosphates from ATP to other nucleotide diphosphates (such as ADP or GDP).

For both the 2 mM and 10 mM ATP treatments, proteins stabilized by ATP were enriched for known ADP and ATP binding proteins (Figure 4.2C). By contrast, proteins destabilized by 10 mM were enriched for ribosomal proteins; however, this effect is likely non-specific given changes in solubility induced by 10 mM that we observed in the control (Figure 4.2C). Lastly, we

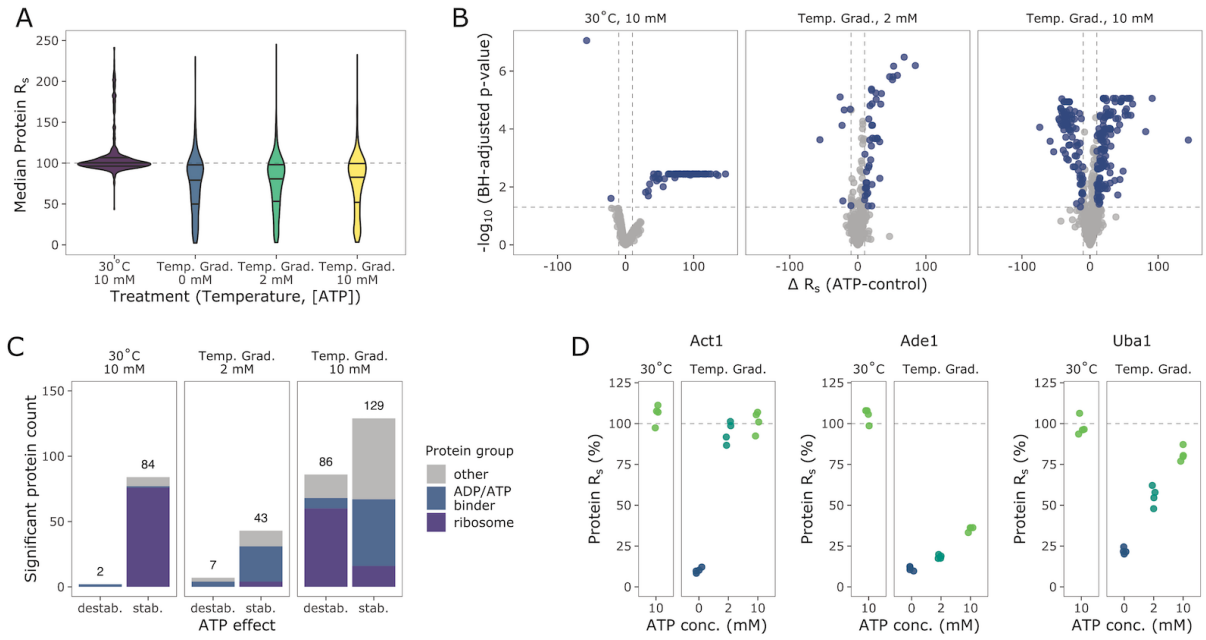


Figure 4.2. Detecting protein-metabolite interactions in a mistranslated proteome.

(A) Distribution of median protein stability (R_s) across different treatment conditions. (B) Volcano plots indicating protein stability differences with and without ATP and associated significance. (C) Total number of proteins with a significant change in stability and their assignment as ADP/ATP binders, ribosomal proteins, or other. (D) Examples of three proteins where ATP causes a shift in stability. All three are known ATP-binding proteins, but show different sensitivities to 2 mM and 10 mM ATP.

plotted three example dose-response curves for known ATP-binding proteins: Act1, Ade1, and Uba1 (Figure 4.2D). Similar to what we observed in wildtype yeast (Figure 4.1C), these proteins illustrate different sensitivities to ATP that were observevd across 5-fold differences in ATP concentration. These data also highlight the utility of probing changes in protein stability across different metabolite concentrations.

Mapping ATP-sensitive glutamate residues in Pgk1

Next, we asked whether we could identify ATP-binding sites within proteins. We compared wildtype peptides with ncAA-containing peptides, looking for differences in ATP-induced changes in protein thermal stability. As an initial benchmark within our dataset, we focused on glutamate residues in phosphoglycerate kinase (Pgk1). We focused on Pgk1 for several reasons. First, Pgk1 is an ADP and ATP-binding protein, and we observed ATP-induced stabilizing of Pgk1 in the presence of ATP. Second, Pgk1 is an abundant protein in yeast and potentially provides the best coverage of substitution effects. Lastly, at least one glutamic acid residue (Glu342) is known to be important for Pgk1's interaction with nucleotides.

In total, we quantified the effect of 22 fluoroglutamate substitutions on Pgk1 stability. In general, most residues were tolerant to fluoroglutamate, with only five substitutions showing a change in Pgk1 stability in the absence of ATP (Figure 4.3, heatmap). However, incubating these lysates with different concentrations of ATP unmasked an intolerance of several residues that were previously tolerant. Almost all of the observed changes were significantly destabilizing compared to their wildtype counterparts (Figure 4.3).

For example, we captured the peptide that spans the only glutamate known in Pgk1 to mediate binding with ATP (peptide: TIVWNGPPGVF**E**FEK, spanning Glu342) (Figure 4.3A). We

measured three different fluoroglutamate-containing versions of this peptide (single incorporation at Glu342, single incorporation at Glu344, and double incorporation at Glu342 and Glu344). In the absence of ATP, only the double substitution at Glu342,344 significantly destabilized Pgk1, while substitutions at Glu342 or Glu344 had no effect on Pgk1 stability. However, substitutions at Glu342 (both the single and double incorporated peptide) led to significantly less stabilization by 2 mM ATP compared to wildtype and single substitutions at Glu344, strongly suggesting that Glu342 is important for ATP binding. One possible explanation for this decreased stabilization is that adding a fluorine group to Glu342 (as fluoroglutamate) decreases the electron density to maintain a strong interaction with ATP (relative to wildtype). Of note, incorporation of fluoroglutamate at 342 or 344 did not impact overall Pgk1 stability (see the 0 mM condition), indicating that these sites are only moderately contributing to overall Pgk1 stability.

In addition to identifying metabolite-binding sites using ncAAs, we also identified sites outside of the known ATP-binding pocket in Pgk1 that are (1) important for Pgk1 stability; or (2) sensitive only at certain ATP concentrations. For example, fluoroglutamate substitutions at Glu356 destabilized Pgk1 but did not alter ATP-induced stabilization of Pgk1 (Figure 4.3B). Conversely, fluoroglutamate substitutions at Glu200 did not alter Pgk1 stability in the absence of ATP, but affected ATP-induced stabilization of Pgk1 (Figure 4.3C). While Glu200 is outside the nucleotide-binding pocket of Pgk1, this residue is within the “hinge” region, a critical molecular juncture for Pgk1 and its catalytic lifecycle. These data suggest that this glutamate residue may be important for mediating conformational changes resulting from ATP binding.

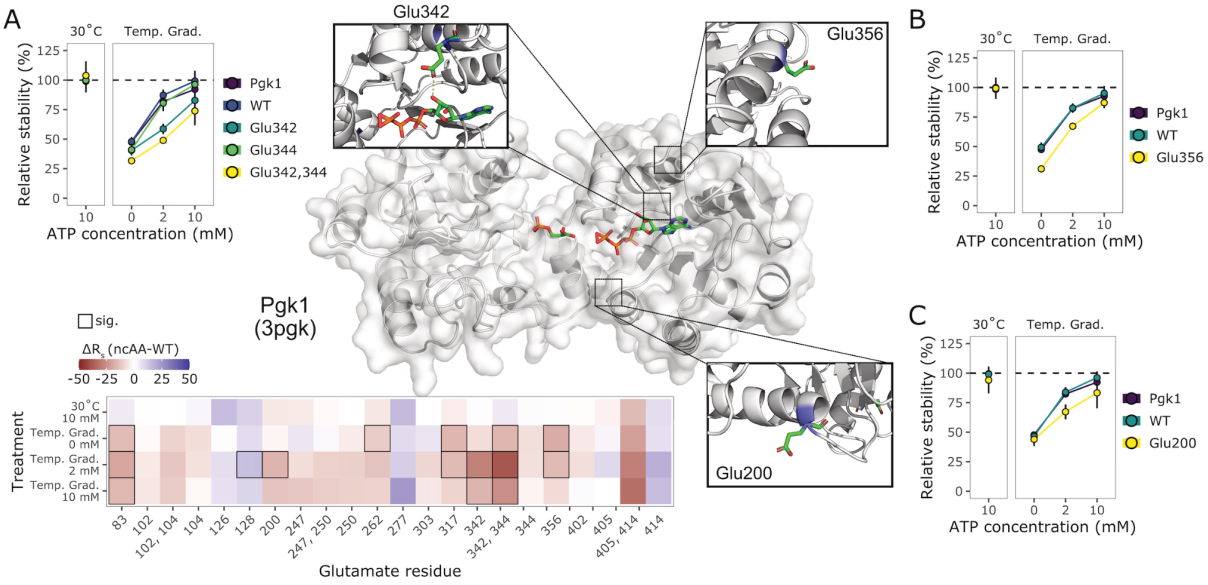


Figure 4.3. Mapping ATP-sensitive residues within phosphoglycerate kinase.

The effect of fluoroglutamate substitutions at (A) Glu342 and Glu344, (B) Glu356, and (C) Glu200 on Pgk1 stability and ATP-dependent stabilization of Pgk1. Bottom panel: heatmap showing change in Pgk1 stability across four treatment conditions. Fill color represents the ΔR_s relative to the matching wildtype peptides. Black borders represent sites where fluoroglutamate had a statistically significant effect on Pgk1 stability (BH-adjusted p-value < 0.05). Glutamate residues containing two residues listed are cases where we quantified the ncAA-containing peptide with fluoroglutamate at both positions.

Identifying ATP-sensitive residues proteome-wide

Lastly, we assessed the dose-dependent behavior of fluoroglutamate substitutions proteome-wide. After stringent filtering, we quantified dose-dependent change in stability for 434 fluoroglutamate substitutions across 116 yeast proteins, with 129 of these substitutions occurring in 29 known ADP- or ATP-binding proteins (Figure 4.4A). Globally, we quantified 108 different substitutions with a significant effect on protein thermal stability in at least one of the conditions tested (Figure 4.4C). Notably, the 30˚C control sample with 10 mM ATP was the only condition without any substitutions significantly altering stability (Figure 4.4B).

We systematically tracked changes in thermal stability for the 108 substitutions that had a significant effect in at least one condition (Figure 4.4C). A few patterns emerged from this kind of analysis. First, the addition of increasing concentrations of ATP to the lysate not only increased protein stability, but also reduced the number of substitutions that had a destabilizing effect on protein stability (Figure 4.4C, destabilizing substitutions). Conversely, stabilizing substitutions were less interpretable, with several substitutions having a stabilizing effect, but only at specific ATP concentrations (Figure 4.4C, stabilizing substitutions). Second, many substitutions had a significant effect on protein stability that was ATP-dependent, while only a small fraction of substitutions had a systematic effect on protein stability across all concentrations of ATP (Figure 4.4C, tracking non-significant substitutions).

A deeper look at three example substitutions in three different ATP-sensitive proteins (Figure 4.4D-F) reveals dose-dependent changes in protein stability that mirror some of the patterns we observed in Pgk1 (Figure 4.3). For example, we detected substitutions that significantly altered protein stability regardless of ATP concentration, such as fluoroglutamate incorporation at Glu178 in glycerol-3-phosphate phosphatase (Figure 4.4D), as shown previously in Chapter 3

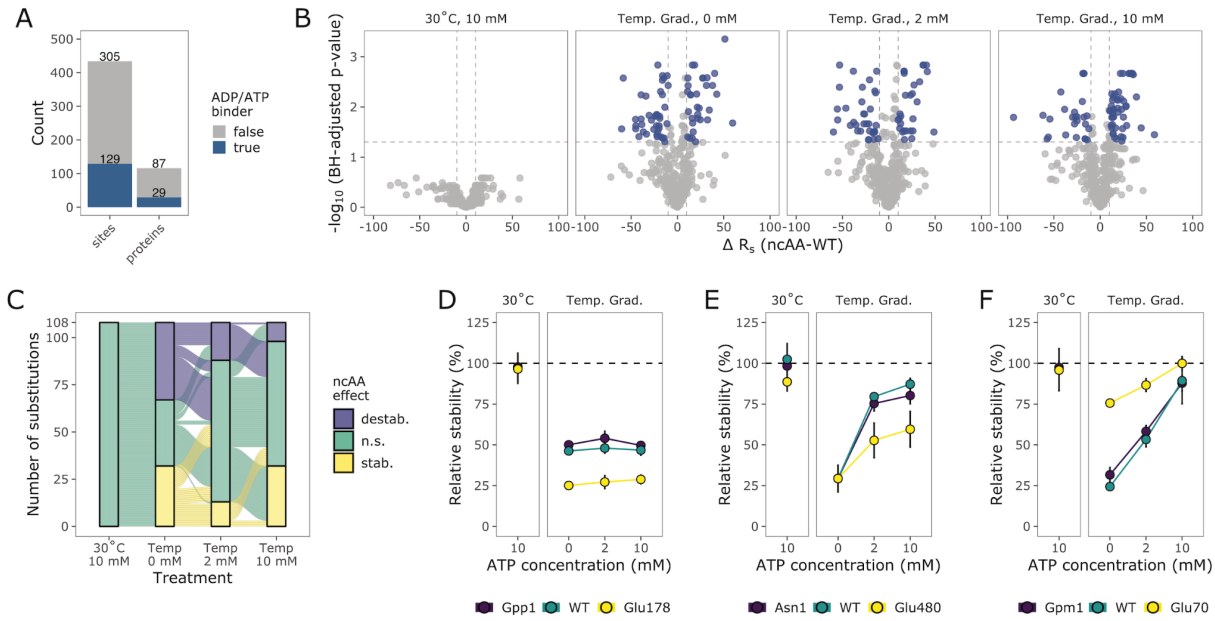


Figure 4.4. Proteome-wide discovery of ATP-sensitive glutamate residues.

(A) Total number of glutamate residues (sites) and proteins quantified in our assay. Fill color indicates residues found in ADP/ATP binding proteins. (B) Volcano plots illustrating the effect of fluoroglutamate substitutions on protein thermal stability. (C) Alluvial plot tracking the ATP-dependent effects of fluoroglutamate substitutions on protein thermal stability. Fill colors represent the effect of a substitution. The lines in between bar plots represent the change in substitution effect between the different conditions. (D-E) Three examples of fluoroglutamates substitutions within three different proteins that significantly alter protein stability in at least one of the tested ATP concentrations.

(Figure 3.3). We also detected substitutions that did not alter protein stability in the absence of ATP, but were destabilizing in the presence of ATP. For example, substitutions at Glu480 in asparagine synthetase (Asn1) did not affect Asn1 stability until ATP was added to the lysate, mirroring some of the patterns observed with Glu200 in Pgk1 (Figure 4.3C). Lastly, we observed substitutions in a novel ATP-sensitive protein, phosphoglycerate mutase (Gpm1), that may pinpoint specific residues moderating its sensitivity. For example, substitutions at Glu70 significantly stabilized Gpm1, regardless of ATP concentration. However, the relative change in stability across concentrations was diminished compared to wildtype Gpm1, suggesting that residue is important for ATP sensitivity of Gpm1.

4.4 Discussion

In this study, we describe a novel proteomics approach to identify residues important for small molecule binding at a proteome-wide scale. This method involves first generating a collection of protein variants through mistranslation with non-canonical amino acids (ncAAs) using Miro (Rodriguez-Mias et al. 2022). Second, lysates of these mistranslated proteomes are incubated with different concentrations of a small molecule and then analyzed by a high-throughput thermal stability assay to quantify differences in protein stability across conditions. We first optimized this stability assay for small molecule binding in yeast and then showed that we can detect ATP-binding proteins in both wildtype yeast (Figure 4.1) and in mistranslated proteomes (Figure 4.2). We identified glutamate residues known to mediate interactions with ATP using ATP-dependent ncAA sensitivity as the primary readout (Figure 4.3). Lastly, we showcase several examples of ATP-sensitive glutamate residues across the yeast proteome that may be involved in binding ATP, either directly (binding site) or indirectly (allosteric site).

As we show here, ncAAs substitutions provide unique advantages compared to conventional cognate amino acid substitutions when mapping protein-small molecule interactions. Using ncAAs enables precise chemical changes that target residue features important for protein-small molecule interactions, such as residue side-chain pK_a or side-chain length, potentially affecting the docking behavior of a metabolite just enough to detect a change in stability. In our case, the incorporation of fluoroglutamate at Glu342 in Pgk1 decreased ATP-dependent stabilization of Pgk1, supporting a role for Glu342's charge distribution or acidity in mediating interactions with ATP.

Additionally, introducing targeted chemical changes with ncAA substitutions may offer an opportunity to probe protein allostery without disrupting conformational protein ensembles (Tang and Fenton 2017). These ncAA substitutions open the door to potentially mapping allosteric residues, pinpointing allosteric binding sites of small molecules, and highlighting residues involved in small molecule-induced molecular motions that enable small molecule binding. We speculate that residues involved in ATP-induced conformational changes may explain a portion of residues where we observed significant changes in stability only in the presence of ATP (Glu 200 in Pgk1, Figure 4.3C; Glu480 in Asn1, Figure 4.4E).

Taken in context, the method presented here is unbiased in mapping residues with direct contact with small molecules or residues important for mediating these interactions in an unbiased manner. We expect that coupling this workflow with multiple ncAA substitutions and integrating these effects across conditions may help identify clusters of residues in 3-dimensional space that are sensitive to small molecules, which will illuminate novel binding pockets on the surface of proteins. We also envision this approach helping to illuminate paths of allostery within proteins by mapping residue sensitivity, very similar to what has been observed with mutational libraries.

Scaling these approaches to multiple mistranslated proteomes, small molecules, concentrations, and temperatures will enable generalized principles of allostery across proteomes. Additionally, sensitivity maps that converge ncAA sensitivity and small molecule sensitivity will pave the way for comparative genomics to characterize the evolution of allosteric regulation within individual proteins and protein families, a feat that has been difficult given the evolvable nature of allostery and poor resolution of small molecule binding sites within proteins. These approaches will enable integrated genome interpretation across species, helping annotate the potential consequences of standing genetic variation in yeast, and potentially clinical variation in humans. Lastly, these approaches can be ported to studying therapeutically-relevant small molecule binding events, enabling the proteome-wide mapping of therapeutic agents at single amino acid residue resolution, to help predict disease resistance mechanisms and to advance the development of therapeutic allosteric modulators.

4.5 Methods

4.5.1 Yeast strains

All experiments here were done with the *Saccharomyces cerevisiae* haploid strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0), a direct descendant from FY2, which is itself a direct descendant of S288C.

4.5.2 Generating mistranslated proteomes

Cultures of BY4741 were grown overnight at 30˚C in SCM containing 6.7 g/L yeast nitrogen base, 2 g/L of synthetic complete mix minus lysine, 2% glucose. These cultures were used to seed four fresh 60 mL cultures at OD_{600} 0.025, left growing at 30˚C. At OD_{600} 0.15,

isotopically-heavy lysine was supplemented at 0.872 mM final concentration, along with fluoroglutamate at 80 mg/L. Cultures were harvested at OD_{600} ~ 1 by centrifugation in 50 mL Falcon tubes at 8,000 × g, 4˚C, for 10 min. Supernatant was decanted and yeast pellets were washed by resuspension in 1 mL ice-cold sterile water and centrifugation in 2 mL screw cap tubes at 21,000 × g, 4˚C, for 10 min. Yeast pellets were washed one more time by resuspension in 1 mL ice-cold sterile water, split into two separate 2 mL screw cap tubes (500 µL each) at 21,000 × g, 4˚C, for 10 min; supernatant was decanted and pellets were snap-frozen in liquid nitrogen and stored at −80˚C.

4.5.3 Modified PISA

Frozen yeast cell pellets were resuspended in 450 µL of non-denaturing lysis buffer (25 mM HEPES pH 7.5, 75 mM NaCl) containing 0.25x protease inhibitors (Pierce) on ice. Cells were lysed by bead beating with 0.5 mm 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 × g for 1 min to remove the beads and then at 21,000 × g for 10 min at 4°C to remove cell debris. To bring all protein extracts to the same concentration, a Bradford Assay was performed and an additional amount of lysis buffer was added to bring protein extracts to 1.5 mg/mL.

For each biological replicate, cell extracts were aliquoted into two strips of PCR tubes (1×8 for the temperature gradient and 1×8 for the 30°C) dispensing 10 µL of protein extract per tube. All samples were initially equilibrated to 30°C for 5 min. Temperature gradient samples were subjected to 51.5°C, 52.7°C, 54.3°C, 55.5°C, one tube to each temperature, for 5 min. In parallel, controls were subjected to an additional 30.0°C temperature treatment for 5 min. All samples were cooled down to room temperature for 5 min. For each replicate, temperature gradient samples were pooled into one tube and 30°C controls were pooled into a separate tube

prior to centrifugation at 17,100 × g for 60 min at 4°C. The soluble protein fraction for the temperature gradient and 30.0˚C controls were mixed with denaturing lysis buffer (1:1, 25 mM HEPES pH 8.9, 75 mM NaCl, 9 M Urea). The 30.0˚C controls were then diluted one more time, this time by combining 1:1 with equal mixture of non-denaturing lysis buffer and denaturing lysis buffer (with buffer composition around 25 mM HEPES pH 8.2, 75 mM NaCl, 4.5 M Urea). Protein concentration was measured with a BCA assay.

Table 4.1. Temperatures used for optimizing stability assay on wildtype yeast

Temperature block	Temperatures
T1	48.0˚C, 49.2˚C, 50.7˚C, 52.3˚C
T2	49.2˚C, 50.7˚C, 52.3˚C, 53.7˚C
T3	50.7˚C, 52.3˚C, 53.7˚C, 55.2˚C
T4	52.3˚C, 53.7˚C, 55.2˚C, 56.8˚C
T5	53.7˚C, 55.2˚C, 56.8˚C, 58.0˚C

4.5.4 Proteomics sample preparation and desalting

Protein samples were reduced with 5 mM dithiothreitol (DTT) for 30 min at 55°C, followed by alkylation with 15 mM iodoacetamide for 30 min at room temperature in the dark. The alkylation reaction was quenched with 5 mM DTT at room temperature for 15 min. The pH was adjusted to 8.5 with 1 M HEPES pH 8.9 and 25 µg of sample was digested overnight with Lysyl endopeptidase (LysC; Wako Chemicals) at a 1:25 enzyme:substrate ratio. Digestion was quenched with 10% trifluoroacetic acid (TFA) to a final concentration of 1.5% and pH ~2-3.

Peptide samples were desalted by solid-phase extraction over 2 mg Oasis HLB 96-well μElution Plates using vacuum pressure. Packing material was conditioned with 200 μL methanol, 3 × 200 μL 100% acetonitrile, 200 μL 75% acetonitrile, 0.5% acetic acid, 200 μL 50% acetonitrile, 0.5% acetic acid, 3 × 200 μL of 0.1% TFA. Peptides were then loaded under low vacuum pressure, washed with 3 × 200 μL 0.1% TFA and 200 μl 0.5% acetic acid. Peptides were eluted into PCR tubes with 50 μl of 50% acetonitrile, 0.5% acetic acid and 50μL 75% acetonitrile, 0.5% acetic acid, and aliquoted into 15 μg aliquots for downstream labeling. All samples were lyophilized by vacuum centrifugation and stored at −80°C before labeling with tandem mass tags (TMTs).

4.5.5 Peptide labeling and desalting

Peptide aliquots (20 μg) were resuspended in 20 μl of 100 mM HEPES pH 8.2, 30% acetonitrile, and labeled with 5 μl of TMT10plex solution (equivalent to 50 μg of TMT10plex). Reactions were incubated at room temperature for 1 hour followed by a 15 min quench with 2.5 μl of 5% hydroxylamine solution followed at room temperature. Samples were then pooled and acidified to pH 2-3 using a final concentration of 1% TFA. Excess acetonitrile was removed by brief vacuum centrifugation followed by desalting over a 50 mg Sep-Pak tC_{18} cartridge (Waters). Packing material was washed with 1 mL methanol, 3 × 1 mL 100% acetonitrile, 1 mL 75% acetonitrile, 0.5% acetic acid, 1 mL 50% acetonitrile, 0.5% acetic acid, and equilibrated with 3 × 1 mL 0.1% TFA. Peptides were then loaded by gravity, washed with 3 × 1 mL 0.1% TFA and 1 mL 0.5% acetic acid. Peptides were eluted with 750 μL of 50% acetonitrile, 0.5% acetic acid and 750 μL 75% acetonitrile, 0.5% acetic acid. Samples were vortexed, dried by vacuum centrifugation, and stored at −80°C before acquisition.

4.5.6 Mass spectrometry data acquisition

Lyophilized TMT-labeled peptide samples were resuspended in 5% ACN, 5% formic acid and subjected to liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Duplicate injections of unfractionated samples for each sample were collected on an Orbitrap Eclipse mass spectrometer (ThermoFisher Scientific) coupled to an EASY-nLC 1200 LC (ThermoFisher Scientific). Peptides were loaded on a 100 µm x 3 cm trap column packed with 3 µm C18 beads (Dr. Maisch) and separated on a 50˚C-heated 35 cm analytical column packed with 1.9 µm C18 beads (Dr. Maisch) using a 160-minute gradient of 80% acetonitrile, 0.1% formic acid. Peptides were analyzed online using data-dependent acquisition using an MS3-based method with real time search and 3-second cycle time. First, MS1 data were collected using the Orbitrap (120,000 resolution; maximum injection time 50 ms; AGC 4e5, mass range 400-1600 m/z, charge states 2-6, dynamic exclusion 30 seconds). MS2 scans of the most intense precursors were performed in the ion trap with CID fragmentation (isolation window 0.8 Da, rapid, NCE 35%, maximum injection time 50 ms; AGC 1e4). An online real-time search algorithm (Schweppe et al. 2020) was used to search spectra against a yeast FASTA database (static modifications: TMT10plex on N-terminus and lysines; variable modifications: mass shift associated with ncAA, heavy lysine, methionine oxidation) with maximum 1 missed cleavage and 3 variable modifications. MS2 scans that passed RTS thresholds were sent for MS3 quantification. The top 10 matching fragment ions were isolated using synchronous precursor selection (McAlister et al. 2014) followed by HCD fragmentation and collected for an MS3 scan in the Orbitrap (resolution of 60,000, NCE of 55%, maximum injection time of 120 ms, isolation window 0.8 Da, and AGC of 1e5). For one of the replicate injections (for wildtype yeast used for the optimization PISA samples), turboTMT was used for the MS3 acquisition.

4.5.7 Mass spectrometry data processing

Raw files were converted to mzML formats using msconvert (version), and MS/MS spectra were searched against a target/decoy protein sequence database using Comet (v2019.01.02) (Eng,

Jahan, and Hoopmann 2013). *Saccharomyces cerevisiae* (orf_trans_all.fasta downloaded from the Saccharomyces Genome Database in 2014). 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 0.6 Da 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 fluoroglutamate (+17.99057813134). Dynamic SILAC samples were searched with light lysine (K0) and heavy (K8, +8.0142 Da) variable modifications in binary mode. TMT-labeled samples were searched with constant modification (+229.1629 Da) on lysines and peptide N-termini. Search results were filtered with Percolator (Percolator version 3.01) (Käll et al. 2007) 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.

4.5.8 Quantifying peptide-level and protein-level relative stability

All scan-level TMT reporter ion intensities were first corrected for differences in sample loading by summing reporter ion intensities across each of the TMT channels and applying correction factors so that summed intensities were the same across all channels. A key difference with this workflow: since the proteomic composition was known to be different between the 30˚C samples and the temperature gradient-treated samples, sample loading normalization was applied separately for each group and each ATP concentration. Lastly, to account for potential systematic differences in the distribution of relative stability, a correction factor was applied to each mistranslated proteome such that the dominant peak of non-melting proteins all aligned at an R_s value equal to 100%. The relative stability for each peptide was quantified using the

equation below. Lastly, protein relative stability was determined as the median R_s value for all peptides quantified for that protein, with a requirement of at least two unique peptides quantified per protein.

$$R_s = \frac{Reporter\ ion\ intensity_{temperature\ gradient}}{Reporter\ ion\ intensity_{30^{\circ}C\ control} \times 2} \times 100$$

4.5.9 Identifying stability-altering substitutions

To identify substitutions that significantly alter protein thermal stability, we compared the R_s values of ncAA-containing peptides against wildtype peptides using a paired t-test in R and limma. Specifically, we first required that all comparisons were between fully tryptic peptides containing a heavy lysine. Second, to consolidate from peptide-level to site-level quantifications, we required that all versions of peptide mapping to the same site be seen in the ncAA and wildtype form (i.e. if one of the versions is a methionine oxidized peptide, then the oxidized version must have been quantified for both the ncAA and wildtype peptide). Then, after ensuring overlap between ncAA and wildtype peptides, we collapsed peptides into site-level quantifications by taking the median R_s value for each replicate. We then compared the ncAA R_s value with the wildtype R_s value using a paired t-test in R with the limma package. All p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg correction. Statistical testing, p-values, and p-value corrections were done separately for each mistranslated to control for sample-specific effects on protein stability. Lastly, to determine an effect size cutoff, we required that ΔR_s of ncAA substitutions be greater than the median standard deviation ΔR_s across mistranslated proteomes between wildtype peptides and their corresponding protein-level R_s values.

4.5.10 Data availability

The raw data analyzed here will be posted to PRIDE upon submission of this manuscript to a journal.

4.5.11 Code availability

All code used to generate the analyses above will be uploaded to GitLab upon submission of this manuscript to a journal.

Chapter 5. Identifying substrates of a protease proteome-wide

This Chapter is based on unpublished observations

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

5.1 Summary

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, in HEK293T cells, we separately overexpressed wildtype NSP5, as well as controls of catalytically inactive NSP5, and GFP. Following this overexpression, we subjected proteomes 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 that, only in the presence of catalytically active NSP5, had altered protein turnover or thermal stability. Proteins with the LQ|[AS] motif that demonstrated an altered protein turnover tended to have faster turnover in the presence of NSP5, supporting the idea that NSP5-cleaved substrates generally are destabilized and have increased degradation upon cleavage. Using protein-level or peptide-level readouts, we identified candidate NSP5 substrates, many of which aligned with known NSP5 substrates and proteins containing the NSP5 motif. In combination with N-terminomics, our protein thermal stability and protein turnover assays can unbiasedly catalog NSP5's protease substrates. The methods presented here enable unprecedented functional insight into the consequences of a protein cleavage event on the protein and its cleaved products.

5.2 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 the 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 (Stukalov et al. 2021), proteome (Stukalov et al. 2021; Bojkova et al. 2020; Selkrig et al. 2021; Bouhaddou et al. 2020; Klann et al. 2020), ubiquitome (Stukalov et

al. 2021), phosphoproteome (Stukalov et al. 2021; Bouhaddou et al. 2020; Klann et al. 2020), protein interactome (Stukalov et al. 2021; Gordon et al. 2020; Laurent et al. 2020), translatome (Bojkova et al. 2020), and proteome thermal stability (Selkrig et al. 2021). Despite vaccines, SARS-CoV-2 infection still evades our immune system and propagates to others, creating a need for effective antivirals to combat active infection. One of the leading candidates for therapeutic target is SARS-CoV-2's main protease NSP5 (M^{pro}; chymotrypsin-like protease: $3CL^{pro}$), which is essential for viral replication.

The SARS-CoV-2 genome contains ORF1 which is transcribed and then translated into two polyproteins that must be cleaved by NSP5 and a papain-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 (Jin et al. 2020; Zhang et al. 2020) has been explored to prevent viral propagation. NSP5, whose sequence and function is conserved across coronaviruses (Roe et al. 2021; Flynn et al. 2022), demonstrates catalytic specificity for substrates containing the putative LQ|[AS] motif (Pablos et al. 2021).

Many proteomics methods have been employed to identify NSP5 protein interactions and protease substrates among 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) (Gordon et al. 2020). To improve the sensitivity of detecting NSP5 interactors, N-terminomics MS approaches, in the context of viral infection (B. Meyer et al. 2021) or cell lysates with dosed recombinant NSP5 (Pablos et al. 2021), have enabled the discovery of hundreds of viral and host neo-N-terminus peptides. In a N-terminomics approach like TAILS (Kleifeld et al. 2010, 2011), 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 (Pablos et al. 2021). Although hundreds of NSP5 cleavage events have been discovered, extensive follow-up experiments would be required to determine the impact of the NSP5 cleavage on the substrate's protein function. NSP5 protein cleavages could also 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 (Zecha et al. 2018) and thermal stability (Smith et al. 2021) 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.

Here, we assayed protein thermal stability and protein turnover proteome-wide in HEK293T cells overexpressing wildtype NSP5, catalytically inactive NSP5 C145A variant, or GFP. We observed global proteome changes in protein turnover and thermal stability that 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 that 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.

5.3 Results

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

Here, 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 wildtype NSP5 protease, catalytically inactive NSP5 C145A, or GFP (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 were digested into peptides and analyzed by MS/MS (Figure 5.1a). 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 consolidated 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, the

interpretation would be that NSP5 protease is either directly cleaving the protein substrate and thereby altering its turnover or is indirectly acting to modulate the protein's turnover.

Alternatively, we could 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. After a protein is cleaved, its products should contain specific peptides that reflect the turnover of the protein and the cleavage products. We could identify a protein cleavage event when the cleaved protein products have different turnovers. This difference 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 leveraged 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 5.1b). 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 result suggests the promoter system likely enables exceedingly faster synthesis rates compared to what is naturally observed across the HEK293T during the 39 hour overexpression.

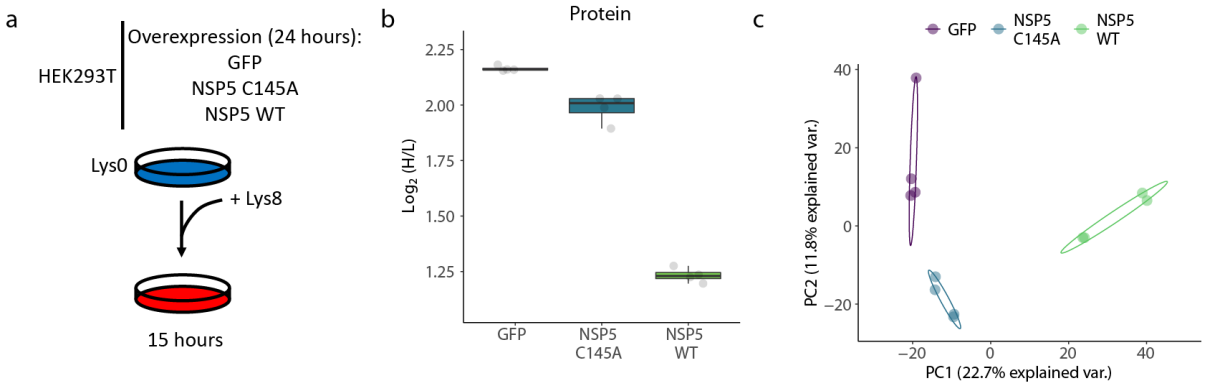


Figure 5.1: Dynamic SILAC to measure HEK293T protein turnover.

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).

Across the proteome, we captured ~4,000 unique human proteins across all replicates and all overexpression conditions (Appendix C Supplementary Figure 5.1a). 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 5.2). 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 5.1c). The replicates for the same overexpressed protein clustered closely together, and replicates of different overexpressed proteins separated. In particular, the variance in principal component 1 (PC1) which explained 22% of the total variance, separated NSP5 wildtype replicates from GFP and NSP5 C145A replicates, highlighting that protease activity likely contributes unique differences in protein turnover across the proteome.

NSP5 protease activity modulates host protein turnover

A method has not been available to explicitly explore the functional impacts of the critical NSP5 protease on the host proteome. The dynamic SILAC data allowed us to explore whether NSP5 activity modulated 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 5.2a, Appendix C Supplementary Figure 5.3). We observed very few proteins with significantly altered protein turnover when we compared wildtype NSP5 to catalytically inactive NSP5, suggesting that NSP5 interaction events likely have limited impact on protein turnover. We observed greater than 100 proteins with altered protein turnover when comparing NSP5 wildtype and GFP conditions, suggesting that NSP5 activity plays a role in modulating protein turnover across the proteome. When we compared NSP5 wildtype to

inactive NSP5 C145A conditions, we observed fewer 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 NSP5 wildtype could be attributed to the proteins being a NSP5 protease substrate and/or binding partners of substrates. 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 resulting in 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 5.2b). In alignment with our Limma analysis, very few proteins with altered ΔR_{TO} lie on the diagonal (line with slope=1) and deviate from the origin, suggesting that few proteins had 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} arose 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 NSP5 wildtype protease (Appendix C Supplementary Figure 5.4a). 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 5.4b). 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 their sequence. In theory, NSP5 motif-containing proteins should be more likely to be protease substrates of NSP5. When mapped to our ANOVA significant calls, 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 5.2b). This result supports our hypothesis that NSP5 protease cleavage of protein substrates accelerates their protein turnover, likely due to destabilization and increased degradation of its antecedent 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 NSP5 wildtype protease was expressed (Figure 5.2c-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 5.2f). Collectively, the protein turnover at the protein-level revealed that NSP5 protease activity can modulate proteome-wide protein turnover (R_{TO}) changes. Many faster R_{TO} proteins during NSP5 overexpression contain the putative NSP5 motif, potentially implicating them as NSP5 protease substrates.

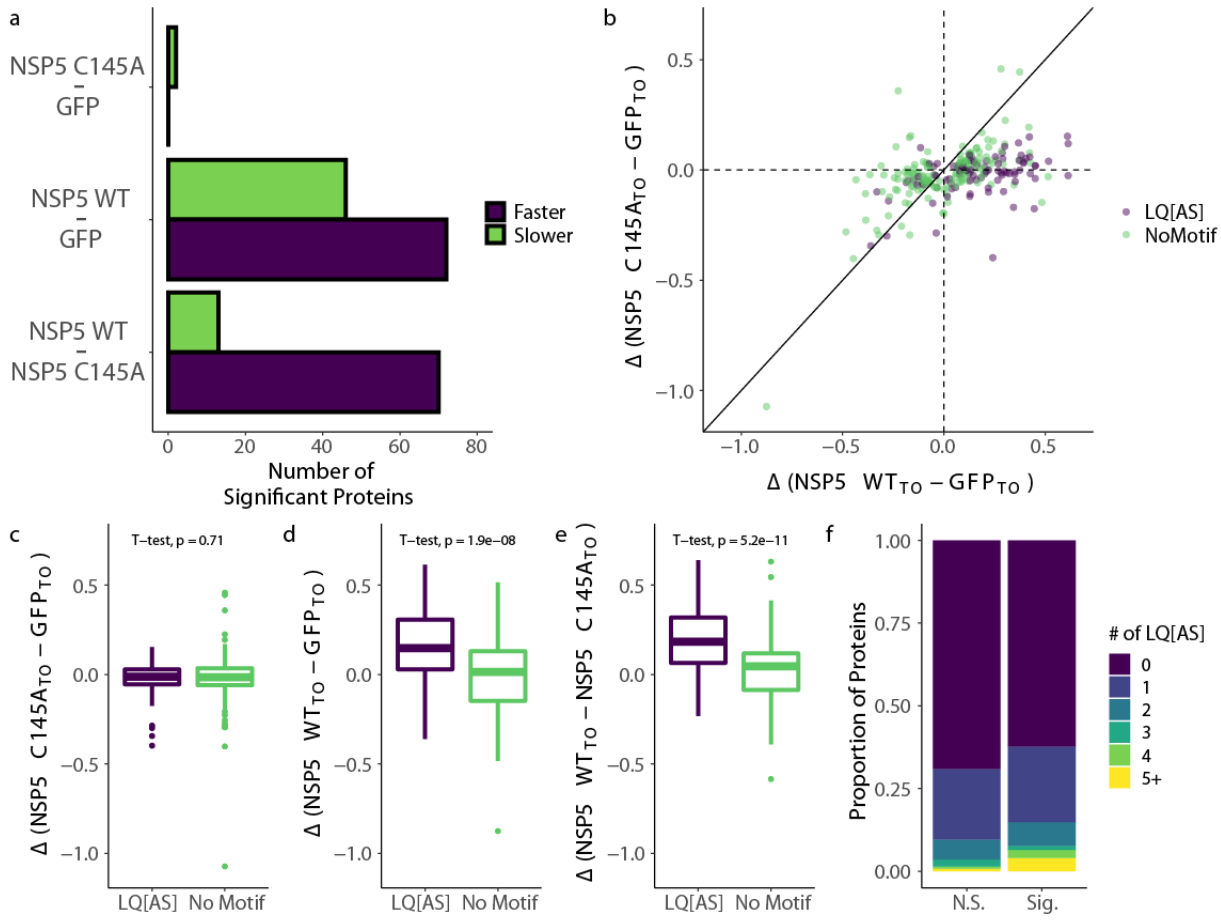


Figure 5.2: NSP5 protease activity modulated changes in protein R_{TO} and the faster R_{TO} proteins associated with the NSP5 motif.

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.

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 for 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 incubating lysates with non-denaturing detergent and benzonase, followed by aggregate removal by centrifugation. The non-denatured protein fraction from each channel was 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 5.5), 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 5.3a). Interestingly, the increased stability of catalytically-dead NSP5 lines up with prior observations that point mutants

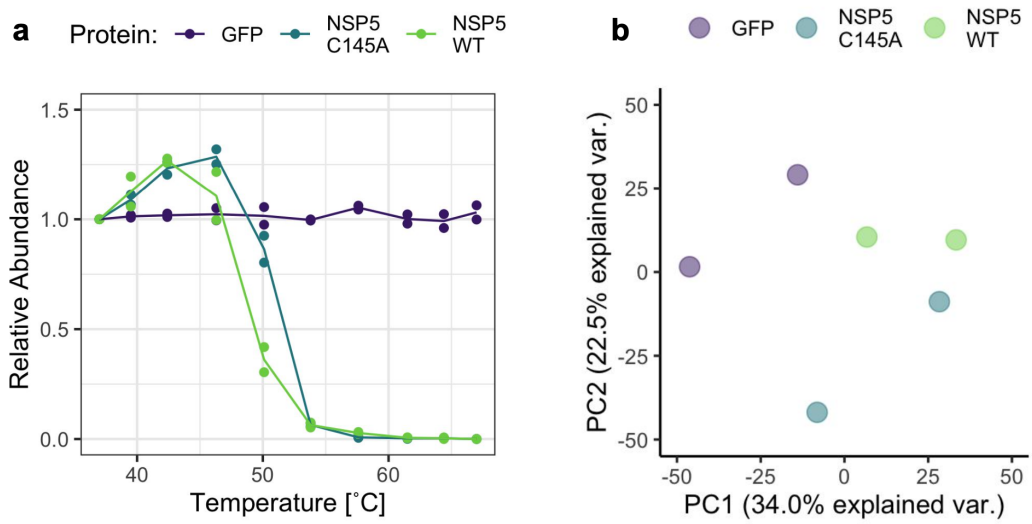


Figure 5.3: Crude Thermal Proteome Profiling to measure NSP5-dependent changes in protein stability.

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).

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) (Ward et al. 1982), 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.

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 principal component analysis, where we observed broad separation of samples that clustered based on which enzyme was overexpressed (Figure 5.3b). 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) (Childs et al. 2019). Specifically, we focused on the 2,570 proteins that showed 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

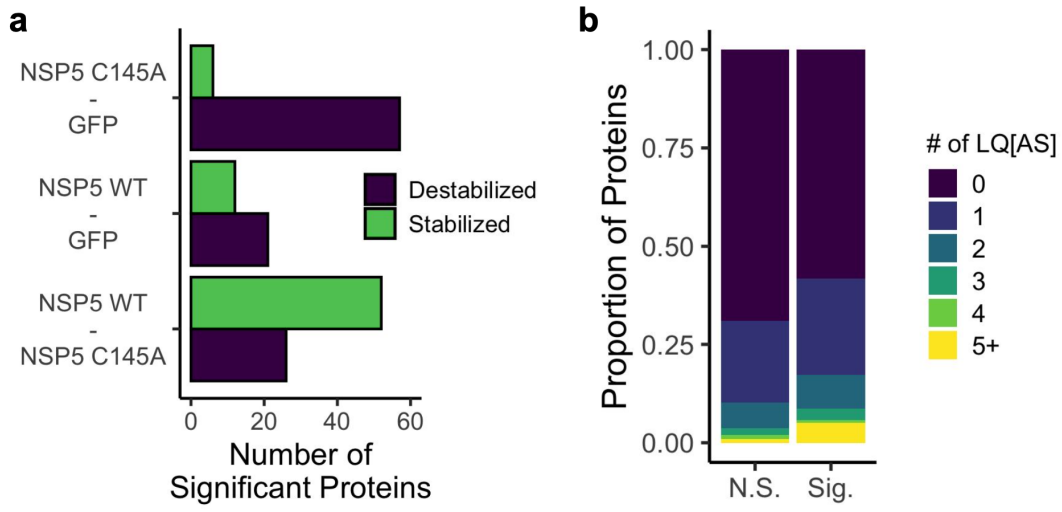


Figure 5.4: Proteins with altered stability due to NSP5 overexpression contain the NSP5 motif.

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.

NSP5 compared to 52 stabilized in wildtype NSP5) (Figure 5.4a), 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 5.4b), 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.

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 due to the activity of NSP5 protease could be substrates. Thus, we expected to observe substantial overlap between known NSP5 substrates and proteins with altered turnover and thermal stability. In this section, we 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 prioritize 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. 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 (Pablos et al. 2021; B. Meyer et al. 2021; Moustaqil et al. 2021). Thus, we leveraged the collective curated list for comparison.

First, we asked whether any of these known targets of NSP5 correspond to proteins that have altered protein thermal stability during wildtype NSP5 overexpression. From a protein-level perspective, we quantified changes in stability for 113 proteins out of a possible 244 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 proteins with changes in thermal stability (Figure 5.5a).

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 5.5b-c). The change in thermal stability was similar across the different peptides quantified within OAT (Figure 5.6b), resulting in a reproducible stability change at the protein-level (Figure 5.6c). 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.

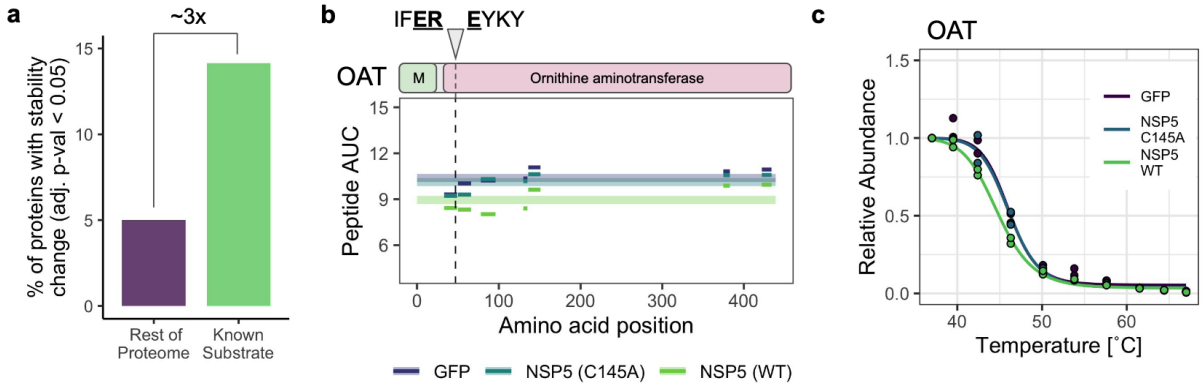


Figure 5.5: Protein-level thermal stability changes of known NSP5 substrates.

(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) is 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 wildtype 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 enrichment is in agreement with the notion presented previously that the cause of the significant alterations in turnover of some proteins is due to their being NSP5 protease substrates.

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 wildtype NSP5 protease overexpression. The R_{TO} change was observed to the same extent for all its peptides across the length of the protein (Figure 5.6a), resulting in a reproducible turnover change at the protein-level (Figure 5.6b). EIF4G1 has been observed in multiple studies to bind viral RNA prominently after 24 hours of SARS-CoV-2 infection (Schmidt et al. 2020; Kamel et al. 2021; Labeau et al. 2022). Labeau *et al.* (2022) followed-up on many of the viral RNA human binding proteins (RBP), including EIF4G1, for their functional role during SARS-CoV-2 infection. In this experiment, siRNA-mediated knockdown of EIF4G1 and other RBPs was performed in A549-ACE2 cells, and viral titers and viral replication was then assessed following SARS-CoV-2 infection. For EIF4G1 knockdown cells, there was little to no observed effect on viral replication upon infection; however, viral titers were over 200% greater from 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 titers. 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 titers.

We also observed increased protein turnover for the polypyrimidine tract binding protein (PTBP1) for all its peptides (Figure 5.6c) and at the protein-level (Figure 5.6d). 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.* (2021) 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. 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 eliminate 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 (Pablos et al. 2021). 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.

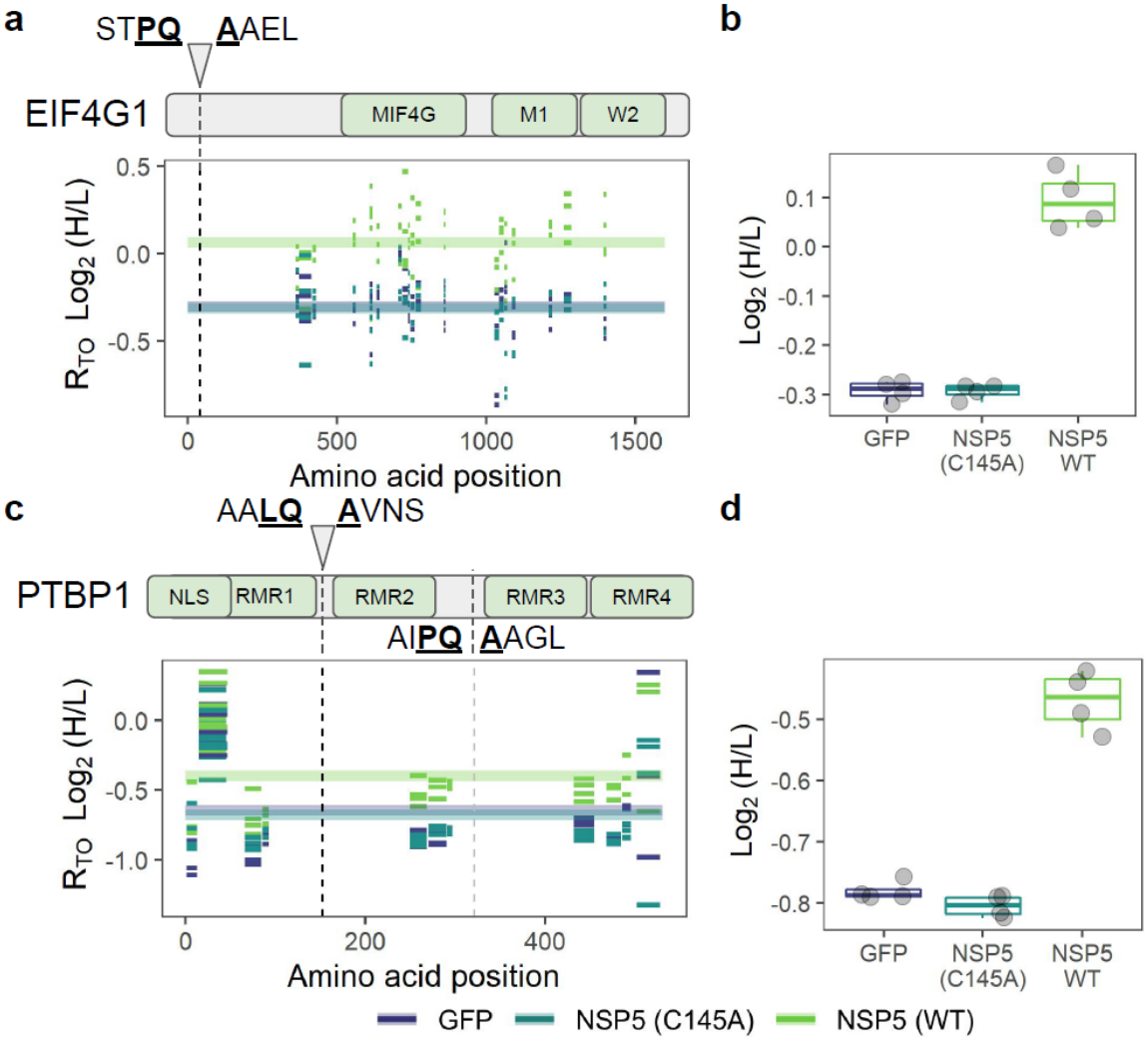


Figure 5.6. Protein-level faster R_{TO} in known NSP5 substrates.

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.

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 phenomenon, we first focus on two proteins that showed cleavage site-dependent changes in stability and turnover: SEPTIN9 and MAGE-D2 (Figure 5.7). Both of these proteins are high-confidence substrates of NSP5 in the HEK293T proteome (Pablos et al. 2021)) 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. Both of these proteins showed a shift in peptide-level behaviors that coincided 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 were more stable and had 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 were more stable and had a slower turnover compared to the peptides on the C-terminal side of the cleavage site.

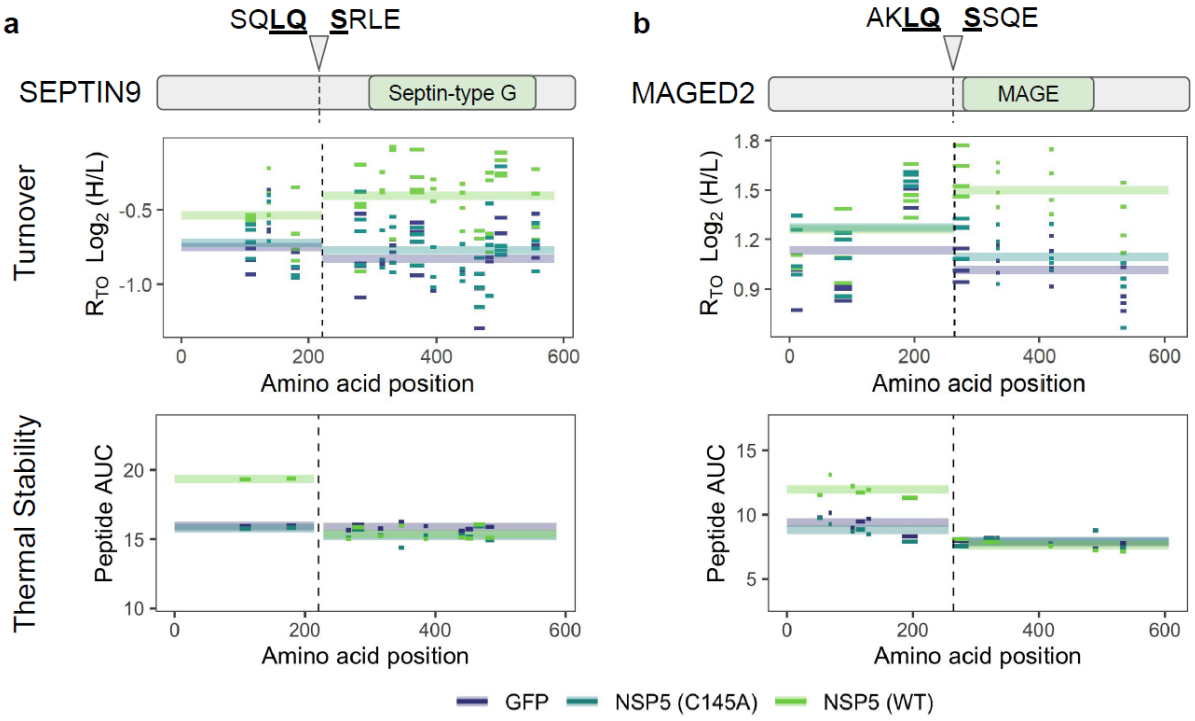


Figure 5.7. Peptide-level readouts for turnover and stability locate the specific cleavage sites for known NSP5 substrates.

(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 full-length protein. As a result, both dynamic SILAC and TPP offer opportunities to capture cleavage events, given that both thermal stability and turnover, while intrinsically related to protein structure and fold, will inherently be 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. Similar to SEPTIN9 and MAGE-D2, we observed a breakpoint in the peptide-level R_{TO} values that coincide with the known cleavage site at position 452 only in the wildtype NSP5 overexpression condition. The C-terminus of the protein demonstrated an accelerated protein turnover, which is likely due to an altered function of the C-terminus comparted to the full length protein or the N-terminal cleavage product. Pablos *et al.* (2021) 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. We also captured the known NSP5 cleavage event at position 34 in the SAICAR synthetase portion of the multi-functional protein ADE2 (PAICS). This cleavage 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 (B. Meyer et al. 2021). A siRNA knockdown of PAICS did not significantly reduce viral titers; 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, the small peptide N-terminus released upon NSP5 cleavage likely behaves differently than its full length protein and its C-terminal cleavage product.

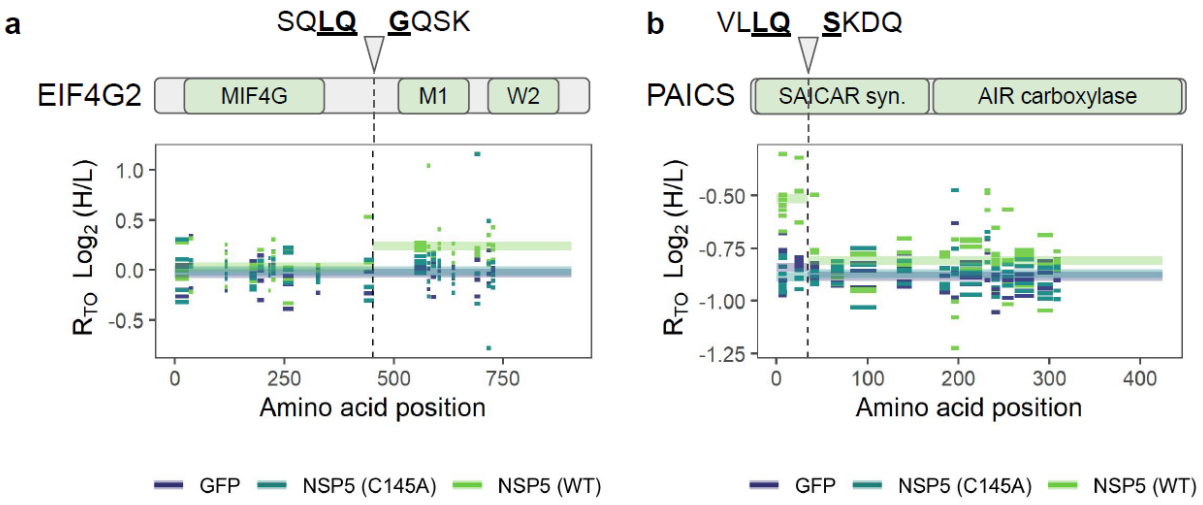


Figure 5.8. Known NSP5 substrates, EIF4G2 and PAICS, have altered protein turnover for their cleaved products.

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 emerged 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 5.9) and the U4/U6-U5 tri-snRNP pre-spliceosome associated protein PRPF3 (Figure 5.10). For RPL4, there is one potential NSP5 cut site, the LQA motif at position 362. In our TPP assay, we captured 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 observed a marked shift in thermal stability for peptides derived from the C-terminus that is specific to wildtype NSP5 (Figure 5.8a). Interestingly, while this C-terminal region is not captured in crystal structures of the ribosome, the region spanning the LQA motif is crystalized (Figure 5.8b). This region is at the very edge of the ribosome, solvent accessible, and therefore likely accessible to wildtype NSP5. RPL4 overexpression has been shown to increase the efficiency of viral translation for viruses requiring frameshifts (Green et al. 2012), 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.

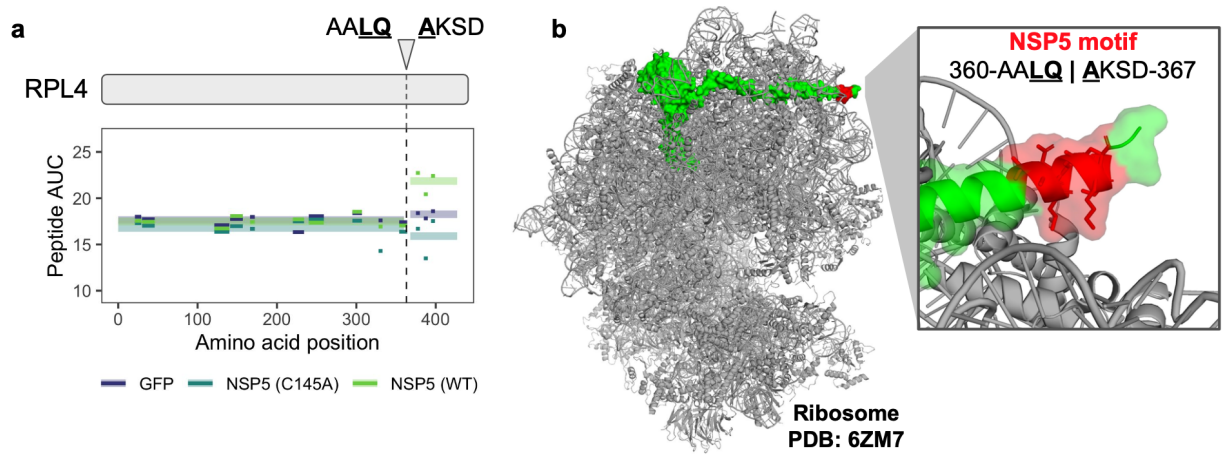


Figure 5.9. The ribosomal subunit RPL4 is a potential substrate of NSP5.

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.

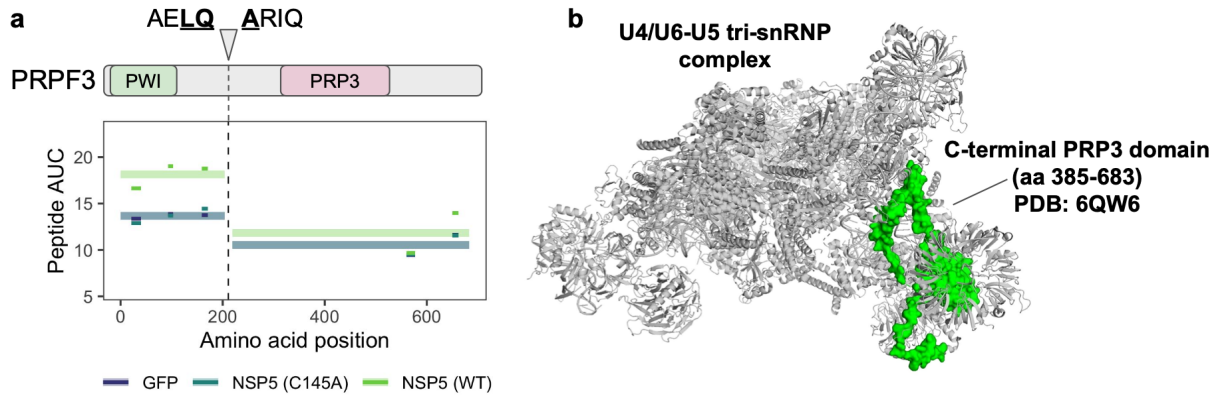


Figure 5.10. The pre-mRNA splicing protein PRPF3 is a potential substrate of NSP5.

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.

The second protein whose cleavage may have relevant biological consequences is the spliceosomal-associated protein PRPF3 (Figure 5.10). SARS-CoV-2 has already been shown to suppress global mRNA splicing through targeting U1/U2 RNAs by NSP16 (Banerjee et al. 2020). Additionally, knocking down PRPF3 has been shown to increase splicing defects in neuronal cells (Schaffert et al. 2004). 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, these data suggest NSP5 cleavage 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.

5.4 Discussion

The SARS-CoV-2 main viral protease, NSP5, has been extensively studied for its essential role of cleaving a large viral polyprotein into its many functional protein units, a process essential for viral propagation. Beyond viral proteins, N-terminomics studies (Pablos et al. 2021; B. Meyer et al. 2021) have revealed that NSP5 can target over 100 human proteins as protease substrates, many of which are relevant during infection. However, less is known about the 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 NSP5 protease activity. By comparing HEK293T cells that overexpress wildtype NSP5 to those that overexpress GFP or catalytically inactive NSP5 C145A, we captured over a hundred proteins with altered protein turnover and/or thermal stability attributed specifically to NSP5

protease activity. Our assay was sensitive, and enabled the characterization of NSP5 protease functions apart from its other functional roles.

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 was present. In addition, 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 that 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 candidateq substrates of NSP5.

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 identified 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 between 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, these 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, the presence of an NSP5 motif

and certain directional changes in the functional properties can help prioritize between the two scenarios. Second, the absence of altered thermal stability or turnover of a protein does not exclude the possibility that a cleavage event occurred due to NSP5 protease. The lack of observed changes in turnover or stability could be due to biological and technical reasons. For instance, a technical consideration is that some proteins are not amenable to our thermal stability and protein turnover assays. 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 for us to capture a reliable turnover readout. These proteins likely result in false negatives or missed proteins in our analysis. A biological consideration is that some proteins that are cleaved and are amenable to the assays might not have a change in thermal stability or turnover, and thus the assays do no provide functional support for the cleavage. 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, a better statistical framework is needed to reproducibly call breakpoints and give confidence in our NSP5 cleavage calls.

We can 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 posit that 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 hypothesis is supported by the overlap in N-terminomic studies and our study for the protein PTBP1, for which the functional impact of 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-terminus 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 criterion 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.

5.5 Methods

5.5.1 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.

5.5.2 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. (Leutert et al. 2019) study.

Peptides were desalted and fractionated on StageTips (Rappsilber, Mann, and Ishihama 2007) 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.

5.5.3 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.

5.5.4 TPP 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 (Leutert et al. 2019) 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.

5.5.5 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.

5.5.6 Offline peptide fractionation for TPP samples

Peptides were fractionated using a pentafluorophenyl (PFP) reverse-phase fractionation (Grassetti, Hards, and Gerber 2017), 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).

5.5.7 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.

5.5.8 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 (Eng et al. 2015; Eng, Jahan, and Hoopmann 2013) (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 (Nesvizhskii et al. 2003). TMT reporter ion intensities were corrected for isotopic interference.

5.5.9 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.

5.5.10 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. 2018) study. 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.

5.5.11 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) (Childs et al. 2019) 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).

5.5.12 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 (Cox and Mann 2008) (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 ($R_{TO}=\log_2(\Sigma(\text{heavy})/\Sigma(\text{light}))$ (Zecha et al. 2022). 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 (Ritchie et al. 2015) (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 (Tyanova et al. 2016) 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.

Chapter 6: Functional interrogation of proteomes at single amino acid resolution

Protein biochemistry is in the midst of a revolution, and mass spectrometry-based proteomics is an essential tool to fully realizing the potential of this revolution. The methods presented in this dissertation, and the observations that have emerged as a result of their development, are only a small part of the larger movement in the field toward mapping protein functions at single amino acid resolution, and offer an opportunity to address the problem of elucidating protein sequence-function relationships.

The overarching theme of my thesis work has been to interrogate sequence-function relationships from a proteomics perspective. There were two main goals from the outset of this project. The first goal was to develop a technology that increased the scale of mutational scanning-type experiments from single proteins to entire proteomes. To accomplish this goal, I helped develop Miro, a proteomics platform that generates protein containing amino acid substitutions across entire proteomes and measures the effects of substitutions using biochemical selections with a mass spectrometry-based readout. To further develop Miro, I screened a panel of 26 non-canonical amino acids for proteome-wide incorporation in yeast, all of which were curated for their potential for incorporation based on *in vitro* (Hartman, Josephson, and Szostak 2006; Hartman et al. 2007) or *in vivo* (Richmond 1962; Cowie et al. 1959) evidence. Even within this set of analogs, incorporation varied drastically, with fluorinated amino acids emerging as the analogs most effective at bypassing translational quality control in yeast. Interestingly, the relationship between analog toxicity and proteome-wide incorporation was, at-best, complicated; while most toxic ncAAs incorporated in yeast, several analogs were toxic with no evidence of incorporation, and conversely, some mistranslating ncAAs were not toxic (at the concentrations we tested). I intentionally selected one of these “non-toxic” ncAAs,

thioproline, to assess its effects on protein stability in Chapter 3. While this analog was less impactful than its structural neighbor, azetidine, I still found close to 100 substitutions in yeast that significantly altered thermal stability, the majority of which occurred at proline residues insensitive to azetidine, indicating that there may be significant value in probing the effects of any mistranslating ncAA on the molecular phenotypes of proteins, regardless of its inherent toxicity on cell viability and biology. This finding has future implications for larger chemical libraries and for porting these systems into mammalian cells.

The experiments presented here also suggest that proteome-wide incorporation of ncAAs is indiscriminate; genes that are considered “essential” and “non-essential” in yeast (at least when grown in standard laboratory conditions, which were used here) were equally susceptible to mistranslation. However, variant coverage was biased towards abundant proteins and proteins with slower half-lives. Future advances in instrumentation, data acquisition, and data processing should alleviate some of this bias towards abundance. However, the bias towards slower turnover proteins hints at some intrinsic biological constraints. Endogenous protein quality control, even in minimally mistranslated proteomes, may prevent some extremely damaging substitutions from being quantified within proteins, and may warrant additional manipulation of protein quality control machinery to achieve complete coverage. On a final note, the analogs I screened in yeast were also screened in *Escherichia coli* by another colleague, Dr. Bianca Ruiz, who found similar patterns and levels of proteome-wide mistranslation. These screening data add additional support to the generalizability of Miro to other biological systems (such as mammalian cells).

The second goal at the outset of this project was to develop and apply modular proteomics assays on mistranslated proteomes. On the one hand, this goal included adapting currently-available proteomics assays to be compatible with mistranslated proteomes and

peptide-level readouts. In Chapter 2, I showcased the adaptation and application of Thermal Proteome Profiling in an azetidine-mistranslated proteome. On the other hand, there was motivation to continue developing new methods to enable novel discoveries, insights, or experimental scale, and to extend these methods to other collections of protein variants, such as proteins with post-translational modifications, or more conventional mutational libraries, where these methods may provide a more mechanistic understanding of variant effects. In Chapters 3 and 4, I developed and modified a previously-existing stability assay to measure protein relative stability in high-throughput, and showed how this approach could be coupled with small molecules to identify ATP binding sites proteome-wide, in addition to identifying residues that may be important for allostery within ATP-binding proteins.

The proteomics assays described in this dissertation provide a snapshot of the type of information we can glean by coupling protein variants with biochemical selections. Most of my efforts in this space have been geared towards thermal stability assays, given that it is an incredibly rich molecular phenotype. For one, protein stability is intrinsically tied to a protein's fold and structure. Thus, changes in thermal stability caused by an amino acid substitution may pinpoint residues important for protein fold and structure. For another, protein thermal stability is acutely sensitive to changes in protein functions, such as changes in interactions with binding partners (Becher et al. 2018), small molecules, (Savitski et al. 2014), or changes in enzymatic activity (André Mateus et al. 2020; Perrin et al. 2020). Unsurprisingly, we detected several instances of residues at interaction interfaces that altered protein stability in Chapter 3. However, an added benefit to protein thermal stability as a property to assay is that these methods can be easily expanded to include secondary selections and subsequently to probe changes in stability in this context. For example, in Chapter 4 I showcased how including metabolites or small molecules in this workflow helped identify residues involved in mediating interactions with proteins. One could imagine including other types of biomolecules in this

workflow to probe targeted questions of protein biology, such as: including motif-containing oligonucleotides of DNA or RNA to determine which proteins bind a particular motif; spiking in protein domains to define which proteins across the proteome are recognized by particular domains; or using peptides containing ncAA-encoded PTMs to map which protein-protein interactions are mediated by these PTMs. Simply put, thermal stability assays are both phenotypically rich measurements *and* are highly modular workflows with plenty of room for user- and question-specific customization. Case in point, in Chapter 5, I illustrated how thermal stability workflows can be adapted to identify protease substrates in an unbiased manner.

Future methodological improvements

Having reached the end of the first round of development for these technologies, what is left to be done and improved upon? I highlight several outstanding questions and briefly discuss some of the key improvements needed to address these questions below.

I. Are we missing the most damaging substitutions?

Variant coverage is still a major challenge and constraint. Even in the best of circumstances presented here (i.e., in a deeply fractionated proteome from a relatively “simple” organism, acquired on state-of-the-art instrumentation with real time search and multiplexing capabilities) I still only captured ~18% of the possible substitutions across the detected yeast mistranslated proteome; this fraction does not even take into account proteins that we miss in this workflow, such as membrane proteins and insoluble nuclear proteins. Additionally, we quantified the effects of substitutions in only ~700 out of ~6500 possible yeast proteins. Of course, not all of these proteins are expressed in standard laboratory growth conditions, nor are they all compatible with biochemical selections that require proteins to be soluble in non-denaturing

buffers, but proteome coverage is clearly an important bottleneck that will benefit from alternative functional selections and workflows targeting different subsets of the proteome, such as selections for solubility. Improvement in this area should help clarify the observed systematic "tolerance" of ncAA substitutions in Chapter 3. Given the significant correlation between variant coverage and protein and peptide abundance, and the current lack of enrichment strategies for mistranslated peptides (not accounting for click chemistry-based enrichment), continued advances in intelligent data acquisition strategies (Schweppe et al. 2020) will be crucial, with a particular focus on methods that do not rely on observing the peptides-of-interest in the initial precursor scan (i.e. MS1 scan).

II. Should we increase proteome-wide mistranslation?

Alternatively, increasing proteome-wide mistranslation by supplementing media with a higher concentration of ncAA may offer a feasible solution to this problem. Dialing up mistranslation should increase the abundance of ncAA-containing peptides, which would increase coverage of the mistranslated proteome using current instrumentation. However, one important consideration with increasing proteome-wide mistranslation is whether or not epistasis has a significant confounding effect at higher mistranslation rates. For one, the host cell will mount a stress response at higher mistranslation levels, which would presumably increase the fraction of the proteome actively monitored by molecular chaperones. For another, higher mistranslation rates increase the number of proteins with two, three, sometimes four ncAA incorporation events, exponentially expanding the combinatorial protein variant space, which would be subsequently decoupled by enzymatic digestion. Presumably, epistasis can still be mitigated, but I advise extreme caution before concluding too much from proteomes with a higher mistranslation frequency; it is not clear at this point whether some proteins would be more susceptible than others, and whether or not specific analogs would be more likely to generate epistatic effects.

III. What functional selections are most informative?

The last area where there is still plenty of room growth is in the types of functional selections compatible with mistranslated proteomes and protein variants. Several established proteomics assays already exist that could feasibly be adapted for mistranslated proteomes, such as assays for protein-protein interactions (Affinity Purification-MS, Size Exclusion Chromatography-MS, crosslinking-MS, Protein Correlation Profiling), subcellular localization (LOPIT, hyperLOPIT), catalytic activity (Activity-Based Proteome Profiling), turnover (dynamic SILAC), solubility (Solubility Proteome Profiling, Ionic Proteome Integral Solubility Alteration), and accessibility (Limited Proteolysis, Protein Painting). Additionally, as we saw with protein thermal stability, some of these assays may be extended through modifications to the workflow. However, which functional selections are most informative for different facets of protein biology, and how many functional selections are needed if the end-goal is to classify residues as functional or non-functional, are questions that will need to be answered with time. These are questions that should become clearer as we continue to apply these assays to different mistranslated proteomes and different cellular and organismal contexts, and as the proteomics community as a whole begins to apply these selections proteome-wide for functional proteomics questions.

IV. Moving towards peptide-level readouts of protein function

A theme that I explored only implicitly here is the role of peptide-level readouts in mapping sequence-function relationships. From the experiments presented here and from my own personal experience in the lab working with these datasets, precise and accurate peptide-level measurements of protein properties (e.g. stability and turnover) are essential to tackling challenging questions in protein biology. For the field of proteomics as a whole, collapsing peptide-level measurements (even measurements of abundance) to quantify “proteins”, while

convenient in certain applications, is likely concealing *a lot* of underlying protein dynamics related to proteoforms (Plubell et al. 2021). For any given protein in any given sample, a plethora of molecular permutations may exist due to PTMs, splicing, and cleavages, among other molecular changes, that drive diversity. Many of these biological drivers are embedded in the peptides that span these sites (Plubell et al. 2021), as we saw when measuring the effects of the SARS-CoV-2 protease NSP5 on protein turnover and stability; we could identify substrates of the enzyme by looking for peptide-level changes in turnover or stability in regions flanking the cleavage sites. Protein-level quantification still has its place in proteomics. However, if protein-level quantification is required at all, there has to be care and caution on which peptides should be chosen for consolidation, as differences in the peptides acquired due to sample differences, batch effects, or instrument methods (e.g. DDA) may confound or be misleading to the true underlying biology in a system. A move towards less biased instrument methods, such as Data-Independent Acquisition, may mitigate some of these issues, but still fails to address the biggest constraint - which is that peptides are the biological molecules quantified at the end of the day, not proteins.

Future applications

Where else will the technologies and ideas presented in this thesis have an impact? I want to briefly conclude by looking at two areas of applications.

I. What are the consequences of other forms of sequence variation?

Amino acid substitutions are just a subset of the possible diversity across the proteome. Post-translational modifications represent an old, yet exciting frontier that is garnering significant interest in the last few years due to advances in quantitative proteomics and sample throughput.

Will probing the effects of PTMs on protein biochemical properties yield novel biological insights into protein function? Potentially. Early signs suggest these technologies and applications should provide valuable insights from coupling protein turnover or thermal stability with collections of phosphorylated or ubiquitinated proteins (Smith et al. 2021; Potel et al. 2021; Zecha et al. 2018, 2022; J. X. Huang et al. 2019).

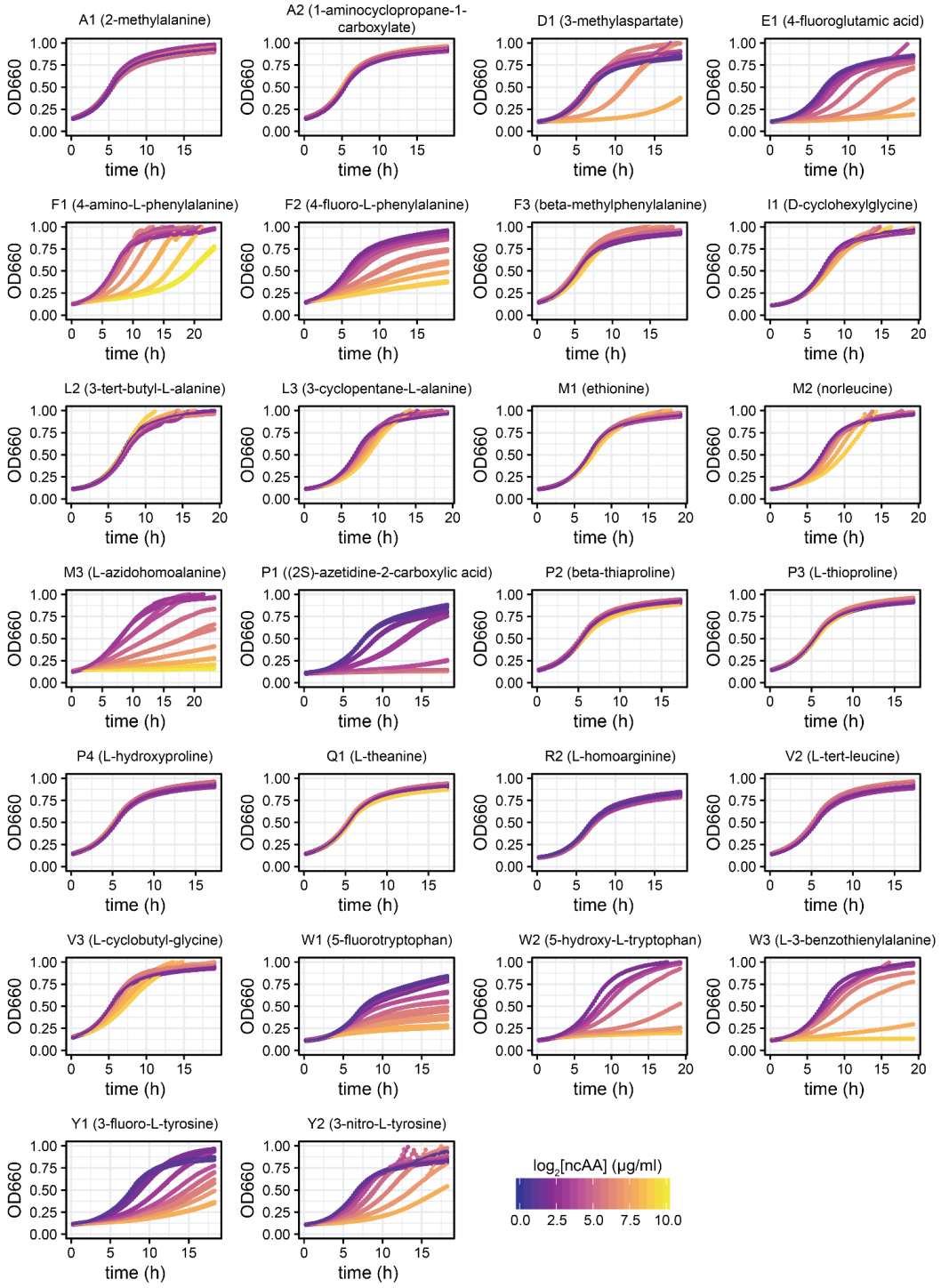
II. What will the convergence of genomics and proteomics look like in the coming decade?

Several clues can be found in recently-developed technologies that suggest increased harmonization moving forward. For one, peptide barcoding is a promising technology that has yet to be explored to its full potential (Egloff et al. 2019). The few applications so far have done a good job at characterizing collections of mutant proteins. However, they have yet to be used for multidimensional molecular phenotyping purposes. I imagine this will change in the near future. For another, developments in adjacent fields are establishing inroads for proteomics in areas where it can have an immediate impact without needing to change much current technology. For example, proteomics, even in the absence of barcodes, could be coupled to collections of designed proteins to characterize synthetic protein properties, such as oligomerization, protein-protein interactions, or thermal stability and solubility. As long as these proteins are diverse in their primary amino acid sequence, the peptides they generate should match the complexity of a conventional bottom-up proteomics sample. As another case in point, conventional mutational libraries are hampered by low diversity and high similarity in sequence space. However, continuous evolution and directed evolution platforms, such as OrthoRep (Ravikumar et al. 2018) and PACE (Esvelt, Carlson, and Liu 2011) generate proteins that are several mutation steps away from the “ancestral” starting protein sequence. Some of these technologies have even been used to generate diverse collections of protein orthologs from a common ancestor that maintain the ancestral function, but have evolved ancillary functions for synthetic biology purposes (Rix et al. 2020). Given the degree of sequence variation within

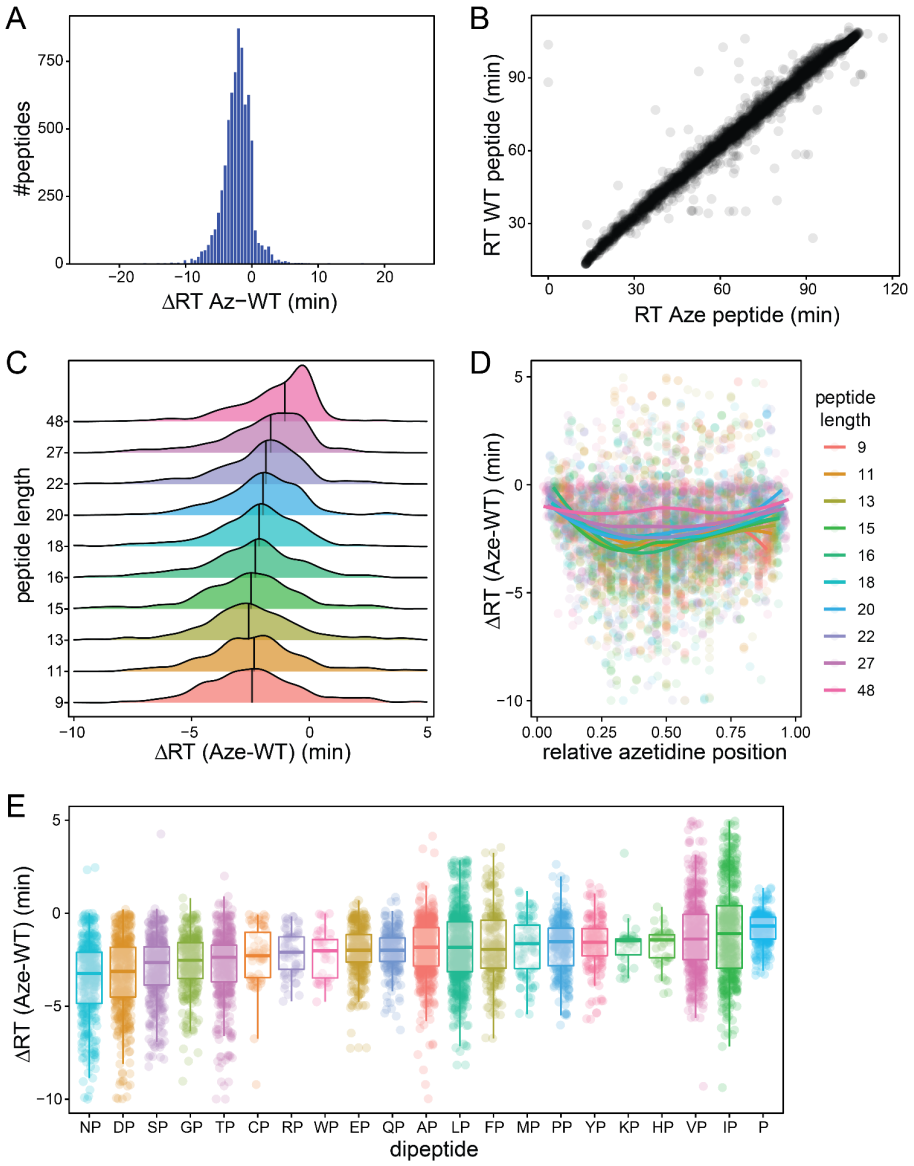
these proteins, mass spectrometry-based proteomics assays offer a unique opportunity to characterize the functions of these orthologs in high-throughput. In turn, coupling these multidimensional profiling maps of diverse protein orthologs with computational advances and machine learning will yield fundamental insights into protein evolution and sequence-function relationships.

Concluding remarks

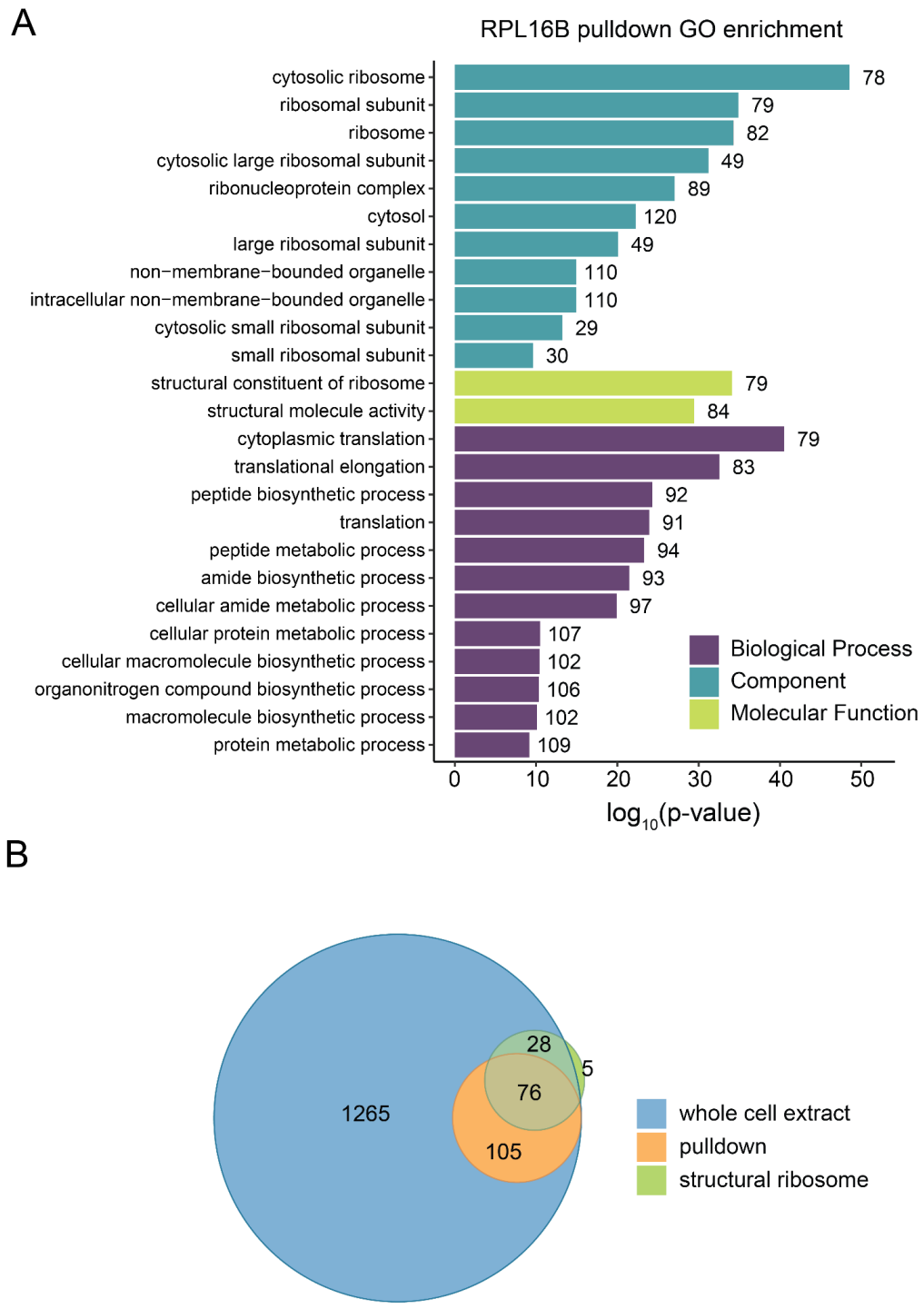
Protein biochemistry and sequence-function relationships have been at the core of scientific inquiry for decades, and are now front-and-center due to rapid technological innovations. In turn, scientists have been able to leverage our understanding of protein biology to manipulate and engineer the world around us. Meat alternatives, CAR-T cell therapies, antivirals for COVID-19, and prenatal genetic testing are modern day examples of this impact. How long will this revolution last in protein biochemistry? Given the fundamental role of proteins in genome, organelle, cellular, tissue, and organismal function, I imagine we will see continued progress and impacts throughout this century. Over the next ~80 years, we may see transformative technologies around customized therapies for individuals based on genetics, engineering of synthetic organisms using orthogonal translation systems, and the design and evolution of proteins to tackle climate change. A continued quest to understand protein sequence, structure, and function will be essential to making these futures become a reality.



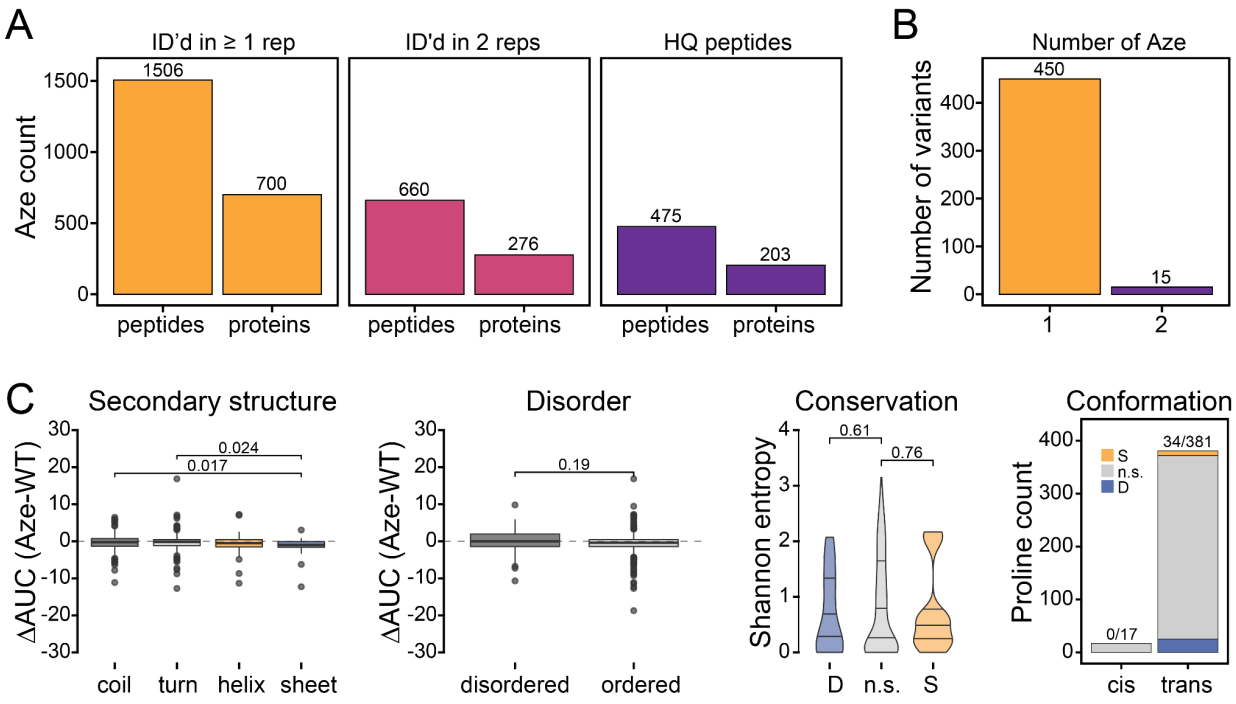
Supplemental Figure 2.1. Toxicity screen for a library of ncAA. Growth curves for S.cerevisiae cultures in the presence of various concentrations of ncAAs (see Supplementary Table 1 for amino acid structures and concentrations).



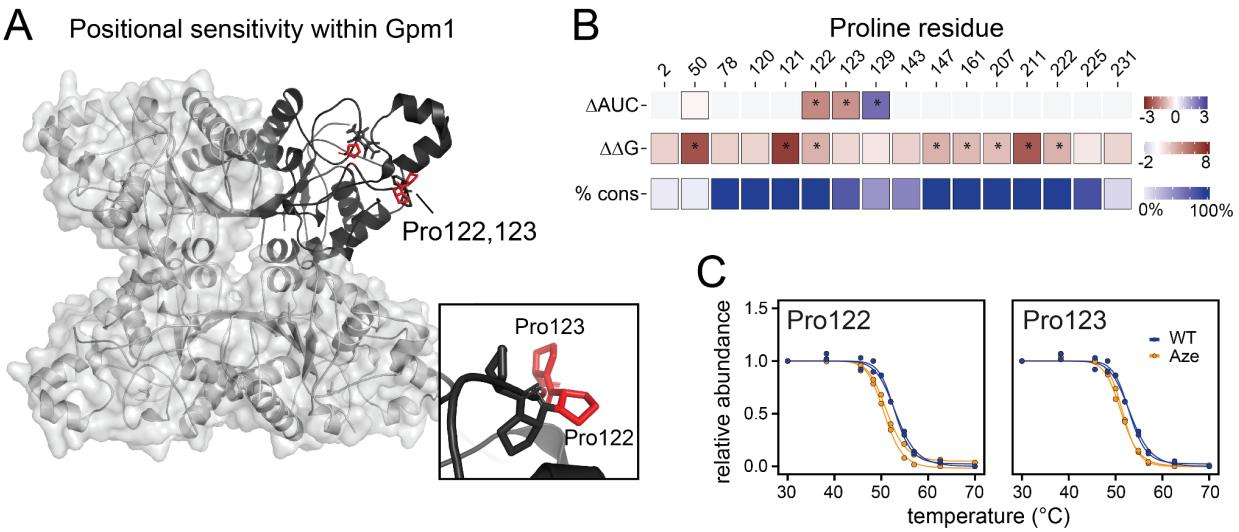
Supplemental Figure 2.2. Effects of azetidine-2-carboxylic acid alters cis-trans equilibrium in peptides and alters chromatographic retention time. A) Histogram of retention time differences between wild type and azetidine-2-carboxylic acid containing peptide. B) Scatterplot showing correlation between chromatographic retention times between s.cerevisiae wild type peptides and their azetidine-2-carboxylic containing counterparts. C) Distribution of retention time differences between wild type and azetidine-2-carboxylic acid containing peptide binned by peptide length. D) Scatter plot of retention time differences between wild type and azetidine-2-carboxylic acid containing peptides and relative position of misincorporated proline site binned and colored by peptide length. E) Boxplot of retention time differences between wild type and azetidine-2-carboxylic acid containing peptide and binned by dipeptide motif where Aze misincorporation occurs.



Supplemental Figure 2.3. Proteins in the ribosome pulldown. A)Significant Gene ontology terms enriched in ribosome purification. B) Venn diagram displaying the number and overlap of protein identifications for whole cell lysate and ribosome pulldown eluate fractions, as well as the coverage for the ribosome.

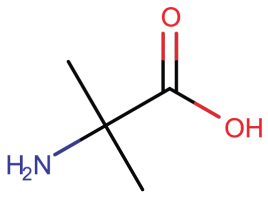
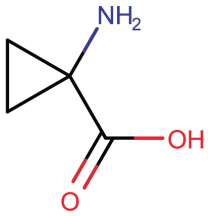
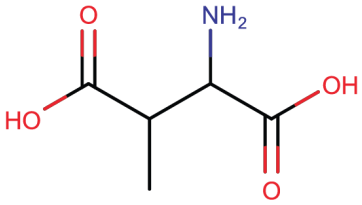
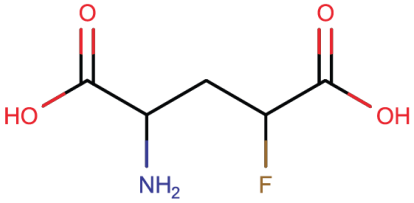


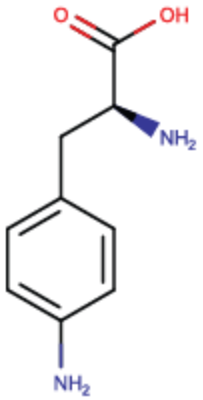
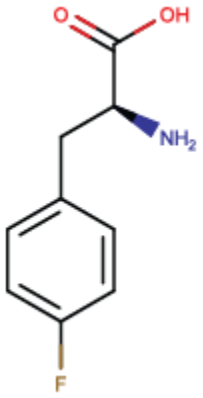
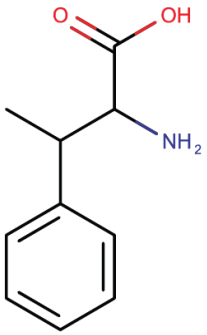
Supplemental Figure 2.4: Thermal proteome profiling on yeast lysates with azetidine-2-carboxylic acid misincorporation. A) Summary statistics of identified unique peptides containing azetidine-2-carboxylic acid and corresponding proteins across two replicate thermal proteome profiling experiments. B) The number of singly- and doubly-modified azetidine substitutions detected. C) Associations between structural and evolutionary features with azetidine sensitivity. From left to right: secondary structure (Coil: 162; Turn: 109; Helix: 96; Sheet: 31), structure disorder predictions (Disordered: 41; Ordered: 409), evolutionary conservation measured as raw Shannon entropy values (Destabilizing: 23; Non-significant: 358; Stabilizing: 9), and proline conformation (Cis: 17; Trans: 381). Listed p-values above each comparison are from Wilcoxon-ranked sum test.

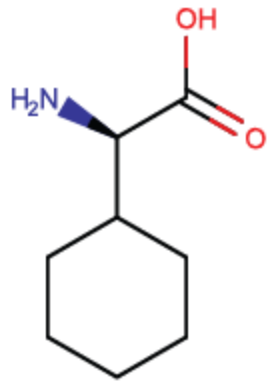
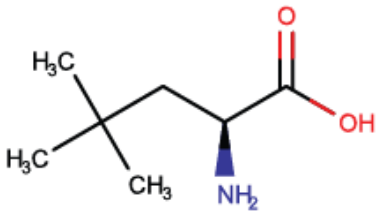
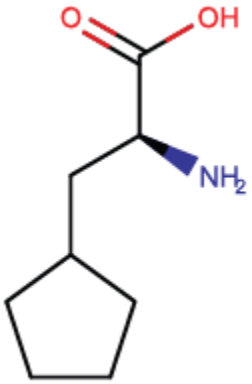
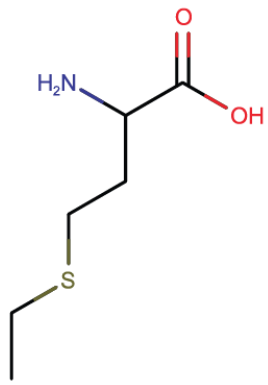


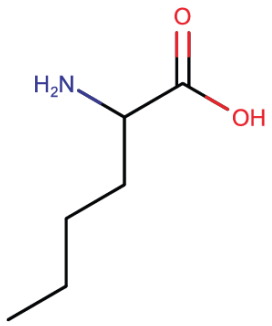
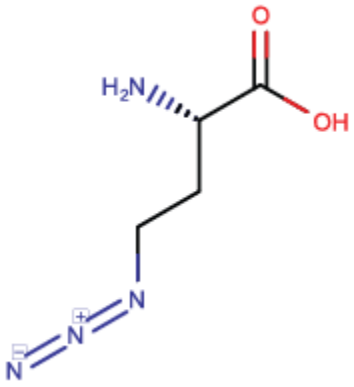
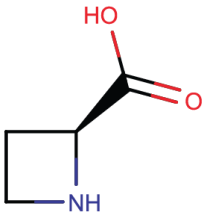
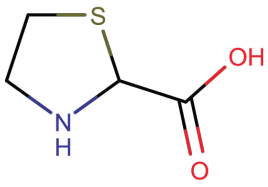
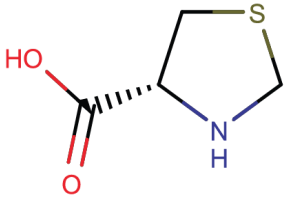
Supplemental Figure 2.5. A conserved polyproline region is important for Gpm1 stability in yeast. A) Crystal structure (PDB ID: 1BQ4) of tetrameric yeast phosphoglycerate mutase (Gpm1) with the positions of destabilizing substitutions highlighted in red. B) Positional sensitivity map of Gpm1, which includes differences in melting curves (ΔAUC), predicted effects of natural substitutions (ΔΔG), and residue conservation across other eukaryotic Gpm1 orthologs. C) Melting curves for two proline-to-azetidine substitutions within a conserved polyproline helix in Gpm1 that decrease protein thermal stability.

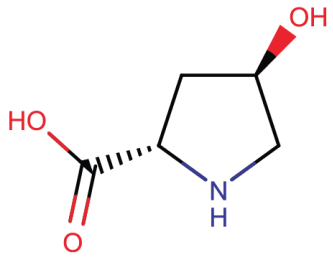
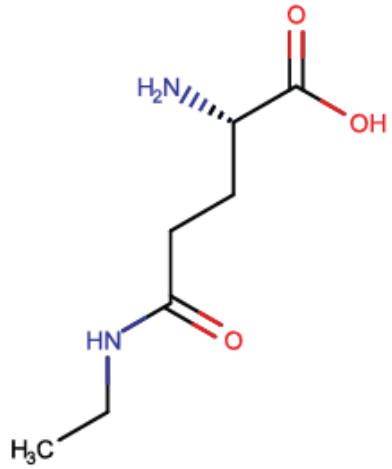
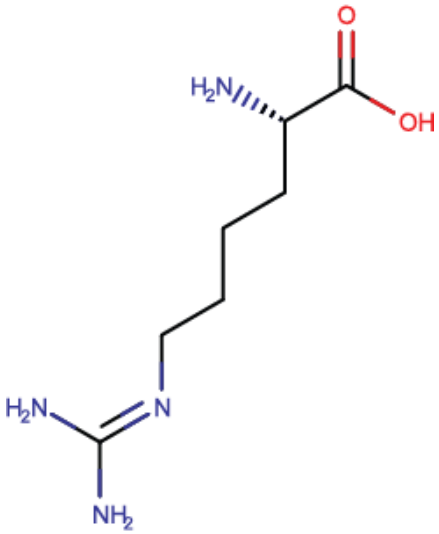
Supplemental Table 2.1. Library of ncAAs with structures and concentrations used in the toxicity and incorporation screens.

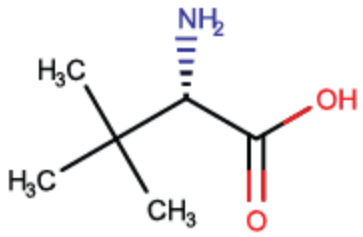
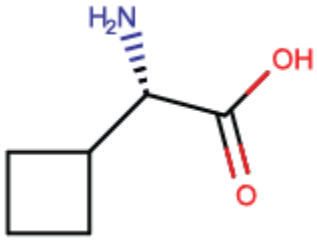
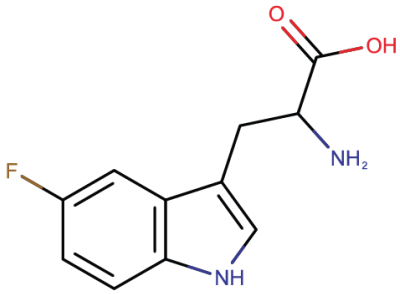
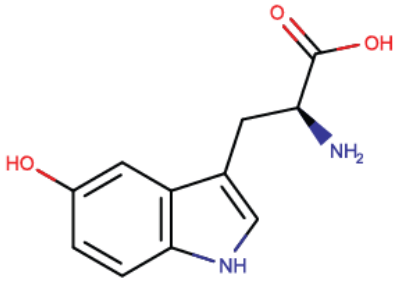
Symbol	Name	Structure	Growth screen concentrations (µg/ml)	Incorporation concentration (µg/ml)
A1	2-methylalanine		3.91 7.81 15.63 31.25 62.5 125 250 500	500
A2	1-aminocyclopropane-1-carboxylic acid		3.91 7.81 15.63 31.25 62.5 125 250 500	500
D1	3-methylaspartic acid		1.96 3.91 7.81 15.63 31.25 62.5 125 250	100
E1	4-fluoroglutamic acid		1.96 3.91 7.81 15.63 31.25 62.5 125 250	80

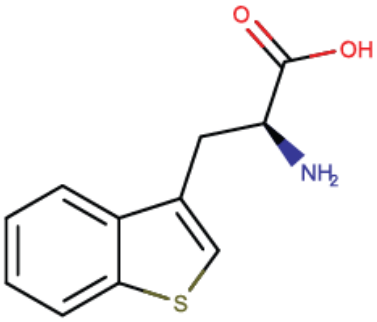
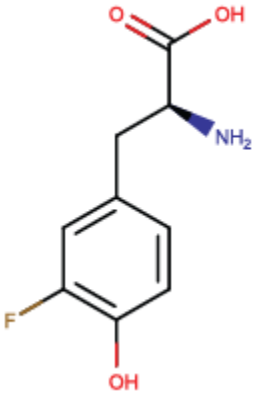
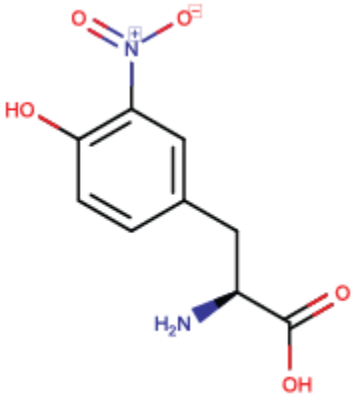
F1	4-amino-L-phenylalanine		7.81 15.63 31.25 62.5 125 250 500 1000	250
F2	4-fluoro-L-phenylalanine		3.91 7.81 15.63 31.25 62.5 125 250 500	30
F3	beta-methylphenylalanine		3.91 7.81 15.63 31.25 62.5 125 250 500	500

I1	D-cyclohexylglycine		3.91 7.81 15.63 31.25 62.5 125 250 500	1000
L2	3-tert-butyl-L-alanine		3.91 7.81 15.63 31.25 62.5 125 250 500	1000
L3	3-cyclopentane-L-alanine		7.81 15.63 31.25 62.5 125 250 500 1000	1000
M1	DL-ethionine		3.91 7.81 15.63 31.25 62.5 125 250 500	1000

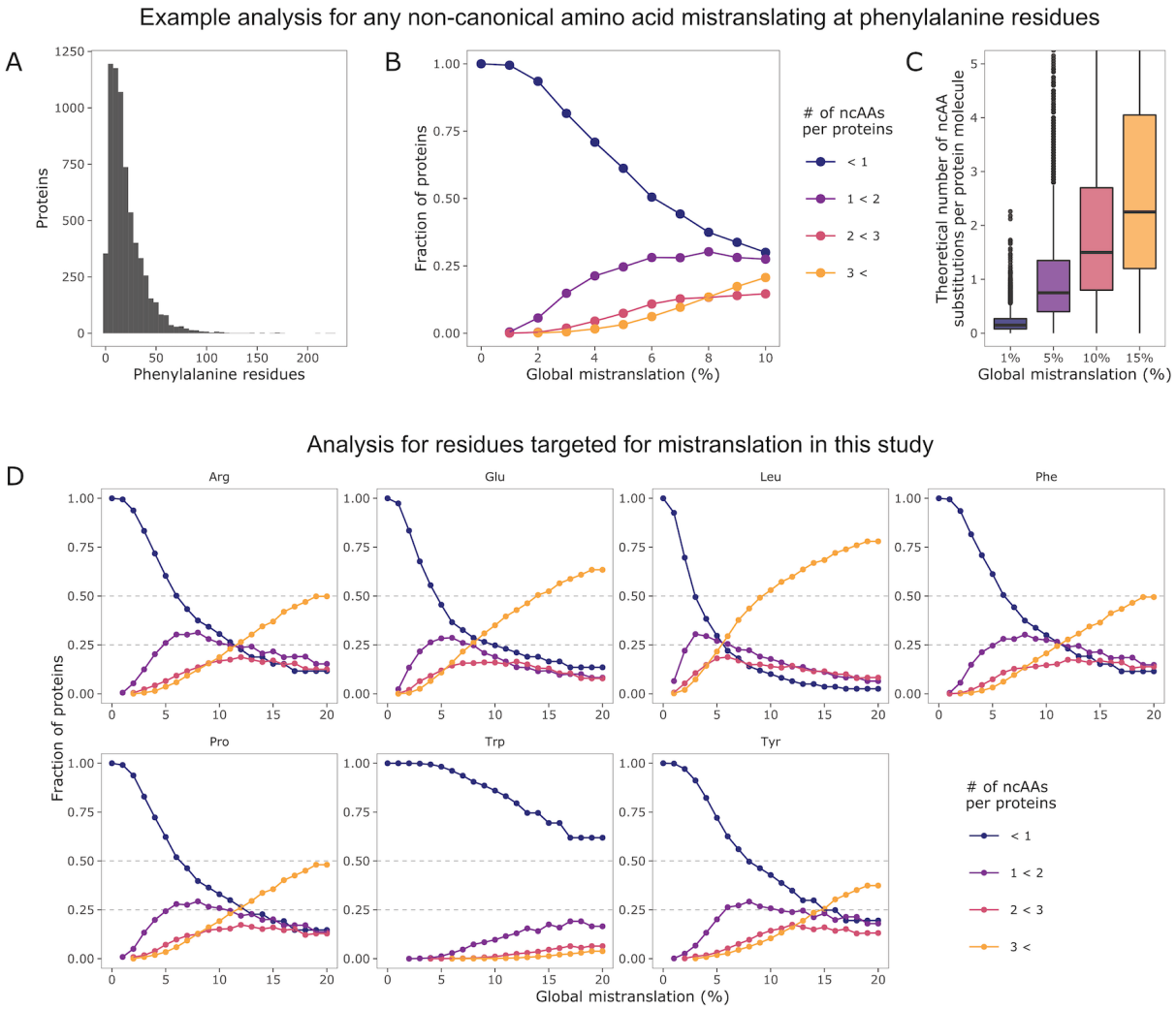
M2	DL-norleucine		7.81 15.63 31.25 62.5 125 250 500 1000	1000
M3	L-azidohomoalanine		7.81 15.63 31.25 62.5 125 250 500 1000	80
P1	(2S)-azetidine-2-carboxylic acid		1.96 3.91 7.81 15.63 31.25 62.5 125 250	100
P2	beta-thiaproline		3.91 7.81 15.63 31.25 62.5 125 250 500	500
P3	L-thioproline		3.91 7.81 15.63 31.25 62.5 125	500

			250 500	
P4	L-hydroxyproline		3.91 7.81 15.63 31.25 62.5 125 250 500	500
Q1	L-theanine		3.91 7.81 15.63 31.25 62.5 125 250 500	500
R2	L-homoarginine		1.96 3.91 7.81 15.63 31.25 62.5 125 250	500

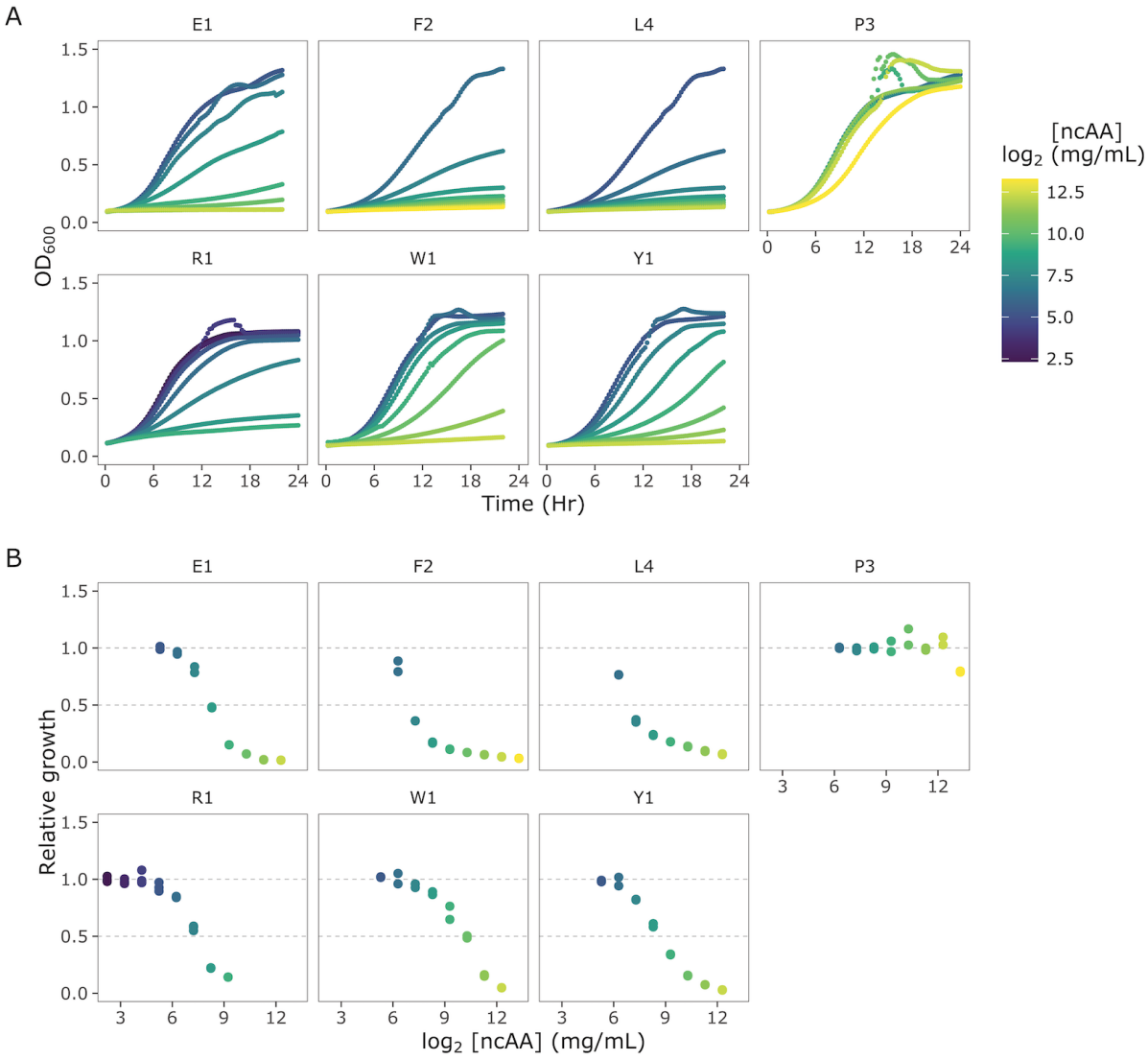
V2	L-tert-leucine		3.91 7.81 15.63 31.25 62.5 125 250 500	500
V3	L-cyclobutyl-glycine		3.91 7.81 15.63 31.25 62.5 125 250 500	250
W1	5-fluorotryptophan		1.96 3.91 7.81 15.63 31.25 62.5 125 250	15
W2	5-hydroxy-L-tryptophan		3.91 7.81 15.63 31.25 62.5 125 250 500	30

W3	L-3-benzothienylalanine		3.91 7.81 15.63 31.25 62.5 125 250 500	80
Y1	3-fluoro-L-tyrosine		1.96 3.91 7.81 15.63 31.25 62.5 125 250	5
Y2	3-nitro-L-tyrosine		1.96 3.91 7.81 15.63 31.25 62.5 125 250	60

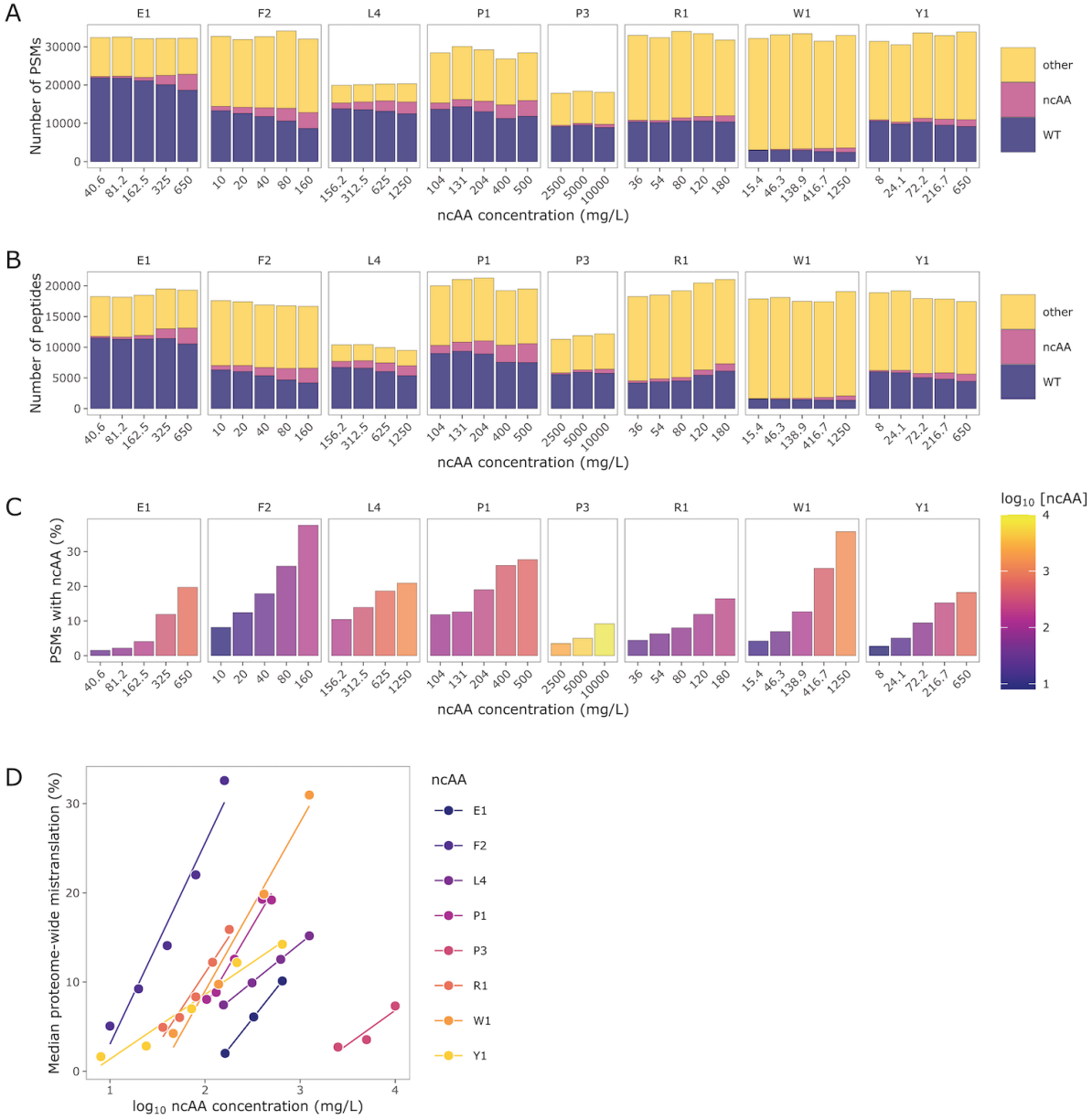
APPENDIX B: SUPPLEMENTAL MATERIAL FOR CHAPTER 3



Supplemental Figure 3.1. Relationship between global mistranslation and protein-level incorporation for the entire yeast proteome. (A) The distribution of phenylalanine residue counts in the amino acid sequences of the yeast proteome. (B) The relationship between theoretical global mistranslation percentage and the fraction of proteins across the yeast proteome that would incorporate a specific amount of ncAA substitutions per round of protein synthesis. All data is for a hypothetical phenylalanine analog. (C) The relationship between theoretical global mistranslation rate and the theoretical number of ncAA substitutions per protein molecule synthesized. Each point is a single protein and the anticipated number of ncAA substitutions events for a hypothetical phenylalanine analog. (D) Fraction of proteome that is mistranslated at different global mistranslation percentages for each of the seven residues targeted in our study.

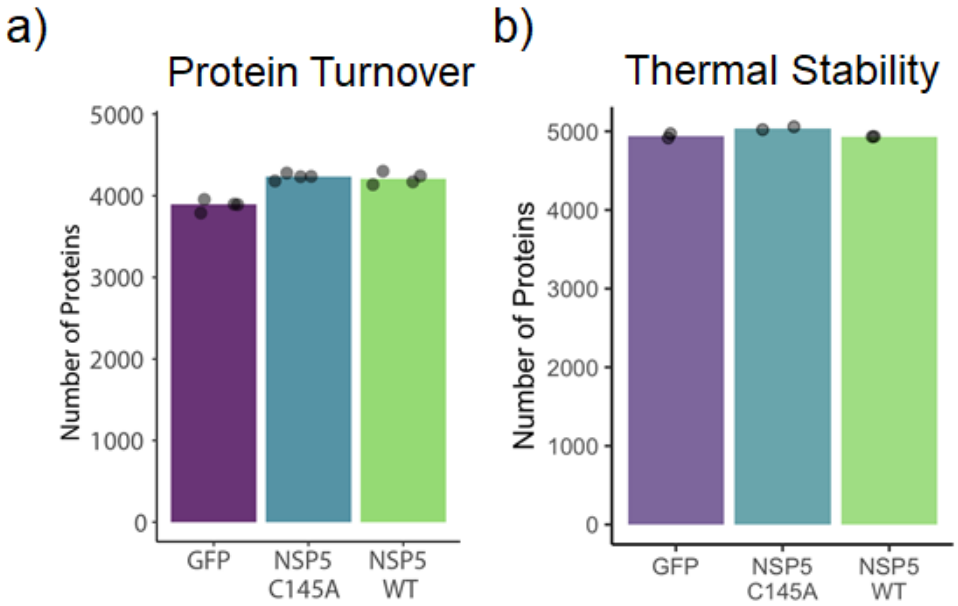


Supplemental Figure 3.2. Dose-dependent toxicity and IC50 estimation for seven ncAAs. (A) Growth curves for yeast grown in synthetic complete media in the presence of one of seven different ncAAs (listed above each plot) at different concentrations (represented by the color). Azetidine (P1) was not included in this initial screen. (B) Dose-response relationships between ncAA concentration and relative area under the growth curve, labeled as relative growth.

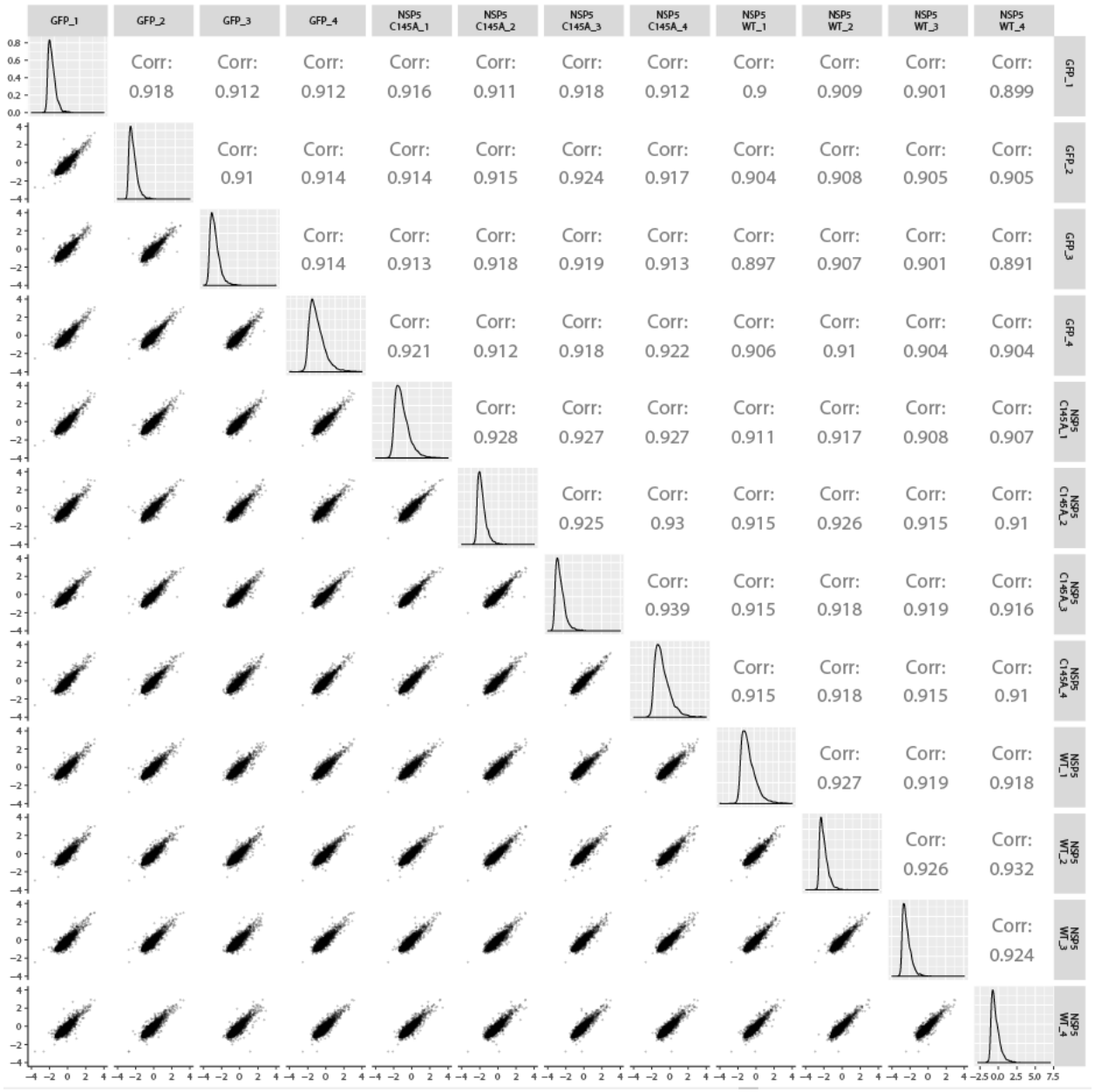


Supplemental Figure 3.3. Proteome-wide mistranslation is tunable and increases linearly. (A-B) Total number of (A) PSMs and (B) unique peptides per mistranslated proteome across eight different ncAAs. PSMs and peptides labeled as: (1) ncAA contain the substitution of interest; (2) WT contain the cognate amino acid that is targeted for mistranslation by the indicated ncAA; (3) other does not contain a ncAA substitution or a possible site for substitution. (C) The percent of PSMs that contain a ncAA relative to the total number of possible PSMs (ncAA PSM count divided by ncAA plus WT PSM count). (D) Dose-response relationship between ncAA concentration and median proteome-wide mistranslation, as determined by MS1 signal intensity for ncAA-containing peptides and WT peptides in each mistranslated proteome.

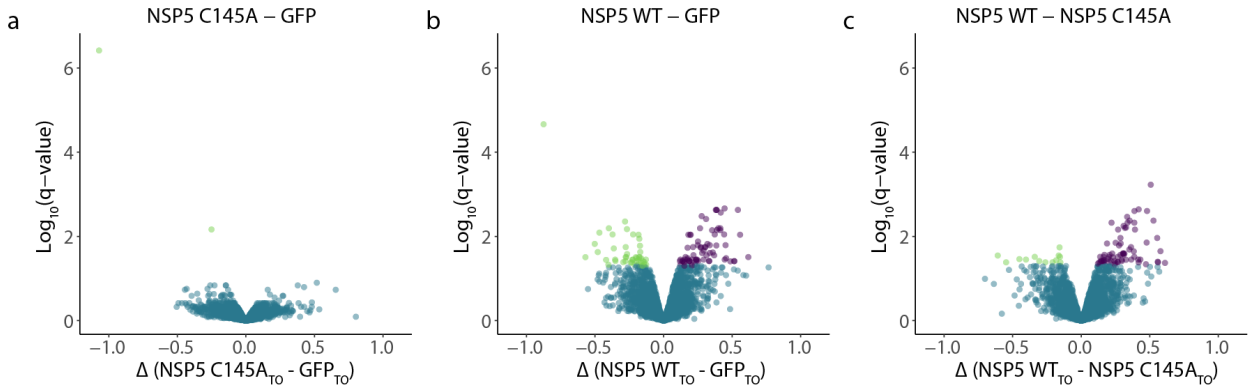
APPENDIX C: SUPPLEMENTAL MATERIAL FOR CHAPTER 5



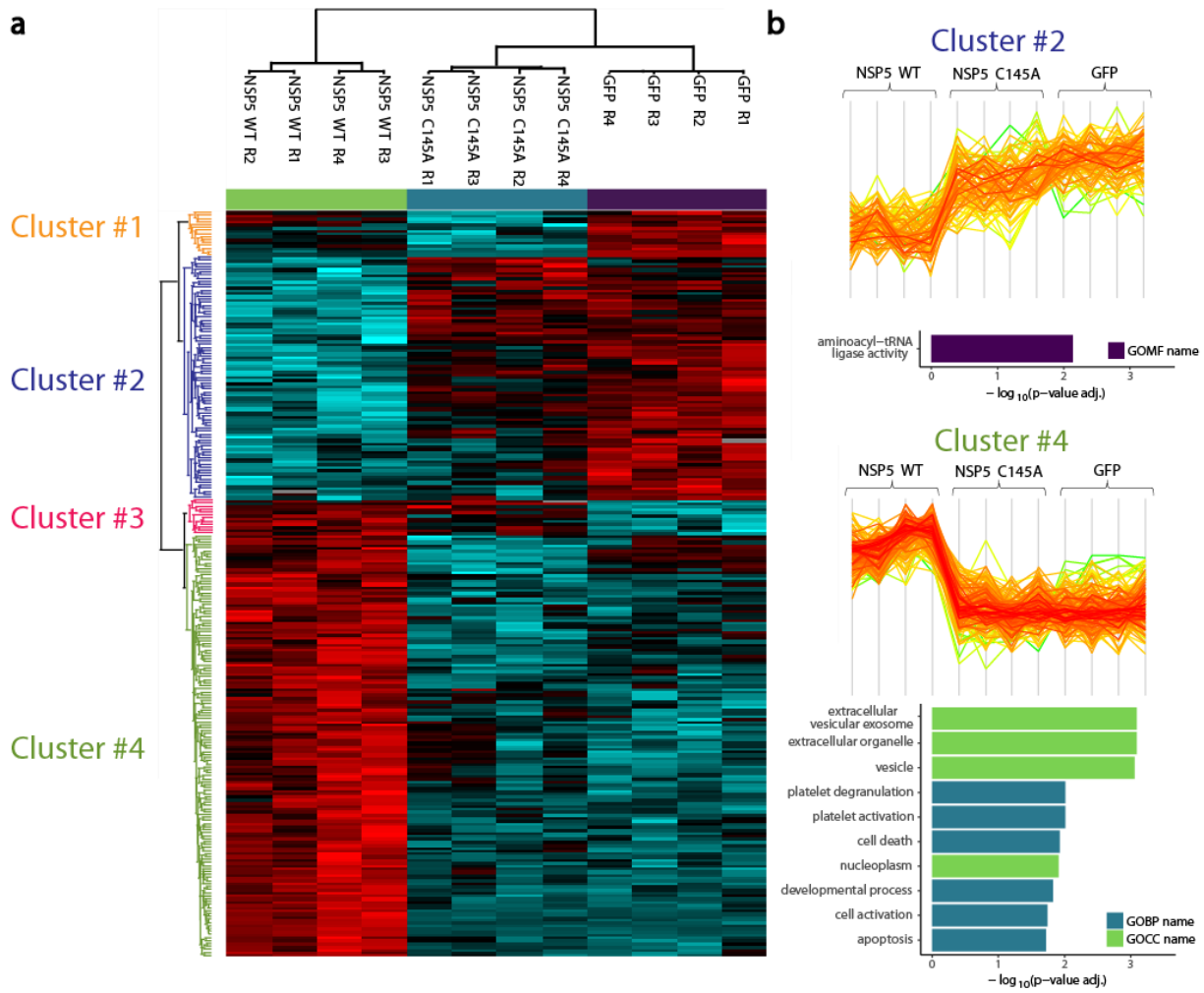
Supplementary Figure 5.1: Protein identifications for protein turnover and protein thermal stability assay. a) For protein turnover, bar plot of protein identifications of HEK293T proteomes overexpressing GFP (purple), NSP5 C145A (blue), and NSP5 wildtype (green) with points representing identifications for each replicate. **b)** Same as (**a**) for protein thermal stability assay (thermal proteome profiling: TPP).



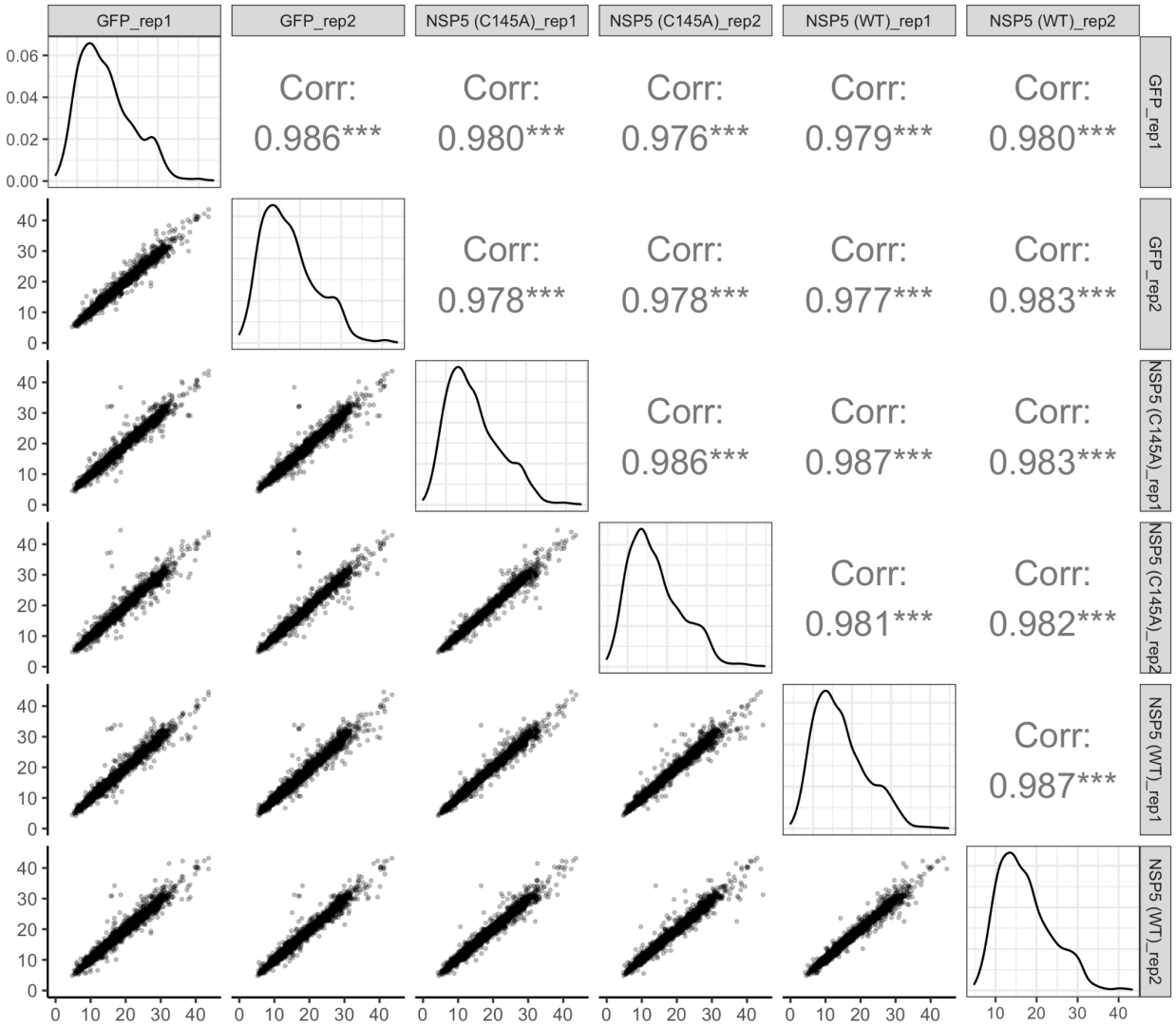
Supplementary Figure 5.2: Dynamic SILAC replicate reproducibility. a) The lower triangle contains scatter plots comparing R_{TO} for all pairwise replicates (within and across protein overexpression conditions). Each point represents a protein. The density plots along the diagonal are for each replicate's R_{TO} distribution across conditions. The upper triangle contains the Pearson Correlation R for all pairwise replicates across all the overexpression conditions.



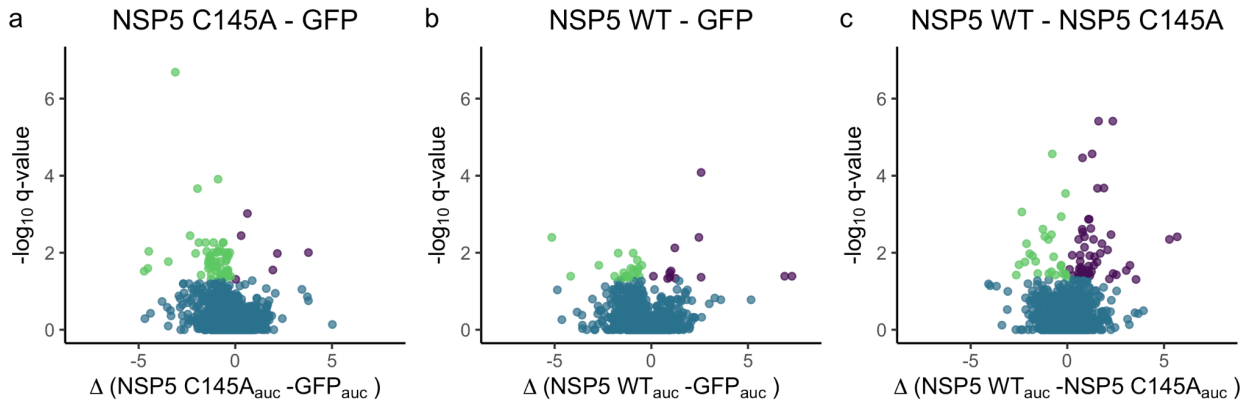
Supplementary Figure 5.3: Pairwise Limma analysis of overexpression conditions. a) Pairwise statistical comparison between NSP5 C145A and GFP with Limma. The x-axis designates the difference in median replicate R_{TO} between the conditions or ΔR_{TO}. The y-axis designates the negative log10 of the 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 5.4: 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 5.5: 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 5.6: 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.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Anfinsen, C. B. 1973. “Principles That Govern the Folding of Protein Chains.” *Science* 181 (4096): 223–30.

Anfinsen, C. B., E. Haber, M. Sela, and F. H. White Jr. 1961. “The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain.” *Proceedings of the National Academy of Sciences of the United States of America* 47 (September): 1309–14.

Anfinsen, C. B., R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll. 1954. “Studies on the Gross Structure, Cross-Linkages, and Terminal Sequences in Ribonuclease.” *The Journal of Biological Chemistry* 207 (1): 201–10.

Anfinsen, Christian B. 1959. *The Molecular Basis of Evolution*. Wiley.

Arnez, J. G., A. C. Dock-Bregeon, and D. Moras. 1999. “Glycyl-tRNA Synthetase Uses a Negatively Charged Pit for Specific Recognition and Activation of Glycine.” *Journal of Molecular Biology* 286 (5): 1449–59.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. “Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.” *Science* 373 (6557): 871–76.

Banerjee, Abhik K., Mario R. Blanco, Emily A. Bruce, Drew D. Honson, Linlin M. Chen, Amy Chow, Prashant Bhat, et al. 2020. “SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses.” *Cell* 0 (0). https://doi.org/10.1016/j.cell.2020.10.004.

Baum, B. J., L. S. Johnson, C. Franzblau, and R. F. Troxler. 1975. “Incorporation of L-Azetidine-2-Carboxylic Acid into Hemoglobin in Rabbit Reticulocytes in Vitro.” *The Journal of Biological Chemistry* 250 (4): 1464–71.

Beausoleil, Sean A., Judit Villén, Scott A. Gerber, John Rush, and Steven P. Gygi. 2006. “A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization.” *Nature Biotechnology* 24 (10): 1285–92.

Becher, Isabelle, Amparo Andrés-Pons, Natalie Romanov, Frank Stein, Maike Schramm, Florence Baudin, Dominic Helm, et al. 2018. “Pervasive Protein Thermal Stability Variation during the Cell Cycle.” *Cell* 173 (6): 1495–1507.e18.

Bekker-Jensen, Dorte B., Ana Martínez-Val, Sophia Steigerwald, Patrick Rüther, Kyle L. Fort, Tabiwang N. Arrey, Alexander Harder, Alexander Makarov, and Jesper V. Olsen. 2020. “A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients.” *Molecular & Cellular Proteomics: MCP* 19 (4): 716–29.

Beltrao, Pedro, Véronique Albanèse, Lillian R. Kenner, Danielle L. Swaney, Alma Burlingame, Judit Villén, Wendell A. Lim, James S. Fraser, Judith Frydman, and Nevan J. Krogan. 2012. “Systematic Functional Prioritization of Protein Posttranslational Modifications.” *Cell* 150 (2): 413–25.

Berg, Matthew D., Yanrui Zhu, Julie Genereaux, Bianca Y. Ruiz, Ricard A. Rodriguez-Mias, Tyler Allan, Alexander Bahcheli, Judit Villén, and Christopher J. Brandl. 2019. “Modulating Mistranslation Potential of tRNASer in Saccharomyces Cerevisiae.” *Genetics* 213 (3): 849–63.

Bludau, Isabell, Max Frank, Christian Dörig, Yujia Cai, Moritz Heusel, George Rosenberger, Paola Picotti, Ben C. Collins, Hannes Röst, and Ruedi Aebersold. 2021. “Systematic Detection of Functional Proteoform Groups from Bottom-up Proteomic Datasets.” *Nature Communications* 12 (1): 3810.

Bojkova, Denisa, Kevin Klann, Benjamin Koch, Marek Widera, David Krause, Sandra Ciesek, Jindrich Cinatl, and Christian Münch. 2020. “Proteomics of SARS-CoV-2-Infected Host Cells Reveals Therapy Targets.” *Nature* 583 (7816): 469–72.

Bouhaddou, Mehdi, Danish Memon, Bjoern Meyer, Kris M. White, Veronica V. Rezelj, Miguel Correa Marrero, Benjamin J. Polacco, et al. 2020. “The Global Phosphorylation Landscape of SARS-CoV-2 Infection.” *Cell* 182 (3): 685–712.e19.

Budisa, Nediljko. 2006. *Engineering the Genetic Code: Expanding the Amino Acid Repertoire for the Design of Novel Proteins*. John Wiley & Sons.

Cavarelli, J., B. Delagoutte, G. Eriani, J. Gangloff, and D. Moras. 1998. “L-Arginine Recognition by Yeast Arginyl-tRNA Synthetase.” *The EMBO Journal* 17 (18): 5438–48.

Cervettini, Daniele, Shan Tang, Stephen D. Fried, Julian C. W. Willis, Louise F. H. Funke, Lucy J. Colwell, and Jason W. Chin. 2020. “Rapid Discovery and Evolution of Orthogonal Aminoacyl-tRNA Synthetase-tRNA Pairs.” *Nature Biotechnology* 38 (8): 989–99.

Childs, Dorothee, Karsten Bach, Holger Franken, Simon Anders, Nils Kurzawa, Marcus Bantscheff, Mikhail M. Savitski, and Wolfgang Huber. 2019. “Nonparametric Analysis of Thermal Proteome Profiles Reveals Novel Drug-Binding Proteins.” *Molecular & Cellular Proteomics: MCP* 18 (12): 2506–15.

Christiano, Romain, Nagarjuna Nagaraj, Florian Fröhlich, and Tobias C. Walther. 2014. “Global Proteome Turnover Analyses of the Yeasts S. Cerevisiae and S. Pombe.” *Cell Reports* 9 (5): 1959–65.

Cirino, Patrick C., Kimberly M. Mayer, and Daisuke Umeno. 2003. “Generating Mutant Libraries Using Error-Prone PCR.” *Methods in Molecular Biology* 231: 3–9.

Cong, Qian, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. 2019. “Protein Interaction Networks Revealed by Proteome Coevolution.” *Science* 365 (6449): 185–89.

Cornish, Virginia W., David Mendel, and Peter G. Schultz. 1995. “Probing Protein Structure and Function with an Expanded Genetic Code.” *Angewandte Chemie* 34 (6): 621–33.

Courbet, A., J. Hansen, Y. Hsia, N. Bethel, Y-J Park, C. Xu, A. Moyer, et al. 2022. “Computational Design of Mechanically Coupled Axle-Rotor Protein Assemblies.” *Science* 376 (6591): 383–90.

Cowie, D. B., G. N. Cohen, E. T. Bolton, and H. de Robichon-Szulmajster. 1959. “Amino Acid Analog Incorporation into Bacterial Proteins.” *Biochimica et Biophysica Acta* 34 (July): 39–46.

Cox, Jürgen, and Matthias Mann. 2008. “MaxQuant Enables High Peptide Identification Rates, Individualized P.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification.” *Nature Biotechnology* 26 (12): 1367–72.

Cox, Jürgen, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, Matthias Selbach, Jesper V. Olsen, and Matthias Mann. 2009. “A Practical Guide to the MaxQuant Computational Platform for SILAC-Based Quantitative Proteomics.” *Nature Protocols* 4 (5): 698–705.

Cunningham, B. C., and J. A. Wells. 1989. “High-Resolution Epitope Mapping of hGH-Receptor Interactions by Alanine-Scanning Mutagenesis.” *Science* 244 (4908): 1081–85.

Dai, Lingyun, Tianyun Zhao, Xavier Bisteau, Wendi Sun, Nayana Prabhu, Yan Ting Lim, Radoslaw M. Sobota, Philipp Kaldis, and Pär Nordlund. 2018. “Modulation of Protein-Interaction States through the Cell Cycle.” *Cell* 173 (6): 1481–94.e13.

Delagoutte, B., D. Moras, and J. Cavarelli. 2000. “tRNA Aminoacylation by Arginyl-tRNA Synthetase: Induced Conformations during Substrates Binding.” *The EMBO Journal* 19 (21): 5599–5610.

Després, Philippe C., Alexandre K. Dubé, Motoaki Seki, Nozomu Yachie, and Christian R. Landry. 2020. “Perturbing Proteomes at Single Residue Resolution Using Base Editing.” *Nature Communications* 11 (1): 1–13.

Doolittle, R. F. 1994. “Convergent Evolution: The Need to Be Explicit.” *Trends in Biochemical Sciences* 19 (1): 15–18.

Egloff, Pascal, Iwan Zimmermann, Fabian M. Arnold, Cedric A. J. Hutter, Damien Morger, Lennart Opitz, Lucy Poveda, et al. 2019. “Engineered Peptide Barcodes for in-Depth Analyses of Binding Protein Libraries.” *Nature Methods* 16 (5): 421–28.
Eng, Jimmy K., Michael R. Hoopmann, Tahmina A. Jahan, Jarrett D. Egertson, William S. Noble, and Michael J. MacCoss. 2015. “A Deeper Look into Comet—Implementation and Features.” *Journal of the American Society for Mass Spectrometry* 26 (11): 1865–74.
Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. 2013. “Comet: An Open-Source MS/MS Sequence Database Search Tool.” *Proteomics* 13 (1): 22–24.
Esvelt, Kevin M., Jacob C. Carlson, and David R. Liu. 2011. “A System for the Continuous Directed Evolution of Biomolecules.” *Nature* 472 (7344): 499–503.
Faure, Andre J., Júlia Domingo, Jörn M. Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. 2022. “Mapping the Energetic and Allosteric Landscapes of Protein Binding Domains.” *Nature* 604 (7904): 175–83.
Faust, Ofrah, Meital Abayev-Avraham, Anne S. Wentink, Michael Maurer, Nadinath B. Nillegoda, Nir London, Bernd Bukau, and Rina Rosenzweig. 2020. “HSP40 Proteins Use Class-Specific Regulation to Drive HSP70 Functional Diversity.” *Nature*, November. https://doi.org/10.1038/s41586-020-2906-4.
Feng, Yuehan, Giorgia De Franceschi, Abdullah Kahraman, Martin Soste, Andre Melnik, Paul J. Boersema, Patrizia Polverino De Laureto, Yaroslav Nikolaev, Ana Paula Oliveira, and Paola Picotti. 2014. “Global Analysis of Protein Structural Changes in Complex Proteomes.” *Nature Biotechnology* 32 (10): 1036–44.
Feng, Yuehan, Giorgia De Franceschi, Abdullah Kahraman, Martin Soste, Andre Melnik, Paul J. Boersema, Patrizia Polverino de Laureto, Yaroslav Nikolaev, Ana Paula Oliveira, and Paola Picotti. 2014. “Global Analysis of Protein Structural Changes in Complex Proteomes.” *Nature Biotechnology* 32 (10): 1036–44.
Fersht, Alan R. 1977. “Editing Mechanisms in Protein Synthesis. Rejection of Valine by the Isoleucyl-tRNA Synthetase.” *Biochemistry* 16 (5): 1025–30.
Fersht, A. R., and C. Dingwall. 1979. “Evidence for the Double-Sieve Editing Mechanism in Protein Synthesis. Steric Exclusion of Isoleucine by Valyl-tRNA Synthetases.” *Biochemistry* 18 (12): 2627–31.
Findlay, Gregory M., Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 2018. “Accurate Classification of BRCA1 Variants with Saturation Genome Editing.” *Nature* 562 (7726): 217–22.
Flynn, Julia M., Neha Samant, Gily Schneider-Nachum, David T. Barkan, Nese Kurt Yilmaz, Celia A. Schiffer, Stephanie A. Moquin, Dustin Dovala, and Daniel N. A. Bolon. 2022. “Comprehensive Fitness Landscape of SARS-CoV-2 Mpro Reveals Insights into Viral Resistance Mechanisms.” *bioRxiv*. https://doi.org/10.1101/2022.01.26.477860.
Fowden, L., and M. H. Richmond. 1963. “Replacement of Proline by Azetidine-2-Carboxylic Acid during Biosynthesis of Protein.” *Biochimica et Biophysica Acta* 71 (January): 459–61.
Fowler, Douglas M., Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. 2010. “High-Resolution Mapping of Protein Sequence-Function Relationships.” *Nature Methods* 7 (9): 741–46.
Fowler, Douglas M., and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of Protein Science.” *Nature Methods* 11 (8): 801–7.
Frazer, Jonathan, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. 2021. “Disease Variant Prediction with Deep Generative Models of Evolutionary Data.” *Nature*, October, 1–5.
Fukai, S., O. Nureki, S. Sekine, A. Shimada, J. Tao, D. G. Vassylyev, and S. Yokoyama. 2000. “Structural Basis for Double-Sieve Discrimination of L-Valine from L-Isoleucine and L-Threonine by the Complex of tRNA(Val) and Valyl-tRNA Synthetase.” *Cell* 103 (5):

793–803.
Gaetani, Massimiliano, Pierre Sabatier, Amir A. Saei, Christian M. Beusch, Zhe Yang, Susanna L. Lundström, and Roman A. Zubarev. 2019. “Proteome Integral Solubility Alteration: A High-Throughput Proteomics Assay for Target Deconvolution.” *Journal of Proteome Research* 18 (11): 4027–37.
Galperin, M. Y., D. R. Walker, and E. V. Koonin. 1998. “Analogous Enzymes: Independent Inventions in Enzyme Evolution.” *Genome Research* 8 (8): 779–90.
Garrett, Meghan E., Jared Galloway, Helen Y. Chu, Hannah L. Itell, Caitlin I. Stoddard, Caitlin R. Wolf, Jennifer K. Logue, et al. 2021. “High Resolution Profiling of Pathways of Escape for SARS-CoV-2 Spike-Binding Antibodies.” *Cell* 0 (0). https://doi.org/10.1016/j.cell.2021.04.045.
Gelperin, Daniel M., Michael A. White, Martha L. Wilkinson, Yoshiko Kon, Li A. Kung, Kevin J. Wise, Nelson Lopez-Hoyo, et al. 2005. “Biochemical and Genetic Analysis of the Yeast Proteome with a Movable ORF Collection.” *Genes & Development*. https://doi.org/10.1101/gad.1362105.
Gerhart, J. C., and A. B. Pardee. 1962. “The Enzymology of Control by Feedback Inhibition.” *The Journal of Biological Chemistry* 237 (March): 891–96.
Geyer, Philipp E., Florian M. Arend, Sophia Doll, Marie-Luise Louiset, Sebastian Virreira Winter, Johannes B. Müller-Reif, Furkan M. Torun, et al. 2021. “High-Resolution Serum Proteome Trajectories in COVID-19 Reveal Patient-Specific Seroconversion.” *EMBO Molecular Medicine* 13 (8): e14167.
Geyer, Philipp E., Nils A. Kulak, Garwin Pichler, Lesca M. Holdt, Daniel Teupser, and Matthias Mann. 2016. “Plasma Proteome Profiling to Assess Human Health and Disease.” *Cell Systems* 2 (3): 185–95.
Gilbert, W., and B. Muller-Hill. 1966. “ISOLATION OF THE LAC REPRESSOR.” *Proceedings of the National Academy of Sciences* 56 (6): 1891–98.
Gordon, David E., Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M. White, Matthew J. O’Meara, et al. 2020. “A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing.” *Nature* 583 (7816): 459–68.
Grant, Barry J., Ana P. C. Rodrigues, Karim M. ElSawy, J. Andrew McCammon, and Leo S. D. Caves. 2006. “Bio3d: An R Package for the Comparative Analysis of Protein Structures.” *Bioinformatics* 22 (21): 2695–96.
Grassetti, Andrew V., Rufus Hards, and Scott A. Gerber. 2017. “Offline Pentafluorophenyl (PFP)-RP Prefractionation as an Alternative to High-pH RP for Comprehensive LC-MS/MS Proteomics and Phosphoproteomics.” *Analytical and Bioanalytical Chemistry* 409 (19): 4615–25.
Greenfield, N., and G. D. Fasman. 1969. “Computed Circular Dichroism Spectra for the Evaluation of Protein Conformation.” *Biochemistry* 8 (10): 4108–16.
Green, Lisa, Brian Houck-Loomis, Andrew Yueh, and Stephen P. Goff. 2012. “Large Ribosomal Protein 4 Increases Efficiency of Viral Recoding Sequences.” *Journal of Virology* 86 (17): 8949–58.
Halabi, Najeeb, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. 2009. “Protein Sectors: Evolutionary Units of Three-Dimensional Structure.” *Cell* 138 (4): 774–86.
Hanna, Ruth E., Mudra Hegde, Christian R. Fagre, Peter C. DeWeirdt, Annabel K. Sangree, Zsofia Szegletes, Audrey Griffith, et al. 2020. “Massively Parallel Assessment of Human Variants with Base Editor Screens.” https://doi.org/10.1101/2020.05.17.100818.
———. 2021. “Massively Parallel Assessment of Human Variants with Base Editor Screens.” *Cell* 184 (4): 1064–80.e20.
Hartman, Matthew C. T., Kristopher Josephson, Chi-Wang Lin, and Jack W. Szostak. 2007. “An Expanded Set of Amino Acid Analogs for the Ribosomal Translation of Unnatural Peptides.” *PloS One* 2 (10): e972.

Hartman, Matthew C. T., Kristopher Josephson, and Jack W. Szostak. 2006. “Enzymatic Aminoacylation of tRNA with Unnatural Amino Acids.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (12): 4356–61.
Hasle, Nicholas, Anthony Cooke, Sanjay Srivatsan, Heather Huang, Jason J. Stephany, Zachary Krieger, Dana Jackson, et al. 2020. “High-Throughput, Microscope-Based Sorting to Dissect Cellular Heterogeneity.” *Molecular Systems Biology* 16 (6): e9442.
Hess, Gaelen T., Laure Frésard, Kyuho Han, Cameron H. Lee, Amy Li, Karlene A. Cimprich, Stephen B. Montgomery, and Michael C. Bassik. 2016. “Directed Evolution Using dCas9-Targeted Somatic Hypermutation in Mammalian Cells.” *Nature Methods* 13 (12): 1036–42.
Hsu, Chloe, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. 2022. “Learning Protein Fitness Models from Evolutionary and Assay-Labeled Data.” *Nature Biotechnology*, January, 1–9.
Huang, Jun X., Gihoon Lee, Kate E. Cavanaugh, Jae W. Chang, Margaret L. Gardel, and Raymond E. Moellering. 2019. “High Throughput Discovery of Functional Protein Modifications by Hotspot Thermal Profiling.” *Nature Methods* 16 (9): 894–901.
Huang, Kuan-Lin, Adam D. Scott, Daniel Cui Zhou, Liang-Bo Wang, Amila Weerasinghe, Abdulkadir Elmas, Ruiyang Liu, et al. 2021. “Spatially Interacting Phosphorylation Sites and Mutations in Cancer.” *Nature Communications* 12 (1): 1–13.
Huang, Po-Ssu, Scott E. Boyken, and David Baker. 2016. “The Coming of Age of de Novo Protein Design.” *Nature*. https://doi.org/10.1038/nature19946.
Huber, Kilian V. M., Karin M. Olek, André C. Müller, Chris Soon Heng Tan, Keiryn L. Bennett, Jacques Colinge, and Giulio Superti-Furga. 2015. “Proteome-Wide Drug and Metabolite Interaction Mapping by Thermal-Stability Profiling.” *Nature Methods* 12 (11): 1055–57.
Humphrey, Sean J., S. Babak Azimifar, and Matthias Mann. 2015. “High-Throughput Phosphoproteomics Reveals in Vivo Insulin Signaling Dynamics.” *Nature Biotechnology* 33 (9): 990–95.
Imami, Koshi, Miha Milek, Boris Bogdanow, Tomoharu Yasuda, Nicolai Kastelic, Henrik Zauber, Yasushi Ishihama, Markus Landthaler, and Matthias Selbach. 2018. “Phosphorylation of the Ribosomal Protein RPL12/uL11 Affects Translation during Mitosis.” *Molecular Cell* 72 (1): 84–98.e9.
Iqbal, Sumaiya, Eduardo Pérez-Palma, Jakob B. Jespersen, Patrick May, David Hoksza, Henrike O. Heyne, Shehab S. Ahmed, et al. 2020. “Comprehensive Characterization of Amino Acid Positions in Protein Structures Reveals Molecular Effect of Missense Variants.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (45): 28201–11.
Isom, Daniel G., Carlos A. Castañeda, Brian R. Cannon, Priya D. Velu, and Bertrand García-Moreno E. 2010. “Charges in the Hydrophobic Interior of Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (37): 16096–100.
Jelesarov, I., and H. R. Bosshard. 1999. “Isothermal Titration Calorimetry and Differential Scanning Calorimetry as Complementary Tools to Investigate the Energetics of Biomolecular Recognition.” *Journal of Molecular Recognition: JMR* 12 (1): 3–18.
Jiang, Yuxiang, Tal Ronnen Oron, Wyatt T. Clark, Asma R. Bankapur, Daniel D’Andrea, Rosalba Lepore, Christopher S. Funk, et al. 2016. “An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy.” *Genome Biology* 17 (1): 184.
Jin, Zhenming, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, et al. 2020. “Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors.” *Nature* 582 (7811): 289–93.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Kabsch, W., and C. Sander. 1983. “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.” *Biopolymers* 22 (12): 2577–2637.

Käll, Lukas, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. 2007. “Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets.” *Nature Methods* 4 (11): 923–25.

Kamel, Wael, Marko Noerenberg, Berati Cerikan, Honglin Chen, Aino I. Järvelin, Mohamed Kammoun, Jeffrey Y. Lee, et al. 2021. “Global Analysis of Protein-RNA Interactions in SARS-CoV-2-Infected Cells Reveals Key Regulators of Infection.” *Molecular Cell* 81 (13): 2851–67.e7.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kern, Dorothee, Mike Schutkowski, and Torbjörn Drakenberg. 1997. “Rotational Barriers of Cis/trans Isomerization of Proline Analogues and Their Catalysis by Cyclophilin.” *Journal of the American Chemical Society* 119 (36): 8403–8.

Klann, Kevin, Denisa Bojkova, Georg Tascher, Sandra Ciesek, Christian Münch, and Jindrich Cinatl. 2020. “Growth Factor Receptor Signaling Inhibition Prevents SARS-CoV-2 Replication.” *Molecular Cell*, August. https://doi.org/10.1016/j.molcel.2020.08.006.

Kleifeld, Oded, Alain Doucet, Ulrich auf dem Keller, Anna Prudova, Oliver Schilling, Rajesh K. Kainthan, Amanda E. Starr, Leonard J. Foster, Jayachandran N. Kizhakkedathu, and Christopher M. Overall. 2010. “Isotopic Labeling of Terminal Amines in Complex Samples Identifies Protein N-Termini and Protease Cleavage Products.” *Nature Biotechnology* 28 (3): 281–88.

Kleifeld, Oded, Alain Doucet, Anna Prudova, Ulrich auf dem Keller, Magda Gioia, Jayachandran N. Kizhakkedathu, and Christopher M. Overall. 2011. “Identifying and Quantifying Proteolytic Events and the Natural N Terminome by Terminal Amine Isotopic Labeling of Substrates.” *Nature Protocols* 6 (10): 1578–1611.

Kong, Andy T., Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. 2017. “MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–based Proteomics.” *Nature Methods* 14 (5): 513–20.

Kuhlman, Brian, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. 2003. “Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.” *Science*. https://doi.org/10.1126/science.1089427.

Kulak, Nils A., Garwin Pichler, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. “Minimal, Encapsulated Proteomic-Sample Processing Applied to Copy-Number Estimation in Eukaryotic Cells.” *Nature Methods* 11 (3): 319–24.

Labeau, Athéna, Luc Fery-Simonian, Alain Lefevre-Utile, Marie Pourcelot, Lucie Bonnet-Madin, Vassili Soumelis, Vincent Lotteau, Pierre-Olivier Vidalain, Ali Amara, and Laurent Meertens. 2022. “Characterization and Functional Interrogation of the SARS-CoV-2 RNA Interactome.” *Cell Reports* 39 (4): 110744.

Laurent, Estelle M. N., Yorgos Sofianatos, Anastassia Komarova, Jean-Pascal Gimeno, Payman Samavarchi Tehrani, Dae-Kyum Kim, Hala Abdouni, et al. 2020. “Global BioID-Based SARS-CoV-2 Proteins Proximal Interactome Unveils Novel Ties between Viral Polypeptides and Host Factors Involved in Multiple COVID19-Associated Mechanisms.” *bioRxiv*. https://doi.org/10.1101/2020.08.28.272955.

Lawrence, Robert T., Elizabeth M. Perez, Daniel Hernández, Chris P. Miller, Kelsey M. Haas, Hanna Y. Irie, Su-In Lee, C. Anthony Blau, and Judit Villén. 2015. “The Proteomic Landscape of Triple-Negative Breast Cancer.” *Cell Reports* 11 (6): 990.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation

in 60,706 Humans." *Nature* 536 (7616): 285–91.
Leman, Julia Koehler, Pawel Szczerbiak, P. Douglas Renfrew, Vladimir Gligorijevic, Daniel Berenberg, Tommi Vatanen, Bryn C. Taylor, et al. 2022. "Sequence-Structure-Function Relationships in the Microbial Protein Universe." *bioRxiv*. https://doi.org/10.1101/2022.03.18.484903.
Leutert, Mario, Ricard A. Rodríguez-Mias, Noelle K. Fukuda, and Judit Villén. 2019. "R2-P2 Rapid-Robotic Phosphoproteomics Enables Multidimensional Cell Signaling Studies." *Molecular Systems Biology* 15 (12): e9021.
Li, Jiaming, Jonathan G. Van Vranken, Joao A. Paulo, Edward L. Huttlin, and Steven P. Gygi. 2020. "Selection of Heating Temperatures Improves the Sensitivity of the Proteome Integral Solubility Alteration Assay." *Journal of Proteome Research* 19 (5): 2159–66.
Li, Jiaming, Jonathan G. Van Vranken, Laura Pontano Vaites, Devin K. Schweppe, Edward L. Huttlin, Chris Etienne, Premchendar Nandhikonda, et al. 2020. "TMTpro Reagents: A Set of Isobaric Labeling Mass Tags Enables Simultaneous Proteome-Wide Measurements across 16 Samples." *Nature Methods* 17 (4): 399–404.
Li, Xiyan, Tara A. Gianoulis, Kevin Y. Yip, Mark Gerstein, and Michael Snyder. 2010. "Extensive In Vivo Metabolite-Protein Interactions Revealed by Large-Scale Systematic Analyses." *Cell* 143 (4): 639–50.
Ludwig, Christina, Ludovic Gillet, George Rosenberger, Sabine Amon, Ben C. Collins, and Ruedi Aebersold. 2018. "Data-Independent Acquisition-Based SWATH-MS for Quantitative Proteomics: A Tutorial." *Molecular Systems Biology* 14 (8): e8126.
Manoharan, T. Herbert, Andrew M. Gulick, Peter Reinemer, Heini W. Dirr, Robert Huber, and William E. Fahl. 1992. "Mutational Substitution of Residues Implicated by Crystal Structure in Binding the Substrate Glutathione to Human Glutathione S-Transferase π." *Journal of Molecular Biology* 226 (2): 319–22.
Mateus, André, Johannes Hevler, Jacob Bobonis, Nils Kurzawa, Malay Shah, Karin Mitosch, Camille V. Goemans, et al. 2020. "The Functional Proteome Landscape of Escherichia Coli." *Nature*, December. https://doi.org/10.1038/s41586-020-3002-5.
Mateus, Andre, Mikhail M. Savitski, and Ilaria Piazza. 2021. "The Rise of Proteome-Wide Biophysics." *Molecular Systems Biology* 17 (7): e10442.
Matreyek, Kenneth A., Lea M. Starita, Jason J. Stephany, Beth Martin, Melissa A. Chiasson, Vanessa E. Gray, Martin Kircher, et al. 2018. "Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing." *Nature Genetics* 50 (6): 874–82.
Matsuzaki, Yusei, Wataru Aoki, Takumi Miyazaki, Shunsuke Aburaya, Yuta Ohtani, Kaho Kajiwara, Naoki Koike, et al. 2021. "Peptide Barcoding for One-Pot Evaluation of Sequence–function Relationships of Nanobodies." *Scientific Reports* 11 (1): 1–13.
McAlister, Graeme C., David P. Nusinow, Mark P. Jedrychowski, Martin Wühr, Edward L. Huttlin, Brian K. Erickson, Ramin Rad, Wilhelm Haas, and Steven P. Gygi. 2014. "MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes." *Analytical Chemistry*. https://doi.org/10.1021/ac502040v.
McCallister, Chelsea, Matthew C. Siracusa, Farzaneh Shirazi, Dimitra Chalkia, and Nikolas Nikolaidis. 2015. "Functional Diversification and Specialization of Cytosolic 70-kDa Heat Shock Proteins." *Scientific Reports* 5 (March): 9363.
McHarg, J., S. M. Kelly, N. C. Price, A. Cooper, and J. A. Littlechild. 1999. "Site-Directed Mutagenesis of Proline 204 in the 'Hinge' Region of Yeast Phosphoglycerate Kinase." *European Journal of Biochemistry / FEBS* 259 (3): 939–45.
Meier, Florian, Philipp E. Geyer, Sebastian Virreira Winter, Juergen Cox, and Matthias Mann. 2018. "BoxCar Acquisition Method Enables Single-Shot Proteomics at a Depth of 10,000 Proteins in 100 Minutes." *Nature Methods* 15 (6): 440–48.
Melamed, Daniel, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. 2013. "Deep Mutational Scanning of an RRM Domain of the Saccharomyces Cerevisiae

poly(A)-Binding Protein." *RNA* 19 (11): 1537–51.
Meyer, Bjoern, Jeanne Chiaravalli, Stacy Gellenoncourt, Philip Brownridge, Dominic P. Bryne, Leonard A. Daly, Arturas Grauslys, et al. 2021. "Characterising Proteolysis during SARS-CoV-2 Infection Identifies Viral Cleavage Sites and Cellular Targets with Therapeutic Potential." *Nature Communications* 12 (1): 1–16.
Meyer, Michael J., Juan Felipe Beltrán, Siqi Liang, Robert Fragoza, Aaron Rumack, Jin Liang, Xiaomu Wei, and Haiyuan Yu. 2018. "Interactome INSIDER: A Structural Interactome Browser for Genomic Studies." *Nature Methods* 15 (2): 107–14.
Miettinen, Teemu P., Julien Peltier, Anetta Härtlova, Marek Gierliński, Valerie M. Jansen, Matthias Trost, and Mikael Björklund. 2018. "Thermal Proteome Profiling of Breast Cancer Cells Reveals Proteasomal Activation by CDK4/6 Inhibitor Palbociclib." *The EMBO Journal* 37 (10). https://doi.org/10.15252/embj.201798359.
Moustaqil, Mehdi, Emma Ollivier, Hsin-Ping Chiu, Sarah Van Tol, Paulina Rudolffi-Soto, Christian Stevens, Akshay Bhumkar, et al. 2021. "SARS-CoV-2 Proteases PLpro and 3CLpro Cleave IRF3 and Critical Modulators of Inflammatory Pathways (NLRP12 and TAB1): Implications for Disease Presentation across Species." *Emerging Microbes & Infections* 10 (1): 178–95.
Needham, Elise J., Janne R. Hingst, Benjamin L. Parker, Kaitlin R. Morrison, Guang Yang, Johan Onslev, Jonas M. Kristensen, et al. 2021. "Personalized Phosphoproteomics Identifies Functional Signaling." *Nature Biotechnology*, December. https://doi.org/10.1038/s41587-021-01099-9.
Nesvizhskii, Alexey I., Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. "A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry." *Analytical Chemistry* 75 (17): 4646–58.
Noren, C. J., S. J. Anthony-Cahill, M. C. Griffith, and P. G. Schultz. 1989. "A General Method for Site-Specific Incorporation of Unnatural Amino Acids into Proteins." *Science* 244 (4901): 182–88.
Ochoa, David, Andrew F. Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A. Kleefeldt, et al. 2020. "The Functional Landscape of the Human Phosphoproteome." *Nature Biotechnology* 38 (3): 365–73.
Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics." *Molecular & Cellular Proteomics: MCP* 1 (5): 376–86.
Pablos, Isabel, Yoan Machado, Hugo C. Ramos de Jesus, Yasir Mohamud, Reinhild Kappelhoff, Cecilia Lindskog, Marli Vlok, et al. 2021. "Mechanistic Insights into COVID-19 by Global Analysis of the SARS-CoV-2 3CLpro Substrate Degradome." *Cell Reports*, October, 109892.
Perona, John J., and Andrew Hadd. 2012. "Structural Diversity and Protein Engineering of the Aminoacyl-tRNA Synthetases." *Biochemistry* 51 (44): 8705–29.
Perrin, Jessica, Thilo Werner, Nils Kurzawa, Anna Rutkowska, Dorothee D. Childs, Mathias Kalxdorf, Daniel Poeckel, et al. 2020. "Identifying Drug Targets in Tissues and Whole Blood with Thermal-Shift Profiling." *Nature Biotechnology* 38 (3): 303–8.
Piatigorsky, J., and G. Wistow. 1991. "The Recruitment of Crystallins: New Functions Precede Gene Duplication." *Science* 252 (5009): 1078–79.
Piazza, Ilaria, Karl Kochanowski, Valentina Cappelletti, Tobias Fuhrer, Elad Noor, Uwe Sauer, and Paola Picotti. 2018. "A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication." *Cell* 172 (1-2): 358–72.e23.
Plesa, Calin, Angus M. Sidore, Nathan B. Lubock, Di Zhang, and Sriram Kosuri. 2018. "Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes." *Science* 359 (6373): 343–47.

Plubell, Deanna L., Lukas Käll, Bobbie-Jo Webb-Robertson, Lisa Bramer, Ashley Ives, Neil L. Kelleher, Lloyd M. Smith, Thomas J. Montine, Christine C. Wu, and Michael J. MacCoss. 2021. “Can We Put Humpty Dumpty Back Together Again? What Does Protein Quantification Mean in Bottom-up Proteomics?” *bioRxiv*. https://doi.org/10.1101/2021.01.25.428175.

Potel, Clément M., Nils Kurzawa, Isabelle Becher, Athanasios Typas, André Mateus, and Mikhail M. Savitski. 2020. “Impact of Phosphorylation on Thermal Stability of Proteins.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.01.14.903849.

———. 2021. “Impact of Phosphorylation on Thermal Stability of Proteins.” *Nature Methods* 18 (7): 757–59.

Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. “A Large-Scale Evaluation of Computational Protein Function Prediction.” *Nature Methods* 10 (3): 221–27.

Rappsilber, Juri, Yasushi Ishihama, and Matthias Mann. 2003. “Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics.” *Analytical Chemistry* 75 (3): 663–70.

Rappsilber, Juri, Matthias Mann, and Yasushi Ishihama. 2007. “Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips.” *Nature Protocols* 2 (8): 1896–1906.

Ravikumar, Arjun, Garri A. Arzumanyan, Muaeen K. A. Obadi, Alex A. Javanpour, and Chang C. Liu. 2018. “Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds.” *Cell* 175 (7): 1946–57.e13.

Rawlings, N. D., and A. J. Barrett. 1993. “Evolutionary Families of Peptidases.” *Biochemical Journal* 290 (Pt 1) (February): 205–18.

Reinhard, Friedrich B. M., Dirk Eberhard, Thilo Werner, Holger Franken, Dorothee Childs, Carola Doce, Maria Fälth Savitski, et al. 2015. “Thermal Proteome Profiling Monitors Ligand Interactions with Cellular Membrane Proteins.” *Nature Methods* 12 (12): 1129–31.

Richmond, M. H. 1962. “The Effect of Amino Acid Analogues on Growth and Protein Synthesis in Microorganisms.” *Bacteriological Reviews* 26 (December): 398–420.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47–e47.

Rix, Gordon, Ella J. Watkins-Dulaney, Patrick J. Almhjell, Christina E. Boville, Frances H. Arnold, and Chang C. Liu. 2020. “Scalable Continuous Evolution for the Generation of Diverse Enzyme Variants Encompassing Promiscuous Activities.” *Nature Communications* 11 (1): 5644.

Rodriguez-Mias, Ricard A., Kyle N. Hess, Bianca Y. Ruiz, Ian R. Smith, Anthony S. Barente, Stephanie M. Zimmerman, Yang Y. Lu, William S. Noble, Stanley Fields, and Judit Villén. 2022. “Proteome-Wide Identification of Amino Acid Substitutions Deleterious for Protein Function.” *bioRxiv*. https://doi.org/10.1101/2022.04.06.487405.

Roe, Molly K., Nathan A. Junod, Audrey R. Young, Dia C. Beachboard, and Christopher C. Stobart. 2021. “Targeting Novel Structural and Functional Features of Coronavirus Protease nsp5 (3CLpro, Mpro) in the Age of COVID-19.” *The Journal of General Virology* 102 (3). https://doi.org/10.1099/jgv.0.001558.

Savitski, Mikhail M., Friedrich B. M. Reinhard, Holger Franken, Thilo Werner, Maria Fälth Savitski, Dirk Eberhard, Daniel Martinez Molina, et al. 2014. “Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome.” *Science* 346 (6205): 1255784.

Savitski, Mikhail M., Nico Zinn, Maria Faelth-Savitski, Daniel Poeckel, Stephan Gade, Isabelle Becher, Marcel Muelbaier, et al. 2018. “Multiplexed Proteome Dynamics Profiling Reveals Mechanisms Controlling Protein Homeostasis.” *Cell* 173 (1): 260–74.e25.

Sayed, Y., L. A. Wallace, and H. W. Dirr. 2000. “The Hydrophobic Lock-and-Key Intersubunit

Motif of Glutathione Transferase A1-1: Implications for Catalysis, Ligandin Function and Stability." *FEBS Letters* 465 (2-3): 169–72.
Schaffert, Nina, Markus Hossbach, Rainer Heintzmann, Tilmann Achsel, and Reinhard Lührmann. 2004. "RNAi Knockdown of hPrp31 Leads to an Accumulation of U4/U6 Di-snRNPs in Cajal Bodies." *The EMBO Journal* 23 (15): 3000–3009.
Schmidt, Nora, Caleb A. Lareau, Hasmik Keshishian, Sabina Ganskih, Cornelius Schneider, Thomas Hennig, Randy Melanson, et al. 2020. "The SARS-CoV-2 RNA–protein Interactome in Infected Human Cells." *Nature Microbiology* 6 (3): 339–53.
Schwartz, Michael H., and Tao Pan. 2017. "Function and Origin of Mistranslation in Distinct Cellular Contexts." *Critical Reviews in Biochemistry and Molecular Biology* 52 (2): 205–19.
Schweppe, Devin K., Jimmy K. Eng, Qing Yu, Derek Bailey, Ramin Rad, Jose Navarrete-Perea, Edward L. Huttlin, Brian K. Erickson, Joao A. Paulo, and Steven P. Gygi. 2020. "Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics." *Journal of Proteome Research* 19 (5): 2026–34.
Selkrig, Joel, Megan Stanifer, André Mateus, Karin Mitosch, Inigo Barrio-Hernandez, Mandy Rettel, Heeyoung Kim, et al. 2021. "SARS-CoV-2 Infection Remodels the Host Protein Thermal Stability Landscape." *Molecular Systems Biology* 17 (2): e10188.
Sellick, Christopher A., and Richard J. Reece. 2003. "Modulation of Transcription Factor Function by an Amino Acid: Activation of Put3p by Proline." *The EMBO Journal* 22 (19): 5147–53.
Sharon, Eilon, Shi-An A. Chen, Neil M. Khosla, Justin D. Smith, Jonathan K. Pritchard, and Hunter B. Fraser. 2018. "Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing." *Cell* 175 (2): 544–57.e16.
Shendure, Jay, and Joshua M. Akey. 2015. "The Origins, Determinants, and Consequences of Human Mutations." *Science* 349 (6255): 1478–83.
Shendure, Jay, and Hanlee Ji. 2008. "Next-Generation DNA Sequencing." *Nature Biotechnology* 26 (10): 1135–45.
Shuker, S. B., P. J. Hajduk, R. P. Meadows, and S. W. Fesik. 1996. "Discovering High-Affinity Ligands for Proteins: SAR by NMR." *Science* 274 (5292): 1531–34.
Smith, Ian R., Kyle N. Hess, Anna A. Bakhtina, Anthony S. Valente, Ricard A. Rodríguez-Mias, and Judit Villén. 2021. "Identification of Phosphosites That Alter Protein Thermal Stability." *Nature Methods*.
Socolich, Michael, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. 2005. "Evolutionary Information for Specifying a Protein Fold." *Nature* 437 (7058): 512–18.
Sondermann, H., C. Scheufler, C. Schneider, J. Hohfeld, F. U. Hartl, and I. Moarefi. 2001. "Structure of a Bag/Hsc70 Complex: Convergent Functional Evolution of Hsp70 Nucleotide Exchange Factors." *Science* 291 (5508): 1553–57.
Song, Hyebin, Bennett J. Bremer, Emily C. Hinds, Garvesh Raskutti, and Philip A. Romero. 2020. "Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning." *Cell Systems* 0 (0). https://doi.org/10.1016/j.cels.2020.10.007.
Sridharan, Sindhuja, Nils Kurzawa, Thilo Werner, Ina Günthner, Dominic Helm, Wolfgang Huber, Marcus Bantscheff, and Mikhail M. Savitski. 2019. "Proteome-Wide Solubility and Thermal Stability Profiling Reveals Distinct Regulatory Roles for ATP." *Nature Communications* 10 (1): 1155.
Starita, Lea M., Nadav Ahituv, Maitreya J. Dunham, Jacob O. Kitzman, Frederick P. Roth, Georg Seelig, Jay Shendure, and Douglas M. Fowler. 2017. "Variant Interpretation: Functional Assays to the Rescue." *American Journal of Human Genetics* 101 (3): 315–25.
Starita, Lea M., David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. 2015.

"Massively Parallel Functional Analysis of BRCA1 RING Domain Variants." *Genetics* 200 (2): 413–22.
Stelter, Philipp, Ferdinand M. Huber, Ruth Kunze, Dirk Flemming, André Hoelz, and Ed Hurt. 2015. "Coordinated Ribosomal L4 Protein Assembly into the Pre-Ribosome Is Regulated by Its Eukaryote-Specific Extension." *Molecular Cell* 58 (5): 854–62.
Stukalov, Alexey, Virginie Girault, Vincent Grass, Ozge Karayel, Valter Bergant, Christian Urban, Darya A. Haas, et al. 2021. "Multilevel Proteomics Reveals Host Perturbations by SARS-CoV-2 and SARS-CoV." *Nature* 594 (7862): 246–52.
Süel, Gürol M., Steve W. Lockless, Mark A. Wall, and Rama Ranganathan. 2003. "Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins." *Nature Structural Biology* 10 (1): 59–69.
Swaney, Danielle L., Pedro Beltrao, Lea Starita, Ailan Guo, John Rush, Stanley Fields, Nevan J. Krogan, and Judit Villén. 2013. "Global Analysis of Phosphorylation and Ubiquitylation Cross-Talk in Protein Degradation." *Nature Methods* 10 (7): 676–82.
Tan, Chris Soon Heng, Ka Diam Go, Xavier Bisteau, Lingyun Dai, Chern Han Yong, Nayana Prabhu, Mert Burak Ozturk, et al. 2018. "Thermal Proximity Coaggregation for System-Wide Profiling of Protein Complex Dynamics in Cells." *Science* 359 (6380): 1170–77.
Tang, Qingling, and Aron W. Fenton. 2017. "Whole‐protein Alanine‐scanning Mutagenesis of Allostery: A Large Percentage of a Protein Can Contribute to Mechanism." *Human Mutation* 38 (9): 1132–43.
Tien, Matthew Z., Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. 2013. "Maximum Allowed Solvent Accessibilites of Residues in Proteins." *PloS One* 8 (11): e80635.
Tran, Duc T., Jagat Adhikari, and Michael C. Fitzgerald. 2014. "StableIsotope Labeling with Amino Acids in Cell Culture (SILAC)-Based Strategy for Proteome-Wide Thermodynamic Analysis of Protein-Ligand Binding Interactions." *Molecular & Cellular Proteomics: MCP* 13 (7): 1800–1813.
Tsai, F. H., C. G. Overberger, and R. Zand. 1990. "Synthesis and Peptide Bond Orientation in Tetrapeptides Containing L-Azetidine-2-Carboxylic Acid and L-Proline." *Biopolymers* 30 (11-12): 1039–49.
Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome." *Nature* 596 (7873): 590–96.
Tyanova, Stefka, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y. Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. 2016. "The Perseus Computational Platform for Comprehensive Analysis of (prote) Omics Data." *Nature Methods* 13 (9): 731–40.
Wagih, Omar, Marco Galardini, Bede P. Busby, Danish Memon, Athanasios Typas, and Pedro Beltrao. 2018. "A Resource of Variant Effect Predictions of Single Nucleotide Variants in Model Organisms." *Molecular Systems Biology* 14 (12): e8430.
Wang, Harris H., Farren J. Isaacs, Peter A. Carr, Zachary Z. Sun, George Xu, Craig R. Forest, and George M. Church. 2009. "Programming Cells by Multiplex Genome Engineering and Accelerated Evolution." *Nature* 460 (7257): 894–98.
Ward, William W., Hugh J. Prentice, Amy F. Roth, Chris W. Cody, and Sue C. Reeves. 1982. "SPECTRAL PERTURBATIONS OF THE AEQUOREA GREEN-FLUORESCENT PROTEIN." *Photochemistry and Photobiology*. https://doi.org/10.1111/j.1751-1097.1982.tb02651.x.
Wistow, G. J., J. W. Mulders, and W. W. de Jong. 1987. "The Enzyme Lactate Dehydrogenase as a Structural Protein in Avian and Crocodilian Lenses." *Nature* 326 (6113): 622–24.
Wistow, G., and J. Piatigorsky. 1987. "Recruitment of Enzymes as Lens Structural Proteins." *Science* 236 (4808): 1554–56.
Wittmann, Bruce J., Yisong Yue, and Frances H. Arnold. 2021. "Informed Training Set Design

Enables Efficient Machine Learning-Assisted Directed Protein Evolution." *Cell Systems*, August. https://doi.org/10.1016/j.cels.2021.07.008.
Wu, Chongde, Qian Ba, Dayun Lu, Wenxue Li, Barbora Salovska, Pingfu Hou, Torsten Mueller, et al. 2020. "Global and Site-Specific Effect of Phosphorylation on Protein Turnover." *Developmental Cell*, November. https://doi.org/10.1016/j.devcel.2020.10.025.
Zecha, Jana, Wassim Gabriel, Ria Spallek, Yun-Chien Chang, Julia Mergner, Mathias Wilhelm, Florian Bassermann, and Bernhard Kuster. 2022. "Linking Post-Translational Modifications and Protein Turnover by Site-Resolved Protein Turnover Profiling." *Nature Communications* 13 (1): 1–14.
Zecha, Jana, Chen Meng, Daniel Paul Zolg, Patroklos Samaras, Mathias Wilhelm, and Bernhard Kuster. 2018. "Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics." *Molecular & Cellular Proteomics: MCP* 17 (5): 974–92.
Zhang, Linlin, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. 2020. "Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors." *Science* 368 (6489): 409–12.
Zheng, Lei, Ulrich Baumann, and Jean-Louis Reymond. 2004. "An Efficient One-Step Site-Directed and Site-Saturation Mutagenesis Protocol." *Nucleic Acids Research* 32 (14): e115.
Zhou, X. Z., O. Kops, A. Werner, P. J. Lu, M. Shen, G. Stoller, G. Küllertz, M. Stark, G. Fischer, and K. P. Lu. 2000. "Pin1-Dependent Prolyl Isomerization Regulates Dephosphorylation of Cdc25C and Tau Proteins." *Molecular Cell* 6 (4): 873–83.
Zhou, Yuexin, Shiyou Zhu, Changzu Cai, Pengfei Yuan, Chunmei Li, Yanyi Huang, and Wensheng Wei. 2014. "High-Throughput Screening of a CRISPR/Cas9 Library for Functional Genomics in Human Cells." *Nature* 509 (7501): 487–91.
Zimmerman, Stephanie M., Yoshiko Kon, Alayna C. Hauke, Bianca Y. Ruiz, Stanley Fields, and Eric M. Phizicky. 2018. "Conditional Accumulation of Toxic tRNAs to Cause Amino Acid Misincorporation." *Nucleic Acids Research* 46 (15): 7831–43.