Dissecting Specificity of Short Linear Motifs in Protein Quality Control

Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

vorgelegt von

Susmitha Shankar

geboren am 16.02.1995

in Bhopal, India

Mainz, April 2024

Dekan: Prof. Dr. Eckhard Thines

1. Gutachter: [redacted]

2. Gutachter: [redacted]

Tag der mündlichen Prüfung: 10.04.2024

Preface

I hereby declare that I am the sole author and composer of this report and that no other source or learning aids other than those listed have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I hereby also declare that this report has not been prepared for another examination, either wholly or excerpts thereof. Throughout this study, I have received support for understanding experimental designs and building computational pipelines from [redacted]. [redacted], [redacted], [redacted] and [redacted] from IMB, Mainz, Germany, performed the experiments. The processing pipeline for deep mutational scanning was created by [redacted] from [redacted].

Place, Date Signature

Table of Contents

1 List of publications

Kong, K.-Y.E., **S. Shankar**, F. Rühle, and A. Khmelinskii. 2023. Orphan quality control by an SCF ubiquitin ligase directed to pervasive C-degrons. *Nat. Commun.* 14:8363. doi:10.1038/s41467-023-44096-z.

Shankar, S., I. Bhandari, D.T. Okou, G. Srinivasa, and P. Athri. 2021. Predicting adverse drug reactions of two-drug combinations using structural and transcriptomic drug representations to train an artificial neural network. *Chem. Biol. Drug Des.* 97:665–673. doi:10.1111/cbdd.13802.

Shankar, S., and S. Thangam. 2019. Integrated System for easier and effective Drug Information. *Biomed. Pharmacol. J.* 12:1069–1077. doi:10.13005/bpj/1736.

2 Summary

Cellular processes hinge on protein interactions with genetic materials, enzymes, and modifiers. Protein can interact with other proteins via structured domains or Short Linear Motifs (SLiM). SLiMs, typically 3 to 10 amino acids long, govern diverse functions, including protein quality control (PQC). By ensuring correct localization, degradation, and protein complex assembly, the PQC intricately preserves protein homeostasis, a critical determinant of cellular health. Despite their significance in maintaining proteostasis, SLiMs and their role in various protein quality control pathways are yet to be discovered entirely. In this work, I built tools to identify SLiMs in PQC, acting as localization signals and degrons in degradation pathways.

Degrons are found extensively at protein termini from bacteria to mammals. Though extensively studied, our understanding of the prevalence and specificity of degrons at termini is still incomplete. Here, I built a pipeline to analyze the Deep mutational scanning (DMS) experiments that help dissect the specificity of degrons. The pipeline first performs the quality assessment of DMS experiments. Then, it performs downstream analysis to infer degron motifs, using simple visualization for shorter peptides and interpretable machine learning for longer peptides.

To systematically study the N-terminal amino acid specificity, I employed simple visualization techniques after assessing the quality of the experiment. This is performed on the stability profile from DMS of N-terminal diresidue constructs in the *H. sapiens* cell line constructs. Furthermore, to study the eukaryotic C-degron pathway in an unbiased fashion, I applied Interpretable deep learning on the stability profile of over 40k random peptides in yeast C-terminomes to infer degron motifs. From the ~10% of putative C-degron peptides in this library, I found 21 potential C-degron motifs. Combining the results from machine learning, mutagenesis, and genetic screens reveals that the F-box substrate receptor of SCF ubiquitin ligase, Das1, targets ~40% of degrons and recognizes at least five distinct but overlapping motifs.

SLiMs also play a vital role as protein localization signals. Biophysical properties of various localization signals help in their correct localization. However, which features of the localization signals are crucial for targeting is poorly understood. In this work, I investigate how various biophysical properties of transmembrane domains(TMD) in Tail-anchored proteins help in localization. Analyzing the localization of TMD-variants reveals combinations of biophysical properties that help in compartment-specific localization. For example, short, low hydrophobic TMD-variants with positively charged C-terminals are more prone to mitochondrial localization. In contrast, hydrophobic TMD-variants with positively charged N-terminals tend to localize in PM. To extract the general but quantitative rules for localization, I created a random forest, which revealed the importance of the combined effect of hydrophobicity, flanking charge, length and cysteine content of TMD-variants in distinguishing between organelle-specific localization.

Overall, this work demonstrates the pipeline for dissecting specificity of SLiMs in PQC. The computational pipeline for dissecting degrons is scalable for different lengths of proteins and for any biophysical features at any protein termini. This pipeline could be used for analysis of future DMS experiments. Furthermore, the pipeline for dissecting the localization signals could be utilized to design new organelle-specific targeting signals.

3 Zusammenfassung

Zelluläre Prozesse hängen von Proteininteraktionen mit genetischem Material, Enzymen und Modifikatoren ab. Proteine können mit anderen Proteinen über strukturierte Domänen oder kurze lineare Motive (SLiM) interagieren. SLiMs, die typischerweise 3 bis 10 Aminosäuren lang sind, steuern vielfältige Funktionen, einschließlich der Proteinqualitätskontrolle (PQC). Durch die Sicherstellung korrekter Lokalisierung, Degradierung und Montage von Protein-Komplexen bewahrt die PQC-Maschinerie auf komplexe Weise die Proteostase, einen kritischen Determinanten für die Zellgesundheit. Trotz ihrer Bedeutung für die Aufrechterhaltung der Proteostase sind SLiMs und ihre Rollen in verschiedenen Proteinqualitätskontrollwege noch nicht vollständig verstanden. In dieser Arbeit entwickelte ich Werkzeuge zur Identifizierung von SLiMs in der PQC, wobei der Schwerpunkt auf der Analyse von Lokalisierungssignalen und Degradierungssignalen oder Degrons mit Deep Mutational Scanning (DMS)-Experimenten lag, bei denen Tausende von Sequenzvarianten parallel experimentell getestet werden können.

Degrons können an Proteinenden bei Bakterien bis hin zu Säugetieren gefunden werden. Trotz umfangreicher Untersuchungen ist unser Verständnis für die Häufigkeit und Spezifität von Degrons an den Enden noch unvollständig. Hier baute ich eine Pipeline zur Analyse von DMS-Experimenten auf, die dazu beiträgt, die Spezifität der Degrons zu analysieren. Die Pipeline bewertet zunächst die Qualität der DMS-Experimente. Anschließend führt sie eine nachgelagerte Analyse durch, um Degron-Motive zu erfassen, wobei einfache Visualisierungen für kürzere Peptide und interpretierbares maschinelles Lernen für längere Peptide verwendet werden.

Um kurze N-terminale Degrons systematisch zu untersuchen, verwendete ich einfache Visualisierungstechniken nach der Qualitätsbewertung der Experimente. Diese Analyse wurde an einem DMS-Experiment durchgeführt, das die Stabilität von N-terminalen Dipeptid-Konstrukten in einer menschlichen Zelllinie profilte. Darüber hinaus habe ich zur Untersuchung eukaryotischer C-Degron-Pfade auf unvoreingenommene Weise interpretierbares Deep Learning angewendet, um Degron-Motive aus DMS-Experimenten zu erfassen, die die Stabilität von über 50000 zufälligen Peptiden in der Bäckerhefe profilieren. Aus den ~10% der vermeintlichen C-Degron-Peptide in dieser Bibliothek fand ich 21 potenzielle C-Degron-Motive. Die Kombination der Ergebnisse aus maschinellem Lernen, Mutagenese und genetischen Screens zeigt, dass der F-Box-Substratrezeptor der SCF-Ubiquitin-Ligase, Das1, ~40% der Degrons abzielt und potenziell mindestens fünf weitere verschiedene, aber überlappende Motive erkennt.

SLiMs spielen auch eine wichtige Rolle als Protein-Lokalisierungssignale. Die physikalischen Eigenschaften verschiedener Lokalisierungssignale scheinen ihre Interaktionen zu bestimmen und letztendlich die Proteinlokalisierung zu definieren. Es ist jedoch unklar, welche Merkmale der Lokalisierungssignale für die Zielerfassung wichtig sind. In dieser Arbeit untersuchte ich, wie verschiedene physikalische Eigenschaften von Transmembrandomänen (TMDs) in

schwanzverankerten Proteinen bei der Lokalisierung helfen. Die Analyse der Lokalisierung von über 800 TMD-Varianten enthueallte Kombinationen von physikalischen Eigenschaften, die bei der kompartiment-spezifischen Lokalisierung helfen. Zum Beispiel neigen kurze, wenig hydrophobe TMD-Varianten mit positiv geladenen C-Termini eher zur mitochondrialen Lokalisierung. Im Gegensatz dazu neigen hydrophobe TMD-Varianten mit positiv geladenen N-Termini dazu, sich in der Plasmamembran zu lokalisieren. Um allgemeine, aber quantitative Regeln zur Bestimmung der Lokalisierung zu extrahieren, verwendete ich Random Forests, die die Bedeutung des kombinierten Effekts von Hydrophobizität, flankierender Ladung, Länge und Cystein-Gehalt von TMD-Varianten bei der Unterscheidung zwischen organell-spezifischen Lokalisierungen zeigten.

Insgesamt demonstriert diese Arbeit Pipelines und ihre Anwendungen zur Analysieren der Spezifität von SLiMs in der PQC. Die Rechenpipeline zur Analysieren von Degrons ist für verschiedene Längen von Proteinen und für beliebige physikalische Merkmale an beliebigen Proteinenden skalierbar. Diese Pipeline könnte für die Analyse zukünftiger DMS-Experimente verwendet werden. Darüber hinaus könnte die Pipeline zur Analysieren der Lokalisierungssignale genutzt werden, um neue organell-spezifische Zielsignale zu entwerfen.

4 Introduction

Most proteins perform cellular functions by interacting with other molecules in their surroundings. These interactions are facilitated by modules embedded in their sequence, which assist in binding with ligands, enzymes, other proteins, or genomic materials. Proteins can interact with their partners using either their globular domain or Short Linear Motifs (SLiMs). A SLiM is often referred to by several other names, such as Linear Motif (LM) or Molecular Recognition Features (MoRF) (Mohan et al., 2006; Davey et al., 2012). SLiMs differ from other parts of proteins based on their short length and comparatively low sequence complexity (Tompa et al., 2009). SLiMs are generally 3-10 amino acids (AA) long and are mostly found in the intrinsically disordered regions (IDRs) of the proteins. SLiMs are highly flexible and lack any tertiary conformations. The flexibility in its structure enables SLiMs to adapt and interact with many binding partners. Upon binding, most SLiMs resolve into alpha or beta structures (Davey et al., 2012). SLiMs help mediate many protein-protein interactions, one of which is called Domain-Motif Interaction (DMI). Domain-motif interaction is when the domain of a protein interacts with the SLiM of the interacting protein. An example of DMI is the interaction of SLiM from the Deficient in Utilization of Glutathione (Dug3) protein with the substrate recognition component of Skp, Cullin, F-Box containing complex (SCF) E3 ligase, Das1 protein in *S. cerevisiae* (Figure 1(a,b)). AlphaFold model of Das1 interacting with Dug3 shows that the SLiM at the C-terminus of Dug3 enters the binding pocket of Das1, creating an alpha helix at the interaction site. It is estimated that about 15-40% of the protein-protein interactions in cells are DMI and are enriched in signaling networks (Edwards and Palopoli, 2015). Their short length and flexible nature allow SLiMs to function as dynamic regulatory elements within large protein structures of the signaling network. SLiMs also modulate other diverse cellular processes, such as trafficking and post-translational modification (PTMs) (Van Roey et al., 2014).

SLiMs can be found in rapidly evolving regions of proteins, which can help them become part of multiple cellular processes. Unlike other regions of proteins, SLiMs can evolve from random peptides after multiple point mutations. Most SLiM interactions with the domain of the interacting proteins are weak and transient. Some point mutations in the SLiM can increase the strength of the bond it creates with the domain of the binding proteins. Conversely, mutation at the key residue can also have a deleterious effect on the SLiMs, disrupting the entire molecular pathway (Dinkel et al., 2014). As a result of constant gain and loss of mutations on SLiMs over evolutions, most of them are not conserved. For instance, less than 5% of motifs in the Eukaryotic Linear Motif (ELM) database, which is the most comprehensive resource for experimentally validated motifs (Dinkel et al., 2014), are conserved between *S. cerevisiae* and *H. sapiens*. However, functionally important

motifs remain conserved across many species. For example, the PIP-box motif of Flap-endonuclease 1 (Fen1), which helps in its binding with Proliferating Cell Nuclear Antigen (Pcna), is conserved across all species (Bruning and Shamoo, 2004).

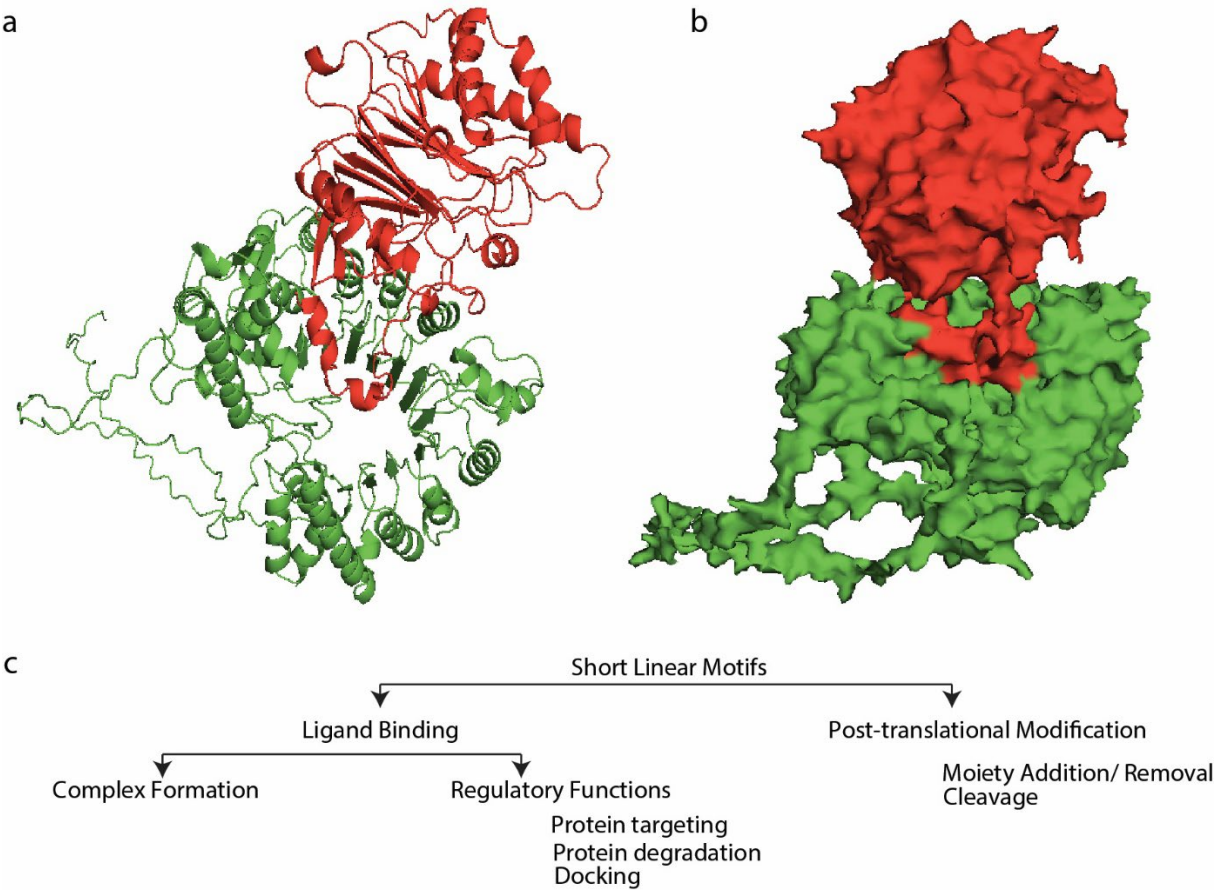

Figure 1: Short Linear Motif in protein-protein interaction. *Protein-protein interaction of Das1-Dug3. Dug3 protein interacts with Das1 using the SLiMs. The C-terminally located SLiM from the Dug3 enters into the pocket of the Das1 protein. The structure is shown in (a) ribbon structure and (b) surface structure. (c) SLiMs can be classified based on functions, such as ligand binding or regulating post-translational modification.*

Overall, SLiMs represent essential elements of protein structure that mediate the protein-protein interactions and regulate various cellular functions, including Protein Quality Control. The study of SLiMs can yield valuable insights into understanding the specificity of these complex processes and holds promise to uncover novel therapeutics.

4.1 Motifs in Protein Quality Control

Due to their physiological flexibility, SLiMs facilitate many processes, including targeting proteins to correct sub-cellular locations, regulatory activities of the proteins, regulating post-translational modifications (PTMs), protein assembly and degradation (Dinkel et al., 2014). (Van Roey et al., 2014) reviews various functions and mechanisms through which

SLiMs regulate cellular processes. Based on their functions, SLiMs can either be ligand motifs mediating protein binding or post-translational modification motifs mediating PTMs.

Another critical factor in maintaining cellular health is the correct localization of proteins to their specific subcellular compartments. Correct localization of proteins to their compartments helps them perform their functions. When mislocalized, these proteins cannot perform their destined functions and, at times, become toxic to the cell. Most proteins are produced in the cytosol and then targeted to their destined compartment. SLiMs in the proteins, called targeting signals, predominantly encompassing certain biophysical properties help the nascent proteins to be targeted to their destined compartment either co-translationally or post-translationally (Hegde and Zavodszky, 2019). Mutation to the targeting signals mislocalizes the proteins to the wrong compartments and plays a role in many cancers and neurodegenerative diseases (Pidasheva et al., 2005; Guo et al., 2014). Though essential to cellular health, organelle-specific targeting signals are yet to be fully understood.

In this work, I explored the role of SLiMs in protein localization and degradation. Chapter 2 entails the dissection of specific SLiMs at protein N and C-termini involved in protein degradation. I also discuss the factors responsible for the degradation of discovered motifs, attempting to picture a mechanism of such degradation. Chapter 3 deals with understanding the factors responsible for the organelle-specific localization of so-called tail-anchored (TA) proteins. Various biophysical properties and amino acid sequences are studied in detail to find patterns for all major compartments, with a focus on the plasma membrane (PM), mitochondria, and endoplasmic reticulum (ER).

5 Dissecting Specificity of SLiMs in selective protein degradation

Short Linear Motifs (SLiMs) are crucial in maintaining protein homeostasis by participating in protein quality control (PQC). One of the cellular processes that SLiMs take part in is selective protein degradation by the ubiquitin-proteasome system (UPS). The SLiMs discussed in this study that help recognize the proteins to be targeted for degradation via the UPS are referred to as degron. In this chapter, I systematically explore specific SLiMs at protein N- and C-termini involved in protein degradation. I also discuss the factors responsible for the degradation of discovered motifs, attempting to picture a mechanism of such degradation.

5.1 Introduction

5.1.1 Ubiquitin-proteasome system

Selective protein degradation by the ubiquitin-proteasome system (UPS) maintains protein homeostasis by removing abnormal and excess proteins from the cell. UPS has many substrates and is involved in many regulatory processes. The substrates include tumor suppressors such as p53 (do Patrocinio et al., 2022) and growth regulators like c-Fos and c-Jun, to name a few (Tsurumi et al., 1995). Via the degradation of short-lived proteins in regulatory pathways, UPS appears to be part of many cellular processes, such as signal transduction and cell cycle control (Voutsadakis, 2012; Reed, 2006). Due to their relevance in maintaining cellular health, aberrations in the components of this system have implications in several diseases. For this reason, several therapeutic advances are focused on the UPS as the center. (Schwartz and Ciechanover, 1999; Zhang et al., 2020). For instance, ubiquitin-mediated proteasome degradation of cyclin partners and kinase inhibitors (CKI) with the help of two E3 ligases - APC or Skp1-Cul2, helps in the activation of cyclin-dependent kinases (CDK), which thereby helps in regulation of cell cycle progression (Zou and Lin, 2021). Inefficient proteolytic control of CKI leads to uncontrolled cellular proliferation and tumorigenesis. Thus, targeting E3 ligases for cell proliferation is considered a novel therapeutic target for cancer treatment (Bielskienė et al., 2015; Wertz and Wang, 2019; Bulatov and Ciulli, 2015; Rodriguez et al., 2020).

Abnormal or excess proteins are marked for degradation by tagging them with ubiquitin. Ubiquitin is a highly conserved 76 amino acid long protein that is covalently attached to a target protein in the process of ubiquitination. Ubiquitination, a post-translational modification, starts when a single ubiquitin is attached to the lysine residue of the substrate by a cascade of the following steps (Figure 2(a)):

- The Ubiquitin-activating enzymes (E1 enzymes) activate the ubiquitin by linking itself to the carboxyl end of the C-terminal glycine of ubiquitin, forming a thioester bond in an energy-dependent manner.

- The activated ubiquitin moiety is transferred from the E1 enzyme to the cysteine of the Ubiquitin-conjugating enzymes (E2 enzyme).
- The E3 ligases transfer the activated ubiquitin moiety from the E2 enzyme to the lysine of the substrate protein. In certain cases, ubiquitin is transferred to amino acids C, S, or T in the substrates.
- Additionally, ubiquitination is a reversible process in which the attached ubiquitin can be released from the target protein with the help of deubiquitinating enzymes (DUB).

The human genome encodes around 2 E1 enzymes, 50 E2 enzymes, and more than 700 E3 ligases (Park et al., 2020). There are around 100 encoded DUBs in humans, 11 of which are unlikely to display ubiquitin protease activity (Nijman et al., 2005).

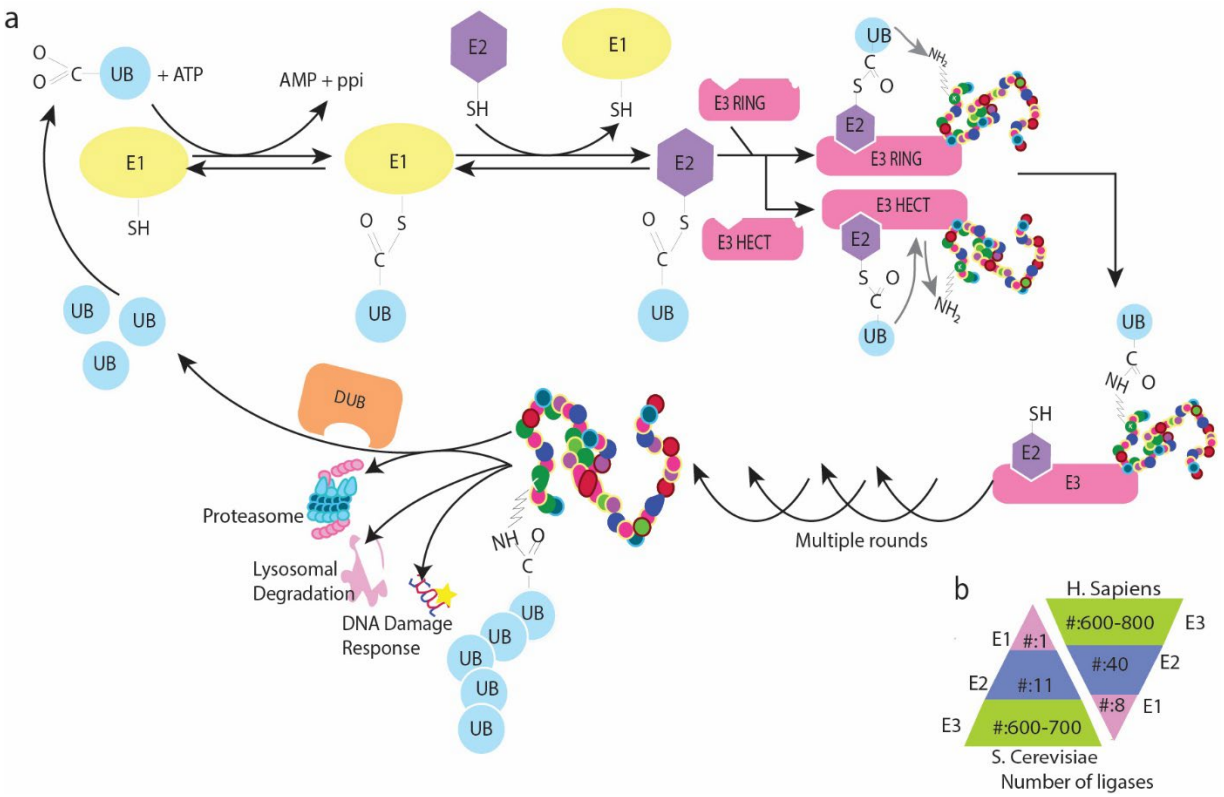

Figure 2: Ubiquitin Proteasome System *(a) Mechanism of ubiquitin-proteasome system. Ubiquitin is transferred to the substrate to be degraded by the cascading effect of E1, E2, and E3 enzymes. Ubiquitination is a reversible process in which the DUBs can deubiquitinate the substrate, UB: Ubiquitin, DUB: deubiquitinating enzymes (b) Number of components from ubiquitination system in* S. cerevisiae *and H. Sapiens*

In eukaryotic cells, monoubiquitination is widely regulated and serves various cellular functions, such as protein stability, localization, and signaling (Nakagawa and Nakayama 2015). Ubiquitin can also create multiple types of polymeric chains by ubiquitinating any of the seven internal lysine residues (K6, K11, K27, K29, K33, K48, and K63) or α-amino group of N-terminal methionine of ubiquitin. The number of ubiquitins and types of linkage of polyubiquitin chains form the

"ubiquitin code" and direct various functions in the cellular environment after being deciphered by the ubiquitin-binding proteins (Kwon and Ciechanover 2017; Komander and Rape 2012; Yau and Rape 2016). Additionally, ubiquitin can also be conjugated with other ubiquitin-like modifiers, such as SUMO (Hendriks et al. 2014), (Sriramachandran and Dohmen 2014), NEDD and phosphate molecules (Wang, Zhu, and Xu 2017).

Among the various roles of ubiquitination in cells, one of the predominant functions is selective degradation by UPS, targeting approximately 80-90% of proteins depending on physiological state and cell type (Hershko and Ciechanover, 1998). Among these polyubiquitin chains, K48 and K11-linked polyubiquitin chains serve as the most potent signals for degradation by the proteasome (Kwon and Ciechanover, 2017). Once taken to the proteasome, the polyubiquitinated substrates are unfolded and cleaved into smaller peptides by the ATPases, with DUBs releasing the ubiquitin into the free cytosolic pool.

5.1.2 Short Linear Motifs in ubiquitin-proteasome system

The human genome encodes over 700 E3 ligases (J. Park, Cho, and Song 2020). Most of these are classified into three categories – 1. Really interesting new finger protein (RING), 2. Homologous to E6-AP Carboxyl Terminus (HECT) and 3. U-Box proteins. E3s recognize the substrates and mark them with ubiquitin. The ubiquitinated substrates have many fates, one of which is degradation by 26S proteasome.

What are the molecular features that E3 ligases recognize in the ubiquitinated substrates that indicate degradation? These features, called degrons (A. Varshavsky 1991), are short linear motifs (SLiMs), about 3 to 10 AA long, preferentially in disordered regions (Timms and Koren, 2020). Degrons are typically transferable; that is, degrons from an unstable protein when substituted in the otherwise stable protein, can destabilize the stable protein. This property of transferability has further facilitated the discovery of degrons by low throughput and high throughput technologies (Ella, Reiss, and Ravid 2019).

Degrons can reside anywhere in the protein. There are multiple reasons why the termini of proteins might be particularly fertile ground for degron motifs:

- Sequence at the termini can acquire more post-translational processing events and regulatory modifications, which are found to be predominantly important in degradation pathways (Timms and Koren, 2020).
- Protein termini need not be evolutionarily constrained to confer a three-dimensional state (Timms and Koren, 2020).
- Protein termini are more disordered than internal regions in humans and *S. cerevisiae* (Figure 3). The region of the protein is defined as disordered if the amino acid has an IUPred score of more than 0.5. IUPred is a computational tool for predicting intrinsically

disordered regions (IDRs) within protein sequences (Mészáros et al., 2018). It employs an algorithm based on the physicochemical properties of amino acids to assess the likelihood of disorderliness in a given protein region. IUPred analysis for disordered regions in protein termini versus internal region reveals that approximately 15% of amino acids are disordered in termini in *S. cerevisiae* and 10% in *H. sapiens*. About 18% of yeast proteome and ~22% of human proteome are disordered.

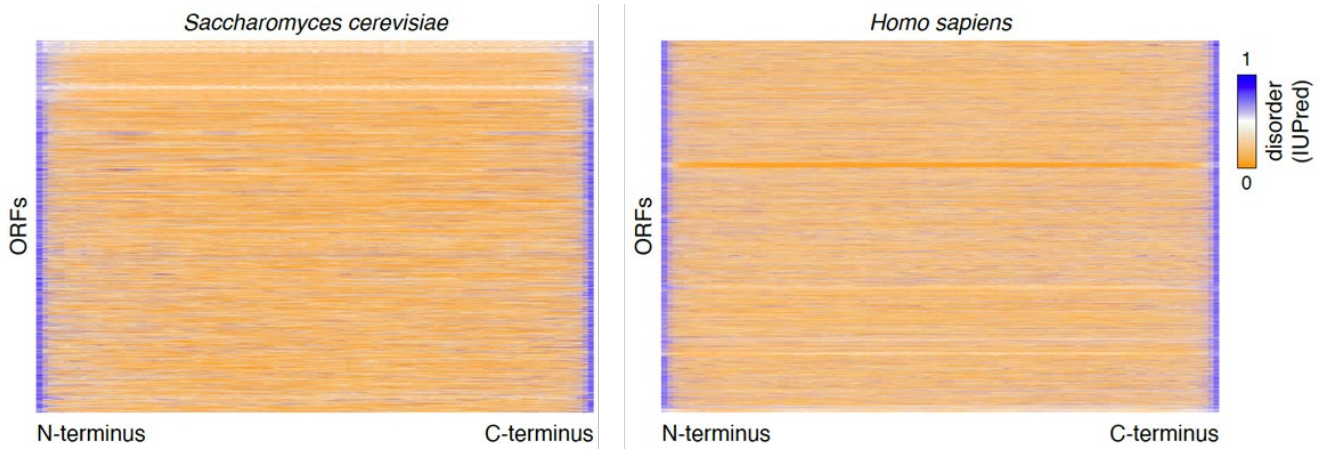

Figure 3: Distribution of disorder in S. Cerevisiae and H. Sapiens*. The disorder rate is calculated using IUPRED using the "short" type prediction for predicting terminal disorder. Positions of amino acids are normalized. Blue color indicates higher disorder. ORFs are arranged alphabetically in their UniProtID, such that ORFs forming similar complexes are in sequence.*

5.1.3 Pathways in ubiquitin-proteasome degradation

Since the discovery of the Arg/N-degron pathways in 1986 (Bachmair et al., 1986; Bachmair and Varshavsky, 1989), multiple other N-degron pathways have been identified. Recent developments in oligonucleotide synthesis and sequencing made it possible to create and sequence millions of DNA simultaneously. Advanced bioinformatics pipelines have accelerated the discovery of degrons and their association with their respective E3 ligases, as discussed in subsequent sections. This has helped to unveil the intricacies of their mechanism. This section lists a few notable N- and C-degron pathways.

5.1.3.1 N-degron pathway

The first to be discovered, the N-degron pathways have been deeply studied over the years. Proteins with N-terminal degradation signals are recognized by N-recognins or chaperons before being sent to the 26S proteasome. The importance of N-degron pathways is extensively discussed in many reviews (Leboeuf et al., 2020; Varshavsky, 2019). Some prominent N-degron pathways are shortly pointed out in this section.

Arg/N-degron pathway

The Arg/N-degron pathway in yeast and humans recoganized the positively charged (R, K, H) or large hydrophobic (W, L, F, Y, I) N-terminal residues for degradation by the proteasome. Additionally, proteins with N-terminal N or Q residues may undergo processing by the N-terminal amidase Nta1 in yeast and NTAN1 and NTAQ1 in humans, resulting in N-termini containing D or E residues, which are then arginylated by the N-terminal arginyl-transferase ATE1, leading to the generation of N-termini starting with an arginine and thereby becoming targets for UBR E3 ligases. Additionally, in humans, cysteine at the N-terminal can also be targeted for degradation by the Arg/N-degron pathway. UBR1 in yeast also can recognize N-termini, where the initiator methionine remains unacetylated and is followed by a large hydrophobic residue (Figure 4(a)).

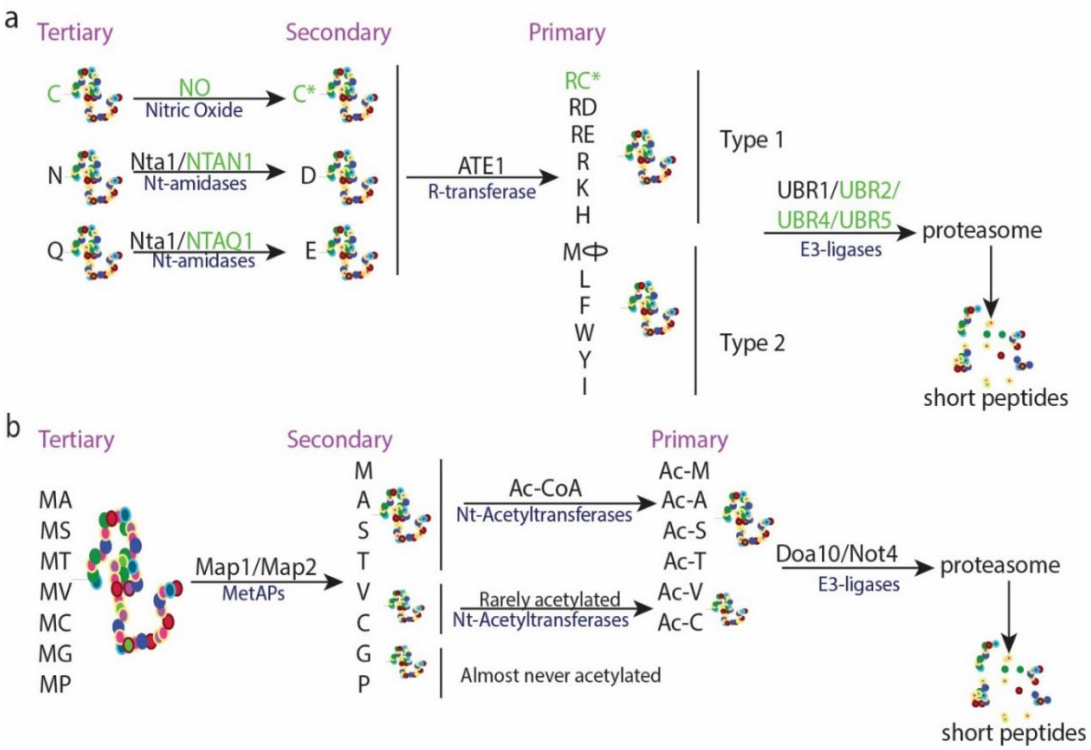

Figure 4: N-degron pathway *(a) Arg/N-degron pathway in humans and yeast. Labels in green are part of the Arg/N-degron pathway in humans, while in black, are part of the Arg/N-degron pathway in yeast and humans. (b) Ac/N-degron pathway*

Ac/N-degron pathway

Acetylation of proteins is one of the most prominent PTM in eukaryotic proteins. About 60% of *S. Cerevisiae* proteins and 80% of human proteins are acetylated (Aksnes et al., 2016) and take an active role in many processes, including degradation by UPS (Hwang et al., 2010). Acetylation of the N-terminus of some proteins in yeast plays a key role in their degradation through the Ac/N-degron pathway. The initiator methionine (iMet) is cleaved off protein N-termini by methionine

aminopeptidases (MetAPs) if the second residue is one of the six amino acids (A, S, T, C, V, G, and P). Acetyltransferases then acetylate the exposed N-terminus of the protein. The acetylated protein can be recognized and ubiquitinated by two E3 ligases – MARCH6 or Doa10, a ring E3 ligase in the ER membrane, and Not4, a component of the CCr4-Not complex (Figure 4(b)). In some instances, large hydrophobic residues (F, W, Y, L, and I) and acidic residues (N, Q, D, and E) are also acetylated by NatC and NatB, respectively, before being targeted to degradation. The regulator of G protein Signaling 2 (Rgs2) in mammals, which helps in blood pressure regulation, is a well-characterized substrate of this pathway (Park et al., 2015).

However, since the deletion of Nt-acetyltransferases does not substantially affect the protein stability profile in *S. Cerevisiae*, the Ac/N-degron pathway is unlikely to have a significant role in protein stability regulation (Kats et al. 2018).

Gly/N-degron pathway

The N-terminal glycine of target substrates can be recognized by E3 ligases Cul2 in conjunction with two substrate receptors – ZYG11B and ZER1 (Figure 5(a)). Although both substrate receptors share 29% sequence similarity, ZYG11B degrons consist of two amino acids starting with G, whereas ZER1 degrons comprise longer sequences. The crystal structure of ZYG11B and ZER1 use their armadillo repeats (ARM) to create a cavity that accommodates the first four amino acids of the substrate (Yan et al., 2021). The Gly/N-degron pathway can degrade proteins that lack N-myristoylation, thus overseeing the quality control of protein N-myristoylation (Timms et al., 2019).

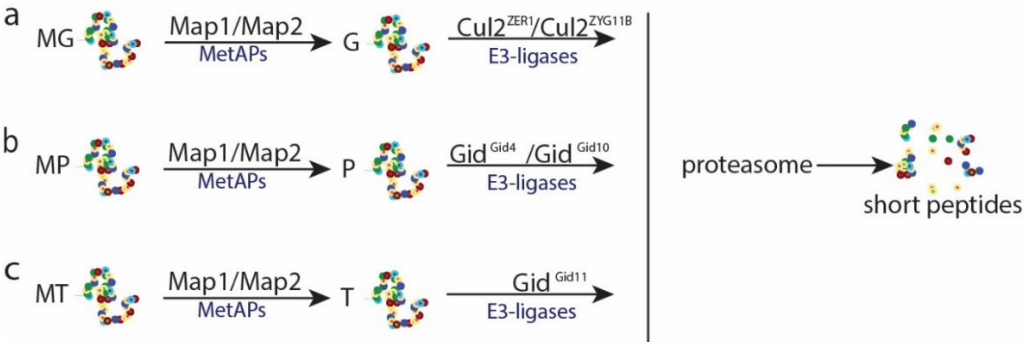

Figure 5: N-degron pathway *(a) Gly/N-degron pathway: N-terminal glycine is recognized by CRL E3 ligase with the help of two substrate receptors – ZYG11B or ZER1. (b) Pro/N-degron pathway: The N-terminal proline in gluconeogenesis enzymes, namely Mdh2, Fbp1, Icl1, and Pck1, are recognized by Gid through Gid4 or Gid11 substrate receptors. (c) Thr/N-degron pathway The N-terminal threonine after iMet cleavage is recognized by Gid E3 ligase using Gid11 as substrate receptor. The schematic presented here is in yeast.*

Pro/N-degron pathway

The GID (glucose-induced degradation-deficient) complex, a multi-subunit E3 ubiquitin ligase, is evolutionarily conserved in eukaryotes (Francis et al., 2013). In yeast, it acts as the N-recognin of

the Pro/N-degron pathway, targeting specific gluconeogenic enzymes such as fructose-1,6-bisphosphatase Fbp1, malate dehydrogenase Mdh2, phosphoenolpyruvate carboxykinase Pck1, and isocitrate lyase Icl1 for degradation by the proteasome upon transition from ethanol to glucose as the carbon source (Regelmann et al., 2003; Chen et al., 2017). These substrates harbor N-degrons characterized by a proline residue at the N-terminal or second position, which is recognized by the receptor subunit Gid4 (Chen et al., 2017) (Figure 5(b)). Gid4 expression is upregulated during the carbon source switch, ensuring timely suppression of gluconeogenesis (Santt et al., 2008; Menssen et al., 2018). Additionally, the GID E3 complex likely possesses other functions as its core components are present under diverse conditions, and another substrate receptor, Gid10, is induced by starvation or osmotic stress (Melnykov et al., 2019).

Thr/N-degron pathway

Another degradation pathway for gluconeogenesis enzymes involves the degradation of substrates with threonine at the N-terminus (Figure 5(c)). One of the most prominent substrates of this pathway is the nucleotidase PHosphate Metabolism (Phm8). After the initiator methionine is cleaved by MetAPs, the exposed T is recognized by Gid11 and targeted for proteasomal degradation (Kong et al., 2021).

5.1.3.2 C-degron pathways

The first C-degron was discovered in 1990 in bacteria when the protein was tagged with a C-terminal sequence – ANDENYALAA. The terminal sequence acts as a C-degron, targeting protein for degradation by the proteasome-like bacterial protease ClpXP (Keiler et al., 1996). Later in 2018, Elledge and Yen's laboratories discovered a large set of degron in the human C-terminome, calling them C-degron and the pathways helping in degradation as “DesCEND”(destruction via C-end degron) (Lin et al., 2018; Koren et al., 2018). C-terminal degrons can arise through various mechanisms, including premature terminal translation or cleavage by caspases. In either case, the protein has a chance to refold in a way that shields the C-degron. When exposed, the C-degrons are targeted by the E3 ligases, predominantly CRL E3 ligases, for degradation by the 26S proteasome. A few known C-degron pathways are listed below:

Gly/C-degron pathway

Various substrate receptors of CRL E3 ligases target the C-terminal glycine. The Cul2 complex uses three substrate adaptors of the Kelch family, each recognizing different motifs with amino acid G at the C-terminus. $Cul2^{KHLDC2}$ have a preference towards –GG motifs, $Cul2^{KHLDC3}$ recognize –RG and -KG, while $Cul2^{KHLDC10}$ prefer –AG, –WG, and –PG (Koren et al. 2018; Lin et al. 2018). Substrate adaptor KHLDC2 is a six-bladed propeller with a deep central pocket. A strong bond is formed between KHLDC2 and G at -2 of the substrate. KHLDC2 could tolerate only amino acids G and A at the C-terminus (Rusnac et al., 2018). Structurally, other members of the Kelch family (KLHDC1,

KLHDC3, and KLHDC10) are similar to KLHDC2 and might employ similar specificity for motifs. Apart from the Cul2 E3 ligase, the Cul5 E3 ligase also recognizes the –GG motif using the KHLDC1 substrate adaptor (Okumura et al., 2020)(Figure 6).

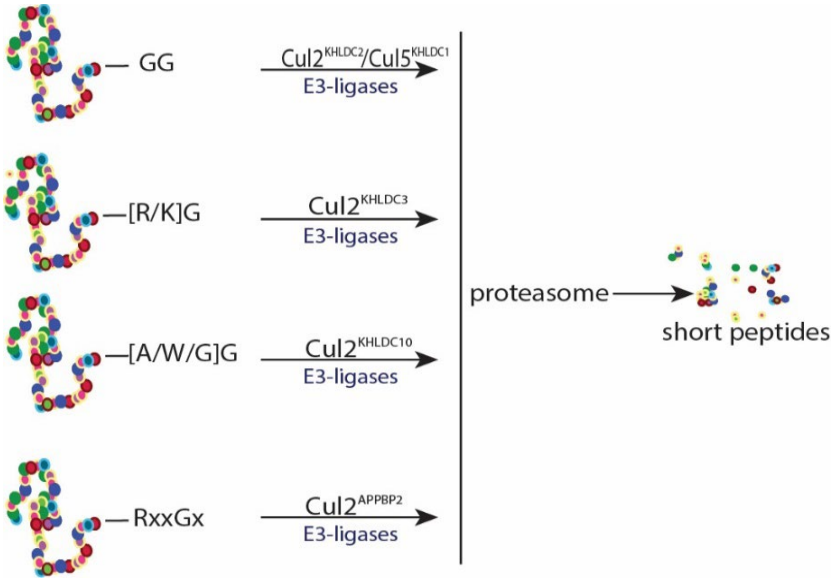

***Figure 6**: C-degron pathway – Gly/C-degron pathway in mammals. CRL E3 ligase recognizes the C-terminal glycine with the help of some members of Kelch domains or TPR repeats.*

Other C-terminal glycine motifs recognized by Cul2 with substrate adaptor APPBP2 are RxxG or RxxxG), where x can be any amino acid (Koren et al. 2018; Lin et al. 2018) (Figure 6). APPBP2 substrate adaptor is rather flexible due to its tetratricopeptide repeats (TPR) in terms that the motifs could be within the 10 positions of the C-terminus, with glycine at the second (-2) or third position (-3).

Arg/C-degron pathway

Other than amino acid G at the C-terminus, Cul2 can also recognize amino acid A with the help of the FEM1 family substrate adaptor. FEM1-family has Ankyrin repeats that create a degron binding pocket. Three of the substrate adaptors from the FEM1 family– FEM1A, FEM1B, and FEM1C, help recognize C-terminal arginine. These adapters show variability in degron specificity (Figure 7(a)). For example, while FEM1C could recognize R at -2 and -3. In some substrates, some FEM1 substrates hold internal degron (Lin et al. 2018).

Additional C-degron pathways

Apart from Cul2 and Cul5, Cul4 can also recognize C-degron motifs using two substrate adaptors

- TRPC4AP: This substrate receptor recognizes –Rxxx motif using its armadillo-like repeats (Figure 7(b)) (Koren et al. 2018)

- DCAF12: This substrate receptor recognizes –Ex motif using its WD40 repeats. The preference is much stronger for the –EE motif than just E at -2 (Figure 7(c)) (Koren et al. 2018).

Additionally, a member of the U-box family of atypical RING E3 ligases, CHIP, targets aspartate at -1 and a plethora of other residues that are the product of caspase cleavage (Figure 7(d)) (Ravalin et al., 2019).

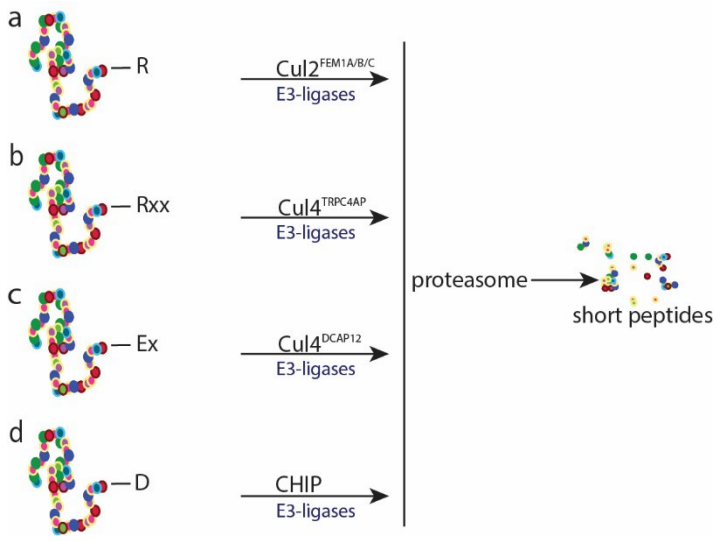

Figure 7: Additional C- degron pathway*. The schematic describes the pathways through which some C-terminal arginine are targeted by CRL E3 ligases, (a) Cul2 and (b) Cul4 in mammals. (c) Cul4 also recognizes C-terminal glutamic acid for degradation. (d) C-terminal aspartic acid is recognized by CHIP E3 ligases for degradation.*

5.1.4 Relevance of UPS in health and disease

Since the UPS is involved in numerous cellular processes, such as cell cycle division, DNA damage response, and stress response, to name a few, dysregulation of any component can be detrimental to health. Dysregulation of ubiquitination and UPS function has been strongly implicated in the pathogenesis of various neurodegenerative diseases (Zheng et al., 2016; Hegde and Upadhya, 2007), cancers (Aliabadi et al., 2021) and aging (Löw, 2011).

Numerous studies have investigated the effect of homeostasis of UPS components in aging. One example is the suppression of increased lifespan when the E3 ligase WW Domain-Containing Protein 1 (Wwp1) is overexpressed (Carrano et al., 2009). The experiments were performed under ad libitum feeding and dietary conditions. Thus, under specific nutritional conditions, it was established that E3 ligases are essential for age longevity. Moreover, (Hughes et al., 2022) review the significance of ubiquitin ligases in regulating muscle protein degradation and maintaining muscle mass during aging.

Dysregulation of UPS components also affects age-related diseases such as Parkinson's, Angelman's disease, and Alzheimer's. Angelman syndrome (AS) is a rare neurodevelopmental

disorder primarily caused by mutations or deletions in the maternal copy of the Ube3A gene located on chromosome 15q11-13 (Matsuura et al., 1997; Kishino et al., 1997). The Ube3A gene encodes the E3A E3 ubiquitin ligase. Another example of UPS affecting neurodegenerative disease is its effect on Parkinson's Disease (PD). Genetic mutations in the E3 ligase, parkin (Parkn2), compromise its activity and link parkin to dopaminergic neuronal survival. Additionally, a missense mutation (I93M) in a deubiquitinase, Uchl1, is also linked to reduced ubiquitin hydrolase activity in vitro, contributing to dopaminergic neuronal death (Ham et al., 2021).

Overall, UPS dysregulation affects various cellular processes, including aging, and has been linked to neurodegenerative diseases. Detrimental effects of mutations in UPS components highlight the critical role of UPS in maintaining cellular homeostasis and health.

5.1.5 Existing work on degron discovery

The first degron discovered in 1986 by the Varshavsky lab used the Colorometric method, which uses β-Galactosidase (β-Gal) to select the reporter (Gilon et al., 1998; Bachmair et al., 1986). In the colorometric method, candidate degron sequences are fused in-frame with the reporter protein coding sequence to create reporter fusion constructs. These constructs are then transfected or expressed in cells or organisms for analysis. Cells expressing the reporter fusion constructs are treated with compounds or subjected to conditions that modulate protein degradation pathways. Changes in the reporter signal, such as color change or luminescence, are monitored over time using colorimetric or luminescent assays. Candidate degron sequences that exhibit altered reporter signals upon treatment are further validated and characterized. Though versatile, colorimetric assays may lack the sensitivity or specificity required to detect subtle changes in protein degradation or discriminate between different degron sequences, making validation using complementary approaches essential (Krohn, 2011).

Later, degron studies were made using growth assays, which were more sensitive and quantitatively more accurate than the colorimetric assays (Gilon et al., 1998). Growth assays typically involve the monitoring of cell proliferation or viability over time in response to experimental manipulations that modulate protein degradation rates. In this, cells expressing degron fusion constructs are cultured under appropriate conditions and treated with compounds or subjected to conditions that modulate protein degradation pathways. Control cells expressing the wild-type protein or non-degradable mutants are included for comparison. Cell growth or viability is monitored over time using various assays, such as cell counting or live-cell imaging. Changes in growth kinetics or cell viability between experimental conditions are analyzed to assess the impact of degron-mediated protein degradation. Growth assays provide a quantitative measurement of cell growth, enabling precise characterization of degron-mediated effects. Moreover, growth assays can detect subtle changes in protein stability and degradation kinetics, allowing for the identification of degron sequences with varying efficacies. Due to their

advantages in quantitative studies, precision, and higher sensitivity to colorometric assay, growth assays have also been widely used in advanced deep mutational scanning. However, growth assays may be susceptible to off-target effects or indirect consequences of protein degradation.

With the advent of high throughput synthesis of oligonucleotide and next-generation sequencing, it became easier to test multiple peptides parallel for the presence of degrons in workflows such as deep mutational scanning (DMS). Global Protein Stability (Yen et al., 2008) is a deep mutational scanning approach in which peptides are fused to enhanced green fluorescence protein (EGFP), followed by an internal ribosome entry site (IRES) and a second fluorescent protein *Discosoma* sp. red fluorescent protein (DsRed). The mRNA (EGFP-IRES-DsRed) is transcribed from a single promotor. The ratio DsRed /EGFP denotes the protein stability and can be quantified using FACS. Further studies used variants of GPS to profile the protein stability due to the first 23 amino acids of primary isoforms of all human proteins, with and without iMet, using lentiviral expression (Timms et al. 2019). The study revealed that highly hydrophobic and positively charged N terminus leads to protein instability in proteins with cleaved iMet. Another DMS approach, called Multiplexed Protein Stability (MPS) (Kats et al., 2018), used a tandem-fluorescent protein timer (tFT), which is the fusion of slow maturing mCherry and fast maturing sfGFP to quantify the stability of thousands of proteins. This method was used to define the specificity of degrons in yeast N-termini. It revealed hydrophobicity and not Nt-acetylation as the key determinant of protein degradation via N-degron pathways. GPS profiling was also further used to identify the diverse set of C-degron in human proteome for the first time in 2018 (Koren et al. 2018).

Simultaneous to advanced in vivo degron studies, sophisticated computational pipelines are also developed to study degrons. Prior studies have used machine learning to understand degron motifs from GPS of mutated sequences. QCDPred (Johansson et al., 2023) uses the yeast-based GPS for 326 proteins, tiled into peptides of 17 amino acids, to predict internal degrons using logistic regression on the amino acid composition of peptides. The GPS data for the yeast stability is categorized into two labeled classes for the prediction – stable and unstable, based on the Protein Stability Index. Through a 10-fold cross-validation process, the logistic model utilized a ridge regularization penalty with a strength set at 0.001. Subsequently, the model underwent retraining using all labeled data, yielding an AUC value of 0.85. They found the unstable peptides to be predominantly hydrophobic and the negatively charged amino acids to counterbalance the instability.

Another model, deepDegron (Tokheim et al., 2021), predicts the degron motifs from human N- and C-terminal GPS profiling datasets. The input dataset contained the protein stability profile for 24-mer (with iMet) and 23-mer peptide sequences from N- and C-termini, respectively, from the human proteome. The peptides were categorized as stable and unstable based on their stability profile. A fully connected neural network was trained separately for predicting instability in N-

terminal peptides and C-terminal peptides. In both the cases, two separate models are created using models - 1. amino acid sequence and 2. amino acid composition -the difference of which indicated degron potency, to understand specific sequence motifs, than biophysical property playing an important role in mediating substrate recognition of UPS components. The model performed well with an AUC of 0.96 and 0.93, respectively. The model was further used to explore driver mutations in cancer.

Recent studies have also adopted more sophisticated mechanisms to represent proteins using the encoder-decoder method. For example, ProteinBERT (Brandes et al., 2022) uses the architecture of the Bidirectional Encoder Representations from Transformers (BERT) model. BERT is a popular model used in natural language processing tasks. ProteinBERT adopts the transformer architecture, which consists of multiple layers of self-attention mechanisms and feedforward neural networks. This architecture allows the model to capture long-range dependencies and contextual information within protein sequences. to perform two tasks: 1. bidirectional language modeling for protein sequences on 106 million proteins. The representation includes the local character level space and the global while sequence level space 2. GO annotation prediction to capture protein function on 45,000 proteins. After pre-training, the model can be fine-tuned on specific downstream tasks using supervised learning. By fine-tuning labeled datasets for tasks such as protein classification, protein-protein interaction prediction, or protein structure prediction, ProteinBERT can adapt its representations to the task at hand and achieve state-of-the-art performance. These models could eventually capture additional context-aware protein representation using transfer learning and could further increase performance in phenotype prediction.

Complex models with high performance built to predict protein instability could be used to define de novo degrons, which could be further experimentally validated (Martínez-Jiménez et al., 2020). However, most models use only sequence-based information for the prediction. The effect of biophysical properties, apart from hydrophobicity, on protein degradation is yet to be explored in detail. Additionally, though the deep learning models predict instability accurately, they are limited in inference due to their “black-box” nature, which could limit our understanding of defining new degron motifs clearly.

5.2 Aim

Short Linear Motifs, called degrons, play a crucial role in maintaining protein homeostasis by helping E3 ligases recognize the proteins to be degraded via UPS. Degrons are present in all eukaryotes, from yeast to humans. Despite their importance, degrons, their functions, and the various protein quality control pathways they are involved in are yet to be discovered entirely. With the advent of high throughput synthesis of oligonucleotide and next-generation sequencing, advanced workflows such as DMS have eased the creation of protein stability profiles for millions of peptides parallelly, which can then be used to dissect the degrons in the unstable peptides. In this study, I developed a methodology to analyze DMS experiments aimed at unraveling the specificity of degrons. The pipeline begins with assessing the quality of DMS experiments, followed by downstream analysis to deduce degron motifs. This involves employing straightforward visualization for shorter peptides and employing interpretable machine-learning techniques for longer peptides. The pipeline to analyze the DMS experiment is demonstrated using two studies:

1. Systematic study of the N-terminal amino acid specificity using the stability profile from DMS of N-terminal di-residue constructs in the *H. sapiens* cell line.

2. Systematic study of eukaryotic C-degron pathways in an unbiased fashion using stability profile of over 50k random peptides and yeast C-terminome to infer degron motifs.

5.3 Methods

5.3.1 Experimental setup

A systematic study of factors influencing the degradation of proteins is conducted using the DMS approach, which allows the simultaneous measurement of the stability of thousands of protein variants. A general outline of the DMS approach used in this study is schematically presented in Figure 8 and as described:

1. Using an oligonucleotide pool, pooled libraries made of variable regions fused with fluorescent proteins are constructed. Two fluorescent proteins that differ in their maturation rates are taken to measure the protein stability.
2. Individual libraries are then sorted into bins of varying protein stability using fluorescence-activated cell sorting (FACS) based on readouts of fluorescent proteins.
3. The variable regions in each stability bin are then barcoded by PCR and subjected to pooled high-throughput DNA sequencing.
4. The sequences are then processed and aligned to the reference genome of the model organism using PEAR and Bowtie2 algorithms as part of the NGSPipe2go pipeline before deconvoluting the reads per protein variant. Furthermore, the distribution of sequencing reads for each protein variant is used to estimate the stability of the protein variant, defined as the Protein Stability Index (PSI). PSI is a quantitative indication of the stability of the peptide of interest. Lower PSI is indicative of unstable peptides. Thus, peptides with low PSI could be indicative of the presence of degron. The pipeline was built by [redacted]. The pipeline can be accessed at https://github.com/Khmelinskii-Lab/Das1_C-degrons/tree/main/NGSpipe2goMPSprofiling.

Keeping the general protocol of DMS the same as described above, two different sets of studies are conducted as described.

To map the degradation signals at protein N-termini in mammalian cell line

Even though extensively studied, there are many gaps to be bridged in our understanding of protein degradation via N-degron pathways in humans. While certain amino acids at the N-terminus have been implicated in N-degron pathways, the specific sequence motifs and the enzymes that recognize them are still to be completely understood. Thus, to systematically identify the amino acid specificity at the N-terminus of proteins in humans for degradation, a GPS experiment on the library of the di-residue variable region was conducted by [redacted] and [redacted] in our lab.

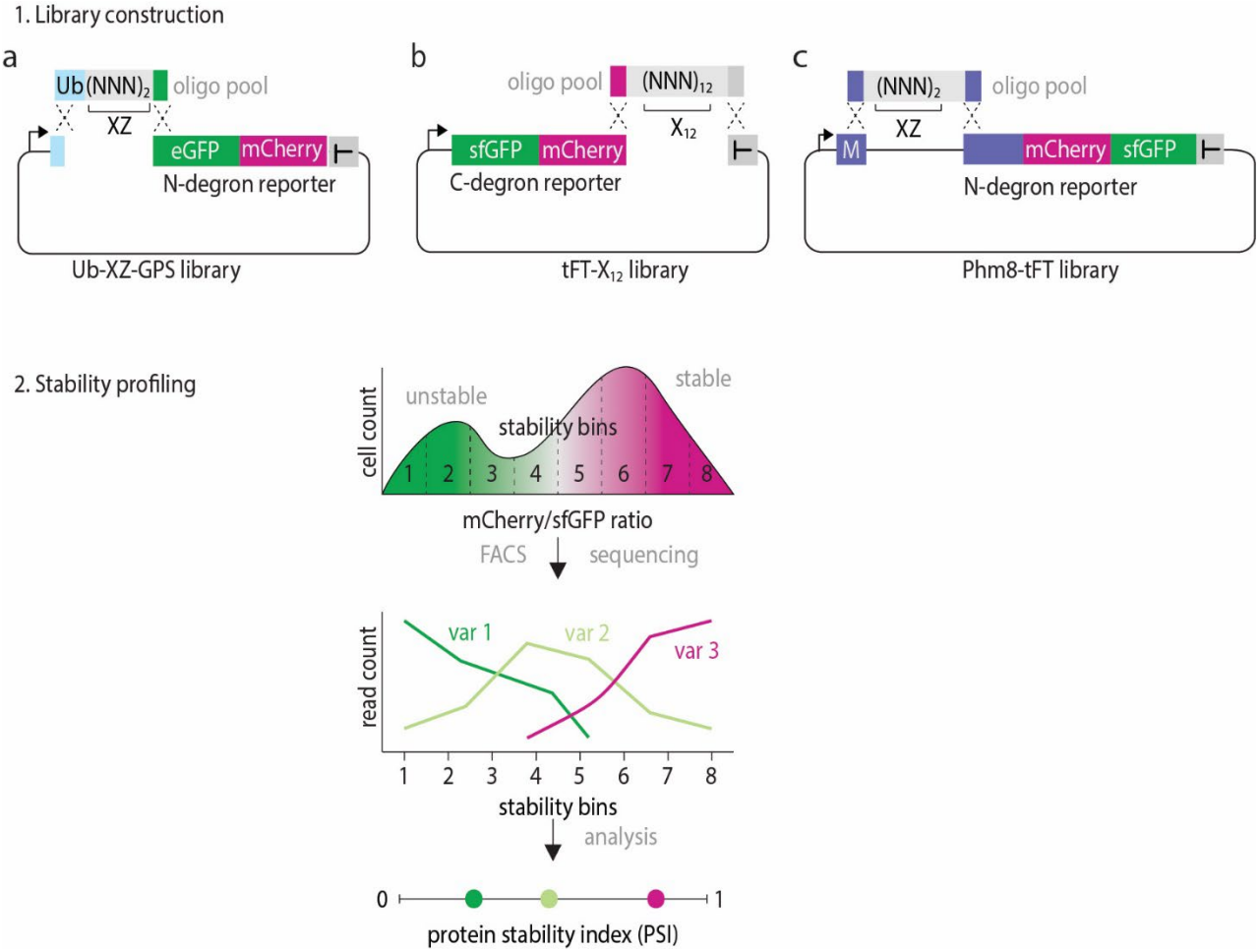

Figure 8: GPS experiments used in this study to systematically decipher N- and C- terminal degrons in yeast and humans. *Briefly, the experiments could be in two steps: 1. Library construction: Degenerate oligonucleotides encoding random peptides of defined length are cloned into a variable region in the N- and C- terminal of the reporter plasmid by homologous recombination. (a) Library to dissect specificity of degrons in human N-termini. The variable region encompasses all combinations of amino acids for di-residue peptides that have ubiquitin moiety at the N-terminus, and eGFP-P2A-mCherry reporter in the C-terminal is constructed in Flp-In T-REx 293 cells. The library is referred to as Ub-XZ-GPS library (b) Library construction for systematic study of C-degron in yeast. Random peptides of 12 amino acids are cloned into the variable region at the C-terminus of the tFT reporter. The library is referred to as the tFT-X_{12} library. (c) Library to study the amino acid specificity of N-terminal Phm8 protein. Two amino acids after the initiator methionine are mutated to all combinations in the Phm8 protein and tagged to tFT-tag at its C-terminus. 2. Stability profiling: The stability of each sequence variant in the pool is then determined. The library pool is sorted into stability bins using the fluorescence readout, followed by deep sequencing to determine the frequency of each variant across bins. The distribution of sequencing reads for each variant is summarized by a weighted average of reads.*

For this, a pooled library, called Ub-XZ-GPS, was created with all the combinations of two amino acids located at the N-terminus. Di-residues were fused with a reporter containing green fluorescent protein (eGFP) and red fluorescent protein (mCherry) with P2A sequence in between them to get separate fluorescent proteins after translation.

The ubiquitin fusion technique was used to expose di-residues at the N-terminus upon expression. In this technique, ubiquitin moiety was located before di-residues and cleaved by endogenous

deubiquitinating enzymes, exposing the N-terminal region of interest. The library was transduced into the Flp-In T-REx 293 cell line. The library is then sorted using FACS into 8 stability bins based on the ratio of eGFP to mCherry. At least 10^5 cells are sorted into bins in each of the replicates. Cells from each bin are then grown, DNA is extracted, barcoded by PCR for NGS, and processed for protein stability using the NGSPipe2Go pipeline (Figure 8, library construction (a)).

To map the degradation signals at protein C-termini in yeast cell line

While N-degrons are extensively studied, our understanding of C-degrons is still limited. Thus, to systematically study C-degrons in yeast in an unbiased fashion, multiplexed protein stability profile (MPS) is performed to screen libraries of random peptides and yeast C-terminome for putative degrons. MPS is a modified version of the GPS which allows the simultaneous measurement of the stability of thousands of protein variants. Using an oligonucleotide pool, pooled libraries of 12 amino acid long tandem Fluorescent timer (tFT)-tagged peptides are constructed. The tFT tag used in all the experiments in the studies here is a variant of mCherry/sfGFP. tFT-peptide libraries can be sorted into eight stability bins according to the mCherry/sfGFP ratio using fluorescence-activated cell sorting (FACS), followed by deep DNA sequencing to identify the peptide sequences present in each bin. The distributions of sequencing reads are finally summarized in the form of a protein stability index (PSI) for each peptide, scaled between 0 (unstable) and 1 (stable), using NGSPipe2GO. This library is hereafter referred to as the tFT-X_{12} library (Figure 8, library construction (b)). The experiments are conducted by [redacted] in our lab.

5.3.2 Analysis of N-terminal degrons in mammalian cell line

To investigate N-terminal amino acid specificity in humans, GPS on the Ub-XZ-GPS library Figure 8, library construction (a)) is conducted in three replicates, where X and Z are any two amino acids. Each of these is individually analyzed for the quality of the experiments. PSIs for each construct from the three replicates are then combined using linear regression using the “limma” package (Ritchie et al., 2015), with a false discovery rate (FDR) for each peptide calculated by the Benjamin-Hochberg procedure (Benjamini and Hochberg, 1995). The Ub-XZ-GPS is a library of smaller peptides – two amino acids long, for which inference of the putative degrons could be simply recognized by simple visualization techniques, such as heatmaps. Visual analysis of the PSIs for Ub-XZ-GPS constructs is created using the “ggplot2” package in R.

5.3.3 Analysis of C-degron in random peptide library in yeast

The sequencing data after the NGSPipe2Go pipeline is further analyzed to infer the degrons from the tFT-X_{12} library. The library is constructed with two replicates. Only sequence variants with more than 10 reads across all stability bins were considered. The downstream analysis on the tFT-

X_{12} library was performed using replicate 1, in which 46152 unique peptides were identified with more than 10 sequencing reads. Further investigation on the corresponding E3 for the tFT-X_{12} degron library was done on combined replicates, merged using linear regression in the "limma" R package ((Ritchie et al., 2015), with a false discovery rate (FDR) for each peptide calculated by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

5.3.3.1 Deep neural network and performance

Visualizations such as heatmaps can detect patterns that depict motifs in smaller constructs, up to 2 amino acids long. However, it becomes harder to interpret longer sequences using simple visualization techniques. A Fully Connected Neural Network (FCNN) has been developed to predict unstable peptides from the MPS experiments to address this issue. An example of this pipeline is shown using the tFT-X_{12} library.

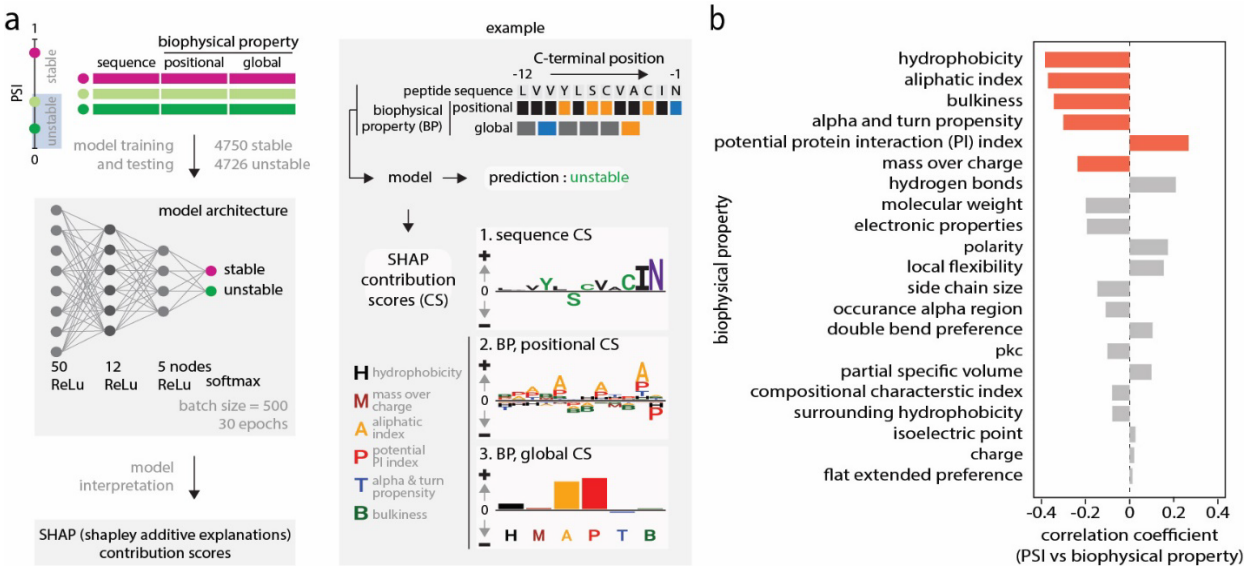

Figure 9: Creation of computational pipeline *(a) Workflow of the deep neural network model trained on the tFT-X_{12} library to classify peptides into stable or unstable groups, followed by model interpretation using SHAP contribution scores (CS). The input data, the model output and its interpretation with SHAP are exemplified using one peptide (right panel). SHAP contribution scores for the example peptide are separated by type for clarity (sequence and biophysical properties per position—represented as logos, global biophysical properties represented in a bar chart). (b) – Pearson correlation coefficients between PSI and global biophysical properties for replicate 1. Red, biophysical properties with the highest absolute correlations included in the deep neural network model from (a).*

Peptides from replicate 1 of the tFT-X_{12} library were categorized as unstable if PSI < 0.5 and stable otherwise. This threshold of 0.5 for PSIs to categorize the peptides based on stability was chosen based on the distribution of PSIs in the library, which is unimodal, centered at 0.668 ± 0.046 (median ± median absolute deviation (mad)) and skewed towards low PSIs. Out of 46152 peptides, 4726 (~10.2%) are unstable. To avoid class bias, a subset of 4750 randomly selected stable peptides and all 4726 peptides are used to create the prediction model (Figure 9(a)).

The peptides are represented by their sequences and biophysical properties – both positional and global. The primary FCNN uses six biophysical properties from the "Peptides" package in R(Osorio et al., 2015) – Hydrophobicity (Kyte and Doolittle, 1982), Aliphatic Index (Ikai, 1980), Bulkiness (Liang et al., 2008), Mass Over Charge, Potential protein Interaction Index (Boman, 2003) and Alpha and turn propensity. These properties are chosen based on their high correlation to PSI (Figure 9(b)). Overall, each peptide is represented by a vector of 318 features, comprising 240 for sequence (12 positions * 20 AA per position), 72 (12 positions * 6 biophysical properties), and six global biophysical properties).

The dataset was divided into (a) a training set, consisting of 3750 each of unstable and stable peptides, (b) a validation set, consisting of 20% of the training set taken at random, and (c) 4 test sets with peptides, excluded from the training set. Three test sets had no peptides in common (Test sets 1–3), and a fourth test set was created by combining test sets 1–3. These sets were used to check the model performance in seen data (training and validation) as well as unseen data (test sets). The four test sets were used to ensure that the assessment of the model performance was not affected by the selection of the test set.

The neural network is fully connected and comprises five layers. It includes three internal layers, each with 50, 12, and 5 nodes. All the internal layers use a dropout rate of 0.3 and the rectified linear unit (ReLU) activation function. The second and fourth internal layers have L1 and L2 regularizers of 0.001, respectively. The output layer uses the softmax activation function. The model uses binary cross-entropy as a loss function, Adam optimizer, with a batch size of 500. It runs for 30 epochs, and the parameters for the model were selected using the brute force method to ensure optimal performance and eliminate over- or underfitting of the data (Figure 9(a)). This model is labeled as "primary model" and is used for downstream analysis to extract motifs. The model is evaluated using six metrics – Binary Cross entropy, Accuracy, Recall, Precision, AUC-ROC, and Mean-Squared Error, as defined below.

TP refers to True Positives. TP are unstable peptides correctly classified as unstable.
TN refers to True Negatives. TN are stable peptides correctly classified as stable.
FP refers to False Positives. FP are stable peptides predicted as unstable.
FN refers to False Negative. FN are unstable peptides predicted as stable.

Loss: The loss defined is binary cross entropy. Binary cross entropy is a loss function that helps evaluate the dissimilarity between the predicted probability distribution and the true label. Not only does this help in evaluating the model performance, but it is also helpful in internal model optimization. For each peptide i with actual class y[i] (y[i] = 1 if unstable, 0 otherwise),

$$Binary\ cross\ entrophy = \frac{1}{N}\sum_{i=1}^{N} -(y[i] * \log(p[i]) + (1 - p[i]) * \log(1 - p[i]))$$

Where p[i] is the probability of predicting class 1 (unstable in our case) and (1-p[i]) is the probability of predicting class 0 (stable in our case).

Accuracy: Accuracy measures the number of correct predictions over the total number of predictions.

$$\text{Accuaracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Recall: Recall or True Positive Rate (TPR) is the evaluation metric that measures the percentage of peptides correctly classified as unstable with respect to the peptides classified as unstable. Recall provides the sensitivity of the model, i.e., the probability that the actual unstable will test as an unstable peptide.

$$\text{Recall/TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision: Precision or specificity helps evaluate how good the model is at predicting unstable peptides. For a balanced class, high precision means a lower false positive. Precision is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

AUC-ROC: AUC-ROC, or Area under the Receiver Operating Characteristics curve, compares the True Positive Rate with the False Positive Rate and indicates classification problems at different thresholds. Intuitively, a model with a high AUC suggests that the model can classify unstable peptides as unstable and stable ones as stable. The curve is a plot of TPR versus FPR, where TPR is as defined in Recall and FPR is:

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

$$\text{AUC} - \text{ROC} = \frac{1 + \text{TPR} - \text{FPR}}{2}$$

Mean Squared Error: Mean Squared Error (MSE) is one of the simplest loss functions that indicates the actual difference between the predictions by the model and the actual class. For a peptide i, with class y[i] and prediction ¥[i], MSE is calculated as:

$$\text{MSE} = \frac{1}{N} + \sum_{i=1}^{N} (\text{y[i]} + \text{¥[i]})$$

Apart from the primary model described above, I also created six additional models based on different feature combinations to understand each feature's effect independently. To include the information on all the stable peptides, I took all the 41,426 stable peptides from replicate 1 of the tFT-X_{12} library and overrepresented the unstable peptides so that the classes are balanced. All 23 biophysical properties from the Peptides package in R to create the models (Figure 9(b)). The models are:

1. Model based only on AA sequence: The model is trained using amino acid sequence. The peptides are represented by a one-hot encoded vector with 20 amino acids for each of the 12 positions.
2. Model based only on positional biophysical properties: The model takes only the positional biophysical properties as input. 23 biophysical properties represent each position in the peptide. Thus, a vector of 276 length represents the peptides (12 positions * 23 properties = 276)
3. Model based only on global biophysical properties: The model takes global biophysical properties as input to train the model. Each peptide is represented by 23 biophysical properties.
4. Model based on AA sequence and positional biophysical property: The model takes two inputs – 1. amino acid sequence and 2. Positional biophysical property. Each peptide consists of a one-hot encoded amino acid vector and vector of 23 biophysical properties for 12 positions.
5. Model based on AA sequence and global biophysical property: The model takes two inputs – 1. amino acid sequence and 2. Global biophysical property. Each peptide consists of a one-hot encoded amino acid vector and vector of 23 biophysical properties.
6. Model based on AA sequence, positional biophysical, and global biophysical property: The model takes three inputs – 1. amino acid sequence, 2. 23 Positional biophysical properties for 12 positions, and 3. 23 global biophysical properties.

Apart from the metrics mentioned earlier, the F1-score was used to evaluate the model performance. F1-score is the harmonic mean of precision and recall and is defined as follows:

$$\mathrm{F1-score} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

5.3.3.2 Interpretation of deep neural network

The fully connected neural network helps in the accurate prediction of protein stability. However, to illustrate what factors contribute to predicting the stability of peptides, I mounted a SHapely Additive explanation (SHAP)(Lundberg and Lee, 2017) on the model. SHAP is a cooperative game theory technique that helps understand the contribution of individual features in predictive models. The average of all the dataset's actual class labels is considered the “reference prediction”. The prediction is extracted for all different permutations of features, called “actual model prediction”. The feature importance is based on the average of the difference between the actual model prediction and a reference prediction. The importance score for each feature per peptide was obtained from SHAP on 4110 correctly predicted unstable peptides from replicate 1 of the tFT-X_{12} library. The number of Monte Carlo samples was set to the default of 100 in the “iml” R package (Molnar, 2018). The feature importance score is referred to as the “Contribution Score” (CS) for the given feature towards the instability of the peptide. SHAP was mounted on the

primary model and each of the additional six models with the same parameter to find the CS due to each of the properties independently. SHAP provides a comprehensive understanding of the importance of features and their interaction effects, enabling users to gain valuable insights into model behavior and decision-making processes.

5.3.3.3 Defining C-degron motifs

SHAP on the correctly predicted unstable peptides gives a CS for each feature for each of the peptides. We used the clustering on CS for each peptide to create representative motifs by identifying common features recognized by the model that cause instability in peptides. Since the SHAP score created a high-density cluster, a simple k-means approach (Lloyd, 1982) is used for clustering. The number of clusters is chosen using the Elbow method, which uses the Within-Cluster Sum of Square (WSS) method. For the maximum number of cluster K_{max}, the elbow method plots WSS for each K from 1 to K_{max}. For cluster C of m peptides with centroid feature values (C1, C2, C3,..Cn) and peptide CS CSi represented as CSi1, CSi2, CSi3,.CSin, WSS is defined:

$$WSS(C) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} (Ci - CSji)^2 \right)^{1/2}$$

The sum of all WSS for all the clusters from 1 to K gives the WSS value for k number of clusters. The maximum value of WSS is when K = 1, and it saturates parallel to the x-axis as we increase the number of clusters, creating an elbow. The value of k for which the WSS saturates is taken as an optimum number of clusters. WSS was performed on the CS by the sequences, and biophysical properties for all 4110 correctly predicted unstable peptides, with the maximum number of clusters chosen to be 150. The optimal number of clusters chosen for clustering CS is 56. These clusters consisted of predominantly important internal positions, and some clusters had no particularly interesting patterns from the features. To define the C-degron motifs, 21 clusters with at least one mean SHAP sequence contribution score greater than 0.05 at positions −5 to −1 were considered. For these potential C-degrons, sequence and position-specific biophysical property logos were generated, with positive values reflecting positive contributions to the unstable prediction. Mean contribution scores of global biophysical properties were displayed as bar graphs (Example in Figure 14(b)).

5.3.4 Analysis of Doa10, Skp1 and Das1 substrates

Out of 21 C-degron clusters, three of the predominant clusters had ΦN at the C-terminal, Φ = [ILMV]. To systematically understand the factors responsible for the degradation of ΦN at the C-terminal, 10 peptides from 4726 unstable peptides are chosen. These peptides were either the

least hydrophobic or had the most sequencing reads in the ΦN group. To evaluate the importance of amino acids, the C-terminus of the peptides is capped before performing MPS in WT. Additionally, saturation mutagenesis is conducted to understand the importance of another amino acid at all of the 12 positions.

Fluorescent measurement suggested the involvement of only one non-essential gene- Das1, an F-box substrate receptor of SCF (Skp1, cullin, F-box), predominantly working with the Cdc34 ubiquitin-conjugating enzyme. Capping and saturation mutagenesis of tFT- ΦN is further conducted in *das1Δ*, *cdc53-1* mutant, and *cdc34-1* mutant with temperature sensitive allele at 37°. For all the experiments, sequences with reads more than 10 are only considered.

Furthermore, to systematically decipher the relevance of these components on C-degrons, MPS profiling is performed on the tFT-X_{12} degron library in *das1Δ*, *cdc34-1*, *cdc53-1*, and *doa10Δ* mutants. These libraries contained all putative degrons from the tFT-X_{12} library and a set of stable peptides as controls. Experiments were performed by [redacted] in multiple replicates (3 replicates of WT at 30°, 3 replicates of tFT-X_{12} degron library in *das1Δ* at 30°, 2 replicates of tFT-X_{12} degron library in *doa10Δ* at 30°, 2 replicates each of WT, *das1Δ*, *cdc34-1* and *cdc53-1* at 37°). After ensuring the reproducibility of the experiments, the PSI of the replicates was combined using linear regression with FDR for each peptide calculated by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) using the "limma" package(Ritchie et al., 2015) in R. In the tFT-X_{12} degron library, peptides are categorized into 3 categories based on the following conditions:

- stabilized (s): PSI [WT] 0.58 & PSI [mutant] – PSI [WT] > 0.1
- partially stabilized (p): PSI [WT] < 0.5 & PSI [mutant] < 0.58 & PSI [mutant] – PSI [WT] > 0.1
- not affected (n): PSI [mutant] – PSI [WT] > 0.1 & p-adj > 0.05 | PSI [mutant] – PSI [WT] < 0.1,

where mutant refers to *das1Δ* or *doa10Δ* and p-adj refers to the adjusted p-value post-regression from the replicates.

5.3.5 Analysis of yeast C-terminome

To identify endogenous substrates of SCF^{Das1}, the C-termini of all yeast proteins (C-terminome) were surveyed for Das1 degrons. MPS profiling of a yeast C-terminome library consisted of 12 AA long C-termini of 6793 annotated yeast open reading frames (ORFs) fused to the tFT. Sequences with sequencing reads of more than 10 are only considered. The endogenous Das1 degron in yeast C-terminome library is categorized into 3 categories:

- stabilized (s): PSI [WT] < 0.5 & PSI [*das1Δ*] > 0.58 & PSI [*das1Δ*] – PSI [WT] > 0.1;
- partially stabilized (p): PSI [WT] < 0.5 & PSI [*das1Δ*] < 0.58 & PSI [*das1Δ*] – PSI [WT] > 0.1;

- not affected (n): PSI [WT] < 0.5 & PSI [*das1Δ*] – PSI [WT] > 0.1 & p-adj > 0.05 | PSI [WT] < 0.5 & PSI [*das1Δ*] – PSI [WT] < 0.1

where p-adj refers to the adjusted p-value post-regression from the replicates.

To evaluate the relevance of endogenous Das1 substrates from yeast C-terminome, I checked the type of ORF and performed a Gene Ontology analysis. Yeast C-terminome were mapped to the corresponding ORF information taken from the Swiss-Prot-reviewed set of *S. Cerevisiae* proteins in UniProt(UniProt Consortium, 2023). Classification of 6611 yeast open reading frames by type (verified, uncharacterized and dubious) and the corresponding slim gene ontology terms were retrieved from YeastMine (Balakrishnan et al., 2012).

5.4 Results

5.4.1 Quality assessment pipeline of DMS experiments shows good quality N- and C-degron library

The deep sequencing experiments employed in this study, such as MPS and GPS, utilize pooled libraries of constructs with variable regions categorized into distinct stability bins based on fluorescence ratios. Subsequently, these libraries undergo deep sequencing, with sequencing reads summarized as Protein Stability Index (PSI). Given the complexity of this pipeline, there exist numerous potential sources of errors, encompassing both technical and biological factors such as PCR cycle inaccuracies and issues arising from deconvolution. These errors have the potential to compromise the integrity of downstream analyses. Therefore, assessing the quality of MPS/GPS experiments before proceeding with downstream analyses is imperative to ensure reliability and confidence in the obtained results. This quality assessment not only aids in mitigating technical inconsistencies but also offers insights into pertinent biological phenomena, such as the impact of constructs on cell fitness and toxicity, which may inadvertently affect experimental outcomes. The implemented pipeline for evaluating the quality of MPS/GPS experiments in this study comprehensively addresses technical noise, assesses the efficacy of deconvolution, and evaluates for unwanted stop-codon mutations, thus enhancing the robustness of the experimental findings (Figure 10(a)).

Percentage of reads with stop codons.

Cells have a small fluorescence of their own. Low mCherry and sfGFP (or eGFP) signals would thus represent the auto-fluorescence of the cell. This would also be true for the reporters with an unintentional stop codon mutation as the variable region. We are only interested in sequences that do not have any stop codon. These are expected to have higher mCherry and sfGFP intensity than that of the stop codon. Thus, the selection of an optimal threshold for mCherry and sfGFP indicates low background noise and low stop codon mutations (Figure 10 (a)) 1. Percentage of Stop Codon). For instance, For the Ub-XZ-GPS library, Less than 11 % of di-residues are stop codons in all the three replicates in the Ub-XZ-GPS library in the Flp-In T-REx 293 cell line, which ensures that the selection of the optimum threshold of mCherry and sfGFP (Figure 10(b)).

The higher percentage of stop codons also indicates probable biological phenomena. For instance, in a different experiment on the effect of N-terminal amino acid in protein stability of PHosphate Metabolism Phm8 in yeast, [redacted] performed the MPS on overexpressed, mutated Phm8 (mutating the first two N-terminal amino acids after iMet, Figure 8, library construction (c)). In this library, ~70% of reads were stop-codons. Most stop codons in this experiment also have high sequencing reads, indicating probably their heavy duplication in the PCR cycle. This high

percentage of stop codon reads was found due to high cell toxicity upon Phm8 overexpression (Figure 10(c)).

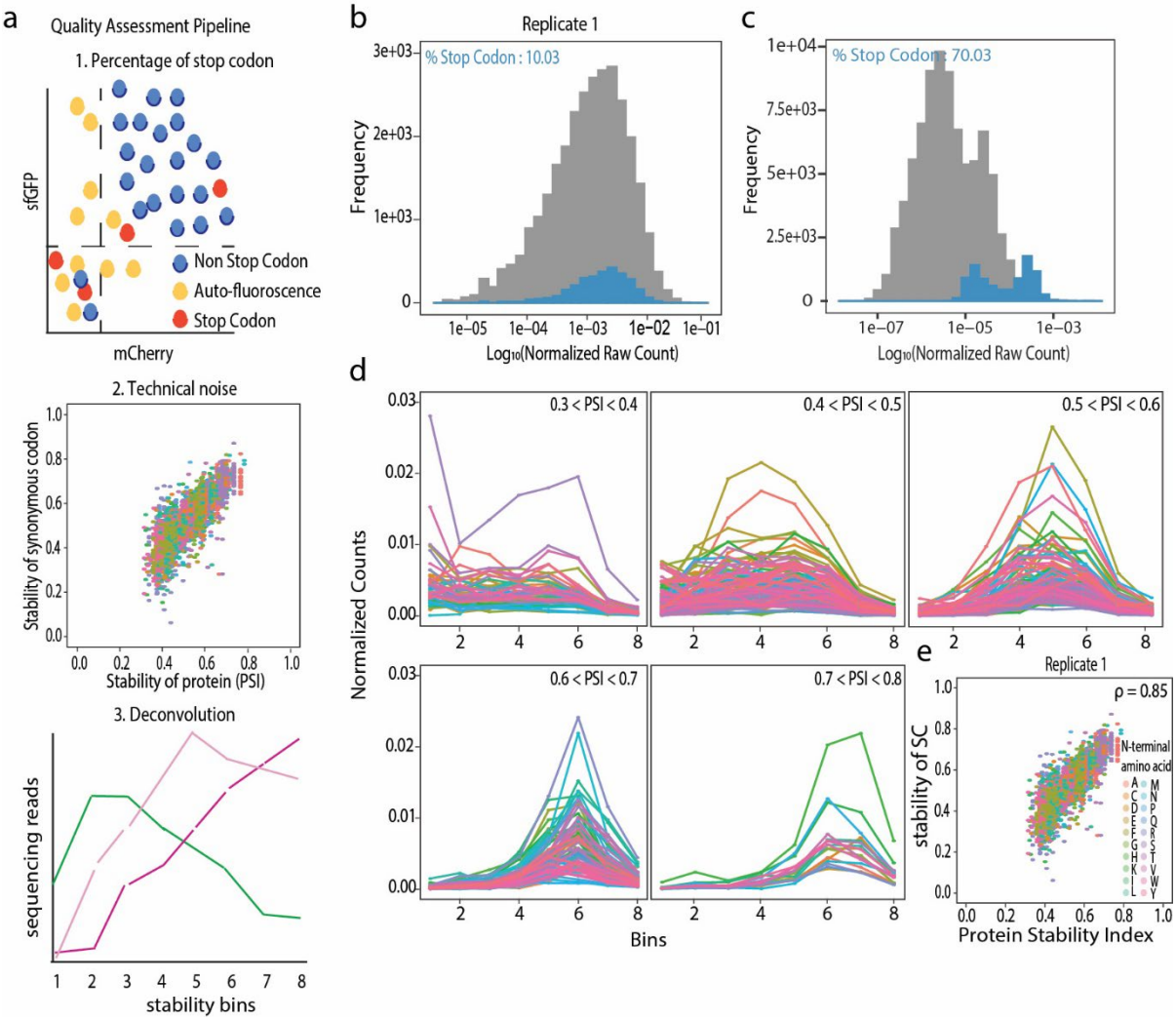

Figure 10: Quality assessment pipeline shows good quality N-degron. *(a) Components of Quality assessment pipeline. The quality assessment evaluates the percentage of stop codons, compares the stability of codons with proteins, and evaluates deconvolution. (b) Variation of frequency of normalized raw count in replicate 1 of UB-XZ-GPS library. The normalized raw count is reported in Log10 format. Blue indicates the region of the stop codon. Data is shown only for replicate 1. Similar trends are seen for other replicates as well. (c) Variation of frequency of normalized raw count in replicate 1 of PhM8-MPS library in yeast. The normalized raw count is reported in Log10 format. Blue indicates the region of the stop codon. (d) Variation of normalized count per stability bin for deconvoluted translation. Each colored line shows one translated peptide. Translations are grouped according to the PSI interval range. Data is shown only for replicate 1. Similar trends are seen for other replicates as well. (e) Variation of protein stability index with the stability of synonymous codon stability. Translations are color-coded based on the N terminal residue, ρ: Pearson correlation coefficient. Data is shown only for replicate 1. Similar trends are seen for other replicates as well.*

Technical noise

An amino acid can be translated using different sets of codons during the process of translation in the cell. For example, both AAA and AAG are translated to amino acid K. Here, I refer to these codons that translate to the same amino acid as "synonymous codons". Codon usage can affect protein folding (Liu, 2020) and possibly protein stability. In the MPS experiment, the sequencing reads are converted to PSI by aggregating the stability of the synonymous codons they are translated from. To evaluate the effect of codons on protein stability, I compared the variation of stability of synonymous codons with the stability of the protein. In case the codon does not affect the protein stability, the variation would mostly be linear. Large deviation of the two stability might also be indicative of noise due to many technical reasons, such as the need for optimization in PCR cycles. For the Ub-XZ-GPS library, we find that the variation of the two stabilities is mostly linear, with a correlation of 0.84, indicating that the choice of codons does not significantly affect the stability and substantially low noise in our library (Figure 10(e))

Deconvolution of the pooled library from MPS

Deconvolution of the oligonucleotide pool contributes to read count variation for individual regions of interest. Deconvolution would ideally lead to an unimodal distribution of sequence reads for each peptide of interest across stability bins. To evaluate the pattern of deconvolution for peptides in the library, variation of sequencing reads across bins per peptide is plotted. However, with the increase in peptide length, the number of combinations of peptides increases exponentially. For example, for a peptide with two AA, the number of combinations is 400 (20^2) combinations, which increases to 8000 (20^3) when the peptide of interest is 3 AA long, and so on. This plotting and checking individually would be laborious. On the other hand, plotting all the peptides in the library into one graph would make it messy and hard to understand any anomaly. So, I divided the translations into 10 groups based on PSI intervals and plotted the sequencing reads per bin for the group. With the increase in the PSI group, read counts should be higher for higher bins. For the Ub-XZ-GPS library, Higher bins have higher read counts for stable translations (Higher PSI) (Figure 10(d)). We see similar variation in quality for yeast tFT-X_{12} library.

5.4.2 N-degron library shows a close relation to the Arg/N-degron pathway

To dissect the N-terminal AA specificity in humans, GPS on the library of di-residue after ubiquitin moiety, tagged to eGFP-P2A-mCherry, is performed in replicates of three to extract PSI (Ub-XZ-GPS, Figure 8, library construction (b)). The experiments are replicable, as shown by the high correlation among the PSIs of the constructs replicated in the library (Figure 11(a)). The correlation between any two replicates in the library is above 0.9. These PSIs from the replicates are then linearly regressed to obtain representative PSIs for each of the 400 constructs in the library. Variation of the PSI for the construct by amino acids at each position shows a close relation to the Arg/N degrons pathway (Figure 11(b)). In the Arg/N-degron pathway, destabilizing amino

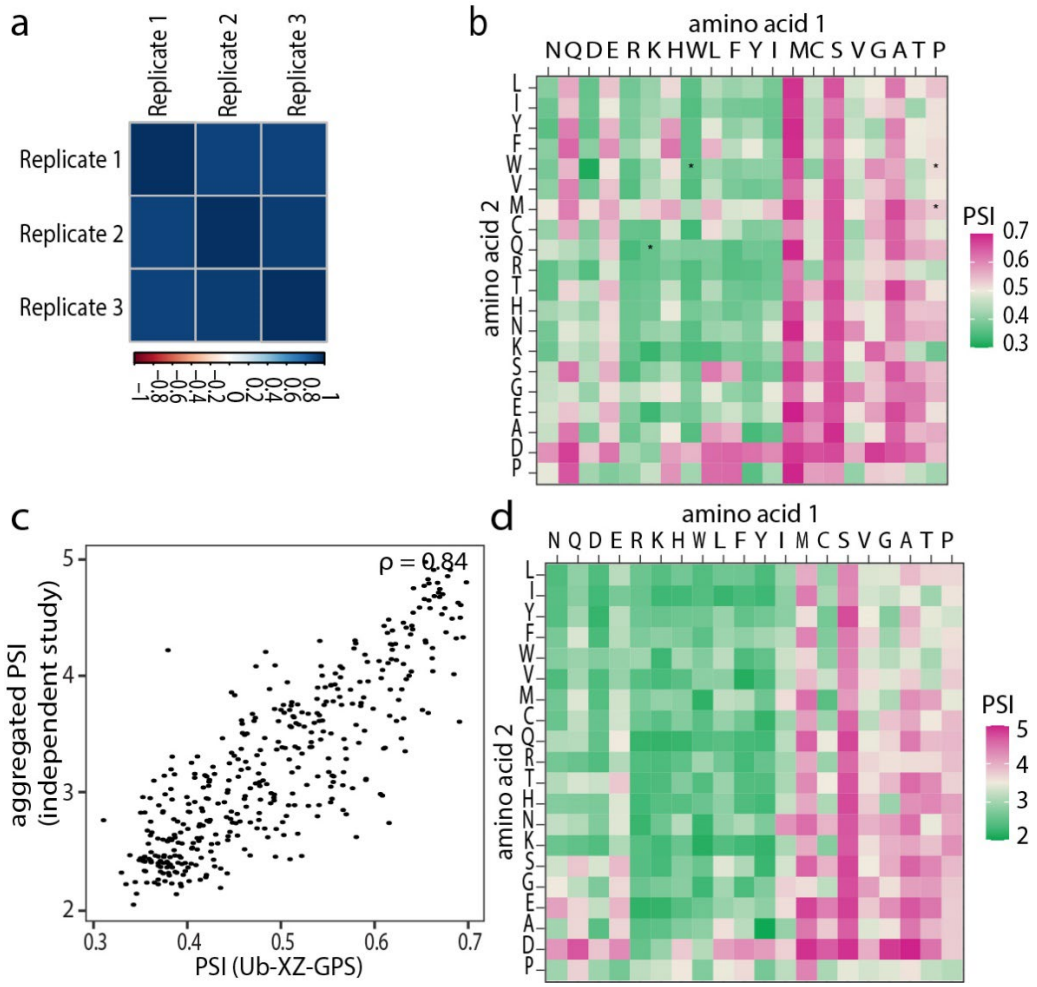

Figure 11: N-degron library closely relates to Arg/N-degron pathway*: (a) Correlation among the PSI of different replicates in Ub-XZ-GPS library. Correlation is calculated using Pearson correlation (b) Heatmap depicting PSI for Ub-XZ-GPS library after merging replicates. *: high deviation in PSI among replicates (c) Variation of aggregated PSIs for construct grouped by N-termini di-residue from all 23000 primary isoforms of human proteins from (Timms et al., 2019) with PSIs of constructs in Ub-XZ-GPS library, ρ: Pearson correlation coefficient (d) Heatmap depicting of aggregated PSIs for construct grouped by N-termini di-residue from all isoforms of human proteins from (Timms et al., 2019).*

acids are grouped into the primary amino acid group (L, F, W, Y, I, R, K, H), the secondary amino acid group (C, D, E), or the tertiary amino acid group (N, Q) (Figure 4(a))(Bachmair et al., 1986). In the Ub-XZ-GPS library, we find that the constructs with the primary amino acid group of Arg/N-degron pathway at the N-terminus are highly unstable, mostly irrespective of the amino acids at the second position (Figure 11(b)). An exception to this is the presence of D at the second position, which has a stabilizing effect on the construct. However, the presence of positively charged amino acids, lysine and arginine, at the N-terminus has a destabilizing effect on the construct, even if there is D at the second position.

While the presence of D at the N-terminus destabilizes the construct as expected (with construct "DW" being the most unstable in the library), other secondary amino acid group members from

the Arg/N-degron pathway, E at the N-terminus stabilizes the construct (Figure 11(b)). A similar trend is noticed for the N-terminal amino acid from the tertiary amino acid group of the Arg/N degrons pathway in the Ub-XZ-GPS library, where N at the N-terminus leads to instability (except when there is a D at the second position), and Q stabilizes the construct, in contrast to our expectation from Arg/N-degron pathway. This could have many explanations. One of which is that processing by NTAQ1 is blocked, that the effect of N-amidase NTAQ1 for peptides ending with Q and ATE1 for peptides ending with E is somehow blocked, suggesting potential roles of other potential factors to degrade these peptides. The constructs specific to the Ac/N-degrons pathways are comparatively stable (M, A, S). Except for the 'PS' construct, all others grouped with the N-termini amino acids in the Ac/N-degrons pathway are comparatively stable in the Ub-XZ-GPS library.

In an independent study, GPS on the first 23 amino acids, without iMet, of the primary isoform(s) of all human protein, cloned to Ub-GPS expression vector is performed, resulting in the PSIs of ~23000 peptides (Timms et al., 2019). I grouped these peptides based on the first two amino acids at the N-termini and then aggregated the PSI per group. The high correlation of the two experiments (Ub-XZ-GPS and data from Timms) shows that for most peptides, the N-termini of most peptides harbor adequate potential to elucidate their stability (Figure 11(c)). Additionally, the variation of aggregated PSI from the study shows a similar trend as seen in the Ub-XZ-GPS library (Figure 11(d)). The peptides beginning with the amino acid from the Ac/N-degrons pathway are stable, while peptides beginning with the amino acids in the Arg/N-degrons pathway are degraded (except peptides beginning with Q and E). For all the N-terminus amino acids from the Arg/N-degron pathway, except for R, K and H, D at the second position stabilizes the peptides. This independent study further confirms the effect seen in the Ub-XZ-GPS library (Figure 11(b)).

5.4.3 C-terminal degrons in yeast show a preference for hydrophobic peptides.

To systematically identify the C-degron in *S. cerevisiae* in an unbiased fashion, MPS profiling of the tFT-X_{12} library, consisting of random 12 amino acid-long peptides at the C-terminus of the reporter, is performed to extract the protein stability (Figure 8(b)) in the replicates of 2. The experiments were replicable (with the Pearson correlation coefficient of 0.82). Further analysis is conducted in replicate 1 of the tFT-X_{12} library. The library has a uniform composition of amino acids across all positions (Figure 12(a)). Hereafter, the 12 amino acids in the X_{12} peptides are numbered from −12 (the most proximal to the tFT) to −1 (the most C-terminal). Over 10% of peptides (4726 out of 46152 detected sequences) were putative degrons (Figure 12(b)). Putative degrons are defined using a threshold of 0.5 as described in the Methods. Grouping all peptides by the C-terminal amino acid showed that a larger fraction of peptides ending in one of three amino acids [C, F, V] were unstable (Figure 12(b)). Moreover, putative degrons were, on average, enriched in hydrophobic amino acids [C, F, I, L, M, V, W, Y] in all positions (Figure 12(d)),

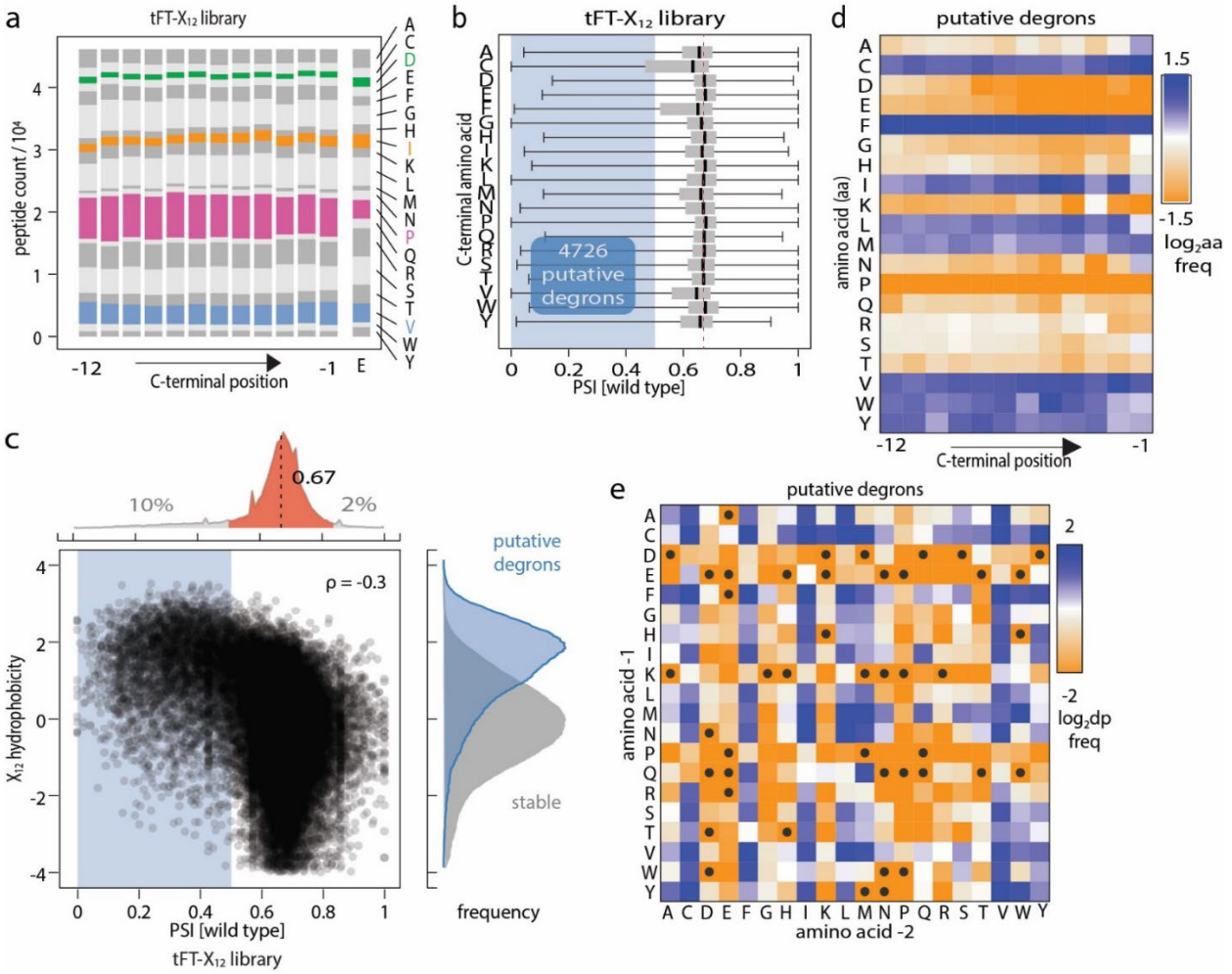

***Figure 12: C-terminal degrons are hydrophobic in nature**: (a) Amino acid frequency in the tFT-X_{12} library. Number of X_{12} peptides with each amino acid per C-terminal position. The expected distribution (E) based on equal frequency of all codons is shown for comparison. (b) Distribution of PSIs in the tFT-X_{12} library determined by MPS profiling. Peptides were grouped by the C-terminal amino acid. Blue region marks putative degrons (PSI < 0.5). For all boxplots hereafter, centerlines mark the medians, box limits indicate the 25th and 75th percentiles, and whiskers extend to the minimum and maximum value in each group. (c) Correlation between peptide hydrophobicity (Kyte-Doolittle hydropathy scale) and PSI in the tFT-X_{12} library. Blue region marks putative degrons. ρ, spearman correlation coefficient. The median ± 3.7 x median absolute deviation range (Methods) is highlighted in red. Hydrophobicity distributions for the putative degrons and stable peptides (right) (d) Relative amino acid frequency (log2 transformed) per position in the 4726 putative degrons from the tFTX_{12} library, normalized to the relative frequency in the whole library. Blue – amino acid enrichment, orange – depletion in the putative degrons (e) Relative frequency of dipeptide motifs (dp) at C-termini of 4726 putative degrons from the tFT-X_{12} library, normalized to the relative frequency in the whole library. Motifs absent from the C-termini of the putative degrons are marked (●).*

suggesting hydrophobicity to be an important determinant in the tFT-X_{12} library. Supporting this notion, peptide hydrophobicity showed a moderate negative correlation with the PSIs in the library (Figure 12(c)), consistent with prior studies of terminal degrons both in yeast and human cells(Hasenjäger et al., 2023; Koren et al., 2018; Kats et al., 2018; Mashahreh et al., 2022). Moreover, most putative degron peptides in the tFT-X_{12} library with [C, I, L, F, V] at the second position from the C-terminus (amino acid -2) tend to destabilize the peptide (Figure 12(e)).

5.4.4 Interpretable fully connected neural network helps in classification and inference of C-degrons

Although hydrophobic regions are prevalent (Figure 12(c)), there are many other interesting trends in putative degrons. For instance, amino acids N and A are enriched at the C-terminus (Figure 12(d)).

Thus, to study the sequence specificity of the putative degrons in tFT-X_{12}, I used a deep learning approach mounted with Shapely Additive explanation (SHAP) for interpretability. The model is trained on replicate 1 of the tFT-X_{12} library. Apart from the sequence information for these peptides, six biophysical properties that correlated to the PSIs from the tFT-X_{12} library were used (mentioned in Methods). To find the effect of amino acids of the peptide (referred to as sequence) and biophysical property (both positional and overall) on the stability of the peptide, I trained a fully connected neural network on ~4500 peptides of each class to classify them into stable and unstable classes, based on all these properties taken simultaneously (Figure 9(a)). The fully connected neural network has 5 layers, with 3 internal layers of 50, 12 and 5 nodes. Each internal layer had a dropout of 0.3 and a rectified linear unit (reLU) activation function. The second and fourth internal layers had L1 and L2 regularizers of 0.001. The output layer had a softmax activation function. The model had binary cross-entropy as a loss function, Adam optimizer, batch size of 500 and runs in 30 epochs. The classification model performed well with a training accuracy of 0.85 and a training AUC of 0.93. Consistent performance is seen across all four test and validation datasets created out of replicate 1 of the tFT-X_{12} library (Figure 13(a)).

Next, I assessed the performance of the model in replicate 2. Replicate 2 has 47876 peptides with read counts over 10, with 4837 having PSI < 0.5. Approximately 87% of unstable peptides and 89% of stable peptides are correctly predicted. In addition, replicate 1 and replicate 2 of the tFT-X_{12} library are linearly regressed to obtain the combined data for replicates of tFT-X_{12}, which consist of 43095 peptides. In the combined dataset, approximately 90% of unstable peptides (3973 of the 4412) are predicted correctly by the primary model. The performance of models on the three datasets and the corresponding confusion matrix is shown in Table 1.

Table 1: Performance of the model on different datasets

	loss	accuracy	AUC	precision	recall	MSE
Replicate 1 (model performance)	0.32	0.85	0.93	0.85	0.85	0.09
Replicate 1(test set 4)	0.36	0.82	0.91	0.82	0.82	0.11
Replicate 2	0.30	0.89	0.94	0.89	0.89	0.09
Combined	0.29	0.89	0.95	0.89	0.89	0.09

Combined refers to the dataset after linear regression in the replicates of tFT-X_{12}.

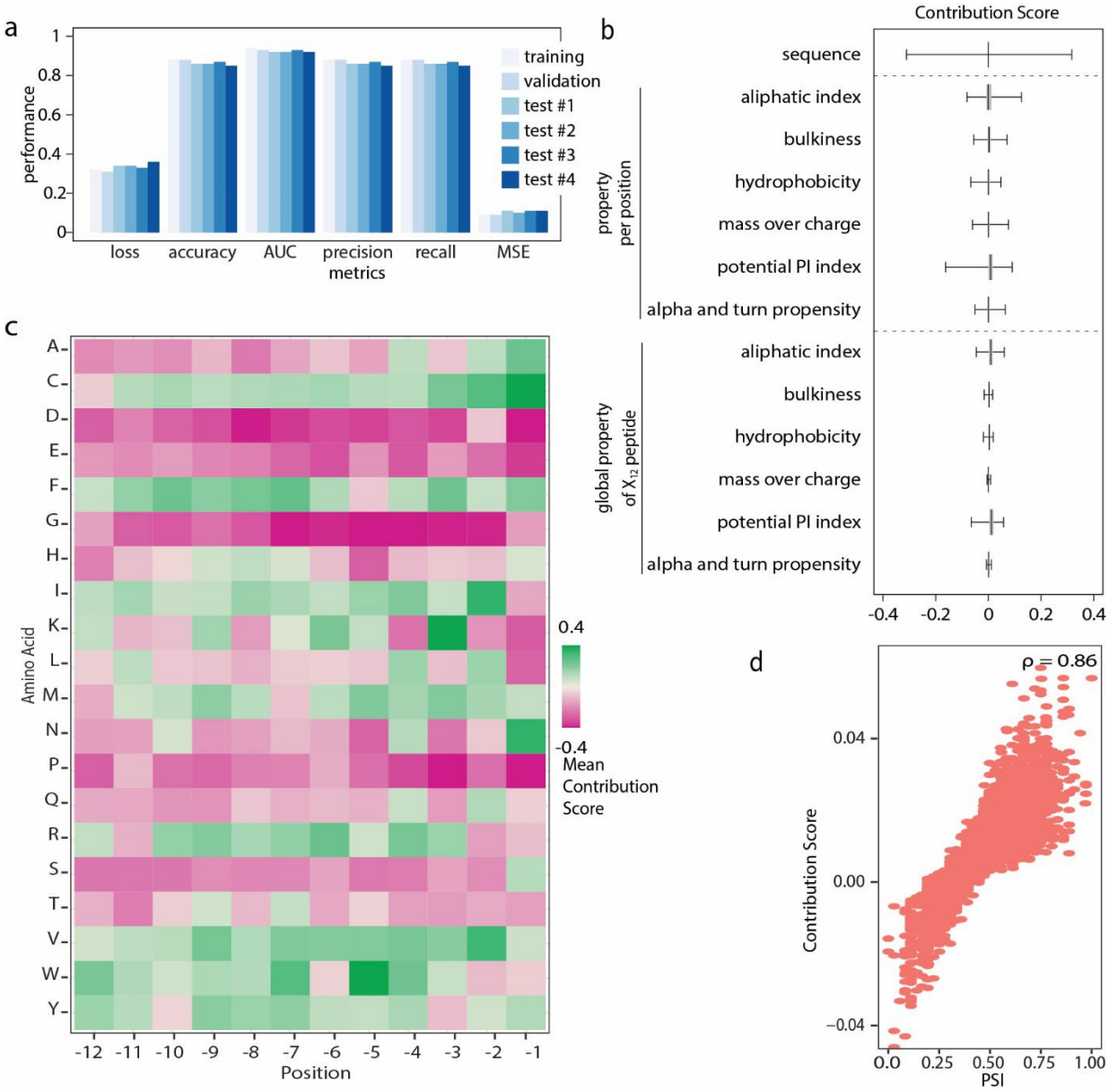

Figure 13: Interpretable Fully Connected neural network helps in classification and inference of C-degrons*: (a) Performance metrics for the deep neural network model from Figure 9, trained on the tFT-X_{12} library to classify peptides into stable or unstable groups. AUC is the area under the receiver operating characteristic curve; MSE is the mean squared error. (b) Distributions of SHAP contribution scores for 4110 peptides classified as unstable by the deep neural network model (c) Variation of mean contribution score from 4110 unstable peptides for each amino acid at each position. Higher CS has the potential to cause degradation. (d) Variation of contribution score with the aliphatic index for 4110 unstable peptides, ρ: Pearson Correlation Coefficient*

Though the model predicts the stability of the peptides in the tFT-X_{12} library with high performance across varied datasets, I also wanted to understand what features help the model predict the stability class. Thus, to illuminate the "black-box" nature of the model, a Shapely Additive explanation (SHAP) is mounted onto the model described above. SHAP contribution

scores for the 4110 correctly classified putative degrons showed that the peptide sequence contributed the most to the prediction (Figure 13(b)).

Detailed analysis of the mean SHAP score by the amino acids at each position reveals the favorability of C and N at the C-terminus in the set of putative degrons (Figure 13(b)). Additionally, SHAP indicates the importance of hydrophobic residues (I, F, L and V) in the second position from the C-terminus. Other interesting findings include the importance of K at -3, W at -5 and A at -1 and -2. One of the most common C-degron motifs is the C-terminal G, with CRL complexes $Cul2^{KLHDC2}$, $Cul2^{KLHDC3}$, and $Cul2^{KLHDC10}$, exhibiting distinct preferences for recognizing –GG, –[R, K]G, and –[A, W, P]G motifs respectively. Additionally, $Cul5^{KLHDC1}$ can recognize C-terminal –GG degrons (Timms and Koren, 2020). Thus, we expected a positive SHAP contribution by G at the C-terminus of the putative degrons. However, for most unstable peptides, G is not noticed as an important contributing factor for instability at either the C-terminus (position -1) or at the second position from the C-terminus (position -2) in the tFT-X_{12} screen. Meanwhile, apart from Kelch-family substrate receptors, the $Cul2^{APPBP2}$ complex also recognizes C-terminal glycine degrons but with a preference for RxxG or RxxxG motifs, where x could be any amino acid (Timms and Koren, 2020). Though not many, for the peptides where G contributes positively to instability, I also observe a higher contribution score for R at -3 (Figure 13(c)), which lies in line with the existing results. Apart from the sequence and biophysical properties - both positional and global, the aliphatic index, in particular, also contributed to the classification (Figure 13(b, d)). Out of six global biophysical properties, only the potential PI index is anti-correlated with PSI (correlation coefficient -0.87) (Figure 13(d)). The potential PI index is scored using the function from (Boman, 2003). It proposes a protein interaction index from a protein's amino acid sequence, which sums solubility values for all residues, potentially estimating a peptide's ability to bind to membranes or other proteins. The anti-correlation of PSI with the contribution score by the potential PI index might suggest that the unstable peptides have a lower potential to bind to other proteins that could stabilize by probable complex formation.

5.4.5 Clustering of SHAP helps us understand C-degron motifs

Individual SHAP scores for peptides help understand the effect of sequence and biophysical properties in predicting stability. To identify degron motifs, I clustered the SHAP score using k-means. The number of clusters chosen was 56 with the help of the WSS Score, as described in the methods. Number of peptides in clusters vary from 4 to 279. The clusters are represented by a representative peptide, which is the most similar to the rest of the peptides in the cluster (Figure 14(a)). Similarity between the peptide sequences of the cluster is calculated using local alignment using the Smith-Waterman algorithm (Smith and Waterman, 1981).

Of 56 clusters, 21 have a mean sequence contribution score suggestive of C-degron. These clusters have at least one mean SHAP sequence contribution score greater than 0.05 at positions

−5 to −1. The total number of peptides in these 21 clusters is 1476, comprising 35.9 % of putative degrons. Figure 14(b) shows the contribution score based on sequence and biophysical properties for each cluster. These clusters show the importance of asparagine and cysteine at the C-terminus. Also, the dominance of hydrophobic residues (I, L, M and V) at -2, cysteine at -2 and tryptophan at -5 is consistent with our previous result (Figure 13(a))

Another interesting C-degron cluster is the xxKIN motif, where x can be any amino acid (Figure 14 (b), cluster c31). Peptides in this cluster are more hydrophilic than any other cluster. Interestingly, for all the peptides with C-terminal KIN in tFT-X_{12}, the PSI < 0.5, suggesting KIN to be a strong C-degron motif.

Table 2: Performance of additional models

Metrics	Sequential[1]	Positional Biophysical[2]	Global Biophysical[3]	Sequential + Positional Biophysical[4]	Sequential + Global Biophysical[5]	All[6]
Loss	0.13	0.1	0.33	0.15	0.15	0.18
Accuracy	0.959	0.97	0.8	0.96	0.96	0.95
Precision	0.948	0.96	0.83	0.95	0.95	0.94
Recall	0.97	0.97	0.89	0.97	0.97	0.968
AUC	0.98	0.994	0.93	0.98	0.98	0.984
F1-Score	0.96:0, 0.97:1	0.96:0, 0.97:1	0.86:0, 0.87:1	0.96:0, 0.96:1	0.97:0, 0.97:1	0.96:0, 0.96:1

The feature descriptions corresponding to the models are described in Method. The model number corresponding to the method is superscripted in the table cell.

For all the C-degron clusters, the majority of the clusters have minimal variation in contribution score for hydrophobicity, Mass over charge and bulkiness (Figure 14(b)). However, the contribution score is comparatively higher for the aliphatic index and potential protein interaction index of the clusters. Apart from the C-degron clusters, the clusters also indicate the presence of internal degrons in the library (Figure 15(b), cluster c13)

Though the primary model described above suggests the features that contribute to peptide instability when all properties are taken together, I also wanted to investigate how individual features – sequence and each biophysical property – both positional and global, contribute to the prediction separately. This could also reduce any unforeseen bias on one feature due to another. Additionally, I underrepresented the stable class peptide by taking only a subset of ~4000 peptides from ~41000 stable peptides to avoid class bias in the primary model. Thus, to incorporate all the data from the tFT-X_{12} library, I took all of the stable peptides to build the six independent models. I created models separately and then ensemble them to test for contribution by sequence and biophysical properties individually, as described in the Methods section.

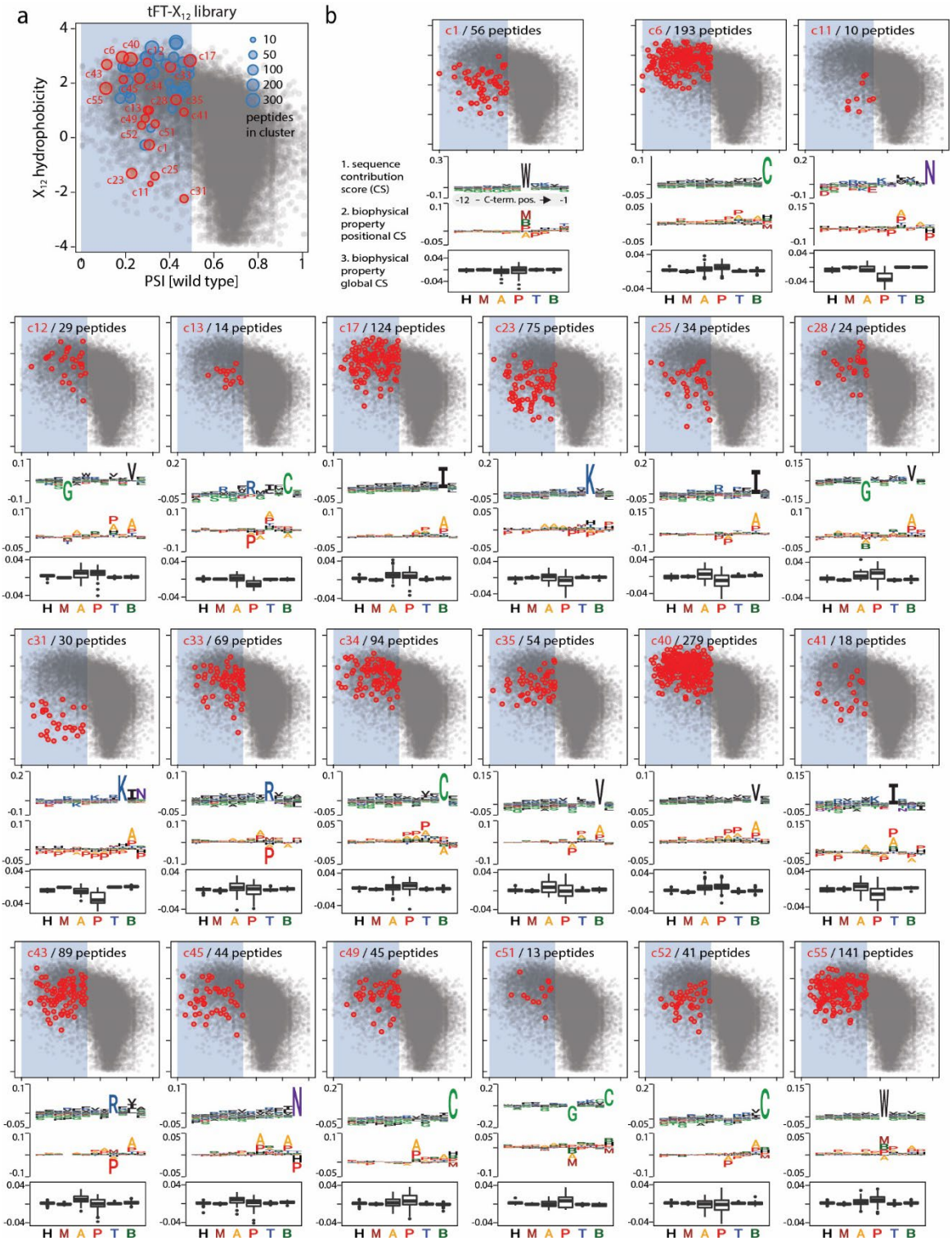

Figure 14: Clustering of SHAP helps us understand C-degrons motifs*: (a) Distribution of putative degrons in the tFT-X_{12} library into 56 groups by k-means clustering of vectors of SHAP contribution scores. Clusters with at least one mean sequence contribution score > 0.05 at positions -5 to -1 are highlighted. (b) Description of clusters highlighted in a. Top, hydrophobicity (y-axis) and PSI (x-axis) for all peptides in a cluster (axes as in a), compared to all peptides in the tFT-X_{12} library (grey). Bottom, mean SHAP contribution scores of the peptides in the cluster. In the box plots, centerlines mark the medians, box limits indicate the 25th and 75th percentiles, and whiskers extend to 1.5xIQR (interquartile range) in each group. Axis units and labels are omitted for simplicity.*

Table 2 shows the performance of all the models. Models based on individual properties (like the model just based on amino acid sequence or positional biophysical property) perform better than the ensemble of the models. The model based only on amino acid sequence or only positional biophysical property performs comparatively better than the primary model with all the properties taken together. This could be because the input data in the models contains all the stable peptides, thus enhancing the information given to the model for prediction. Additionally, the model based only on the biophysical property uses all 23 biophysical properties, unlike the primary model that uses the six biophysical properties correlated with the PSIs. The model built on the 23 global biophysical properties performs worst. This suggests that amino acid composition plays a pivotal role in understanding protein stability for most peptides. Additionally, the F1-score of both stable and unstable classes are comparable in all the models, suggesting no class-based bias from the models. It is also noticed that the ensemble of the model performs similarly to the individual models, suggesting that using sequences with biophysical properties may be redundant. However, we still use the ensemble of all the models for further analysis since it is easier to interpret the effect of biophysical properties on stability this way.

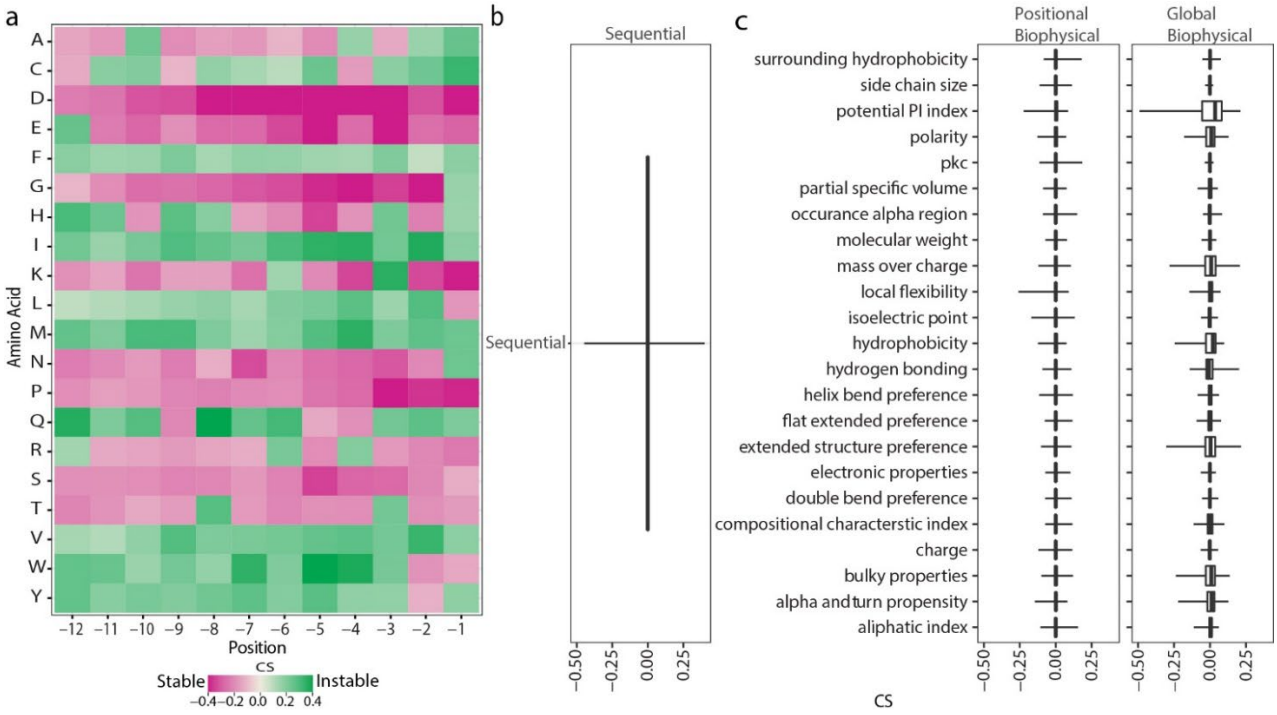

Figure 15: Variation of SHAP score based on models based on individual properties*: (a) Variation of mean Contribution score from 4658 correctly predicted unstable peptide for each amino acid at each position by model created based on sequence only. (b,c) Distributions of SHAP contribution scores for 4658 correctly predicted unstable peptides by separate models – (b) sequence, (c) positional biophysical property (left), and global biophysical property(right).*

The ensemble of models is built with individual properties to gather the patterns from each model and culminate them to extract meaningful patterns to decipher the factors contributing to

instability. All the models are given equal weight. The variation of the mean contribution score per amino acid at each position shows a similar trend as the primary model. However, G at the C-terminus is seen to be important when taken only in sequence as a property for degradation (Figure 15(a)). Other interesting amino acids for predicting the instability of the peptides, which were not captured in the primary model, include T at -3 and R at -4 and -6. Comparing the contribution score by sequence and biophysical property – both positional and biophysical suggests that the sequence of amino acids still plays a crucial role in predicting protein stability (Figure 15(b,c)). However, SHAP score variation of the global biophysical property shows that some properties not directly correlated to PSIs also help predict peptide stability. This set of properties includes hydrogen bonds and extended structure preference (Figure 15(c)). Overall, the ensemble models of individual properties yield better ease in understanding the effect of biophysical properties on protein stability.

5.4.6 Das1 recognizes C-terminal ϕN motifs in yeast

Clustering of the SHAP contribution score from the primary model gave 21 C-degron clusters. Since most degrons are hydrophobic with degradation pathways known for some (Koren et al., 2018), we focused on clusters containing some of the least hydrophobic degrons in the tFT-X_{12} library. We focused on factors that help in targeting them to degradation. These clusters show a strong contribution towards instability for C-terminal N and constitute 87 peptides (Figure 14 (b), c11, c31 and c45). Large hydrophobic amino acids Φ = [ILMV], but not [FW] at position -2, also appeared to contribute towards instability in these clusters, suggesting ΦN as a potential C-degron motif. Of 87 peptides with N at the C-terminus, 56 had Φ at the second position (-2). Further supporting this analysis, the IN, LN, MN and VN dipeptides were enriched at the C-termini of the 4726 putative degrons in the tFT-X_{12} library (Figure 12(e)). Moreover, mean PSIs of peptides with C-terminal IN, LN and VN were among the least when compared against the internal positions.

To identify the machinery targeting ΦN C-degron, ten unique C-degrons with peptides ending in IN, LN, MN and VN which satisfy at least one of the criteria: lowest PSI, least hydrophobic or with most sequencing reads in the replicate 1 of tFT-X_{12} library were chosen to identify the UPS factor targeting these peptides to degradation. A representative stable construct is also taken as a control to study the degradation factors (Figure 16(a)). The chosen representative peptides were C-degrons, as they degraded once the translation was blocked with the help of cyclohexamide chase (CHX) experiments (Kong, Shankar et al., 2023). The potent peptides from the CHX experiments, IN2, LN1. LN3 and VN2 were crossed with 138 non-essential UPS components using synthetic genetic array (SGA) semi-automated crossing. Fluorescence measurement of the resultant colony highlighted only DAS1 as a non-essential UPS gene that was involved in the

turnover of the four peptides. Knocking out Das1 completely stabilized IN2, LN1, LN3 and VN2, while not affecting the

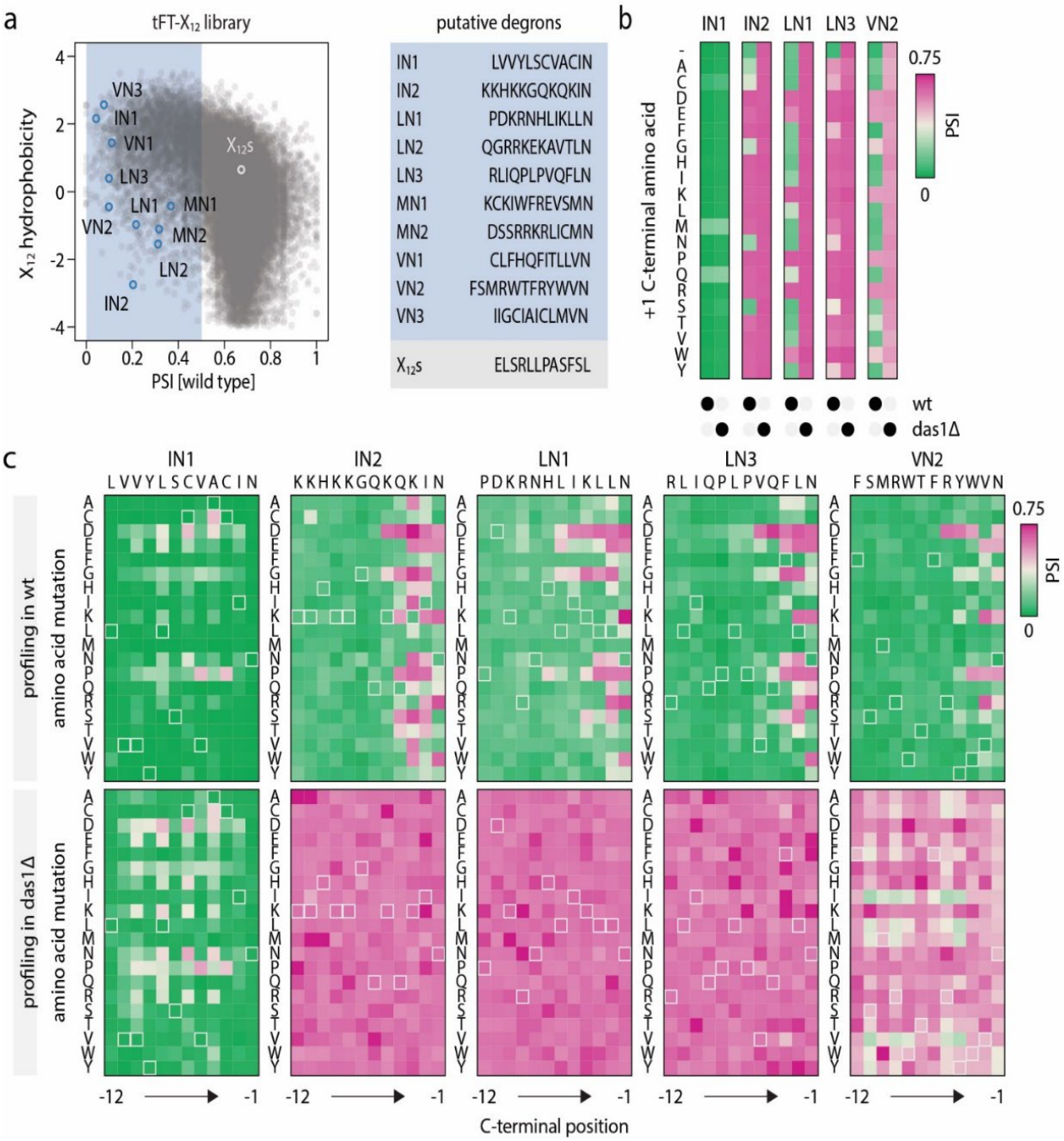

Figure 16: ΦN motifs are recognized by Das1*: (a) Representative C-terminal ΦN peptides labeled in the variation of hydrophobicity versus PSI from MPS replicate 1 of tFT-X₁₂ library. The sequence of the representative peptides is mentioned (right). Stable control is also labeled in white. (b) Variation of PSI after C-terminal capping of representative peptides with ΦN at C-terminus in WT and das1Δ genetic background. The peptides are labeled in (a). (c) Saturation mutagenesis after mutating all positions of representative peptides with ΦN at C-terminus in WT and Das1Δ genetic background.*

stable peptide Das1 is an F-box substrate receptor, conserved in yeasts, of the SCF (Skp1, cullin, F-box) ubiquitin ligase. The SCF works predominantly with the Cdc34 ubiquitin-conjugating enzyme (Seol et al., 2001; Kus et al., 2004; Finley et al., 2012).

To understand the specificity of SCF^{Das1}, capping experiments were performed by adding AA to the C-terminus (+1 position) of the constructs, with IN1 considered as a control. IN1 remained unstable irrespective of the AA added to the C-terminus. The addition of any AA at the C-terminus of IN2 and VN3 stabilized the peptide, suggesting that they are C-degron and get degraded in a DAS1-dependent manner. In contrast, most amino acids at the +1 position did not affect the Das1-dependent turnover of tFT-LN1 and tFT-VN2. Despite these differences, there were trends common to the four Das1 degrons: capping the C-terminus of the ΦN motif with one of six amino acids (D, E, K, R, P and W) stabilize the peptide and persistence of Das1-dependent turnover upon addition of one of the four amino acids (A, C, N and S) (Figure 16(b)).

To investigate the effect of amino acids at positions other than the C-terminal, saturation mutagenesis of the ΦN mentioned above (IN2, LN1, LN3 and VN2) representative peptide is conducted. In saturation mutagenesis, each amino acid is substituted with all possible amino acids at that position to understand the sequence requirement for Das1 recognition. MPS profiling of the saturation mutagenesis library is then conducted and revealed IN1 remained unstable and Das1-independent irrespective of the mutated position. For the IN2, LN1, LN3 and VN2 constructs, mutations of amino acids at positions -12 to -6 had essentially no effect on their turnover, consistent with the notion that these peptides are C-degrons. Consistent with the results of capping experiments, mutation of C-terminal N to any of the six AA (D, E, K, R, P and W) stabilized the peptides. Multiple amino acids, including charged ones (D, E, K and R) and (G, P and S), were disallowed at positions -5 to -2. Consistent with the SHAP analysis, amino acids (I, L, M, V) were preferred at position -2 compared to (F or W). Irrespective of mutation at any position, Das1 KO stabilized the peptides except I and V at the internal positions of VN2 (Figure 16(c)).

Overall, the most important non-essential UPS gene that targets the ΦN is the SCF^{Das1}, with high sequence specificity to the C-terminal positions. Mutagenesis of the representative ΦN peptides reveals that six specific amino acids (D, E, K, R, P, and W) at the C-terminus stabilized the peptides while (A, C, N, and S) did not affect degradation. Additionally, positions -12 to -6 seem to have less effect on protein stability.

5.4.7 Effect of QC factor on orphan C-degron

The four representative ΦN peptides from the least hydrophobic SHAP-based clusters were Das1-dependent. The apparent promiscuity of Das1 suggested by the mutagenesis is thus far of these four ΦN C-degrons. Therefore, to conduct a systematic survey of Das1-degron, MPS profiling of

putative degrons in the tFT-X12 library with different genetic backgrounds (*das1Δ, cdc34-1, cdc53-1* and *doa10Δ*) was performed as discussed in this section.

5.4.7.1 Doa10 predominantly recognizes arginine at C-terminal ends

The library of the tFT-X_{12} degrons consisted of many hydrophobic degrons. Hydrophobic degrons are known to be targeted by Doa10. MPS profiling of the tFT-X_{12} degron library in doa10Δ is also conducted. The Doa10 is a membrane-embedded E3 ubiquitin ligase and is involved in endoplasmic reticulum-associated degradation (ERAD) (Swanson et al., 2001; Krshnan et al., 2022). Doa10 is responsible for rapidly degrading membrane proteins with misfolded or unassembled states. Doa10 targets a variety of hydrophobic degrons, including N-termini of secretory proteins, C-termini of tail-anchored proteins and other hydrophobic sequences with transmembrane domain-like properties (Ravid et al., 2006; Maurer et al., 2016). Apart from the degrons in the tFT-X_{12} library, 274 stable peptides from tFT-X_{12} are also taken as a control for the study.

MPS profiling of multiple replicates per mutant showed good reproducibility (Figure 17(a)), and the stable control peptides were not affected in any mutant, as expected (Figure 17(b)).

The doa10Δ completely stabilized only 37 constructs from the tFT-X_{12} degron library (Figure 17 (c)). These degrons had high overall hydrophobicity, similar to the rest of the unaffected tFT-X_{12} degron library (Figure 17(d)) and, consistently, were enriched in the hydrophobic amino acids F and L (Figure 17(e)). 21 out of the 37 doa10Δ stabilized peptides had R at the C-terminus (Figure 17 (e, g)). To test if the R at the C-terminus is a key factor for degradation, MPS profiling of capped variants of five out of the 37 Doa10 degron is conducted (R1 to R4 with R at -1, R5 with R at -2 (Figure 17(c)). A capping experiment reveals that adding a single amino acid to the C-terminus had negligible impact on the Doa10-dependent turnover of tFT-tagged R1 to R4 peptides, indicating that these are not C-degrons (Figure 17(f)). Further, saturation mutagenesis on the peptides, such that all positions are mutated with all amino acids, reveals that key determinants of Doa10-based degradation are mostly hydrophobic residues at the C-terminal end (position -5 to -1)(F, I, L, M and V) (Figure 17(h)). Interestingly, replacing the -1 arginine with one of three amino acids (A, C and N) in R1-R4 peptides yielded degrons that were largely Doa10-independent. In contrast, the R5 peptide appeared to be a C-degron with strict requirements across the C-terminal positions and that remained unstable but Doa10-independent with most amino acids at the +1 position.

Overall, from this experiment, I conclude that deep mutational scanning on degrons from the tFT-X_{12} library has peptides targeted by substrate receptors or E3 ligases other than Das1. Although Doa10 substrates have enrichment of C-terminal arginine, saturation mutagenesis and capping experiments reveal that these tested peptides may not be C-degron as other amino acids at -1 other than R also have a destabilizing effect on the tested peptides.

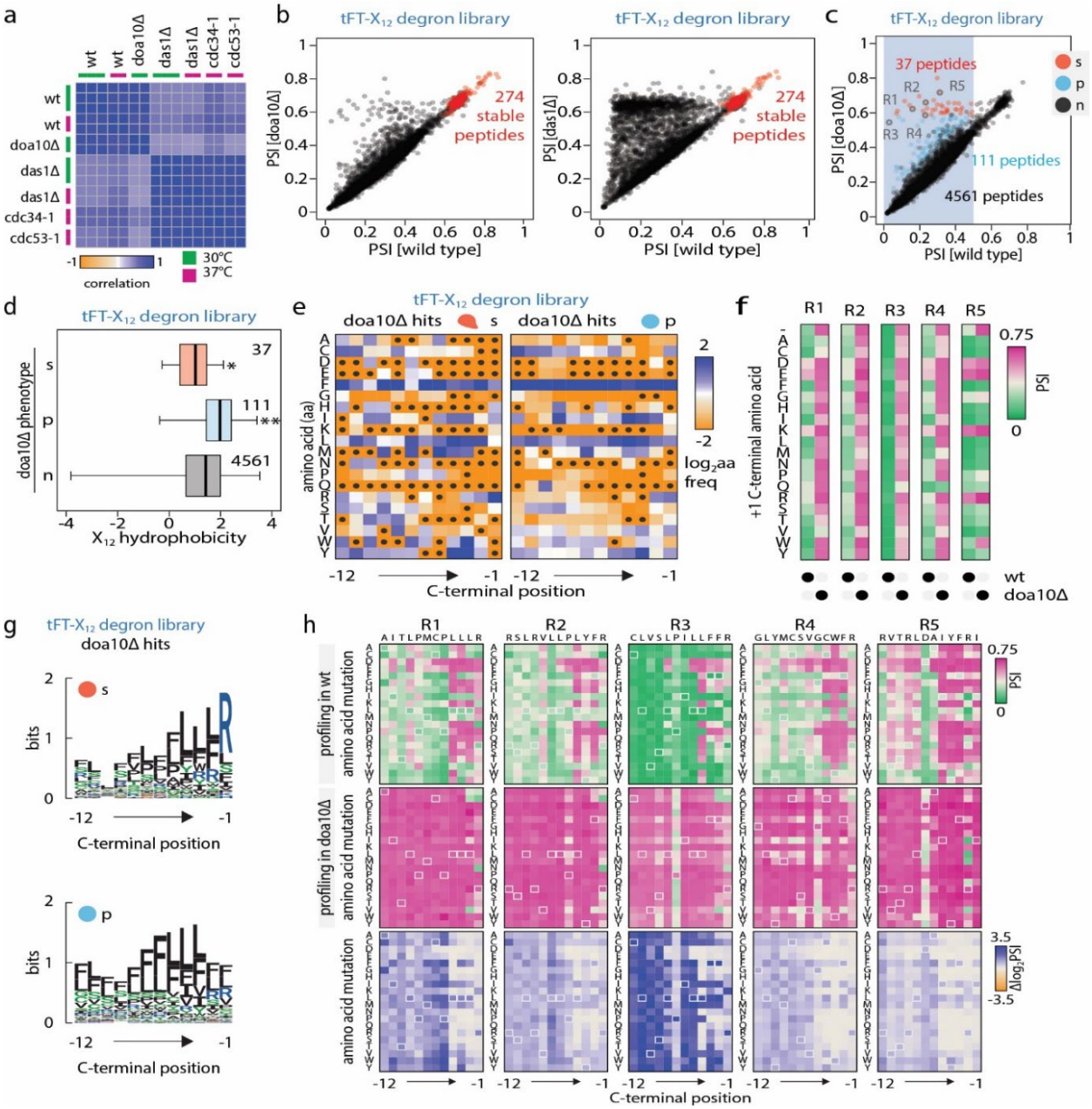

Figure 17: Doa10 predominantly recognizes arginine at C-terminal ends: *(a) Heatmap displaying Pearson correlation coefficients of PSIs from multiple replicates from tFT-X_{12} C-degron library in indicated genetic background. (b) Variation of PSI from tFT-X_{12} C-degron library in WT versus doa10Δ genetic background (left) and WT versus das1Δ genetic background (right). Stable controls are indicated in red. (c) Variation of PSI from MPS performed on tFT-X_{12} C-degron library with stable control (in (b)) in WT and doa10Δ genetic background. Doa10-dependent peptides are in the blue region (PSI[WT]< 0.5). These are further sub-classified into three groups – s: stabilized upon Doa10Δ, p: partially stabilized by doa10Δ , and n: not affected by doa10Δ. Number of peptides in each class are also indicated. (d) Variation of hydrophobicity for the classes mentioned in (c). Centerlines mark the medians, box limits indicate the 25th and 75th percentiles, and whiskers extend to the minimum and maximum value in each group. *p = 0.012, **p = 4 × 10−10 in a one-sided Mann-Whitney U-test. (e) Variation of relative amino acid frequency in Doa10-dependent constructs (s and p category in (c)), normalized to the relative frequency in the tFT-X_{12} library. Amino acids absent per position in Doa10-dependent constructs are marked with black circles (f) Variation of PSI after C-terminal capping of representative peptides with arginine at C-terminus in WT and doa10Δ genetic background. The peptides are labeled in (c). (g) Sequence logo of Doa10 dependent degron (top) stabilized (s) group; (bottom) partially stabilized (p) group*

(h) Saturation mutagenesis after mutating all positions of representative peptides with R at C-terminus in WT and doa10Δ.

5.4.7.2 SCF^{Das1} recognizes atleast five distinct but overlapping C-degrons motifs in yeast

Das1 is an F-box substrate receptor of SCF ubiquitin ligase that predominantly works with the Cdc34 ubiquitin-conjugating enzyme (Finley et al., 2012). Thus, to understand if Cdc34 and the SCF subunit cdc53 are involved in the Das1-dependent turnover, MPS on tFT-X_{12} degrons under *das1Δ, cdc34-1* mutant and *cdc53-1* mutant conditions are conducted. MPS on *cdc34-1* mutant and *cdc53-1* mutant conditions are at restrictive temperatures (37 °C).

MPS profiling of the tFT-X_{12} degron library under *das1Δ*, *cdc34-1* and *cdc53-1* mutant conditions shows a high correlation among the stability of their constructs (Figure 18(a)). This suggests that Das1 recognizes most degrons in the library targeted by the SCF ubiquitin ligase. This starkly contrasts with observations in human cells, where cullin-RING ubiquitin ligases target a variety of C-degrons using receptor subunits dedicated to different C-degron motifs (Lin et al., 2018; Koren et al., 2018). Approximately 36% of putative degrons from the tFT-X_{12} degron library are Das1-dependent, with 20% of constructs completely stabilized (Figure 18(b)). The peptides completely stabilized by *das1*Δ are hereby indicated as Das1-degrons. Das1-degrons were enriched in three AA residues (C, M and N) at the C-terminus and were depleted of six amino acids (D, E, K, R, P and T) (Figure 18 (d)).

They also show enrichment of hydrophobic amino acid at -2. This is consistent with the saturation mutagenesis of the four representative ΦN C-degrons (Figure 16(c)), further highlighting the broad specificity of Das1. Moreover, Das1-degrons are less hydrophobic than the rest of the peptides in the tFT-X_{12} degron library (Figure 18(c)). The partially stabilized Das1-dependent degron is enriched with hydrophobic amino acid from position -12 to -1 and tends to be more hydrophobic than the Das1-degron. (Figure 18(c,d)).

Apart from the enrichment of [ILM] at -1, Das1 degrons were also enriched in W at -5 (Figure 18 (d)), which suggested the possibility of a separate degron motif. Thus, to test the importance of W at -5, saturation mutagenesis is performed on three representative peptides (W1, W2 and W3) from the Das1-degron set. These peptides had a G (W1 and W3) or a S (W2) at position -1 (Figure 18(b)). The capping experiment revealed that these peptides are C-degrons, as adding any AA to the C-terminus stabilized the peptides (Figure 18 (e)). Only serine at +1 preserved the Das1-dependent turnover of W1 and W3 constructs. Furthermore, saturation mutagenesis of W1, W2 and W3 revealed that mutation to any six amino acids (D, E, K, R. P and W) at the C-terminus stabilized the peptides (Figure 18 (f)). Moreover, mutation of W at -5 to any other amino acid also

stabilized these proteins, highlighting the importance of W at -5. Additionally, ΦN at C-terminus seems not to affect the peptide turnover, suggesting W at -5 to be a separate degron.

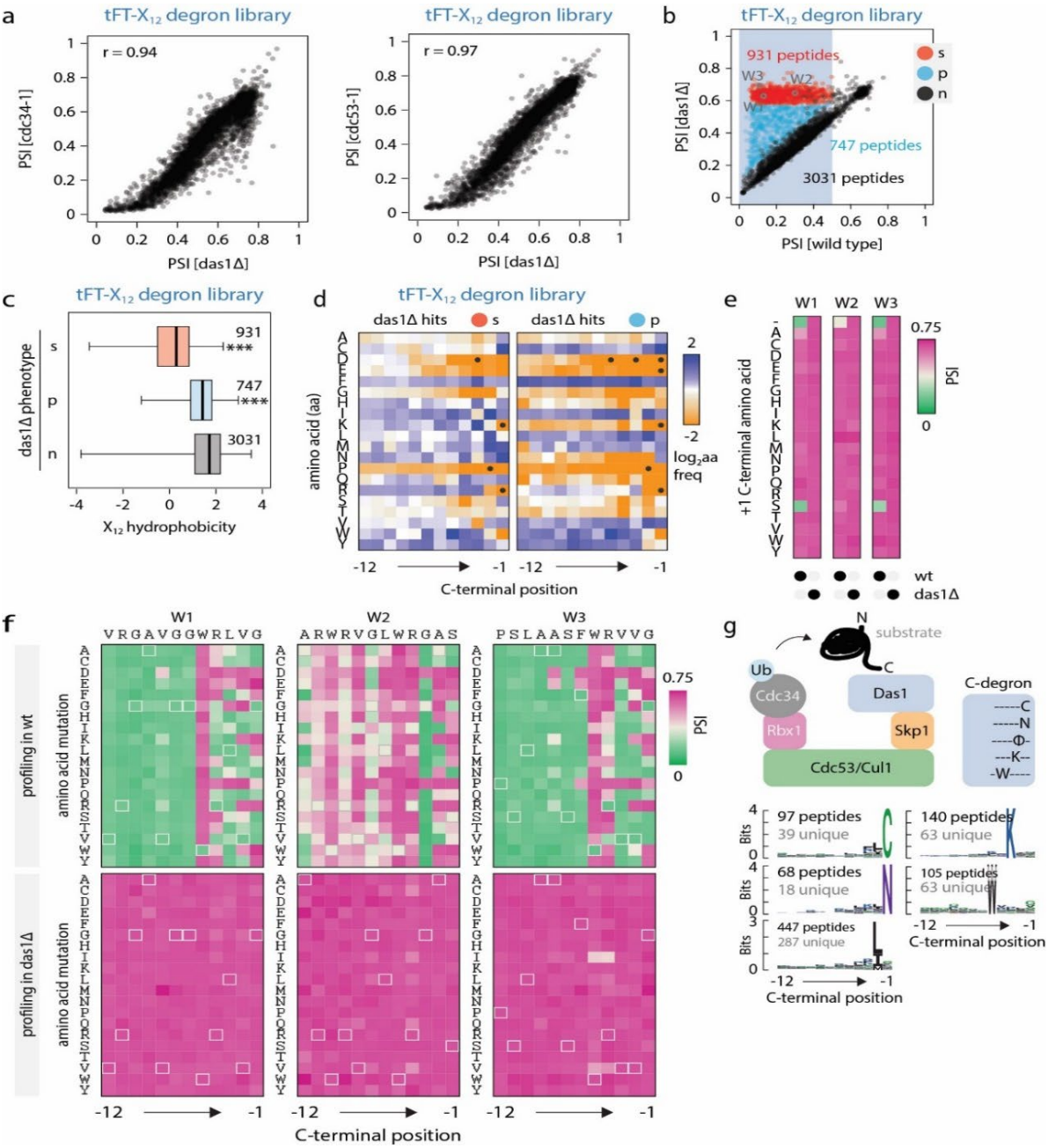

Figure 18: Das1 recognizes atleast five distinct but overlapping degron motifs*: (a) Variation of PSI from MPS profiling of tFT-X₁₂ C-degron in das1Δ versus cdc34-1(left) and das1Δ versus cdc53-1(right). R, spearman correlation coefficient. (b) Variation of PSI from MPS profiling of tFT-X₁₂ C-degron library for das1Δ versus WT. PSIs are linearly fit for each of the corresponding replicates. The blue region indicates the region of putative degron (PSI[WT] < 0.5). The region holding putative degrons is classified into 3 categories: s: stabilized by das1Δ, p: partially stabilized by das1Δ and n: having no effect. The number of peptides in each category is also labeled. (c) Variation of hydrophobicity for the categories mentioned in (b), ***p < 2.2 × 10−16 in a one-sided Mann-Whitney U-test (d) Variation of relative amino acid frequency in Das1-dependent constructs (s and p category in (b)), normalized to the relative frequency in the tFT-X₁₂ library. Amino acids absent per position in Das1-dependent constructs are marked with black circles (e) Variation of PSI after capping the C-terminus of representative peptides with W at -5. The representative peptides are*

marked in (b) in WT and das1Δ. (f) Saturation mutagenesis after mutating all positions of representative peptides with W at -5 in WT and das1Δ. (g) (Top) Model of substrate recognition by SCF^{Das1} via C-degrons (left) and potential C-degron motifs (right). (Bottom) sequence logos for the five motifs, based on the 931 Das1 degrons from b. Number of peptides forming and unique to each logo are indicated in black and grey, respectively

Overall, these experiments conclude that Das1 recognizes most degrons targeted by SCF. The Das1-degrons create several motifs, such as W at -5, K at -3 or ΦN, where Φ = [ILM], to name a few. Approximately 70% of the Das1-degrons fall under the category with W at -5, K at -3, [ILM] at -2 and/or [CN] at -1. Only 21 degrons simultaneously had three or more of these preferred amino acids, indicating distinct but overlapping motifs (Figure 18(g)). Together, these are termed as Das1 C-degron motifs.

5.4.7.3 Yeast Das1 degron shows specificity similar to tFT-X_{12} of random peptides

From earlier analysis, Das1 has broad specificity. Thus, to identify endogenous substrates of SCF^{Das1} in *S. cerevisiae*, MPS profiling on the yeast C-terminome of 6793 annotated yeast ORF was conducted in WT and *das1Δ* conditions. For this, 12 AA from the yeast C-terminome was fused to the tFT tag and tested for protein stability. Experiments were conducted in replicates of two, and both replicates showed a high correlation in WT and *das1Δ* (Figure 19(a)). In the WT conditions, 458 of these 6793 ORFs were unstable (PSI < 0.5). Testing for Das1-dependent degron reveals that 54% are targeted by Das1, in which 43% were completely stabilized (yDas1-degron) (Figure 19 (b)). These degrons were less hydrophobic than the ones not affected by das1Δ and had properties similar to Das1 degron in tFT-X_{12} library – 1. lower overall hydrophobicity compared to the other yeast C-degrons (Figure 19 (c)), 2. depletion of P, W and charged amino acids at position - 1 and 3. enrichment of W at -5, K at -3, I or L at -2 and C or N at -1 (Figure 19(f)).

Approximately 25% of Das1 degrons in the yeast C-terminome originated from dubious ORFs, which are unlikely to encode functional proteins (Figure 19(d)), and the remaining degrons corresponded to proteins with a variety of different subcellular localizations and functions (Figure 19(g)).

Considering Das1 is a cytosolic protein (Weill et al., 2018), it may play a role in the degradation of proteins mislocalized to the cytosol. To test this hypothesis, ORFs with yDas1-Degron were N-terminally tagged with mNeonGreen-mCherry timer (tFT) at endogenous levels or overexpressed and are tested for stability. The overexpression of these tFT-tagged proteins leads to the production of unnecessary or abnormal proteins. 18 of 74 overexpressed tFT-tagged ORFs stabilized in *das1Δ*, out of which 12 are subunits of complex formation (Kong, Shankar et al., 2023). To test the hypothesis that SCF^{Das1} is involved in the degradation of orphan protein complex subunits, we focused on two proteins from this set - the autophagy kinase Atg1 and the RNA polymerase I subunit Rpa12. Both these proteins interact with their binding partner via their C-terminal domains. The C-terminal domain of Atg1 and Rpa12 interacts with Atg13 and polymerase I active sites, respectively (Noda and Fujioka, 2015; Girbig et al., 2022). To test if the

C-terminal amino acid of these proteins is important for their degron function, MPS profiling of constructs with capped C-terminus of these proteins is conducted. The capping experiments on the C-termini of Atg1 (Atg-12_-1) and Rpa12 (Rpa12-12_-1) in isolation reveal that they are C-degron as they do not tolerate any other AA at the C-termini (Figure 20(a)). Additionally, they are degraded in a Das1-dependent manner.

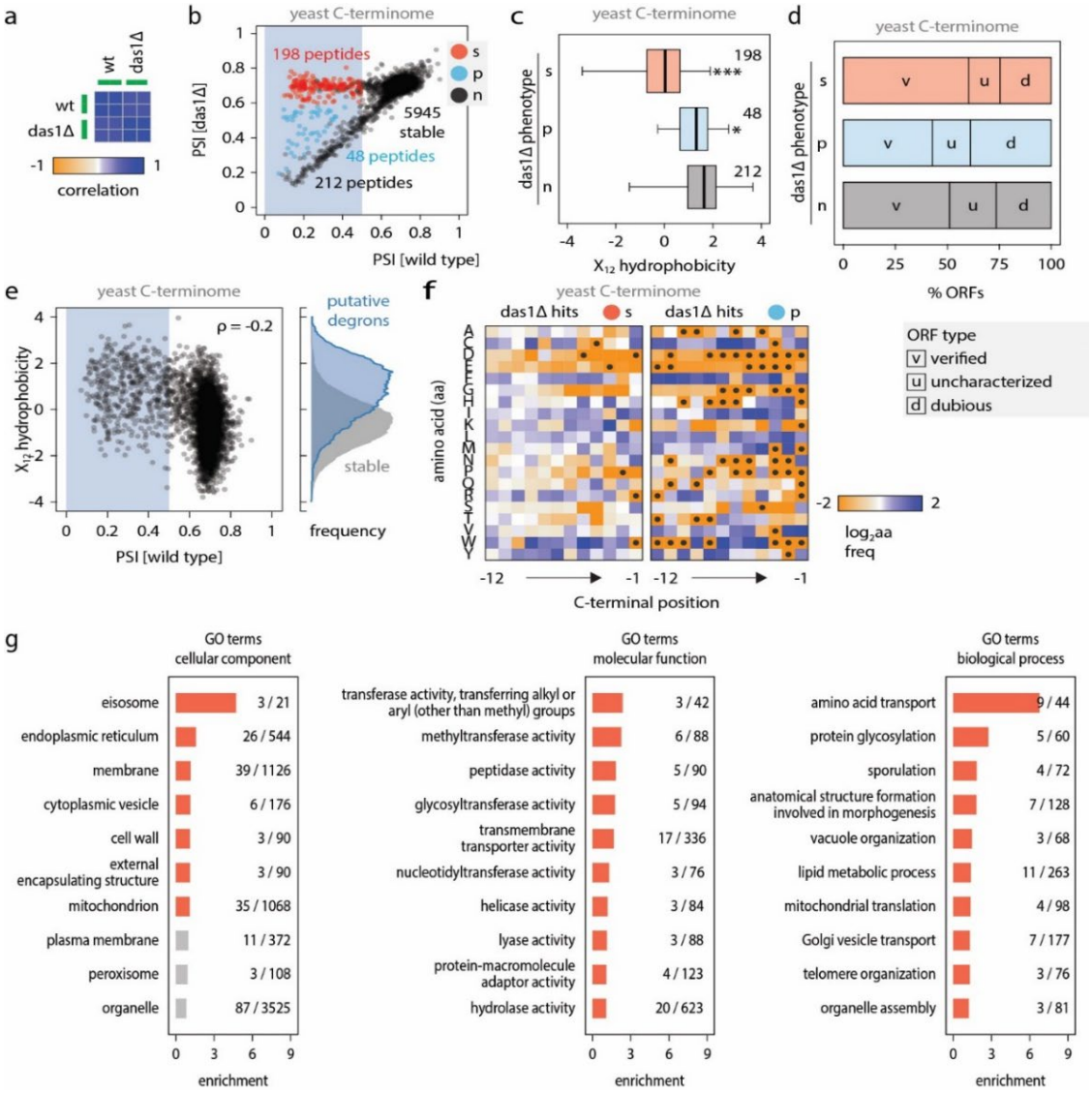

Figure 19: Yeast Das1 degrons show specificity similar to tFT-X$_{12}$. *(a) Heatmap of Pearson correlation coefficients between PSIs of replicates of yeast C-terminome library in WT and das1Δ backgrounds. (b) Variation of linearly fitted PSIs from two replicates of MPS on C-terminome library in wild type and das1Δ backgrounds, s: stabilized, p: partially stabilized and n: no effect. (c) Variation of hydrophobicity within unstable constructs (blue region), PSI [WT] < 0.5. Number of peptides is mentioned in the plot and the peptides are categorized as mentioned in (b). Centerlines mark the medians, box limits indicate the 25th and 75th percentiles, and whiskers extend to the minimum and maximum value in each group, *p = 0.012, ***p < 2.2 × 10−16 in a one-sided Mann-Whitney U-test. (d) Percentage of types of ORF in each category of SCF^{Das1} based unstable constructs (blue region in (b)) from yeast C-terminome MPS library. (e) Variation of hydrophobicity with PSI calculated from MPS of yeast C-terminome, blue region marks putative degrons. (Right) Hydrophobicity distribution of two classes of peptides – putative degrons (PSI < 0.5) in blue and stable constructs in grey. (f) Relative amino acid frequency in Das1-dependent constructs from (b), normalized to the relative frequency in the C-terminome library. Amino acids absent per position in Das1-dependent constructs are marked with*

black circles. (g) Gene Ontology analysis for SCF^{Das1} C-degron substrates (s – 198 peptides from (b)) from yeast C-terminome library. For each GO term, the number of ORFs in the set of Das1 C-degrons and the number of ORFs in the yeast genome are indicated. Only the top 10 GO terms by enrichment are shown: red – enrichment, grey – no enrichment in the set of Das1 C-degrons

Saturation mutagenesis on Atg1 and Rpa12 C-termini reveals broad specificity of Das1, with most of the degradation determining factor at the C-terminus (Figure 20(b)). Further experiments on individual full-length protein constructs using two mutants, capped C-terminus with L and mutated C-terminus with K, both revealed that the introduction of this mutation stabilized the protein to the level of *das1Δ.* This indicates that Das1 recoganized the C-terminus of Atg1 and Rpa12.

Overall, SCF^{Das1} plays a role in degrading various orphan protein complex subunits in the cytosol, of which Atg1 and Rpa12 are tested here. Using mutagenesis, we show that the proteins are recoganized by Das1 via C-terminus. The study also reveals that Atg1 and Rpa12 undergo Das1-dependent degradation only when overexpressed. This suggests that SCF^{Das1} recognizes the molecules when they fail to assemble in complexes. Thus, overexpressed Atg1 and Rpa12 degradation must be abolished if their binding partners are also overexpressed, as shown in (Kong, Shankar et al., 2023). Combining all these points, I conclude that SCF^{Das1} recognizes orphan complex subunits via C-degrons.

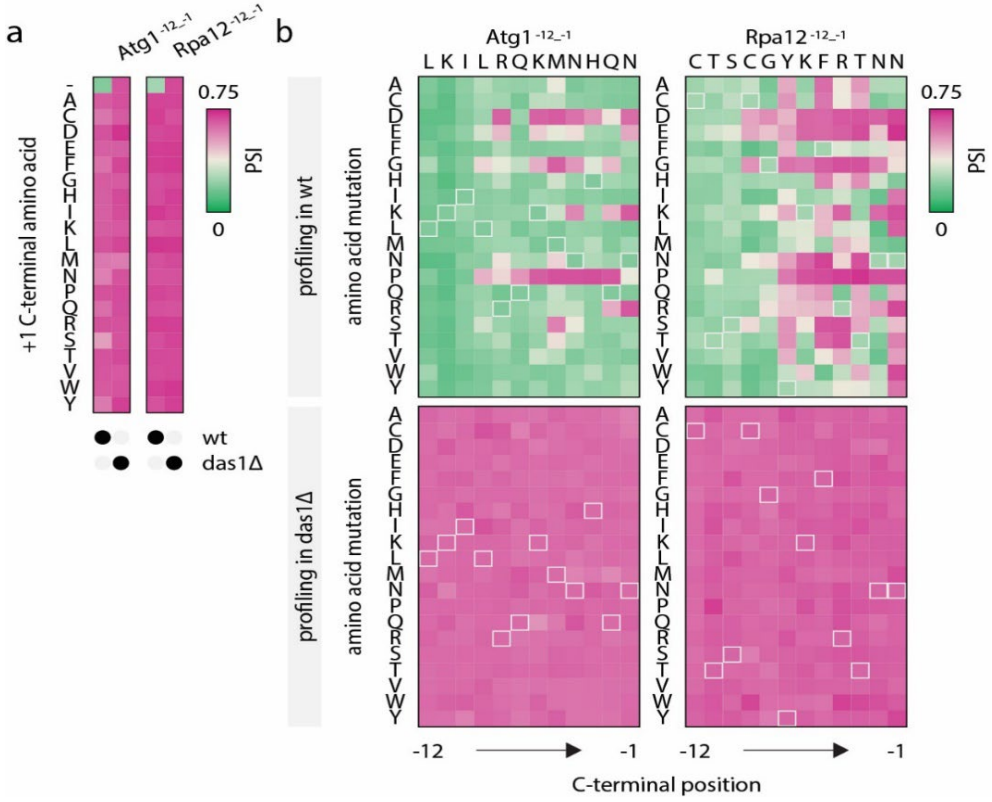

Figure 20: Saturation mutagenesis on Atg1 and Rpa12 C-terminome reveals SCF^{Das1} based C-terminal degradation*: (a) Heatmap displaying Psi variation after - terminal Capping of Atg1 and RPa12 C-terminus (b) Heatmap displaying saturation mutagenesis of 12 AA of C-terminus of Atg1 and Rpa12 in WT and Das1*

5.5 Discussion

Despite the importance of selective protein degradation in maintaining cellular homeostasis, the degrons, their role in various degradation pathways and the molecular factors that mediate the degradation still need to be better understood for various substrates. Recent advancements in sequencing paved the path to sophisticated techniques such as deep mutational scanning (DMS) that could ease and fasten the experiments required to decipher these degrons. Moreover, computational pipelines, varying from simple visualization to complex artificial intelligence, offer a supporting hand in accelerating the discovery of degron and the mechanism behind the degradation. In this work, I built a pipeline to analyze deep mutational scanning (DMS) experiments that help dissect the specificity of degrons. I used the Global Protein Stability (GPS) experiments dataset in the *H. Sapiens* cell line and the Multiplexed Protein Stability (MPS) experiment dataset for *S. cerevisiae*. The pipeline first performs the quality assessment of DMS experiments. Then, it performs downstream analysis to infer degron motifs, using simple visualization for shorter peptides and interpretable machine learning for longer peptides.

The quality assessment pipeline checks for the percentage of stop codons, deconvolution and technical noise. This pipeline is demonstrated using the GPS on the library of di-residue after ubiquitin moiety, tagged to eGFP-P2A-mCherry at the N-termini of the human cell line. The quality assessment pipeline shows a lower percentage of stop codons for the library, suggesting that the mCherry and eGFP threshold is optimally chosen during FACS sorting. Moreover, the pipeline also assesses proper deconvolution and low technical noise for this library. The assessment pipeline could be suggestive of both the technical noise, such as proper selection of fluorescence intensity threshold during FACS sorting, or can also point to possible biological nuances. For example, by utilizing MPS profiling to investigate N-terminal amino acid specificity in PhM8 proteins, our analysis revealed a notable increase in stop codons within the assessment pipeline. Subsequent investigation unveiled that the elevated proportion of stop codon reads stemmed from pronounced cellular toxicity induced by Phm8 overexpression.

After assessing the quality of the experiment, the next step is to infer degrons using downstream analysis. Downstream analysis to decipher degron motifs uses either simple visualization techniques for shorter constructs or interpretable deep learning for longer constructs. To systematically interpret the N-terminal amino acid specificity in a human cell line, I employed simple visualization techniques such as heatmaps on the stability profile from DMS of N-terminal di-residue constructs in the *H. Sapiens* cell line constructs (Ub-XZ-GPS). Analysis of the Ub-XZ-GPS shows the preference for the amino acids from the primary amino acid group of the Arg/N-degron pathway for degradation. I also see that the amino acids Q and E from the tertiary amino acid group and the secondary amino acid group of the Arg/N-degron pathway do not facilitate degradation. One hypothesis could be that the presence of these AA at the N terminus leads to

their acetylation, thereby preventing the action of Nt-deaminases and R transferases and, hence, degradation (Varland et al., 2023; Kats et al., 2018). Additionally, the analysis shows various trends in mammalian N-degron pathways that have not been previously established, such as the presence of V at the N-terminus facilitating degradation in mammalian cells. Additionally, the presence of D at the second position somehow stabilized the peptide irrespective of the N-terminus. The relevance of these trends in selective protein degradation is still to be experimentally validated.

Furthermore, to systematically study the eukaryotic C-degron, we surveyed C-degrons in an unbiased fashion in yeast. The library consisted of 12 AA long peptides tagged with tFT at their N-terminal (tFT-X_{12}). The library revealed that ~10% of the random peptides from the tFT-X_{12} library contain putative degrons. Analyzing the distribution of amino acid sequences in the degron library, we do not initially see any strong sequence specificity. Most of these degrons are hydrophobic, in line with already existing work (Ravid and Hochstrasser, 2008; Mashahreh et al., 2022; Johansson et al., 2023), which indicates that biophysical properties, in addition to sequence, would play a role in understanding degradation. Thus, I included known biophysical properties and peptide sequences to understand the collection of degrons. Six of the biophysical properties with correlation to the protein stability are used to create the model. I applied interpretable deep learning in this library to understand various features (amino acid sequence, biophysical property) that contribute to instability. A fully connected neural network is trained to predict if the protein is unstable and thereby contains any putative degron. The performance of the model on different datasets – four test sets, one validation set, and an independent replicate dataset indicates high and consistent confidence in the model prediction, with a high training accuracy of 0.85.

Although the model predicts well if the peptide of interest would be unstable, the “black-box” nature of the deep learning model fails to help in direct interpretation. To understand the factors that would have contributed to the degradation of peptides, SHAP is mounted onto the model that reveals the contribution of each factor in prediction. SHAP gives the contribution score (CS) for each factor that helped in correct prediction, including CS for each position of peptide, CS for positional biophysical properties and CS for each overall biophysical property. Although biophysical properties contribute to degradation, the peptide sequence contributes more towards instability in the tFT-X_{12} library. Variation of mean CS of sequence over positions also highlights the importance of C-terminal A, C and N. Hydrophobic residues (I, V, M) contribute to instability at all positions. Few known degrons, such as Rxx, are also noticeable (Lin et al. 2018). Among hydrophobic residues, it is interesting that L contributes to instability only if present at -2 or -4. Analysis of biophysical properties reveals that hydrophobicity contributes more to determining instability. Other biophysical properties, such as aliphatic index and potential PI interaction, also affect the degradation of peptides.

To define C-degron motifs, clustering of the contribution score for the correctly predicted 4110 peptides helped define 21 C-degron clusters. These motifs show a clear preference for sequence information. Using the motifs from the 21 C-degron clusters, new unstable peptides could be designed based on the similarity of the peptides in the cluster. For example, from clusters c1, c2, and c31, a random peptide with high similarity to peptides in all three clusters, RWCFSGRIVFVC, is predicted to be unstable with high confidence. Artificially simulated unstable peptides can be computationally analyzed for functional importance in the cell and then be further synthesized and validated in the lab for studying various pathways in protein quality control (Tokheim et al., 2021).

The primary model mentioned above helps to understand the simultaneous contribution of amino acid features and biophysical features to protein instability. To test the importance of each feature separately, models with individual properties, either using only amino acid sequence, only positional biophysical properties or only global biophysical properties, are created. These additional models contain 23 biophysical properties. The individual models are then ensembled. Models solely focused on either amino acid sequence or positional biophysical properties demonstrate better performance than composite models incorporating all properties. Although ensemble models perform similarly to individual ones, suggesting redundancy in combining sequences with biophysical properties, they are still utilized for further analysis due to their facilitation of interpreting the effect of biophysical properties on stability.

The current model captures the degrons in the random tFT-X_{12} library. The model uses fixed-length peptide sequences for prediction. Using additional GPS datasets of multiple lengths along with the tFT-X_{12}, such as (Mashahreh et al., 2023; Kats et al., 2018), annotated with the organism information, could enrich the input dataset. Furthermore, additional post-translational modifications along with biophysical properties could be added to evaluate how PTMs lead to peptide instability. More sophisticated models with enriched datasets, such as ProteinBERT (Brandes et al. 2022) and TAPE (Rao et al. 2019), could help better predict the unstable peptides. These modes are language-based models and are openly available for usage. Apart from using these models to infer degrons, the models can combine the degrons with known E3 ligases to search for interacting interfaces and evaluate the proteome for additional substrates (Shu et al., 2023).

Further looking deeper into patterns in the tFT-X_{12} degron library, the depletion of glycine at the C-terminus is observed. SHAP contribution to the sequence-based prediction model indicates the importance of glycine at the C-terminal, which is also shown in the Gly/C-degron pathway, in which the Cul2 targets the C-degron glycine for degradation. Using saturation mutagenesis, we found that when asparagine in ΦN degrons is mutated to glycine, it does not affect the degradation. However, glycine is depleted from the C-terminome of eukaryotes and yeast (Yeh et al., 2021; Koren et al., 2018), which could be due to the evolutionary shaping of the C-terminome

by continuous recognition of such motifs. This might indicate the role other ubiquitin-like proteins have in the degradation of such peptides.

Furthermore, combining machine learning, mutagenesis, and genetic screens, results reveal that a single F-box substrate receptor of SCF ubiquitin ligase, Das1, targets ~40% of degrons in tFT-X_{12}. Systematic analysis of Das1-dependent C-degrons reveals that the peptides are comparatively less hydrophobic. Saturation mutagenesis of representative Das1 degron indicated certain amino acids strongly disfavored at the C-terminus, including proline at positions -5 to -1. An independent study on recognition and degradation of C-terminal Hac1u intron sequence, generated as a result of accidental translation of unspliced HAC1 mRNA, reveals that mutation of leucine at -2 to proline or tyrosine at -5 to proline leads to degradation by SCF^{Das1} (Di Santo et al., 2016), in line with our analysis. In addition to the disfavored AA, we find AA preferred for SCF^{Das1}-based degradation. This includes 5 C-degron motifs – W at -5, K at -3, (I, L or M) at -2 and (C/N) at -1. These combinations rarely occur in the tFT-X_{12} library or the yeast C-terminome, indicating potentially overlapping C-degron motifs.

Furthermore, a study to elucidate the endogenous substrates of SCF^{Das1} in *S. cerevisiae*, MPS profiling on the yeast C-terminome of 6793 annotated yeast ORF was conducted in WT, and *das1Δ* conditions were conducted. 54% of degrons from yeast C-terminome are recognized for degradation by a single ubiquitin ligase, SCF, using the F-box substrate receptor Das1. Furthermore, Gene Ontology analysis on yeast C-degron reveals that Das1 substrates are involved in various functions and processes, including autophagy, peroxisome biogenesis, transcription, translation, redox metabolism and regulation of gene expression. The role of SCF^{Das1} in the degradation of orphan subunits of protein complexes is studied using two proteins – Atg1 and Rpa12. Using mutagenesis, the study shows that Das1 recognizes the proteins via the C-terminus, which is in line with the study (Hasenjäger et al., 2023). The study also reveals that Atg1 and Rpa12 undergo Das1-dependent degradation only when overexpressed. This suggests that SCF^{Das1} recognizes the molecules when they fail to assemble in complexes. Thus, overexpressed Atg1 and Rpa12 degradation must be abolished if their binding partners are also overexpressed, as shown in (Kong, Shankar et al., 2023). Combining all these points, I conclude that SCF^{Das1} recognizes orphan complex subunits via C-degrons.

5.6 Conclusion

Sophisticated computer algorithms, along with experimental methods, could help accelerate the degron discovery and factors affecting degradation machinery. In this work, I show how simple visualization techniques and complex deep learning models could be used to decipher the degrons using MPS profiling. Interpretability could further enhance our chances of de novo degron discovery, which could be experimentally validated. The survey of C-degron in yeast random library reveals ~10% of peptides to encode degron, out of which 40% are degraded by SCF^{Das1}. Using the example of the C-degron tFT-X_{12} library, we find that though sequence still plays a pivotal role in the degradation of proteins, the analysis indicates the importance of biophysical properties such as aliphatic index or hydrogen bonding towards understanding degradation, other than hydrophobicity. Thus, further study into the biophysical property and structural phenotype could help in a clearer understanding of degradation mechanics. The pipeline and the downstream analysis could be taken as a first step towards a more refined prediction model, encompassing the deciphering of degron motifs, potential E3 ligases and chaperons and the mechanism of degradation. More properties, such as potential energy requirement and known PTMs, could also be added to elucidate a complete picture. Moreover, sophisticated models such as encoders and transformers with interpretability could also help in enhancing our understanding of degradation machinery.

5.7 Code availability

All codes for the processing pipeline and downstream analysis are in
https://github.com/Susmitha-Nair/SLiMsInDegradation.git

5.8 References

Aksnes, H., A. Drazic, M. Marie, and T. Arnesen. 2016. First things first: Vital protein marks by N-terminal acetyltransferases. *Trends Biochem. Sci.* 41:746–760.

Aliabadi, F., B. Sohrabi, E. Mostafavi, H. Pazoki-Toroudi, and T.J. Webster. 2021. Ubiquitin–proteasome system and the role of its inhibitors in cancer therapy. *Open Biol.* 11. doi:10.1098/rsob.200390.

Bachmair, A., D. Finley, and A. Varshavsky. 1986. In vivo half-life of a protein is a function of its amino-terminal residue. *Science*. 234:179–186.

Bachmair, A., and A. Varshavsky. 1989. The degradation signal in a short-lived protein. *Cell*. 56:1019–1032.

Balakrishnan, R., J. Park, K. Karra, B.C. Hitz, G. Binkley, E.L. Hong, J. Sullivan, G. Micklem, and J. Michael Cherry. 2012. YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford)*. 2012. doi:10.1093/database/bar062.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57:289–300.

Bielskienė, K., L. Bagdonienė, J. Mozūraitienė, B. Kazbarienė, and E. Janulionis. 2015. E3 ubiquitin ligases as drug targets and prognostic biomarkers in melanoma. *Medicina (Kaunas)*. 51:1–9.

Boman, H.G. 2003. Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.* 254:197–215.

Brandes, N., D. Ofer, Y. Peleg, N. Rappoport, and M. Linial. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 38:2102–2110.

Bulatov, E., and A. Ciulli. 2015. Targeting Cullin-RING E3 ubiquitin ligases for drug discovery: structure, assembly and small-molecule modulation. *Biochem. J.* 467:365–386.

Carrano, A.C., Z. Liu, A. Dillin, and T. Hunter. 2009. A conserved ubiquitination pathway determines longevity in response to diet restriction. *Nature*. 460:396–399.

Chen, S.-J., X. Wu, B. Wadas, J.-H. Oh, and A. Varshavsky. 2017. An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes. *Science*. 355:eaal3655.

Di Santo, R., S. Aboulhouda, and D.E. Weinberg. 2016. The fail-safe mechanism of post-transcriptional silencing of unspliced HAC1 mRNA. *Elife*. 5. doi:10.7554/elife.20069.

Ella, H., Y. Reiss, and T. Ravid. 2019. The hunt for degrons of the 26S proteasome. *Biomolecules*. 9:230.

Finley, D., H.D. Ulrich, T. Sommer, and P. Kaiser. 2012. The ubiquitin-proteasome system of Saccharomyces cerevisiae. *Genetics*. 192:319–360.

Francis, O., F. Han, and J.C. Adams. 2013. Molecular phylogeny of a RING E3 ubiquitin ligase, conserved in eukaryotic cells and dominated by homologous components, the muskelin/RanBPM/CTLH complex. *PLoS One*. 8:e75217.

Gilon, T., O. Chomsky, and R.G. Kulka. 1998. Degradation signals for ubiquitin system proteolysis in Saccharomyces cerevisiae. *EMBO J.* 17:2759–2766.

Girbig, M., A.D. Misiaszek, and C.W. Müller. 2022. Structural insights into nuclear transcription by eukaryotic DNA-dependent RNA polymerases. *Nat. Rev. Mol. Cell Biol.* 23:603–622.

Ham, S.J., D. Lee, W.J. Xu, E. Cho, S. Choi, S. Min, S. Park, and J. Chung. 2021. Loss of UCHL1 rescues the defects related to Parkinson's disease by suppressing glycolysis. *Sci. Adv.* 7:eabg4574.

Hasenjäger, S., A. Bologna, L.-O. Essen, R. Spadaccini, and C. Taxis. 2023. C-terminal sequence stability profiling in Saccharomyces cerevisiae reveals protective protein quality control pathways. *J. Biol. Chem.* 299:105166.

Hegde, A.N., and S.C. Upadhya. 2007. The ubiquitin-proteasome pathway in health and disease of the nervous system. *Trends Neurosci.* 30:587–595.

Hendriks, I.A., R.C.J. D'Souza, B. Yang, M. Verlaan-de Vries, M. Mann, and A.C.O. Vertegaal. 2014. Uncovering global SUMOylation signaling networks in a site-specific manner. *Nat. Struct. Mol. Biol.* 21:927–936.

Hershko, A., and A. Ciechanover. 1998. The ubiquitin system. *Annu. Rev. Biochem.* 67:425–479.

Hughes, D.C., L.M. Baehr, D.S. Waddell, A.P. Sharples, and S.C. Bodine. 2022. Ubiquitin ligases in longevity and aging skeletal muscle. *Int. J. Mol. Sci.* 23:7602.

Hwang, C.-S., A. Shemorry, and A. Varshavsky. 2010. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*. 327:973–977.

Ikai, A. 1980. Thermostability and aliphatic index of globular proteins. *J. Biochem.* 88:1895–1898.

Johansson, K.E., B. Mashahreh, R. Hartmann-Petersen, T. Ravid, and K. Lindorff-Larsen. 2023. Prediction of quality-control degradation signals in yeast proteins. *J. Mol. Biol.* 435:167915.

Kats, I., A. Khmelinskii, M. Kschonsak, F. Huber, R.A. Knieß, A. Bartosik, and M. Knop. 2018. Mapping degradation signals and pathways in a eukaryotic N-terminome. *Mol. Cell*. 70:488-501.e5.

Keiler, K.C., P.R. Waller, and R.T. Sauer. 1996. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*. 271:990–993.

Kishino, T., M. Lalande, and J. Wagstaff. 1997. UBE3A/E6-AP mutations cause Angelman syndrome. *Nat. Genet.* 15:70–73.

Komander, D., and M. Rape. 2012. The ubiquitin code. *Annu. Rev. Biochem.* 81:203–229.

Kong, K.-Y.E., B. Fischer, M. Meurer, I. Kats, Z. Li, F. Rühle, J.D. Barry, D. Kirrmaier, V. Chevyreva, B.-J. San Luis, M. Costanzo, W. Huber, B.J. Andrews, C. Boone, M. Knop, and A. Khmelinskii. 2021. Timer-based proteomic profiling of the ubiquitin-proteasome system reveals a substrate receptor of the GID ubiquitin ligase. *Mol. Cell*. 81:2460-2476.e11.

Kong, K.-Y.E., S. Shankar, F. Rühle, and A. Khmelinskii. 2023. Orphan quality control by an SCF ubiquitin ligase directed to pervasive C-degrons. *Nat. Commun.* 14:8363.

Koren, I., R.T. Timms, T. Kula, Q. Xu, M.Z. Li, and S.J. Elledge. 2018. The eukaryotic proteome is shaped by E3 ubiquitin ligases targeting C-terminal degrons. *Cell*. 173:1622-1635.e14.

Krohn, R.I. 2011. The colorimetric detection and quantitation of total protein. *Curr. Protoc. Cell Biol.* Appendix 3:3H.

Krshnan, L., M.L. van de Weijer, and P. Carvalho. 2022. Endoplasmic reticulum-associated protein degradation. *Cold Spring Harb. Perspect. Biol.* 14:a041247.

Kus, B.M., C.E. Caldon, R. Andorn-Broza, and A.M. Edwards. 2004. Functional interaction of 13 yeast SCF complexes with a set of yeast E2 enzymes in vitro. *Proteins*. 54:455–467.

Kwon, Y.T., and A. Ciechanover. 2017. The Ubiquitin Code in the Ubiquitin-Proteasome System and Autophagy. *Trends Biochem. Sci.* 42:873–886.

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.

Leboeuf, D., M. Pyatkov, T.S. Zatsepin, and K. Piatkov. 2020. The Arg/N-degron pathway-A potential running back in fine-tuning the inflammatory response? *Biomolecules*. 10:903.

Liang, G., G. Chen, W. Niu, and Z. Li. 2008. Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. *Chem. Biol. Drug Des.* 71:345–351.

Lin, H.-C., C.-W. Yeh, Y.-F. Chen, T.-T. Lee, P.-Y. Hsieh, D.V. Rusnac, S.-Y. Lin, S.J. Elledge, N. Zheng, and H.-C.S. Yen. 2018. C-terminal end-directed protein elimination by CRL2 ubiquitin ligases. *Mol. Cell*. 70:602-613.e3.

Liu, Y. 2020. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal.* 18:145.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*. 28:129–137.

Löw, P. 2011. The role of ubiquitin-proteasome system in ageing. *Gen. Comp. Endocrinol.* 172:39–43.

Lundberg, S., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *arXiv [cs.AI]*.

Martínez-Jiménez, F., F. Muiños, E. López-Arribillaga, N. Lopez-Bigas, and A. Gonzalez-Perez. 2020. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer*. 1:122–135.

Mashahreh, B., S. Armony, K.E. Johansson, A. Chappleboim, N. Friedman, R.G. Gardner, R. Hartmann-Petersen, K. Lindorff-Larsen, and T. Ravid. 2022. Conserved degronome features governing quality control associated proteolysis. *Nat. Commun.* 13:7588.

Mashahreh, B., S. Armony, and T. Ravid. 2023. YGPS-P: A yeast-based peptidome screen for studying quality control-associated proteolysis. *Biomolecules*. 13. doi:10.3390/biom13060987.

Matsuura, T., J.S. Sutcliffe, P. Fang, R.-J. Galjaard, Y.-H. Jiang, C.S. Benton, J.M. Rommens, and A.L. Beaudet. 1997. De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat. Genet.* 15:74–77.

Maurer, M.J., E.D. Spear, A.T. Yu, E.J. Lee, S. Shahzad, and S. Michaelis. 2016. Degradation signals for ubiquitin-proteasome dependent cytosolic protein quality control (CytoQC) in yeast. *G3 (Bethesda)*. 6:1853–1866.

Melnykov, A., S.-J. Chen, and A. Varshavsky. 2019. Gid10 as an alternative N-recognin of the Pro/N-degron pathway. *Proc. Natl. Acad. Sci. U. S. A.* 116:15914–15923.

Menssen, R., K. Bui, and D.H. Wolf. 2018. Regulation of the Gid ubiquitin ligase recognition subunit Gid4. *FEBS Lett.* 592:3286–3294.

Mészáros, B., G. Erdős, and Z. Dosztányi. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46:W329–W337.

Molnar, C. 2018. iml: An R package for Interpretable Machine Learning. *J. Open Source Softw.* 3:786.

Nakagawa, T., and K. Nakayama. 2015. Protein monoubiquitylation: targets and diverse functions. *Genes Cells*. 20:543–562.

Nijman, S.M.B., M.P.A. Luna-Vargas, A. Velds, T.R. Brummelkamp, A.M.G. Dirac, T.K. Sixma, and R. Bernards. 2005. A genomic and functional inventory of deubiquitinating enzymes. *Cell*. 123:773–786.

Noda, N.N., and Y. Fujioka. 2015. Atg1 family kinases in autophagy initiation. *Cell. Mol. Life Sci.* 72:3083–3096.

Okumura, F., Y. Fujiki, N. Oki, K. Osaki, A. Nishikimi, Y. Fukui, K. Nakatsukasa, and T. Kamura. 2020. Cul5-type ubiquitin ligase KLHDC1 contributes to the elimination of truncated SELENOS produced by failed UGA/Sec decoding. *iScience*. 23:100970.

Osorio, D., P. Rondón-Villarreal, and R. Torres. 2015. Peptides: A package for data mining of antimicrobial peptides. *R J.* 7:4.

Park, J., J. Cho, and E.J. Song. 2020. Ubiquitin–proteasome system (UPS) as a target for anticancer treatment. *Arch. Pharm. Res.* 43:1144–1161.

Park, S.-E., J.-M. Kim, O.-H. Seok, H. Cho, B. Wadas, S.-Y. Kim, A. Varshavsky, and C.-S. Hwang. 2015. Control of mammalian G protein signaling by N-terminal acetylation and the N-end rule pathway. *Science*. 347:1249–1252.

do Patrocinio, A.B., V. Rodrigues, and L. Guidi Magalhães. 2022. P53: Stability from the ubiquitin-proteasome system and specific 26S proteasome inhibitors. *ACS Omega*. 7:3836–3843.

Ravalin, M., P. Theofilas, K. Basu, K.A. Opoku-Nsiah, V.A. Assimon, D. Medina-Cleghorn, Y.-F. Chen, M.F. Bohn, M. Arkin, L.T. Grinberg, C.S. Craik, and J.E. Gestwicki. 2019. Specificity for latent C termini links the E3 ubiquitin ligase CHIP to caspases. *Nat. Chem. Biol.* 15:786–794.

Ravid, T., and M. Hochstrasser. 2008. Diversity of degradation signals in the ubiquitin-proteasome system. *Nat. Rev. Mol. Cell Biol.* 9:679–690.

Ravid, T., S.G. Kreft, and M. Hochstrasser. 2006. Membrane and soluble substrates of the Doa10 ubiquitin ligase are degraded by distinct pathways. *EMBO J.* 25:533–543.

Reed, S.I. 2006. The ubiquitin-proteasome pathway in cell cycle control. *Results Probl. Cell Differ.* 42:147–181.

Regelmann, J., T. Schüle, F.S. Josupeit, J. Horak, M. Rose, K.-D. Entian, M. Thumm, and D.H. Wolf. 2003. Catabolite degradation of fructose-1,6-bisphosphatase in the yeast Saccharomyces cerevisiae: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol. Biol. Cell*. 14:1652–1663.

Ritchie, M.E., B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, and G.K. Smyth. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47.

Rodriguez, S., C. Abundis, F. Boccalatte, P. Mehrotra, M.Y. Chiang, M.A. Yui, L. Wang, H. Zhang, A. Zollman, R. Bonfim-Silva, A. Kloetgen, J. Palmer, G. Sandusky, M. Wunderlich, M.H. Kaplan, J.C. Mulloy, G. Marcucci, I. Aifantis, A.A. Cardoso, and N. Carlesso. 2020. Therapeutic targeting of the E3 ubiquitin ligase SKP2 in T-ALL. *Leukemia*. 34:1241–1252.

Rusnac, D.-V., H.-C. Lin, D. Canzani, K.X. Tien, T.R. Hinds, A.F. Tsue, M.F. Bush, H.-C.S. Yen, and N. Zheng. 2018. Recognition of the diglycine C-end degron by CRL2KLHDC2 ubiquitin ligase. *Mol. Cell*. 72:813-822.e4.

Santt, O., T. Pfirrmann, B. Braun, J. Juretschke, P. Kimmig, H. Scheel, K. Hofmann, M. Thumm, and D.H. Wolf. 2008. The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism. *Mol. Biol. Cell*. 19:3323–3333.

Schwartz, A.L., and A. Ciechanover. 1999. The ubiquitin-proteasome pathway and pathogenesis of human diseases. *Annu. Rev. Med.* 50:57–74.

Seol, J.H., A. Shevchenko, A. Shevchenko, and R.J. Deshaies. 2001. Skp1 forms multiple protein complexes, including RAVE, a regulator of V-ATPase assembly. *Nat. Cell Biol.* 3:384–391.

Shu, Y., Y. Hai, L. Cao, and J. Wu. 2023. Deep-learning based approach to identify substrates of human E3 ubiquitin ligases and deubiquitinases. *Comput. Struct. Biotechnol. J.* 21:1014–1021.

Smith, T.F., and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–197.

Sriramachandran, A.M., and R.J. Dohmen. 2014. SUMO-targeted ubiquitin ligases. *Biochim. Biophys. Acta Mol. Cell Res.* 1843:75–85.

Swanson, R., M. Locher, and M. Hochstrasser. 2001. A conserved ubiquitin ligase of the nuclear envelope/endoplasmic reticulum that functions in both ER-associated and Matalpha2 repressor degradation. *Genes Dev.* 15:2660–2674.

Timms, R.T., and I. Koren. 2020. Tying up loose ends: the N-degron and C-degron pathways of protein degradation. *Biochem. Soc. Trans.* 48:1557–1567.

Timms, R.T., Z. Zhang, D.Y. Rhee, J.W. Harper, I. Koren, and S.J. Elledge. 2019. A glycine-specific N-degron pathway mediates the quality control of protein N-myristoylation. *Science*. 365:eaaw4912.

Tokheim, C., X. Wang, R.T. Timms, B. Zhang, E.L. Mena, B. Wang, C. Chen, J. Ge, J. Chu, W. Zhang, S.J. Elledge, M. Brown, and X.S. Liu. 2021. Systematic characterization of mutations altering protein degradation in human cancers. *Mol. Cell*. 81:1292-1308.e11.

Tsurumi, C., N. Ishida, T. Tamura, A. Kakizuka, E. Nishida, E. Okumura, T. Kishimoto, M. Inagaki, K. Okazaki, and N. Sagata. 1995. Degradation of c-Fos by the 26S proteasome is accelerated by c-Jun and multiple protein kinases. *Mol. Cell. Biol.* 15:5682–5687.

UniProt Consortium. 2023. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res.* 51:D523–D531.

Varland, S., R.D. Silva, I. Kjosås, A. Faustino, A. Bogaert, M. Billmann, H. Boukhatmi, B. Kellen, M. Costanzo, A. Drazic, C. Osberg, K. Chan, X. Zhang, A.H.Y. Tong, S. Andreazza, J.J. Lee, L. Nedyalkova, M. Ušaj, A.J. Whitworth, B.J. Andrews, J. Moffat, C.L. Myers, K. Gevaert, C. Boone, R.G. Martinho, and T. Arnesen. 2023. N-terminal acetylation shields proteins from degradation and promotes age-dependent motility and longevity. *Nat. Commun.* 14:6774.

Varshavsky, A. 1991. Naming a targeting signal. *Cell*. 64:13–15.

Varshavsky, A. 2019. N-degron and C-degron pathways of protein degradation. *Proc. Natl. Acad. Sci. U. S. A.* 116:358–366.

Voutsadakis, I.A. 2012. The ubiquitin-proteasome system and signal transduction pathways regulating Epithelial Mesenchymal transition of cancer. *J. Biomed. Sci.* 19:67.

Wang, Z., W.-G. Zhu, and X. Xu. 2017. Ubiquitin-like modifications in the DNA damage response. *Mutat. Res.* 803–805:56–75.

Weill, U., I. Yofe, E. Sass, B. Stynen, D. Davidi, J. Natarajan, R. Ben-Menachem, Z. Avihou, O. Goldman, N. Harpaz, S. Chuartzman, K. Kniazev, B. Knoblach, J. Laborenz, F. Boos, J. Kowarzyk, S. Ben-Dor, E. Zalckvar, J.M. Herrmann, R.A. Rachubinski, O. Pines, D. Rapaport, S.W. Michnick, E.D. Levy, and M. Schuldiner. 2018. Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods*. 15:617–622.

Wertz, I.E., and X. Wang. 2019. From discovery to bedside: Targeting the ubiquitin system. *Cell Chem. Biol.* 26:156–177.

Yan, X., Y. Li, G. Wang, Z. Zhou, G. Song, Q. Feng, Y. Zhao, W. Mi, Z. Ma, and C. Dong. 2021. Molecular basis for recognition of Gly/N-degrons by CRL2ZYG11B and CRL2ZER1. *Mol. Cell*. 81:3262-3274.e3.

Yau, R., and M. Rape. 2016. The increasing complexity of the ubiquitin code. *Nat. Cell Biol.* 18:579–586.

Yeh, C.-W., W.-C. Huang, P.-H. Hsu, K.-H. Yeh, L.-C. Wang, P.W.-C. Hsu, H.-C. Lin, Y.-N. Chen, S.-C. Chen, C.-H. Yeang, and H.-C.S. Yen. 2021. The C-degron pathway eliminates mislocalized proteins and products of deubiquitinating enzymes. *EMBO J.* 40:e105846.

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

Zhang, X., S. Linder, and M. Bazzaro. 2020. Drug development targeting the ubiquitin-proteasome system (UPS) for the treatment of human cancers. *Cancers (Basel)*. 12:902.

Zheng, Q., T. Huang, L. Zhang, Y. Zhou, H. Luo, H. Xu, and X. Wang. 2016. Dysregulation of ubiquitin-proteasome system in neurodegenerative diseases. *Front. Aging Neurosci.* 8:303.

Zou, T., and Z. Lin. 2021. The involvement of ubiquitination machinery in cell cycle regulation and cancer progression. *Int. J. Mol. Sci.* 22:5754.

6 Dissecting Specificity of Short Linear Motifs in protein localization

6.1 Introduction

6.1.1 Targeting proteins for correct localization

Most proteins are synthesized in the cytosol by the ribosomes. Nearly two-thirds of these proteins are then transported to various sub-cellular compartments, where they perform their specialized functions (Juszkiewicz and Hegde, 2018; Hegde and Zavodszky, 2019; Fu et al., 2018). For certain proteins, Short Linear Motifs (SLiMs) can act as signals that target them to specific subcellular compartments or structures within the cell. These SLiMs that transport the proteins to their destined compartments are in the form of a localization signal. These motifs often contain specific amino acid sequences, biophysical properties or a site of post-translational modification (PTM), recognized by various cellular machineries responsible for protein trafficking, sorting, and localization, thereby regulating diverse cellular processes, including signal transduction, gene expression, and metabolism. Understanding the sequence features and recognition mechanisms of these motifs provides valuable insights into the mechanisms underlying protein localization and cellular organization. In this chapter, I use computational methods to dissect the SLiMs responsible for protein localization.

A fundamental step in the correct targeting of proteins to their sub-cellular location is to recognize the targeting signal in the nascent protein. The fidelity of the protein organization depends on the correctness of the targeting signals and the specificity with which the cognate receptors of these signals target them to correct localization. Mutations in the targeting signals lead to protein mislocalization (Pidasheva et al., 2005). Additionally, due to the similarities in the targeting signals for different compartments, mislocalization for a proportion of nascent proteins is inevitable even under optimal conditions (Hegde and Zavodszky, 2019). Proteins mislocalized in the wrong compartments are prone to inappropriate interactions, misfunction or aggregation, thereby playing a role in various diseases (Guo et al., 2014). Thus, it is important to understand the mechanics of correct localization by understanding the nature of the localization signals.

Most localization signals are degenerate sequence motifs, with more specificity towards biophysical properties than amino acid sequences (Hegde and Zavodszky, 2019). As a result, the discovery of a targeting signal becomes much more complex as the binding site of the targeting factors that mediate the localization must be capable of recognizing diverse sequences.

6.1.1.1 Localization signals for various compartments

Proteins synthesized in the cytosol are targeted to the destined compartments to fulfill their function through their respective targeting mechanisms. These proteins are recognized by the recognition factors via targeting signals. Efficient targeting requires correct targeting signals, the absence of which affects cellular homeostasis. This section describes the commonly found protein localization signals in various compartments.

Mitochondrial proteins often contain presequences that are mostly cleavable once the protein is translocated correctly to the compartment. These mitochondrial targeting signals are stretches of up to 50 AA, predominantly at the protein termini. They have alternating positive AA and hydrophobic AA, forming positively charged amphipathic α helices, with the groups of AA facing opposite in the helix. The hydrophobic surface is recognized by the Translocase of Outer Membrane 20 (Tom20), and Tom22 recognizes the positively charged surface. These presequences are cleaved off by the Mitochondrial Processing Peptidases (MPP) once the protein translocates to mitochondria (Chacinska et al., 2009). There are also other classes of mitochondrial signals that are not cleaved after translation. For example, proteins destined to be targeted to the outer mitochondrial membrane have signal clusters at protein termini and internal region, typically consisting of an α-helical transmembrane segment flanked by positively charged residues. The signals do not follow a unique insertion pathway into the outer membrane (Chacinska et al., 2009). Some groups of mitochondrial signals are indicated in Figure 23 (a)).

Secretory and soluble proteins targeted to the Endoplasmic Reticulum (ER) have a 12-30 AA long targeting signal that does not have amino acid sequence specificity but has common biophysical properties. The targeting sequence consists of a positive N-terminus followed by a hydrophobic core and polar C-terminus (rich in G and P). These signals are recognized by the Signal Recognition Particle (SRP) during translation (Pool, 2022). Another predominantly studied ER targeting signal includes the "HDEL" sequence (or "KDEL" sequence in mammals) at the protein C-terminus. This sequence helps in retrieving the proteins from the downstream compartments in the secretory pathways, mediated by p26 KDEL and COPI coatomer structure (Stornaiuolo et al., 2003).

The most studied localization signal is the nuclear localization signal (NLS). The first NLS was "PKKKRKV," found in the mutant of large T-antigen of simian virus 40 (SV40), which was later discovered to be part of other nuclear proteins as well (Adam et al., 1989). In the review by (Lu et al., 2021), NLS is categorized as classical localization signals (cNLS) and non-classical localization signals (ncNLS) based on their residue composition. The cNLS are enriched in R and K. Two most common cNLS motifs are K(K/R)X(K/R) (Bradley et al., 2007) or R/K$(X)_{10\text{-}12}$KRXK (Nguyen Ba et al., 2009), where X can be any amino acid. The ncNLS are not enriched in K or R as the cNLS. The most common ncNLS studied is the proline-tyrosine (PY) NLS. It majorly constitutes two clusters of motif in the peptide - a hydrophobic or basic N-terminus and a (R/K/H)$X_{2\text{-}5}$PY motif at the C-

terminus, where x can be any amino acid. Apart from the cNLS and ncNLS, some additional NLS discovered in recent years are not sequence-specific. For example, the signal transducers and activators of transcription 1 (Stat1) do not have a cNLS. However, upon dimerization, each subunit contributes a basic residue that forms a cNLS and mediates the nuclear transport (Lange et al., 2007b).

Proteins can be targeted to peroxisomes by recognition of two signals - PTS1 (Peroxisomal targeting signal 1) and PTS2 (Peroxisomal targeting signal 2). The PTS1 signal is located at the C-terminus. It consists of amino acid L at the extreme C-terminal position (position -1), a positively charged residue (K, R or H) at the penultimate position (position -2), and a small uncharged side chain (S, A or C) at position -3 (Brocard and Hartig, 2006; El Magraoui et al., 2019). After the remodeling of its tetratricopeptide repeat (TPR)-domain, the peroxisomal biogenesis factor 5 (Pex5) receptor recognizes peroxisomal proteins with this signal motif.

In contrast, the PTS2 signal can be longer, up to 10 amino acids. Recent studies have shown a consensus PTS2 motif: (R/K)(L/V/I)xxxxx(H/Q)(L/A), with the most common PTS2 consensus being R(L/V/I/Q)xx(L/V/I/H)(L/S/G/A)x(H/Q)(L/A), where x could be any amino acid. The peroxisomal biogenesis factor 7 (Pex7) receptor recognizes the PTS2 signal and targets the proteins to peroxisomes (Lazarow 2006, Waterham et al. 1994).

6.1.1.2 Role of transmembrane domains in understanding protein localization

Integral membrane proteins are permanently embedded to the cell membrane and serve various functions, including transporting proteins across the membrane. Transmembrane proteins are a type of integral membrane proteins that have transmembrane domains (TMDs) spanning across the lipid bilayer. TMDs consist of 16-30 hydrophobic amino acids forming an α-helical structure, which provides a stable anchoring of the protein in the lipid bilayer. TMDs are considered a separate class of localization signals.

Transmembrane proteins could be signal-anchored, short, multipass or tail-anchored (Ott and Lingappa, 2002)(Figure 21). Signal-anchored proteins are a type of transmembrane protein that utilizes its initial transmembrane domain (TMD) both as a signal sequence for targeting to the membrane and as a stop-transfer sequence to halt translocation. During protein synthesis, the N-terminal TMD of these proteins can be recognized by the signal recognition particle (SRP) and inserted into the endoplasmic reticulum (ER) co-translationally. Type I signal-anchored proteins possess a cleavable N-terminal signal peptide in addition to the TMD. Conversely, type II signal-anchored proteins have the TMD located in the middle of the protein sequence, while type III proteins have the N-terminal TMD. Transmembrane domain (TMD) proteins with short cytoplasmic tails at both ends are termed small or short TMD proteins, respectively. Multipass TMD proteins can contain several TMD domains. Furthermore, transmembrane proteins with a

signal-anchored domain at the extreme C-terminus are referred to as C-terminally anchored proteins or TA proteins. TA proteins are known for their diverse functions, including carrying out enzymatic and regulatory roles in cellular metabolism, aiding in protein localization, and facilitating membrane traffic (Borgese et al., 2003).

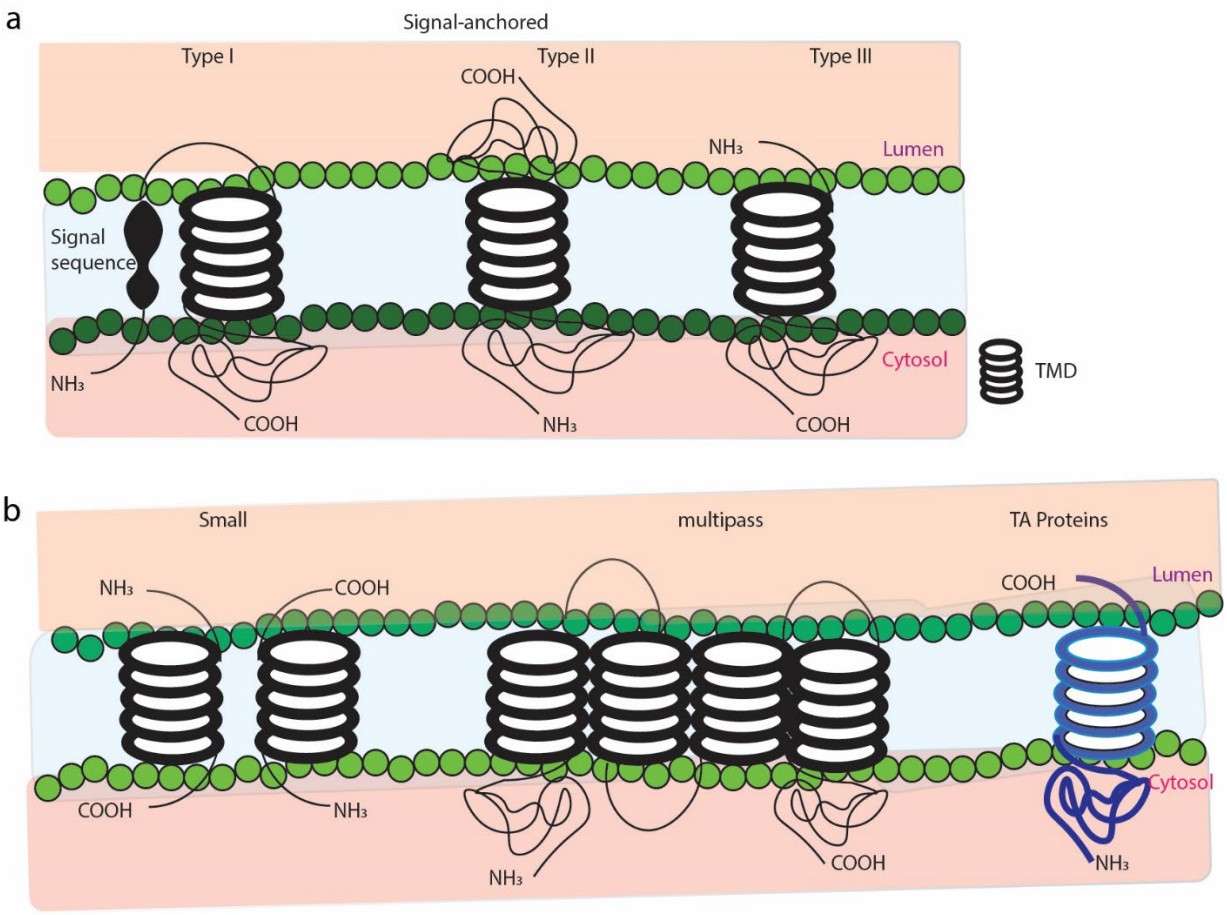

Figure 21: Classification of transmembrane proteins based on their topology *(a) Signal-anchored transmembrane proteins are classified into Type I, Type II or Type III. (b) Proteins that do not harbor signal-anchored TMDs are of small type or multipass. Proteins with TMDs at the C-terminus are TA-proteins and are the focus of this work.*

The TA-proteins are known to localize in ER, mitochondria, PM, peroxisome, nucleus and vacuole. Because of their ability to localize in different compartments, we have used them in this study to identify factors responsible for organelle-specific localization. Specific biophysical properties of the TMDs of the TA protein help them to localize them in their respective compartment. For instance, the TMDs of ER-resident TA proteins have high hydrophobicity (Wang et al., 2010), while the TMD of mitochondrial TA proteins have low hydrophobicity, low helical content and are flanked by positively charged AA (Costello et al., 2017). Change in the biophysical properties of these TMDs leads to their mislocalization. For example, the absence of a C-terminal positive charge in the mitochondrial protein Fis1 led to its mislocalization to the ER (Habib et al., 2003; Keskin et al., 2017). The addition of a positive charge to ER-resident TA-protein, Bos1, slowed its

insertion into ER, indicating that the GET pathway might disfavor positive charge to localize in ER (Rao et al., 2016a). PM-resident TA-proteins have longer TMDs to span through the membrane. Several computational studies also show the relation between hydrophobicity, volume, length and charge of TMD proteins to their localizations (Sharpe et al., 2010). For example, high cysteine content in the C-terminal TMD of TA-proteins can lead to PM localization (Venancio and Aravind, 2010). This subset of proteins was labeled as Cysteine-rich TMD proteins or Cystm, and are conserved across eukaryotes. The absence of the cysteine group can mislocalize the protein from PM (Joshi et al., 2020; Mir and León, 2014; Pereira Mendes et al., 2021).

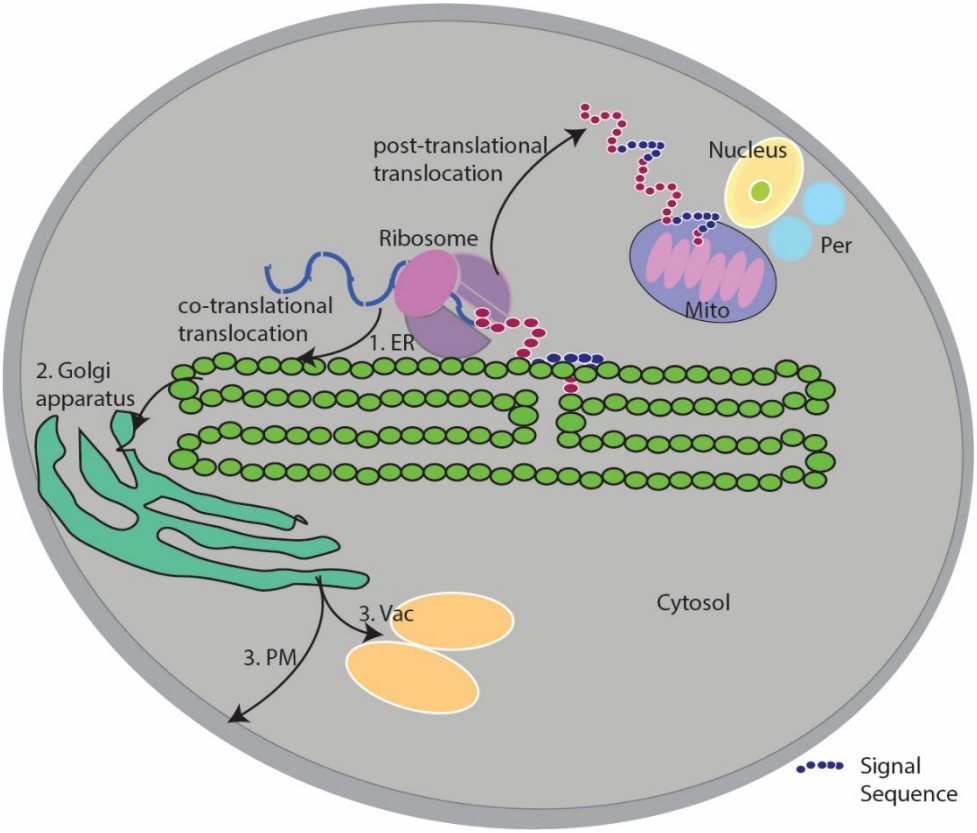

Figure 22: Protein targeting pathways in cells*. Proteins are synthesized in the cytosol. These proteins are then transported to destined pathways either co-translationally, post-translationally or by vesicular trafficking. Proteins targeted to ER can be either co-translationally or post-translationally. ER-destined proteins with N-terminal signal sequences are predominantly targeted co-translationally, while ones with C-terminal signal sequences are targeted post-translationally. Once targeted to ER, proteins are targeted to the Golgi apparatus via vesicular trafficking. From the Golgi apparatus, the protein is translocated to the vacuole or PM or exported to the cell exterior. Proteins destined for the nucleus, peroxisome and mitochondria are targeted post-translationally.*

Due to their ability to localize in multiple compartments, mutations in the TMDs of the TA proteins could be used to study the properties and mechanism of protein localization in different compartments. The known mechanism of translocations to different compartments is discussed in the next section.

6.1.2 Compartment-specific protein localization

Proteins, once synthesized at the cytosol, are directed to their appropriate compartment either co-translationally or post-translationally (Hegde and Keenan, 2022; Shurtleff et al., 2018). A schematic diagram of the protein localization mechanism is depicted in Figure 22.

6.1.2.1 Localization to mitochondria

Roughly 800 proteins in yeast and 1500 proteins in humans are mitochondrial proteins (Calvo et al., 2016; Morgenstern et al., 2017). They are synthesized in the cytosol. These proteins are in their nascent unfolded state and thus require the assistance of many chaperones and other folding factors once they enter mitochondria.

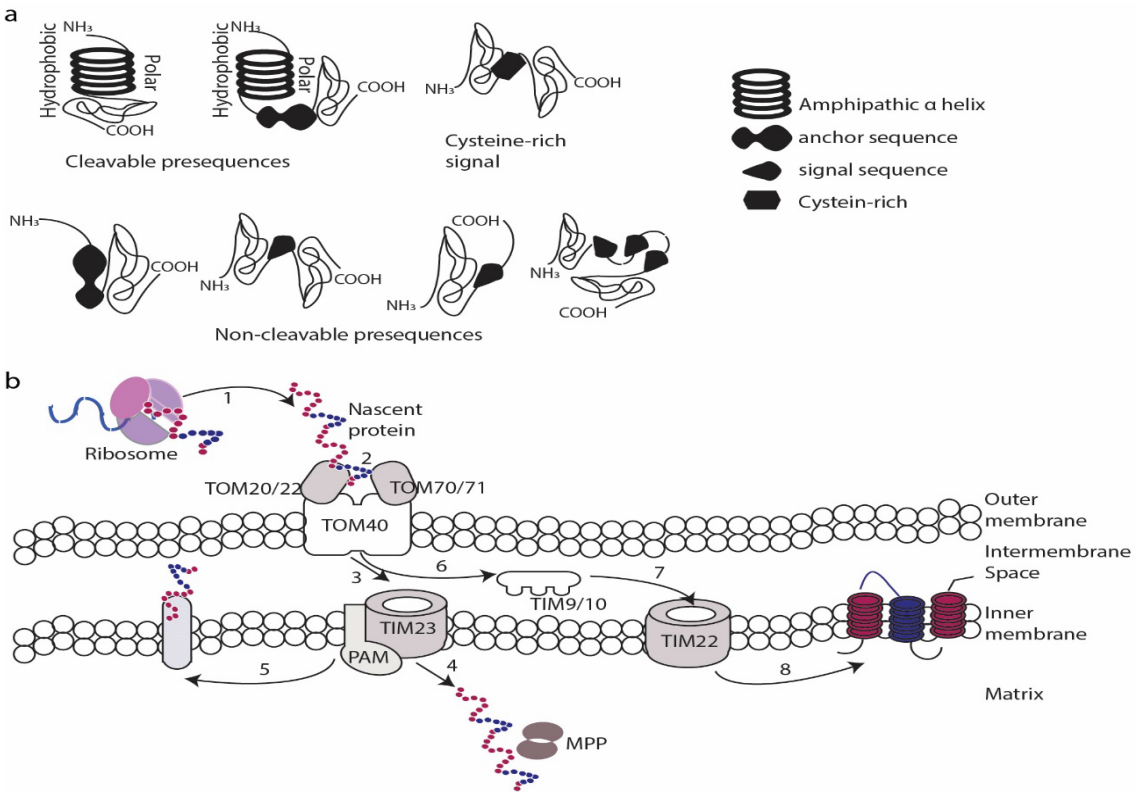

Figure 23: Protein localization in mitochondria*. (a) Types of mitochondrial localization signal (MLS). MLS have cleavable presequence, cleaved after the nascent protein localizes into mitochondria. The cleavable sequence has an amphipathic α helix with alternate hydrophobic and positive charge. This set of sequences may contain an anchor sequence followed by the α helix. Non-cleavable signal sequence either has an N-terminal anchor sequence or has a signal sequence in the internal or C-terminal. The nascent protein may have multiple signal sequences in the internal region. Additionally, the signal sequence may also be cysteine-rich. (b) General mechanism of protein localization in mitochondria. Mitochondrial translocation happens post-translationally in unfolded form. The nascent protein in the presence of Hsp70/90 heat shock proteins is targeted to TOM40 and is recognized by substrate receptor TOM20/22. The protein could then have multiple routes it can take. Proteins destined for the mitochondrial matrix are imported to the matrix by TIM23 and PAM. Once inserted into the matrix, the mitochondrial processing peptidase cleaves off the presequence. The proteins are then folded into the correct conformation. Error in the folding leads to degradation of these proteins. Proteins in the mitochondrial inner membrane are laterally inserted into the lipid phase (step 5).*

Another route for inner membrane targeting is when the hydrophobic precursor proteins are recognized by TIM9/10 and passed to TIM22.

A majority of mitochondria proteins are targeted to respective sub-mitochondrial compartments through the classical import pathway, which uses cleavable presequence in protein to be targeted (Figure 23 (b)). The proteins are first targeted to the mitochondrial outer membrane by the Translocase of Outer Membrane (TOM) complex. For this, the presequence in the nascent protein is recognized by Tom20/22 and translocated through the Tom40 pore. Proteins destined for the mitochondrial inner membrane are targeted to the mitochondrial inner membrane, for which the Tom22 binds with the translocase of the inner membrane protein Tim23. Tim50, a Tim23 subunit exposed to intermembrane space (IMS), translocates the protein from the outer membrane to the inner membrane. Once inside the inner membrane, the PAM machinery (presequence translocase-associated motor), with Hsp70 chaperon as its active player in the mitochondrial matrix, guides the protein into the matrix and prevents backsliding to the inner membrane (Backes and Herrmann, 2017). Following translocation, the presequence is cleaved by the mitochondrial processing peptidase (MPP), and Hsp70 mediates correct protein folding (Burkhart et al., 2015).

6.1.2.2 Localization to ER

The transport of precursor proteins into and across the ER membrane is a highly conserved process in eukaryotic cells. These proteins can be transported to the ER either co-translationally or post-translationally. The translocation process of the protein to the ER is described in this section.

Co-translation targeting of ER-protein

Once an N-terminal targeting signal for ER proteins emerges from the ribosomal exit tunnel, it can be recognized by the SRP. Translation at this point is halted, and the ribosome-nascent chain complex (RNC) is directed to the ER membrane with the help of the cognate SRP receptor (Gilmore et al., 1982; Rapoport, 2007; Walter et al., 1981; Zimmermann et al., 2011). The SRP not only protects the hydrophobic polypeptide from degradation during translation but also targets the RNC to the ER membrane. For this, SRP first guides the RNC to channel Sec61 (Görlich and Rapoport, 1993). After the GTP hydrolysis, the ribosome is dissociated, and the nascent protein, along with SRP, is inserted into the Sec61 channel (Mitra and Frank, 2006). In the next steps, the membrane proteins diffuse laterally from the Sec61 complex into the bilayer. Chaperons such as BiP can ensure a unidirectional transport of nascent protein through the Sec61 channel to retain these proteins in the ER lumen (Alder et al., 2005). Their interaction with the nascent protein is mediated by the J domain of ER proteins such as Sec63 in an energy-dependent manner (Lyman and Schekman, 1995). Once the nascent protein enters the ER lumen, the signal sequence is

cleaved off by the signal peptidase complex (Linxweiler et al., 2017). The translocated nascent protein is then correctly folded into its conformational state.

Post-translation targeting of ER-protein

Proteins post-translationally targeted to the ER are mostly TA proteins with a hydrophobic TMD at the C-terminus (Hegde and Keenan, 2022). They are mostly recruited to ER via the GET pathway in yeast or the TRC40/GET pathway in mammals (Denic, 2012; Guna et al., 2018; Schuldiner et al., 2008; Wang et al., 2010). The TMD of the newly synthesized proteins is recognized by the Sgt2 chaperone in yeast (SGTA in mammals) near the ribosome. The TA-Sgt2 chaperone is handed over to the Get3 complex after mediation by the Get4/5 scaffolding complex. The Get1/2 complex binds to the TMD binding site in Get3 and enables the release of the protein to the ER membrane. Post the release of protein to the ER membrane, Get3 is recycled to the cytosol to help in the next rounds of targeting (Hegde and Keenan, 2022). In this process of targeting the protein to the ER membrane, at times, the proteins also get ubiquitinated (Culver and Mariappan, 2021). There could be mainly two reasons for ER proteins to be ubiquitinated during their translocation from cytosol to the ER: 1. to prevent the enzymatic actions or potential unwanted interactions of soluble domains of TA protein (McQuown et al., 2021) and 2. to avoid the interaction of the TA protein with the unfolding enzymes such as p97 ATPases to facilitate their folding and maturation (Culver et al., 2022).

6.1.2.3 Localization to the nucleus

Cargo proteins produced in the cytoplasm and destined for the nucleus contain NLS and must be transported efficiently to the nuclear envelope post-translationally. This process is facilitated by members of the importin superfamily, consisting of importin αs and importin βs. The Ran protein facilitates nuclear transport by providing energy. Ran protein, due to its asymmetric distribution in the cell, acts like a molecular switch that controls compartment-specific transport of cargo protein. The importin α recognizes the NLS of the cargo protein, which in turn forms a trimer with importin β1 (cNLS-importin α- importin β1). The importin β1 directs importin α to the nuclear pore complex (NPC). Once the cNLS-importin α- importin β1 trimer is inside the nucleus, RanGTP binding causes a conformational change in importin β1, thus releasing the cNLS-importin α. Further, nucleoporins in the NPS and CAS help in the dissociation of cNLS cargo into the nucleus. Once the cargo is in the nucleus, importin α is exported out of the nucleus with the help of CAS. The importin β1-RanGTP complex returns to the cytoplasm, where it is hydrolyzed to RanGDP and is ready for the next round of import (Lange et al., 2007a). Additionally, proteins lacking cNLS follow other mechanisms for nuclear import, as is the case of the nuclear import of proteins with PY-NLS (Terry and Wente, 2009). However, unlike the localization signals for the ER or mitochondria, the NLS can remain intact post-translocation. Thus, there may be many continuous rounds of nucleus-cytoplasmic transports.

6.1.3 Protein quality control for mislocalized proteins

When proteins localize to an incorrect compartment, they can be recognized by protein quality control (PQC) machinery. The PQC for mislocalized proteins either redirects them to their correct compartment or degrade them. Several protein quality factors continuously monitor the translation, targeting, insertion, folding and assembly. When the proteins mislocalize in the wrong compartments due to the change in the cellular environment, they can misfold to expose some structural or biophysical motifs, which can lead to degradation. The major pathway for the degradation of such proteins is the ubiquitin-proteasome system (UPS).

6.1.3.1 Quality control factors in cytosol

The ER-resident proteins can mislocalize to cytosol when they have mutated or absent signal sequences or have a failure in targeting machinery. These are recognized and eventually degraded. For example, when the ER signal is mutated in the protein carboxypeptidase Y and the vacuolar proteinase A, it leads to its mislocalization in the cytosol (Prasad et al., 2010). There are two known mechanisms for their removal from the cytosol. Upon mislocalization, the proteins misfold and are recognized by the Stress-Seventy subfamily A protein (Ssa1) and Yeast dnaJ (Ydj1). Further, they are ubiquitinated by the Ubr1 E3 ligase in the cytosol (Park et al., 2007). An alternate path is that these misfolded proteins are transported to the nucleus by the combined action of Ssa1, Ydj1, SseE1 and Sis1 and then ubiquitinated in the nucleus by San1 E3 ligase (Heck et al., 2010).

Not all mislocalized proteins in the cytosol misfold. Membrane protein is one such group of proteins that, when mislocalized in the cytosol, do not misfold but aggregate and expose their TMDs. For ER targeting proteins such as Trc40, these proteins are either re-localised to the ER membrane or degraded by the action of ribosome-associated chaperone BAG6. If not correctly re-localized, Bag6 directs the mislocalized ER-resident TA protein for ubiquitination by Rnf126 (Ring Finger Protein 126) E3 ligase (Rodrigo-Brenni et al., 2014; Mariappan et al., 2010).

6.1.3.2 Quality control factors in mitochondria

TA proteins can localize to multiple compartments. However, due to limited fidelity, the proteins tend to mislocalize even under optimal cellular conditions. When correctly targeted to peroxisomes, the Peroxisomal Membrane Protein 15 (Pex15) interacts with Pex3, hiding the hydrophobicity patch recognized for quality control and the positive charge at the C-terminus (Li et al., 2019; Castanzo et al., 2020). However, when mistargeted to the mitochondrial outer membrane, Pex15 is orphaned by the lack of Pex3 and is recognized by the AAA-ATPase Msp1. The mistargeted Pex15 is extracted by Msp1 to the cytosol, where it gets ubiquitinated by Doa10 E3 ligase along with Ubc6/7 and is targeted for degradation (Dederer et al., 2019). Alternatively, the extracted Msp1 substrates could be re-localized to the ER membrane. However, when not relocated into the correct compartment, the proteins are ubiquitinated by Doa10 E3 ligase,

extracted by another AAA-ATPase Cdc48, and targeted for proteasomal degradation (Matsumoto et al., 2019). Msp1 also plays a key role in clearing mitochondrial import intermediate from clogged TOM complexes (Weidberg and Amon, 2018).

6.1.3.3 Quality control factors in membrane

Several different mechanisms degrade proteins mislocalized in the ER. One such mechanism involves BiP, a member of the Hsp7- heat shock family. BiP plays a role in protecting against unwanted degradation of the immunoglobulin heavy chains until they partner with the light chain; otherwise, it commits them to degradation (Lee et al., 1999). BiP also plays a key role in recognizing the mistargeted TCR-defined complex subunit and extracting them for proteasomal degradation. It binds with TMD of unassembled T cell receptor (TCR) α-chains when it is mistargeted to the ER lumen and targets it to the proteasomal degradation via ERAD-associated E3 ubiquitin-protein ligase Hrd1 E3 ligase (Call et al., 2002; Bonifacino et al., 1990).

Positive charge in the TMD of the TCR α-chains are usually masked by the TCR β-chains. If this interaction fails, TCR α-chains are ubiquitinated by the Glycoprotein 78 (Gp78) E3 ligase, followed by the intramembrane cleavage by intramembrane serine Rhomboid Protease (Rhbdl4) and the transfer from ER to the cytosol by the Cdc48/p97 before proteasomal degradation. Another subunit of TCR complex CD3-δ is ubiquitinated by Gp78 and Trim13/Rfp2 E3 ligases before being proteasomally degraded (Lerner et al., 2007; Fleig et al., 2012).

One of the most common pathways for targeting mislocalized proteins in ER is through ER-associated degradation (ERAD). Proteins misfolded or mistargeted to the ER are recognized by the ER chaperones, like BiP, by their exposed TMDs or the hydrophobic stretch. They are ubiquitinated by Doa10 or Hrd1 E3 ligases before being extracted from ER to the cytosol by Cdc48 (Ruggiano et al., 2014).

Due to some shared features of localization signals of ER and nucleus and close positions of subcellular compartments, some ER proteins may occasionally leak into the nuclear membrane. Quality control with the nuclear membrane is maintained by the Asi E3 ligase in yeast, which is composed of Asi1 and Asi3. The ASI complex functions along with Ubc6/7 and Cdc48 for targeting unwanted proteins in nuclear membranes such as Erg11 and Sec61 (Khmelinskii et al., 2014; Foresti et al., 2014).

Apart from direct degradation, the mistargeted proteins are also given a second chance for correct relocation to their compartment. For example, when the mitochondrial inner membrane protein Oxa1 is mistargeted to ER, it is redirected to mitochondria with the help of chaperone Djp1 (Hansen et al., 2018). The inverse also holds. Some ER-destined TA proteins, when mislocalized to mitochondria, are also relocated correctly to the ER (Matsumoto et al., 2019). This relocation serves additional PQC mechanics for correct localization apart from degradation.

6.1.4 Effect of protein mislocalization on health

The mislocalization of nascent proteins and their aggregation in the wrong compartment is known to cause several diseases, including cancer, neurodegenerative disorder and aging (Wang and Li, 2014; Hoover et al., 2010; Szczesny et al., 2003). Aging may lead to protein mislocalization due to the deterioration of the NPC and reduced efficiency of mitochondrial import (D'Angelo et al., 2009). Moreover, aging is correlated with the loss of multiple protein complexes, including the ribosome, proteasome, and nuclear pore, due to the deregulation of the translation machinery. In various experimental models of Alzheimer's, Parkinson's, and Huntington's diseases, mitochondrial import is inhibited. The expression of Ubqn1, which is involved in the degradation of some mislocalized proteins in the cytosol, appears to be limiting in Alzheimer's and Huntington's, as its overexpression can alleviate cytotoxicity and disease symptoms (Di Maio et al., 2016). All these mechanisms increase the load of mislocalized and orphan proteins in aging or neurodegeneration, thus reducing the capacity of protein quality control systems. This eventually leads to proteostasis collapse and cell death.

When a protein is found in the wrong place, it can cause toxicity, due to many reasons, such as protein aggregation or gain of function. Studies have shown that the altered location of transcription factors, such as NF-kB, Atf2, Creb, p53, E2F transcription factor, and Nrf2, can lead to the commitment of cell death in various neurodegenerative diseases (Chu et al., 2007; Hung and Link, 2011).

Tumorigenesis is frequently associated with protein mislocalization, as observed for many proteins with oncogenic, tumor-suppressive, or other functions. Cancer cells are subjected to constant challenges to their proteome integrity and, therefore, show increased dependency on protein folding and quality control systems to sustain their rapid growth. Several inhibitors of the HSP90 chaperone are being tested for their anti-cancer efficacy with some promising results (Liu et al., 2019). Further understanding of the mechanisms and quality control factors that handle mislocalized and orphan proteins could help identify novel targets and strategies for anti-cancer therapy.

6.1.5 Existing work on deciphering protein localization

Several experimental and computation methods have been developed over the years to study the localization of the proteins. Some of the recent prominent work for understanding SLiMs helping in protein localization are discussed in this section.

6.1.5.1 Experimental methods to understand protein localization

Experimental methods to understand protein localization involve a variety of techniques. These techniques provide insights into subcellular compartments where the proteins reside. Methods based on quantitative mass spectrometry allow the identification of proteins across subcellular fractions (Itzhak et al., 2016; Orre et al., 2019). Subcellular fractionation involves isolating different cellular organelles or compartments using centrifugation-based techniques. By separating organelles based on their size, density, or other physical properties, proteins in

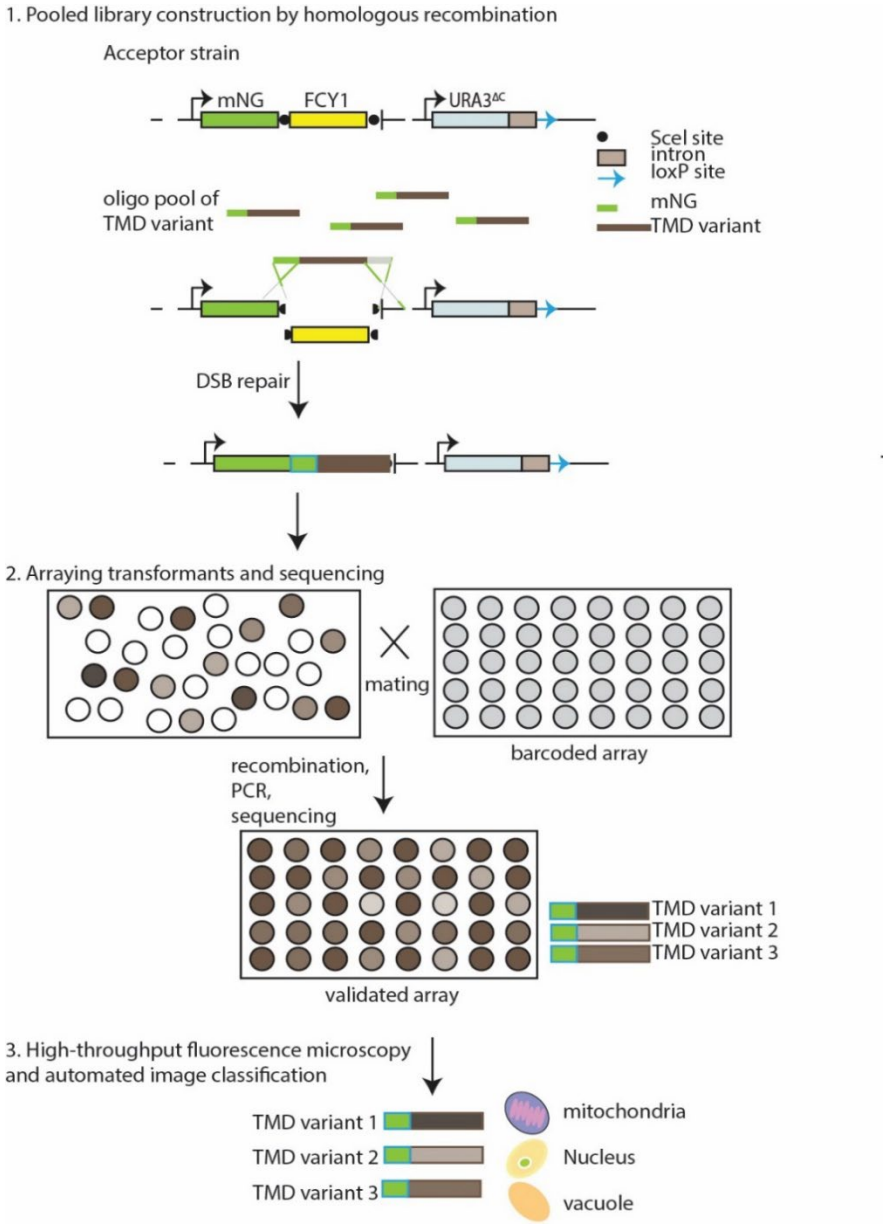

Figure 24: Schematic diagram of the experimental workflow of iPAL*: Pooled variants that differ only in the TMDs of TA proteins are created and introduced to the acceptor strain. The pooled library is further created into an array library of TMD variants by deep sequencing, such that each well encompasses one TMD variant. The localization of the variant is then observed using microscopy and processed using automated image analysis.*

different cellular fractions are analyzed, providing insights into their localization. Techniques based on labeling of proteins using fluorescently tagged antibodies or electron-dense markers coupled with antibodies have enabled the visual estimation of protein localization within a single cell using immunofluorescence microscopy or immunoelectron microscopy (Chong et al., 2015; Osafune and Schwartzbach, 2007). Moreover, dynamic protein localization can also be studied using Live-cell imaging, in which fluorescently tagged proteins or organelle-specific markers can be tracked over time using advanced microscopy systems, providing dynamic insights into protein trafficking, localization changes, and interactions (Liu et al., 2015).

Modern methods to study phenotypes encompass high-content screening, which consists of high-throughput microscopy and automated image analysis to understand the relation between protein sequences and the visual phenotype, such as their localization. Typically, the high-content screen is performed in array format such that the effect of variants, such as mutation of amino acids, can be tested in separate wells of the array in parallel (Mattiazzi Usaj et al., 2016). Compared to this method, the pooled format to understand the effect of perturbation is much more cost-effective and has high throughput. Pools of variants are created with the phenotype under study linked to selectable readouts, such as the expression of a fluorescent reporter. The cells with these phenotypes are then collected using fluorescence-activated cell sorting (FACS). Further, targeting DNA or RNA sequencing is used to identify the variants in these cells.

Recent approaches of high-content pooled screens completely bypass the need for selectable readouts to be linked to visual phenotype. For example, a modified version of Fluorescence in situ hybridization (FISH), called immunofluorescence in situ hybridization (immuno-FISH) or protein-FISH, is employed to decipher protein localization using probes that specifically bind to protein targets (Meng et al., 2018). These probes are typically antibodies conjugated to fluorescent dyes. The antibodies are first bound to the target proteins, with excess antibodies washed out. Next, the labeled proteins are visualized using a fluorescence microscope, and the images are analyzed to determine the localization pattern. In another method, such as the Visual Cell Sorting method, cells are engineered to express a photoconvertible fluorescent protein like Dendra2, serving as a phenotypic marker (Hasle et al., 2020). Subsequent automated imaging and analysis led to the targeted photoconversion of Dendra2 from green to red fluorescence, specifically in cells exhibiting the desired phenotype. Following this, the entire cell population is sorted based on the photoconversion state using fluorescence-activated cell sorting. However, careful consideration is needed regarding the dilution of the photoconverted signal over time due to cell division, particularly in rapidly dividing organisms like bacteria and yeast.

For some protein localization studies, the best option would be to have an array of variants of a localization signal (or different localization signals), but dealing with arrayed libraries is usually very laborious. With the goal of combining the ease of the pooled library and the efficiency of the

arrayed library for screening, a deep mutational scanning approach to dissect protein localization signals, called imaging pooled-to-arrayed libraries, iPAL, was created (Figure 24,(Ivanova, 2022)).

In iPAL, an oligonucleotide pool of libraries of protein variants that differ only in potential localization signal, such as Transmembrane Domains (TMDs) of Tail-Anchored (TA) proteins, are constructed and introduced into an acceptor locus. The acceptor strain consists of a mNeonGreen (mNG) fluorescence reporter, a selection marker – FCY1, flanked by SceI cut sites and a terminator. Further, the acceptor strain is transformed with an oligo pool and a plasmid encoding SceI endonuclease, which induces double-strand breaks (DSBs) at the integration locus. The selection marker in the acceptor strain is efficiently replaced with the oligos encoding TMDs by homologous recombination (HR), leading to the expression of the mNG-TMD variant, referred to in this study as the TMD variant. Using deep sequencing, pooled libraries are converted into verified arrayed libraries (Smith et al., 2017), followed by high-throughput fluorescence microscopy to determine the localization of each variant.

6.1.5.2 Computational prediction of subcellular localization of proteins

Experimental methods to identify protein localization are laborious, time-consuming and resource-extensive. In addition, experimental and computational approaches for identifying protein localization complement each other, with experimental annotations often serving as reference points for computational methods. Computational models are trained using these validated datasets to predict the localization of additional proteins. The cost-effective, automated, and high-throughput nature of computational methods makes them valuable for large-scale characterization of protein subcellular locations. Several reviews discuss the developments in protein localization prediction in many organisms. For example, (Gardy and Brinkman, 2006) focus on bacterial protein localization prediction. (Shen et al., 2020) reviews methods for protein localization in humans. Protein prediction uses multiple types of data. (Jiang et al., 2021) discusses the features, algorithms, and tools for protein prediction in great detail. The prediction tools for sub-cellular localization can be mainly divided into four based on the input they take- 1. overall amino acid composition, 2. amino acid sequence of the targeting signal, 3. sequence homology, and 4. hybrid or all those as mentioned above.

Since experimentally, the most commonly studied localization signal is the NLS, large-scale experimental datasets were developed. Prediction tools specialized for understanding nuclear localization have since been developed. For instance, PredictNLS is a method specialized in recognizing nuclear proteins based on a collection of nuclear localization sequences. These methods also have indicated that NLS need not be at the N-terminus of the protein but could be even in the internal region (Cokol et al., 2000).

Recent computational prediction methods involve sophisticated machine-learning approaches. Machine learning plays an important role in the development and implementation of complex underlying biological models, which can also be seen from the frequent application within this field. As methods for protein classification become more accurate, it becomes increasingly important to examine the small proportion of misclassified proteins, as they may be players of interface organelles. The study of these could help identify the mechanics of protein localization and the chaperones that help mediate translocation smoothly. Additionally, to improve the model accuracy, it is necessary to continually update methods and gather new datasets for training prediction models.

Advanced algorithms such as support vector machines (SVM) (Kumar et al., 2017), the Markov chain model and artificial neural networks (Almagro Armenteros et al., 2019; Savojardo et al., 2018) are used to predict the localization of proteins with high accuracy. CELLO2GO is a web server designed for predicting protein subcellular localization and providing functional gene ontology annotation (Yu et al., 2014). CELLO2GO integrates multiple machine-learning classifiers to accurately predict the subcellular localization of proteins, offering users a user-friendly interface for streamlined analysis. The tool also incorporates functional gene ontology annotation to enhance the understanding of protein functions and cellular processes. A modified version of SVM, called multiple kernel learning, is used for predicting subcellular localization (Hasan et al., 2017). By integrating information from multiple data sources, such as amino acid sequences, evolutionary information, and physicochemical properties, the proposed method enhances the accuracy of localization prediction. Through MKL, different types of data are combined effectively, allowing for improved discrimination between subcellular compartments. Artificial neural networks also help in predicting localization. SCLpred (Kaleel et al., 2021) is a novel method for predicting protein subcellular localization using N-to-1 neural networks. SCLpred utilizes a neural network architecture capable of handling multiple input sequences and predicting their corresponding subcellular localizations simultaneously. By leveraging features extracted from protein sequences, such as amino acid composition and physicochemical properties, SCLpred achieves accurate localization predictions across multiple subcellular compartments.

6.2 Aim

Short Linear Motifs play a key role in protein localization. These could either be due to the specificity of amino acids or biophysical properties. However, SLiMs and the features of proteins that help in localization, are not yet completely understood.

One of the groups of integral membrane proteins is tail-anchored (TA) proteins. They contain a C-terminal transmembrane domain that, in most TA proteins, targets them to correct subcellular localization. Because of their ability to localize in different compartments, TA proteins can be used to identify factors responsible for organelle-specific localization.

In this work, I investigate how various biophysical properties of transmembrane domains (TMD) in Tail-anchored proteins help in localization. The aim of the project is fulfilled with two sub-goals:

1. To perform computational analysis on the biophysical properties of TMD variants of TA proteins to find determinants of sub-cellular localization

2. To create the machine-learning model for the interpretation of rules for sub-cellular localization based on biophysical properties

6.3 Methods

6.3.1 Data collection using iPAL

Proteins are targeted to their correct compartment due to the presence of localization signals in their sequences. Tail-anchored (TA) proteins are membrane proteins that carry **α**-helical Transmembrane Domain (TMD) at their C-terminus, which enables them to localize to the ER, mitochondria, PM and several other compartments. For some TA proteins, TMD domains alone can act as a localization signal. To understand the specificity of localization signals, I analyzed the localization of TMD variants for 59 TA proteins and their relation with various biophysical properties. The variants are created by adding flanking regions, N- and C-terminal flanking charge, and elongating the TMD length by addition of V and A. Additionally, to assess the importance of the AA sequence, the library of shuffled TMDs is also created. These localizations are found using (imaging pooled-to-array library) iPAL, in which variants are created using DMS and localizations are found using fluorescence microscopy (Figure 24, (Ivanova, 2022)).

Overall, the dataset contains 864 TMD variants from 59 unique TA proteins. Out of these, 6 TA proteins (Ybr016w, Ydr034w-b, Ybr056w-a, Ydr210w, Ydl012c) are cysteine-rich based on high cysteine content in their TMDs. The dataset consists of 8 localization classes: Endoplasmic Reticulum (ER), mitochondria, cytosol, vacuole, puncta, Plasma Membrane (PM), PM+ and ambiguous. For some TMD variants, localization signals are spread in PM, ER and vacuole and are labeled as PM+. Additionally, for some TMD variants, localization signals are found in several compartments and are labeled as ambiguous.

6.3.2 Analysis of TMD properties

The physical and biophysical properties of TMD variants are analyzed to understand their effect on localization. TMD variants contain TMD from the TA proteins along with the fusions. The fusions could be additional flanking regions, charged flanking residues at the N- and C-terminal or the addition of amino acids V or A. For each TMD variant in the dataset, the charge of the flanking residues, hydrophobicity, length, fraction of V and L in the N-half and the C-half of TMD, and volume is calculated from the "Peptide" package in R (Osorio et al., 2015). The hydrophobicity and volume are calculated using the "KyteDooLittle" scale (Kyte and Doolittle, 1982), the "Lehninger" scale (Boyle, 2005) and the "Pointus" scale, respectively. The effect of one biophysical property on another for each of the localization classes is inferred using simple visualization techniques such as dot plots, density plots and automated density plots in R (Wilkinson, 2011). Density plots are smoothened with 100 grid points in all directions.

Apart from the analysis of individual variant groups, analysis was also done by taking all variants simultaneously to understand the mechanics of localization. To understand the impact of amino acids at certain positions for localization into a compartment, we calculated the enrichment of amino acids per position for each compartment. For amino acid A at position P in localization class L,

$$Enrichment\ (A, P, L) = \log 2 \frac{relative\ frequency\ of\ A\ at\ P\ for\ L}{relative\ frequency\ of\ A\ at\ P\ for\ all\ L}$$

Where relative frequency is defined as:

$$relative\ frequency = \frac{\text{number of TMDs with A at P}}{Total\ number\ of\ TMDs}$$

Also, charge and hydrophobicity per position for TMD variants are calculated to understand if biophysical properties play a role in some specific regions to help in localization.

6.3.3 Classification of localization using biophysical properties

Simple two-dimensional visualization helps in understanding the effect of two properties on localization at a time. With the hypothesis that multiple biophysical properties of the proteins play a role in targeting to their sub-cellular localization simultaneously, I created an interpretable classification model using random forest. Several biophysical properties are first considered for the model - hydrophobicity, charge, polarity, hydrogen bonding, alpha and turn propensity, length, cysteine fraction, volume and charge of the flanking residue of the TA variants. Since polarity, hydrogen bonding, and alpha and turn propensity are related to hydrophobicity, they are derived variables and might influence the performance of the random forest model. For the final model, I took only the independent properties – hydrophobicity, charge of N and C-flanking regions, and length of TMD variants. Each of the models is created with 250 trees. The performance of the model is measured by the Out-of-bag (OOB) error rate and class error rate for each localization. The OOB estimate of error rate is a useful measure to discriminate between different random forest classifiers and is defined by:

$$OOB\ error\ rate = 1 - \frac{\text{True Positive}}{Total\ number\ of\ sample}$$

The first model is created for classifying all sub-cellular localizations. Since multiple localization classes share localization signals with similar biophysical properties, for some classes, error rates are high compared to others, meaning random forest does not perform well to classify them. A low number of TMD fusions in the class may also result in a poor classification rate. The final model is created for classifying mitochondria, ER and two classes of PM. Two separate models are created for datasets with and without cysteine-rich datasets.

The importance of property on effective classification is measured by its Mean Decrease Accuracy. Mean decrease accuracy represents a drop in model accuracy if a property is removed. The rules for correct localization are then found by plotting trees from random forests.

6.4 Results

6.4.1 Biophysical properties of TMD variants help understand their localization

Biophysical properties of various localization signals can affect where they localize. To test which biophysical properties can help in localization to a specific compartment, I analyzed the localization of TMD variants from 59 TA proteins. The TMD variants can be categorized as:

1. TMD variants based on flanking lengths (Figure 25(a))
2. TMD variants with different N and C terminal charges (Figure 25(b))
3. TMD variant with different lengths (Figure 25(c))

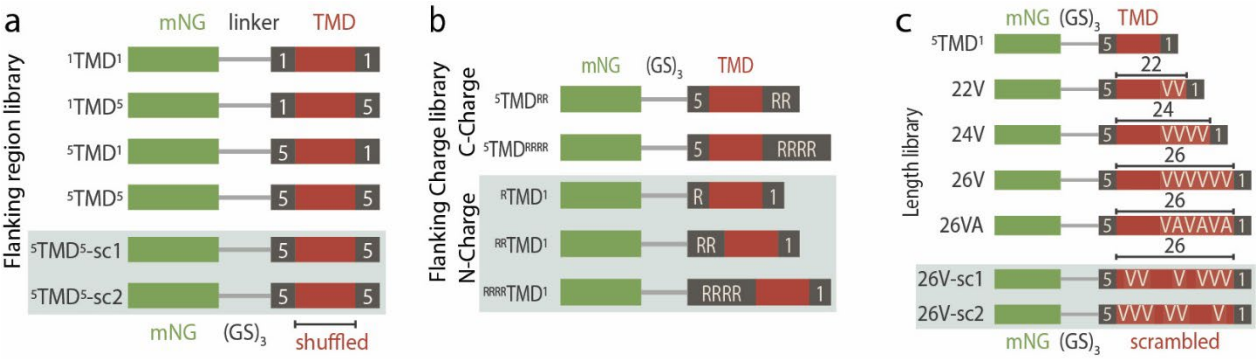

Figure 25: Libraries of TMD variants used to dissect SLiMs that help in localization*: Three libraries with TMD variants varying in flanking region, flanking charge and overall TMD variant length are created using deep mutational scanning: (a) Flanking region library: Six variants of TMD from 59 TA proteins are created with varied flanking regions. A single residue before and after the TMD (^{1}TMD^{1}), a single residue before and 5 residues after the TMD (^{1}TMD^{5}), 5 residues before and a single residue after the TMD (^{5}TMD^{1}), 5 residues before and 5 residues after the TMD (^{5}TMD^{5}) and with randomly shuffled TMD sequence (^{5}TMD^{5}-sh1 and ^{5}TMD^{5}-sh2), were fused with the mNeonGreen (mNG. (b) Flanking Charge library: Two variants with flanking positive C-Charge: (^{5}TMD^{RR}) and 4 (^{5}TMD^{RRRR}), where R is added to C-terminal of TMD for 59 yeast TA proteins, and 5 variants with flanking positive, N-Charge library: 1 arginine at N-terminus (^{R}TMD^{1}), 2 arginine at N-terminus (^{RR}TMD^{1}) and 4 arginine at N-terminus (^{RRRR}TMD^{1}). (c) TMD length library: The TMD sequences from 59 yeast TA proteins with 5 native residues before and a single native residue after the TMD (^{5}TMD^{1}) were elongated by adding valine (V) and alanine (A) residues to the C-terminal part of the TMD to the total length of 22 (22V), 24 (24V) and 26 (26V and 26VA) AA. For the longest version of the TMD (26V), two combinations of randomly shuffled elongated TMD sequences (26V-sh1 and 26V-sh2) were created.*

These variants are created using deep mutational scanning, and the localization is determined using microscopy and automated image analysis. Downstream analysis is conducted by analyzing the relation of the biophysical properties of the TMD variants to their subcellular localization.

6.4.1.1 Analysis of TMD variants with different flanking regions

The context around the TMD is crucial in determining their localization for some TA proteins. Thus, to understand how the flanking length of the TMD variants can affect the localization, five variants of TMDs with different N- and C-flanking lengths are created, and their localization is

found using microscopy. The variants have either one or five flanking amino acids at the N- or C-terminus. There are 4 TMD variants created in this set – 1. ^{1}TMD^{1}: 1 flanking residue before and after the TMD, 2. ^{1}TMD^{5}: 1 flanking residue before the TMD and 5 flanking residue after the TMD, 3. ^{5}TMD^{1}: 5 flanking residue before and 1 flanking residue after the TMD, 4. ^{5}TMD^{5}: 5 flanking residues before and after the TMD. To evaluate the amino acid specificity of TMD variants in localization, two additional variants are created after shuffling 5TMD5, labeled ^{5}TMD^{5}-sh1 and ^{5}TMD^{5}-sh2 Figure 25(a)). This section discusses the effect of the biophysical properties of these TMD variants with varied flanking regions on localization.

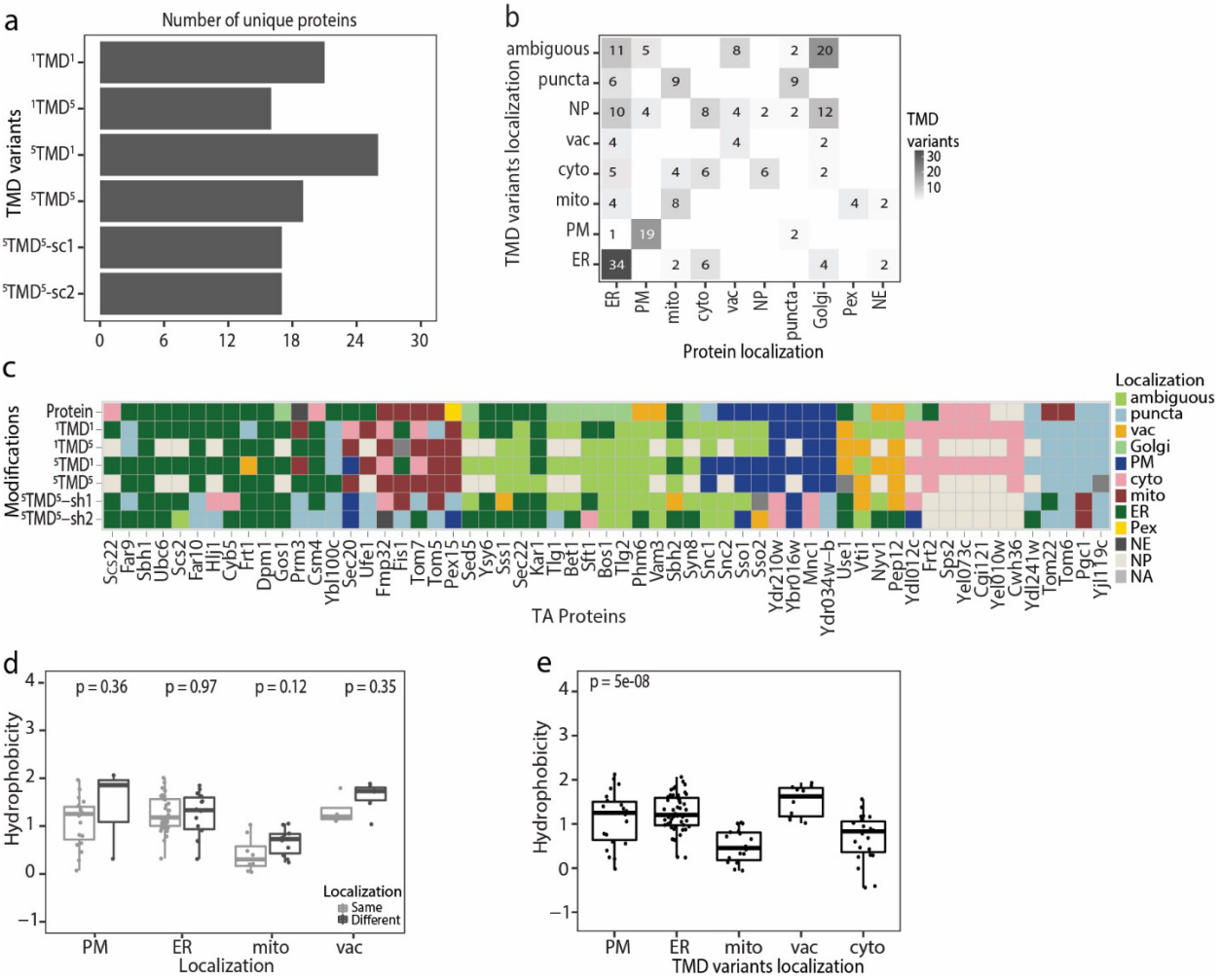

Figure 26: The biophysical property of TMD variants with flanking regions helps in localization. *The TMD variants in this library include ^{1}TMD^{1}, ^{1}TMD^{5}, ^{5}TMD^{1},^{5}TMD^{5},^{5}TMD^{5}-sh1, ^{5}TMD^{5}-sh2 (a) Number of TMD variants with flanking regions showing similar localization as corresponding full-length TA proteins, (b) Comparison of observed localization of all TMD variants with flanking regions and full-length TA protein localization, (c) Localization of full-length TA proteins and TMD variants. ER, endoplasmic reticulum; mito, mitochondria; mixed, several compartments; PM, plasma membrane; vac, vacuole; cyto, cytosol; NE, nuclear envelope; NA, protein variants, which were not identified in the experiment; NP, protein variants without C-terminal extension, which were not included in the experiment. (d) Comparison of hydrophobicity per localization for TMD variants with the same localization as TA proteins. p-value*

from the student t-test (e) Variation of hydrophobicity of TMD variants per localization.p, p-value from ANOVA test. (g) Comparison of hydrophobicity of shuffled TMD fusions with 5TMD5 per localization. p, p-value from student t-test

To verify whether the transmembrane domains (TMDs) effectively target TA (tail-anchored) proteins to their intended subcellular compartments, we analyze the localization patterns of TMD variants alongside the published localization data of TA proteins in the dissertation. (Ivanova, 2022). 71% of the TMD variants (^{1}TMD^{1}) have localization the same as their corresponding TA proteins, suggesting that TMDs alone can act as localization signals for TA proteins (Figure 26(a)). Most of these TMD variants were membrane-bound, as seen from the TMD variants localized in ER, PM and mitochondria (Figure 26(b), ER, PM and mitochondria). It can also be seen that most TMD variants of all of the Golgi and Peroxisomal TA proteins localize in ER and mitochondria, or their signal cannot be well defined to one specific compartment (ambiguous or puncta). This might indicate the presence of extra localization signals for Golgi and Peroxisomal proteins in regions other than TMDs.

Next, I compared the localization for individual TMD variants from 59 TA proteins (Figure 26(c). 61% of TMD variants are anchored at ER, mitochondria, PM, vacuole or several membrane compartments together (ambiguous). 7 TA proteins showed puncta localization upon 24 different TMD fusions, 3 (Ydl241w, Pgc1, Yjl119c) of which had puncta localization. Out of 59, 10 TA proteins were found to localize in Golgi, 1 in NE and 1 in the peroxisome. TMD variants of these were not correctly classified, which is consistent with our analysis in Figure 26(b). TMD variants of 17 TA proteins did not change compartments upon the addition of flanking residues to the N or C-terminal. These mainly were ER proteins.

The higher hydrophobicity observed in TMD variants that exhibit similar localization to their respective TA proteins, in contrast to TMD variants localizing to different compartments, suggests a role for biophysical properties like hydrophobicity in dictating the localization of these variants (Figure 26(d)). Thus, I checked the variation of hydrophobicity of TMD variants for different subcellular compartments (Figure 26(e)). The variation of hydrophobicity for different compartments reveals that separate subcellular compartments have distinguishable hydrophobicity profiles. For example, mitochondrial TMD fusions have considerably lower hydrophobicity than any other membrane-bound TMDs or cytosol. Additionally, Plasma-membrane localized TMD fusions can be separated into two groups – very high hydrophobicity or very low hydrophobicity.

To evaluate if the localization signals for these TMD variants depend only on hydrophobicity or if any other factors influence the localization, I analyzed the localization of the shuffled TMD variant for any amino acid specificity. The shuffling of the sequence is done on ^{5}TMD^{5} variants. The number of amino acids and TMD length are kept constant in the two shuffled sets before the variant creation. However, the degree of randomization could vary in the two sets created. Prediction by Phobius predictor(Käll et al., 2004) ensured that the shuffled sequences possess

TMD properties, with 95% of the shuffled sequences being TMD. TMD fusions of 5 unique TA proteins did not change localization upon shuffling, out of which 3 were ER, and 2 were localized in multiple membrane compartments (Figure 26(c)). Overall, 73% of TMD variants showing ER and ambiguous localizations retained their localization irrespective of shuffling. In comparison, 80% of mitochondria and PM localized TMD fusions changed to completely different compartments, indicating some compartments have high specificity towards the sequence.

Taken together, TMDs alone can act as localization signals for some TA proteins, depending on the compartments. However, the length of the flanking regions, the TMD sequence, and the biophysical properties, such as hydrophobicity, play a key role in their correct localization.

6.4.1.2 Analysis of TMD variants with charged flanking regions

Analysis of the TMD variants of flanking regions between ^{1}TMD^{1} and ^{5}TMD^{5} reveals approximately 70% of TMD variants change localization upon the addition of flanking residues (Figure 26(c)). The addition of flanking regions to the C-terminus of the TMD variants (^{1}TMD^{1} versus ^{1}TMD^{5},^{5}TMD^{5}) changes or retains the localization to mitochondria for variants of 6 TA proteins (Sec20, Fmp32, Fis1, Tom7, Tom5 and Pex15), suggesting the C-terminal flanking region holds some properties for mitochondrial localization. On the other hand, the addition of flanking regions on the N-terminus of the TMD variants (^{1}TMD^{1}, ^{5}TMD1,^{5}TMD^{5}) changes or retains the localization to PM for variants for 9 TA proteins, suggesting the importance of N-terminal flanking region for PM localization for these variants. Comparing the lengths of the variants with varied flanking regions from section 6.4.1.1 with respect to their localizations shows shorter TMDs tend to localize to mitochondria (Figure 27(a)). PM-localized TMDs can be categorized into two groups based on length – 1. Longer length TMD fusions; 2. Shorter length cysteine-rich TMD fusion. Moreover, a higher positive mean charge is noted in the N-terminal flanking region for a subset of PM-localized TMD variants with varied flanking regions compared to other localizations (Figure 27(b)). Conversely, mitochondrial-localized TMD variants with flanking regions exhibit increased positive charge in the C-terminal flanking region.

Thus, to understand what properties of the N- and C-terminal affects the localization, 5 TMD variants with flanking positive charges are created and analyzed for localization: 1. ^{5}TMD^{RR}: 5 flanking residues before the TMD and 2 R after the TMD, 2. ^{5}TMD^{RRRR}: 5 flanking residue before the TMD and 4 R after the TMD, 3. ^{R}TMD^{1}: 1 R before the TMD and a flanking residue after the TMD, 4. ^{RR}TMD^{1}: 2 R before the TMD and a flanking residue after the TMD, 5. ^{RRRR}TMD^{1}: 4 R before the TMD and a flanking residue after the TMD (Figure 25(b)). Arginine was chosen over lysine to avoid the creation of a ubiquitination site on TMD fusions. This section discusses the effect of the charged flanking regions of the TMD variants on localization.

Compared to full-length TA protein localization, adding high charge at the C-terminus of TMD variants of 8 TA proteins changes their localization to mitochondria (Figure 27(c)). For example,

adding 4 R to the C-terminus of the TMD variant of Far9 and Far10 changed its localization from ER to mitochondria. Overall, 26 TMD variants with added C-terminal R (^{1}TMD^{1}, ^{5}TMD^{5}, ^{5}TMD^{RR}, ^{5}TMD^{RRRR}) localized in mitochondria. All these variants had low hydrophobicity (lower than 2.5, Figure 27(d)). On the other hand, the addition of 4 R at the N-terminus of TMD variants of 18 of 53 TA proteins changed localization to PM or PM+ compared to ^{1}TMD^{1} (Figure 27(c)). 70% of these variants had high hydrophobicity (greater than 2.7, Figure 27(e)). Interestingly, out of the five mitochondrial TA proteins, the addition of N-terminal charge in their TMD variants changes the localization of 4 variants to puncta or cytosol.

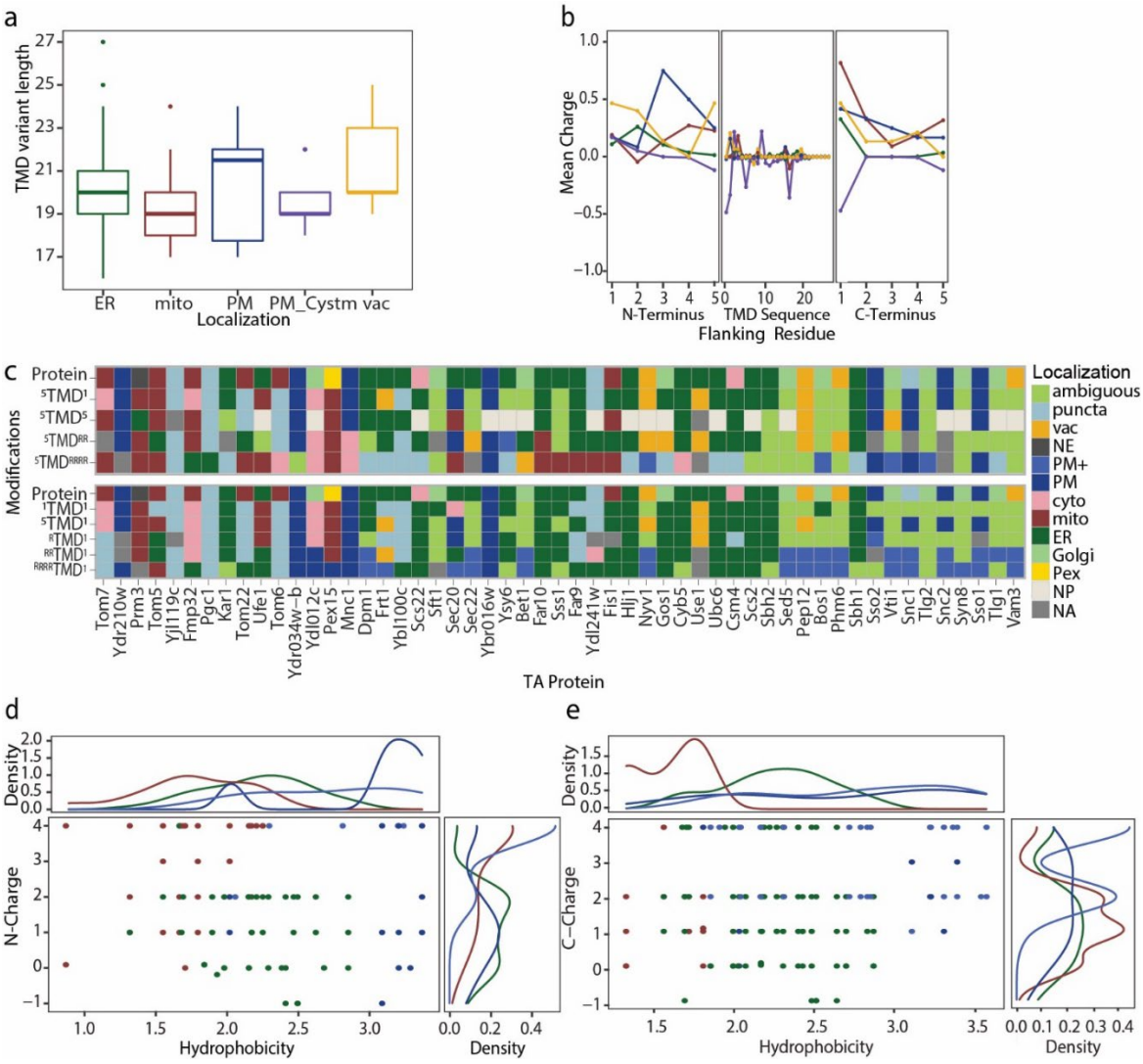

Figure 27: Flanking C-terminal and N-terminal positive charge helps in mitochondrial and PM localization. *(a) Variation of TMD length for different localization in TMD fusion library from Figure 26 (a). (b) Variation of mean charge for TMD variants at N-flanking region, TMD sequence and C-flanking region from TMD variant library in* ***Figure 26*** *(a). (c) Localization of TMD variants with added flanking charge (Figure (c)). ER, endoplasmic reticulum; mito, mitochondria; mixed, several compartments; PM, plasma membrane; vac, vacuole; cyto, cytosol; NE, nuclear envelope; NA, protein variants, which were not identified in the experiment; NP, protein variants without C-terminal extension, which were not included in the experiment. (d, e) Dot plots for comparison of TMD hydrophobicity with the*

*charge of C- (d) or N-terminal (e) library from **Figure 27**(c). The top curves show the distribution of hydrophobicity and the curves on the right show charge distribution.ER, endoplasmic reticulum; mito, mitochondria; mixed, several compartments; PM, plasma membrane; PM+, PM, ER and vacuole; vac, vacuole; cyto, cytosol; NE, nuclear envelope; NA, protein variants, which were not identified in the experiment.*

Taken together, the results suggest low hydrophobicity and positive charge at the C-terminus of TMD variants help in mitochondrial localization. In contrast, high hydrophobicity and positive charge at their N-terminus enables PM localization.

6.4.1.3 Analysis of TMD variants with varied lengths

TMD variants with varied flanking regions show variation in length based on localization (Figure 27(a)). Most of these TMD variants have a length of 18-20 AA. The TMD variants with varied flanking regions with mitochondrial localization are comparatively shorter than other variants in this group. Most ER-localized TMD variants with varied flanking regions have a moderate length, with few exceptionally longer than others. Additionally, the TMD variants with varied flanking residues localized in PM can be grouped into two categories: (a) longer variants (21-24 AA) or short variants (18-19 AA). The shorter TMD variants are enriched in C, while the longer ones are enriched in V. Existing research shows that the V-enriched, highly hydrophobic fungal proteins are known to localize in PM (Sharpe et al., 2010). These results suggest that for certain TMDs, their length affects localizations. Thus, to test the effect length on the localization of TMD variants, four variants with elongated lengths are created and analyzed for localization (Figure 25(c)). Out of these, three variants are elongated by the addition of V at the C-Terminal of TMD while keeping the total TMD length to 22 (22V), 24 (24V) and 26 (26V). Another group of TMD variants is one with alternating VA amino acids at the TMD C-terminus, with a constant length of 26 (26VA). To understand if the position of V plays a role in localization, the 26V TMD variant is shuffled to form two additional variants. The localization of all these variants is then studied using microscopy and automated image analysis. In this section, I present the downstream analysis for the relation between the localizations and the length of TMD length variants.

Comparison of the localizations of variants of TMDs with each other reveals that 60% of variants are localized in ER, PM or vacuole upon elongation (Figure 28(a)). Variants from 13 out of 53 TA proteins localized to PM or PM+ from other compartments upon elongation in at least one of the length variants. Similarly, TMD length variants of 18 and 11 TA proteins localized to ER and vacuole from other compartments upon elongation. TMD variants of Golgi protein BET1 localize to vacuole upon increased V fraction. Also, there are no TMD length variants that localize to mitochondria upon elongation to more than 24 AA, in line with our previous result that mitochondrial proteins are shorter in length. Upon further analysis of the length library, we see that the PM-localized TMD length variants are longer and more hydrophobic than the length variants in mitochondria or ER (Figure 28(b)). PM+ or PM localized elongated TMD variants have higher N-terminal charge than variants from other classes from this group (Figure 28(c)), which

aligns with previous observations that higher N-terminal charge helps in PM localization. This also suggests that localization to the membrane-bound compartment happens due to a combination of multiple properties and not by a single property. Hydrophobicity per position of shuffled TMD length variants does not show a trend that is very different from the non-shuffled ones, suggesting that localization is driven by the V fraction and not the position of V in the TMD (Figure 28(d)).

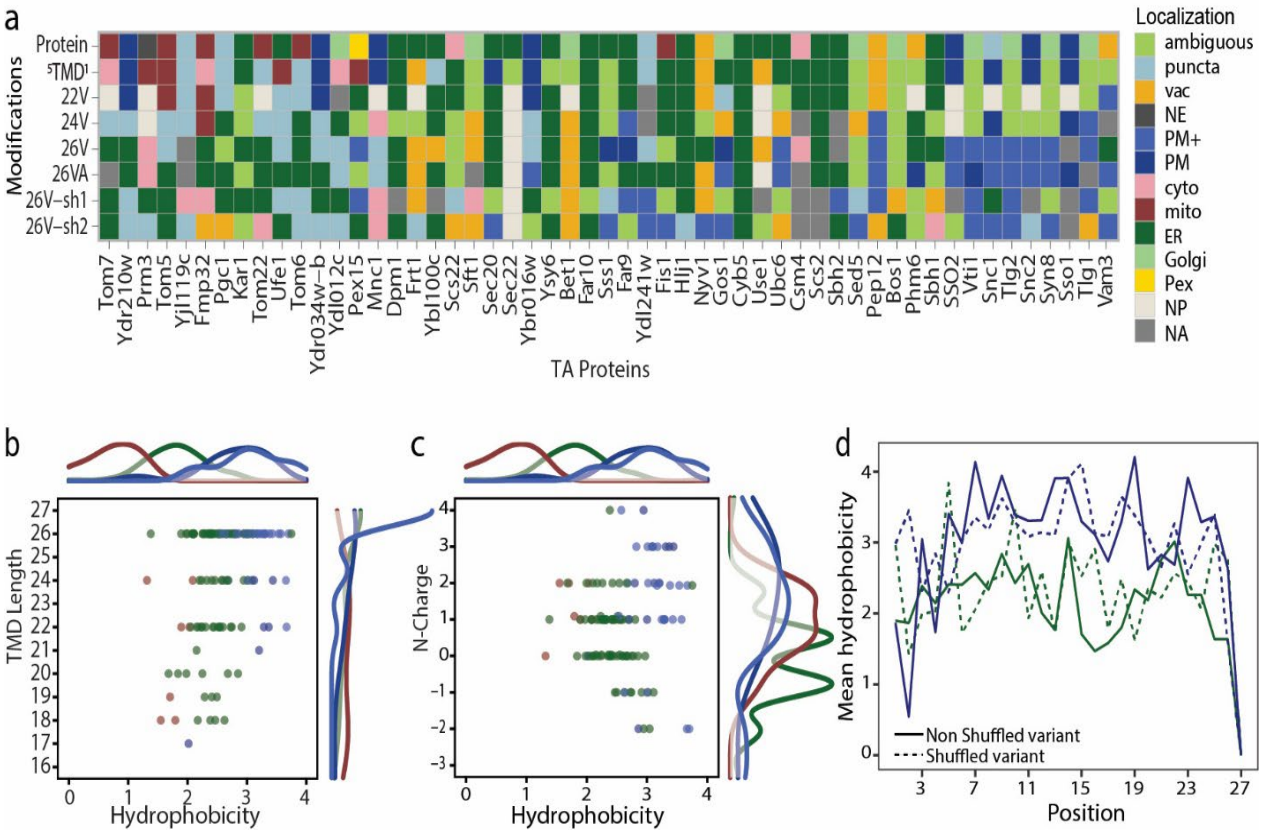

Figure 28: Extension of TMD Length helps in PM, ER and vacuole localization. *The analysis in this panel is for variants: 22V, 24V, 26V, 26VA and two shuffled variants- 26V-sh1 and 26V-sh2. (a) Localization of TMD length variants (Figure 28 (a)). ER, endoplasmic reticulum; mito, mitochondria; mixed, several compartments; PM, plasma membrane; PM+, PM, ER and vacuole; vac, vacuole; cyto, cytosol; NE, nuclear envelope; NA, protein variants, which were not identified in the experiment, (b, c) Dot plots for comparison of TMD hydrophobicity with TMD length (b) or N-terminal charge (c) library from Figure 28 (a). The top curves show the distribution of hydrophobicity, and the curves on the right show TMD length and N charge. (d) Variation of hydrophobicity per position for ER and PM for shuffled and non-shuffled variants.*

Overall, the length plays a crucial role in determining the localization of the TMD variants. Furthermore, while the position of V within TMD length variants appears to have no impact on PM and ER localization, the V fraction notably affects their localization in these compartments.

6.4.2 Plasma Membrane localized TMD variants are enriched in N-terminal aspartic acid and glutamic acid

For a comprehensive understanding of what causes a TMD to localize in a particular compartment, I performed the analysis of different properties on the combined dataset, which contained 864 TMD variants with varied flanking regions (Figure 25(a)), charged flanking regions (Figure 25(b)), and elongated lengths (Figure 25(c)). Comparing the amino acid per position of PM localized TMD variants to all variants from other compartments reveals the enrichment of D and E at the first two positions from the N-terminus of TMD variants (Figure 29(a)). Amino acid C is enriched at all positions in the PM-localized TMD variants, which could be the short cysteine-rich class for PM TMD variants, as discussed earlier. V are enriched more in the C-terminal of PM-localized TMD variants (Figure 29(a)).

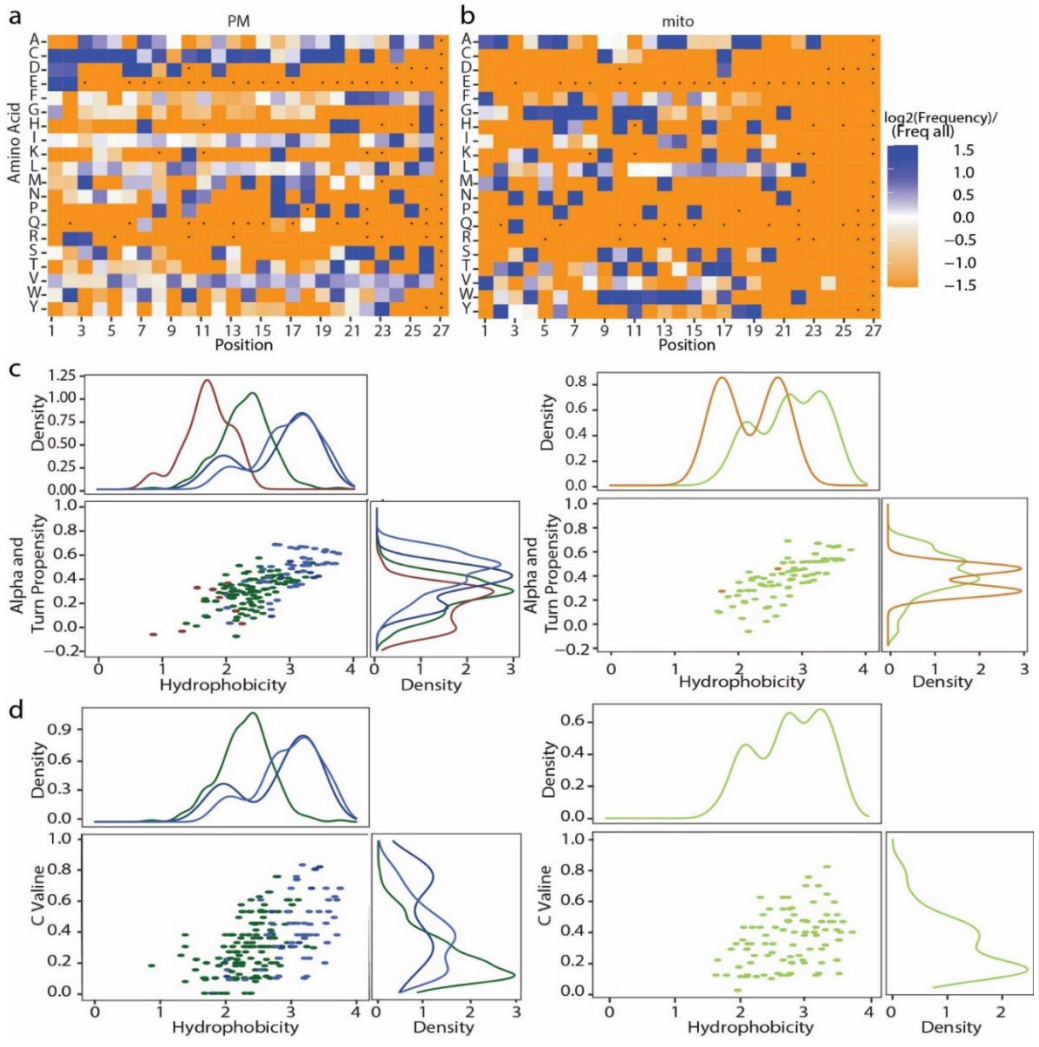

Figure 29: PM-localized TMDs are enriched in N-terminal glutamic acid and aspartic acid*. (a) Normalized Frequency of AA per position for PM localization. (b) Normalized Frequency of AA per position for mitochondria localization. (c,d) Dot plots with density overheads showing the localization of mNG-TMD fusions and the correlation between TMD GRAVY and Alpha and turn propensity (c) and fraction of valine at C-terminal (d). The top curves show the distribution of hydrophobicity, and the curves on the right show Alpha and turn propensity (c) and fraction of valine*

at C-terminal (d) distribution.ER, endoplasmic reticulum; mito, mitochondria; mixed, several compartments; PM, plasma membrane; PM+, PM and vacuole; vac, vacuole; cyto, cytosol; NE, nuclear envelope; NA, protein variants, which were not identified in the experiment; NP, TMD variants with the TMD length more than 22 residues, which were not included in the experiment.

Amino acid enrichment patterns in mitochondria differ significantly from PM. While amino acids A and M are depleted at the N-terminus of the PM-localized TMD variants, they are enriched in mitochondrial TMD variants. Mitochondrial TMD variants are short in length, with no high enrichment of V at their C-terminus. This result was also noted from the length library, where TMD length elongation leads to almost no localization in mitochondria. Mitochondrial TMD variants also have enrichment of W from position 9 to 15 (Figure 29(b)). Apart from lower hydrophobicity compared to PM-localized TMD variants, mitochondrial TMD variants also have lower alpha and turn propensity (Figure 29(c)), indicating the role of other biophysical properties in the localization of TMD variants. 26% of the TMD variants from all libraries localize in the ER. The ER TMD variants show no clear trend in amino acid preference per position. However, they show a lower preference for V at their C-terminus and moderate hydrophobicity, compared to high hydrophobicity and C-terminus V preference of PM-localized TMD variants (Figure 29(d)).

Overall, PM and mitochondrial localized TMD variants have specific AA preferences, while ER-localized TMD variants show trends in biophysical properties.

6.4.3 Combination of multiple biophysical properties of TMD variants helps in understanding their localization

Multiple biophysical properties help in the localization of TMD variants to their respective compartments. To decipher biophysical property specificity for correct localization, I calculated hydrophobicity, AA composition, asymmetry in V and L content, charge of flanking regions and residue volume for all the 864 TMD variants.

Clear trends could be for localizations based on the C-terminal flanking charge, N-terminal flanking Charge, length and hydrophobicity of TMD variants (Figure 30). Mitochondrial TMD variants have positive C-terminal flanking charge and moderate N-terminal flanking charge (Figure 30(a)). They are usually short (mainly less than 20 AA) and have lower hydrophobicity than TMD variants of other membrane organelles. ER-localized TMD variants, on the other hand, require moderate C-terminal and N-terminal flanking charges. However, they are of varied length and moderate hydrophobicity (Figure 30(b).

Two distinct trends are seen in PM-localized TMD variants – 1. Hydrophobic TMD variants, and 2. Cysteine-rich TMD variants (Figure 30(c,d)). Both categories show very distinctive properties. While PM-localized TMD variants are longer (between 22 to 26 AA), cysteine-rich PM TMD variants are shorter (predominantly 19 AA). Cysteine-rich PM TMD variants also have considerably lower N-terminal and C-terminal flanking charges. Although both mitochondrial

TMD variants and cysteine-rich PM TMD variants have low hydrophobicity, they are distinguishable based on the C-terminal flanking charge. Additionally, mitochondrial TMD variants are not enriched in amino acid C).

Overall, these results suggest that a combination of hydrophobicity and the charged flanking residues ensures the targeting of mitochondrial and PM-localized TMD variants. ER targeting requires the TMD variants of medium hydrophobicity and a fraction of TMD variants of medium and high hydrophobicity destined for residence at the vacuole, and Golgi can migrate from the ER to these organelles by a vesicular trafficking route.

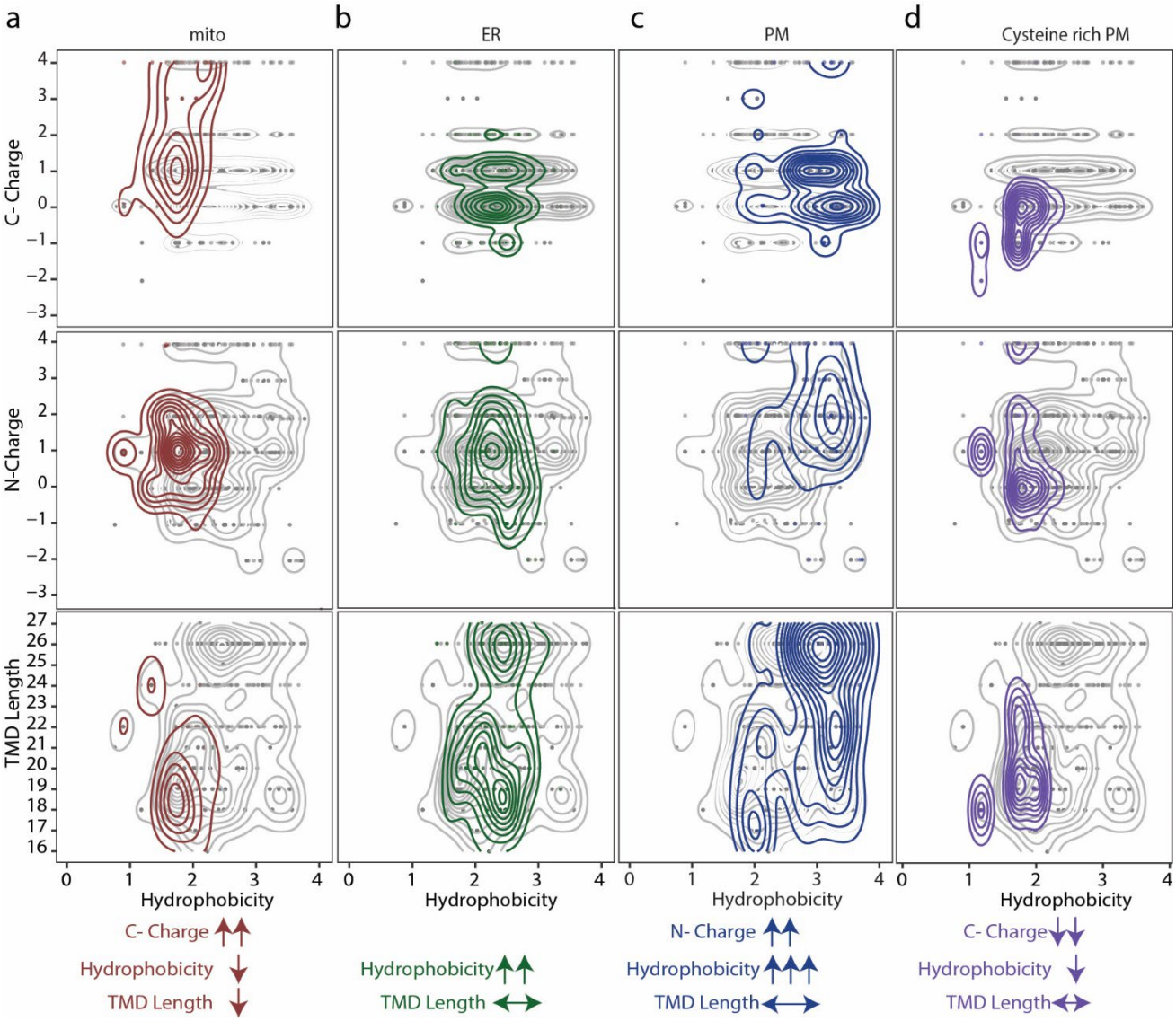

***Figure 30*: Biophysical properties of the TMDs are important for protein localization.** *(a,b,c,d) Automated density plot showing the variation of C-terminal Charge, N-terminal Charge and TMD Length versus hydrophobicity for mitochondria (a), ER(b), PM (c) and cysteine-rich PM (d)*

6.4.4 Random forest helps in predicting the TMD variant localizations

Simple visualization techniques, such as density plots and heatmaps, could only help in analyzing two properties at a time and their effect on localization. Considering that multiple properties interplay to target a protein to its compartment, I used Random Forest to draw out the rules for localization using localization data for all of the 864 TMD variants. Upon comparison of the properties with another, I see polarity and hydrogen bonding are anti-correlated to hydrophobicity, while alpha and turn propensity are correlated to the hydrophobicity of the TMD variants (Figure 31(a)). Avoiding the derived features, I created the Random Forest using C-terminal flanking charge, N-terminal flanking Charge, TMD variant's lengths, hydrophobicity and cysteine fraction in the TMD variants. Two Random Forest models are created – 1. without cysteine-rich TMD variants and 2. with cysteine-rich TMD variants.

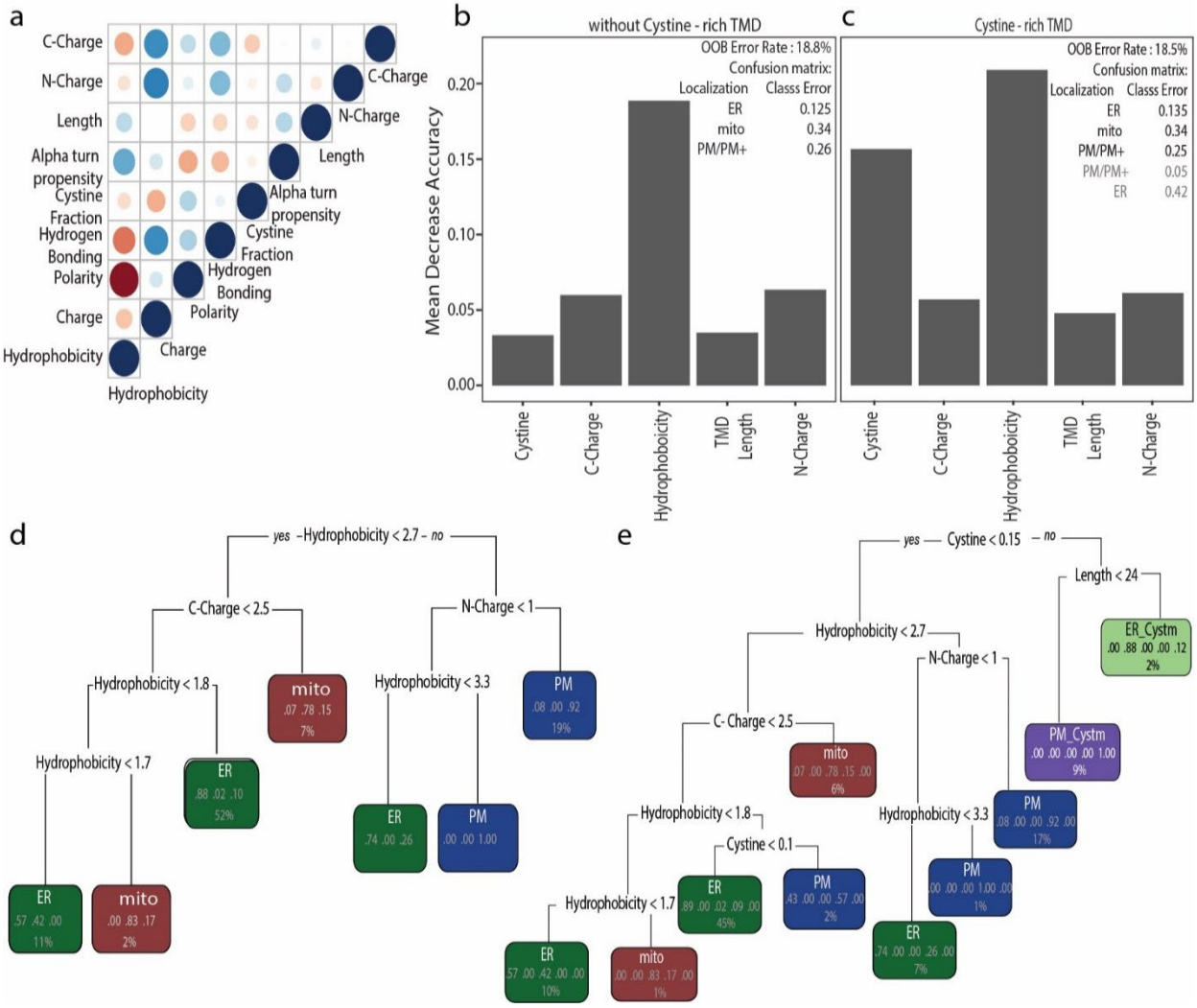

Figure 31: Random Forest helps in predicting TMD localization*. (a) Correlation of multiple biophysical properties of TMD (b,c) Effect of biophysical property on TMD localization. The effect is measured with a mean decrease in accuracy for TMD localization without cysteine (b) and with cysteine (c). (d,e) The tree generated by rules of random forest for the model without cysteine (d) and with cysteine (e).*

Model for TMD variants without cysteine-rich domains classifies TMD variants to their compartments with 82% accuracy, with the best prediction for ER-localized TMD variants. Hydrophobicity plays a key role in identifying correct localization. Mitochondrial localization is predicted with a 0.34 class error rate, possibly because of fewer samples (Figure 31(b)). We see irrespective of N-terminal flanking charge, TMD variants are localized in PM if they have high hydrophobicity (hydrophobicity > 3.3 if N-terminal charge is lower than 1; hydrophobicity > 2.7 otherwise). TMD variants with very low hydrophobicity (between 1.7 and 1.8) and lower C-terminal flanking charge (less than 2.5) localize in mitochondria, while comparatively hydrophobic TMD variants (less than 2.7) but with high C-terminal charge (more than 2.5) also tend to localize in mitochondria. Unlike mitochondrial TMD variants, both hydrophobicity and C-terminal flanking charge are low for ER-localized TMD variants. For high hydrophobicity (between 2.7 and 3.3), if the N-terminal charge is low (less than 1), TMDs localize in ER (Figure 31(c)).

For TMDs enriched in cysteine, classification is made with 82% accuracy and shows the importance of hydrophobicity and cysteine fraction in TMD for correct classification. Cysteine-rich PM and PM+ are classified with an error rate of 0.05. There were only 7 examples of cysteine-rich ER TMD variants, which could have resulted in an error rate of 0.42 (Figure 31(c)). Cysteine-rich PM localized TMD variants have high cysteine fraction (over 0.15) and shorter TMD variants (length < 24, Figure 31(e)).

Overall, with the analysis, we see that random forest could help us generate rules to find quantitatively which TMDs could localize in which compartment.

6.5 Discussion

Transmembrane domains (TMDs) in tail-anchored (TA) proteins are crucial for their correct localization, and mutations in these domains lead to degradation. The biophysical properties of some TMDs help in the correct localization of TA proteins. However, the specificities of localization signals in TMDs towards sub-cellular localization are not fully understood. In this study, iPAL, a deep mutational scanning approach, is used to dissect localization signals for variants of TMDs. Using deep mutational scanning, variants with varied flanking regions, flanking charges and elongated lengths are created, and their localization is first studied through microscopy and later through automated image analysis. In more than 27% of unique proteins, TMD variants localize in the same compartment as their corresponding full-length TA proteins, irrespective of their flanking residue. 68% of the TMD variants are membrane-localized, confirming that TMD alone can act as a localization signal for most TA proteins.

Approximately 22.5% of all TMD variants show the same localization as TA proteins upon the addition of flanking residues to the N and C-terminal, most of which are ER-localized. The change in hydrophobicity is reflected when the TMD variants localize to another compartment upon the addition of flanking regions, indicating the effect of hydrophobicity on the TMD localization. In general, mitochondrial TMD variants are less hydrophobic than variants of other cellular compartments, and the mitochondrial TMD variants localized to other compartments have higher hydrophobicity than the ones localized to mitochondria. Additionally, mitochondrial and PM TMD variants are more susceptible to AA sequence change than ER-localized TMD variants. 73% of TMD variants showing ER and ambiguous localizations retained their localization irrespective of shuffling, while 80% of mitochondria and PM localized TMD variants changed to completely different compartments.

PM localization shows two trends – positive N-terminal charge when hydrophobic and lower N-terminal charge when cysteine-rich. Compared to their corresponding full-length TA protein localization, all PM TMD variants localize correctly upon the addition of an N-terminal flanking charge. Positive C-terminal flanking region helps in mitochondrial localization, which correlates with the previously published data for Fis1 protein – removing or mutating R and K within Fis1 C-terminus led to the anchoring of Fis1 at ER (Rao et al., 2016b).

Besides charge and hydrophobicity, clear trends in TMD variant length affecting localization indicate shorter mitochondrial TMD variants (18-20 AA) and longer PM TMD variants (21-24 AA). TMD elongation studies reveal 60% of variants localized in ER, PM or vacuole upon elongation. The PM-localized TMD variants are longer and more hydrophobic than variants in mitochondria or ER. No TMD variants mislocalize to mitochondria upon elongation to more than 24 AA.

A combined analysis of all TMD variants reveals a preference for certain amino acids for correct localization. While mitochondrial localization prefers four amino acids (A, F, M and V) at the N-terminus, PM prefers amino acids C, D and E. At the second position from the N-terminus, mitochondrial localization prefers four amino acids (A, M, T and Y), while PM localized TMD variants prefer (C, D, E and R). C, acid and E are depleted in the mitochondrial N-terminus. Although R is enriched in the second and third positions of PM-localized TMD variants, it is not enriched at any position in mitochondrial TMD variants. Additionally, mitochondria have lower alpha turn propensity compared to ER or PM, while ER has less V fraction at the C-terminal half of TMD variants than PM. Taken together, sub-cellular localization of TMD variants results from multiple biophysical properties and AA specificity.

Variation of C-terminal flanking charge, N-terminal flanking charge, length and hydrophobicity by simple visualization suggest a preference for high C-terminal charge and low hydrophobicity of short-length TMD variants towards mitochondria. PM localization requires high hydrophobicity of the TMD and high positive N-terminal flanking charge or short cysteine-rich modules of low hydrophobicity. These features might be related to the differences in the membrane thickness and lipid composition for the various organelles (Sharpe et al., 2010; Mitra et al., 2004; Krumpe et al., 2012).

Similar trends can be seen when random forest is used to decipher quantitative rules for correct localization. Overall, PM localization is favored for higher N-flanking charge and high hydrophobicity (> 2.7). A lower N-flanking charge (< 1.1) could still result in PM localization if hydrophobicity is very high (> 3.3). A high C fraction in the TMD variants and a comparatively shorter length could also help in PM localization. Since PM localization has multiple rules, it could be suggestive of multiple pathways of PM targeting. High C-terminal flanking charge and low hydrophobicity help in mitochondrial localization.

6.6 Conclusion

Taken together, biophysical properties such as hydrophobicity, length charge and AA sequence play key roles in targeting the TMD to correct localization, which could target TA protein to their compartment. Although a small set of biophysical properties is analyzed, these already suggest the interplay of many biophysical properties of TMD and flanking regions to correct targeting. The use of advanced computational algorithms such as transfer learning, which might enhance the sample size and thus give better classification for all compartments, might help in creating synthetic TMDs for each localization. This inference could already be useful to create useful experimental hypotheses to decipher pathways and mechanics of how proteins localize and what quality control factors are involved if they are mislocalized.

6.7 Code availability

All codes for the study are available at https://github.com/Susmitha-Nair/SLiMsInLocalization.git.

6.8 References

Adam, S.A., T.J. Lobl, M.A. Mitchell, and L. Gerace. 1989. Identification of specific binding proteins for a nuclear location sequence. *Nature*. 337:276–279. doi:10.1038/337276a0.

Alder, N.N., Y. Shen, J.L. Brodsky, L.M. Hendershot, and A.E. Johnson. 2005. The molecular mechanisms underlying BiP-mediated gating of the Sec61 translocon of the endoplasmic reticulum. *J. Cell Biol.* 168:389–399. doi:10.1083/jcb.200409174.

Almagro Armenteros, J.J., K.D. Tsirigos, C.K. Sønderby, T.N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37:420–423. doi:10.1038/s41587-019-0036-z.

Backes, S., and J.M. Herrmann. 2017. Protein translocation into the intermembrane space and matrix of mitochondria: Mechanisms and driving forces. *Front. Mol. Biosci.* 4:83. doi:10.3389/fmolb.2017.00083.

Bonifacino, J.S., P. Cosson, and R.D. Klausner. 1990. Colocalized transmembrane determinants for ER degradation and subunit assembly explain the intracellular fate of TCR chains. *Cell*. 63:503–513. doi:10.1016/0092-8674(90)90447-m.

Borgese, N., S. Colombo, and E. Pedrazzini. 2003. The tale of tail-anchored proteins. *J. Cell Biol.* 161:1013–1019. doi:10.1083/jcb.200303069.

Boyle, J. 2005. Lehninger principles of biochemistry (4th ed.): Nelson, D., and Cox, M. *Biochem. Mol. Biol. Educ.* 33:74–75. doi:10.1002/bmb.2005.494033010419.

Bradley, K.J., M.R. Bowl, S.E. Williams, B.N. Ahmad, C.J. Partridge, A.L. Patmanidi, A.M. Kennedy, N.Y. Loh, and R.V. Thakker. 2007. Parafibromin is a nuclear protein with a functional monopartite nuclear localization signal. *Oncogene*. 26:1213–1221. doi:10.1038/sj.onc.1209893.

Brocard, C., and A. Hartig. 2006. Peroxisome targeting signal 1: is it really a simple tripeptide? *Biochim. Biophys. Acta*. 1763:1565–1573. doi:10.1016/j.bbamcr.2006.08.022.

Burkhart, J.M., A.A. Taskin, R.P. Zahedi, and F.-N. Vögtle. 2015. Quantitative profiling for substrates of the mitochondrial presequence processing protease reveals a set of nonsubstrate proteins increased upon proteotoxic stress. *J. Proteome Res.* 14:4550–4563. doi:10.1021/acs.jproteome.5b00327.

Call, M.E., J. Pyrdol, M. Wiedmann, and K.W. Wucherpfennig. 2002. The organizing principle in the formation of the T cell receptor-CD3 complex. *Cell*. 111:967–979. doi:10.1016/s0092-8674(02)01194-7.

Calvo, S.E., K.R. Clauser, and V.K. Mootha. 2016. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* 44:D1251-7. doi:10.1093/nar/gkv1003.

Castanzo, D.T., B. LaFrance, and A. Martin. 2020. The AAA+ ATPase Msp1 is a processive protein translocase with robust unfoldase activity. *Proc. Natl. Acad. Sci. U. S. A.* 117:14970–14977. doi:10.1073/pnas.1920109117.

Chacinska, A., C.M. Koehler, D. Milenkovic, T. Lithgow, and N. Pfanner. 2009. Importing mitochondrial proteins: machineries and mechanisms. *Cell*. 138:628–644. doi:10.1016/j.cell.2009.08.005.

Chong, Y.T., J.L.Y. Koh, H. Friesen, S. Kaluarachchi Duffy, M.J. Cox, A. Moses, J. Moffat, C. Boone, and B.J. Andrews. 2015. Yeast proteome dynamics from single cell imaging and automated analysis. *Cell*. 161:1413–1424. doi:10.1016/j.cell.2015.04.051.

Cokol, M., R. Nair, and B. Rost. 2000. Finding nuclear localization signals. *EMBO Rep.* 1:411–415. doi:10.1093/embo-reports/kvd092.

Costello, J.L., I.G. Castro, F. Camões, T.A. Schrader, D. McNeall, J. Yang, E.-A. Giannopoulou, S. Gomes, V. Pogenberg, N.A. Bonekamp, D. Ribeiro, M. Wilmanns, G. Jedd, M. Islinger, and M. Schrader. 2017. Predicting the targeting of tail-anchored proteins to subcellular compartments in mammalian cells. *J. Cell Sci.* doi:10.1242/jcs.200204.

Culver, J.A., X. Li, M. Jordan, and M. Mariappan. 2022. A second chance for protein targeting/folding: Ubiquitination and deubiquitination of nascent proteins. *Bioessays*. 44. doi:10.1002/bies.202200014.

Culver, J.A., and M. Mariappan. 2021. Deubiquitinases USP20/33 promote the biogenesis of tail-anchored membrane proteins. *J. Cell Biol.* 220. doi:10.1083/jcb.202004086.

D'Angelo, M.A., M. Raices, S.H. Panowski, and M.W. Hetzer. 2009. Age-dependent deterioration of nuclear pore complexes causes a loss of nuclear integrity in postmitotic cells. *Cell*. 136:284–295. doi:10.1016/j.cell.2008.11.037.

Dederer, V., A. Khmelinskii, A.G. Huhn, V. Okreglak, M. Knop, and M.K. Lemberg. 2019. Cooperation of mitochondrial and ER factors in quality control of tail-anchored proteins. *Elife*. 8. doi:10.7554/eLife.45506.

Denic, V. 2012. A portrait of the GET pathway as a surprisingly complicated young man. *Trends Biochem. Sci.* 37:411–417. doi:10.1016/j.tibs.2012.07.004.

Di Maio, R., P.J. Barrett, E.K. Hoffman, C.W. Barrett, A. Zharikov, A. Borah, X. Hu, J. McCoy, C.T. Chu, E.A. Burton, T.G. Hastings, and J.T. Greenamyre. 2016. α-Synuclein binds to TOM20 and inhibits mitochondrial protein import in Parkinson's disease. *Sci. Transl. Med.* 8:342ra78-342ra78. doi:10.1126/scitranslmed.aaf3634.

El Magraoui, F., R. Brinkmeier, T. Mastalski, A. Hupperich, C. Strehl, D. Schwerter, W. Girzalsky, H.E. Meyer, B. Warscheid, R. Erdmann, and H.W. Platta. 2019. The deubiquitination of the PTS1-import receptor Pex5p is required for peroxisomal matrix protein import. *Biochim. Biophys. Acta Mol. Cell Res.* 1866:199–213. doi:10.1016/j.bbamcr.2018.11.002.

Fleig, L., N. Bergbold, P. Sahasrabudhe, B. Geiger, L. Kaltak, and M.K. Lemberg. 2012. Ubiquitin-dependent intramembrane rhomboid protease promotes ERAD of membrane proteins. *Mol. Cell*. 47:558–569. doi:10.1016/j.molcel.2012.06.008.

Foresti, O., V. Rodriguez-Vaello, C. Funaya, and P. Carvalho. 2014. Quality control of inner nuclear membrane proteins by the Asi complex. *Science*. 346:751–755. doi:10.1126/science.1255638.

Fu, X., C. Liang, F. Li, L. Wang, X. Wu, A. Lu, G. Xiao, and G. Zhang. 2018. The rules and functions of nucleocytoplasmic shuttling proteins. *Int. J. Mol. Sci.* 19. doi:10.3390/ijms19051445.

Gardy, J.L., and F.S.L. Brinkman. 2006. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4:741–751. doi:10.1038/nrmicro1494.

Gilmore, R., G. Blobel, and P. Walter. 1982. Protein translocation across the endoplasmic reticulum. I. Detection in the microsomal membrane of a receptor for the signal recognition particle. *J. Cell Biol.* 95:463–469. doi:10.1083/jcb.95.2.463.

Görlich, D., and T.A. Rapoport. 1993. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell*. 75:615–630. doi:10.1016/0092-8674(93)90483-7.

Guna, A., N. Volkmar, J.C. Christianson, and R.S. Hegde. 2018. The ER membrane protein complex is a transmembrane domain insertase. *Science*. 359:470–473. doi:10.1126/science.aao3099.

Guo, H., Y. Xiong, P. Witkowski, J. Cui, L.-J. Wang, J. Sun, R. Lara-Lemus, L. Haataja, K. Hutchison, S.-O. Shan, P. Arvan, and M. Liu. 2014. Inefficient translocation of preproinsulin contributes to pancreatic β cell failure and late-onset diabetes. *J. Biol. Chem.* 289:16290–16302. doi:10.1074/jbc.M114.562355.

Habib, S.J., A. Vasiljev, W. Neupert, and D. Rapaport. 2003. Multiple functions of tail-anchor domains of mitochondrial outer membrane proteins. *FEBS Lett.* 555:511–515. doi:10.1016/s0014-5793(03)01325-5.

Hansen, K.G., N. Aviram, J. Laborenz, C. Bibi, M. Meyer, A. Spang, M. Schuldiner, and J.M. Herrmann. 2018. An ER surface retrieval pathway safeguards the import of mitochondrial membrane proteins in yeast. *Science*. 361:1118–1122. doi:10.1126/science.aar8174.

Hasan, M.A.M., S. Ahmad, and M.K.I. Molla. 2017. Protein subcellular localization prediction using multiple kernel learning based support vector machine. *Mol. Biosyst.* 13:785–795. doi:10.1039/c6mb00860g.

Hasle, N., A. Cooke, S. Srivatsan, H. Huang, J.J. Stephany, Z. Krieger, D. Jackson, W. Tang, S. Pendyala, R.J. Monnat Jr, C. Trapnell, E.M. Hatch, and D.M. Fowler. 2020. High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Mol. Syst. Biol.* 16:e9442. doi:10.15252/msb.20209442.

Heck, J.W., S.K. Cheung, and R.Y. Hampton. 2010. Cytoplasmic protein quality control degradation mediated by parallel actions of the E3 ubiquitin ligases Ubr1 and San1. *Proc. Natl. Acad. Sci. U. S. A.* 107:1106–1111. doi:10.1073/pnas.0910591107.

Hegde, R.S., and R.J. Keenan. 2022. The mechanisms of integral membrane protein biogenesis. *Nat. Rev. Mol. Cell Biol.* 23:107–124. doi:10.1038/s41580-021-00413-2.

Hegde, R.S., and E. Zavodszky. 2019. Recognition and degradation of mislocalized proteins in health and disease. *Cold Spring Harb. Perspect. Biol.* 11:a033902. doi:10.1101/cshperspect.a033902.

Hoover, B.R., M.N. Reed, J. Su, R.D. Penrod, L.A. Kotilinek, M.K. Grant, R. Pitstick, G.A. Carlson, L.M. Lanier, L.-L. Yuan, K.H. Ashe, and D. Liao. 2010. Tau mislocalization to dendritic spines mediates synaptic dysfunction independently of neurodegeneration. *Neuron*. 68:1067–1081. doi:10.1016/j.neuron.2010.11.030.

Itzhak, D.N., S. Tyanova, J. Cox, and G.H.H. Borner. 2016. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*. 5. doi:10.7554/elife.16950.

Ivanova, E. 2022. Dissecting tail-anchored protein localization signals with iPAL (imaging pooled-to-array library) scanning. Johannes Gutenberg-Universität Mainz.

Jiang, Y., D. Wang, W. Wang, and D. Xu. 2021. Computational methods for protein localization prediction. *Comput. Struct. Biotechnol. J.* 19:5834–5844. doi:10.1016/j.csbj.2021.10.023.

Joshi, J.R., V. Singh, and H. Friedman. 2020. Arabidopsis cysteine-rich trans-membrane module (CYSTM) small proteins play a protective role mainly against heat and UV stresses. *Funct. Plant Biol.* 47:195. doi:10.1071/fp19236.

Juszkiewicz, S., and R.S. Hegde. 2018. Quality control of orphaned proteins. *Mol. Cell*. 71:443–457. doi:10.1016/j.molcel.2018.07.001.

Kaleel, M., L. Ellinger, C. Lalor, G. Pollastri, and C. Mooney. 2021. SCLpred-MEM: Subcellular localization prediction of membrane proteins by deep N-to-1 convolutional neural networks. *Proteins*. 89:1233–1239. doi:10.1002/prot.26144.

Käll, L., A. Krogh, and E.L.L. Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338:1027–1036. doi:10.1016/j.jmb.2004.03.016.

Keskin, A., E. Akdoğan, and C.D. Dunn. 2017. Evidence for amino acid snorkeling from a high-resolution,*in vivo*analysis of Fis1 tail-anchor insertion at the mitochondrial outer membrane. *Genetics*. 205:691–705. doi:10.1534/genetics.116.196428.

Khmelinskii, A., E. Blaszczak, M. Pantazopoulou, B. Fischer, D.J. Omnus, G. Le Dez, A. Brossard, A. Gunnarsson, J.D. Barry, M. Meurer, D. Kirrmaier, C. Boone, W. Huber, G. Rabut, P.O. Ljungdahl, and M. Knop. 2014. Protein quality control at the inner nuclear membrane. *Nature*. 516:410–413. doi:10.1038/nature14096.

Krumpe, K., I. Frumkin, Y. Herzig, N. Rimon, C. Özbalci, B. Brügger, D. Rapaport, and M. Schuldiner. 2012. Ergosterol content specifies targeting of tail-anchored proteins to mitochondrial outer membranes. *Molecular Biology of the Cell*. 23:3927–3935. doi:10.1091/mbc.E11-12-0994.

Kumar, R., B. Kumari, and M. Kumar. 2017. Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. *Mitochondrion*. doi:10.1016/j.mito.2017.10.004.

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132. doi:10.1016/0022-2836(82)90515-0.

Lange, A., R.E. Mills, C.J. Lange, M. Stewart, S.E. Devine, and A.H. Corbett. 2007a. Classical nuclear localization signals: Definition, function, and interaction with importin α. *J. Biol. Chem.* 282:5101–5105. doi:10.1074/jbc.r600026200.

Lange, A., R.E. Mills, C.J. Lange, M. Stewart, S.E. Devine, and A.H. Corbett. 2007b. Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* 282:5101–5105. doi:10.1074/jbc.R600026200.

Lee, Y.-K., J.W. Brewer, R. Hellman, and L.M. Hendershot. 1999. BiP and immunoglobulin light chain cooperate to control the folding of heavy chain and ensure the fidelity of immunoglobulin assembly. *Mol. Biol. Cell*. 10:2209–2219. doi:10.1091/mbc.10.7.2209.

Lerner, M., M. Corcoran, D. Cepeda, M.L. Nielsen, R. Zubarev, F. Pontén, M. Uhlén, S. Hober, D. Grandér, and O. Sangfelt. 2007. The RBCC gene*RFP2*(*Leu5*) encodes a novel transmembrane E3 ubiquitin ligase involved in ERAD. *Mol. Biol. Cell*. 18:1670–1682. doi:10.1091/mbc.e06-03-0248.

Li, L., J. Zheng, X. Wu, and H. Jiang. 2019. Mitochondrial AAA-ATPase Msp1 detects mislocalized tail-anchored proteins through a dual-recognition mechanism. *EMBO Rep.* 20. doi:10.15252/embr.201846989.

Linxweiler, M., B. Schick, and R. Zimmermann. 2017. Let's talk about Secs: Sec61, Sec62 and Sec63 in signal transduction, oncology and personalized medicine. *Signal Transduct. Target. Ther.* 2:17002. doi:10.1038/sigtrans.2017.2.

Liu, K., J. Chen, F. Yang, Z. Zhou, Y. Liu, Y. Guo, H. Hu, H. Gao, H. Li, W. Zhou, B. Qin, and Y. Wang. 2019. BJ-B11, an Hsp90 inhibitor, constrains the proliferation and invasion of breast cancer cells. *Front. Oncol.* 9:1447. doi:10.3389/fonc.2019.01447.

Liu, Z., L.D. Lavis, and E. Betzig. 2015. Imaging live-cell dynamics and structure at the single-molecule level. *Mol. Cell*. 58:644–659. doi:10.1016/j.molcel.2015.02.033.

Lu, J., T. Wu, B. Zhang, S. Liu, W. Song, J. Qiao, and H. Ruan. 2021. Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Commun. Signal.* 19:60. doi:10.1186/s12964-021-00741-y.

Lyman, S.K., and R. Schekman. 1995. Interaction between BiP and Sec63p is required for the completion of protein translocation into the ER of Saccharomyces cerevisiae. *J. Cell Biol.* 131:1163–1171. doi:10.1083/jcb.131.5.1163.

Mariappan, M., X. Li, S. Stefanovic, A. Sharma, A. Mateja, R.J. Keenan, and R.S. Hegde. 2010. A ribosome-associating factor chaperones tail-anchored membrane proteins. *Nature*. 466:1120–1124. doi:10.1038/nature09296.

Matsumoto, S., K. Nakatsukasa, C. Kakuta, Y. Tamura, M. Esaki, and T. Endo. 2019. Msp1 clears mistargeted proteins by facilitating their transfer from mitochondria to the ER. *Mol. Cell*. 76:191-205.e10. doi:10.1016/j.molcel.2019.07.006.

Mattiazzi Usaj, M., E.B. Styles, A.J. Verster, H. Friesen, C. Boone, and B.J. Andrews. 2016. High-content screening for quantitative cell biology. *Trends Cell Biol.* 26:598–611. doi:10.1016/j.tcb.2016.03.008.

McQuown, A.J., D. Reif, and V. Denic. 2021. A TRCky TA protein delivery service snubs the UPS. *J. Cell Biol.* 220. doi:10.1083/jcb.202103196.

Meng, C., X. Zhao, and J. Lao. 2018. A modified immunofluorescence in situ hybridization method to detect long non-coding RNAs and proteins in frozen spinal cord sections. *Exp. Ther. Med.* 15:4623–4628. doi:10.3892/etm.2018.6046.

Mir, R., and J. León. 2014. Pathogen and circadian controlled 1 (PCC1) protein is anchored to the plasma membrane and interacts with subunit 5 of COP9 signalosome in Arabidopsis. *PLoS One*. 9:e87216. doi:10.1371/journal.pone.0087216.

Mitra, K., and J. Frank. 2006. A model for co-translational translocation: ribosome-regulated nascent polypeptide translocation at the protein-conducting channel. *FEBS Lett.* 580:3353–3360. doi:10.1016/j.febslet.2006.05.019.

Mitra, K., I. Ubarretxena-Belandia, T. Taguchi, G. Warren, and D.M. Engelman. 2004. Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *PNAS*. 101:4083–4088. doi:10.1073/PNAS.0307332101.

Morgenstern, M., S.B. Stiller, P. Lübbert, C.D. Peikert, S. Dannenmaier, F. Drepper, U. Weill, P. Höß, R. Feuerstein, M. Gebert, M. Bohnert, M. van der Laan, M. Schuldiner, C. Schütze, S. Oeljeklaus, N. Pfanner, N. Wiedemann, and B. Warscheid. 2017. Definition of a high-confidence mitochondrial proteome at quantitative scale. *Cell Rep.* 19:2836–2852. doi:10.1016/j.celrep.2017.06.014.

Nguyen Ba, A.N., A. Pogoutse, N. Provart, and A.M. Moses. 2009. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics*. 10:202. doi:10.1186/1471-2105-10-202.

Orre, L.M., M. Vesterlund, Y. Pan, T. Arslan, Y. Zhu, A. Fernandez Woodbridge, O. Frings, E. Fredlund, and J. Lehtiö. 2019. SubCellBarCode: Proteome-wide mapping of protein localization and relocalization. *Mol. Cell*. 73:166-182.e7. doi:10.1016/j.molcel.2018.11.035.

Osafune, T., and S.D. Schwartzbach. 2007. Intracellular protein localization by immunoelectron microscopy. *In* Protein Targeting Protocols, Second Edition. Humana Press, New Jersey. 407–416.

Osorio, D., P. Rondón-Villarreal, and R. Torres. 2015. Peptides: A package for data mining of antimicrobial peptides. *R J.* 7:4. doi:10.32614/rj-2015-001.

Ott, C.M., and V.R. Lingappa. 2002. Integral membrane protein biosynthesis: why topology is hard to predict. *J. Cell Sci.* 115:2003–2009. doi:10.1242/jcs.115.10.2003.

Park, S.-H., N. Bolender, F. Eisele, Z. Kostova, J. Takeuchi, P. Coffino, and D.H. Wolf. 2007. The cytoplasmic Hsp70 chaperone machinery subjects misfolded and endoplasmic reticulum import-incompetent proteins to degradation via the ubiquitin–proteasome system. *Mol. Biol. Cell*. 18:153–165. doi:10.1091/mbc.e06-04-0338.

Pereira Mendes, M., R. Hickman, M.C. Van Verk, N.M. Nieuwendijk, A. Reinstädler, R. Panstruga, C.M.J. Pieterse, and S.C.M. Van Wees. 2021. A family of pathogen-induced cysteine-rich transmembrane proteins is involved in plant disease resistance. *Planta*. 253. doi:10.1007/s00425-021-03606-3.

Pidasheva, S., L. Canaff, W.F. Simonds, S.J. Marx, and G.N. Hendy. 2005. Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. *Hum. Mol. Genet.* 14:1679–1690. doi:10.1093/hmg/ddi176.

Pool, M.R. 2022. Targeting of proteins for translocation at the endoplasmic reticulum. *Int. J. Mol. Sci.* 23:3773. doi:10.3390/ijms23073773.

Prasad, R., S. Kawaguchi, and D.T.W. Ng. 2010. A nucleus-based quality control mechanism for cytosolic proteins. *Mol. Biol. Cell*. 21:2117–2127. doi:10.1091/mbc.e10-02-0111.

Rao, M., V. Okreglak, U.S. Chio, H. Cho, P. Walter, and S.-O. Shan. 2016a. Multiple selection filters ensure accurate tail-anchored membrane protein targeting. *Elife*. 5. doi:10.7554/elife.21301.

Rao, M., V. Okreglak, U.S. Chio, H. Cho, P. Walter, and S.O. Shan. 2016b. Multiple selection filters ensure accurate tail-anchored membrane protein targeting. *eLife*. 5:e21301. doi:10.7554/ELIFE.21301.

Rapoport, T.A. 2007. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*. 450:663–669. doi:10.1038/nature06384.

Rodrigo-Brenni, M.C., E. Gutierrez, and R.S. Hegde. 2014. Cytosolic quality control of mislocalized proteins requires RNF126 recruitment to Bag6. *Mol. Cell*. 55:227–237. doi:10.1016/j.molcel.2014.05.025.

Ruggiano, A., O. Foresti, and P. Carvalho. 2014. Quality control: ER-associated degradation: protein quality control and beyond. *J. Cell Biol.* 204:869–879. doi:10.1083/jcb.201312042.

Savojardo, C., P.L. Martelli, P. Fariselli, and R. Casadio. 2018. DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*. 34:1690–1696. doi:10.1093/bioinformatics/btx818.

Schuldiner, M., J. Metz, V. Schmid, V. Denic, M. Rakwalska, H.D. Schmitt, B. Schwappach, and J.S. Weissman. 2008. The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell*. 134:634–645. doi:10.1016/j.cell.2008.06.025.

Sharpe, H.J., T.J. Stevens, and S. Munro. 2010. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell*. 142:158–169. doi:10.1016/j.cell.2010.05.037.

Shen, Y., Y. Ding, J. Tang, Q. Zou, and F. Guo. 2020. Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform.* 21:1628–1640. doi:10.1093/bib/bbz106.

Shurtleff, M.J., D.N. Itzhak, J.A. Hussmann, N.T. Schirle Oakdale, E.A. Costa, M. Jonikas, J. Weibezahn, K.D. Popova, C.H. Jan, P. Sinitcyn, S.S. Vembar, H. Hernandez, J. Cox, A.L. Burlingame, J.L. Brodsky, A. Frost, G.H.H. Borner, and J.S. Weissman. 2018. The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *Elife*. 7. doi:10.7554/elife.37018.

Stornaiuolo, M., L.V. Lotti, N. Borgese, M.-R. Torrisi, G. Mottola, G. Martire, and S. Bonatti. 2003. KDEL and KKXX retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and Golgi complex. *Mol. Biol. Cell*. 14:889–902. doi:10.1091/mbc.e02-08-0468.

Szczesny, B., T.K. Hazra, J. Papaconstantinou, S. Mitra, and I. Boldogh. 2003. Age-dependent deficiency in import of mitochondrial DNA glycosylases required for repair of oxidatively damaged bases. *Proc. Natl. Acad. Sci. U. S. A.* 100:10670–10675. doi:10.1073/pnas.1932854100.

Terry, L.J., and S.R. Wente. 2009. Flexible gates: dynamic topologies and functions for FG nucleoporins in nucleocytoplasmic transport. *Eukaryot. Cell*. 8:1814–1827. doi:10.1128/EC.00225-09.

Venancio, T.M., and L. Aravind. 2010. CYSTM, a novel cysteine-rich transmembrane module with a role in stress tolerance across eukaryotes. *Bioinformatics*. 26:149–152. doi:10.1093/bioinformatics/btp647.

Walter, P., I. Ibrahimi, and G. Blobel. 1981. Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to in-vitro-assembled polysomes synthesizing secretory protein. *J. Cell Biol.* 91:545–550. doi:10.1083/jcb.91.2.545.

Wang, F., E.C. Brown, G. Mak, J. Zhuang, and V. Denic. 2010. A chaperone cascade sorts proteins for posttranslational membrane insertion into the endoplasmic reticulum. *Mol. Cell*. 40:159–171. doi:10.1016/j.molcel.2010.08.038.

Wang, X., and S. Li. 2014. Protein mislocalization: mechanisms, functions and clinical applications in cancer. *Biochim. Biophys. Acta*. 1846:13–25. doi:10.1016/j.bbcan.2014.03.006.

Weidberg, H., and A. Amon. 2018. MitoCPR—A surveillance pathway that protects mitochondria in response to protein import stress. *Science*. 360:eaan4146. doi:10.1126/science.aan4146.

Wilkinson, L. 2011. Ggplot2: Elegant graphics for data analysis by WICKHAM, H. *Biometrics*. 67:678–679. doi:10.1111/j.1541-0420.2011.01616.x.

Yu, C.-S., C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, and C.-H. Lu. 2014. CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PLoS One*. 9:e99368. doi:10.1371/journal.pone.0099368.

Zimmermann, R., S. Eyrisch, M. Ahmad, and V. Helms. 2011. Protein translocation across the ER membrane. *Biochim. Biophys. Acta Biomembr.* 1808:912–924. doi:10.1016/j.bbamem.2010.06.015.

7 Conclusion

The work entails the recognition of SLiMs (degrons and targeting signals), which are important for Protein Quality Control. Short degron motifs could be interpreted using simple visualization techniques. For longer sequences, sophisticated methods are employed while maintaining good predictability to decipher the motifs. Models such as fully connected neural networks hold a good capacity to predict degrons with high accuracy. Porting SHAP onto the model provided how much the sequence and biophysical properties of each amino acid and position contributed towards degradation. This thereby helped in defining sequence motifs. Using the interpretable deep learning model, C-terminal and internal degradation motifs are recognized. Furthermore, the corresponding E3 ligases and the important factors are found using the pipeline, leading to 5 Das1 motifs. These are experimentally validated as well. Using simple visualization techniques, biophysical properties that highlight targeting into specific compartments are found. Additionally, Random Forest is used to decipher the rules for organelle-specific targeting.

8 References

Bruning, J.B., and Y. Shamoo. 2004. Structural and thermodynamic analysis of human PCNA with peptides derived from DNA polymerase-δ p66 subunit and flap endonuclease-1. *Structure*. 12:2209–2219. doi:10.1016/j.str.2004.09.018.

Davey, N.E., K. Van Roey, R.J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T.J. Gibson. 2012. Attributes of short linear motifs. *Mol. Biosyst.* 8:268–281. doi:10.1039/c1mb05231d.

Dinkel, H., K. Van Roey, S. Michael, N.E. Davey, R.J. Weatheritt, D. Born, T. Speck, D. Krüger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L.B. Chemes, J. Glavina, I.E. Sánchez, F. Diella, and T.J. Gibson. 2014. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* 42:D259-66. doi:10.1093/nar/gkt1047.

Edwards, R.J., and N. Palopoli. 2015. Computational prediction of short linear motifs from protein sequences. *In* Methods in Molecular Biology. Springer New York, New York, NY. 89–141.

Guo, H., Y. Xiong, P. Witkowski, J. Cui, L.-J. Wang, J. Sun, R. Lara-Lemus, L. Haataja, K. Hutchison, S.-O. Shan, P. Arvan, and M. Liu. 2014. Inefficient translocation of preproinsulin contributes to pancreatic β cell failure and late-onset diabetes. *J. Biol. Chem.* 289:16290–16302. doi:10.1074/jbc.M114.562355.

Hegde, R.S., and E. Zavodszky. 2019. Recognition and degradation of mislocalized proteins in health and disease. *Cold Spring Harb. Perspect. Biol.* 11:a033902. doi:10.1101/cshperspect.a033902.

Mohan, A., C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, and V.N. Uversky. 2006. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362:1043–1059. doi:10.1016/j.jmb.2006.07.087.

Pidasheva, S., L. Canaff, W.F. Simonds, S.J. Marx, and G.N. Hendy. 2005. Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. *Hum. Mol. Genet.* 14:1679–1690. doi:10.1093/hmg/ddi176.

Timms, R.T., and I. Koren. 2020. Tying up loose ends: the N-degron and C-degron pathways of protein degradation. *Biochem. Soc. Trans.* 48:1557–1567. doi:10.1042/BST20191094.

Tompa, P., M. Fuxreiter, C.J. Oldfield, I. Simon, A.K. Dunker, and V.N. Uversky. 2009. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*. 31:328–335. doi:10.1002/bies.200800151.

Van Roey, K., B. Uyar, R.J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T.J. Gibson, and N.E. Davey. 2014. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* 114:6733–6778. doi:10.1021/cr400585q.

9 Appendix I

9.1 List of packages used

Programming Language: R
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

Locale:

1. LC_COLLATE=English_United States.1252
2. LC_CTYPE=English_United States.1252
3. LC_MONETARY=English_United States.1252
4. LC_NUMERIC=C
5. LC_TIME=English_United States.1252

Attached base packages:

1. stats4
2. grid
3. stats
4. graphics
5. grDevices utils
6. datasets
7. methods
8. base

Attached packages:

1. CORElearn_1.56.0
2. cowplot_1.1.1
3. dummies_1.5.6
4. varImp_0.4
5. party_1.3-9
6. strucchange_1.5-2
7. sandwich_3.0-2
8. zoo_1.8-10

9. modeltools_0.2-23
10. mvtnorm_1.1-3
11. measures_0.3
12. randomForest_4.6-14
13. reshape2_1.4.4
14. patchwork_1.1.2
15. reshape_0.8.9
16. dplyr_1.0.7
17. ggpubr_0.4.0
18. corrplot_0.92
19. Peptides_2.4.4
20. seqinr_4.2-8
21. stringi_1.7.6
22. stringr_1.4.1
23. plotly_4.10.2
24. ggplot2_3.3.6

9.2 Abbreviations

AA: Amino acid

cNLS: classical Nuclear Localization Signal

Cystm: Cysteine-rich transmembrane protein

CS: Contribution Score

DMS: Deep Mutational Scanning

DNA: deoxyribonucleotide

FACS: Fluorescence-activated cell sorting

GET: guided entry of TA protein

IM: Inner Membrane

iMet: initiator Methionine

GPS: Global Protein Stability

iPAL: imaging pooled-to-arrayed libraries

MPP: Mitochondrial Processing Peptidases

MetAPs: methionine aminopeptidases

MLS: Mitochondrial Localization Signal

NLS: Nuclear Localization Signal

NPC: Nuclear Pore Complex

ER: Endoplasmic reticulum

ERAD: ER-associated degradation

FACS: Fluorescence-Activated Cell Sorting

MPS: Multiplexed Stability Profiling

ncNLS: non-classical Nuclear Localization Signal

OM: Outer Membrane

OOB: Out-of-bag

PCR: Polymerase chain reaction

potential PI index: potential Protein Interaction index

PM: Plasma Membrane

PQC: Protein Quality Control

PSI: Protein Stability Index

PTM: Post-translational modifications

QC: Quality Control

RNC: Ribosome-Nascent Chain

SRP: Signal Recognition Pathway

SCF: Skp, Cullin, F-Box containing complex

SHAP: Shapely Additive explanation

SLiMs: Short Linear Motifs

TA: Tail-anchored

TCR: T-Cell Receptor

TOM: Translocase of Outer Membrane

TIM: Translocase of Inner Membrane

TMD: Transmembrane domain

UPS: Ubiquitin Proteasome System

Amino Acids

A: Alanine

C: Cysteine

D: Aspartic acid

E: Glutamic acid

F: Phenylalanine

G: Glycine

H: Histidine

I: Isoleucine

K: Lysine

L: Leucine

M: Methionine

N: Asparagine

P: Proline

Q: Glutamine

R: Arginine

S: Serine

T: Threonine

V: Valine

W: Tryptophan

Y: Tyrosine

10 Appendix II

10.1 Acknowledgement

10.2 Curriculum Vitae