# Mechanistic Studies of Tail-anchored Membrane Protein Targeting to the ER

Thesis by
Michelle Yen Fry

In Partial Fulfillment of the Requirements for the
Degree of
Doctorate in Philosophy


CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended July 26th, 2021

Michelle Yen Fry
ORCID: 0000-0002-3209-5492

# ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Professor Bil Clemons. Under his mentorship I have matured both as a scientist and as a functioning human. He fostered an environment of encouragement and scientific freedom that motivated me. Not only did he provide me with the resources to learn new techniques, but he was patient with me as I struggled to grasp them. I am grateful that he agreed to take me on as a student and wouldn't be on my current trajectory without him.

I would like to thank my committee members: Dr. Shu-ou Shan, Dr. Pamela Bjorkman, and Dr. Rebecca Voorhees, for challenging me, especially early on in my career. The advice garnered from them help push my research forward and strengthen my findings. As women, they provided sound advice moving forward in my career.

My scientific career began at Brandeis University in the lab of Dr. Christopher Miller, he took in a freshman biophysics major and nurtured my curiosity for understanding protein function. Dr. Janice Robertson was the first person to mentor me and taught me how to do pretty much everything from balancing centrifuges to using pipettes. Without her, I would likely be doing math problems somewhere. Later, I discovered enzymes and crystallography in Dr. Douglas Theobald's lab under the mentorship of Dr. Jeffrey Boucher.

My growth and experience would have been incomplete without the support and teachings of other Clemons Lab members. Especially Shyam Saladi, who has helped me with everything related to computers as well as provided advice and support throughout my time at Caltech. GET team members – Dr. Geoffrey Lin, Dr. Amanda Mock, Dr. Aye Myatt Thinn, Alex Barbato, and Victor Ruiz, and even those who I've never overlapped with, Dr. Justin Chartron and Dr. Harry Gristick, helped me navigate the world of TA protein. To the initial cryo-EM team: Dr. Nadia Riera and Dr. Hyun Gi Yun, it was a pleasure stumbling through and learning this technique with you. The crystallography work presented here wouldn't have been possible without the help and constant support of Ailiena Maggiolo and Dr. Jens Kaiser. Members of the Rees lab – Dr. Naima Sharaf, Allen Lee, and Dr. ChengCheng Fan, have also been indispensable to my experience at Caltech. I wouldn't have been able to do single particle cryo-EM without wisdom from Dr. Alasdair McDowall, Dr. Andrey Malyutin, and Dr. Songye Chen. Reeti Gulati, Rita

Askenfield, and Vanessa Mechem were incredible undergrads who not only helped propel my research forward, but also taught me how to be a mentor.

To my friends, Dr. Daria Ameri, Dr. Danielle Dressler, Elle Braun, Stephanie Chang, and Morgan and Steven Gray, whose encouragement never faltered. Ally Rennell, thank you for proofreading my writing.

To my husband, Ryan: this thesis is as much your accomplishment as it is mine. Your constant support, proofreading, and encouragement when things got hard is why this got done. Thank you for dealing with the unpredictable schedule and nontraditional hours science has brought to our lives.

Lastly, and most importantly, I have to thank my family. My parents, Binh Tran and Colin Fry – you two have always encouraged my curiosity from buying me my first microscope to never getting angry when I broke all our electronics – taking them apart to see how they were made. Mom, you are the toughest person I know and my perseverance comes from you and your mantra "No choice." Kevin Fry, having you as a younger brother has always pushed me. Thank you for helping me sort out my data and ways to interpret the numbers. You'll be my statistician for life.

# ABSTRACT

The successful biogenesis – synthesis, delivery, and insertion into designated membranes – of membrane proteins is a crucial cellular process. One particular class of membrane proteins, tail-anchored (TA) proteins have a single transmembrane domain (TMD) that this located at their C-termini and are targeted to membranes post-translationally. Multiple pathways have been identified to target TA proteins to the ER membranes, but designated pathways for targeting TA proteins to the mitochondria remain elusive. The most well understood ER TA protein pathway is the Guided Entry of Tail-anchored proteins (GET) pathway, consisting of six (fungal) or seven (metazoans) proteins, SGTA, Get1-5, and Bag6 (metazoans only), has nearly been studied exclusively in Opisthokants (fungi and metazoans). Here we employed a combination of x-ray crystallography, cryo-electron microscopy, computational modeling, cellular biology, fluorescent imaging, and bioinformatics in order to understand the underlying factors that regulate the targeting of these TA proteins to their correct membranes. Our work reveals that ER-bound TA proteins tend to have a hydrophobic face whereas mitochondria-bound TA proteins contain a charge following their TMD. This finding corroborates our observation that the first component of the GET pathway to interact with TA proteins, SGTA, falls in a category of other hydrophobic segment binding domains, dubbed STI1-domains. Structures presented here demonstrate that the overall structure of Get3 is conserved in organisms as distant as Excavats and Opistokonts, and slight conformational changes in the ATPase allows the described chaperone cascade of the GET pathway to progress. Together these results refine the model for TA protein targeting to the ER membrane.

# PUBLISHED CONTENT AND CONTRIBUTIONS

Fry, Michelle Y, Vladimíra Najdrová, et al. (2021). "The conformational changes during the catalytic cycle of the tail-anchor targeting chaperone Get3 based on structures from a human pathogen". *Submitted*.
Contributions: M.Y. Fry collected and refined X-ray diffraction data sets and collected and processed the single particle cryo-electron microscopy data sets. M.Y. Fry participated in the design and execution of biochemical experiments.

Fry, Michelle Y, Shyam M Saladi, and William M Clemons Jr (2021). "The STI1-domain is a flexible alpha-helical fold with a hydrophobic groove". In: *Protein Science* 30.4, pp. 882–898. DOI: 10.1002/pro.4049.
Contributions: M.Y. Fry contributed to identifying STI1-domain containing proteins, the analysis of the amino acid distribution, and interpreting the computational models.

Fry, Michelle Y, Shyam M Saladi, Alexandre Cunha, et al. (2021). "Sequence-based features that are determinant for tail-anchored membrane protein sorting in eukaryotes". In: *Traffic* 22.9, pp. 306–318. DOI: 10.1111/tra.12809.
Contributions: M.Y. Fry designed and executed all *in vivo* imaging experiments and participated in bioinformatical analyses and image processing.

Lin, Ku-Feng et al. (Jan. 2021). "Molecular basis of tail-anchored integral membrane protein recognition by the cochaperone Sgt2". In: *Journal of Biological Chemistry* 296. DOI: 10.1016/j.jbc.2021.100441.
Contributions: M.Y. Fry designed and executed pull-down experiments pertaining to the identification the minimal binding domain and verification the computational model of Sgt2.

Fry, Michelle Y and William M Clemons Jr (2018). "Complexity in targeting membrane proteins". In: *Science* 359.6374, pp. 390–391. DOI: 10.1126/science.aar5992.
Contributions: M.Y. Fry wrote the manuscript.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

*Chapter 1*

# INTRODUCTION

Before animals, plants, and bacteria, there were "protocells" – RNA and proteins encapsulated by lipids. Over the course of history, these protocells have evolved into the cells that make up the complex organisms we know today (Schrum, T. F. Zhu, and Szostak, 2010). Organisms are comprised of cells which are primarily proteins, DNA, and RNA surrounded by a lipid-bilayer composed of fatty acid chains. Eurkaryotes, including mammals and plants, are more complex than prokaryotes, *i.e.* bacteria, as they contain membrane-bound compartments or organelles within the cell.

For a cell, making hydrophobic integral membrane proteins (IMPs) is a complicated but critical process. All proteins are made by ribosomes in the cytosol, but IMPs, accounting for ~30% of the proteins encoded in the eukaryotic genome, must also be properly delivered to and inserted into their designated membranes, a process known as targeting. This is important because hydrophobic IMPs are rapidly degraded in the cytoplasm to prevent aggregation, which can lead to broad disruptions in cellular homeostasis. Due to the number and diversity of IMPs, the process of identifying and targeting relies on pathways that often overlap in function. The information for targeting, typically stored in hydrophobic alpha-helical transmembrane domain (TMD) signals in the IMPs, is recognized by factors that then ferry the IMP clients to the destined lipid bilayer.

The precise information encoded in molecular signals important for targeting continues to be elusive. Historically, decoding known signals into detailed rules has proven difficult given their great variation and the lack of sequence motifs. Despite our inability to define these rules, cellular chaperones accurately recognize the various signals to sort clients into their distinct cellular destinations.

## 1.1 Targeting tail-anchored proteins

IMP targeting is dominated by the secretory (SEC) pathway with the signal recognition particle (SRP) being the central targeting factor (Guna and Hegde, 2018). For the majority of IMPs, SRP binds the N-terminal signal sequence as it emerges from a ribosome. The subsequent nascent protein-SRP complex is delivered to the

endoplasmic reticulum (ER) membrane for co-translational insertion via the SEC translocon, an ER-protein complex that acts as a conduit for the insertion of IMPs (Shao and Hegde, 2011a). However, not all IMPs can access the SEC pathway.

Tail-anchored (TA) proteins, found across eukaryotic cellular compartments, are a large class of these SEC-independent IMPs and are involved in a variety of roles including vesicle trafficking, protein translocation, quality control, and apoptosis (reviewed in (Krogh et al., 2001; Fry and Clemons, 2018; Guna and Hegde, 2018)). Marked by a single TMD near their C-termini, TA proteins account for approximately 2% of the genome. The TMD acts as the signal that is recognized to target these TA proteins to membranes, yet it remains hidden within the ribosomal exit tunnel at the end of translation. This necessitates post-translational targeting of the newly synthesized protein primarily to either the ER membrane or the outer mitochondria membrane (OMM).

The first pathway identified specifically for TA protein targeting, dubbed the Guided Entry of TA protein (GET) pathway, delivers TA proteins to the ER with the ATPase Get3 as the central targeting factor (Shao and Hegde, 2011a; Stefanovic and Hegde, 2007; Schuldiner, Metz, et al., 2008). Consisting of five proteins, Sgt2 and Get1-5, the GET pathway is responsible for targeting ER TA proteins with extremely hydrophobic TMDs. The co-chaperone Sgt2 first captures TA proteins from the yeast heat-shock protein 70 (Hsp70), Ssa1, and with the aid of Get4 and Get5, transfers the client to an ATP-bound Get3 (Cho and Shan, 2018; Chartron, Clemons, and Suloway, 2012). Upon TA protein binding and ATP hydrolysis, Get4/5 disassociates from Get3 and an ER membrane bound Get1/2 complex releases the TA protein from Get3 and drives its insertion into the ER membrane.

Conformational changes in Get3 driven by nucleotide and binding partners regulates the GET pathway, resulting in the insertion of TA proteins into the ER membrane. In the nucleotide free (apo) state, Get3 adopts an open conformation (Chio, Chung, et al., 2017). Once nucleotide (ATP) binds, Get3 transitions into a closed conformation that is recognized by a Get4/5 complex (Gristick et al., 2014). Get4 inhibits ATP hydrolysis of Get3 and recruits Sgt2, facilitating the hand off of TA protein to Get3 (Rome, Chio, et al., 2014). Once TA protein binds and ATP hydrolysis occurs, Get4/5 disassociates from Get3; this is critical as Get4 and Get2 have overlapping binding sites on Get3. The Get3/TA protein complex is driven apart by the Get1/2 complex, opening up Get3 (McDowell et al., 2020). While work structurally characterizing Get3 in these different conformational states have been successful,

a key missing component is characterization of the conformational changes in the Get3/TA complex in the post-hydrolysis state that cause Get4 to disassociate and the TA protein to continue on its route to the ER membrane.

Most of the work dedicated to the GET pathway has been limited to Opisthokonts (humans and yeast), Archaea, and plants, with little known about the pathway in other organisms including human pathogens. Deletion of GET pathway components in yeast results in a variety of phenotypes including, but not limitied to defects in Golgi-to-ER targeting, sensitivity to metal ions, and effects on the protein degradation machinery, which have been attributed to the mislocalization of TA proteins (Stefanovic and Hegde, 2007; Schuldiner, Metz, et al., 2008). In yeast, the localization of TA proteins vary in their sensitivity to the loss of GET proteins – some TA proteins, such as Sed5, remain in the cytosol whereas others, such as Sec61 and Emp47, localize correctly in membranes (Schuldiner, Metz, et al., 2008; Rivera-Monroy et al., 2016). Consequently, there must be alternative pathway(s) to traffic a subset of TA proteins.

### 1.2 The numerous routes to the ER membrane

Recently, two new pathways capable of targeting TA proteins to the ER membrane have been discovered, the SRP-independent (SND) pathway and one that uses the ER membrane complex (EMC) as an insertase. The first alternative pathway identified was the SND pathway, which has the ability to traffic several IMPs to the ER membrane, including both TA proteins and multi-pass IMPs, although the SND pathway may play a secondary role for some (Aviram, Costa, et al., 2016). Snd1, the first component of the SND pathway, interacts with the ribosome and possibly the nascent chain while the membrane bound Snd2 and Snd3 interact with the translocon complex. In the absence of the GET pathway, the SND pathway targets ER TA proteins with TMDs towards the center of the nascent chain.

The third targeting pathway for ER TA proteins uses the EMC as an insertase. In a recent report, Guna and colleagues demonstrated that human Get3 (*Hs*Get3) fails to bind to TA proteins with relatively low hydrophobicity within their TMDs. These proteins instead are delivered by calmodulin and inserted into the ER membrane by the EMC (Guna, Volkmar, et al., 2018), a ten-subunit complex. Interestingly, the same series of beautiful genetic experiments that first discovered GET components, also found a group of proteins that formed EMC (Schuldiner, Collins, et al., 2005; Jonikas et al., 2009). For TA proteins with moderately hydrophobic TMDs, both

the GET pathway and EMC can facilitate insertion.

The requirement of the EMC for the targeting of perhaps as much as 50% of ER-bound TA proteins suggests that the observed effects from the deletion of the EMC, such as the accumulation of misfolded IMPs, could indirectly result from failed TA protein insertion, similar to defects arising from deletion of the GET pathway. For example, a number of ER-bound TA proteins are involved in lipid synthesis and ER-associated degradation. Although other factors are not required, the lack of a direct interaction between calmodulin and the EMC suggests there may be accessory factors that remain to be determined.

These overlapping pathways, dependent on either hydrophobicity or signal positioning, highlight the diversity in these proteins and illuminate the difficulty of identifying a common characteristic of ER-destined TMDs. The TMDs of TA proteins vary in length and hydrophobicity, which likely explains the dependence on different pathways.

### 1.3 Classifying TA proteins

The identification of multiple routes for targeting TA proteins to the ER membrane challenges how we previously differentiated mitochondria- from ER-bound TA proteins (Wattenberg and Lithgow, 2001; Rao et al., 2016; Guna and Hegde, 2018). To date, while the cellular components involved in mitochondrial-bound TA protein targeting remain unclear (Aviram, Costa, et al., 2016; Guna, Volkmar, et al., 2018; Fry and Clemons, 2018). How the various pathways discriminate between clients remains an open question.

From exploration of targeting information within the TMD and in the C-terminal residues following the TMD of TA proteins general patterns have been observed. ER-bound TA proteins tend to have more hydrophobic TMDs (Guna and Hegde, 2018; Guna, Volkmar, et al., 2018; Rao et al., 2016; Wattenberg and Lithgow, 2001) while some mitochondria-bound TA proteins are amphipathic (Wattenberg and Lithgow, 2001). Positive charges following the TMD of TA proteins appear to prevent insertion into the ER membrane regardless of TMD hydrophobicity (Rao et al., 2016). Throughout these previous works, the ability of these rules to separate ER- and mitochondrial-bound TA proteins at-large has not been assessed, so their broader applicability is still unclear. With multiple pathways with overlapping clients, understanding the factors within clients recognized for targeting is critical.

When I first began studying the GET pathway, we were primarily focusing on

Opisthokont GET homologs. With the identification of alternative targeting pathways for ER-bound TA proteins we began to ask the question: what is encoded in TA proteins that are recognized by all these factors to ensure correct targeting? As time progressed we began investigating how GET components, mainly Get3 and Sgt2, interact with TA proteins and protect the hydrophobic TMDs from the aqueous cytosol across different eukaryotic supergroups. The work presented here addresses these inquiries. Chapter 2 presents a new and more inclusive metric for classifying ER- and mitochondria-bound TA proteins which was verified through a combination of computational analyses and live-cell imaging. In Chapter 3, the first molecular model of the TA protein binding domain in Sgt2 is proposed as well as biochemical insights into the binding mechanism of the co-chaperone. Chapter 4 chronicles how the molecular model of Sgt2-C domain lead to the development of a new definition for a protein fold domain, dubbed the STI1-domain. Three structures of Get3 from the supergroup Excavata in three distinct nucleotide states are described in Chapter 5. These structures are the first structures of a protist GET component and reveal a previously unseen conformation of Get3. Together these structures complete the catalytic cycle of Get3 in *Giardia intestinalis*. Comparisons between the fungi, mammalian, and protist Get3 are made throughout this text. Finally, Chapter 6 concludes this thesis by reviewing the finding discussed in the previous chapters and future experiments to further our understanding of client recognition in the GET pathway as well as how conformational changes in Get3 drive protein targeting.

*Chapter 2*

# DETERMINING FEATURES ENCODED IN TA PROTEINS THAT ENSURE CORRECT TARGETING

*Adapted from:*

Fry, Michelle Y et al. (2021). "Sequence-based features that are determinant for tail-anchored membrane protein sorting in eukaryotes". In: *Traffic* 22.9, pp. 306–318. DOI: 10.1111/tra.12809.

M.Y. Fry designed and executed of all *in vivo* imaging experiments and participated in bioinformatical analyses, and image processing.

## Abstract

The correct targeting and insertion of tail-anchored (TA) integral membrane proteins is critical for cellular homeostasis. TA proteins are defined by a hydrophobic transmembrane domain (TMD) at their C-terminus and are targeted to either the ER or mitochondria. Derived from experimental measurements of a few TA proteins, there has been little examination of the TMD features that determine localization. As a result, the localization of many TA proteins are misclassified by the simple heuristic of overall hydrophobicity. Because ER-directed TMDs favor arrangement of hydrophobic residues to one side, we sought to explore the role of geometric hydrophobic properties. By curating TA proteins with experimentally determined localizations and assessing hypotheses for recognition, we bioinformatically and experimentally verify that a hydrophobic face is the most accurate singular metric for separating ER and mitochondria-destined yeast TA proteins. A metric focusing on an eleven residue segment of the TMD performs well when classifying human TA proteins. The most inclusive predictor uses both hydrophobicity and C-terminal charge in tandem. This work provides context for previous observations and opens the door for more detailed mechanistic experiments to determine the molecular factors driving this recognition.

## 2.1 Introduction

Biogenesis of membrane proteins is an essential, yet complicated, process necessary for maintaining cellular homeostasis. Synthesized by ribosomes in the cytosol, membrane proteins account for approximately a third of the proteome and must be targeted to specified membranes (reviewed in (Krogh et al., 2001; Fry and Clemons, 2018; Guna and Hegde, 2018)). A hydrophobic alpha-helical stretch, often a transmembrane domain (TMD), encodes this information and its position within an open reading frame dictates the cellular machinery responsible for its recognition and targeting (Guna and Hegde, 2018). While computational methods have refined the ability to detect and predict cellular localization of these integral membrane proteins over time (Almagro Armenteros et al., 2017), the precise molecular signals continue to be elusive. Historically, decoding known signal sequences into detailed rules has proven difficult given their great variation and the lack of sequence motifs—thus these signals are often discussed at a high level, e.g. hydrophobic alpha-helical stretches. Despite the inability to define these rules, cellular chaperones accurately recognize the various signals to sort clients into their distinct cellular destinations.

Here, we attempt to address one class of membrane proteins, tail-anchored (TA) proteins, found across cellular compartments and involved in a variety of roles including vesicle trafficking, protein translocation, quality control, and apoptosis (reviewed in (Fry and Clemons, 2018; Borgese, Colombo, and Pedrazzini, 2003; Chartron, Clemons, and Suloway, 2012; Rabu et al., 2009a)). TA proteins are marked by a single TMD near their C-terminus and account for approximately 2% of the genome (Kutay, Hartmann, and Rapoport, 1993; Denic, 2012; Wattenberg and Lithgow, 2001; Chartron, Clemons, and Suloway, 2012). Due to the position of their signal sequence, TA proteins are translated by the ribosome and then post-translationally targeted primarily to the endoplasmic reticulum (ER) or outer mitochondrial membrane. The TMD and C-terminal residues following have been demonstrated to be necessary and sufficient for correct targeting in many experimental contexts (F. Wang, Brown, et al., 2010; Lin et al., 2021). Thus, it is suggested that the information recognized by TA protein targeting pathways is contained within the transmembrane domain and neighboring residues.

The recent identification of a new route for TA proteins to the ER membrane has challenged how we previously differentiated between mitochondria and ER-bound TA proteins (Guna and Hegde, 2018; Rao et al., 2016; Wattenberg and Lithgow, 2001). To date, while the cellular components involved in mitochondrial TA protein

targeting remain unclear, multiple overlapping pathways have been identified for TA protein targeting to the ER membrane (Chartron, Clemons, and Suloway, 2012; Schuldiner, Metz, et al., 2008; Stefanovic and Hegde, 2007; Aviram, Costa, et al., 2016; Guna, Volkmar, et al., 2018; Fry and Clemons, 2018). The first identified and most studied pathway is the Guided Entry of TA protein (GET) pathway (Stefanovic and Hegde, 2007; Schuldiner, Metz, et al., 2008). Consisting of six proteins, Sgt2 and Get1-5, the GET pathway is responsible for targeting ER TA proteins with more hydrophobic TMDs. In yeast, the co-chaperone Sgt2 first captures TA proteins from Ssa1 and, with the aid of Get4 and Get5, transfers the client to the ATPase Get3 that acts as the central targeting factor of the pathway (Chio, Cho, and Shan, 2017; Shao and Hegde, 2011b; Guna and Hegde, 2018; Shao and Hegde, 2011a; Cho and Shan, 2018). An ER membrane bound Get1/2 complex facilitates disassociation of the Get3/TA complex and insertion of the TA protein into the membrane. Recently, Guna and colleagues demonstrated that human Get3 (HsGet3) fails to bind to TA proteins with relatively low hydrophobicity within their TMDs. These proteins instead are inserted into the ER membrane by the ER Membrane Complex (EMC) (Guna, Volkmar, et al., 2018). A ten-subunit complex, the EMC inserts TA proteins delivered by calmodulin. For TA proteins with moderately hydrophobic TMDs, both the GET pathway and EMC can facilitate insertion. A third dedicated pathway capable of targeting TA proteins into the ER membrane is the SRP-independent (SND) pathway (Aviram, Costa, et al., 2016). Snd1, the first component of the SND pathway, interacts with the ribosome and possibly the nascent chain while the membrane bound Snd2 and Snd3 interact with the translocon complex. In the absence of the GET pathway, the SND pathway is capable of targeting ER-bound TA proteins with TMDs further away from their C-termini. These overlapping pathways, dependent on either hydrophobicity or signal sequence positioning, highlight the diversity in these proteins and the difficulty in identifying a common characteristic of ER-destined TMDs (Aviram, Costa, et al., 2016).

General patterns have been observed based on exploration of targeting information within the TMD and the C-terminal residues of TA proteins. ER-bound TA proteins tend to have more hydrophobic TMDs (Guna, Volkmar, et al., 2018; Chitwood et al., 2018; Rao et al., 2016; Wattenberg and Lithgow, 2001) while some mitochondria TA proteins are amphipathic (Wattenberg and Lithgow, 2001). By modifying the positive charge following their TMDs with an example TMD, studies have shown how insertion by the GET pathway into the ER membrane can be impaired (Figueiredo Costa et al., 2018; Rao et al., 2016). Distinction between peroxisomal and mito-

chondria TA proteins have been made based on the charge of their C-terminal tails, whereas mitochondria- and ER-bound TA proteins in mammals are differentiated by a combination of TMD hydrophobicity and C-terminal charge (Costello et al., 2017). A charged tail was overcome by increasing the hydrophobicity of the TMD, directing the mitochondrial TA protein to the ER. Guna and colleagues determine a threshold in total hydrophobicity by modifying a model TMD to delineate clients that are inserted either via the GET or EMC pathways (Guna, Volkmar, et al., 2018). Throughout these previous works, the ability of these rules to separate ER vs mitochondrial TA proteins at-large has not been systematically assessed, so their broader applicability is still unclear.

With multiple pathways with overlapping clients, understanding the factors within clients recognized for targeting is critical. Here we show that formalizing previously suggested criteria, while adequate, are not sufficient for classifying ER-bound TA proteins with moderately hydrophobic TMDs suggested to be clients of the EMC insertase. We demonstrate through computational and experimental methods that classifying TA proteins by the presence of a hydrophobic face in their TMD is more inclusive, properly capturing both ER TMDs with low hydrophobicity and mitochondrial TA proteins in both yeast and humans.

## 2.2 Results

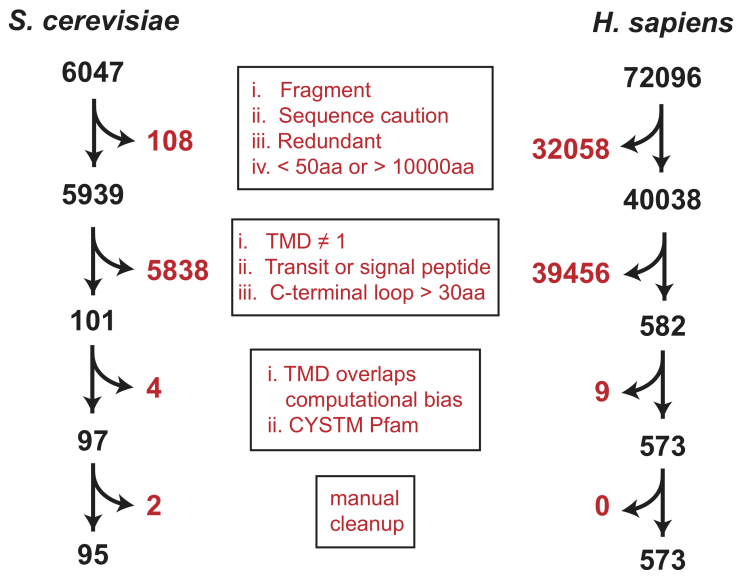
### Curating TA proteins with experimentally determined localizations

In order to screen TA proteins to identify a concise criterium for localization, we first curated a comprehensive set of TA proteins from the yeast proteome pulling together localizations across public repositories and publication-associated datasets. We screened the reference yeast genome from UniProt (Consortium, 2020) for putative TA proteins and filtered for unique genes longer than 50 residues (Fig. 2.1 *A*). Uniprot and TOPCONS2 (Tsirigos et al., 2015) were used to identify proteins with a single TMD within 30 amino acids of the C-terminus (Borgese, Colombo, and Pedrazzini, 2003) that lacked a predicted signal sequence (as determined by SignalP 4.1 (Nielsen, 2017)). While this set encompasses proteins previously predicted as TA proteins (Beilharz et al., 2003; Kalbfleisch, Cambon, and Wattenberg, 2007), it is larger (95 vs 55 or 56) and we believe a more accurate representation of the repertoire of TA proteins (Fig. 2.1*B*). Based on their UniProt-annotated and Gene Ontology Cellular Components (GO CC) localizations (Gene Ontology Consortium, 2021), TA proteins were subcategorized as ER-bound (encompassing labels including cell membrane, Golgi apparatus, nucleus, lysosome, and vacuole membrane), mitochondrial (inner and outer mitochondria membrane (IMM & OMM)), peroxisomal, and unknown (Fig. 2.1 *C*). This set is readily available for future analyses (Table S2.1). The majority of proteins have no annotated cellular localization. Several previously suggested TA proteins are excluded from this new set including, for example, OTOA (otoancorin) that contains a likely signal sequence, FDFT1 (squalene synthase or SQS) with two predicted hydrophobic helices by this method, and YDL012C which has a TMD with very low hydrophobicity (Guna, Volkmar, et al., 2018; Beilharz et al., 2003). This analysis was also applied to the human genome and a list of 573 putative TA proteins was compiled and annotated based on published localizations (Fig. 2.1 *A*-*C*). Like with the yeast list, the human list is larger than previous reports (573 vs 411), and the majority of the proteins have no annotated localization.

### Assessing current metrics for TA classification

To identify factors encoded within TA proteins that ensure correct localization, we began by considering several posited properties including the charge following the TMD, TMD length, and TMD hydrophobicity. Previous reports suggest that the presence of positively charged residues following the TMD of mitochondria-bound TA proteins prevents insertion into the ER membrane (Figueiredo Costa et al., 2018; Rao et al., 2016). The number of positively charged C-terminal residues for

A




B

| | *S. cerevisiae* | *H. sapiens* |
|---|---|---|
| Beilharz, et al. 2003 | 55 | - |
| Kalbfleisch, et al. 2007 | 56 | 411 |
| Here | 95 | 573 |

C

| | *S. cerevisiae*: 95 | | | *H. sapiens*: 573 | | |
|---|---|---|---|---|---|---|
| | Uniprot | GO CC | Final | Uniprot | GO CC | Final |
| ER | 36 | 35 | 43 | 114 | 152 | 164 |
| Mit | 9 | 8 | 10 | 37 | 35 | 45 |
| ER & Mit | 1 | 1 | 2 | 13 | 29 | 18 |
| Peroxisomal | 1 | 3 | 2 | 4 | 7 | 5 |
| Unknown | 48 | 48 | 38 | 405 | 350 | 341 |

Figure 2.1: **Compiling a list of TA proteins from the human and yeast genomes.** *A)* A schematic of the pipeline used to gather TA proteins by filtering the Human and Yeast proteomes for TA proteins. *B)* A comparison of the TA proteins collected for the analyses here versus previous datasets. *C)* Localizations gathered from Uniprot entry Subcellular Localizations (CC) and Gene Ontology Cellular Compartment (GO) annotations. Those with conflicts were resolved by manually parsing the literature to build the final set.

all 95 yeast proteins was calculated, avoiding issues associated with defining the extent of TMDs by counting any charge from the center of the predicted TMD to the C-terminus. No clear separation is observed when plotting TA proteins with known localizations by number of positively charged residues (Fig. 2.2 *A*). As a metric this does a poor job distinguishing between the two; six ER-annotated pro-

teins have a C-terminal positive charge of three or more and one out of the eight mitochondria-annotated proteins has no C-terminal positive charge. Furthermore, neither negative nor net charge of the C-terminal loop separates ER from mitochondrial TA proteins (Fig. 2.2 *B&C*). While modulating the C-terminal positive charge affects localization (Rao et al., 2016), cells do not solely use this signal to specify protein localization. Considering the difference in lipid compositions of the ER and mitochondrial membranes, a signal might be encoded in the TMD lengths, but this metric also fails to separate the two sets (Fig. 2.2*D*).

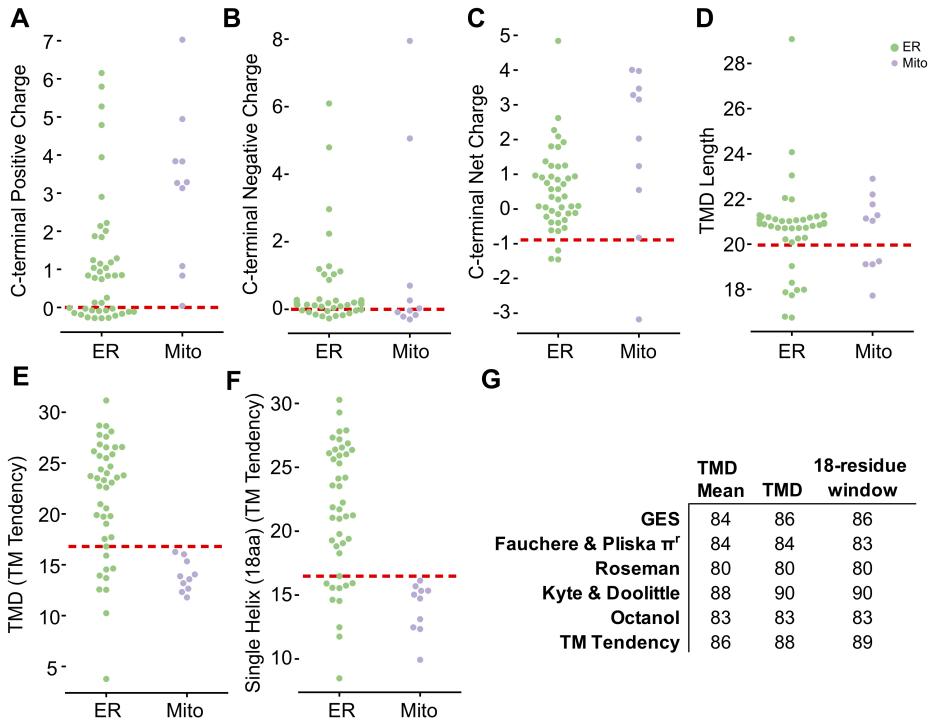



Figure 2.2: **Investigating properties encoded in the C-terminal residues of TA proteins.**
For *A-F*, Jitter plots of property distribution for predicted TA proteins identified as ER (*green*) or mitochondria (*purple*) with the best predictive threshold indicated by a dashed red line. Properties visualized are for the C-terminal number of *(A)* positive residues, *(B)* negative residues, and *(C)* net charge and then for *(D)* TMD length, *(E)* TMD hydrophobicity, and *(F)* maximum hydrophobicity of an 18-residue stretch. *(G)* The AUROC across various hydrophobicity scales for the mean, total, and 18-residue windows of the predicted TMDs.

TMD hydrophobicity is the proposed localization-determining feature of TA proteins in studies thus far (Guna and Hegde, 2018; Rao et al., 2016; Wattenberg and Lithgow, 2001). The TM tendency scale, used here and in past studies with TA targeting (Guna, Volkmar, et al., 2018; Guna and Hegde, 2018), is a statistical

hydrophobicity scale that incorporates both hydrophobicity and helical propensity into a single value assigned to each of the 23 amino acids by using amino acid propensities in TMDs known at the time of its creation (Zhao and London, 2006) (Fig. 2.2*E*). The total hydrophobicity (sum of each residue's hydrophobicity value) of a TMD sufficiently splits ER and mitochondrial proteins but places a significant number of ER-bound TA proteins among mitochondrial-bound TA proteins. In other words, the total hydrophobicity can classify GET pathway clients as ER-bound but fails to identify clients of the EMC insertase that are also ER-bound TA proteins (Guna, Volkmar, et al., 2018). For example, the TMD of squalene synthase, a bona fide EMC client (Guna, Volkmar, et al., 2018), has a lower hydrophobicity than that of model mitochondrial TA protein, Fis1 (Total TM Tendency = 12.5 vs 18.78, respectively). Limiting the hydrophobicity to a single helix stretch, i.e. 18aa, sees no improvement in classification (Fig. 2.2 *F*).

To examine this inability to correctly classify lower hydrophobic ER-bound TA proteins, we comprehensively assess hydrophobicity across a variety of established scales (Eisenberg et al., 1984; Fauchere and Pliska, 1983; Roseman, 1988; Wimley, Creamer, and White, 1996; Zhao and London, 2006) (Fig. 2.2 *G*) and then quantitatively assess predictive power using the receiver operating characteristic (ROC) framework (Swets, Dawes, and Monahan, 2000). An ROC curve captures how well a numerical score separates two categories, here ER vs mitochondria, and whose figure of merit is the area under the curve (AUROC). This is a more accurate representation of prediction than simpler numbers like accuracy and precision, which require setting a specific threshold in a numerical score. A perfect separation gives an AUROC of 100 whereas a random separation results in an AUROC of 50. No matter the hydrophobicity scale used, the total hydrophobicity captures the ER vs mitochondria split to varying extents. In each case, the mean hydrophobicity performs more poorly, yet considering the most hydrophobic 18-residue single-helix stretch results in a slight improvement in predictive ability suggesting that a subset of the helix can explain recognition (Fig. 2.2 *G*).

**TMD residue organization better classifies TA protein localization**

We wondered if TA protein classification could be improved by carefully assessing the hydrophobicity of the TMDs. Data showing that Sgt2 (a co-chaperone in the GET pathway) binds a TMD of a minimal length of 11 residues suggests only a subset of each helix may be necessary to classify localization (Lin et al., 2021). Indeed, the maximum hydrophobicity of segments, specified by the number residues

selected, better classifies ER vs mitochondrial TA proteins across hydrophobicity scales (Fig. 2.3 *A&B*).

Furthermore, it was also reported that TMDs where the most hydrophobic residues cluster to one side of a helical wheel plot (Schiffer and Edmundson, 1967), a 2D representation of an alpha-helix, bind more efficiently to Sgt2 (Lin et al., 2021). We sought to examine if this clustering is a feature of ER-bound TA proteins and absent in mitochondria-bound TA proteins. This clustering we define as a helical wheel face (Wheel Face) and specify a length by the number of residues selected (Fig 2.3 *A&B*). We also extend the face along the sides of the helix, defining a Patch, selecting three of the four residues in a single turn of a helix. Patch geometries are specified by length of the segment considered, i.e. Patch 11 is confined in a 11 segment residues with 9 residues selected (Fig. 2.3 *A&B*). Improvements in classification over the total hydrophobicity metric are seen in several cases (Fig 2.3 *B*, *green*, Fig. 2.4 *B*, *green*, S2.2). The metrics with the best classification capability are Patch 15 (Kyte & Doolittle and TM Tendency), Wheel Face 5 (TM Tendency), and Patch 11 (Kyte & Doolittle) (Fig 2.3 *B*, *dashed red box*). These metrics have an improved AUROC value of 96, 96, 95, and 95, respectively, compared to the TMD hydrophobicity score of 90 (Kyte & Doolittle) and 88 (TM Tendency) (Fig 2.3 *B*). At the best threshold of the ROC curve, these metrics correspond to five, seven, six, and eight miscategorized proteins, respectively. A scatter plot illustrates how these metrics translate to improved separation of ER and mitochondrial TA proteins (Fig. 2.3 *C*).

Other hydrophobic geometries were also explored as potential competing hypotheses: residues in a line (every fourth residue), rectangle (one residue plus two residues two away on either side), or star (two adjacent residues and one residue two away on either side) (Fig. 2.4 *A&B*). As with the Patch geometries, these geometries are specified by the length of the TMD considered. Again, improvements are seen in geometries that present hydrophobic patches, i.e. Rectangle 9 and Star 8, where line geometries rarely improved classification regardless of scale used (Fig. 2.4 *B*). Given the relative dearth of experimental data and the substantial number of hypotheses being tested (geometries and hydrophobicity scales), it is difficult to definitively say if one geometry is the sole deciding factor for localization based only on bioinformatics. Regardless of the hydrophobicity scale used, it is clear that the organization of hydrophobic residues within a TMD is important for targeting TA proteins to their intended membranes.

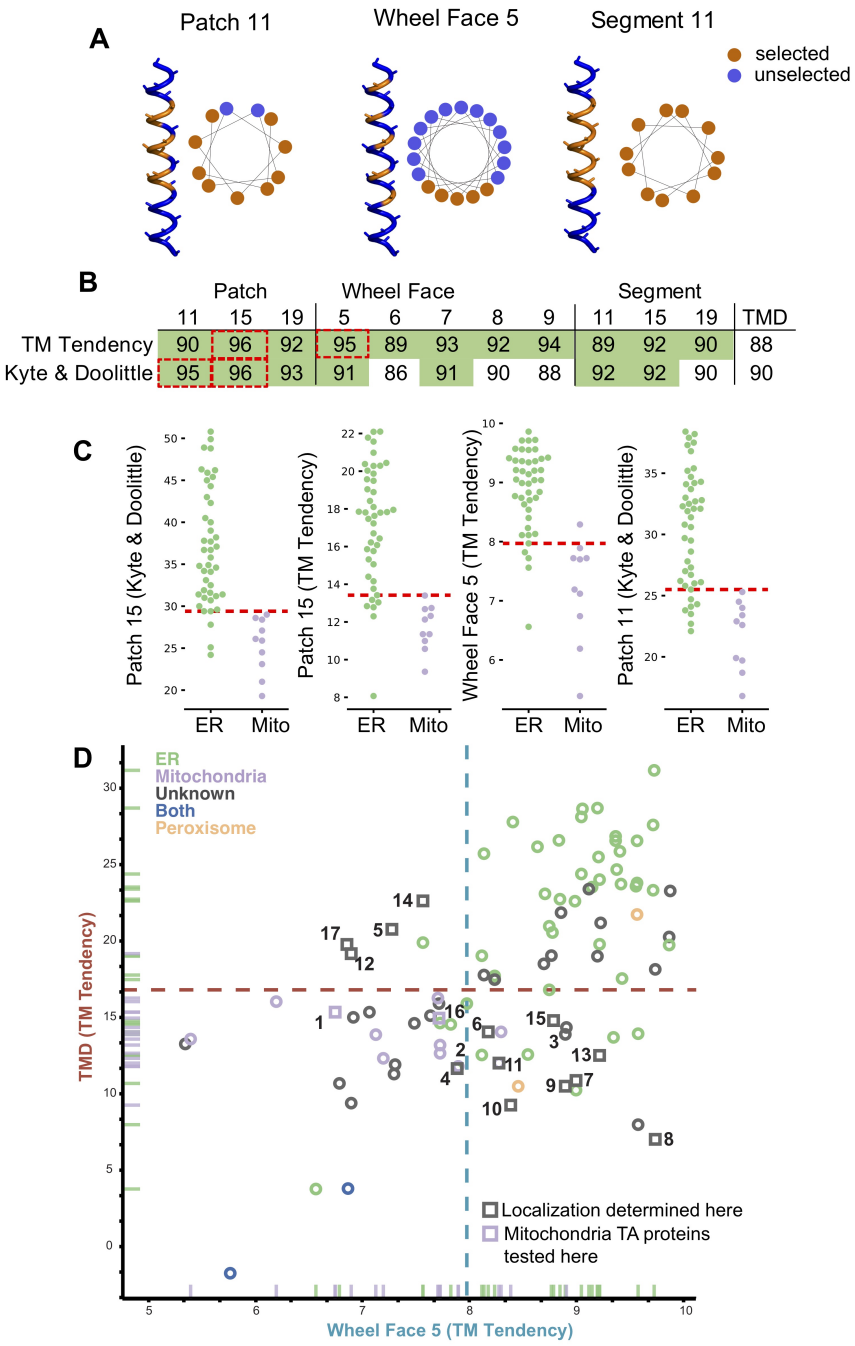
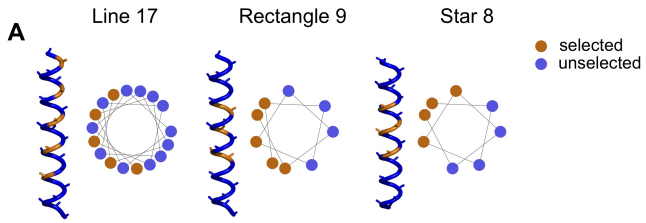



Figure 2.3: **Analyzing different geometries of hydrophobic residues in TMDs to improve classification.**
*A*) $\alpha$-helices and helical wheel plots illustrating the residues selected (*orange*) for each metric tested, patch, wheel face, and segment, showing residues selected and not selected (*blue*) in each analysis. *B)* AUROC values for the metrics illustrated in *(A)* and total hydrophobicity. *C)* Jitter plots as in Fig. 3.3 for the top four hydrophobic metrics: Patch 15 (Kyte & Doolittle), Patch 15 (TM Tendency scale), Wheel Face 5 (TM Tendency scale), and Patch 11 (Kyte & Doolittle). Red dashed line indicates the best predictive threshold. *D)* 2D comparison plot of total hydrophobicity (*y-axis*) and a Wheel Face 5 (TM Tendency scale) (x-axis). TA proteins are colored by localization, ER (*green*), mitochondria (*purple*), Unknown (*grey*), both mitochondria and ER (*blue*), and peroxisome (*orange*). TA proteins selected for experimental determination of localizations are marked squares. Dashed lines indicate best predictive threshold.




B

| | Patch | | | Helical Wheel Face | | | | | | Line | | | Segment | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 11 | 15 | 19 | 4 | 5 | 6 | 7 | 8 | 9 | 9 | 13 | 17 | Rect9 | Star8 | 11 | 15 | 19 | TMD |
| GES | 90 | 94 | 91 | 74 | 87 | 85 | 88 | 89 | 90 | 79 | 89 | 79 | 89 | 87 | 87 | 92 | 85 | 86 |
| Fauchere & Pliska π′ | 86 | 88 | 90 | 79 | 81 | 80 | 83 | 84 | 87 | 55 | 76 | 73 | 91 | 88 | 86 | 87 | 89 | 84 |
| Roseman | 85 | 85 | 88 | 72 | 81 | 81 | 88 | 88 | 88 | 64 | 94 | 68 | 88 | 88 | 83 | 84 | 82 | 80 |
| Octanol | 83 | 88 | 91 | 80 | 81 | 80 | 83 | 85 | 86 | 70 | 75 | 70 | 87 | 86 | 81 | 85 | 86 | 83 |
| Better than total: | 3 | 4 | 4 | 0 | 2 | 1 | 2 | 4 | 4 | 0 | 2 | 0 | 4 | 4 | 3 | 4 | 3 | |

Figure 2.4: **Alternative geometries of hydrophobic residues in TMDs tested for improved classification.**
*A)* Alpha-helices and helical wheel plots to illustrate the residues used for each metric tested, patch, wheel face, and segment, showing residues selected (*orange*) and not selected (*blue*) in each analysis. *B)* AUROC values for the metrics illustrated in *(A)* and total hydrophobicity.

### Testing the localization of unknown TA proteins

We then tested if either face (Wheel Face or Patch), Segment, or TMD hydrophobicity metrics enabled us to predict the localization of unknown TA proteins. To do this we selected a subset of unknown TA proteins, whose localization would be predicted differently by TMD and Wheel Face 5 metrics using the TM Tendency scale (Fig. 2.3 *D*, *numbered grey points*, S2.3). This selection was made because of the strong AUROC and biochemical data suggesting TA protein containing a helical wheel face bind more efficiently to Sgt2. Several in this group have a hydrophobicity less than the previously suggested cut-off for EMC clients (Guna, Volkmar, et al., 2018) (Fig. 2.3 *D*, *lower right quadrant*). Our experimental setup based on that from Rao et al., 2016 – GFP is fused N-terminally to the TMD and C-terminal residues of the unknown TA protein (Fig. 2.5 *A*. *yellow panel*). Localization is determined by overlap with either a BFP-tagged mitochondria pre-sequence that marks the mitochondria (Fig. 2.5 *A*, *cyan panel*) and a tdTomato-tagged Sec63 acting as an ER marker (Fig. 2.5 *A*, *magenta panel*) (Rao et al., 2016). Overlap was determined computationally using two algorithms we developed: one to segment individual cells in brightfield and another to determine which fluorescence probe the GFP overlapped with on a per cell basis (Table S2.3).

This experimental setup and computational analysis were first applied to the known mitochondria proteins Fis1 and Cox26 (Rao et al., 2016; Levchenko et al., 2016;

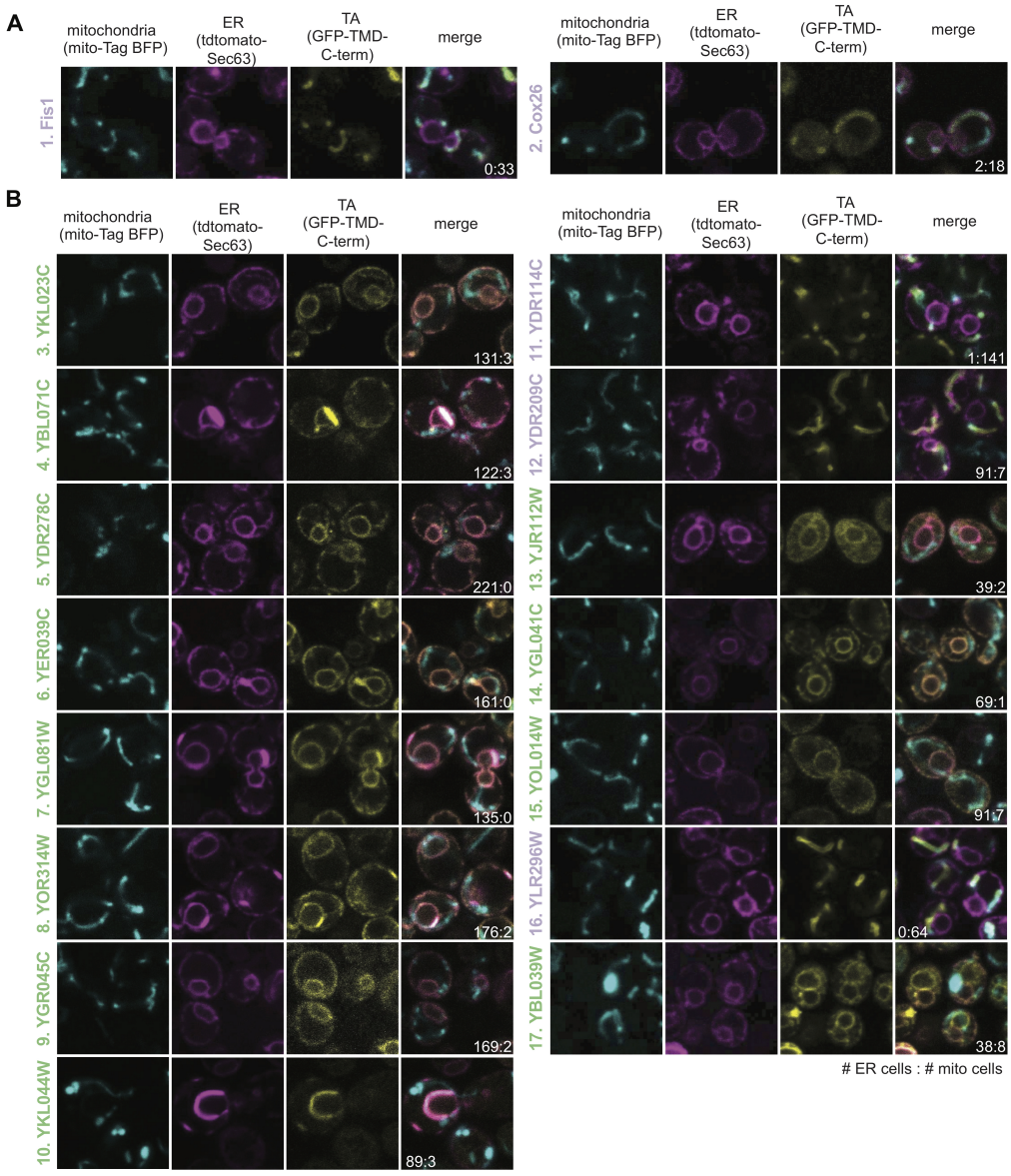



Figure 2.5: **Localization of unknown yeast TA proteins.**
The ER (*magenta panel*) and mitochondria (*cyan panel*) were labeled with tdTomato and BFP, respectively. TA protein localization was visualized by GFP (*yellow panel*) and colocalization was determined by overlap (*merge panel*). The ratio of the number of cells with the TA protein localizing to the ER vs the mitochondria are noted in the merge image. Numbered as in Fig. 3D with labels colored based on their determined localizations: ER (*green*) and mitochondria (*purple*). TA proteins include *(A)* two mitochondrial TA proteins with known localizations and *(B)* 15 with unknown localizations.

Hartley et al., 2018). The analysis correctly determines these proteins to colocalize with BFP, thus correctly classifying them as mitochondria-bound TA proteins (Fig. 2.5 *A*). We then experimentally tested the 15 Unknown TA proteins where 11 localize to the ER, three to the mitochondria, and one to another cellular compartment (Fig. 2.5 *B*, Table 2.1). The localization of this latter TA protein cannot be determined by our experimental setup except to say it does not clearly colocalize with the ER or mitochondria markers visually or through our computational analysis (Table S2.3). The shape of the organelle is consistent with localization to the ER-derived vacuole (Fig. 2.5 *B*, *17*) (Vida and Emr, 1995). In total, we report the first localization of ten previously Unknown TA proteins.

Several datasets report protein localizations in yeast but are not yet, or partially, integrated into bioinformatics databases like Uniprot. One in particular was of use for this study, reporting the localizations assigned by qualitatively accessing the pattern of protein expression in images of 17 TA proteins in the Unknown category (Weill et al., 2018) (Table S4). Coincidentally, a few of these proteins were included in our experimental test set, for a combined 27 new TA proteins with previously unknown localizations (Table S2.3 & S2.4). Of the TA proteins identified by Weill and colleagues, all but one, YKL044W, was confirmed (Table S2.3). Given the ability to mark ER and mitochondria and quantitate colocalization on a per-cell basis, we use the localization determined here throughout our analysis, i.e. YKL044W localizes to the ER. Collectively, we have compiled a list of 27 TA proteins and their localizations that have yet to be integrated into protein databases or reported: 20 ER, six mitochondrial, and one peroxisomal.

**Reassessing classification metrics using newly determined localizations**

The newly determined localizations were compared to the predicted localizations of the best performing hydrophobicity metrics. Total hydrophobicity metrics across all scales only correctly predict 9 or 14 of the 26 ER- and mitochondria-bound TA proteins. Experimental localizations from this work and the Schuldiner Lab (Weill et al., 2018) result in a putative yeast TA protein list with 88% having known localizations (Fig. 2.6 *A*). With most localizations known, comparing metrics based on AUROC values is a good representation of the overall dataset (Table S2.5). The best performing metrics were Wheel Face 7 (TM Tendency) and Wheel Face 5 (TM Tendency), with scores of 89 and 88, respectively vs the TMD hydrophobicity AUROC score of 76 (Table S2.5). These metrics correctly predicted the localization of 19 out of 26 and 17 out of 26, respectively, of the subset of our test set that localized

A

| | Correctly Predicted | | | Final | | Misclassified | | |
|---|---|---|---|---|---|---|---|---|
| | Mit | ER | Total | Rank | AUROC | Mit | ER | Total |
| TM Tendency Wheel Face 7 | 5 | 12 | 17 | 1 | 89 | 0 | 19 | 19 |
| TM Tendency Wheel Face 5 | 4 | 15 | 19 | 2 | 88 | 1 | 18 | 19 |
| Fauchere Pliska Patch 19 | 6 | 12 | 18 | 3 | 87 | 1 | 15 | 16 |
| Kyte & Doolittle Segment 11 | 5 | 12 | 17 | 4 | 87 | 0 | 20 | 20 |
| Fauchere Pliska TMD | 6 | 6 | 12 | 50 | 80 | 4 | 25 | 29 |
| Kyte & Dolittle TMD | 5 | 4 | 9 | 64 | 78 | 6 | 13 | 19 |
| TM Tendency TMD | 5 | 9 | 14 | 78 | 76 | 3 | 21 | 24 |
| | /6 | /20 | /26 | | | /14 | /63 | /77 |

B

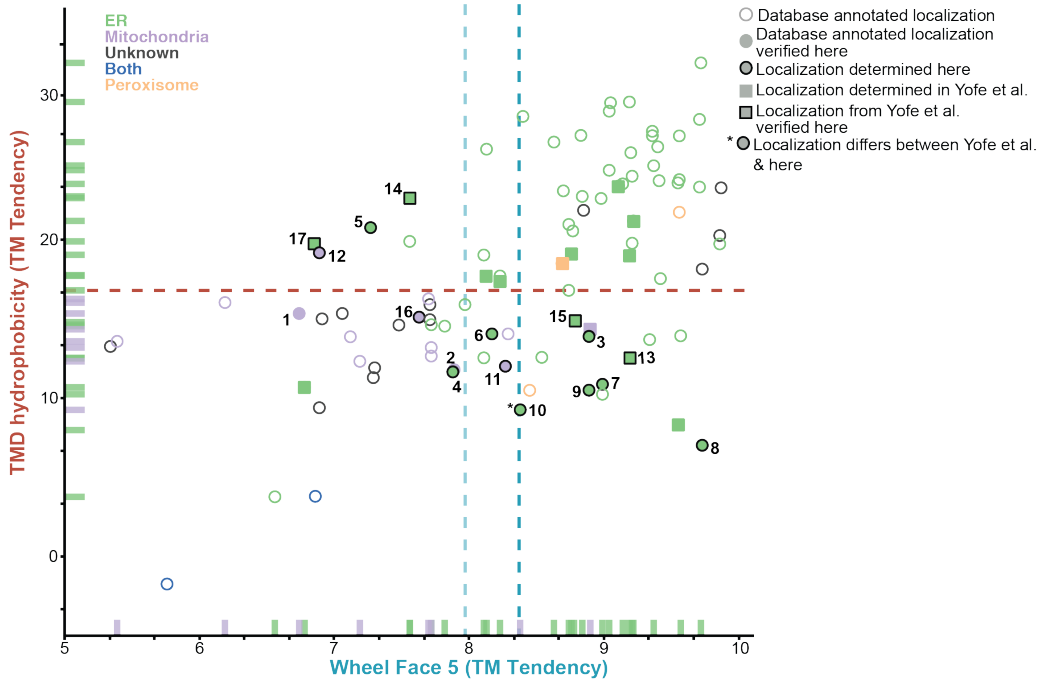



Figure 2.6: **A hydrophobic Wheel Face metric of 5 or 7 residues best separates ER- and mitochondria-bound TA proteins.**
*A)* A ranking of the five best performing hydrophobicity metrics compared to the TMD hydrophobicity metrics of the appropriate hydrophobicity scales (TM Tendency, Fauchere & Pliska, and Kyte& Doolittle). The number of correctly predicted localizations as well as the final AUROC scores are used to assess the effectiveness of each metric. The total number of correctly classified yeast TA proteins is also noted. The two metrics directly compared in the 2D comparison plot in *(B)* are highlighted in blue (TM Tendency, Wheel Face 5, *x-axis*) and red (TM Tendency, TMD, *y-axis*). Hydrophobicities are plotted and TA proteins are colored as they were in Fig. 2.3*D*. Newly determined localizations from Fig. 4 (*black outlined*) and Weill et al., 2018 (*squares*) are filled in with the appropriate colors, ER (*green*), mitochondria (*purple*), and peroxisome (*orange*).

to the ER or mitochondria (Fig. 2.6 *A&B*). A Patch geometry using the Fauchere & Pliska scale performs well when predicting new localizations – correctly predicting 18 of 26 localizations (Fig. 2.6 *A*). Segment metrics performed similarly when predicting new localizations and their AUROC values improved with the inclusion of the new localizations (Fig. 2.6 *A*). In all, metrics focused on the organization of hydrophobic residues within the TMD of TA proteins better predict TA protein localization – the best consider just a five or seven residue face or a fraction of the TMD.

**Expanding this metric to human TA proteins**

We next applied this analysis to the human genome. Using our compiled list of 587 putative human TA proteins, we sought to identify a more inclusive set of criteria for ER- vs mitochondria-bound TA proteins. The best performing hydrophobicity scales in the yeast dataset were TM tendency and Kyte & Doolittle, so the other scales were not further considered with the human dataset. While TMD hydrophobicity metrics correctly capture mitochondria-bound TA proteins, they fail to capture many ER-bound TA proteins (Fig. 2.7 *A*, Table S2.6). Quantitatively assessing all metrics, we see slight improvements in classification with metrics using patches or segments compared to total hydrophobicity (Fig. 2.7 *A&B*, Table S2.6). The metric with the highest AUROC score is Patch 11 (Kyte & Doolittle). Many proteins in our dataset have a single report of their localizations in databases. There is potential for changes to these localizations as seen with many Bcl-2 family members (Fig. 2.7 *B*, *filled blue points*) where there exist multiple reports of these proteins localizing to the ER and/or to the mitochondria. While this may be unique to these TA proteins, as their function to regulating apoptosis is tied in with their transport between the two membranes, some reported localizations may be the product of overexpression. Future work verifying and determining localizations of human TA proteins will likely result in improvements in classification by a metric derived from hydrophobic geometries.

**Determining a two-step criterion for localization determination**

We then tested if combining a hydrophobicity geometry with a C-terminal charge metric resulted in more accurate classification of TA proteins. Costello and colleagues demonstrated in mammals, distinctions between ER, mitochondria, and peroxisomal TA proteins can be made using a combination of charge and TMD hydrophobicity cut-offs (Costello et al., 2017). They suggest mitochondria-bound

A

| | Rank | AUROC | Correctly Classified |
|---|---|---|---|
| Kyte & Doolittle Patch 11 | 1 | 83 | 136 |
| Kyte & Doolittle Segment 19 | 2 | 83 | 151 |
| Kyte & Doolittle Segment 11 | 3 | 82 | 150 |
| TM Tendency Patch 11 | 4 | 82 | 148 |
| TM Tendency Patch 15 | 5 | 82 | 150 |
| Kyte & Dolittle TMD | 8 | 81 | 172 |
| TM Tendency TMD | 15 | 79 | 142 |
| | | | /209 |

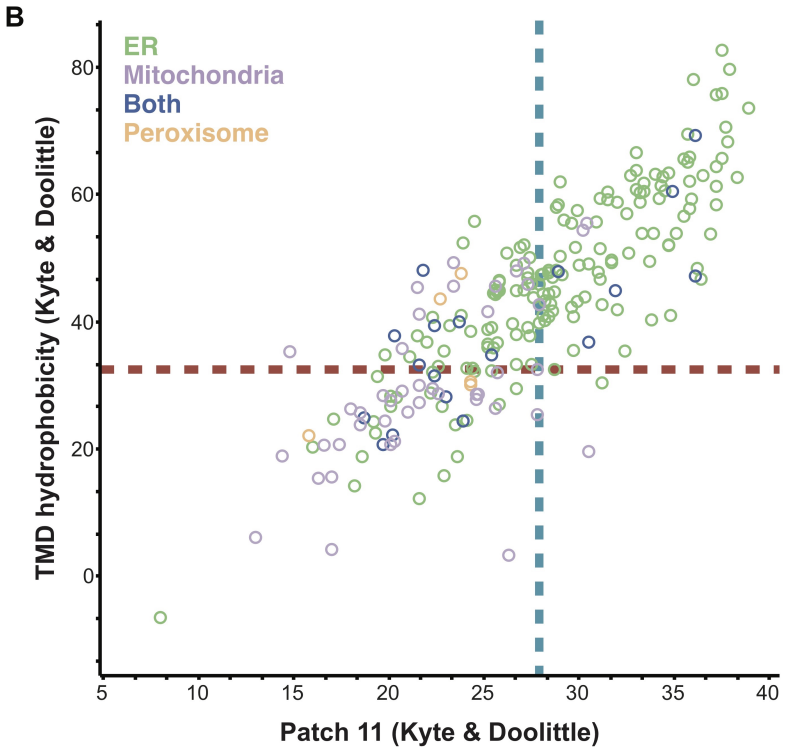



Figure 2.7: **Human ER and mitochondrial TA proteins can be separated by the most hydrophobic 11 residues segment.**
*A)* A table of the with the AUROC values of the best performing hydrophobicity metrics and the overall TMD hydrophobicity, along with their ranking. The number of total misclassified proteins are separated by ER-bound and mitochondria-bound TA proteins. *B)* 2D comparison for the human dataset of TMD hydrophobicity and Patch 11 metrics using the Kyte and Doolittle scale. Hydrophobicities are plotted and TA proteins are colored as in Fig. 2.3 *D*. Unknown TA proteins are not plotted.

TA proteins have tails that are less charged than peroxisomal TA proteins, but more charged than ER-bound TA proteins, which are generally more hydrophobic than mitochondria-bound TA proteins. Previous reports demonstrated the GET pathway fails to insert TA proteins with a sufficiently charged C-terminus (Rao et al., 2016). This selectivity filter was seen at the membrane and cytosolic components were unaffected by the presence of a charge. Perhaps this rejection of TA proteins with a C-terminal charge is seen across all ER targeting pathways in both yeast and humans. To further explore this, we determined anything to be above the hydrophobicity cut-off to be classified as ER-bound and anything below the cut-off to be passed through a charge filter. When analyzing the number of C-terminal positive residues following the TMD of TA proteins that fall below the hydrophobicity cut-off, we find that a benchmark of 3 positive residues best separates ER- and mitochondria-bound TA proteins – mitochondria-bound TA proteins generally contain at least three charged residues. We applied this secondary filter to our best performing yeast metrics (Wheel Face 5 and Wheel Face 7 residues) and the TMD hydrophobicity (Table 2.1). In these cases, the three metrics perform the same, misclassifying 10 TA proteins. Intriguingly, a Patch 15 metric does best, correctly classifying 88% of all yeast TA proteins. A metric utilizing both a helical wheel face and C-terminal charge does slightly better than that using TMD hydrophobicity and charge, but the significance of that improvement is difficult to determine based on this small dataset.

The human dataset is larger, and we sought to apply this tandem metric application to our list of putative TA proteins (Fig. 2.8). Similar to what was observed in the yeast dataset, improvements in classification are seen (Table 1). Interestingly, applying a C-terminal charge sequentially to hydrophobic metrics constrained to a fragment of ~11 residues, either a Patch (TM tendency) or the entire segment (Kyte & Doolittle), and the TMD hydrophobicity metric (Kyte & Doolittle), perform equally well, each misclassifying 38 TA proteins. Most hydrophobicity metrics performed similarly with either scale, suggesting a subset of the TMD is required for correct targeting (Table 2.1). It is clear that in both human and yeast, a combination of hydrophobicity and C-terminal charge filters are necessary for correct classification as was demonstrated in the context of the GET pathway. The hydrophobicity window can be limited to a fraction of the TMD and still perform as well as the entire TMD.

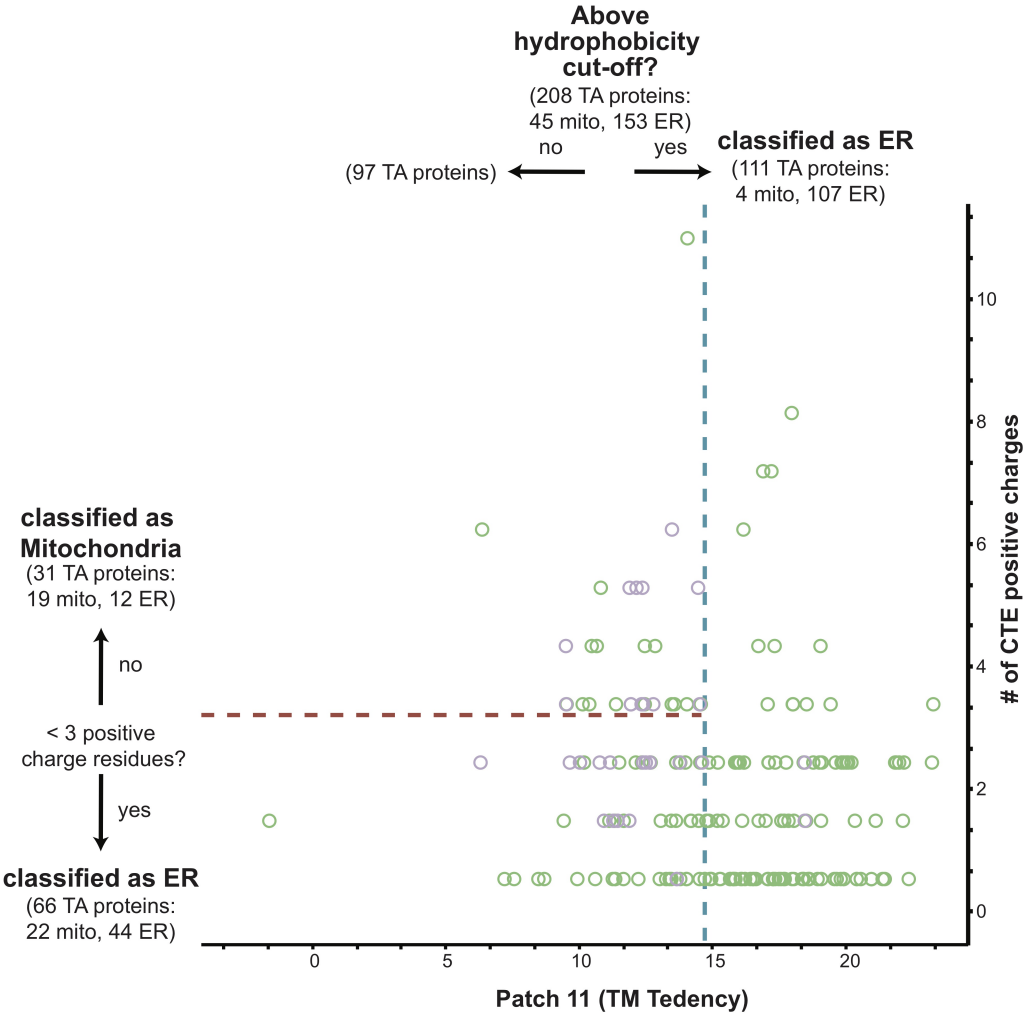



Figure 2.8: **Combining a hydrophobicity and C-terminal charge metric results in a more effective predictor.**
The most hydrophobic eleven amino acid segment of all human TA protein TMDs with known localizations to either the ER (*green*) or mitochondria (*purple*) was calculated using the Kyte & Doolittle scale and plotted along the x-axis. The number of positive charge residues was counted and plotted along the y-axis. The best fit cut-off for the hydrophobicity metric (*blue dotted line*) and charge metric (*red dotted line*) are marked. The number of ER- and mitochondria-bound TA proteins captured in each step is denoted in the corresponding quadrant.

## 2.3 Discussion

Decoding the signaling information in membrane proteins responsible for their correct targeting to cellular membranes is still a mystery. For the class of membrane proteins with a single TMD and no signal peptide, TA proteins, some observations have been made to distinguish between those destined for the ER and those destined for the mitochondria. This report provides an extensive analysis of yeast and human TA proteins to identify a set of criteria to distinguish between ER- and mitochondria-bound TA proteins. This study also includes an expansion of putative TA proteins in both humans and yeast as well as newly determined experimental localization of several yeast TA proteins.

An initial separation by hydrophobicity can be applied to TA proteins, relegating TMDs with high hydrophobicities as ER proteins. A secondary filter can be applied to those below the cut-off classifying TA proteins with at least three charged residues following their TMDs as mitochondria-bound and the rest as ER-bound (Fig. 2.8). This sequential selectivity was noted in the yeast GET pathway (Rao et al., 2016). In this case, it was demonstrated that the cytosolic targeting factors Sgt2 and Get3 bind to optimal TMDs based on a combination of high hydrophobicity and helical propensity. Regardless of hydrophobicity, TA proteins containing a charged C-termini were not inserted into ER microsomes. The analysis here demonstrates that generally ER-bound TA proteins, not just GET clients, lack charges in their C-terminus.

When determining the effectiveness of a hydrophobicity metric alone, metrics that focus on a hydrophobic geometry, a hydrophobic face in yeast and a hydrophobic segment restricted to 11-19 residues in humans, perform better than the hydrophobicity of the entire TMD. Applying the charge filter reveals that total hydrophobicity is as effective as hydrophobic face or segment metrics. Differences in the best performing hydrophobicity metrics between the yeast and human dataset could be explained by the observation that SGTA is more permissive to client binding than Sgt2 (Lin et al., 2021). Collectively, these datasets demonstrate that a fraction of the TMD is necessary and sufficient for correct localization. Interestingly, in the human dataset, some of the best performing metrics are limited to an 11-residue window, concurring with reports that SGTA recognizes TMDs of at least 11 amino acids (Lin et al., 2021).

While biochemical data suggested that clustering hydrophobic residues to one side of a helix increased binding to Sgt2, a co-chaperone in an ER TA protein targeting

pathway, a cellular role of this hydrophobic face remained unclear (Lin et al., 2021). From the bioinformatic analysis and experimental localization data presented here, we demonstrate most yeast ER-bound TA proteins contain a hydrophobic face – made of five to seven adjacent residues along a helical wheel plot. The two components of the GET pathways that directly bind to TA proteins, Sgt2 and Get3, both have binding sites composed of a hydrophobic groove. One could imagine the hydrophobic face in clients buried in the hydrophobic groove of Sgt2 and Get3, enhancing the hydrophobic binding interactions. Perhaps cellular factors involved in targeting TA proteins to the ER recognize this face and future identified ER TA protein binding partners will also feature a helical hand for client binding.

## 2.4 Conclusion

In this work, we provide a comprehensive bioinformatics analysis of naturally occurring TA proteins in the yeast and human genomes. While subtle differences in performance for each geometry metric and hydrophobic scale cannot easily be differentiated by analyzing just wild-type proteins, similar work has helped disentangle the positional dependence of hydrophobicity in the insertion of integral membrane proteins (Hessa, Meindl-Beinker, et al., 2007). Likewise, future work could better define the geometry and hydrophobic scale needed for TA protein targeting by larger scale mutational analyses, perhaps even transforming the question of TA protein targeting into that of sequence selection/enrichment (Fowler and Fields, 2014).

The targeting of TA proteins presents an intriguing and enigmatic problem for understanding the biogenesis of this important class of proteins. How subtle differences in clients modulate the interplay of hand-offs that direct these proteins to the correct membrane remains to be understood. Through *in vivo* imaging of yeast cells and computational analysis, we provide more clarity to client discrimination. A major outcome of this is the clear preference for a hydrophobic face in ER-bound TA proteins of low hydrophobicity. In yeast, this alone is sufficient to predict the destination of a TA protein. In mammals, and likely more broadly in metazoans, while clearly an important component, targets cannot be discrimated by the hydrophobic face alone. For a full understanding, we expect other factors to contribute, reflective of the increased complexity of higher eukaryotes, and perhaps involving more players (Aviram, Costa, et al., 2016).

Interestingly, the classification of protist TA proteins may be simpler than what was

observed in yeast. While the EMC is highly conserved throughout eurkaryotes, none of the complex's 10 subunits have been identified in the protist, *Giardia intestinalis* (Wideman, 2015) and it is speculated that this is the result of a loss in metamonads. As discussed in detail in later chapters, Get3 and several other GET components have been found in organisms outside of the Opisthokonts. Focusing solely on Get3, we see very few differences between the structure of these homologs. As the field expands and begins to study the GET pathway in other eurkaryotes, additional functions may come the light and thus differences in the underlying information encoded in TA proteins for targeting. This study helps prime a foundation to study patterns in TA proteins from other organisms.

## 2.5 Methods

*Assembling a database of putative tail-anchored proteins and their TMDs*

Proteins identified from UniProt (Consortium, 2020) containing a single transmembrane domain within 30 residues of the C-terminus were separated into groups based on their localization reported in UniProt. The topology of all proteins with 3 TMs or fewer was further analyzed using TOPCONS (Tsirigos et al., 2015) to avoid missed single-pass TM proteins. Proteins with a predicted signal peptide (Nielsen, 2017), an annotated transit peptide, problematic cautions, or with a length less than 50 or greater than 1000 residues were excluded. Proteins localized to the ER, golgi apparatus, nucleus, endosome, lysosome, and cell membrane were classified as ER-bound, those localized to the outer mitochondrial membrane were classified as mitochondria-bound, those localized to the peroxisome were classified as peroxisomal proteins, and those with unknown localization were classified as unknown. Proteins with a compositional bias overlapping with the predicted TMD were also excluded. A handful of proteins and their inferred localizations were manually corrected or removed (see notebook and Table S2.1).

*Assessing the predictive power of various hydrophobicity metrics*

We thoroughly examined the metrics relating hydrophobicity, both published and by our own exploration, to better understand their relationship to protein localization. Notably, we recognized that a TMD's hydrophobic moment $< \mu H >$ (Eisenberg et al., 1984) was a poor predictor of localization, e.g. although a Leu18 helix is extremely hydrophobic, it has $< \mu H >= 0$ since opposing hydrophobic residues are penalized in this metric. To address this, we define a metric that capture the presence of a hydrophobic face of the TMD: the maximally hydrophobic cluster on the face.

For this metric we sum the hydrophobicity of residues that orient sequentially on one side of a helix when visualized in a helical wheel diagram. While a range of hydrophobicity scales were predictive using this metric, we selected the TM Tendency scale (Zhao and London, 2006) to characterize the TMDs of putative TA proteins and determined the most predictive window by assessing a range of lengths from 4 to 12 (this would vary from three turns of a helix to six).

By considering sequences with inferred ER or mitochondrial localizations, we calculated the Area Under the Curve of a Receiver Operating Characteristic (AUROC) to assess predictive power. As we are comparing a real-valued metric (hydrophobicity) to a 2-class prediction, the AUROC is better suited for this analysis over others like accuracy or precision (a primer (Swets, Dawes, and Monahan, 2000)). Due to many fewer mitochondrial proteins (i.e. a class imbalance), we also confirmed that ordering hydrophobicity metrics by AUROC was consistent with the ordering produced by the more robust, but less common, Average Precision (see notebook).

*Constructing plasmids for live cell imaging*

A p416ADH-GFP-Fis1 plasmid and a mt-TagBFP described in Rao et al. 2016 were gifted to us from the Walter lab, UCSF (Rao et al., 2016; Okreglak and Walter, 2014) and a Sec63-tdtomato was a gift from Sebastian Schuck, ZMVH, Universitat Heidelberg. TMDs sequences were ordered from Twist Biosciences (San Francisco, CA) with flanking HindIII and XhoI sites. GFP-TMD constructs were made by restriction enzyme digestion (New England Biolabs, USA) of the p416ADH-GFP-Fis1 plasmid and the genes ordered from Twist Biosciences followed by T4 DNA (New England Biolabs, USA) ligation of the template and TMD fragments.

*Live cell imaging*

The yeast strain used are those described in Rao et al. 2016, also a gift from the Walter Lab, UCSF. Strains containing each GFP fused TMD were grown in appropriate selection media. Coverslips were prepped by coating with 0.1mg/mL concavalin A (Sigma, USA) in 0.9% NaCl solution. Cells were immobilized on coverslips at a concentration of 5000 cells/mm$^2$ (plates at 1.8cm$^2$, thus 9x10$^8$ cells/well) and imaged using a Nikon LSM800 (Nikon, Japan). Images were collected at wavelengths 488nm, 514nm, and 581nm and were processed with ImageJ (Schneider, Rasband, and Eliceiri, 2012) and two in-house image processing algorithms.

*Image processing to determine localization*

Yeast cells were segmented using deep learning-based tools. The variable pattern

of DIC images with mixed low and high contrasts for backgrounds and cell bodies (signal variance of each whole image ranging from 67.4 to 2,706.3, a 40x difference – average, median, and standard deviation of signal variance for all images were, respectively 645.6, 563.8, and 419.1) prevented using classical gradient based methods to successfully segment cells. We adopted and compared two contemporary tools, YeastSpotter, a Mask-RCNN method dedicated to yeast cells (Lu et al., 2019), and Cellpose, a generalist method trained on a large pool of cell images (Stringer et al., 2020). Note that the former was not trained on yeast cell images but used a model pre-trained on a larger set of other cell images to build a friendly tool for yeast cell segmentation. Cellpose is a more sophisticated tool whose pre-trained models have learned to segment well based on a myriad of intensity gradient values and image styles. It has shown to achieve high quality segmentation on an extended variety of cell images, including in our yeast cells images, producing superior results when compared to YeastSpotter with the advantage of running faster on GPUs (tested on Nvidia RTX 2080 Ti). We thus exclusively used Cellpose with its cyto pre-trained model to segment yeast cells in all our DIC images. We used maximum intensity projections of up to two or three slices per image stack but mostly a single slice was sufficient to create a single representative image for segmentation. Spurious, tiny, segmented regions whose size were shown to be outliers were automatically removed using an area opening morphological operation.

Individual cells were isolated by applying the mask to the corresponding florescent images of each of the three wavelengths. Masks less than 7.5 $\mu$m$^2$ corresponded to incorrectly identified, incomplete, or out-of-plane cells and were omitted from analysis. Masks were applied to each florescence channel. An empirical threshold was applied to each channel to identify true florescence from background, and the percentage of each cell with co-localized GFP and BFP or GFP and tdTomato was then calculated. Localization was then determined identifying which pair of channels (GFP&BFP vs GFP&tdTomato) had greater overlap, i.e. Overlap$_{GFP\&BFP}$ > Overlap$_{GFP\&tdTomato}$ resulted in a mitochondria annotation. The number of individual cells in each category were counted. Outputs from this algorithm were verified by manually inspecting individual images.

*Code and data availability*

All code employed is available openly at github.com/clemlab/ta_classifier with analysis done in Jupyter Lab/Notebooks using Python 3.6 enabled by Numpy, Pandas, Scikit-Learn, BioPython, bebi103 (Bois, 2020), and Bokeh as well as in Rstu-

dio/Rmarkdown Notebooks enabled by packages within the Tidyverse ecosystem.

## 2.6 Acknowledgements

We thank members of the Clemons lab for support and discussion. We would also like to thank the Caltech Biological Imaging center for the use of their microscopes for our live cell imaging as well as the Caltech Center for Advanced Methods in Biological Image Analysis for aiding in the isolation of individual cells in our images. This work was supported by National Institutes of Health (NIH) Grant R01GM097572 (to WMC), NIH/National Research Service Award Training Grant 5T32GM07616 (to SMS and MYF), and a National Science Foundation Graduate Research fellowship under Grant 1144469 (to SMS).

## 2.7 Tables

| metric | scale | Organism | correct (%) | misclassified mito | misclassified ER | miclassified total |
|---|---|---|---|---|---|---|
| Patch 11 | TM Tendency | *H. sapiens* | 82% | 26 | 12 | 38 |
| TMD | Kyte & Doolittle | *H. sapiens* | 82% | 33 | 5 | 38 |
| Segment 19 | Kyte & Doolittle | *H. sapiens* | 81% | 29 | 9 | 38 |
| Segment 11 | TM Tendency | *H. sapiens* | 81% | 27 | 12 | 39 |
| Segment 15 | TM Tendency | *H. sapiens* | 81% | 25 | 14 | 39 |
| Wheel face 9 | TM Tendency | *H. sapiens* | 81% | 26 | 13 | 39 |
| Wheel face 7 | TM Tendency | *H. sapiens* | 81% | 27 | 13 | 40 |
| Segment 11 | Kyte & Doolittle | *H. sapiens* | 81% | 28 | 12 | 40 |
| Segment 15 | Kyte & Doolittle | *H. sapiens* | 81% | 30 | 10 | 40 |
| Patch 19 | TM Tendency | *H. sapiens* | 80% | 27 | 14 | 41 |
| Patch 15 | TM Tendency | *H. sapiens* | 80% | 30 | 11 | 41 |
| Segment 19 | TM Tendency | *H. sapiens* | 80% | 26 | 15 | 41 |
| Patch 19 | Kyte & Doolittle | *H. sapiens* | 80% | 30 | 12 | 42 |
| Patch 11 | Kyte & Doolittle | *H. sapiens* | 79% | 28 | 15 | 43 |
| Patch 15 | Kyte & Doolittle | *H. sapiens* | 78% | 28 | 17 | 45 |
| TMD | TM Tendency | *H. sapiens* | 78% | 31 | 14 | 45 |
| Wheel face 5 | TM Tendency | *H. sapiens* | 77% | 28 | 19 | 47 |
| Patch 15 | TM Tendency | *S. cerevisiae* | 88% | 7 | 2 | 9 |
| Wheel face 5 | TM Tendency | *S. cerevisiae* | 87% | 6 | 4 | 10 |
| Wheel face 7 | TM Tendency | *S. cerevisiae* | 87% | 6 | 4 | 10 |
| Patch 11 | TM Tendency | *S. cerevisiae* | 87% | 6 | 4 | 10 |
| TMD | TM Tendency | *S. cerevisiae* | 87% | 6 | 4 | 10 |
| Segment 15 | TM Tendency | *S. cerevisiae* | 86% | 6 | 5 | 11 |
| Segment 19 | TM Tendency | *S. cerevisiae* | 86% | 6 | 5 | 11 |
| Patch 19 | TM Tendency | *S. cerevisiae* | 84% | 7 | 5 | 12 |
| Segment 11 | TM Tendency | *S. cerevisiae* | 83% | 6 | 7 | 13 |

Table 2.1: **Best performing hydrophobicity metrics when combined with charge are those restricted to shorter segments of a helix in humans.**
A ranked comparison of the best performing hydrophobicity metric when combined with a C-terminal charge cut-off for both human and yeast TA proteins.

Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | Q16611 | BAK | Both |
| H. sapiens | P10415 | BCL2 | Both |
| H. sapiens | O60238 | BNIP3L | Both |
| H. sapiens | P26678 | PLN | Both |
| H. sapiens | Q12981 | BNIP1 | Both |
| H. sapiens | Q09013 | DMPK | Both |
| H. sapiens | Q969F0 | FATE1 | Both |
| H. sapiens | Q07812 | BAX | Both |
| H. sapiens | Q07817 | BCLX | Both |
| H. sapiens | Q9HD36 | BCLB | Both |
| H. sapiens | Q12983 | BNIP3 | Both |
| H. sapiens | Q9UMX3 | BOK | Both |
| H. sapiens | Q9BWH2 | FUNDC2 | Both |
| H. sapiens | Q07820 | MCL1 BCL2L3 | Both |
| H. sapiens | Q86Y07 | VRK2 | Both |
| H. sapiens | P23763 | VAMP1 | Both |
| H. sapiens | A0A0A0MTJ1 | FKBP8 | Both |
| H. sapiens | I3L3X5 | PLSCR3 | Both |
| H. sapiens | P18031 | PTPN1 | ER |
| H. sapiens | Q13323 | BIK | ER |
| H. sapiens | Q9H305 | CDIP1 | ER |
| H. sapiens | Q5VV42 | CDKAL1 | ER |
| H. sapiens | A4D256 | CDC14C | ER |
| H. sapiens | Q96JN2 | CCDC136 | ER |
| H. sapiens | Q9NXE4 | SMPD4 | ER |
| H. sapiens | Q9H0X9 | ORPL5 | ER |
| H. sapiens | Q9BZF1 | ORP8 | ER |
| H. sapiens | Q9NZM1 | MYOF | ER |
| H. sapiens | Q9HC10 | OTOF | ER |
| H. sapiens | Q9HCU5 | SEC12 | ER |
| H. sapiens | O15162 | PLSCR1 | ER |
| H. sapiens | Q9NRQ2 | PLSCR4 | ER |
| H. sapiens | A0PG75 | PLSCR5 | ER |
| H. sapiens | Q9NRY7 | PLSCR2 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | P50876 | RNF144A | ER |
| H. sapiens | Q9NZ42 | PSENEN | ER |
| H. sapiens | Q96CS7 | EVT2 | ER |
| H. sapiens | Q6ZNB6 | NFXL1 | ER |
| H. sapiens | P17706 | PTPN2 | ER |
| H. sapiens | Q7Z6L0 | PRRT2 | ER |
| H. sapiens | Q9Y6X1 | SERP1 RAMP4 | ER |
| H. sapiens | P61266 | STX1B | ER |
| H. sapiens | Q9UNK0 | STX8 | ER |
| H. sapiens | Q16623 | STX1A | ER |
| H. sapiens | Q8WXE9 | STON2 | ER |
| H. sapiens | Q86Y82 | STX12 | ER |
| H. sapiens | Q7Z699 | SPRED1 | ER |
| H. sapiens | Q9P2W9 | STX18 | ER |
| H. sapiens | Q13190 | STX5 | ER |
| H. sapiens | O15400 | STX7 | ER |
| H. sapiens | P32856 | STX2 | ER |
| H. sapiens | O60499 | STX10 | ER |
| H. sapiens | Q13277 | STX3 | ER |
| H. sapiens | O14662 | STX16 | ER |
| H. sapiens | Q5QGT7 | RTP2 | ER |
| H. sapiens | P59025 | RTP1 | ER |
| H. sapiens | Q9BQQ7 | RTP3 | ER |
| H. sapiens | Q8N205 | SYNE4 | ER |
| H. sapiens | Q6ZMZ3 | SYNE3 | ER |
| H. sapiens | B2RUZ4 | SMIM1 | ER |
| H. sapiens | Q7Z698 | SPRED2 | ER |
| H. sapiens | Q96DX8 | RTP4 | ER |
| H. sapiens | Q96QK8 | SMIM14 | ER |
| H. sapiens | Q14BN4 | SLMAP | ER |
| H. sapiens | Q86T96 | RNF180 | ER |
| H. sapiens | Q8N8N0 | RNF152 | ER |
| H. sapiens | Q12846 | STX4 | ER |
| H. sapiens | O43752 | STX6 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | P60059 | SEC61G | ER |
| H. sapiens | Q14D33 | RTP5 | ER |
| H. sapiens | P60468 | SEC61B | ER |
| H. sapiens | Q8N6R1 | SERP2 | ER |
| H. sapiens | P01850 | TRBC1 | ER |
| H. sapiens | P03986 | TRGC2 | ER |
| H. sapiens | Q629K1 | TRIQK | ER |
| H. sapiens | A0A5B9 | TRBC2 | ER |
| H. sapiens | Q9NSU2 | TREX1 | ER |
| H. sapiens | P01848 | TRAC | ER |
| H. sapiens | B7Z8K6 | TRDC | ER |
| H. sapiens | Q96D59 | RNF183 | ER |
| H. sapiens | P00167 | CYB5A | ER |
| H. sapiens | O75923 | DYSF | ER |
| H. sapiens | P50402 | EMD | ER |
| H. sapiens | O42043 | ERVK-18 | ER |
| H. sapiens | Q52LJ0 | FAM98B | ER |
| H. sapiens | Q9NYM9 | BET1L | ER |
| H. sapiens | P54710 | FXYD2 | ER |
| H. sapiens | O95415 | BRI3 | ER |
| H. sapiens | Q9BXU9 | CALN1 | ER |
| H. sapiens | O15155 | BET1 | ER |
| H. sapiens | Q86V35 | CABP7 | ER |
| H. sapiens | Q01740 | FMO1 | ER |
| H. sapiens | P49326 | FMO5 | ER |
| H. sapiens | P31513 | FMO3 | ER |
| H. sapiens | Q8N8J7 | FAM241A | ER |
| H. sapiens | P31512 | FMO4 | ER |
| H. sapiens | Q9P0K9 | FRRS1L | ER |
| H. sapiens | Q9Y2H6 | FNDC3A | ER |
| H. sapiens | P13164 | IFITM1 | ER |
| H. sapiens | Q01629 | IFITM2 | ER |
| H. sapiens | Q01628 | IFITM3 | ER |
| H. sapiens | Q8TBA6 | GOLGA5 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | O95249 | GOSR1 | ER |
| H. sapiens | Q14789 | GOLGB1 | ER |
| H. sapiens | Q96JJ6 | JPH4 | ER |
| H. sapiens | Q9HDC5 | JPH1 | ER |
| H. sapiens | Q9BR39 | JPH2 | ER |
| H. sapiens | Q8WXH2 | JPH3 | ER |
| H. sapiens | Q99732 | LITAF | ER |
| H. sapiens | O75427 | LRCH4 | ER |
| H. sapiens | Q9Y2L9 | LRCH1 | ER |
| H. sapiens | Q3KP22 | MAJIN | ER |
| H. sapiens | Q86Z14 | KLB | ER |
| H. sapiens | Q9Y6H6 | KCNE3 | ER |
| H. sapiens | P42167 | TMPO | ER |
| H. sapiens | P30519 | HMOX2 | ER |
| H. sapiens | Q8WWP7 | GIMAP1 | ER |
| H. sapiens | Q96F15 | GIMAP5 | ER |
| H. sapiens | P09601 | HMOX1 | ER |
| H. sapiens | Q9UPX6 | MINAR1 | ER |
| H. sapiens | Q8NHP6 | MOSPD2 | ER |
| H. sapiens | P51648 | ALDH3A2 | ER |
| H. sapiens | Q8N2K1 | UBE2J2 | ER |
| H. sapiens | O94966 | USP19 | ER |
| H. sapiens | Q9NZ43 | USE1 | ER |
| H. sapiens | Q9P0L0 | VAP33 | ER |
| H. sapiens | O95159 | ZFPL1 | ER |
| H. sapiens | Q5T7W0 | ZNF618 | ER |
| H. sapiens | O14653 | GOSR2 | ER |
| H. sapiens | P51809 | VAMP7 | ER |
| H. sapiens | Q9BV40 | VAMP8 | ER |
| H. sapiens | Q9UEU0 | VTI1B | ER |
| H. sapiens | Q96AJ9 | VTI1A | ER |
| H. sapiens | O95292 | VAPB | ER |
| H. sapiens | O95183 | VAMP5 | ER |
| H. sapiens | Q15836 | VAMP3 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | O75379 | VAMP4 | ER |
| H. sapiens | Q9Y385 | UBE2J1 | ER |
| H. sapiens | P63027 | VAMP2 | ER |
| H. sapiens | A0A1W2PPG1 | GOSR2 | ER |
| H. sapiens | A0A087WWT2 | NRN1 | ER |
| H. sapiens | E9PLT1 | CD36 | ER |
| H. sapiens | E7ENI6 | ICA1 | ER |
| H. sapiens | E9PN33 | STX3 | ER |
| H. sapiens | X6R383 | SETDB2 | ER |
| H. sapiens | K7EJC8 | GOSR1 | ER |
| H. sapiens | A0A1W2PRH6 | PAX6 | ER |
| H. sapiens | D6RE10 | ELOVL7 | ER |
| H. sapiens | F5H2S3 | P2RX4 | ER |
| H. sapiens | F8WAT4 | PAPOLG | ER |
| H. sapiens | F5H3K6 | SPI1 | ER |
| H. sapiens | E9PE96 | PCLO | ER |
| H. sapiens | D6RF86 | CDH6 | ER |
| H. sapiens | E7ETP9 | LAMP3 | ER |
| H. sapiens | A0A1W2PRF6 | SCARB2 | ER |
| H. sapiens | B4DSN5 | PTPN1 | ER |
| H. sapiens | D6RBD7 | EEF1E1 | ER |
| H. sapiens | A0A087WTJ2 | GIMAP1-GIMAP5 | ER |
| H. sapiens | A0A1W2PS81 | GOSR2 | ER |
| H. sapiens | A0A087WWT0 | JPH4 | ER |
| H. sapiens | A0A0J9YW33 | STX3 | ER |
| H. sapiens | C9JUH5 | SERP1 | ER |
| H. sapiens | H7C410 | GPC1 | ER |
| H. sapiens | U3KQS5 | TATDN1 | ER |
| H. sapiens | G5EA09 | SDCBP | ER |
| H. sapiens | B1AL79 | PKN2 | ER |
| H. sapiens | B7Z5N5 | SMAD2 | ER |
| H. sapiens | I3L3H3 | P2RX1 | ER |
| H. sapiens | A0A087WT82 | GPC6 | ER |
| H. sapiens | F5H895 | DAD1 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | F2Z2S5 | SERP2 | ER |
| H. sapiens | A0A1B0GTF8 | EPB41L3 | ER |
| H. sapiens | E9PCT3 | CAV2 | ER |
| H. sapiens | F8WBE5 | TFRC | ER |
| H. sapiens | K7EQG9 | PTPN2 | ER |
| H. sapiens | A0A0U1RQC9 | TP53 | ER |
| H. sapiens | F8WCE5 | LMLN | ER |
| H. sapiens | A0SDD8 | CLDN16 | ER |
| H. sapiens | B5MCA4 | EPCAM | ER |
| H. sapiens | Q5HY57 | EMD | ER |
| H. sapiens | B4DJ94 | ATP9B | ER |
| H. sapiens | Q86XC5 | TMEM97 | ER |
| H. sapiens | Q9NRY6 | PLSCR3 | Mit |
| H. sapiens | Q7Z419 | RNF144B | Mit |
| H. sapiens | P56378 | MP68 | Mit |
| H. sapiens | P57105 | OMP25 | Mit |
| H. sapiens | Q9P0U1 | TOM7 | Mit |
| H. sapiens | Q8N4H5 | TOM5 | Mit |
| H. sapiens | Q8WWH4 | ASZ1 | Mit |
| H. sapiens | P27338 | MAOB | Mit |
| H. sapiens | Q9BXK5 | BCL2L13 | Mit |
| H. sapiens | O43169 | CYB5B CYB5M | Mit |
| H. sapiens | Q14318 | FKBP8 | Mit |
| H. sapiens | Q9Y3D6 | FIS1 | Mit |
| H. sapiens | Q96I36 | COX14 | Mit |
| H. sapiens | O00198 | HRK | Mit |
| H. sapiens | Q14410 | GK2 GKP2 GKTA | Mit |
| H. sapiens | Q9GZY8 | MFF | Mit |
| H. sapiens | Q7Z434 | MAVS | Mit |
| H. sapiens | Q8IXI1 | RHOT2 | Mit |
| H. sapiens | Q13505 | MTX1 | Mit |
| H. sapiens | Q8IXI2 | RHOT1 | Mit |
| H. sapiens | P21397 | MAOA | Mit |
| H. sapiens | Q14409 | GK3P | Mit |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | Q96IX5 | USMG5 | Mit |
| H. sapiens | A0A087WT64 | MCL1 | Mit |
| H. sapiens | E9PH05 | FAM162A | Mit |
| H. sapiens | A0A0C4DFQ1 | MTX1 | Mit |
| H. sapiens | A0A087WZY2 | TOMM7 | Mit |
| H. sapiens | C9JU26 | ATP5MF | Mit |
| H. sapiens | S4R2X2 | SFXN1 | Mit |
| H. sapiens | H7BXZ6 | RHOT1 | Mit |
| H. sapiens | A0A0A0MS29 | MFF | Mit |
| H. sapiens | J3KNF8 | CYB5B | Mit |
| H. sapiens | P56134 | ATP5J2 | MitIn |
| H. sapiens | O43676 | NDUFB3 | MitIn |
| H. sapiens | O14957 | UQCR11 | MitIn |
| H. sapiens | Q9UDW1 | UQCR10 | MitIn |
| H. sapiens | O95168 | NDUFB4 | MitIn |
| H. sapiens | P09669 | COX6C | MitIn |
| H. sapiens | Q9Y2R0 | COA3 | MitIn |
| H. sapiens | Q96AQ8 | MCUR1 | MitIn |
| H. sapiens | F8WAR4 | CHCHD3 | MitIn |
| H. sapiens | D6R9C3 | COX7A2 | MitIn |
| H. sapiens | A0A087WU07 | MINOS1 | MitIn |
| H. sapiens | A0A087WYS9 | SURF1 | MitIn |
| H. sapiens | C9IZW8 | NDUFB2 | MitIn |
| H. sapiens | O96011 | PEX11B | Pex |
| H. sapiens | Q8NFP0 | PXT1 | Pex |
| H. sapiens | P53816 | PLA2G16 | Pex |
| H. sapiens | Q5T8D3 | ACBD5 | Pex |
| H. sapiens | B7Z2R7 | ACBD5 | Pex |
| H. sapiens | Q9P0B6 | CCDC167 | Unknown |
| H. sapiens | Q8N111 | CEND1 | Unknown |
| H. sapiens | Q8WVX3 | C4orf3 | Unknown |
| H. sapiens | Q9H7X2 | C1orf115 | Unknown |
| H. sapiens | Q6ZSY5 | PPP1R3F | Unknown |
| H. sapiens | Q9UF11 | PLEKHB1 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | Q6ZS82 | RGS9BP | Unknown |
| H. sapiens | Q16821 | PPP1R3A | Unknown |
| H. sapiens | Q9NS64 | RPRM | Unknown |
| H. sapiens | Q6IEE8 | SLFN12L | Unknown |
| H. sapiens | Q9NRQ5 | SMCO4 | Unknown |
| H. sapiens | Q8NCU8 | SMIM37 | Unknown |
| H. sapiens | Q96HG1 | SMIM10 | Unknown |
| H. sapiens | Q71RC9 | SMIM5 | Unknown |
| H. sapiens | P0DL12 | SMIM17 | Unknown |
| H. sapiens | Q8WVI0 | SMIM4 | Unknown |
| H. sapiens | A0A1B0GUA5 | SMIM32 | Unknown |
| H. sapiens | Q96KF7 | SMIM8 | Unknown |
| H. sapiens | Q8TC41 | RNF217 | Unknown |
| H. sapiens | Q9Y228 | TRAF3IP3 | Unknown |
| H. sapiens | Q5JXX7 | TMEM31 | Unknown |
| H. sapiens | Q2MJR0 | SPRED3 | Unknown |
| H. sapiens | L0R6Q1 | SLC35A4 | Unknown |
| H. sapiens | O75920 | SERF1A | Unknown |
| H. sapiens | Q9H4I3 | TRABD | Unknown |
| H. sapiens | A6NCQ9 | RNF222 | Unknown |
| H. sapiens | Q5SWX8 | ODR4 | Unknown |
| H. sapiens | Q8N326 | C10orf111 | Unknown |
| H. sapiens | Q96LL3 | C16orf92 | Unknown |
| H. sapiens | Q6P4D5 | FAM122C | Unknown |
| H. sapiens | Q96D05 | FAM241B | Unknown |
| H. sapiens | Q8N7S6 | ARIH2OS | Unknown |
| H. sapiens | Q8IVJ8 | APRG1 | Unknown |
| H. sapiens | Q86W74 | ANKRD46 | Unknown |
| H. sapiens | Q8WVC6 | DCAKD | Unknown |
| H. sapiens | Q2WGJ9 | FER1L6 | Unknown |
| H. sapiens | Q5RGS3 | FAM74A1 | Unknown |
| H. sapiens | A9Z1Z3 | FER1L4 | Unknown |
| H. sapiens | Q53EP0 | FNDC3B | Unknown |
| H. sapiens | Q96JQ2 | CLMN | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | Q9NPU4 | C14orf132 | Unknown |
| H. sapiens | Q6ZS62 | COLCA1 | Unknown |
| H. sapiens | A1L1A6 | IGSF23 | Unknown |
| H. sapiens | Q8IUY3 | GRAMD2A | Unknown |
| H. sapiens | Q9NWW9 | HRASLS2 | Unknown |
| H. sapiens | Q9UL19 | RARRES3 | Unknown |
| H. sapiens | Q68G75 | LEMD1 | Unknown |
| H. sapiens | P59773 | KIAA1024L | Unknown |
| H. sapiens | Q9HDD0 | HRASLS | Unknown |
| H. sapiens | Q96EZ4 | MYEOV | Unknown |
| H. sapiens | A0A024RCL3 | MICA | Unknown |
| H. sapiens | A0A087WXU0 | RMND1 | Unknown |
| H. sapiens | I3L1J9 | TNFRSF12A | Unknown |
| H. sapiens | E9PLR7 | RNF121 | Unknown |
| H. sapiens | H0YIU3 | RNASEK | Unknown |
| H. sapiens | E7EX18 | MPV17 | Unknown |
| H. sapiens | H0YNW0 | SLC12A1 | Unknown |
| H. sapiens | A0A087WWM7 | MME | Unknown |
| H. sapiens | F5GZV7 | VAMP1 | Unknown |
| H. sapiens | A0A0D9SGD9 | SLFN12L | Unknown |
| H. sapiens | J3KR13 | FOLR2 | Unknown |
| H. sapiens | A0A087X240 | EFNA5 | Unknown |
| H. sapiens | A0A1W2PRR9 | EGFR | Unknown |
| H. sapiens | C9JD05 | FSD1L | Unknown |
| H. sapiens | C9JXZ5 | VAMP8 | Unknown |
| H. sapiens | E9PQR3 | FTH1 | Unknown |
| H. sapiens | J3KNC7 | CYB5A | Unknown |
| H. sapiens | J3KPI8 | GPR139 | Unknown |
| H. sapiens | E9PQY3 | ACP2 | Unknown |
| H. sapiens | E7EPM7 | COQ2 | Unknown |
| H. sapiens | A0A087X286 | CKLF-CMTM1 | Unknown |
| H. sapiens | V9GYT2 | ANKRD29 | Unknown |
| H. sapiens | K7EQB1 | STX8 | Unknown |
| H. sapiens | E9PAR0 | FKBP11 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | J3QS48 | MPDU1 | Unknown |
| H. sapiens | H7BXF4 | SMPD4 | Unknown |
| H. sapiens | F8WDY4 | TMBIM1 | Unknown |
| H. sapiens | J3KN43 | TMEM33 | Unknown |
| H. sapiens | A8MTT8 | ZNF286A | Unknown |
| H. sapiens | C9JYK0 | LRCH4 | Unknown |
| H. sapiens | E5RFY6 | RNF217 | Unknown |
| H. sapiens | E9PFI9 | ZP3 | Unknown |
| H. sapiens | A0A0J9YWK4 | HBB | Unknown |
| H. sapiens | A0A0D9SFF9 | MELK | Unknown |
| H. sapiens | A0A1W2PPL1 | SPIN3 | Unknown |
| H. sapiens | F8WF90 | ARL6IP5 | Unknown |
| H. sapiens | F5H0W1 | DPY19L2 | Unknown |
| H. sapiens | F5H543 | IYD | Unknown |
| H. sapiens | A0A1W2PRW4 | PLA2G16 | Unknown |
| H. sapiens | E9PKL4 | C11orf96 | Unknown |
| H. sapiens | E9PNH0 | OSBPL5 | Unknown |
| H. sapiens | E9PJ90 | HBS1L | Unknown |
| H. sapiens | E9PPZ2 | NPEPPS | Unknown |
| H. sapiens | C9IZ55 | MALL | Unknown |
| H. sapiens | A8MPV4 | MPV17 | Unknown |
| H. sapiens | H0YNT6 | FES | Unknown |
| H. sapiens | F5H1L9 | JPH4 | Unknown |
| H. sapiens | A0A087X175 | SLC38A3 | Unknown |
| H. sapiens | A0A286YEN9 | C5orf60 | Unknown |
| H. sapiens | A0A1W2PP90 | ST3GAL5 | Unknown |
| H. sapiens | H3BUG9 | TMEM202 | Unknown |
| H. sapiens | A0A1W2PQZ3 | HLA-B | Unknown |
| H. sapiens | D6R9K1 | CLDND1 | Unknown |
| H. sapiens | A0A2R8Y7N0 | EPB41 | Unknown |
| H. sapiens | A0A087WWT8 | GDAP1L1 | Unknown |
| H. sapiens | K7EJ34 | RETREG3 | Unknown |
| H. sapiens | A0A0A0MRG8 | BAK1 | Unknown |
| H. sapiens | I3L376 | TVP23B | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | F5GYX3 | SEMA7A | Unknown |
| H. sapiens | Q5VTX9 | IFNLR1 | Unknown |
| H. sapiens | J3KP61 | PLD5 | Unknown |
| H. sapiens | A0A087X0T8 | CADM1 | Unknown |
| H. sapiens | J3KSN8 | SMIM21 | Unknown |
| H. sapiens | F8WAW2 | KIAA0319L | Unknown |
| H. sapiens | F8W1Z3 | CERS5 | Unknown |
| H. sapiens | A0A1B0GW78 | RASGEF1B | Unknown |
| H. sapiens | H7C593 | MFSD1 | Unknown |
| H. sapiens | A0A0G2JM16 | MUC4 | Unknown |
| H. sapiens | E9PM16 | ZNF7 | Unknown |
| H. sapiens | E5RFT6 | LYPLA1 | Unknown |
| H. sapiens | E9PM70 | CYB561D1 | Unknown |
| H. sapiens | C9JQU6 | ARL6IP5 | Unknown |
| H. sapiens | J3QLU8 | PEMT | Unknown |
| H. sapiens | E9PQQ2 | MYB | Unknown |
| H. sapiens | A0A0A0MRG3 | ZNF138 | Unknown |
| H. sapiens | A0A0D9SF04 | CLN3 | Unknown |
| H. sapiens | F8W782 | ADIPOR1 | Unknown |
| H. sapiens | F8WEP4 | CHL1 | Unknown |
| H. sapiens | J3KRT1 | DHX38 | Unknown |
| H. sapiens | B5MEG5 | USP19 | Unknown |
| H. sapiens | D6RHV8 | TMEM175 | Unknown |
| H. sapiens | I3L1G0 | SLC5A11 | Unknown |
| H. sapiens | F8WCS3 | POLR1B | Unknown |
| H. sapiens | A0A140TA65 | CES5A | Unknown |
| H. sapiens | F8WCU3 | SLC30A6 | Unknown |
| H. sapiens | D6R9B4 | CD164 | Unknown |
| H. sapiens | F8VXV4 | SLC48A1 | Unknown |
| H. sapiens | E9PIV8 | CKLF-CMTM1 | Unknown |
| H. sapiens | F8VWE0 | TSPAN31 | Unknown |
| H. sapiens | A0A1B0GUE0 | JAKMIP1 | Unknown |
| H. sapiens | F5H7K7 | LPCAT3 | Unknown |
| H. sapiens | D6RB93 | ZNF451 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | E9PM26 | MS4A7 | Unknown |
| H. sapiens | F8WF83 | SLC9A9 | Unknown |
| H. sapiens | F5H7G2 | RGMA | Unknown |
| H. sapiens | E5RI04 | ANKRD46 | Unknown |
| H. sapiens | D6RJC0 | SLC41A3 | Unknown |
| H. sapiens | E9PJF1 | MYB | Unknown |
| H. sapiens | F8WEW7 | PORCN | Unknown |
| H. sapiens | F8WF33 | ARL6IP5 | Unknown |
| H. sapiens | J3KPT4 | TRABD | Unknown |
| H. sapiens | H0YL57 | RPLP1 | Unknown |
| H. sapiens | E9PMW8 | COP1 | Unknown |
| H. sapiens | A0A1B0GU12 | ATP6AP2 | Unknown |
| H. sapiens | E9PKL6 | OR51E1 | Unknown |
| H. sapiens | K7EPN3 | RAMP2 | Unknown |
| H. sapiens | A0A0G2JQ71 | ZNF66 | Unknown |
| H. sapiens | A0A0A0MTQ3 | CFAP54 | Unknown |
| H. sapiens | A0A0A0MSB7 | CALN1 | Unknown |
| H. sapiens | F8WDW0 | LMBR1 | Unknown |
| H. sapiens | A0A0U1RQZ5 | ENTPD1 | Unknown |
| H. sapiens | D6RI03 | TSPAN17 | Unknown |
| H. sapiens | V9GYR6 | ADPRM | Unknown |
| H. sapiens | J3QKR4 | ICAM2 | Unknown |
| H. sapiens | D6RC55 | OCIAD1 | Unknown |
| H. sapiens | C9JE17 | CCDC136 | Unknown |
| H. sapiens | C9JU31 | CCDC136 | Unknown |
| H. sapiens | F6VI00 | ACOT2 | Unknown |
| H. sapiens | H3BTX6 | ARL6IP1 | Unknown |
| H. sapiens | F8WCI3 | CDK5RAP2 | Unknown |
| H. sapiens | F8WB21 | SYS1 | Unknown |
| H. sapiens | A6NG31 | RARRES3 | Unknown |
| H. sapiens | F8WCL9 | ECE2 | Unknown |
| H. sapiens | A0A1B0GU51 | C14orf132 | Unknown |
| H. sapiens | J3KTR2 | PEMT | Unknown |
| H. sapiens | F8WDI1 | C3orf33 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | F8WDN0 | URGCP | Unknown |
| H. sapiens | E9PN09 | SLC36A4 | Unknown |
| H. sapiens | F8VQZ6 | CERS5 | Unknown |
| H. sapiens | F8W0W6 | SNRPF | Unknown |
| H. sapiens | F5GX39 | TMED2 | Unknown |
| H. sapiens | F8WEN8 | LMBR1 | Unknown |
| H. sapiens | A0A0G2JN91 | NCR1 | Unknown |
| H. sapiens | F8WAW3 | GPR156 | Unknown |
| H. sapiens | D6RDM3 | SLC41A3 | Unknown |
| H. sapiens | D6RBY2 | TMEM33 | Unknown |
| H. sapiens | A0A0A6YYJ0 | MSANTD3-TMEFF1 | Unknown |
| H. sapiens | D6RBP2 | GYPB | Unknown |
| H. sapiens | A0A0G2JMZ5 | UGT2B15 | Unknown |
| H. sapiens | A0A075B785 | RELCH | Unknown |
| H. sapiens | A0A087WZR4 | FCGR3B | Unknown |
| H. sapiens | F8VSK7 | CERS5 | Unknown |
| H. sapiens | F5H0T7 | SLC22A6 | Unknown |
| H. sapiens | B5MC89 | THADA | Unknown |
| H. sapiens | E9PRZ6 | CDC27 | Unknown |
| H. sapiens | F5H5G1 | LSAMP | Unknown |
| H. sapiens | F8W1K4 | CERS5 | Unknown |
| H. sapiens | F6WFR7 | NTM | Unknown |
| H. sapiens | H3BP21 | NFAT5 | Unknown |
| H. sapiens | A0A2R8YF92 | SEL1L2 | Unknown |
| H. sapiens | E9PR36 | MTNR1B | Unknown |
| H. sapiens | F8WB98 | GGCX | Unknown |
| H. sapiens | C9JAX8 | SMIM4 | Unknown |
| H. sapiens | H3BS23 | MOSMO | Unknown |
| H. sapiens | A0A0D9SFD8 | CCDC163 | Unknown |
| H. sapiens | E9PFA2 | WDR17 | Unknown |
| H. sapiens | E7EQN9 | INPP4B | Unknown |
| H. sapiens | A0A087WX97 | BCL2L13 | Unknown |
| H. sapiens | E9PHR9 | PLSCR4 | Unknown |
| H. sapiens | F8WE64 | ELP6 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | X6RLY7 | CACNA2D4 | Unknown |
| H. sapiens | F8WCA0 | VAMP2 | Unknown |
| H. sapiens | G3V5F3 | SCFD1 | Unknown |
| H. sapiens | H3BQA3 | PDPK1 | Unknown |
| H. sapiens | E9PM87 | PTPN22 | Unknown |
| H. sapiens | F8WDI5 | STIMATE | Unknown |
| H. sapiens | F5H4H7 | CLEC12B | Unknown |
| H. sapiens | K7ENK9 | VAMP2 | Unknown |
| H. sapiens | A0A087WT28 | CD200R1L | Unknown |
| H. sapiens | G3V232 | ADSSL1 | Unknown |
| H. sapiens | F8W1N7 | CERS5 | Unknown |
| H. sapiens | E9PRZ2 | PGAP2 | Unknown |
| H. sapiens | A0A182DWE8 | CFAP47 | Unknown |
| H. sapiens | Q8TDQ4 | TMEM222 | Unknown |
| H. sapiens | I3L1D2 | MPDU1 | Unknown |
| H. sapiens | A0A0G2JJ55 | MICA | Unknown |
| H. sapiens | E9PC20 | RAMP1 | Unknown |
| H. sapiens | H0YNL7 | PIGH | Unknown |
| H. sapiens | E9PKT4 | TMEM123 | Unknown |
| H. sapiens | G8JLJ3 | SMIM29 | Unknown |
| H. sapiens | G3V1A8 | LY6G6C | Unknown |
| H. sapiens | A6NGS0 | UBE2J2 | Unknown |
| H. sapiens | F5H3M3 | MANSC1 | Unknown |
| H. sapiens | K4JQN1 | BAX | Unknown |
| H. sapiens | A0A075B778 | ABCA5 | Unknown |
| H. sapiens | A0A1W2PR24 | ST3GAL5 | Unknown |
| H. sapiens | A0A0G2JP96 | LILRA1 | Unknown |
| H. sapiens | M9MML0 | FCGR3A | Unknown |
| H. sapiens | A0A2R8Y694 | SLC19A3 | Unknown |
| H. sapiens | F8WDB3 | ARF4 | Unknown |
| H. sapiens | F8WE00 | MFSD9 | Unknown |
| H. sapiens | J3QS78 | CD7 | Unknown |
| H. sapiens | D6RCD9 | TMEM175 | Unknown |
| H. sapiens | F2Z397 | TMEM184B | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | M0R1X3 | CEACAM8 | Unknown |
| H. sapiens | S4R453 | KCNMA1 | Unknown |
| H. sapiens | Q0P6N6 | NRG4 | Unknown |
| H. sapiens | F2Z2J3 | COA1 | Unknown |
| H. sapiens | I3L1Z6 | ABCC6 | Unknown |
| H. sapiens | F8WCB8 | FTO | Unknown |
| H. sapiens | K7ENB6 | SLC7A10 | Unknown |
| H. sapiens | F5H326 | LDHC | Unknown |
| H. sapiens | E9PKZ1 | SLC16A4 | Unknown |
| H. sapiens | M0R2F1 | KCNN4 | Unknown |
| H. sapiens | G3V5W3 | SOS2 | Unknown |
| H. sapiens | Q4KN23 | KIR3DS1 | Unknown |
| H. sapiens | G3V1I3 | FAM9C | Unknown |
| H. sapiens | F8VRN7 | TMEM116 | Unknown |
| H. sapiens | A0A0G2JJ84 | BTNL2 | Unknown |
| H. sapiens | Q8WZ67 | KLRK1 | Unknown |
| H. sapiens | F5GWC9 | TMEM91 | Unknown |
| H. sapiens | B7Z596 | TPM1 | Unknown |
| H. sapiens | C9JKN6 | THSD7B | Unknown |
| H. sapiens | G3V248 | IFI27L1 | Unknown |
| H. sapiens | G3XAK3 | CLIP4 | Unknown |
| H. sapiens | A1A4Z5 | TRPC7 | Unknown |
| H. sapiens | C9J7K9 | PLSCR1 | Unknown |
| H. sapiens | H3BUX2 | CYB5B | Unknown |
| H. sapiens | S4R3Y8 | TMEM91 | Unknown |
| H. sapiens | F5H5K1 | LRRC37B | Unknown |
| H. sapiens | A0A286YFJ5 | MFSD8 | Unknown |
| H. sapiens | F8W7G1 | CD200 | Unknown |
| H. sapiens | F8VVR0 | MRPL42 | Unknown |
| H. sapiens | A0A087X1Q6 | TARM1 | Unknown |
| H. sapiens | F8WCC4 | C3orf18 | Unknown |
| H. sapiens | K7EQ13 | G6PC3 | Unknown |
| H. sapiens | F8WEV1 | MAATS1 | Unknown |
| H. sapiens | A0A0A0MT53 | CD200R1L | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | E9PHY6 | LRRC8C | Unknown |
| H. sapiens | H7BXH0 | KCTD20 | Unknown |
| H. sapiens | E9PIJ2 | CYB561D1 | Unknown |
| H. sapiens | V9GYC5 | COMMD7 | Unknown |
| H. sapiens | B5MCI6 | MEMO1 | Unknown |
| H. sapiens | E9PQX4 | TDRKH | Unknown |
| H. sapiens | K7ELD9 | SYNGR2 | Unknown |
| H. sapiens | F5H038 | CLEC1A | Unknown |
| H. sapiens | K7EIN4 | TMED1 | Unknown |
| H. sapiens | Q5SNW4 | CLCN6 | Unknown |
| H. sapiens | E9PQJ6 | BET1L | Unknown |
| H. sapiens | F2Z2P5 | ERGIC1 | Unknown |
| H. sapiens | F8WDT4 | SUN3 | Unknown |
| H. sapiens | D6RE04 | PLRG1 | Unknown |
| H. sapiens | J3KST8 | CRLF3 | Unknown |
| H. sapiens | J3KRW3 | CEP95 | Unknown |
| H. sapiens | H3BNZ7 | C16orf95 | Unknown |
| H. sapiens | A2A2E0 | MANBAL | Unknown |
| H. sapiens | E7ETC6 | PDPN | Unknown |
| H. sapiens | A0A0H2UH41 | POTEM | Unknown |
| H. sapiens | I3L380 | ABHD12 | Unknown |
| H. sapiens | H0Y870 | TMEM222 | Unknown |
| H. sapiens | F8WCD4 | TMEM184B | Unknown |
| H. sapiens | A0A0A0MS18 | RAD51B | Unknown |
| H. sapiens | E5RK16 | FAXDC2 | Unknown |
| H. sapiens | A0A087WXA9 | KIZ | Unknown |
| H. sapiens | I3L072 | C17orf80 | Unknown |
| H. sapiens | K7EPU5 | SPRED3 | Unknown |
| H. sapiens | F8W1G5 | RNASEK | Unknown |
| H. sapiens | Q5T4Q8 | CD72 | Unknown |
| H. sapiens | A0A1W2PRT0 | ST3GAL5 | Unknown |
| H. sapiens | H3BU94 | SNAP23 | Unknown |
| H. sapiens | A0A2R8YEW2 | CYSTM1 | Unknown |
| H. sapiens | B3KT51 | TM2D3 | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| H. sapiens | E9PI46 | ABCD4 | Unknown |
| H. sapiens | B7Z863 | SLMAP | Unknown |
| H. sapiens | F8WEN7 | MTFP1 | Unknown |
| H. sapiens | A0A0C4DFN5 | TCTN3 | Unknown |
| H. sapiens | M0QZX7 | ZNF816 | Unknown |
| H. sapiens | Q8N329 | EOGT | Unknown |
| H. sapiens | F2Z2A2 | MFSD9 | Unknown |
| H. sapiens | A0A1W2PQE2 | HLA-B | Unknown |
| H. sapiens | E5RG25 | UBE2W | Unknown |
| H. sapiens | A0A0J9YWY1 | LLCFC1 | Unknown |
| H. sapiens | M0R0R3 | SMIM7 | Unknown |
| H. sapiens | A0AVG3 | TSNARE1 | Unknown |
| H. sapiens | B1ANB7 | MCOLN3 | Unknown |
| H. sapiens | I3L288 | TMEM159 | Unknown |
| H. sapiens | B7Z964 | SLMAP | Unknown |
| H. sapiens | X6R3D1 | HRASLS | Unknown |
| H. sapiens | E7EM61 | SLC19A3 | Unknown |
| H. sapiens | Q3KQS6 | MME | Unknown |
| H. sapiens | B9TX75 | MED24 | Unknown |
| H. sapiens | G5E972 | TMPO | Unknown |
| H. sapiens | F8VV56 | CD63 | Unknown |
| H. sapiens | D6RCL9 | SERF1B | Unknown |
| H. sapiens | B3KT28 | FAF1 | Unknown |
| H. sapiens | G3V1R8 | TMBIM4 | Unknown |
| H. sapiens | G5E9Q6 | PFN2 | Unknown |
| H. sapiens | A0A0C4DGX8 | ATP6AP1 | Unknown |
| H. sapiens | K7EMW4 | NCLN | Unknown |
| H. sapiens | B4DKD2 | ADAM11 | Unknown |
| H. sapiens | E5RGC5 | TVP23C-CDRT4 | Unknown |
| S. cerevisiae | Q03941 | CAB5 | Both |
| S. cerevisiae | Q08215 | PEX15 | Both |
| S. cerevisiae | P25580 | PBN1 | ER |
| S. cerevisiae | P32854 | PEP12 | ER |
| S. cerevisiae | Q05637 | PHM6 | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| S. cerevisiae | Q08931 | PRM3 | ER |
| S. cerevisiae | P39926 | SSO2 | ER |
| S. cerevisiae | P31377 | SYN8 | ER |
| S. cerevisiae | P32867 | SSO1 | ER |
| S. cerevisiae | P31109 | SNC1 | ER |
| S. cerevisiae | P33328 | SNC2 | ER |
| S. cerevisiae | P38247 | SLM4 | ER |
| S. cerevisiae | P43682 | SFT1 | ER |
| S. cerevisiae | Q03322 | TLG1 | ER |
| S. cerevisiae | P52870 | SBH1 | ER |
| S. cerevisiae | Q6Q595 | SCS22 | ER |
| S. cerevisiae | P35179 | SSS1 | ER |
| S. cerevisiae | P40075 | SCS2 | ER |
| S. cerevisiae | P22214 | SEC22 | ER |
| S. cerevisiae | P52871 | SBH2 | ER |
| S. cerevisiae | Q01590 | SED5 | ER |
| S. cerevisiae | P38342 | TSC10 | ER |
| S. cerevisiae | Q12255 | NYV1 | ER |
| S. cerevisiae | P14020 | DPM1 | ER |
| S. cerevisiae | Q08955 | CSM4 | ER |
| S. cerevisiae | Q06001 | FAR10 | ER |
| S. cerevisiae | P22804 | BET1 | ER |
| S. cerevisiae | P25385 | BOS1 | ER |
| S. cerevisiae | P40312 | CYB5 | ER |
| S. cerevisiae | P38736 | GOS1 | ER |
| S. cerevisiae | P32363 | SPT14 | ER |
| S. cerevisiae | P43560 | LAM5 | ER |
| S. cerevisiae | P48353 | HLJ1 | ER |
| S. cerevisiae | Q99332 | FRT1 | ER |
| S. cerevisiae | P32339 | HMX1 | ER |
| S. cerevisiae | Q3E790 | TSC3 | ER |
| S. cerevisiae | Q04338 | VTI1 | ER |
| S. cerevisiae | Q3E842 | YMR122W-A | ER |
| S. cerevisiae | P38216 | YBR016W | ER |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| S. cerevisiae | P53146 | USE1 | ER |
| S. cerevisiae | P38374 | YSY6 | ER |
| S. cerevisiae | Q05899 | YLR297W | ER |
| S. cerevisiae | Q03944 | VPS64 | ER |
| S. cerevisiae | P33296 | UBC6 | ER |
| S. cerevisiae | P41834 | UFE1 | ER |
| S. cerevisiae | Q08959 | PGC1 | Mit |
| S. cerevisiae | P80967 | TOM5 | Mit |
| S. cerevisiae | P53507 | TOM7 | Mit |
| S. cerevisiae | P33448 | TOM6 | Mit |
| S. cerevisiae | P40515 | FIS1 | Mit |
| S. cerevisiae | P39722 | GEM1 | Mit |
| S. cerevisiae | P22289 | QCR9 | MitIn |
| S. cerevisiae | P07255 | COX9 | MitIn |
| S. cerevisiae | P10174 | COX7 | MitIn |
| S. cerevisiae | Q2V2P9 | YDR119W-A | MitIn |
| S. cerevisiae | Q02969 | PEX25 | Pex |
| S. cerevisiae | P38335 | MTC4 | Pex |
| S. cerevisiae | Q02820 | NCE1 | Unknown |
| S. cerevisiae | Q03441 | RMD1 | Unknown |
| S. cerevisiae | P43620 | RMD8 | Unknown |
| S. cerevisiae | Q08559 | FYV12 | Unknown |
| S. cerevisiae | P11927 | KAR1 | Unknown |
| S. cerevisiae | Q08630 | IRC13 | Unknown |
| S. cerevisiae | P0CD97 | YER039C-A | Unknown |
| S. cerevisiae | Q3E828 | YJL127C-B | Unknown |
| S. cerevisiae | Q8TGS8 | YMR105W-A | Unknown |
| S. cerevisiae | Q3E760 | YMR030W-A | Unknown |
| S. cerevisiae | O13511 | YAL065C | Unknown |
| S. cerevisiae | P39563 | YAR064W | Unknown |
| S. cerevisiae | Q2V2Q3 | YBR201C-A | Unknown |
| S. cerevisiae | Q3E743 | YJR112W-A | Unknown |
| S. cerevisiae | P47080 | YJL007C | Unknown |
| S. cerevisiae | P36092 | YKL044W | Unknown |

Continuation of Table S2.1

| Organism | Entry | Gene names | Loc |
|---|---|---|---|
| S. cerevisiae | Q2V2P2 | YKL065W-A | Unknown |
| S. cerevisiae | Q07738 | YDL241W | Unknown |
| S. cerevisiae | Q05612 | YDR278C | Unknown |
| S. cerevisiae | Q03480 | YDR209C | Unknown |
| S. cerevisiae | Q3E750 | YGL041C-B | Unknown |
| S. cerevisiae | Q8TGK1 | YHR213W-B | Unknown |
| S. cerevisiae | Q07074 | YHR007C-A | Unknown |
| S. cerevisiae | A5Z2X5 | YPR010C-A | Unknown |
| S. cerevisiae | P53229 | YGR045C | Unknown |
| S. cerevisiae | Q2V2P3 | YKL023C-A | Unknown |
| S. cerevisiae | Q3E814 | YLL006W-A | Unknown |
| S. cerevisiae | Q12506 | YOR314W | Unknown |
| S. cerevisiae | Q8TGU7 | YBR126W-A | Unknown |
| S. cerevisiae | Q04597 | YDR114C | Unknown |
| S. cerevisiae | P0C268 | YBL039W-B | Unknown |
| S. cerevisiae | Q96VH3 | YCL021W-A | Unknown |
| S. cerevisiae | Q8TGT9 | YGR146C-A | Unknown |
| S. cerevisiae | Q08110 | YOL014W | Unknown |
| S. cerevisiae | Q08734 | YOR268C | Unknown |
| S. cerevisiae | P53156 | YGL081W | Unknown |
| S. cerevisiae | Q05898 | YLR296W | Unknown |
| S. cerevisiae | P38185 | YBL071C | Unknown |

Table S2.1: **Putative TA proteins in yeast and humans.** A combined list of all identified TA proteins in both the human and yeast genomes with their known localization marked as ER (which includes ER, Golgi apparatus, nucleus, cell membrane, vacuole, endosomes, and lysosomes), mitochondria, both (ER and mitochondria), peroxisome, and unknown.

Table S2.2

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Patch 15 | S. cerevisiae | Kyte & Doolittle | 96.28 | 5 | 48 |
| Patch 15 | S. cerevisiae | TM Tendency | 95.81 | 7 | 46 |
| Patch 11 | S. cerevisiae | Kyte & Doolittle | 94.88 | 8 | 45 |
| Wheel Face 5 | S. cerevisiae | TM Tendency | 94.65 | 6 | 47 |
| Wheel Face 9 | S. cerevisiae | TM Tendency | 93.95 | 6 | 47 |
| Patch 15 | S. cerevisiae | GES | 93.95 | 11 | 42 |
| Patch 19 | S. cerevisiae | Kyte & Doolittle | 93.49 | 9 | 44 |
| Wheel Face 7 | S. cerevisiae | TM Tendency | 93.02 | 9 | 44 |
| Wheel Face 8 | S. cerevisiae | TM Tendency | 92.91 | 7 | 46 |
| Segment 15 | S. cerevisiae | Kyte & Doolittle | 92.56 | 12 | 41 |
| Segment 15 | S. cerevisiae | TM Tendency | 92.33 | 10 | 43 |
| Patch 19 | S. cerevisiae | TM Tendency | 91.86 | 8 | 45 |
| Segment 15 | S. cerevisiae | GES | 91.63 | 8 | 45 |
| Segment 11 | S. cerevisiae | Kyte & Doolittle | 91.63 | 10 | 43 |
| Rectangle 9 | S. cerevisiae | Kyte & Doolittle | 91.28 | 12 | 41 |
| Wheel Face 5 | S. cerevisiae | Kyte & Doolittle | 91.05 | 9 | 44 |
| Wheel Face 7 | S. cerevisiae | Kyte & Doolittle | 91.05 | 7 | 46 |
| Rectangle 9 | S. cerevisiae | Fauchere Pliska | 90.93 | 8 | 45 |
| Patch 19 | S. cerevisiae | Octanol | 90.93 | 9 | 44 |
| Rectangle 9 | S. cerevisiae | TM Tendency | 90.93 | 10 | 43 |
| Patch 19 | S. cerevisiae | GES | 90.70 | 10 | 43 |
| Line 13 | S. cerevisiae | Kyte & Doolittle | 90.70 | 8 | 45 |
| Patch 11 | S. cerevisiae | GES | 90.35 | 12 | 41 |
| Star 8 | S. cerevisiae | Kyte & Doolittle | 90.12 | 11 | 42 |
| TMD (18 aa) | S. cerevisiae | Kyte & Doolittle | 90.00 | 16 | 37 |
| TMD | S. cerevisiae | Kyte & Doolittle | 90.00 | 16 | 37 |
| Wheel Face 9 | S. cerevisiae | GES | 89.88 | 11 | 42 |
| Patch 19 | S. cerevisiae | Fauchere Pliska | 89.77 | 8 | 45 |
| Segment 19 | S. cerevisiae | Kyte & Doolittle | 89.77 | 9 | 44 |
| Patch 11 | S. cerevisiae | TM Tendency | 89.77 | 12 | 41 |
| Star 8 | S. cerevisiae | TM Tendency | 89.77 | 9 | 44 |
| Wheel Face 8 | S. cerevisiae | Kyte & Doolittle | 89.53 | 15 | 38 |
| Segment 19 | S. cerevisiae | TM Tendency | 89.53 | 11 | 42 |

Continuation of Table S2.2

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Rectangle 9 | S. cerevisiae | GES | 89.30 | 10 | 43 |
| Wheel Face 8 | S. cerevisiae | GES | 89.19 | 12 | 41 |
| Line 13 | S. cerevisiae | GES | 89.19 | 9 | 44 |
| Wheel Face 6 | S. cerevisiae | TM Tendency | 89.07 | 11 | 42 |
| Segment 19 | S. cerevisiae | Fauchere Pliska | 89.07 | 8 | 45 |
| Segment 11 | S. cerevisiae | TM Tendency | 89.07 | 10 | 43 |
| TMD (18 aa) | S. cerevisiae | TM Tendency | 88.84 | 9 | 44 |
| Line 13 | S. cerevisiae | TM Tendency | 88.72 | 10 | 43 |
| Wheel Face 7 | S. cerevisiae | GES | 88.37 | 16 | 37 |
| Wheel Face 9 | S. cerevisiae | Roseman | 88.37 | 8 | 45 |
| Rectangle 9 | S. cerevisiae | Roseman | 88.37 | 14 | 39 |
| TMD average | S. cerevisiae | Kyte & Doolittle | 88.37 | 9 | 44 |
| Star 8 | S. cerevisiae | Roseman | 88.26 | 9 | 44 |
| Wheel Face 9 | S. cerevisiae | Kyte & Doolittle | 88.14 | 10 | 43 |
| Patch 15 | S. cerevisiae | Fauchere Pliska | 88.14 | 12 | 41 |
| Patch 19 | S. cerevisiae | Roseman | 88.14 | 12 | 41 |
| Patch 15 | S. cerevisiae | Octanol | 88.14 | 11 | 42 |
| Wheel Face 7 | S. cerevisiae | Roseman | 88.02 | 9 | 44 |
| Wheel Face 8 | S. cerevisiae | Roseman | 87.91 | 13 | 40 |
| Star 8 | S. cerevisiae | Fauchere Pliska | 87.91 | 18 | 35 |
| TMD | S. cerevisiae | TM Tendency | 87.67 | 10 | 43 |
| Segment 11 | S. cerevisiae | GES | 87.44 | 13 | 40 |
| Segment 15 | S. cerevisiae | Fauchere Pliska | 87.44 | 12 | 41 |
| Star 8 | S. cerevisiae | GES | 87.21 | 9 | 44 |
| Wheel Face 5 | S. cerevisiae | GES | 87.09 | 8 | 45 |
| Wheel Face 4 | S. cerevisiae | Kyte & Doolittle | 87.09 | 8 | 45 |
| Rectangle 9 | S. cerevisiae | Octanol | 86.63 | 14 | 39 |
| Wheel Face 9 | S. cerevisiae | Fauchere Pliska | 86.51 | 8 | 45 |
| Wheel Face 9 | S. cerevisiae | Octanol | 86.28 | 9 | 44 |
| Segment 19 | S. cerevisiae | Octanol | 86.28 | 11 | 42 |
| TMD average | S. cerevisiae | TM Tendency | 86.28 | 12 | 41 |
| Wheel Face 6 | S. cerevisiae | Kyte & Doolittle | 85.93 | 11 | 42 |
| TMD (18 aa) | S. cerevisiae | GES | 85.93 | 10 | 43 |

Continuation of Table S2.2

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Patch 11 | S. cerevisiae | Fauchere Pliska | 85.93 | 16 | 37 |
| Segment 11 | S. cerevisiae | Fauchere Pliska | 85.81 | 13 | 40 |
| Star 8 | S. cerevisiae | Octanol | 85.81 | 10 | 43 |
| TMD | S. cerevisiae | GES | 85.81 | 10 | 43 |
| Twist 8 | S. cerevisiae | Kyte & Doolittle | 85.70 | 10 | 43 |
| Patch 15 | S. cerevisiae | Roseman | 85.35 | 13 | 40 |
| Segment 15 | S. cerevisiae | Octanol | 85.35 | 12 | 41 |
| Wheel Face 8 | S. cerevisiae | Octanol | 85.23 | 13 | 40 |
| Wheel Face 6 | S. cerevisiae | GES | 84.88 | 16 | 37 |
| Segment 19 | S. cerevisiae | GES | 84.65 | 16 | 37 |
| Twist 8 | S. cerevisiae | Fauchere Pliska | 84.65 | 12 | 41 |
| Patch 11 | S. cerevisiae | Roseman | 84.65 | 15 | 38 |
| Line 9 | S. cerevisiae | Kyte & Doolittle | 84.30 | 9 | 44 |
| Twist 8 | S. cerevisiae | TM Tendency | 84.19 | 14 | 39 |
| Wheel Face 8 | S. cerevisiae | Fauchere Pliska | 84.07 | 10 | 43 |
| Twist 8 | S. cerevisiae | GES | 83.95 | 10 | 43 |
| Segment 15 | S. cerevisiae | Roseman | 83.95 | 16 | 37 |
| TMD average | S. cerevisiae | Fauchere Pliska | 83.95 | 10 | 43 |
| TMD | S. cerevisiae | Fauchere Pliska | 83.95 | 15 | 38 |
| Line 13 | S. cerevisiae | Roseman | 83.72 | 11 | 42 |
| TMD average | S. cerevisiae | GES | 83.72 | 11 | 42 |
| TMD (18 aa) | S. cerevisiae | Fauchere Pliska | 83.49 | 16 | 37 |
| Wheel Face 7 | S. cerevisiae | Fauchere Pliska | 83.26 | 10 | 43 |
| Segment 11 | S. cerevisiae | Roseman | 83.02 | 12 | 41 |
| TMD (18 aa) | S. cerevisiae | Octanol | 83.02 | 12 | 41 |
| TMD | S. cerevisiae | Octanol | 83.02 | 12 | 41 |
| Wheel Face 7 | S. cerevisiae | Octanol | 82.79 | 17 | 36 |
| Line 17 | S. cerevisiae | Kyte & Doolittle | 82.79 | 14 | 39 |
| Patch 11 | S. cerevisiae | Octanol | 82.56 | 12 | 41 |
| TMD average | S. cerevisiae | Octanol | 82.56 | 12 | 41 |
| Segment 19 | S. cerevisiae | Roseman | 82.21 | 15 | 38 |
| Wheel Face 5 | S. cerevisiae | Fauchere Pliska | 81.40 | 9 | 44 |
| Wheel Face 6 | S. cerevisiae | Roseman | 81.28 | 15 | 38 |

Continuation of Table S2.2

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Wheel Face 5 | S. cerevisiae | Octanol | 80.93 | 17 | 36 |
| Wheel Face 5 | S. cerevisiae | Roseman | 80.70 | 13 | 40 |
| Twist 8 | S. cerevisiae | Roseman | 80.70 | 14 | 39 |
| Segment 11 | S. cerevisiae | Octanol | 80.70 | 11 | 42 |
| Line 17 | S. cerevisiae | TM Tendency | 80.70 | 14 | 39 |
| Wheel Face 6 | S. cerevisiae | Fauchere Pliska | 80.47 | 8 | 45 |
| Wheel Face 4 | S. cerevisiae | Octanol | 80.35 | 12 | 41 |
| TMD average | S. cerevisiae | Roseman | 79.88 | 16 | 37 |
| TMD | S. cerevisiae | Roseman | 79.88 | 18 | 35 |
| TMD (18 aa) | S. cerevisiae | Roseman | 79.65 | 18 | 35 |
| Wheel Face 6 | S. cerevisiae | Octanol | 79.42 | 13 | 40 |
| Wheel Face 4 | S. cerevisiae | Fauchere Pliska | 79.30 | 9 | 44 |
| Line 17 | S. cerevisiae | GES | 79.19 | 15 | 38 |
| Line 9 | S. cerevisiae | GES | 78.60 | 12 | 41 |
| Twist 8 | S. cerevisiae | Octanol | 76.86 | 16 | 37 |
| Wheel Face 3 | S. cerevisiae | Kyte & Doolittle | 76.63 | 14 | 39 |
| Line 13 | S. cerevisiae | Fauchere Pliska | 75.81 | 10 | 43 |
| Line 13 | S. cerevisiae | Octanol | 74.88 | 15 | 38 |
| Wheel Face 4 | S. cerevisiae | GES | 74.30 | 10 | 43 |
| Wheel Face 4 | S. cerevisiae | TM Tendency | 74.07 | 6 | 47 |
| Line 9 | S. cerevisiae | TM Tendency | 74.07 | 9 | 44 |
| Line 17 | S. cerevisiae | Fauchere Pliska | 73.60 | 21 | 32 |
| Wheel Face 4 | S. cerevisiae | Roseman | 72.44 | 11 | 42 |
| Wheel Face 3 | S. cerevisiae | Fauchere Pliska | 70.58 | 14 | 39 |
| Line 17 | S. cerevisiae | Octanol | 70.35 | 12 | 41 |
| Line 17 | S. cerevisiae | Roseman | 68.37 | 19 | 34 |
| Wheel Face 3 | S. cerevisiae | TM Tendency | 66.98 | 15 | 38 |
| Wheel Face 3 | S. cerevisiae | GES | 66.05 | 12 | 41 |
| Wheel Face 3 | S. cerevisiae | Roseman | 64.77 | 19 | 34 |
| Line 9 | S. cerevisiae | Roseman | 64.19 | 25 | 28 |
| Wheel Face 3 | S. cerevisiae | Octanol | 63.60 | 24 | 29 |
| Line 9 | S. cerevisiae | Fauchere Pliska | 55.23 | 21 | 32 |
| TMD average | S. cerevisiae | Roseman | 53.02 | 33 | 20 |

Continuation of Table S2.2

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| TMD average | S. cerevisiae | TM Tendency | 51.16 | 18 | 35 |
| TMD average | S. cerevisiae | GES | -51.16 | 19 | 34 |
| TMD length | S. cerevisiae | | -51.28 | 39 | 14 |
| Line 9 | S. cerevisiae | Octanol | -51.40 | 28 | 25 |
| TMD average | S. cerevisiae | Kyte & Doolittle | -53.26 | 19 | 34 |
| TMD average | S. cerevisiae | Fauchere Pliska | -57.21 | 24 | 29 |
| TMD average | S. cerevisiae | Octanol | -61.63 | 13 | 40 |

Table S2.2: **Hydrophobicity geometry metrics perform the best when classifying yeast TA proteins.** A list of all the metrics tested against the yeast genome ranked from highest to lowest AUROC score.

| Fig. 2.5 ref # | Entry | TA name | ER | Mito | total # of cells | ER (%) | Mito (%) | Localization |
|---|---|---|---|---|---|---|---|---|
| 1 | P40515 | Fis1 | 0 | 33 | 33 | 0% | 100% | Mitochondria |
| 2 | Q2V2P9 | Cox26 | 2 | 18 | 20 | 10% | 90% | Mitochondria |
| 3 | Q2V2P3 | YKL023C | 131 | 3 | 134 | 98% | 2% | ER |
| 4 | P38185 | YBL071C | 122 | 3 | 125 | 98% | 2% | ER |
| 5 | Q05612 | YDR278C | 221 | 0 | 221 | 100% | 0% | ER |
| 6 | P0CD97 | YER039C | 161 | 0 | 161 | 100% | 0% | ER |
| 7 | P53156 | YGL081W | 135 | 0 | 135 | 100% | 0% | ER |
| 8 | Q12506 | YOR314W | 176 | 2 | 178 | 99% | 1% | ER |
| 9 | P53229 | YGR045C | 169 | 2 | 171 | 99% | 1% | ER |
| 10 | P36092 | YKL044W | 89 | 3 | 92 | 97% | 3% | ER |
| 11 | Q04597 | YDR114C | 1 | 141 | 142 | 1% | 99% | Mitochondria |
| 12 | Q03480 | YDR209C | 3 | 67 | 70 | 4% | 96% | Mitochondria |
| 13 | Q3E743 | YJR112W | 39 | 2 | 41 | 95% | 5% | ER |
| 14 | Q3E750 | YGL041C | 69 | 1 | 70 | 99% | 1% | ER |
| 15 | Q08110 | YOL014W | 91 | 7 | 98 | 93% | 7% | ER |
| 16 | Q05898 | YLR296W | 0 | 64 | 64 | 0% | 100% | Mitochondria |
| 17 | P0C268 | YBL039W | 38 | 8 | 46 | 83% | 17% | Other |

Table S2.3: **Determined localization of unknown TA proteins in yeast cells.** A list of the experimentally determined localization of 2 known (controls) and 15 unknown TA proteins. Localization is split between mitochondria and ER on a per cell basis and TA proteins are referenced based on the number in Figure 2.5.

| Organism | Entry | Name | Localization |
|---|---|---|---|
| S. cerevisiae | Q08110 | YOL014W | ER |
| S. cerevisiae | P11927 | KAR1 | ER |
| S. cerevisiae | Q07738 | YDL241W | ER |
| S. cerevisiae | Q3E750 | YGL041C-B | ER |
| S. cerevisiae | Q8TGU7 | YBR126W-A | ER |
| S. cerevisiae | P43620 | RMD8 | ER |
| S. cerevisiae | Q07074 | YHR007C-A | ER |
| S. cerevisiae | Q08734 | YOR268C | ER |
| S. cerevisiae | Q3E743 | YJR112W-A | ER |
| S. cerevisiae | Q3E828 | YJL127C-B | Mit |
| S. cerevisiae | P36092 | YKL044W | Mit |
| S. cerevisiae | Q2V2P2 | YKL065W-A | Pex |
| S. cerevisiae | P0C268 | YBL039W-B | ER |
| S. cerevisiae | Q96VH3 | YCL021W-A | ER |
| S. cerevisiae | Q03441 | RMD1 | ER |

Table S2.4: **Localization of unknown TA proteins identified in Weill et al., 2018.** Localization of unknown TA proteins identified in Weill et al., 2018. The reported localization for 17 TA proteins identified in the high-throughput screen performed in Weill et al., 2018.

Table S2.5

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Wheel Face 7 | S. cerevisiae | TM Tendency | 0.89 | 20 | 54 |
| Wheel Face 5 | S. cerevisiae | TM Tendency | 0.88 | 21 | 58 |
| Patch 19 | S. cerevisiae | Fauchere Pliska | 0.87 | 19 | 60 |
| Segment 11 | S. cerevisiae | Kyte & Doolittle | 0.87 | 18 | 56 |
| Wheel Face 9 | S. cerevisiae | TM Tendency | 0.87 | 22 | 56 |
| Wheel Face 6 | S. cerevisiae | TM Tendency | 0.86 | 19 | 52 |
| Segment 15 | S. cerevisiae | Kyte & Doolittle | 0.86 | 18 | 63 |
| Wheel Face 8 | S. cerevisiae | TM Tendency | 0.86 | 29 | 54 |
| Patch 19 | S. cerevisiae | Octanol | 0.86 | 22 | 56 |
| Wheel Face 5 | S. cerevisiae | Kyte & Doolittle | 0.86 | 20 | 59 |
| Wheel Face 7 | S. cerevisiae | Kyte & Doolittle | 0.86 | 18 | 62 |
| Patch 15 | S. cerevisiae | Kyte & Doolittle | 0.86 | 15 | 57 |
| Wheel Face 9 | S. cerevisiae | Fauchere Pliska | 0.86 | 25 | 58 |
| Patch 15 | S. cerevisiae | Fauchere Pliska | 0.85 | 16 | 59 |
| Patch 15 | S. cerevisiae | TM Tendency | 0.85 | 19 | 59 |
| Rectangle 9 | S. cerevisiae | Fauchere Pliska | 0.85 | 20 | 59 |
| Patch 11 | S. cerevisiae | Kyte & Doolittle | 0.85 | 16 | 59 |
| Segment 15 | S. cerevisiae | TM Tendency | 0.84 | 19 | 54 |
| Patch 15 | S. cerevisiae | Octanol | 0.84 | 24 | 58 |
| Segment 19 | S. cerevisiae | Fauchere Pliska | 0.84 | 16 | 62 |
| Star 8 | S. cerevisiae | Kyte & Doolittle | 0.84 | 28 | 60 |
| Wheel Face 5 | S. cerevisiae | GES | 0.84 | 32 | 58 |
| Wheel Face 6 | S. cerevisiae | Kyte & Doolittle | 0.84 | 18 | 59 |
| Rectangle 9 | S. cerevisiae | TM Tendency | 0.84 | 20 | 56 |
| Star 8 | S. cerevisiae | Fauchere Pliska | 0.84 | 26 | 44 |
| Wheel Face 7 | S. cerevisiae | Fauchere Pliska | 0.83 | 19 | 60 |
| Wheel Face 7 | S. cerevisiae | GES | 0.83 | 21 | 64 |
| Segment 11 | S. cerevisiae | Fauchere Pliska | 0.83 | 29 | 53 |
| Rectangle 9 | S. cerevisiae | Kyte & Doolittle | 0.83 | 19 | 50 |
| Segment 15 | S. cerevisiae | Fauchere Pliska | 0.83 | 18 | 60 |
| Wheel Face 5 | S. cerevisiae | Fauchere Pliska | 0.83 | 16 | 59 |
| Rectangle 9 | S. cerevisiae | Octanol | 0.83 | 29 | 57 |
| Line 13 | S. cerevisiae | GES | 0.83 | 21 | 48 |

Continuation of Table S2.5

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Patch 15 | S. cerevisiae | GES | 0.83 | 14 | 48 |
| Rectangle 9 | S. cerevisiae | Roseman | 0.83 | 17 | 54 |
| Wheel Face 7 | S. cerevisiae | Octanol | 0.83 | 24 | 51 |
| Wheel Face 9 | S. cerevisiae | Octanol | 0.83 | 23 | 48 |
| Wheel Face 8 | S. cerevisiae | Fauchere Pliska | 0.82 | 18 | 59 |
| Patch 19 | S. cerevisiae | Roseman | 0.82 | 26 | 56 |
| Wheel Face 7 | S. cerevisiae | Roseman | 0.82 | 28 | 51 |
| Segment 11 | S. cerevisiae | TM Tendency | 0.82 | 24 | 50 |
| Segment 11 | S. cerevisiae | GES | 0.82 | 21 | 50 |
| Wheel Face 9 | S. cerevisiae | Roseman | 0.82 | 28 | 62 |
| Wheel Face 6 | S. cerevisiae | Fauchere Pliska | 0.82 | 33 | 62 |
| Segment 15 | S. cerevisiae | GES | 0.82 | 25 | 57 |
| Line 13 | S. cerevisiae | TM Tendency | 0.82 | 19 | 55 |
| Wheel Face 4 | S. cerevisiae | Fauchere Pliska | 0.82 | 18 | 56 |
| Patch 11 | S. cerevisiae | Fauchere Pliska | 0.82 | 17 | 46 |
| TMD average | S. cerevisiae | Fauchere Pliska | 0.82 | 25 | 58 |
| Wheel Face 8 | S. cerevisiae | Roseman | 0.82 | 31 | 52 |
| Wheel Face 8 | S. cerevisiae | Octanol | 0.82 | 22 | 56 |
| Line 13 | S. cerevisiae | Kyte & Doolittle | 0.81 | 40 | 52 |
| Patch 11 | S. cerevisiae | TM Tendency | 0.81 | 23 | 53 |
| Rectangle 9 | S. cerevisiae | GES | 0.81 | 24 | 54 |
| Wheel Face 9 | S. cerevisiae | GES | 0.81 | 23 | 51 |
| Star 8 | S. cerevisiae | TM Tendency | 0.81 | 21 | 59 |
| Wheel Face 6 | S. cerevisiae | GES | 0.81 | 19 | 59 |
| Wheel Face 8 | S. cerevisiae | Kyte & Doolittle | 0.81 | 28 | 65 |
| Star 8 | S. cerevisiae | GES | 0.81 | 23 | 58 |
| Wheel Face 9 | S. cerevisiae | Kyte & Doolittle | 0.81 | 28 | 60 |
| Wheel Face 8 | S. cerevisiae | GES | 0.81 | 25 | 48 |
| Segment 19 | S. cerevisiae | Kyte & Doolittle | 0.81 | 23 | 57 |
| Twist8 | S. cerevisiae | Kyte & Doolittle | 0.81 | 23 | 62 |
| Twist8 | S. cerevisiae | Fauchere Pliska | 0.81 | 17 | 61 |
| Patch 11 | S. cerevisiae | GES | 0.81 | 19 | 54 |
| Patch 19 | S. cerevisiae | TM Tendency | 0.81 | 23 | 58 |

Continuation of Table S2.5

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Segment 19 | S. cerevisiae | TM Tendency | 0.80 | 20 | 56 |
| Segment 19 | S. cerevisiae | Octanol | 0.80 | 19 | 57 |
| Star 8 | S. cerevisiae | Roseman | 0.80 | 28 | 64 |
| Star 8 | S. cerevisiae | Octanol | 0.80 | 23 | 66 |
| TMD | S. cerevisiae | Fauchere Pliska | 0.80 | 29 | 48 |
| Wheel Face 4 | S. cerevisiae | Octanol | 0.80 | 22 | 60 |
| TMD (18aa) | S. cerevisiae | Fauchere Pliska | 0.80 | 20 | 46 |
| Line 13 | S. cerevisiae | Roseman | 0.80 | 26 | 57 |
| Segment 15 | S. cerevisiae | Octanol | 0.80 | 19 | 56 |
| Wheel Face 6 | S. cerevisiae | Octanol | 0.79 | 14 | 54 |
| Patch 19 | S. cerevisiae | Kyte & Doolittle | 0.79 | 15 | 51 |
| Patch 15 | S. cerevisiae | Roseman | 0.79 | 24 | 54 |
| Wheel Face 4 | S. cerevisiae | Kyte & Doolittle | 0.79 | 15 | 39 |
| Patch 11 | S. cerevisiae | Octanol | 0.79 | 22 | 63 |
| Wheel Face 6 | S. cerevisiae | Roseman | 0.79 | 16 | 48 |
| Segment 11 | S. cerevisiae | Octanol | 0.79 | 16 | 40 |
| Patch 19 | S. cerevisiae | GES | 0.78 | 27 | 47 |
| TMD | S. cerevisiae | Kyte & Doolittle | 0.78 | 19 | 45 |
| TMD (18aa) | S. cerevisiae | Kyte & Doolittle | 0.78 | 24 | 50 |
| Patch 11 | S. cerevisiae | Roseman | 0.78 | 26 | 49 |
| TMD average | S. cerevisiae | Kyte & Doolittle | 0.78 | 33 | 51 |
| Wheel Face 5 | S. cerevisiae | Octanol | 0.78 | 21 | 50 |
| TMD average | S. cerevisiae | Octanol | 0.78 | 18 | 51 |
| Segment 15 | S. cerevisiae | Roseman | 0.77 | 24 | 47 |
| Line 9 | S. cerevisiae | GES | 0.77 | 21 | 61 |
| Wheel Face 5 | S. cerevisiae | Roseman | 0.77 | 31 | 43 |
| Line 17 | S. cerevisiae | Fauchere Pliska | 0.77 | 20 | 52 |
| TMD | S. cerevisiae | Octanol | 0.77 | 42 | 51 |
| TMD (18aa) | S. cerevisiae | Octanol | 0.77 | 21 | 50 |
| Segment 11 | S. cerevisiae | Roseman | 0.77 | 28 | 57 |
| TMD (18aa) | S. cerevisiae | TM Tendency | 0.77 | 20 | 55 |
| Line 17 | S. cerevisiae | TM Tendency | 0.76 | 13 | 49 |
| Segment 19 | S. cerevisiae | GES | 0.76 | 30 | 45 |

Continuation of Table S2.5

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Line 17 | S. cerevisiae | Kyte & Doolittle | 0.76 | 36 | 44 |
| TMD | S. cerevisiae | TM Tendency | 0.76 | 24 | 53 |
| TMD average | S. cerevisiae | TM Tendency | 0.76 | 25 | 46 |
| Wheel Face 4 | S. cerevisiae | TM Tendency | 0.76 | 22 | 69 |
| Line 17 | S. cerevisiae | GES | 0.76 | 15 | 42 |
| Line 9 | S. cerevisiae | Kyte & Doolittle | 0.76 | 18 | 62 |
| Line 13 | S. cerevisiae | Fauchere Pliska | 0.75 | 21 | 58 |
| Segment 19 | S. cerevisiae | Roseman | 0.75 | 21 | 13 |
| Line 9 | S. cerevisiae | TM Tendency | 0.75 | 19 | 55 |
| Twist8 | S. cerevisiae | TM Tendency | 0.75 | 21 | 61 |
| Line 13 | S. cerevisiae | Octanol | 0.75 | 35 | 55 |
| Twist8 | S. cerevisiae | GES | 0.74 | 23 | 59 |
| Wheel Face 4 | S. cerevisiae | Roseman | 0.74 | 29 | 50 |
| Wheel Face 3 | S. cerevisiae | Fauchere Pliska | 0.74 | 25 | 59 |
| TMD average | S. cerevisiae | Roseman | 0.74 | 18 | 39 |
| Wheel Face 4 | S. cerevisiae | GES | 0.74 | 20 | 66 |
| TMD | S. cerevisiae | Roseman | 0.74 | 35 | 49 |
| TMD (18aa) | S. cerevisiae | Roseman | 0.74 | 19 | 48 |
| Line 17 | S. cerevisiae | Octanol | 0.73 | 18 | 61 |
| TMD (18aa) | S. cerevisiae | GES | 0.73 | 23 | 56 |
| TMD | S. cerevisiae | GES | 0.73 | 18 | 52 |
| TMD average | S. cerevisiae | GES | 0.73 | 31 | 54 |
| Wheel Face 3 | S. cerevisiae | TM Tendency | 0.72 | 27 | 62 |
| Wheel Face 3 | S. cerevisiae | Kyte & Doolittle | 0.71 | 20 | 39 |
| Twist8 | S. cerevisiae | Octanol | 0.70 | 21 | 55 |
| Line 9 | S. cerevisiae | Roseman | 0.70 | 33 | 43 |
| Wheel Face 3 | S. cerevisiae | Roseman | 0.70 | 22 | 51 |
| Line 17 | S. cerevisiae | Roseman | 0.70 | 28 | 64 |
| Twist8 | S. cerevisiae | Roseman | 0.68 | 21 | 51 |
| Wheel Face 3 | S. cerevisiae | GES | 0.67 | 26 | 62 |
| Wheel Face 3 | S. cerevisiae | Octanol | 0.67 | 33 | 47 |
| Line 9 | S. cerevisiae | Fauchere Pliska | 0.64 | 19 | 49 |
| Line 9 | S. cerevisiae | Octanol | 0.59 | 28 | 36 |

Continuation of Table S2.5

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| TMD len | S. cerevisiae | | -0.53 | 57 | 20 |

Table S2.5: **Metrics using a helical wheel geometry are the best predictors for localization of unknown TA proteins.** A list of the metrics used ranked by performance over the entire yeast dataset (old and new localizations included) with number of correctly predicted TA proteins listed.

Table S2.6

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Patch 11 | H. sapiens | Kyte & Doolittle | 82.63 | 73 | 136 |
| Segment 19 | H. sapiens | Kyte & Doolittle | 82.61 | 58 | 151 |
| Segment 11 | H. sapiens | Kyte & Doolittle | 82.06 | 59 | 150 |
| Patch 11 | H. sapiens | TM Tendency | 81.74 | 61 | 148 |
| Patch 15 | H. sapiens | TM Tendency | 81.51 | 59 | 150 |
| Patch 15 | H. sapiens | Kyte & Doolittle | 81.23 | 74 | 135 |
| Segment 11 | H. sapiens | TM Tendency | 80.75 | 66 | 143 |
| TMD | H. sapiens | Kyte & Doolittle | 80.72 | 37 | 172 |
| Patch 19 | H. sapiens | TM Tendency | 80.59 | 48 | 161 |
| TMD (18 aa) | H. sapiens | Kyte & Doolittle | 80.53 | 36 | 173 |
| Segment 15 | H. sapiens | Kyte & Doolittle | 79.78 | 59 | 150 |
| Patch 19 | H. sapiens | Kyte & Doolittle | 79.55 | 61 | 148 |
| Segment 15 | H. sapiens | TM Tendency | 79.24 | 79 | 130 |
| TMD | H. sapiens | TM Tendency | 79.17 | 67 | 142 |
| TMD (18 aa) | H. sapiens | TM Tendency | 78.91 | 67 | 142 |
| Wheel Face 8 | H. sapiens | Kyte & Doolittle | 78.90 | 49 | 160 |
| TMD (avg) | H. sapiens | Kyte & Doolittle | 78.89 | 49 | 160 |
| Wheel Face 9 | H. sapiens | Kyte & Doolittle | 78.88 | 62 | 147 |
| Segment 19 | H. sapiens | TM Tendency | 78.40 | 80 | 129 |
| Wheel Face 9 | H. sapiens | TM Tendency | 77.47 | 74 | 135 |
| Wheel Face 8 | H. sapiens | TM Tendency | 77.23 | 61 | 148 |
| Wheel Face 7 | H. sapiens | Kyte & Doolittle | 77.19 | 75 | 134 |
| Wheel Face 7 | H. sapiens | TM Tendency | 76.99 | 81 | 128 |
| Rectangle 9 | H. sapiens | TM Tendency | 76.76 | 63 | 146 |
| Wheel Face 6 | H. sapiens | Kyte & Doolittle | 76.52 | 81 | 128 |
| Wheel Face 6 | H. sapiens | TM Tendency | 76.44 | 65 | 144 |
| Twist 8 | H. sapiens | Kyte & Doolittle | 75.75 | 74 | 135 |
| Star 8 | H. sapiens | TM Tendency | 75.71 | 74 | 135 |
| Wheel Face 5 | H. sapiens | Kyte & Doolittle | 75.55 | 86 | 123 |
| Twist 8 | H. sapiens | TM Tendency | 75.37 | 80 | 129 |
| TMD (avg) | H. sapiens | TM Tendency | 75.06 | 79 | 130 |
| Rectangle 9 | H. sapiens | Kyte & Doolittle | 74.84 | 73 | 136 |
| Wheel Face 4 | H. sapiens | Kyte & Doolittle | 72.90 | 87 | 122 |

Continuation of Table S2.6

| metric | Organism | scale | AUROC score | misclassified | correctly classified |
|---|---|---|---|---|---|
| Line 9 | H. sapiens | TM Tendency | 72.76 | 58 | 151 |
| Wheel Face 5 | H. sapiens | TM Tendency | 70.75 | 98 | 111 |
| Wheel Face 4 | H. sapiens | TM Tendency | 70.65 | 70 | 139 |
| Line 13 | H. sapiens | Kyte & Doolittle | 70.51 | 53 | 156 |
| Line 9 | H. sapiens | Kyte & Doolittle | 69.50 | 59 | 150 |
| Star 8 | H. sapiens | Kyte & Doolittle | 69.32 | 73 | 136 |
| Line 17 | H. sapiens | Kyte & Doolittle | 69.28 | 61 | 148 |
| Line 13 | H. sapiens | TM Tendency | 69.00 | 93 | 116 |
| Line 17 | H. sapiens | TM Tendency | 67.77 | 64 | 145 |
| Wheel Face 3 | H. sapiens | TM Tendency | 65.30 | 55 | 154 |
| Wheel Face 3 | H. sapiens | Kyte & Doolittle | 64.40 | 104 | 105 |
| TMD length | H. sapiens | | 59.14 | 145 | 64 |
| CTE negative charge | H. sapiens | | -61.06 | 45 | 164 |
| CTE net charge | H. sapiens | | -62.47 | 164 | 45 |
| CTE positive charge | H. sapiens | | -72.68 | 47 | 162 |

Table S2.6: **Hydrophobic geometry metrics better classify human TA proteins than total TMD hydrophobicity metrics.** A list of all the metrics tested against the human genome ranked from highest to lowest AUROC score.

*Chapter 3*

# THE CLIENT-BINDING DOMAIN OF THE COCHAPERONE SGT2 HAS A HELICAL-HAND STRUCTURE THAT BINDS A SHORT HYDROPHOBIC HELIX

*Adapted from:*

Lin, Ku-Feng et al. (Jan. 2021). "Molecular basis of tail-anchored integral membrane protein recognition by the cochaperone Sgt2". In: *Journal of Biological Chemistry* 296. DOI: 10.1016/j.jbc.2021.100441.

M.Y. Fry designed and executed pull-down experiments pertaining to the identification the minimal binding domain and verification the computational model of Sgt2.

## Abstract

The targeting and insertion of tail-anchored (TA) integral membrane proteins (IMP) into the correct membrane is critical for cellular homeostasis. The fungal protein Sgt2, and its human homolog SGTA, binds hydrophobic clients and is the entry point for targeting of ER-bound TA proteins. Here we reveal molecular details that underlie the mechanism of Sgt2 binding to clients. We establish that the Sgt2 C-terminal region is flexible but conserved and sufficient for client binding. A molecular model for this domain reveals a helical hand forming a hydrophobic groove, consistent with a higher affinity for client TMDs with hydrophobic faces and a minimal length of 11 residues. This work places Sgt2 into a broader family of TPR-containing co-chaperone proteins.

## 3.1 Introduction

An inherently complicated problem of cellular homeostasis is the biogenesis of hydrophobic IMPs which are synthesized in the cytoplasm and must be targeted and inserted into a lipid bilayer. Accounting for ~25% of transcribed genes (Pieper et al., 2013), IMPs are primarily targeted by cellular signal binding factors that recognize a diverse set of hydrophobic $\alpha$-helical signals as they emerge from the ribosome (Aviram and Schuldiner, 2017; Shao and Hegde, 2011b; Guna and Hegde, 2018). One important class of IMPs are tail-anchored (TA) proteins whose hydrophobic signals are their single helical transmembrane domain (TMD) located near the C-terminus and are primarily targeted post-translationally to either the ER or mitochondria (Kutay, Hartmann, and Rapoport, 1993; Hegde and Keenan, 2011; Denic, 2012; Wattenberg and Lithgow, 2001; Chartron, Clemons, and Suloway, 2012). In the case of the canonical pathway for ER-destined TA proteins, each is first recognized by homologs of mammalian SGTA (small glutamine tetratricopeptide repeat protein) (Chio, Cho, and Shan, 2017; Hegde and Keenan, 2011; Guna and Hegde, 2018; Shao and Hegde, 2011a). Common to all signal binding factors is the need to recognize, bind, and then hand off a hydrophobic helix. How such factors can maintain specificity to a diverse set of hydrophobic clients that must subsequently be released remains an important question.

Homologs of Saccharomyces cerevisiae Sgt2 (ySgt2) and Homo sapiens SGTA (referred to here as hSgt2 and collectively Sgt2 for simplicity), are involved in a variety of cellular processes regarding the homeostasis of membrane proteins including the targeting of TA proteins (Chartron, Clemons, and Suloway, 2012; Chartron, Gonzalez, and Clemons, 2011; F. Wang, Brown, et al., 2010; Simon et al., 2013), retrograde transport of membrane proteins for ubiquitination and subsequent proteasomal degradation (Y. Xu, Cai, et al., 2012), and regulation of mislocalized membrane proteins (MLPs) (Wunderley et al., 2014; Pawel Leznicki and High, 2012). Among these, the role of Sgt2 in the primary pathways responsible for targeting TA proteins to the endoplasmic reticulum (ER) is best characterized, i.e. the fungal Guided Entry of Tail-anchored proteins (GET) or the mammalian Transmembrane Recognition Complex (TRC) pathway. In the GET pathway, Sgt2 functions by binding a cytosolic TA protein then transferring the TA protein to the ATPase chaperone Get3 (human homolog is also Get3) with the aid of the heteromeric Get4/Get5 complex (human Get4/Get5/Bag6 complex) (F. Wang, Brown, et al., 2010; Gristick et al., 2014; Mock et al., 2015; F. Wang, Whynot, et al., 2011). In this process, TA protein binding to Sgt2, after hand-off from Hsp70, is proposed

as the first committed step to ensure that ER TA proteins are delivered to the ER membrane while mitochondrial TA proteins are excluded (Shao and Hegde, 2011b; F. Wang, Brown, et al., 2010; Cho and Shan, 2018). Subsequent transfer of the TA protein from Sgt2 to the ATP bound Get3 induces conformational changes in Get3 that trigger ATP hydrolysis, releasing Get3 from Get4 and favoring binding of the Get3-TA protein complex to the Get1/2 receptor at the ER leading to release of the TA protein into the membrane (Stefer et al., 2011; Rome, Chio, et al., 2014; Vilardi, Lorenz, and Dobberstein, 2011; Yamamoto and Sakisaka, 2012; Schuldiner, Metz, et al., 2008). Deletions of yeast GET genes (i.e. get1Δ, get2Δ, or get3Δ) cause cytosolic aggregation of TA proteins dependent on Sgt2 (Schuldiner, Metz, et al., 2008; Kiktev et al., 2012).

In addition to targeting TA proteins, there is evidence hSgt2 promotes degradation of IMPs through the proteasome by cooperating with the Bag6 complex, a heterotrimer containing Bag6, hGet4, and hGet5, which acts as a central hub for a diverse physiological network related to protein targeting and quality control (Mock et al., 2015; Y. Xu, Y. Liu, et al., 2013; Rodrigo-Brenni, Gutierrez, and Hegde, 2014; Hessa, Sharma, et al., 2011). The Bag6 complex can associate with ER membrane-embedded ubiquitin regulatory protein UbxD8, transmembrane protein gp78, proteasomal component Rpn10c, and an E3 ubiquitin protein ligase RNF126 thereby connecting hSgt2 to ER associated degradation (ERAD) and proteasomal activity. Depletion of hSgt2 significantly inhibits turnover of ERAD IMP clients and elicits the unfolded protein response (Wunderley et al., 2014). Furthermore, the cellular level of MLPs in the cytoplasm could be maintained by co-expression with hSgt2, which possibly antagonize ubiquitination of MLPs to prevent proteasomal degradation (Y. Xu, Cai, et al., 2012; Pawel Leznicki and High, 2012). These studies demonstrate an active role of hSgt2 in triaging IMPs in the cytoplasm and the breadth of hSgt2 clients including TA proteins, ERAD, and MLPs all harboring one or more TMD. Roles for hSgt2 in disease include polyomavirus infection (Dupzyk et al., 2017), neurodegenerative disease (Kiktev et al., 2012; Long et al., 2012), hormone-regulated carcinogenesis (Trotta et al., 2013; Buchanan et al., 2007), and myogenesis (H. Wang, Q. Zhang, and D. Zhu, 2003), although the underlying molecular mechanisms are still unclear.

The architecture of Sgt2 includes three structurally independent domains that define the three different interactions of Sgt2 (Fig. 3.1 *A*) (Chartron, VanderVelde, and Clemons, 2012; Chartron, Gonzalez, and Clemons, 2011; Liou and C. Wang, 2005;

Cziepluch et al., 1998; Callahan et al., 1998). The N-terminal domain forms a homo-dimer composed of a four-helix bundle with 2-fold symmetry that primarily binds to the ubiquitin-like domain (UBL) of Get5/Ubl4A for TA IMP targeting (Chartron, VanderVelde, and Clemons, 2012; Winnefeld et al., 2006) or interacts with the UBL on the N-terminal region of Bag6 (Darby et al., 2014) where it is thought to initiate downstream degradation processes (Y. Xu, Cai, et al., 2012; Y. Xu, Y. Liu, et al., 2013; Rodrigo-Brenni, Gutierrez, and Hegde, 2014). The central region comprises a co-chaperone domain with three repeated TPR motifs arranged in a right handed-superhelix forming a 'carboxylate clamp' for binding the C-terminus of heat-shock proteins (HSP) (Chartron, Gonzalez, and Clemons, 2011; Dutta and Tan, 2008). The highly conserved TPR domain was demonstrated to be critical in modulating propagation of yeast prions by recruiting HSP70 (Kiktev et al., 2012) and may associate with the proteasomal factor Rpn13 to regulate MLPs (Leznicki et al., 2015). More recently, it was demonstrated that mutations to residues in the TPR domain which prevent Hsp70 binding impair the loading of TA proteins onto ySgt2 (Cho and Shan, 2018), consistent with a direct role of Hsp70 in TA IMP targeting via the TPR domain. The C-terminal methionine-rich domain of Sgt2 is responsible for binding to hydrophobic clients such as TA proteins (F. Wang, Brown, et al., 2010; Liou and C. Wang, 2005). Other hydrophobic segments have been demonstrated to interact with this domain such as the membrane protein Vpu (viral protein U) from human immunodeficiency virus type-1 (HIV-1) , the TMD of tetherin (Waheed et al., 2016), the signal peptide of myostatin (H. Wang, Q. Zhang, and D. Zhu, 2003), and the N-domain of the yeast prion forming protein Sup35 (Kiktev et al., 2012). All of these studies suggest that the C-terminus of Sgt2 binds broadly to hydrophobic stretches, yet structural and mechanistic information for client recognition is lacking.

In this study, we provide the first structural characterization of the C-domains from Sgt2 (Sgt2-C) and show that, in the absence of client, it is relatively unstructured. We demonstrate that a conserved region of the C-domain, defined here as $C_{cons}$, is sufficient for client binding. Analysis of the $C_{cons}$ sequence identifies six amphipathic helices whose hydrophobic residues are required for client binding. Based on this, we computationally generate an ab initio structural model that is validated by point mutants and disulfide crosslinking. Artificial clients are then used to define the properties within clients critical for binding to Sgt2-C. The results show that Sgt2-C falls into a larger STI1 family of TPR-containing co-chaperones and allow us to propose a mechanism for client binding.

## 3.2 Results

### The flexible Sgt2-C domain

Based on sequence alignment (Fig. 3.1 *A*), the Sgt2-C contains a conserved core of six predicted helices flanked by unstructured loops that vary in length and sequence. Previous experimental work suggested that this region is particularly flexible, as this domain in the Aspergillus fumigatus homolog is sensitive to proteolysis (Chartron, Gonzalez, and Clemons, 2011). Similarly, for ySgt2-TPR-C, the sites sensitive to limited proteolysis primarily occur within the loops flanking the conserved helices (Fig. 3.1 *A*, *red arrows* & Fig. 3.2 *B*). This flexible nature of the C-domain likely contributes to its anomalous passage through a gel-filtration column where Sgt2-C elutes much earlier than the similarly-sized, but well-folded, Sgt2 TPR-domain (Fig.3.1 *B*), as is typical for unstructured proteins (Graether, 2019). The larger hydrodynamic radius matches previous small-angle X-ray scattering measurement of the ySgt2 TPR-C domain that indicated a partial unfolded characteristic in a Kratky plot analysis. The circular dichroism (CD) spectra for both homologs suggests that the C-domain and a predicted six $\alpha$-helical methionine-rich region of Sgt2-C (Fig. 3.1 *A*), hereafter referred to as Sgt2-$C_{cons}$, largely assume a random-coil conformation, with 40-45% not assignable to a defined secondary structure category (Fig. 3.1 *C*, 3.2 *A*) (Luo and Baldwin, 1997). The well-resolved, sharp, but narrowly dispersed chemical shifts of the backbone amide protons in $^{1}$H-$^{15}$N HSQC spectra of Sgt2-C (Fig. 3.1 *D,E*) and Sgt2-$C_{cons}$ (Fig. 3.2 *B,C*), indicate a significant degree of backbone mobility, similar to natively unfolded proteins (Dyson and Wright, 2004) and consistent with results seen by others (Martínez-Lumbreras et al., 2018), further highlighting the lack of stable tertiary structure (Chartron, Gonzalez, and Clemons, 2011). Taken all together, Sgt2-C appears to be a flexible domain.

### The conserved region of the C-domain is sufficient for client binding

We then asked if the flexible Sgt2-C is the site of client binding in the co-chaperone and if so, where within this domain is the binding region. During purification Sgt2-C is susceptible to proteolytic activity being cut at several specific sites (Fig. 3.1*A*). Proteolysis occurred primarily at $Leu_{327}$ and in the poorly conserved N-terminal region (between $Asp_{235}$-$Gly_{258}$). Given the intervening region (ySgt2 $Gly_{258}$-$Leu_{327}$) is conserved (Fig. 3.1*A*), it and the corresponding region in hSgt2 may mediate client binding (Fig. 3.3 *A, grey*). To test this, we established a set of his-tagged Sgt2 constructs of various lengths (Fig. 3.3*C* & 3.4*A*). These Sgt2-C truncations were co-expressed with an MBP-tagged client, Sbh1, and binding

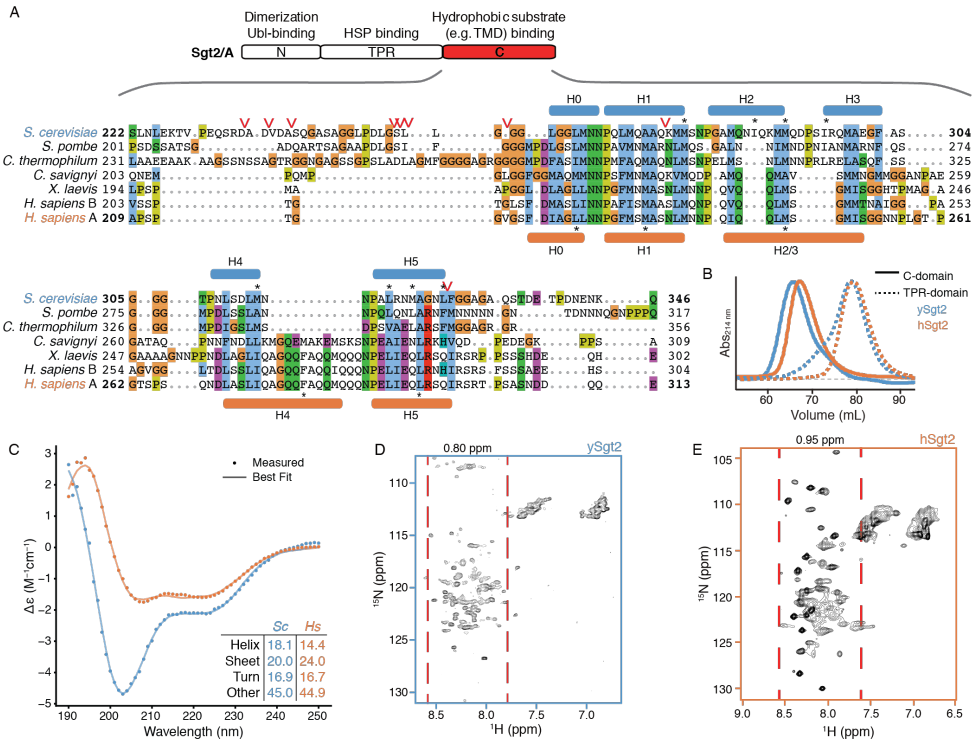



Figure 3.1: **Structural characteristics of free Sgt2 C-domain.**
*A)* Top, Schematic of the domain organization of Sgt2. Below, representative sequences from a large-scale multiple sequence alignment of the C domain: fungal Sgt2 from *S. cerevisiae*, *S. pombe*, and *C. thermophilum* and metazoan Sgt2 from *C. savignyi*, *X. laevis*, and *H. sapiens*. Protease susceptible sites on ySgt2-C identified by mass spectrometry are indicated by red arrowheads. Predicted helices of ySgt2 (*blue*) and hSgt2 (*orange*) by Jpred (Drozdetskiy et al., 2015) and/or structure prediction are shown. Blue/orange color scheme for ySgt2/hSgt2 is used throughout the text. Residues noted in the text are highlighted by an asterisk. *B)* Overlay of size-exclusion chromatography traces of ySgt2-C (*blue line*), hSgt2-C (*orange line*), ySgt2-TPR (*blue dash*), and hSgt2-TPR (*orange dash*). Traces are measured at 214nm, baseline-corrected and normalized to the same peak height. *C)* Far UV CD spectrum of 10$\mu$M of purified ySgt2-C (*blue*) and hSgt2-C (*orange*) at RT with secondary structure decomposition from BestSel (Micsonai et al., 2015). *D)* $^{1}$H-$^{15}$N HSQC spectrum of ySgt2-C at 25°C. The displayed chemical shift window encompasses all N-H resonances from both backbone and side chains. The range of backbone amide protons, excluding possible side-chain $NH_2$ of Asn/Gln, is indicated by pairs of red dashed lines. *E)* As in *(D)* for hSgt2-C at 25°C.

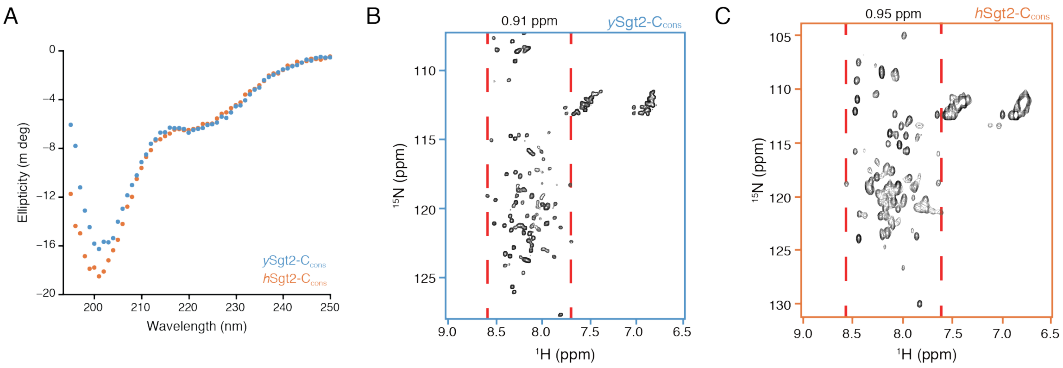



Figure 3.2: **Structural characteristics of free Sgt2 C-domain.**
*A)* CD spectra as in Fig. 3.1*C* for the conserved C-terminal domains of ySgt2 (*blue*) and hSgt2 (*orange*). NMR spectra as in Fig. 3.4 (*D*&*E*) for ySgt2-C$_{cons}$ (*B, blue*) and hSgt2-C$_{cons}$ (*C, orange*).

was detected by the presence of captured TA proteins in nickel elution fractions (Fig. 3.3*B*). The TA protein Sbh1 is the yeast homolog of the mammalian Sec61$\gamma$, a component of the ER-resident Sec translocon. While the relative efficiency of MBP-Sbh1 capture cannot be assessed in this assay due to differences in total protein levels (Fig. 3.4*B*), we can demonstrate the ability of a given construct to bind to the client. As previously seen (F. Wang, Brown, et al., 2010), we confirm that Sgt2-TPR-C alone is sufficient for capturing a client (Fig. 3.3*C*). As one might expect, the C-domain was also sufficient for binding the client. Interestingly, Sgt2-C$_{cons}$ is sufficient for binding to Sbh1. Even a minimal region of the last 5 helices (referred to as $\Delta$H0) also captures Sbh1 (Fig. 3.3*C*). The predicted helices in Sgt2-C$_{cons}$ are amphipathic and their hydrophobic faces may be used for client binding (Fig. 3.3*D*).

Each of the six helices in Sgt2-C$_{cons}$ was mutated to replace the larger hydrophobic residues with alanines, dramatically reducing the overall hydrophobicity. For all of the helices, alanine replacement of the hydrophobic residues significantly reduces binding of Sbh1 to Sgt2-C (Fig. 3.3*E & F*). While these mutants expressed at similar levels to the wild-type sequence, one cannot rule out that some of these changes may affect the tertiary structure of this domain. In general, these results imply that these amphipathic helices are necessary for client binding since removal of the hydrophobic faces disrupts binding. The overall effect on binding by each helix is different, with mutations in helices 1-3 having the most dramatic reduction in binding suggesting that these are more crucial for Sgt2-client complex formation. It is also worth noting, as this is a general trend, that hSgt2 is more resistant to mutations that affect binding (Fig. 3.3*F*) than ySgt2, which likely reflect different

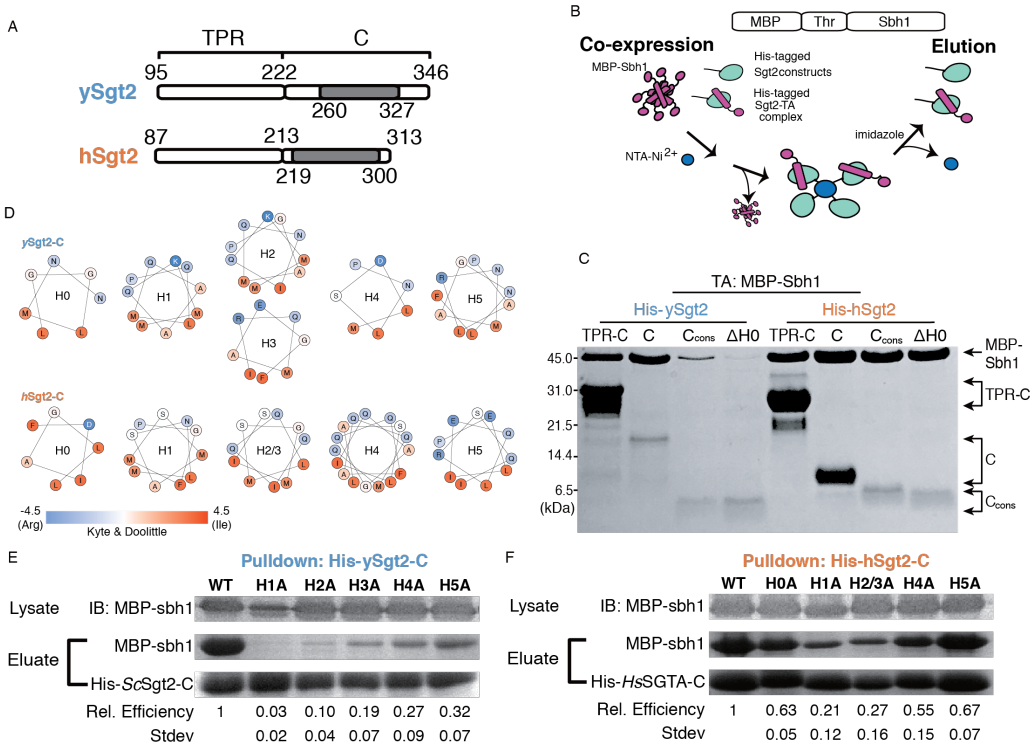



Figure 3.3: **The minimal binding region of Sgt2 for client binding.**
*A)* Diagram of the protein truncations tested for client binding that include the TPR-C domain, C-domain (C), $C_{cons}$, and $C_{cons}$ ΔH0 (ΔH0) from ySgt2 and hSgt2. The residues corresponding to each domain are indicated, and grey blocks highlight the $C_{cons}$ region. *B)* Schematic of capture experiments of MBP-tagged Sbh1 separated by a thrombin (Thr) cleavage site (MBP-Sbh1) by Sgt2 variants. After co-expression, cell pellets are lysed and NTA-$Ni^{2+}$ is used to capture his-tagged Sgt2-TPR-C. *C)* Tris-Tricine-SDS-PAGE gel (Schägger, 2006) of co-expressed and purified MBP-Sbh1 and his-tagged Sgt2 truncations visualized with Coomassie Blue staining. *D)* Helical wheel diagrams of predicted helices (see Fig. 3.1*A*) in the $C_{cons}$ domain of ySgt2 and hSgt2. Residues are colored by the Kyte and Doolittle hydrophobicity scale (Kyte and Doolittle, 1982). *E)* All of the hydrophobic residues (L, I, F, and M) in a predicted helix (H0, H1, etc.) are replaced with alanines and tested for the ability to capture MBP-Sbh1. Protein levels were quantified by Coomassie staining. Relative binding efficiency of MBP-Sbh1 by ySgt2 C-domain (ySgt2-C) variants was calculated relative to total amount of ySgt2-C captured (MBP-Sbh1/Sgt2-C) then normalized to the wild-type ySgt2-C. Experiments were performed 3-4 times and the standard deviations are presented. Total expression levels of the MBP-Sbh1 were similar across experiments as visualized by immunoblotting (IB) of the cell lysate. *F)* As in *(E)* but for hSgt2.

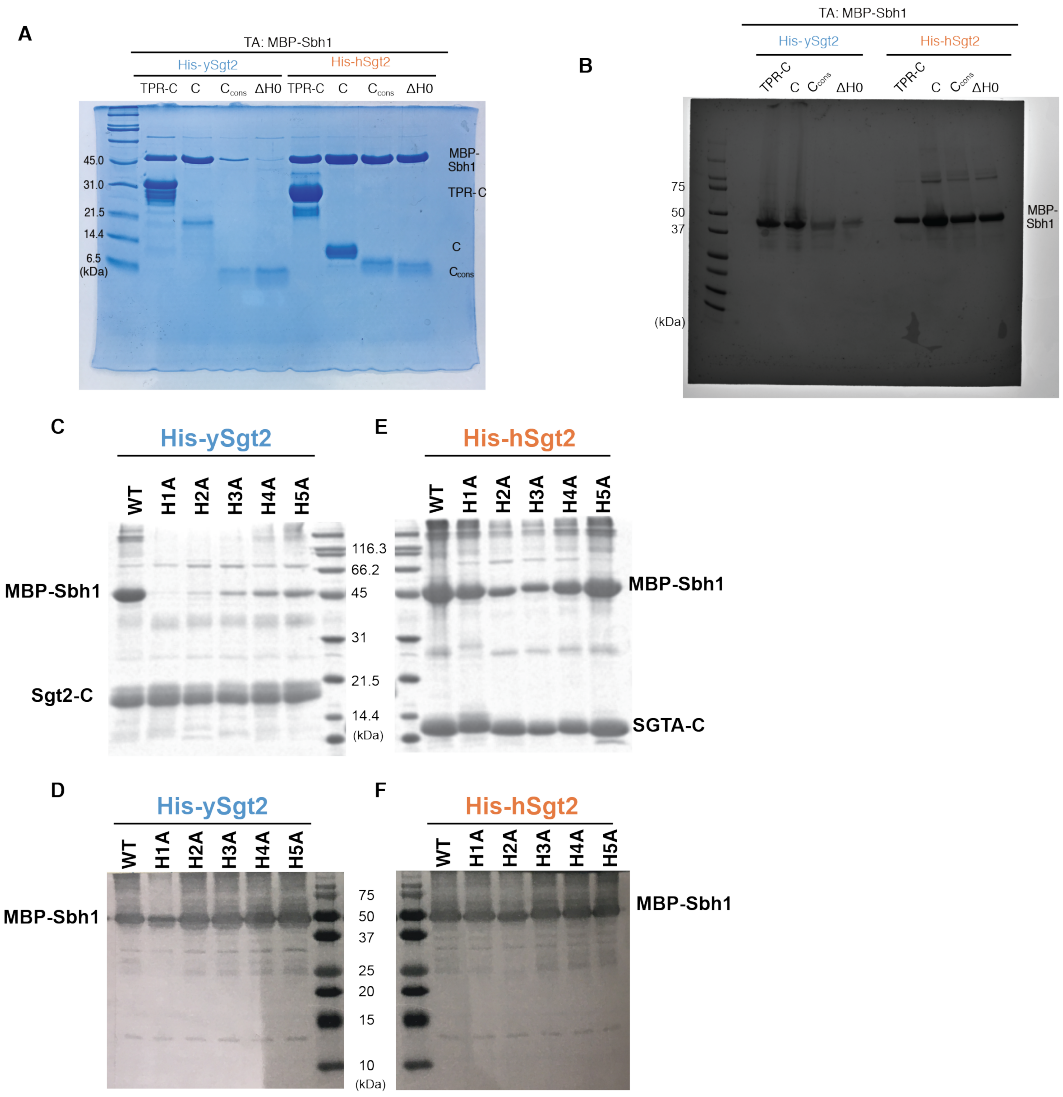



Figure 3.4: **Identification of minimal binding region of Sgt2.**
*A)* The full image of the gel in Fig. 3.3*C*. *B)* An anti-MBP western blot of the lysate from the which the complexes in *(A)* were purified from. The load concentrations were normalized based on the total optical density of the cells when harvested.

thresholds for binding.

### Molecular modeling of Sgt2-C domain

Despite the need for a molecular model, the C-domain has resisted structural studies, likely due to the demonstrated inherent flexibility. Based on the six conserved $\alpha$-helical amphipathic segments (Fig. 3.1*A*) that contain hydrophobic residues critical for client binding (Fig. 3.3*D*-*E*), we expect some folded structure to exist. Therefore, we performed ab initio molecular modeling of Sgt2-C using a variety of prediction

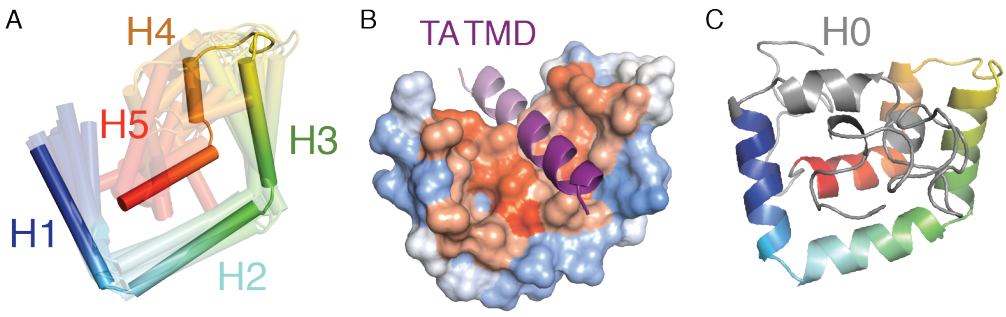



Figure 3.5: **A structural model for Sgt2-C$_{cons}$.**
*A)* The top 10 models of the ySgt2-C$_{cons}$ generated by the template-free algorithm Quark (D. Xu and Y. Zhang, 2012) are overlaid with the highest scoring model in solid. Models are color-ramped from N- (*blue*) to C-terminus (*red*). *B)* A model of ySgt2-C$_{cons}$ (surface colored by Kyte-Doolittle hydrophobicity) bound to a TMD (*purple helix*) generated by rigid-body docking through Zdock (Pierce et al., 2014). The darker purple corresponds to an 11 residue stretch. *C)* The entire ySgt2-C from the highest scoring model from Quark (C$_{cons}$ in rainbow with the rest in grey) highlighting H0 and the rest of the flexible termini that vary considerably across models.

methods resulting in a diversity of putative structures [48-52]. As expected, all models showed buried hydrophobic residues as this is a major criterion for in silico protein folding. Residues outside the ySgt2-C$_{cons}$ region adopted varied conformations consistent with their expected higher flexibility. Pruning these N- and C-terminal regions to focus on the ySgt2-C$_{cons}$ region (Fig. 3.6*A*) revealed a potential binding interface for a hydrophobic client. Examples are seen in Quark models (1, 4,& 6 shown), Robetta 1 & 2, and I-TASSER 2 & 3, whereas others models had no clearly distinguishable groove. Given the intrinsic flexibility of the Sgt2-C domain, it is possible that models without a groove are found in the non-TMD bound structural ensemble.

For a working model of TMD-bound ySgt2-C, we chose the highest scoring Quark structures where a general consistent architecture is seen (Fig. 3.5*A*) (D. Xu and Y. Zhang, 2012). The overall model contained a potential client binding site, a hydrophobic groove formed by the amphipathic helices. The groove is approximately 15 Å long, 12 Å wide, and 10 Å deep, which is sufficient to accommodate three helical turns of an $\alpha$-helix, ~11 amino acids (Fig. 3.5*B*).

To validate the model, we interrogated the accuracy of the predicted structural arrangement by determining distance constraints from crosslinking experiments.

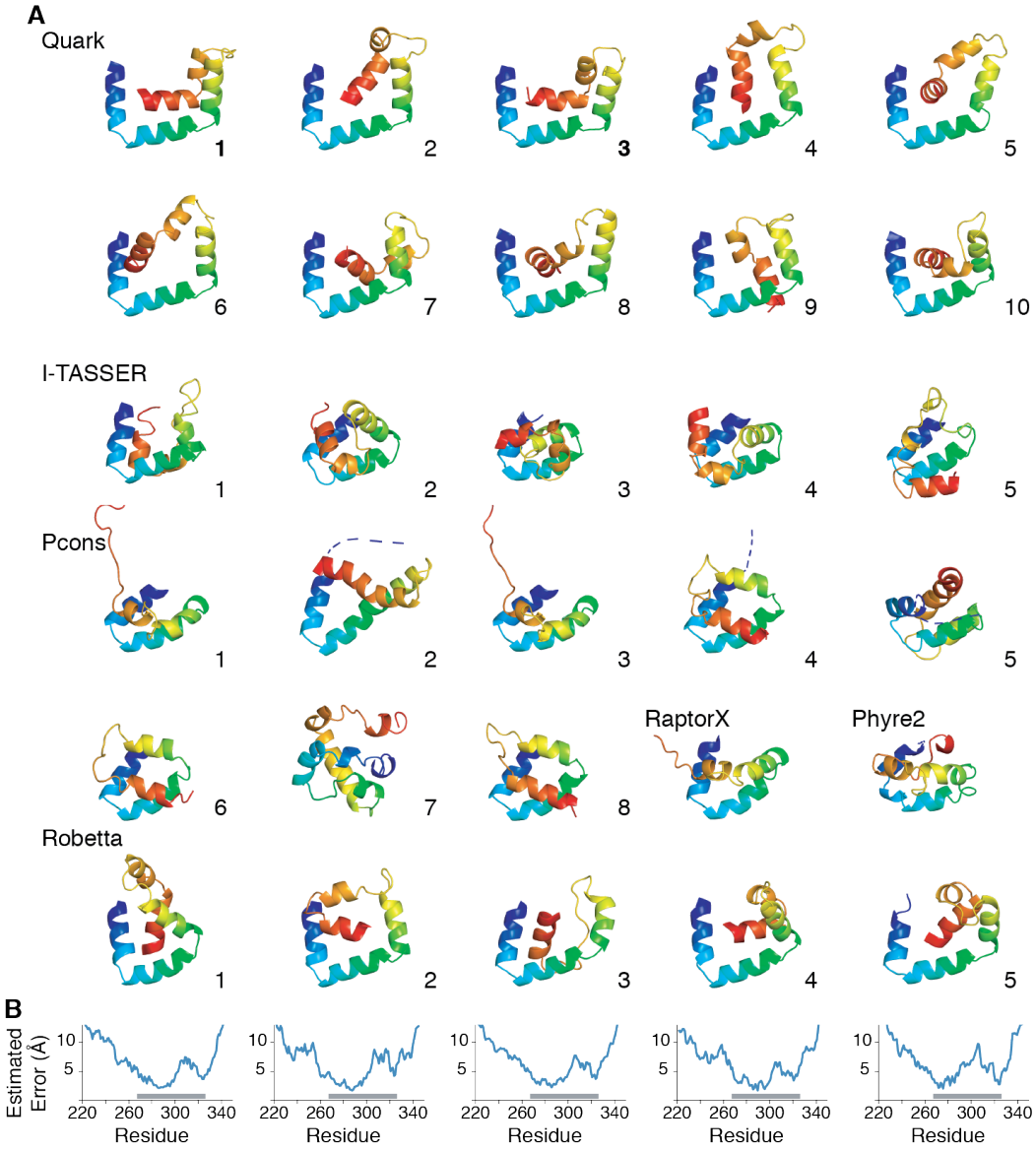
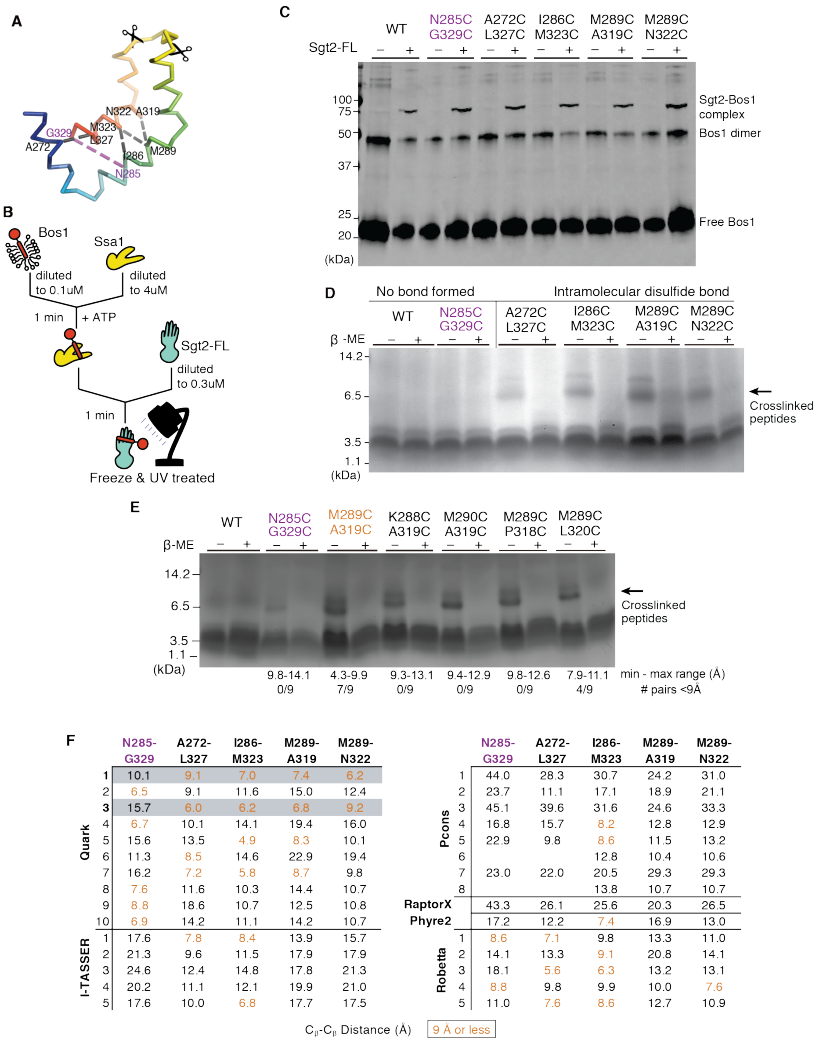



Figure 3.6: **Structural models across prediction methods.**
*A)* Predictions from Quark, I-TASSER, Pcons, Phyre2, RaptorX, and Robetta. Methods produce between 5 and 10 models. *B)* Robetta provides a residue-wise estimated error in Angstroms; this is shown below the corresponding models with a grey bar indicating the $C_{cons}$ region.

We selected four pairs of residues in close spatial proximity and one pair far apart based on the Quark models (Fig. 3.7*A*). Calculating a C$\beta$-C$\beta$ distance between residue pairs for each model (Fig. 3.7*F*), the Quark models 1 and 3 were the most consistent with an expected distance of 9Å or less for the close pairs. In all alternative models, the overall distances are much larger and should not be expected to form disulfide bonds *in vitro* if they represent a TMD-bound state. For Robetta, a number of the models have pairs of residues within 9Å and Robetta's per-residue error estimate suggests relatively high confidence in the $C_{cons}$ region (Fig. 3.6*B*).

As a control, we first confirmed that the cysteine-mutant pairs do not affect the function of ySgt2. We utilized an in vitro capture assay where a yeast Hsp70 homolog Ssa1 loaded with a TA protein, Bos1, delivers the client to ySgt2 (Cho and Shan, 2018; Chio, Chung, et al., 2019; Shao, Rodrigo-Brenni, et al., 2017) (Fig.3.7*C*). Purified Ssa1 is mixed with detergent solubilized strep-tagged Bos1-TMD (a model ER TA protein) that contained a p-benzoyl-l-phenylalanine (BPA) labeled residue, $Bos1_{BPA}$, and diluted to below the critical micelle concentration resulting in soluble complexes of $Bos1_{BPA}$/Ssa1. Full-length ySgt2 variants were each tested for the ability to capture $Bos1_{BPA}$ from Ssa1. After the transfer reaction, each was UV-treated to generate Bos1 crosslinks. Successful capture of the TA proteins by ySgt2 was detected for all cysteine variants using an anti-strep Western blot and the appearance of a $Bos1_{BPA}$/ySgt2 crosslink band, suggesting the mutations do not affect the structure or function of ySgt2 (Fig. 3.7*C*).

We and others have demonstrated that a monomeric Sgt2 is sufficient for binding to clients (F. Wang, Brown, et al., 2010). For the distance experiment, each of the cysteine-mutant pairs was made in the more stable monomeric variant ySgt2-TPR-C. Each variant was coexpressed with an artificial client – a cMyc-tagged BRIL (small, 4-helix bundle protein used in previous work to aid in the crystallization of GPCRs (Chun et al., 2012)) with a C-terminal TMD consisting of eight leucines and three alanines, denoted 11[L8], and purified via nickel-affinity chromatography in reducing buffer (Fig. 3.8*A*). All of the ySgt2 mutants bound to the client and behaved similar to the wild-type (cysteine-free) further suggesting the mutants did not perturb the native structure (Fig. 3.8*B*). For disulfide crosslink formation, each eluate was oxidized, digested using the protease Glu-C, and crosslinks were identified by the visualization of a reducing-agent sensitive ~7.7kDa fragment in gel electrophoresis (Fig. 3.7*D*). For both the wild-type construct and in N285C/G329C, where the pairs are predicted from the Quark models to be too distant for disulfide bond formation,




F

| | | N285-G329 | A272-L327 | I286-M323 | M289-A319 | M289-N322 |
|---|---|---|---|---|---|---|
| Quark | 1 | 10.1 | 9.1 | 7.0 | 7.4 | 6.2 |
| | 2 | 6.5 | 9.1 | 11.6 | 15.0 | 12.4 |
| | 3 | 15.7 | 6.0 | 6.2 | 6.8 | 9.2 |
| | 4 | 6.7 | 10.1 | 14.1 | 19.4 | 16.0 |
| | 5 | 15.6 | 13.5 | 4.9 | 8.3 | 10.1 |
| | 6 | 11.3 | 8.5 | 14.6 | 22.9 | 19.4 |
| | 7 | 16.2 | 7.2 | 5.8 | 8.7 | 9.8 |
| | 8 | 7.6 | 11.6 | 10.3 | 14.4 | 10.7 |
| | 9 | 8.8 | 18.6 | 10.7 | 12.5 | 10.8 |
| | 10 | 6.9 | 14.2 | 11.1 | 14.2 | 10.7 |
| I-TASSER | 1 | 17.6 | 7.8 | 8.4 | 13.9 | 15.7 |
| | 2 | 21.3 | 9.6 | 11.5 | 17.9 | 17.9 |
| | 3 | 24.6 | 12.4 | 14.8 | 17.8 | 21.3 |
| | 4 | 20.2 | 11.1 | 12.1 | 19.9 | 21.0 |
| | 5 | 17.6 | 10.0 | 6.8 | 17.7 | 17.5 |

| | | N285-G329 | A272-L327 | I286-M323 | M289-A319 | M289-N322 |
|---|---|---|---|---|---|---|
| Pcons | 1 | 44.0 | 28.3 | 30.7 | 24.2 | 31.0 |
| | 2 | 23.7 | 11.1 | 17.1 | 18.9 | 21.1 |
| | 3 | 45.1 | 39.6 | 31.6 | 24.6 | 33.3 |
| | 4 | 16.8 | 15.7 | 8.2 | 12.8 | 12.9 |
| | 5 | 22.9 | 9.8 | 8.6 | 11.5 | 13.2 |
| | 6 | | | 12.8 | 10.4 | 10.6 |
| | 7 | 23.0 | 22.0 | 20.5 | 29.3 | 29.3 |
| | 8 | | | 13.8 | 10.7 | 10.7 |
| RaptorX | | 43.3 | 26.1 | 25.6 | 20.3 | 26.5 |
| Phyre2 | | 17.2 | 12.2 | 7.4 | 16.9 | 13.0 |
| Robetta | 1 | 8.6 | 7.1 | 9.8 | 13.3 | 11.0 |
| | 2 | 14.1 | 13.3 | 9.1 | 20.8 | 14.1 |
| | 3 | 18.1 | 5.6 | 6.3 | 13.2 | 13.1 |
| | 4 | 8.8 | 9.8 | 9.9 | 10.0 | 7.6 |
| | 5 | 11.0 | 7.6 | 8.6 | 12.7 | 10.9 |

$C_{\beta}$-$C_{\beta}$ Distance (Å) 9 Å or less

Figure 3.7: **Validating the structural model with disulfide bond formation.** Variants of His-ySgt2-TPR-C (WT or cysteine double mutants) were co-expressed with the artificial client, cMyc-BRIL-11[L8]. After lysis, ySgt2-TPR-C proteins were purified, oxidized, then digested by Glu-C protease and analyzed by gel either in non-reducing or reducing buffer. *A)* C$\alpha$ ribbon of ySgt2-$C_{cons}$ color-ramped with various pairs of cysteines highlighted. Scissors indicate protease cleavage sites resulting in fragments less than 3 kDa in size. *B)* A schematic of the transfer of $Bos1_{BPA}$ from Ssa1 to full-length ySgt2 to demonstrate the double cysteine mutants are still functional. *C)* A western blot visualizing cross-linked ySgt2-Bos1 complexes. All samples tested, WT, N285C/G329C, A272C/L327C, I286C/M323C, M289C/A319C, and M289C/N322C, had a higher molecular weight appear after the addition of ySgt2 which corresponds to the size of the cross-linked complex (Fernandez-Patron et al., 1995). For the WT (cys-free) no significant difference was found between samples in non-reducing vs. reducing conditions. All close residue pairs (A272/L327, I286/M323, M289/A319, and M289/N322) show peptide fragments (higher MW) sensitive to the reducing agent and indicate disulfide bond formation (indicated by arrow). A cysteine pair (N285/G329) predicted to be far apart by the model does not result in the higher MW species. *E)* Tris-Glycine SDS-PAGE gel probing the flexibility of ySgt2-$C_{cons}$. All new pairs (K288/A319, M290/A319, M289/P318, M289/L320) show peptide fragments sensitive to the reducing agent (*indicated by arrow*). The range of distances of the eight closest possible rotamer pairs is annotated below. The cysteine pair (N285/G329) shown to be far apart by the model does have a faint higher molecular weight band. (F) C$\beta$-C$\beta$ distances between the residues mutated to cysteines based on various models predicted by the Quark, I-TASSER, PCONS, and Robetta. Cysteine pairs that are 9Å or less colored in orange and are expected to be close enough to form disulfide bonds. Where all five pair distances are consistent with the experiment (4 near and 1 far), the row is shaded in grey.

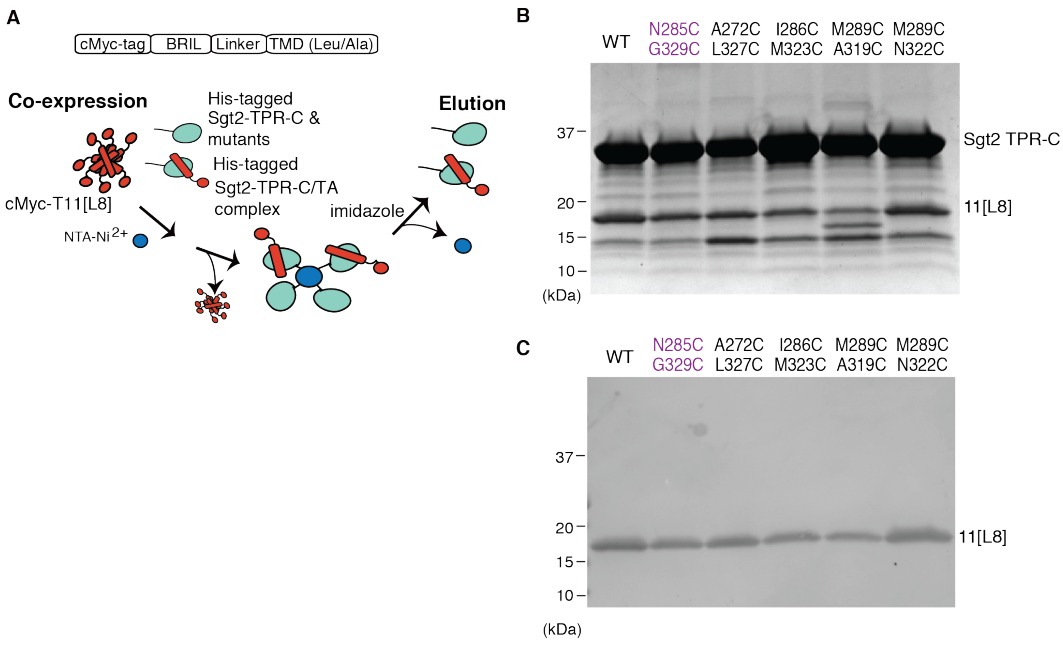



Figure 3.8: **Cysteine mutants are capable of binding to clients.**
*A)* Schematic showing how his-tagged ySgt2-TPR-C and double cysteine mutant constructs were coexpressed with the client 11[L8], and complexes were purified by nickel affinity chromatography. *B)* A coomassie stained SDS-PAGE gel of the elution fractions demonstrates that 11[L8] was present in the elution suggesting double cysteine mutations do not affect client binding. *C)* An anti-cMyc western blot of the fractions represented in the SDS-PAGE gel also demonstrates that 11[L8] was present in all eluates.

no higher molecular weight band was observed. For the remaining pairs that are predicted to be close enough for bond formation, the 7.7kDa fragment was observed in each case and is labile in reducing conditions. Again, these results support the $C_{cons}$ model derived from Quark.

With the four crosslinked pairs as distance constraints, new models were generated using Robetta with a restraint on the corresponding pairs of C$\beta$ atoms less than 9Å (Fig. 3.9*A*). The Robetta models from these runs are similar to the top scoring models from Quark (Fig. 3.5). Satisfyingly, the pair of residues that do not form disulfide crosslinks are generally consistent (Fig. 3.9*B*).

The improvement of the ySgt2 models predicted by Robetta with restraints included encouraged us to generate models for hSgt2-C with constraints. For this, pairs were defined based on sequence alignments of Sgt2 (Fig. 3.1*A*) and used as restraints. The resulting predictions had architectures consistent with the equivalent regions predicted for ySgt2-$C_{cons}$, for example Robetta 4 (Fig. 3.9*C, top*). Although in general the predicted hSgt2 model is similar to that for ySgt2, the region that

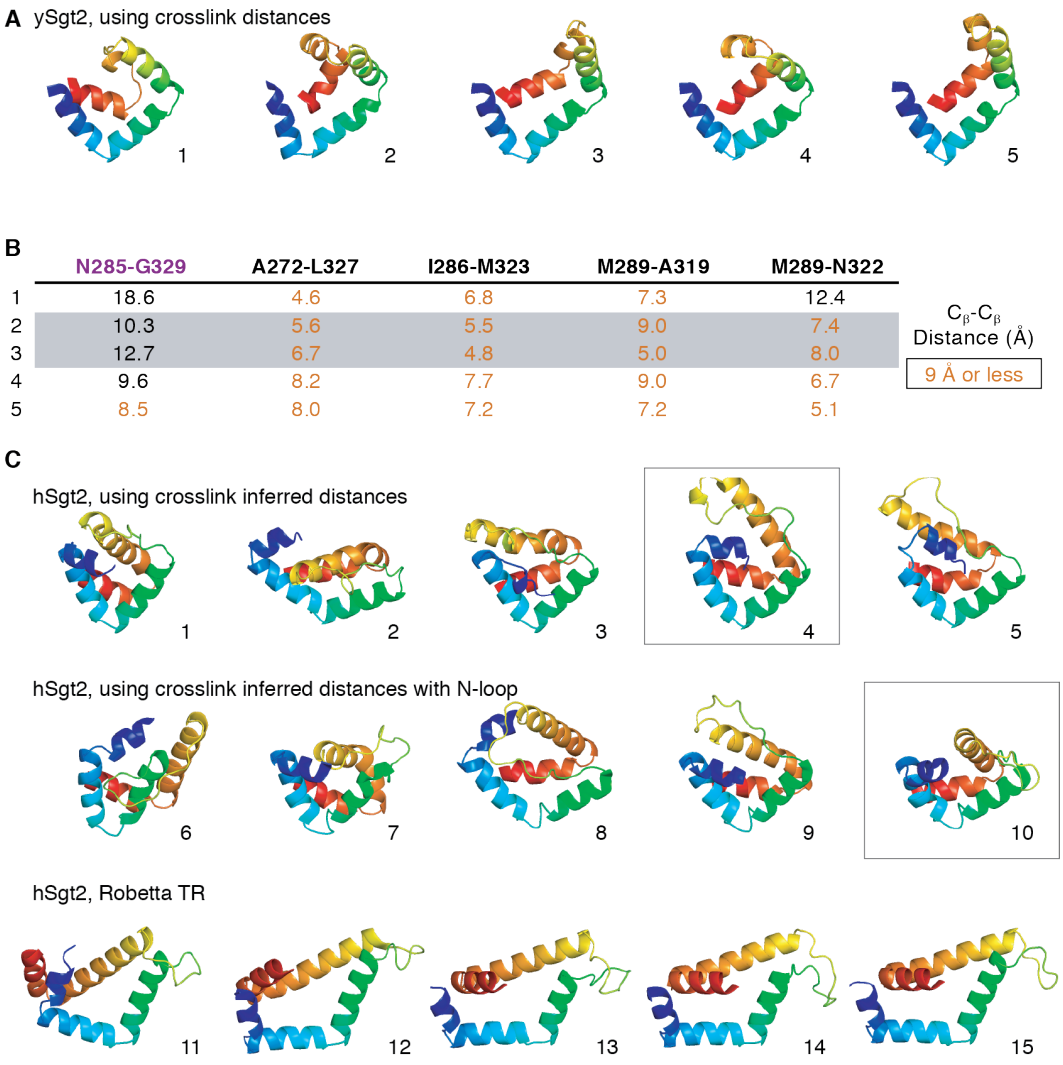



Figure 3.9: **Distance restraints lead to improved ySgt2-C and suggestive hSgt2-C models.**
*A)* Prediction of ySgt2-C using distances from *in vitro* crosslinking. *B)* C$\beta$-C$\beta$ distances between residues probed by *in vitro* disulfide crosslinking for each ySgt2 model. Distances 9Å or less are colored orange. For models where all distances correspond (4 near and 1 far), the row is shaded grey. *C)* Models for hSgt2-C using restraints, adding a N-terminal loop, and via the new Robetta TR method.

corresponds to H2 occupies a position that precludes a clear hydrophobic groove. For ySgt2, the longer N-terminal loop occupies the groove preventing the exposure of hydrophobics to solvent (Fig. 3.5*C, grey*). For hSgt2, the shorter N-terminal loop may not be sufficient to similarly occupy the groove and allowing for the clear hydrophobic hand seen for the ySgt2-C. To correct for this, we replaced the sequence of the N-terminal loop of hSgt2-C with the ySgt2-C loop and ran structure prediction with the pairwise distance restraints. This resulted in a model where the loop occupies the groove and, when pruned away suggests the hydrophobic hand seen in yeast (Fig. 3.9*C, middle boxed*). Of note, we also generated models of hSgt2-C using the most recent Robetta method (transform-restrained) which produces new structures with a groove and similar helical-hand architecture across the board (Fig. 3.9*C, bottom*).

We sought to further test the robustness of our model considering the intrinsic flexibility of Sgt2-C by probing for disulfide bond formation with neighboring residues of one of our crosslinking pairs. While the C$\beta$-C$\beta$ distance puts these adjacent pairs at farther than 9Å, mutating residues to cystines and measuring S-S distances across all possible pairs of rotamers provides a wider interval on possible distances and, therefore, the likelihood a disulfide bond will form (Fig. 3.7*E*). Cysteine mutants were introduced to the residues adjacent to M289 and A319 in ySgt2-TPR-C resulting in four additional pairs: K288C/A319C, M290C/A319C, M289C/P318C, and M289C/L320C. As described previously, these mutants were coexpressed with a client, in this case the cMyc-tag was replaced with an MBP-tag. The MBP-tag on the artificial client allows for tandem amylose- and nickel-affinity chromatography to ensure eluates contained only Sgt2-TPR-C bound to client. Disulfide bond formation was conducted as before and a reductant sensitive band at 7.7kDa is observed for each of these adjacent pairs. While the geometry of each of these C-C pairs might suggest against disulfide bond formation, given the intrinsic flexibility of Sgt2-C, it is not surprising that each of these pairs are able to form disulfide bonds. As before, disulfide bond formation was detected for the M289C/A319C pair. In this new construct, we now see a small amount of disulfide bond formation in the distant N285C/G329C pair, likely an effect of switching to the MBP tag.

**Structural similarity of Sgt2-C domain to STI1-domains**

Attempts to glean functional insight for Sgt2-C from BLAST searches did not reliably return other families or non-Sgt2 homologs making functional comparisons

difficult. A more extensive profile-based search using hidden Markov models from the SMART database (Letunic and Bork, 2017) identified a similarity to domains in the yeast co-chaperone Sti1 (HOP in mammals). First called DP1 and DP2, due to their prevalence of aspartates (D) and prolines (P), these domains have been shown to be required for client-binding by Sti1 (Schmid et al., 2012; Z. Li, Hartl, and Bracher, 2013) and are termed 'STI1'-domains in bioinformatics databases (Letunic and Bork, 2017). In yeast Sti1 and its human homolog HOP (combined will be referred to here as Sti1), each of the two STI1-domains (DP1 and DP2) are preceded by Hsp70/90-binding TPR domains, similar to the domain architecture of Sgt2. Deletion of the second, C-terminal STI1-domain (DP2) from Sti1 *in vivo* is detrimental, impairing native activity of the glucocorticoid receptor (Schmid et al., 2012). *In vitro*, removal of the DP2 domain from Sti1 results in the loss of recruitment of the progesterone receptor to Hsp90 without interfering in Sti1-Hsp90 binding (Nelson, Huffman, and Smith, 2003). These results implicate DP2 in binding of Sti1 clients. In addition, others have noted that, broadly, STI1-domains may present a hydrophobic groove for binding the hydrophobic segments of a client (Schmid et al., 2012; Z. Li, Hartl, and Bracher, 2013). Furthermore, the similar domain organizations (i.e. Sgt2 TPR-C, Sti1 TPR-STI1) and molecular roles could imply an evolutionary relationship between these co-chaperones. Indeed, a multiple sequence alignment of the Sgt2-$C_{cons}$ with several yeast STI1-domains (Fig. 3.10*A*) reveals strong conservation of structural features. H1-H5 of the predicted helical regions in $C_{cons}$ align directly with the structurally determined helices in the DP2 domain of Sti1; this includes complete conservation of helix breaking prolines and close alignment of hydrophobic residues in the amphipathic helices (Schmid et al., 2012).

Based on the domain architecture and homology, a direct comparison between the DP1, DP2, and Sgt2-$C_{cons}$ can be made. A structure of DP2 solved by solution NMR reveals that the five amphipathic helices assemble to form a flexible helical-hand with a hydrophobic groove (Schmid et al., 2012). The lengths of the $\alpha$-helices in this structure concur with those inferred from the alignment in Fig. 3.7*A*. Our molecular model of Sgt2-$C_{cons}$ is strikingly similar to this DP2 structure (Fig. 3.10*B,C*). An overlay of the DP2 structure and our molecular model demonstrates both Sgt2-$C_{cons}$ and DP2 have similar lengths and arrangements of their amphipathic helices (Fig. 3.10*D*). Consistent with our observations of flexibility in Sgt2-$C_{cons}$, Sti1-DP2 generates few long-range NOEs between its helices indicating that Sti1-DP2 also has a flexible architecture (Schmid et al., 2012). We consider this flexibility a feature of

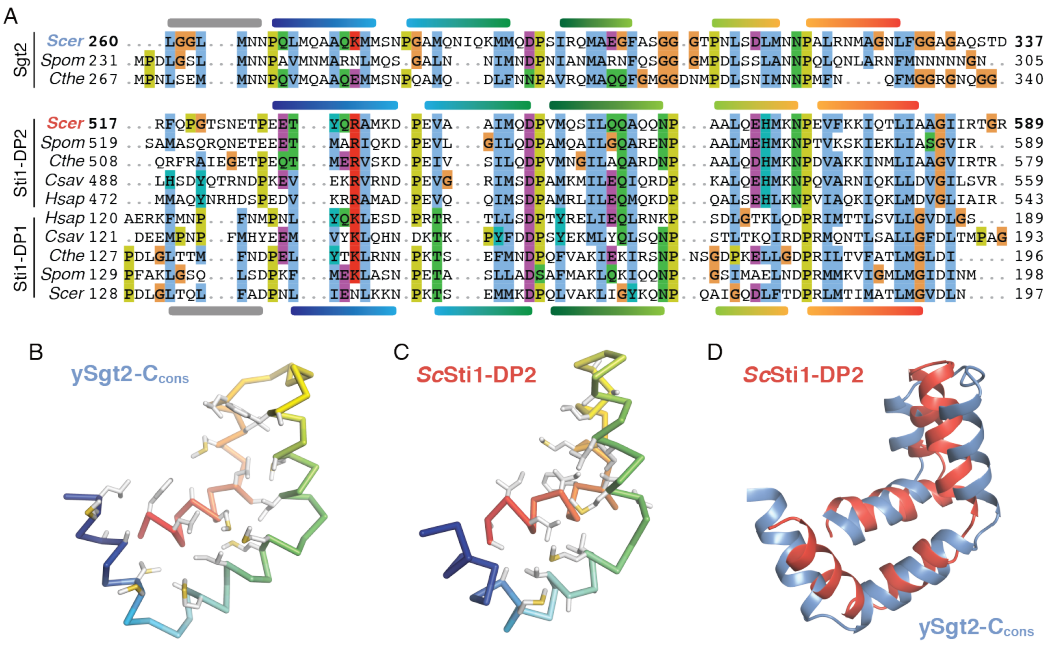



Figure 3.10: **Comparison of STI1-domains and the Sgt2-C$_{cons}$ model.**
*A)* Multiple sequence alignment of Sgt2-C with STI1-domains (DP1, DP2) from Sti1/Hop homologs. Helices are shown based on the Sgt2-C$_{cons}$ model and the *Sc*Sti1-DP1/2 structures. Species for representative sequences are from *S. cerevisiae* (Scer), *S. pombe* (Spom), *C. thermophilum* (Cthe), *C. savignyi* (Csav), and *H. sapiens* (Hsap). *B)* C$\alpha$ ribbon of ScSgt2-C$_{cons}$ color-ramped with large hydrophobic sidechains shown as grey sticks (sulfurs in yellow). *C)* Similar to *(B)* for the solution NMR structure of Sti1-DP2526-582 (PDBID: 2LLW) (Schmid et al., 2012). *D)* Superposition of the Sgt2-C$_{cons}$ (*blue*) and Sti1-DP2526-582 (*red*) drawn as cartoons.

these helical-hands for reversible and specific binding of a variety of clients.

### Binding mode of clients to Sgt2

We examined the Sgt2-C$_{cons}$ surface that putatively interacts with clients by constructing hydrophobic-to-charge residue mutations that are expected to disrupt capture of clients by Sgt2. Similar to the helix mutations in Fig.3.3*E&F*, the capture assay was employed to establish the relative effects of individual mutations. A baseline was established based on the amount of the TA protein Sbh1 captured by wild-type Sgt2-TPR-C. In each experiment, Sbh1 was expressed at the same level; therefore, differences in binding should directly reflect the affinity of Sgt2 mutants for clients. In all cases, groove mutations from hydrophobic to aspartate led to a reduction in client binding (Fig. 3.11). The effects are most dramatic with ySgt2 where each mutant significantly reduced binding by 60% or more (Fig.

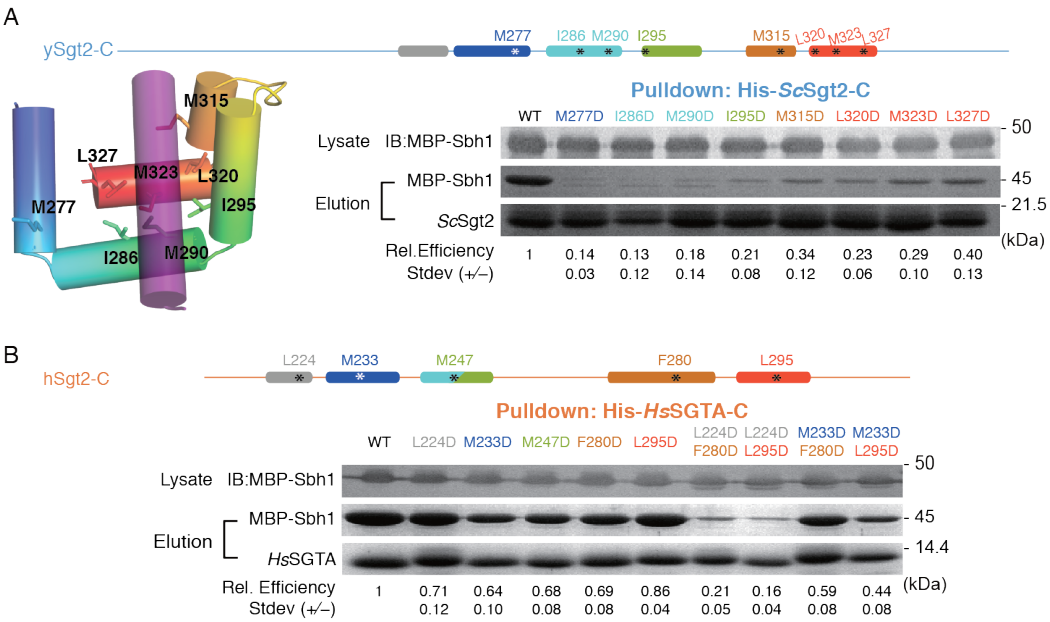



Figure 3.11: **Effects on client binding of charge mutations to the putative hydrophobic groove of Sgt2-C$_{cons}$.**
For these experiments, individual point mutations are introduced into Sgt2-C and tested for their ability to capture Sbh1 quantified as in Fig. 3.3*D*. *A)* For ySgt2-C, a schematic and cartoon model are provided highlighting the helices and sites of individual point mutants both color-ramped for direct comparison. For the cartoon, the docked TMD is shown in purple. Binding of MBP-Sbh1 to his-tagged ySgt2-C and mutants were examined as in Fig. 3.3*E*. Lanes for mutated residues are labeled in the same color as the schematic *B)* Same analysis as in *(A)* for hSgt2-C. In addition, double point mutants are included. Each capture assay was repeated three times.

3.11*A*). While all hSgt2 individual mutants saw a significant loss in binding, the results were more subtle with the strongest a ~36% reduction (M233D, Fig. 3.11*B*). Double mutants were stronger with a significant decrease in binding relative to the individual mutants, more reflective of the individual mutants in ySgt2. As seen before (Fig. 3.3*E&F*), we observe that mutations toward the N-terminus of Sgt2-C have a stronger effect on binding than those later in the sequence, whether single point mutants in the case of ySgt2 or double mutants for hSgt2.

### Sgt2-C domain binds clients with a hydrophobic segment ≥ 11 residues

With a molecular model for ySgt2-C$_{cons}$ and multiple lines of evidence for a hydrophobic groove, we sought to better understand the specific requirements for TMD binding. TMD clients were designed where the overall (sum) and average (mean) hydrophobicity, length, and the distribution of hydrophobic character were varied in

the TMDs. These artificial TMDs, a Leu/Ala helical stretch followed by a Trp, were constructed as C-terminal fusions to the soluble protein BRIL (Fig. 3.12*A*). The total and mean hydrophobicity are controlled by varying the helix-length and the Leu/Ala ratio. For clarity, we define a syntax for the various artificial TMD clients to highlight the various properties under consideration: hydrophobicity, length, and distribution. The generic notation is TMD-length[number of leucines] which is represented, for example, as 18[L6] for a TMD of 18 amino acids containing six leucines.

Our first goal with the artificial clients was to define the minimal length of a TMD to bind to the C-domain. As described earlier in our single point mutation capture assays, captures of his-tagged Sgt2-TPR-C with the various TMD clients were performed. We define a relative binding efficiency as the ratio of captured TMD client by a Sgt2-TPR-C normalized to the ratio of a captured wild-type TA protein by Sgt2-TPR-C. In this case we replaced the TMD in our artificial clients with the native TMD of Bos1 (Bos1$_{TMD}$). The artificial client 18[L13] shows a comparable binding efficiency to Sgt2-TPR-C as that of Bos1$_{TMD}$ (Fig. 3.12*B*). From the helical wheel diagram of the TMD for Bos1, we noted that the hydrophobic residues favored one face of the helix. We explored this 'hydrophobic face' by using model clients that maintained this orientation while shortening the length and maintaining the average hydrophobicity of 18[L13] (Fig. 3.12*B*). Shorter helices of 14 or 11 residues, 14[L10] and 11[L8], also bound with similar affinity to Bos1. Helices shorter than 11 residues, 9[L6] and 7[L5], were not able to bind Sgt2-TPR-C (Fig. 3.12*B*), establishing a minimal length of 11 residues for the helix, consistent with the dimensions of the groove predicted from the structural model (Fig. 3.5).

Since a detected binding event occurs with TMDs of at least 11 amino acids, we decided to probe this limitation further. The dependency of client hydrophobicity was tested by measuring complex formation of Sgt2-TPR-C and artificial TMD clients containing an 11 amino acid TMD with increasing number of leucines (11[Lx]). As shown in Fig. 3.12*C*, increasing the number of leucines monotonically enhances complex formation, echoing previous results (Rao et al., 2016). hSgt2-TPR-C binds to a wider spectrum of hydrophobic clients than ySgt2-TPR-C, which could mean it has a more permissive hydrophobic binding groove, also reflected in the milder impact of Ala replacement and Asp mutations in hSgt2-TPR-C to TMD client binding (Fig. 3.3*F* and Fig. 3.11*B*).

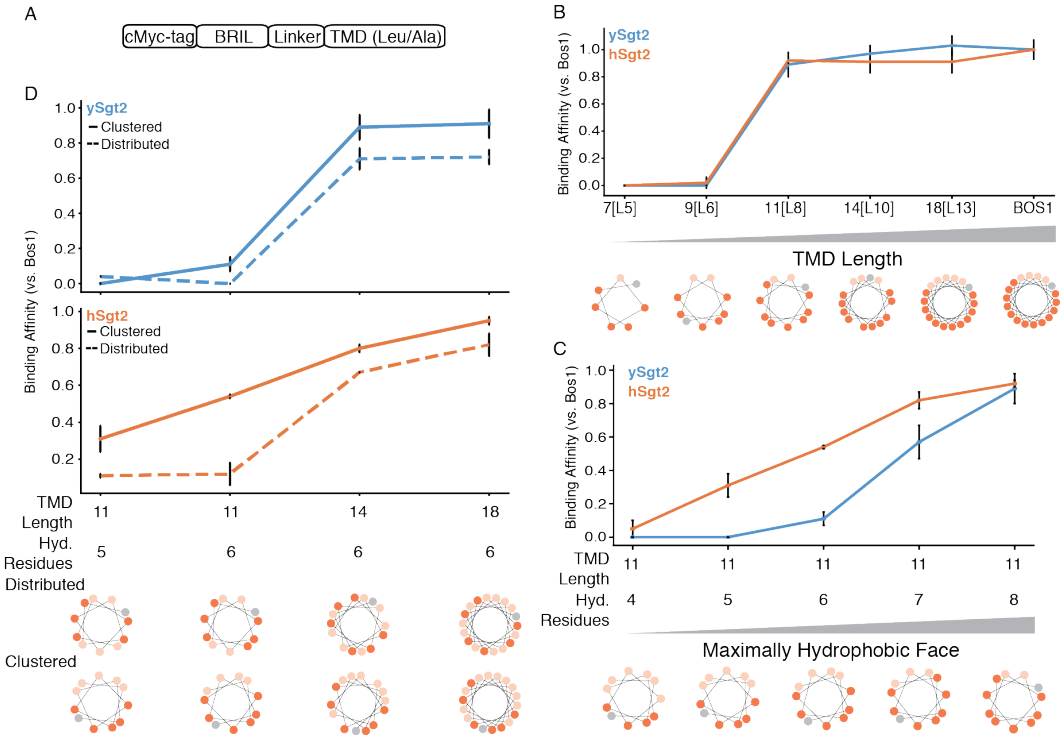



Figure 3.12: **Minimal requirements for client recognition by Sgt2.**
*A)* Schematic of model clients. From capture assays, quantification of complex formation in the eluate is calculated and normalized to that of complexes with $\text{Bos1}_{TMD}$, here defined as relative binding efficiency. *B)* Complex formation of ySgt2 (*blue*) and hSgt2 (*orange*) with the TA protein $\text{Bos1}_{TMD}$ and several artificial clients noted x[Ly], where x denotes the length of the TMD and y denotes the number of leucines in the TMD. The helical wheel diagrams of the TMD of clients here and for subsequent panels with leucines colored in dark orange, alanines colored in pale orange, and tryptophans colored in grey. Each assay was performed four times except for ySgt2-Bos1 and hSgt2-9[L6], which were performed three times. *C)* Complex formation of ySgt2-TPR-C and hSgt2-TPR-C with artificial clients with TMDs of length 11 and increasing numbers of leucine. Capture assays were repeated either two or three times. *D)* Comparison of complex formation of ySgt2-TPR-C and hSgt2-TPR-C with artificial clients of the same lengths and hydrophobicities but differences in the distribution of leucines, i.e. clustered (*solid line*) vs distributed (*dotted line*). Each assay was performed either three or four times.

**Sgt2-C preferentially binds to TMDs with a hydrophobic face**

Next, we address the properties within the TMD of clients responsible for Sgt2 binding. In the case of ySgt2, it has been suggested that the co-chaperone binds to TMDs based on hydrophobicity and helical propensity (Rao et al., 2016). In our system, our artificial TMDs consist of only alanines and leucines which have high helical propensities (Pace and Scholtz, 1998), and despite keeping the helical propensity constant and in a range that favors Sgt2 binding, there is still variation in binding efficiency. For the most part, varying the hydrophobicity of an artificial TMD client acts as expected, the more hydrophobic TMDs bind more efficiently to Sgt2-TPR-C (Fig. 3.12*C*). Our $C_{cons}$ model suggests the hydrophobic groove of ySgt2-C protects a TMD with highly hydrophobic residues clustered to one side (see Fig. 3.5*B*). To test this, various TMD pairs with the same hydrophobicity, but different distributions of hydrophobic residues demonstrates TMD clients with clustered leucines have a higher relative binding efficiency than those with a more uniform distribution (Fig. 3.12*D*). Helical wheel diagrams demonstrate the distribution of hydrophobic residues along the helix (e.g. bottom Fig. 3.12*D*). The clustered leucines in the TMDs create a hydrophobic face which potentially interacts with the hydrophobic groove formed by the Sgt2-$C_{cons}$ region, corresponding to the model in Fig. 3.5*B*.

## 3.3 Discussion

Sgt2, the most upstream component of the GET pathway, plays a critical role in the targeting of TA proteins to their correct membranes along with other roles in maintaining cellular homeostasis. Its importance as the first confirmed selection step of ER versus mitochondrial (Rao et al., 2016) destined TA proteins necessitates a molecular model for client binding. Previous work demonstrated a role for the C-domain of Sgt2 to bind to hydrophobic clients, yet the exact binding domain remained to be determined. Through the combined use of biochemistry, bioinformatics, and computational modeling, we conclusively identify the minimal client-binding domain of Sgt2 and preferences in client binding. Here we present a validated structural model of the Sgt2 C-domain as a methionine-rich helical hand for grasping a hydrophobic helix and to provide a mechanistic explanation for binding a TMD of at least 11 hydrophobic residues.

Identifying the C-domain of Sgt2 as containing a STI1-domain places Sgt2 into a larger context of conserved co-chaperones (Fig. 3.13*A*). In the co-chaperone family, the STI1-domains predominantly follow HSP-binding TPR domains connected by a flexible linker, reminiscent of the domain architecture of Sgt2. As noted above, for

Sti1 these domains are critical for coordinated hand-off between Hsp70 and Hsp90 homologs (Röhl et al., 2015) as well as coordinating the simultaneous binding of two heat shock proteins. Both Sgt2 and the co-chaperone Hip coordinate pairs of TPR and STI1-domains by forming stable dimers via their N-terminal dimerization domains (Coto et al., 2018). With evidence for a direct role of the carboxylate-clamp in the TPR domain of Sgt2 for TA protein-binding now clear (Cho and Shan, 2018), one can speculate that the two TPR domains may facilitate TA protein entry into various pathways that use multiple heat shock proteins.

Computational modeling reveals that a conserved region, sufficient for client binding, forms a five alpha-helical hand which is reminiscent of other proteins involved in membrane protein targeting. Like Sgt2, the signal recognition particle (SRP) contains a methionine-rich domain that binds signal sequences and TMDs. While the helical order is inverted, again five amphipathic helices form a hydrophobic groove that cradles the client signal peptide (Voorhees and Hegde, 2015) (Fig. 3.13*B*). Here once more, the domain has been observed to be flexible in the absence of client (Keenan et al., 1998; Clemons et al., 1999) and, in the resting state, occupied by a region that includes a helix which must be displaced (Voorhees and Hegde, 2015). Another helical-hand example recently shown to be involved in TA IMP targeting is calmodulin where a crystal structure reveals two helical hands coordinating to clasp a TMD at either end (Fig. 3.13*B*) (Tidow and Nissen, 2013). Considering an average TMD of 18-20 amino acids (to span a ~40Å bilayer), each half of calmodulin interacts with about 10 amino acids. This is in close correspondence to the demonstrated minimal 11 amino acids for a TMD client to bind to the monomeric Sgt2-TPR-C. In the context of the full-length Sgt2, one can speculate that the Sgt2 dimer may utilize both C-domains to bind to a full TMD, similar to calmodulin. Cooperation of the two Sgt2 C-domains in client-binding could elicit conformational changes in the complex that could be recognized by downstream factors, such as additional interactions that increase the affinity to Get5/Ubl4A.

Intriguingly, Sgt2-TPR-C preferentially binds to artificial clients with clustered leucines. The hydrophobic groove presented in the computational model provides an attractive explanation for this preference. In order to bind to the hydrophobic groove, a client buries a portion of its TMD in the groove leaving the other face exposed. Clustering the most hydrophobic residues contributes to the hydrophobic effect driving binding efficiency and protecting them from the aqueous environment. Indeed, when focusing on Sgt2's role in TA IMP targeting, GET pathway clients

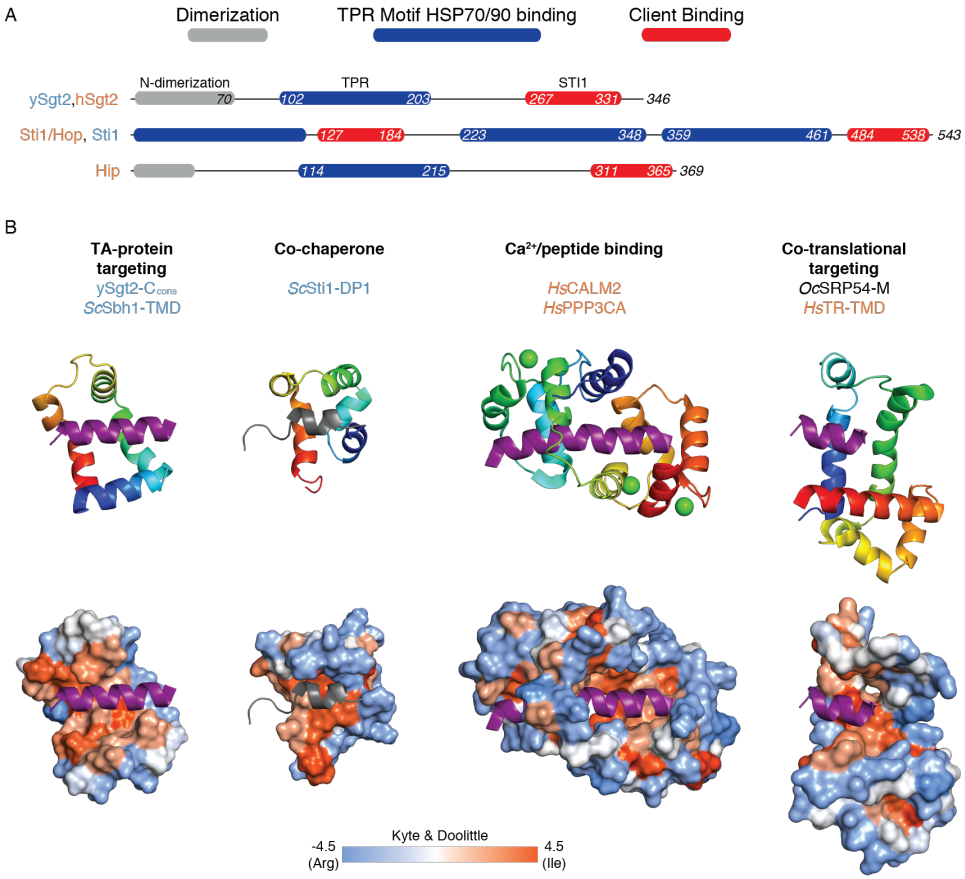



Figure 3.13: **Various domain structures of STI1-domains and other helical-hand containing proteins.**
*A)* The domain architectures of proteins with a STI1-domain were obtained initially from InterPro (Consortium, 2020) and then adjusted as discussed in the text. Each domain within a protein is colored relative to the key. *B)* Structural comparison of various hydrophobic-binding helical-hand protein complexes. For each figure only relevant domains are included. Upper row, color-ramped cartoon representation with bound helices in purple. Lower row, accessible surface of each protein colored by hydrophobicity again with docked helical clients in purple. In order, the predicted complex of ySgt2-$C_{cons}$ and *Sc*Sbh2-TMD, DP1 domain from yeast Sti1 with N-terminus containing H0 in grey (*Sc*Sti1-DP1) (PDBID: 2LLV), human calmodulin (*Hs*CALM2) bound to a hydrophobic domain of calcineurin (*Hs*PPP3CA) (PDBID: 2JZI), and M domain of SRP54 from *Oryctolagus cuniculus* (*Oc*SRP54-M) and the signal sequence of human transferrin receptor (*Hs*TR-TMD) (PDBID: 3JAJ).

have been suggested to be more hydrophobic TMDs than EMC clients (Guna, Volkmar, et al., 2018). Of these, for the most hydrophobic clients, like Bos1, residues on both sides of the TMD could be protected by a pair of C-domains. Alternatively, the unstructured N-terminal loop through H0 could act as a lid surrounding the circumference of the client's TMD. Unstructured regions participating in client binding as well as capping a hydrophobic groove have both suggested in the context of other domains, e.g. with Get3 (Guna and Hegde, 2018). The role for this clustering of hydrophobic residues in client recognition and targeting merits further investigation.

What is the benefit of the flexible helical-hand structure for hydrophobic helix binding? While it remains an open question, it is notable that evolution has settled on similar simple solutions to the complex problem of specific but temporary binding of hydrophobic helices. For all of the domains mentioned above, the flexible helical-hands provide an extensive hydrophobic surface to capture a client-helix—driven by the hydrophobic effect. Typically, such extensive interfaces are between pairs of pre-ordered surfaces resulting in high affinities and slow off rates. These helical hands are required to only engage temporarily, therefore the flexibility offsets the favorable free energy of binding by charging an additional entropic cost for ordering a flexible structure in the client-bound complex. The benefit for clients is a favorable transfer to downstream components in the GET pathway as seen for ySgt2 (Cho and Shan, 2018) and hSgt2 (Shao, Rodrigo-Brenni, et al., 2017). The demonstration that TA protein transfer from hSgt2 to Get3 is twice as fast as disassociation from hSgt2 into solution, perhaps interaction with Get3 leads to conformational changes that further favor release (Shao, Rodrigo-Brenni, et al., 2017). While hSgt2 and ySgt2 share many properties, there are a number of differences between the two homologs that may explain the different biochemical behavior. For the $C_{cons}$-domains, hSgt2 appears to be more ordered in the absence of client as the peaks in its NMR spectra are broader (Fig. 3.2*B,C*). Comparing the domains at the sequence level, while the high glutamine content in the C-domain is conserved it is higher in hSgt2 (8.8% versus 15.2%). The additional glutamines are concentrated in the predicted longer H4 helix (Fig. 3.1*A*). The linker to the TPR domain is shorter compared to ySgt2 while the loop between H3 and H4 is longer. Do these differences reflect different roles? As noted, in every case the threshold for hydrophobicity of client-binding is lower for hSgt2 than ySgt2 (Fig. 3.3*E, 6, & 7*) implying that the mammalian protein is more permissive in client binding. The two C-domains have similar hydrophobicity, so this difference in binding might be due to a lower entropic cost paid by having the hSgt2 C-domain more ordered in the absence of client or the lack

of an unstructured N-terminal loop.

## 3.4 Conclusions

Identification of Sgt2-$C_{cons}$ as a STI1-domain reveals that the database definition for the domain is incomplete. Chapter 4 discusses the development of a more inclusive HMM search definition for STI1-domains and other newly identified STI1-domain containing proteins. Modeling algorithms predict the C-domain of Sgt2 from homologs, including the *G. intestinalis*, to contain the same five $\alpha$-helicial hand as validated here for *Sc*Sgt2-$C_{cons}$. With a similar structure, it is reasonable to surmise protozoan Sgt2 have a similar mechanism and preference as the Opisthokont counterparts. The similarly predicted structure also demonstrates a presence of a STI1-domain in the superfamily Excavata, disclosing the breadth of STI1-domains throughout eurkaryotes.

The targeting of TA proteins presents an intriguing and enigmatic problem for understanding the biogenesis of IMPs. How subtle differences in each client modulates the interplay of hand-offs that direct these proteins to the correct membrane remains to be understood. In this study, we focused on a central player, Sgt2 and its client-binding domain. Through biochemistry and computational analysis, we provided a structural model that adds more clarity to client binding both within and outside of the GET pathway.

## 3.5 Material and Methods

*Plasmid constructs*

MBP-Sbh1 full length, $ySgt_{295-346}$ (ySgt2-TPR-C), $ySgt2_{222-346}$ (ySgt2-C), $ySgt2_{260-327}$ (ySgt2-$C_{cons}$), $ySgt2_{266-327}$ (ySgt2-$\Delta$H0), $hSgt2_{87-313}$ (hSgt2-TPR-C), $hSgt2_{213-313}$ (hSgt2-C), $hSgt2_{219-300}$ (hSgt2-$C_{cons}$), and $hSgt2_{228-300}$ (hSgt2-$\Delta$H0) were prepared as previously described (Chartron, Gonzalez, and Clemons, 2011; Suloway, Rome, and Clemons, 2011). Genes of ySgt2 or hSgt2 variants were amplified from constructed plasmids and then ligated into an pET33b-derived vector with a 17 residue N-terminal hexa-histidine tag and a tobacco etch virus (TEV) protease site. Single or multiple mutations on Sgt2 were constructed by site-direct mutagenesis. Artificial clients were constructed in a pACYC-Duet plasmid with a N-terminal cMyc-tag, BRIL fusion protein (R. Chu et al., 2002), GSS linker, and a hydrophobic C-terminal tail consisting of leucines and alanines and ending with a tryptophan.

*Protein expression and purification*

All proteins were expressed in *Escherichia coli* NiCo21 (DE3) cells (New England BioLabs). To co-express multiple proteins, constructed plasmids were co-transformed as described (Suloway, Rome, and Clemons, 2011). Protein expression was induced by 0.3 mM IPTG at $OD_{600\sim0.7}$ and harvested after 3 hours at 37°C. For structural analysis, cells were lysed through an M-110L Microfludizer Processor (Microfluidics) in lysis buffer (50 mM Tris, 300 mM NaCl, 25 mM imidazole supplemented with benzamidine, phenylmethylsulfonyl fluoride (PMSF), and 10 mM $\beta$-mercaptoethanol ($\beta$ME), pH 7.5). For capture assays, cells were lysed by freeze-thawing 3 times with 0.1 mg/mL lysozyme. To generate endogenous proteolytic products of ySgt2-TPR-C for MS analysis, PMSF and benzamidine were excluded from the lysis buffer. His-tagged Sgt2 and his-tagged Sgt2/TA complexes were separated from the lysate by batch incubation with Ni-NTA resin at 4°C for 1hr. The resin was washed with 20 mM Tris, 150 mM NaCl, 25 mM imidazole, 10 mM $\beta$ME, pH 7.5. The complexes of interest were eluted in 20 mM Tris, 150 mM NaCl, 300 mM imidazole, 10 mM $\beta$ME, pH 7.5.

For structural analysis, the affinity tag was removed from complexes collected after the nickel elution by an overnight TEV digestion against lysis buffer followed by size-exclusion chromatography using a HiLoad 16/60 Superdex 75 prep grade column (GE Healthcare).

Measurement of Sgt2 protein concentration was carried out using the bicinchoninic acid (BCA) assay with bovine serum albumin (BSA) as standard (Pierce Chemical Co.). Samples for NMR and CD analyses were concentrated to 10-15mg/mL for storage at 80°C before experiments.

For the *in vitro* transfer assay, plasmids encoding for the full-length ySgt2 cysteine mutants were transformed into BL21 Star cells (Invitrogen). Cells were grown in 2x yeast-tryptone (2xYT) media and induced with 0.1mM IPTG at an $OD_{600}$ of 0.6 then harvested after 3 hours at 30°C by centrifugation. Cells were lysed in 50mM Tris pH 8.0, 500mM NaCl, 10% glycerol, and 1x BugBuster (Millipore Sigma), supplemented with protease inhibitors (4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (Roche), benzamidine, and $\beta$ME). Full-length his-tagged ySgt2 and cysteine mutants were separated from the lysate by batch incubation with Ni-NTA resin (Qiagen) at 4°C for 1 hour. The resin was washed with 50mM Tris pH 8.0, 500mM NaCl, 10% glycerol, and 25mM imidazole and then the protein was eluted in 50mM Tris pH 8.0, 500mM NaCl, 10% glycerol, and 300mM imidazole. For storage, protein was dialyzed in 25mM K-HEPES pH 7.5, 150mM KOAc, and

20% glycerol at 4°C and then flash frozen in liquid nitrogen. Purified Bos1 with p-benzoyl-l-phenylalanine (BPA) labeled at residue 230 ($Bos1_{BPA}$) and yeast Ssa1 were gifts from the lab of Shu-ou Shan (Caltech).

*NMR Spectroscopy*

$^{15}$N-labeled proteins were generated from cells grown in auto-induction minimal media as described (Studier, 2005) and purified in 20 mM phosphate buffer, pH 6.0 (for ySgt2-C, 10mM Tris, 100mM NaCl, pH 7.5). The NMR measurements of $^{15}$N-labeled Sgt2-C proteins (~0.3-0.5 mM) were collected using a Varian INOVA 600 MHz spectrometer at either 25°C (ySgt2-C) or 35°C (hSgt2-C) with a triple resonance probe and processed with TopSpin™ 3.2 (Bruker Co.).

*CD Spectroscopy*

The CD spectra were recorded at 24°C with an Aviv 202 spectropolarimeter using a 1mm path length cuvette with 10$\mu$M protein in 20mM phosphate buffer, pH 7.0. The CD spectrum of each sample was recorded as the average over three scans from 190/195 to 250nm in 1nm steps. Each spectrum was then decomposed into its most probable secondary structure elements using BeStSel (Micsonai et al., 2015).

*Glu-C digestion of the double cysteine mutants on ySgt2-C*

Complexes of the co-expressed wild type or double cysteine mutated His-ySgt2-TPR-C and the artificial client, 11[L8], with either a cMyc or MBP tag were purified as the other His-Sgt2 complexes described above or initially purified via amylose affinity chromatography before nickel chromatography explained earlier. The protein complexes were mixed with 0.2mM CuSO4 and 0.4mM 1,10-phenanthroline at 24°C for 20 min followed by 50 mM N-ethyl maleimide for 15 min. Sequencing-grade Glu-C protease (Sigma) was mixed with the protein samples at an approximate ratio of 1:30 and the digestion was conducted at 37°C for 22 hours. Digested samples were mixed with either non-reducing or reducing SDS-sample buffer, resolved via SDS-PAGE using Mini-Protean® Tris-Tricine Precast Gels (10-20%, Bio-Rad), and visualized using Coomassie Blue staining.

*In vitro transfer assay of Bos1 from Ssa1 to ySgt2*

The *in vitro* transfer assays were performed as in (Shao, Rodrigo-Brenni, et al., 2017; Chio, Chung, et al., 2019). Specifically, 39$\mu$M $Bos1_{BPA}$ (50mM HEPES, 300mM NaCl, 0.05% LDAO, 20% glycerol) was diluted to a final concentration of 0.1$\mu$M and added to 4$\mu$M Ssa1 supplemented with 2mM ATP (25mM HEPES

pH7.5, 150mM KOAc). After one minute, 0.3$\mu$M of full-length ySgt2 or mutant was added to the reaction. Samples were flash frozen after one minute and placed under a 365nm UV lamp for 2 hours on dry ice to allow for BPA crosslinking.

*Protein immunoblotting and detection*

For western blots, protein samples were resolved via SDS-PAGE and then transferred onto nitrocellulose membranes by the Trans-Blot® Turbo™ Transfer System (Bio-Rad). Membranes were blocked in 5% non-fat dry milk and hybridized with antibodies in TBST buffer (50mM Tris-HCl pH 7.4, 150mM NaCl, 0.1% Tween 20) for 1 hour of each step at 24°C. The primary antibodies were used at the following dilutions: 1:1000 anti-penta-His mouse monoclonal (Qiagen), 1:5000 anti-cMyc mouse monoclonal (Sigma), and a 1:3000 anti-Strep II rabbit polyclonal (Abcam). A secondary antibody conjugated either to alkaline phosphatase (Rockland, 1:8000) or a 800nm fluorophore was employed, and the blotting signals were chemically visualized with either the nitro-blue tetrazolium/5-bromo-4-chloro-3'-indolyphosphate (NBT/BCIP) chromogenic assay (Sigma) or infrared scanner. All blots were photographed and quantified by image densitometry using ImageJ (Schneider, Rasband, and Eliceiri, 2012) or ImageStudioLite (LI-COR Biosciences).

*Quantification of Sgt2-TA complex formation*

The densitometric analysis of MBP-Sbh1 capture by His-Sgt2-TPR-C quantified the intensity of the corresponding protein bands on a Coomassie Blue G-250 stained gel. The quantified signal ratios of MBP-Sbh1/His-Sgt2-TPR-C are normalized to the ratio obtained from the wild-type (WT). Expression level of MBP-Sbh1 was confirmed by immunoblotting the MBP signal in cell lysate. Average ratios and standard deviations were obtained from 3-4 independent experiments.

In artificial client experiments, both his-tagged Sgt2-TPR-C and cMyc-tagged artificial clients were quantified via immunoblotting signals. The complex efficiency of Sgt2-TPR-C with various clients was obtained by

$$E_{complex} = E_{TMD}/T_{TMD}1/E_{capture} \tag{3.1}$$

where $E_{TMD}$ is the signal intensity of an eluted client representing the amount of client co-purified with Sgt2-TPR-C and $T_{TMD}$ is the signal intensity of a client in total lysate which corresponds to the expression yield of that client. Identical volumes of elution and total lysate from different clients experiments were analyzed and quantified. In order to correct for possible variation the amount of Sgt2-TPR-C available for complex formation, $E_{capture}$ represents the relative amount of

Sgt2-TPR-C present in the elution ($E_{Sgt2}$) compared to a pure Sgt2-TPR-C standard ($E_{purified,Sgt2}$).

$$E_{capture} = E_{Sgt2}/E_{purified,Sgt2} \tag{3.2}$$

Each $E_{TMD}$ and $T_{TMD}$ value was obtained by blotting both simultaneously, i.e. adjacently on the same blotting paper. To facilitate comparison between clients, the Sgt2-TPR-C/TA protein complex efficiency $E_{complex,TMD}$ is normalized by Sgt2-TPR-C/Bos1 complex efficiency $E_{complex,Bos1}$.

$$\%Complex = E_{complex,TMD}/E_{complex,Bos1}100 \tag{3.3}$$

*Sequence alignments*

An alignment of Sgt2-C domains was carried out as follows: all sequences with an annotated N-terminal Sgt2/A dimerization domain (PF16546 (El-Gebali et al., 2018), at least one TPR hit (PF00515.27, PF13176.5, PF07719.16, PF13176.5, PF13181.5), and at least 50 residues following the TPR domain were considered family members. Putative C-domains were inferred as all residues following the TPR domain, filtered at 90% sequence identity using CD-HIT (W. Li and Godzik, 2006), and then aligned using MAFFT G-INS-i (Katoh and Standley, 2013). Other attempts with a smaller set (therefore more divergent) of sequences results in an ambiguity in the relative register of H0, H1, H2, and H3 when comparing Sgt2 with SGTA.

Alignments of Sti1 (DP1/DP2) and STI1-domains were created by pulling all unique domain structures with annotated STI1-domains from Uniprot. Where present, the human homolog was selected and then aligned with PROMALS3D (Pei, Kim, and Grishin, 2008). PROMALS3D provides a way of integrating a variety of costs into the alignment procedure, including 3D structure, secondary structure predictions, and known homologous positions. All alignments were visualized using Jalview (Waterhouse et al., 2009).

*Molecular modeling*

Putative models for ySgt2-C were generated with I-TASSER, PCONS, Quark, Robetta (*ab initio* and transform-restrained modes), Phyre2, and RaptorX via their respective web servers (D. Xu and Y. Zhang, 2012; Yang, Yan, et al., 2015; Wallner and Elofsson, 2006; Bradley, Misura, and Baker, 2005; Yang, Anishchenko, et al., 2020a). The highest scoring model from Quark was then chosen to identify putative TA protein binding sites by rigid-body docking of various transmembrane domains

modelled as $\alpha$-helices (3D-HM (Reißer et al., 2014) into the ySgt2-$C_{cons}$ through the Zdock web server (Pierce et al., 2014). Pairwise distances were calculated between C$\beta$ atoms (the closer C$\alpha$ proton on glycine) using mdtraj (McGibbon et al., 2015) . Based on our disulfide crosslinks, new models were predicted using Robetta in *ab initio* mode specifying C$\beta$-C$\beta$ atom distance constraints bounded between 0 and 9 Å.

For hSgt2, using the same set of structure prediction servers above, we were only able to produce a clear structural model using the Robetta transform-restrained mode. We were also unable to generate a reliable model by directly using the ySgt2-C model as a template (Webb and Sali, 2016). To crosslink distance data from ySgt2 as restraints for hSgt2, pair positions were transferred from one protein to the other via an alignment of Sgt2-C domains (excerpt in Fig. 3.1*A*) and ran Robetta *ab initio*. Also, we grafted the N-terminal loop of ySgt2-C on hSgt2-C with the same set of restraints. Images were rendered using PyMOL 2.3 (www.pymol.org).

### 3.6 Acknowledgements

We thank D. G. VanderVelde for assistance with NMR data collection; S. Mayo for providing computing resources; S. Shan, H. J. Cho, Y. Liu, and members of the Clemons lab for support and discussion. We thank J. Mock and A. M. Thinn for comments on the manuscript. This work was supported by the National Institutes of Health (NIH) grants GM105385 and GM097572 (to WMC), NIH/National Research Service Award Training Grant GM07616 (to SMS and MYF), and a National Science Foundation Graduate Research fellowship Grant 1144469 (to SMS).

*Chapter 4*

# STI1-DOMAINS ARE AN ALPHA-HELICAL PROTEIN FOLD

*Adapted from:*

Fry, Michelle Y, Shyam M Saladi, and William M Clemons Jr (2021). "The STI1-domain is a flexible alpha-helical fold with a hydrophobic groove". In: *Protein Science* 30.4, pp. 882–898. DOI: 10.1002/pro.4049.

M.Y. Fry contributed to identifying STI1-domain containing proteins, the analysis on the amino acid distribution, and interpreting the computational models.

## Abstract

STI1-domains are present in a variety of co-chaperone proteins and are required for the transfer of hydrophobic clients in various cellular processes. The domains were first identified in the yeast Sti1 protein where they were referred to as DP1 and DP2. Based on hidden Markov model searches, this domain had previously been found in other proteins including the mammalian co-chaperone SGTA, the DNA damage response protein Rad23, and the chloroplast import protein Tic40. Here, we refine the domain definition and carry out structure-based sequence alignment of STI1-domains showing conservation of five amphipathic helices. Upon examinations of these identified domains, we identify a preceding helix 0 and unifying sequence properties, determine new molecular models, and recognize that STI1-domains nearly always occur in pairs. The similarity at the sequence, structure, and molecular levels likely supports a unified functional role.

## 4.1 Introduction

Hydrophobic stretches that are exposed during protein biosynthesis can aggregate which poses a risk to cellular homeostasis. To avoid this, cells have evolved proteins that bind to and protect hydrophobic segments. Several protein domains exist to assist these proteins and are found in protein families that occur broadly in eukaryotes, here we focus on the STI1-domain. Named for the yeast protein Sti1 (STress Inducible 1) where they were first identified, STI1-domains are referred to as heat-shock chaperonin-binding domains in databases (Letunic and Bork, 2017). Solved structures from Sti1 revealed an alpha-helical domain with five amphipathic helices that present a hydrophobic groove, an likely binding site for hydrophobic segments of a client (Schmid et al., 2012; Z. Li, Hartl, and Bracher, 2013) (Fig. 4.1 *A&B*). In addition to Sti1 homologs, a number of protein families were bioinformatically identified to contain this domain including the co-chaperones HIP (HSP interacting protein) and SGTA (Small Glutamine-rich TPR-containing protein A), the DNA damage response protein Rad23 (RADiation sensitive 23), yeast UBL-UBA family member Dsk2 (Dominant Suppressor of Kar1 2), human KPC2 (Kip1 ubiquitylation-Promoting Complex 2), human ubiquilins (UBQLNs) 1-4, and the plant chloroplast import protein Tic40 (Translocon at the of the Inner envelope membrane of Chloroplasts 40) (Zientara-Rytter and Subramani, 2019; Howe et al., 2019). The identified STI1-domain containing proteins can be broadly classified into two categories: either co-chaperones (homologs of Sti1, HIP, and SGTA and the unique plant Tic40) or adaptors of the ubiquitin proteasome system (AUPS) (homologs of the mammalian Rad23, UBQLNs, KPC2, and yeast Dsk2).

STI1-domain containing proteins have been identified only in eukaryotes. HOP and Rad23 are found throughout eukaryotes including in some protists. SGTA homologs are similarly prevalent, although not readily identifiable in Viridiplantae (Howe et al., 2019). UBQLNs, the closest mammalian relatives to yeast Dsk2, can be found across multicellular eukaryotes (Zientara-Rytter and Subramani, 2019; Howe et al., 2019). Tic40 is restricted to Archaeplastida (algae and land plants). It is likely that more distant homologs for each of these STI1-domain containing proteins exist in taxa that currently seem excluded (Weisman, Murray, and Eddy, 2020).

We consider the co-chaperones first. Broadly, co-chaperones are binding partners for Hsp90 or Hsp70 that enhance the function of these chaperones with a subset directly involved in binding to clients (Caplan, 2003). A first example is the mammalian Sti1-homolog HOP (Hsp70/Hsp90 organizing protein) that coordinates the

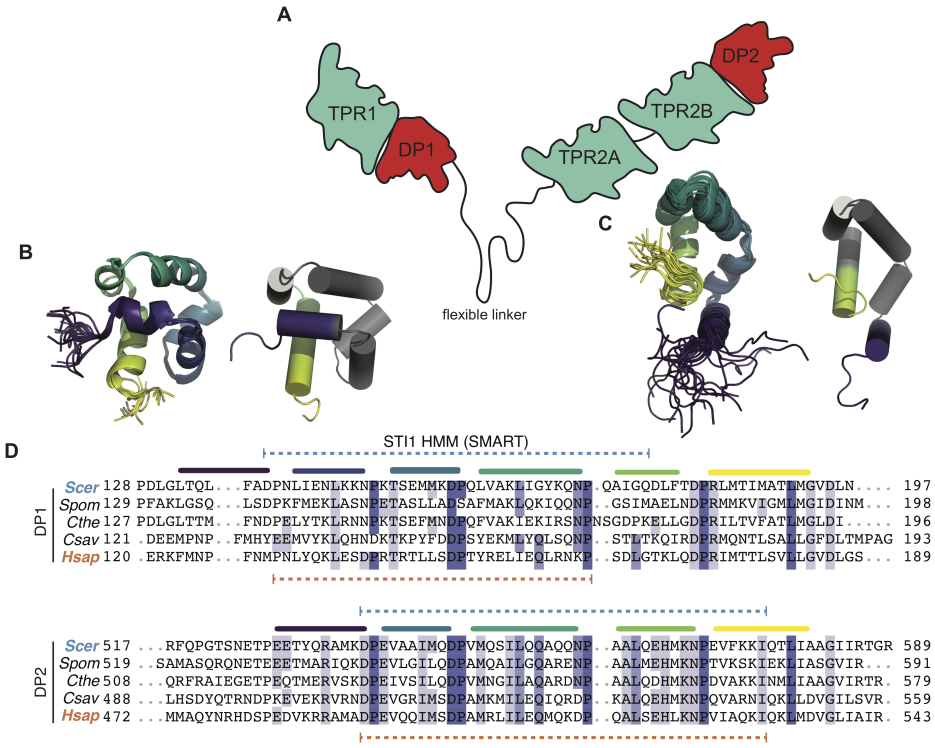



Figure 4.1: **Sequence alignments and structural characterizations of the STI1-domains in HOP.**
*A)* A cartoon representation of the HOP yeast homolog, Sti1, the two STI1-domains (DP2 & DP1) colored in red and three TPR domains in cyan. *B & C)* Ribbon and cylinder cartoon depictions of the structures of *(B)* DP1 (PDBID:2LLV) and *(C)* DP2 (PDBID: 2LLW), colored using Viridis from N- to C-termini (purple to yellow). Grey cylinders indicate residues covered by the SMART STI1 HMM. *D)* Sequence alignments of DP1 (*top*) and DP2 (*bottom*) from HOP homologs from *Saccharomyces cerevisiae* (*Scer*), *Schizosaccharomyces pombe* (*Spom*), *Chaetomium thermophilum* (*Cthe*), *Ciona savignyi* (*Csav*), and *Homo sapiens* (*Hsap*). Predicted helices are annotated above with the SMART STI1-HMM definition marked by dotted lines in blue (*yeast*) and orange (*human*). Names for this and subsequent figures are colored in blue for *S. cerevisiae* and orange for *H. sapiens*.

essential transfer of clients between the chaperones Hsp70 and Hsp90. The abundant chaperone Hsp90 and its homologs are involved in multiple cellular pathways and many clients are first loaded from homologs of the chaperone Hsp70 (Prodromou, 2016). The domain organization of HOP homologs includes two STI1-domains, originally named DP1 and DP2 due to a repeated DP motif (Chen and Smith, 1998; Prapapanich, Chen, and Smith, 1998), that are preceded by Hsp70/90-binding tetratricopeptide-repeat (TPR) domains (Onuoha et al., 2008; Nelson, Huffman, and Smith, 2003; Schmid et al., 2012; Kajander et al., 2009) (Fig. 4.1 *A*). In yeast,

in vivo deletion of the second STI1-domain in Sti1 (DP2) is detrimental, impairing native activity of the glucocorticoid receptor (Schmid et al., 2012). In vitro, removal of DP2 results in the loss of the transfer of the progesterone receptor to Hsp90 (Nelson, Huffman, and Smith, 2003). These results implicate DP2 in client interaction. Besides simply bridging client transfer between the two HSPs, HOP has also been implicated in prion-protein binding (Martins et al., 1997; Zanata et al., 2002; Fonseca et al., 2021). Like HOP, HIP also aids in the transfer of client from Hsp70 to Hsp90 through direct interaction with Hsp70 (Prapapanich, Chen, Toran, et al., 1996; Lässle et al., 1997; Reidy et al., 2018).

For the mammalian SGTA and its homologs, including yeast Sgt2, the STI1-domain directly binds to clients (Lin et al., 2021). SGTA homologs play a number of roles, with the best characterized involving the targeting of tail-anchored (TA) proteins to the ER membrane as a member of the Guided Entry of TA protein (GET) pathway (F. Wang, Brown, et al., 2010; Chartron, Gonzalez, and Clemons, 2011; Chartron, VanderVelde, and Clemons, 2012; Rao et al., 2016; Simon et al., 2013). SGTA has been suggested to play a role in the degradation of mislocalized membrane proteins in conjunction with the protein Bag6 (Hessa, Sharma, et al., 2011; Y. Xu, Cai, et al., 2012; Wunderley et al., 2014; Leznicki et al., 2015; Mock et al., 2015; Rodrigo-Brenni, Gutierrez, and Hegde, 2014). Additionally, SGTA is involved with disease, including polyomavirus infection (Dupzyk et al., 2017), neurodegenerative disease (Kiktev et al., 2012; Long et al., 2012), hormone-regulated carcinogenesis (Trotta et al., 2013; Buchanan et al., 2007), and myogenesis (H. Wang, Q. Zhang, and D. Zhu, 2003), although the underlying molecular mechanisms are still unclear.

The final member of the co-chaperone family is the chloroplast protein Tic40. In Arabidopsis thaliana, Tic40 is found in the inner membrane of the chloroplast and has been suggested to be a co-chaperone for the stroma chaperone complex for protein transport across the inner membrane (Chou, Fitzpatrick, et al., 2003; Bédard, Kubis, et al., 2007). Deleting Tic40 leads to a decrease in the import of precursors into the chloroplast (Kovacheva et al., 2004). Where studied, STI1-domains in each co-chaperone interact with clients (Nelson, Huffman, and Smith, 2003; Z. Li, Hartl, and Bracher, 2013; Mock et al., 2015; Chartron, Gonzalez, and Clemons, 2011; F. Wang, Brown, et al., 2010; Lin et al., 2021; Ko et al., 2004; Chang, Nathan, and Lindquist, 1997; Schmid et al., 2012; Fan et al., 2002), thus due to the role of the STI1 motif in other co-chaperones, the STI1-domain in Tic40 may also interact with clients being transported across the outer chloroplast membrane and into the stroma

(Bédard, Kubis, et al., 2007; Chou, C.-C. Chu, et al., 2006; Chou, Fitzpatrick, et al., 2003; Kovacheva et al., 2004).

The second group of STI1-domain containing proteins, the AUPS, primarily deliver clients to the proteasome. They contain an N-terminal ubiquitin-like (UBL) domain and a C-terminal ubiquitin-associated domain (UBA). One of the earliest identified UBL-containing protein in yeast was Rad23; this protein shuttles some proteins to the proteasome and also protects some clients from degradation by preventing ubiquitin elongation (Watkins et al., 1993; Wade and Auble, 2014; Fishbain et al., 2011; Elsasser et al., 2002; Saeki et al., 2002). Rad23 has also been implicated in nucleotide excision repair as a complex with Rad4 that recognizes DNA damage (Dantuma, Heinen, and Hoogstraten, 2009; Zientara-Rytter and Subramani, 2019). Likewise the fungal Dsk2 acts as an adaptor to target ubiquitin-labeled proteins to the proteasome for degradation Lowe:2006ir. The UBA domain of Dsk2 recognizes the poly-ubiquitin tail on proteins and the UBL domain interacts with the proteasome regulatory subunit, Rpn1. Another UBL-UBA containing adapter protein is KPC2 (Hara et al., 2005), a subunit of the KPC E3 ligase complex where it acts as an adapter for p27 ubiquitination in the G1 phase of the cell cycle (Kotoshiba et al., 2005).

The closest mammalian homologs to Dsk2 are the four ubiquilins, UBQLN-1 to -4 (Zientara-Rytter and Subramani, 2019). While UBQLN-1 is universally expressed and UBQLN-2 & -4 are expressed in most tissues, UBQLN-3 is expressed only in the testes (Conklin et al., 2000; Yuan et al., 2015). The best characterized of these, UBQLN-1, functions similar to Dsk2 and Rad23 by delivering poly-ubiquitinated proteins to the 26S proteasome. Other demonstrated roles for UBQLN-1 are an association with aggregates for delivery to the lysosome for degradation and in the ER-associated degradation (ERAD) pathway (D. Zhang, Raasi, and Fushman, 2008; Seok Ko et al., 2004; El Ayadi et al., 2013; Lim et al., 2009). In UBQLN-1, STI1-domains have been shown to bind to TMDs of mitochondrial membrane proteins and target them to the proteasome for degradation (Itakura et al., 2016). The direct involvement of UBQLNs in client degradation suggests a broader role than simply being shuttling factors (Itakura et al., 2016).

Here, we inspect these identified STI1-domains and the proteins they reside in to clarify the criteria for this domain. Upon examination, there are clear similarities in the structural features of most STI1-domains, while some currently defined STI1-domains are likely misannotated. Based on structure-based sequence alignments

and similarity in predicted secondary structure, we develop a new definition that has allowed the identification of other STI1-domains and clarification of previously misannotated domains. We employ structural prediction methods to model uncharacterized STI1-domains revealing a consistent alpha-helical hand architecture. When considering proteins that contain STI1-domains, we find similar functional roles and domain architecture. In total, this work provides a comprehensive definition of the STI1-domain.

## 4.2 Results

### Amalgamating and examining predicted STI1-domains

A search through protein databases shows a variety of entries with a name or protein domain annotation that includes "Sti1". Close homologs of HOP have "Sti1" or "Sti1-like" in their entry name. These are typically bidirectional BLAST best-hits, i.e. for a new sequence, the top scoring hit in a reference database identifies the original sequence when searched against the set of new sequences (Ward and Moreno-Hagelsieb, 2014). Other entries also have "Sti1" domain annotation. This annotation originates from a hit to the STI1 Hidden Markov Model (HMM) created by the Simple Modular Architecture Research Tool (SMART) database (since 2001) (Letunic and Bork, 2017) and more recently from an HMM in the Pfam v32 database (El-Gebali et al., 2018) (Fig. 4.1 *D*). The STI1-domain is named as such due to the prevalence of Sti1 homologs in the seed sequences of the SMART and Pfam HMMs. HMM-based methods reliably identify homologs of lower sequence similarity and are much more sensitive than sequence-to-sequence searches such as BLAST. HMM-based searching achieves higher sensitivity by using an alignment of query sequences (a "seed") to search a target database. Typically, the set of protein sequence regions that comprise seeds are curated by hand, but the creators of SMART developed an automated method to compile seed protein regions for the categorization of protein domains distinct from those identified by human curators, as in the case of Pfam.

The SMART HMM for the STI1 family provides a useful initial annotation, yet when comparing the HMM to known structures and sequence alignments of STI1-domains, several issues come to light. The first being a partial hit for each DP domain in Sti1 where only four of the five helices are recovered by the SMART HMM (Fig. 4.1 *D*). The last helix of DP1 and the first helix of DP2 are both not covered by the HMM hits in the Sti1 sequence. Accordingly, we sought to understand to what degree the protein regions identified by this HMM actually reflect a single homologous family.

We performed structural alignments (Shindyalov and Bourne, 1998) (Fig. 4.2 *A*) across putative members of this family with experimentally solved structures (Sti1-DP1, Sti1-DP2, Tic40-STI1-II, and Rad23). Sti1-DP1 and Sti1-DP2 have clear homology indicated both by structural similarity (Fig 4.1 *B&C*) and structure-guided sequence alignment using PROMALS3D (Pei, Kim, and Grishin, 2008) (Fig 4.1 *D*). Tic40-STI1-II closely resembles Sti1-DP2 (Fig. 4.2 *B*). For Rad23, the orientation and register of the helices differs from the other domains. This can be visualized with respect to the client binding position where Rad23-STI1 forms the binding-groove with a rotated helix organization compared to the other STI1-domains (Fig. 4.2 *B*). In this orientation, the first helix of Rad23-STI1 aligns to the third helix of Sti1-DP2 and the fourth (last) helix of Rad23-STI1 occupies a position similar to the first two helices of Sti1-DP2. Based on alignments and structures Tic40-STI1 is clearly a member of the STI1-domain family whereas Rad23-STI1 may be erroneously annotated (Fig. 4.2 *A*).

Despite its adequate utility, the SMART definition for STI1-domains can also lead to erroneous annotations of putative domains that have not been structurally char-

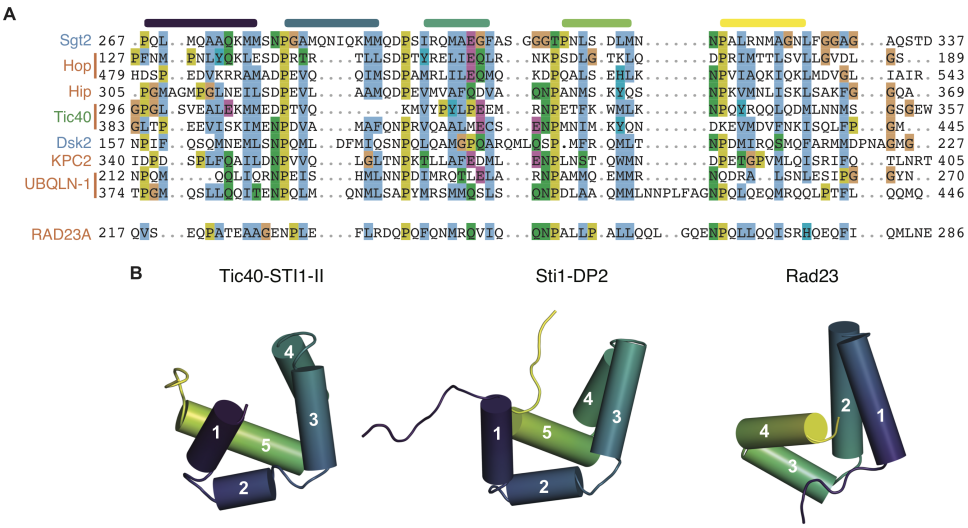



Figure 4.2: **Structure-based sequence alignment of identified STI1-domains.** *A)* Structure-based alignment of identified STI1-domains colored using the ClustalX color scheme (Thompson et al., 1997) with secondary structure elements indicated above the alignment. Names are as in Fig. 1 with the addition of green for *A. thaliana*. *B)* Cartoon representation of the structures of Tic40-STI1-II (PDBID: 2LNM), Sti1-DP2 (PDBID: 2LLW) and Rad23 (PDBID: 1X3W) colored using a viridis color scale from N-terminus (*purple*) to C-terminus (*yellow*). Helices are numbered in white.

acterized. Drawing the stretches of each protein with a hit from the SMART HMM alongside experimental or predicted helical regions, we can separate proper yet partial hits, e.g. Sti1-DP1 and Sti1-DP2, from potentially erroneous ones (Fig. 4.3). An erroneous hit might originate from more than one consecutive hit to the HMM. UBQLN-1, -2, & -4 are predicted by the SMART HMM as having two pairs of abutting STI1-domains, for a total of four (Fig. 4.3, Ubiquilin-1 I&II) (Letunic and Bork, 2017). In each case, the total length of each abutting hit was around 70 residues, not 100 which would be necessary for two STI1-domains considering the length of structurally determined STI1-domains. In addition, the secondary structure prediction within this region showed only seven helices (Fig. 4.2 *A* & 4.3), sufficient for a single domain only. Thus, it is unlikely that there are two pairs of abutting domains.

As discussed earlier, the structure of the SMART identified STI1-domain in Rad23 differs from that of Sti1 DP domains and Tic40-STI1-II, again making it a possible erroneous annotation. The SMART HMM covers the same region as the Pfam XPC-binding HMM, but with a slightly lower score (27 vs 43); both align over a similar number of residues (40 vs 44). Given that the SMART HMM identifies the clear structural homolog Tic40-STI1-II with a similar score of 35, we cannot rule out Rad23 as a member of the STI1 containing family solely based on the SMART HMM score. Since the XPC-binding HMM uses Rad23-STI1 as part of the seed sequences, it could also be possible that the XPC-binding domain is a subfamily of STI1-domains. Although the Rad23 structure in this region appears distinct from other STI1-domain structures, it could be a member of this family based on its score and alignment to the SMART HMM. An HMM with higher specificity could more clearly delineate the difference between Rad23-STI1 and other STI1-domains, which we now aim to define.

**A new definition for the STI1-domain**

In light of these issues with the SMART definition for STI1-domains, we sought to generate an HMM that better defines the full-length of the STI1-domain and thereby more sensitively captures the full breadth of the STI1 family. First, we created an alignment of protein sequences that correspond to both structurally characterized STI1-domains and close homologs then aligned others with constraints from molecular models (see below) and/or secondary structure predictions. As expected, this multiple sequence alignment reveals a strong conservation of structural features (Fig. 4.2 *A*). The predicted helical regions, helices 1-5 (H1-H5), in several proteins

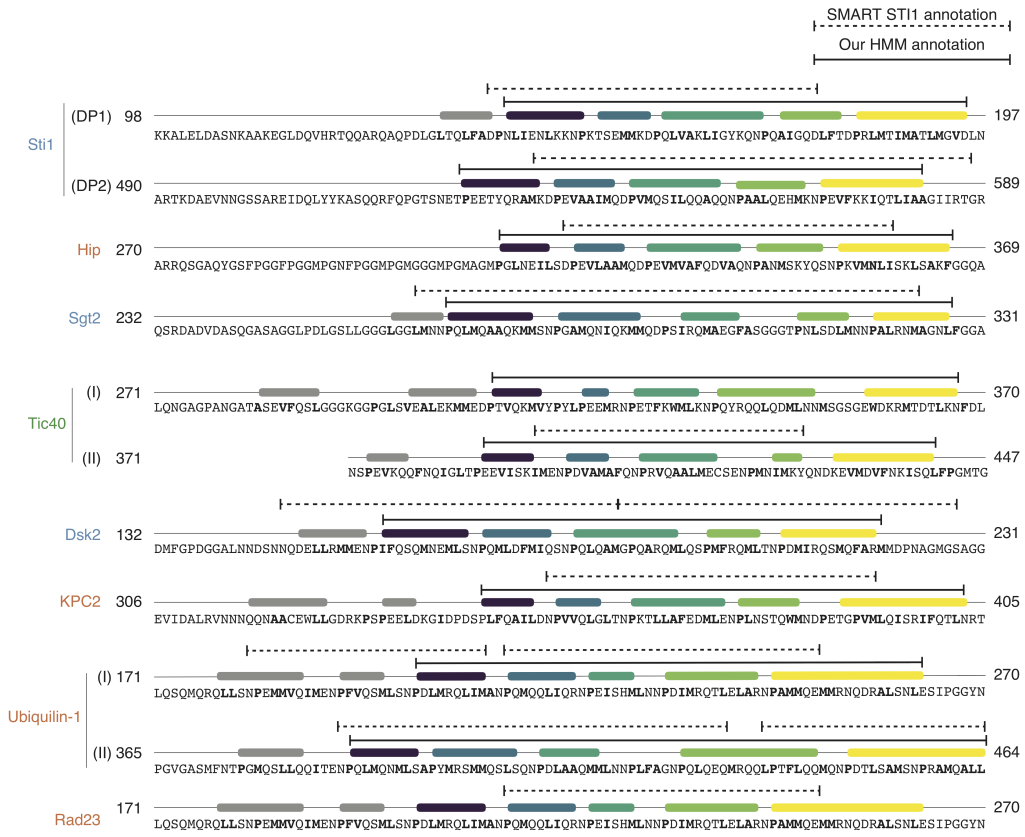



Figure 4.3: **Redefining the STI1-domain model to properly account for number of helices.**
Secondary structure prediction is depicted for each identified STI1-domain. Helices residing within the STI1-domain are colored in viridis (as in Fig. 4.1) and N-terminal helices, H0, are colored in grey. The HMMs are indicated by lines over the secondary structure for each model showing the SMART HMM (*dashed line*) and our new HMM (*solid line*). Hydrophobic residues within the STI1-domains and H0 are indicated in bold.

align directly with the structurally determined helices of DP2 from Sti1 and Tic40-STI1-II. This includes complete conservation of helix breaking prolines and close alignment of hydrophobic residues defining amphipathic helices. The amphipathic nature of these helices appear important for client binding; experiments mutating the hydrophobic faces in these helices in SGTA to less hydrophobic alanine affected binding to tail-anchored protein clients (Lin et al., 2021).

The resulting HMM clears up a number of the issues with the SMART HMM defining STI1-domains. Most identified STI1-domains align well between the two lists with the exception being Rad23 (Fig. 4.2 *A*). Along with other factors discussed later, this suggests Rad23 belongs to a class of STI1-like domains, which could include

proteins like Ddi1 (Trempe et al., 2016). In the case of UBQLNs, the resulting HMM identifies the annotated abutting STI1-domains as a single STI1-domain (Fig. 4.3).

Opposite to what was revealed in the UBQLNs, where a reduction in the number of identified STI1-domains was observed, a second N-terminal STI1-domain is identified by our HMM in Tic40 (Fig. 4.3). Previously this region of Tic40 was suggested to be a TPR domain primarily on the basis of a binding to an anti-TPR1 antibody by western blot (Chou, Fitzpatrick, et al., 2003). While it is reasonable to suspect that Tic40 possesses a TPR domain since TPR domains precede STI1-domains in HOP, HIP, and SGTA, the TPR domain HMMs (El-Gebali et al., 2018) do not suggest a hit in this region. However, this region does produce a hit by our STI1 HMM. The anti-TPR1 antibody was generated against full-length rat TPR1 (F.-H. Liu et al., 1999), and it is possible it lacks specificity for this plant TPR domain. We were unable to identify identical peptides longer than five residues between Tic40 and rat STI1 that could easily explain the cross reactivity. Due to the bioinformatic support for a STI1-domain in this region, we refer to it as Tic40-STI1-I and the structurally solved STI1-domain as Tic40-STI1-II.

**STI1-domains are preceded by an N-terminal helix**

Along with a curated set of STI1-domains, this new HMM reveals a conserved N-terminal sixth helix, hereafter referred to as Helix 0 (H0), which like H1-H5 is also amphipathic (Fig. 4.4). As already noted, UBLQNs were mistakenly characterized as having four STI1-domains. The incorrect annotation was likely because three helices N-terminal to the domain were combined with the five helices of the STI1-domain, which resulted in eight contiguous helices recognized by SMART as two adjacent STI1-domains. The presence of N-terminal helices appears to be a general feature of STI1-domains. Based on secondary structure prediction and structures, it is clear, in most cases, that at least one helix precedes the STI1-domain, the exceptions being Sti1 DP2 and HIP (Fig. 4.3). While the roles of the additional helices are not clear, H0 is well conserved within each protein and are also amphipathic in nature as the other helices (Fig. 4.3 & 4.4).

**Structural similarity between STI1-domains**

With this new list of STI1-domains we inspected their predicted and structurally determined secondary structures. Broadly, these domains share several features including four to five amphipathic helices, as annotated (Fig. 4.2 *A*). For STI1-domains that have been structurally characterized (DP1, DP2, Tic40-STI1-II), the

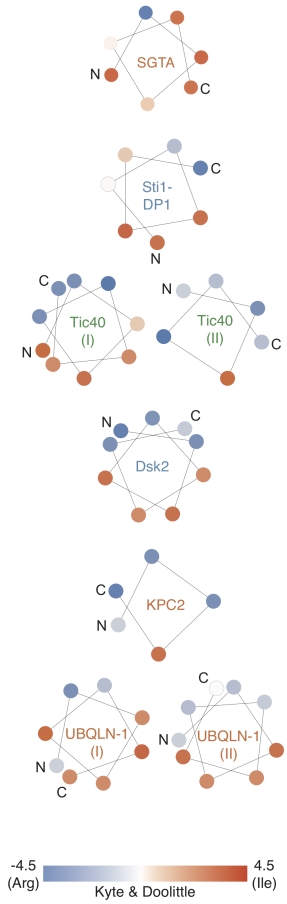



Figure 4.4: **The amphipathic nature of the N-terminal helix preceding STI1-domains.**
Helical wheel plots of the H0 helix immediately preceding identified STI1-domains. Residues making up H0 are represented as circles and colored based on their hydrophobicity using the Kyte & Doolittle scale (Kyte and Doolittle, 1982). The N-terminus (N) and C-terminus (C) of each helix is annotated.

helices assemble into a tertiary structure that resembles a helical-hand forming a hydrophobic groove (Fig. 4.5 *A,B,&C*) and are characterized by structural flexibility (Lin et al., 2021). Though no structures of a STI1-domain from a co-chaperone exist with a client occupying the hydrophobic groove, it presents an appealing pocket for the binding site of hydrophobic segments. The flexibility may contribute to the ability of these domains to specifically bind and then release clients as part of their functional role.

DP1 and DP2 were the first STI1-domains to be structurally characterized. The

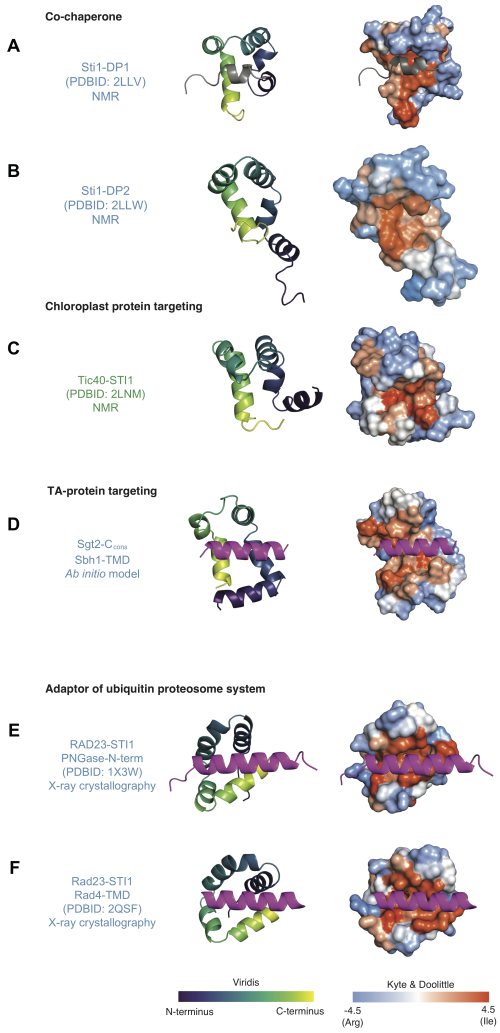



Figure 4.5: **Published structures and models of STI1-domains reveal an alpha-helical hand that forms a hydrophobic groove.**
Structural models of STI1-domains as cartoon and surface hydrophobicity representations: DP1 and DP2 from yeast Sti1 (Sti1-DP1 and Sti1-DP2) (PDBID:2LLV & 2LLW), Tic40 from Arabidopsis thaliana (Tic40-STI1) (PDBID:2LNM), a computational model of the C-domain from the yeast co-chaperone Sgt2, yeast Rad23 (Rad23-STI1) bound to the N-terminus of PNGase (PNGase-N-term) (PDBID: 1X3W) and yeast Rad23 bound to the TMD of yeast Rad4 (Rad4-TMD) (PDBID: 2QSF). Cartoons are colored as in Fig. 4.1 *B* with bound helices in magenta. The surface representation is colored based on hydrophobicity.

two structures have five amphipathic alpha helices arranged like a cupped hand presenting a hydrophobic groove. In the structure of DP1 the hydrophobic hand is occupied by H0, possibly mimicking client binding (Fig. 4.5 *A*, *grey helix*). Both structures of DP1 and DP2 were solved using NMR. When comparing the states from the models for each domain, DP2 appears to be more flexible than DP1, with the N-terminal H0 in DP1 likely stabilizing its core region (Schmid et al., 2012).

Tic40 contains another structurally characterized STI1-domain in the absence of a client occupying the groove. Like DP1 and DP2, Tic40-STI1-II consists of five alpha helices that arrange into a similar helical hand with a hydrophobic groove (Fig. 4.5 *C*). Tic40 is predicted to have an H0 that was not included in the determined structure. Other STI1-domains have remained resistant to structure determination.

Alternative structural methods have been used to characterize other STI1-domains. One domain in particular is the C-terminal domain of fungal homolog of SGTA, Sgt2, which remains recalcitrant to experimental structural determination. Ab initio molecular modeling of Sgt2-C followed by experimental validation of residue-pair distances further suggests that the domain is part of the STI1 family (Lin et al., 2021) (Fig. 4.5 *D*). Residues of a conserved region resolve a potential binding interface for a helical hydrophobic client. Outside this region they adopt varied conformations consistent with expected high flexibility. The working model contains a potential TA protein binding site – a hydrophobic groove formed by the amphipathic helices. The groove is approximately 15 Å long, 12 Å wide, and 10 Å deep, which is sufficient to accommodate three helical turns of an alpha-helix, ~11 amino acids. Like the NMR structures of other STI1-domains found in co-chaperones, the *ab initio* model of Sgt2-C resembles the general STI1-domain structure.

For the STI1-domain containing AUPS proteins, no experimental or *ab initio* structures currently exist. To predict structures of these STI1-domains including those in the UBQLNs we employed the Robetta transform-restrained (TR) tool, a state-of-the-art structure prediction method (Yang, Anishchenko, et al., 2020b) where a deep neural network predicts pairwise residue distances and angles followed by energy minimization. Distinct from template- or fragment-based approaches, Robetta TR generates de novo structures from restraints where structures are not explicitly used. As validation for our domains, we first compare the prediction of multiple STI1-domains by Robetta TR versus the experimentally derived structure. Providing the full-length sequences from *Sc*Sti1 and *At*Tic40, Robetta TR provides a model of the full-length protein (Fig. 4.7 *A*). We isolated the STI1-domains from each and

compare them to the structures solved by NMR (Fig. 4.6 *A* & 4.7). For the top predicted models, DP1, DP2, and Tic40-STI1-II have five helices that assemble into a helical hand (Fig 4.6 *A*, 4.7 *B&C*). These predictions are in close agreement with the NMR derived structures, with the last five helices of the prediction overlaying with the five alpha helices in the NMR structure – supporting that the prediction method can provide data broadly on STI1-domains (Fig. 4.6 *A*, 4.7 *B&C*).

We proceeded to predict the structures of uncharacterized STI1-domains in other co-chaperones. The predicted structure for Tic40-STI1-I has a similar fold to other STI1-domains, the five alpha-helical hand, supporting the new domain definition (Fig. 4.6 *B*). This model further reduces the likelihood that this region contains a TPR domain that are structurally distinct alpha-solenoids. The STI1-domain from human HIP (Fig. 4.6 *C*), the last of the uncharacterized co-chaperone STI1-domains, is similar to the experimentally determined Tic40-STI1-II (Fig. 4.6 *A*), DP2, and DP1 (Fig. 4.5 *A&B*). Across all co-chaperones we observe five helices coming together to form an alpha helical hand.

We next predicted the structures of the uncharacterized STI1-domains in the AUPS family (Dsk2, KPC2, and UBQLNs) (Fig. 4.6 *D-G*). Like HIP, the predicted structure of KPC2 is consistent with the experimentally determined structures of STI1-domains from the co-chaperones. The predicted structure of the Dsk2 and UBQLN-1 also display a variation of the STI1 helical hand observed in the solved structures (Fig. 4.6 *D,F&G*). A helical groove compatible for binding a hydrophobic alpha helix is formed by five helices in each of these cases. Dsk2-STI1 (Fig. 4.6 *D*) and the second STI1-domain of UBQLN-1 (UBQLN-1-STI1-II) differs from the co-chaperone STI1-domains as their helices form the groove in the reverse order (Fig. 4.6 *G*). Unlike the prediction of UBQLN-1-STI1-II, the first STI1-domain (UBQLN-1-STI1-I) forms a nearly enclosed groove, which could accommodate client binding with some rearrangement (Fig. 4.6 *F*). These predictions are models for what these STI1-domains could look like and in both the co-chaperones and AUPS these models suggest that these STI1-domains can form helical hands to accommodate client binding. UBQLNs (UBQLN 1-4) are a particularly important focus of research and experimental work to test these predictions will be broadly useful (Itakura et al., 2016; Deng et al., 2011; Şentürk et al., 2019; Subudhi and Shorter, 2018; Marín, 2014).

The structure of Rad23-STI1 bound to client supports the exclusion of Rad23 from the STI1-domain containing family of proteins (Fig. 4.5 *D&E)*. While several

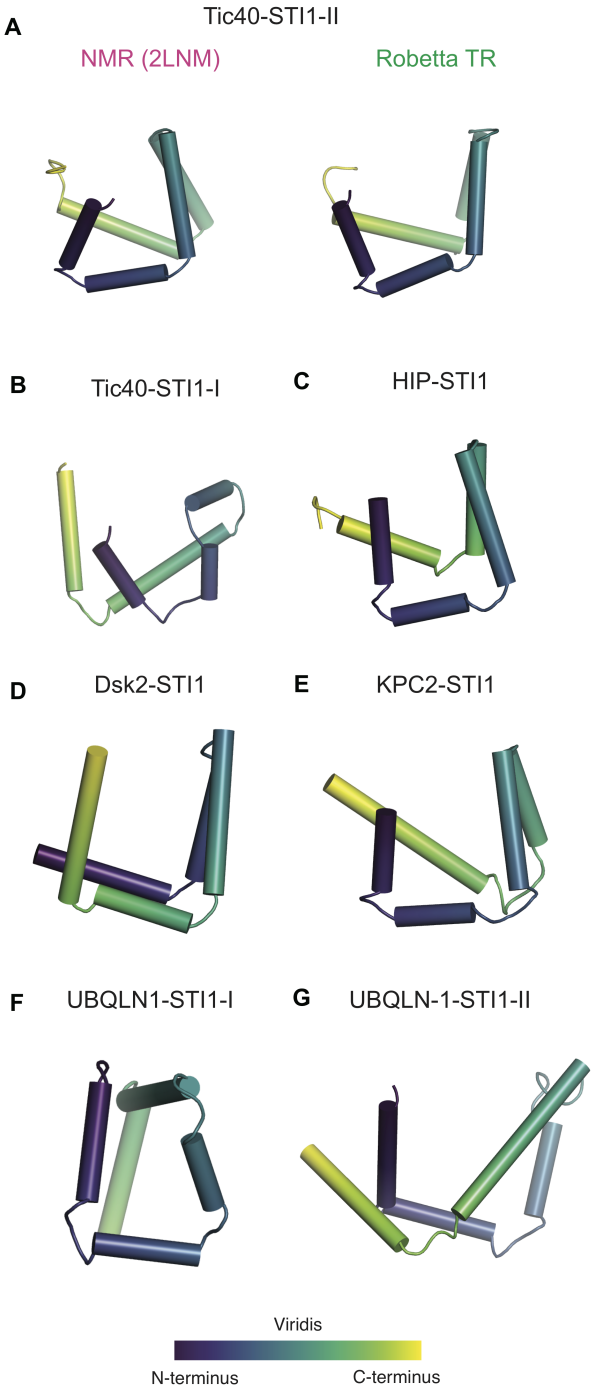



Figure 4.6: **Predicted structures of uncharacterized STI1-domains reveal a hydrophobic groove as seen in the NMR solved structures.**
*A)* A comparison of the Tic40-STI1-II structure model either determined by NMR (left) or predicted using Robetta-TR (right). The predicted models using Robetta-TR of the uncharacterized STI1-domain(s) from *(B)* Tic40 (Tic40-STI1-I), *(C)* HIP, *(D)* Dsk2, *(E)* KCP2, and *(F&G)* UBQLN-1. All structures are colored as in Fig. 4.1 *B*.

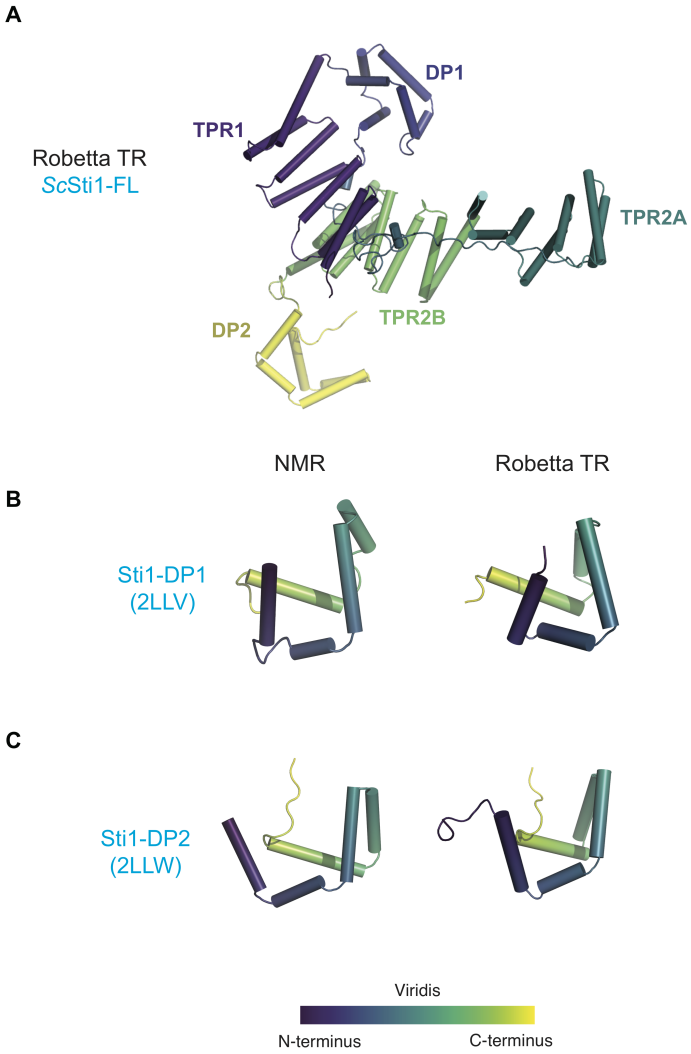



Figure 4.7: **Computational model of *Sc*Sti1.**
*A)* The output model from Robetta-TR from the full-length *Sc*Sti1 sequence input. The model is colored using a viridis color map from N-terminus (*purple*) to C-terminus (*yellow*) with the two DP domains (DP1 and DP2) colored magenta. The five domains of Sti1 (DP1, TPR1, TPR2A, TPR2B, and DP2) are labeled. A comparison of the predicted DP1 *(B)* and DP2 *(C)* domains from Robetta-TR to the NMR-solved structures.

structures of complexes of Rad23-STI1 bound to amphipathic clients show in each that the client-helix binds via a hydrophobic groove, the domain architecture differs from those determined in the co-chaperones and the predicted structures of other STI1-domain containing proteins (Fig. 4.2 *C* & 4.6 *B-G*). Despite Rad23-STI1 being a helical bundle that binds clients similar to the co-chaperones, the absence of a fifth helix supports that Rad23 does not contain a STI1-domain, but is instead a STI1-like domain which also utilizes a hydrophobic groove. It has been observed that the first three helices of the Rad23 XCP domain are structurally similar to the first three helices of the N-terminal domain of the Helical Domain (HDDnt) from the DNA damage inducible 1 protein (Ddi1), but the fourth helix deviates and goes in a different direction (Trempe et al., 2016). Ddi1 differs from other shuttle proteins because of its proteolytic role and interacting partners (Zientara-Rytter and Subramani, 2019). Like the shuttle proteins described above, Ddi1 contains a UBA and UBL domain, but the UBA domain has been lost in mammalian homologs. Ddi1 also contains a retrovirus protease (RVP) fold domain. The UBL domain of Ddi1 has an unusual binding preference, unlike the domain in Rad23 or Dsk2, it does not interact with its UBA domain or Rpn10 and interacts weakly with Rpn1. It has been suggested that Ddi1 may assist Rad23 or Dsk2 instead of acting as a shuttle factor on its own.

Due to its homology with the XCP domain of Rad23, which has been implicated through protein-protein interactions, HDDnt may play a similar role (Trempe et al., 2016). It has also been observed that HDDnt has a similar structure to other DNA binding domains suggesting HDDnt may bind directly to DNA. With structural similarities to various domains, it is reasonable to think Rad23 and Ddi1 both contain STI1-like domains.

**Similarity in the domain structures of STI1 proteins**

When examining the predicted secondary structure of the entirety of STI1-domain containing proteins several common characteristics became clear. Dual STI1-domains are present in both the co-chaperones and AUPS (Fig. 4.8). Within the co-chaperones, two distinct groups emerge — ones that possess two STI1-domains (HOP, Tic40) and those that possess a dimerization domain (SGTA, HIP). As stated previously, HOP contains two STI1-domains separated by multiple TPR domains with both required for efficient client transfer (Schmid et al., 2012). It has been speculated that DP1 and the first TPR domain (TPR1) act as an intermediate in the shuttling of a client from Hsp70 to the TPR2A&B and DP2-bound Hsp90 (Schmid

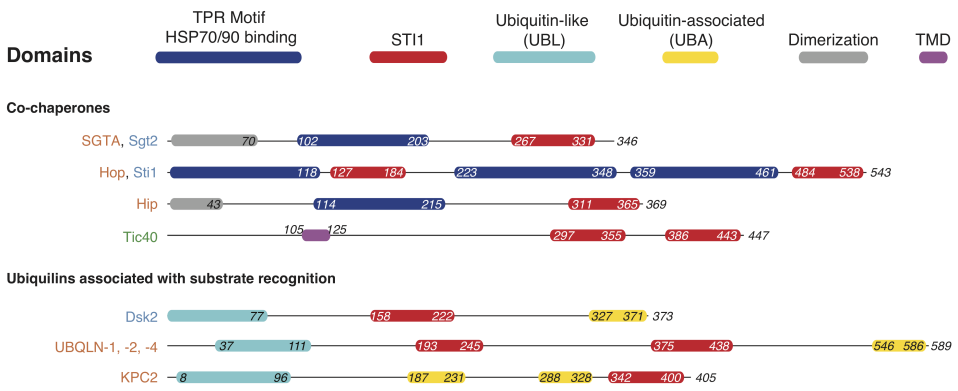



Figure 4.8: **Various domain architectures of STI1-domain containing proteins.** The domain definitions of proteins containing at least one STI1-domain were obtained initially from InterPro (Consortium, 2020) and then adjusted as discussed in the text. Each domain within a protein is colored relative to the key. Numbering where it is not clear is in reference to the human protein. Names are colored as in Fig. 4.1 & 4.2.

et al., 2012; Röhl et al., 2015; Biebl and Buchner, 2019; Lott, Oroz, and Zweckstetter, 2020; Kirschke et al., 2014; Schopf, Biebl, and Buchner, 2017; Röhl, Rohrberg, and Buchner, 2013; Reidy et al., 2018).

Unlike HOP, Tic40 contains two abutting STI1-domains. Given our recent identification of the first STI1-domain, its function has yet to be determined. Found in the chloroplast inner membrane with the STI1-domains in the stroma, the C-terminal domain can be replaced with the STI1-domain from HIP without loss-of-function (Bédard, Trösch, et al., 2017). In HIP the STI1-domain interacts with the leucine-rich chemokine receptor (Fan et al., 2002), the previously proposed TPR domain, now STI1-I, of Tic40 interacts with the leucine-rich mature region of Tic110 (Bédard, Kubis, et al., 2007).

The homodimerization of HIP and SGTA, each containing a single STI1-domain in the monomer, results in a complex with two STI1-domains. Small angle X-ray scattering (SAXS) data has revealed that for both proteins in solution, the dimers (Chartron, Gonzalez, and Clemons, 2011) form elongated, flexible complexes (Z. Li, Hartl, and Bracher, 2013). The elongated form would put maximal distance between the two STI1-domains which are found opposite of the dimerization domains. The fact that the STI1-domain from HIP can functionally replace a STI1-domain from Tic40, both in pairs, but one through dimerization domain and the other encoded in the monomer, suggests the STI1-domains have similar overall functions, the significance of these pairs in the co-chaperones is still unclear.

The presence of a pair of STI1-domains is also observed in UBQLNs. As discussed earlier, UBLQNs contain two STI1-domains and not four as previously thought. This clarification relates UBQLNs to the co-chaperones HOP and Tic40, where each have a STI1-domain pair encoded in their monomer, while still separating them from other AUPS family members which contain a single STI1-domain (Fig. 4.8). The observation of a pair of STI1-domains in UBLQNs and the previously identified active roles in protein targeting and degradation (Itakura et al., 2016) set these proteins apart from the other AUPS family members. The previously defined M domain of UBQLN-1 contains both identified STI1-domains and is responsible for its ability to shield TMDs of mitochondrial membrane proteins from the cytosol and to deliver them for degradation (Itakura et al., 2016). Identifying pairs of STI1-domains in both co-chaperones and UBLQNs, proteins with a known role in protecting TMDs in the cytosol through their STI1-domains, suggests these pairs aid in this role.

While a pair of STI1-domains is a found in both AUPS and co-chaperone proteins, an HSP-binding TPR domain preceding the STI1-domain(s) connected by a flexible linker is exclusively observed in the co-chaperones (Fig. 4.8). These TPR domains have been shown to aid in client hand-off in these proteins. The multiple TPR domains in HOP are used to coordinate simultaneous binding of Hsp70 and Hsp90, facilitating client transfer between the two chaperones (Schmid et al., 2012; Scheufler et al., 2000; Zeytuni and Zarivach, 2012). In contrast, HIP contains a TPR domain that only interacts with Hsp70. Additionally in Sgt2, the TPR domain increases the efficiency of capture of TA proteins by coordinating with a client bound Hsp70 homolog, Ssa1 (Cho and Shan, 2018). While HOP has two TPR domains within a monomer, both SGTA and HIP link two TPR and STI1-domains by forming stable dimers via N-terminal dimerization domains (Coto et al., 2018). For the SGTA and HIP homodimers, a cooperative role between the two copies of each TPR- and STI1-domain remains a possibility.

Differing from the other co-chaperones, the relatively more distant chloroplast Tic40 has its own domain architecture. Previously, the N-terminal STI1-domain was annotated as a TPR domain but, as discussed earlier, bioinformatics and computational models counter this claim. The rest of the protein lacks a clear TPR domain and has an N-terminal TMD. How Tic40 fits mechanistically into this group of co-chaperones is less clear due to it missing a TPR domain and being membrane bound.

**Amino acid distribution in STI1-domains**

STI1-domains were initially described as DP domains due to two repeats of a DPEV motif in HIP (Prapapanich, Chen, and Smith, 1998) and a DPEV and DPAM motif in HOP (Chen and Smith, 1998). From the solved structures (Fig. 4.5) and predicted secondary structure (Fig. 4.3), we see this motif localizing to the N-terminus of helices likely acting as a cap. Aspartate and threonine most frequently occur at the cap of a helix often followed by either a glutamine or proline (Aurora and Rosee, 1998). The role for this motif where found is likely as a stabilizing N-cap accounting for its conservation. When analyzed broadly, a repeat DP motif is not observed in the majority of STI1-domains, it not even found in all HOP homologs.

We were interested if there were common residues overrepresented in STI1-domains, considering that in some cases these domains have been referred to as methionine-rich (Itakura et al., 2016; Lin et al., 2021). To quantify this, we began with compiling a list of homologs for HOP, SGTA, HIP, UBQLN-1, and KPC2 using EnsemblGenomes (Burri and Lithgow, 2003) and then calculated the distribution of amino acids within the STI1-domain. As for Dsk2 homologs, only two hits were found after searching EnsemblGenomes, therefore we omitted Dsk2 from this analysis. Only an overrepresentation of methionine, asparagine, and leucine is observed across all five protein groups (Fig. 4.9). An overrepresentation of methionine has also been observed in other hydrophobic segment binding domains such as the M domain of the signal recognition particle (SRP) (Bernstein et al., 1989; Zopf et al., 1990) and Get3 (Mateja, Szlachcic, et al., 2009; Suloway, Chartron, et al., 2009). As discussed previously, Rad23 likely does not contain a STI1-domain. We applied the same analysis of the amino acid distribution in the previous annotated STI1-domain of Rad23 and found that, unlike verified STI1-domains, methionine is not overrepresented (Fig. 4.9). Overall, this analysis reveals that over representation of methionine, asparagine, and leucine is a feature of STI1-domains.

## 4.3 Discussion

STI1-domains have been predicted in a number of proteins essential for protein biogenesis. Here we explicate a definition for STI1-domains and curate a list of STI1-domain containing proteins through structure-based sequence alignments, validating some previously predicted domains as well as identifying new ones. Solved structures and computational models reveal STI1-domains consist of five to six helices organized into a helical hand with a hydrophobic groove. Upon close inspection, STI1-domain containing proteins can be classified into two families –

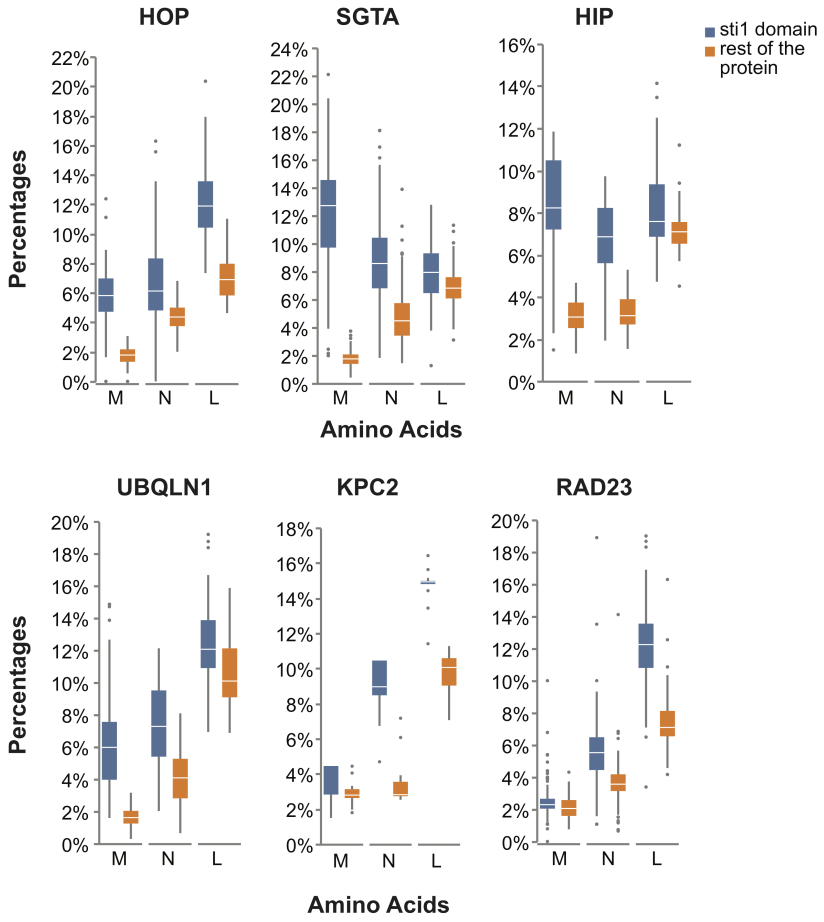



Figure 4.9: **Methionine, asparagine, and leucine are overrepresented in STI1-domains.**
Box plots of the prevalence of methionine (M), asparagine (N), and leucine (L) in homologs of HOP (176 sequences), SGTA (155 sequences), HIP (39 sequences), UBQLN-1 (95 sequences), KPC2 (11 sequences), and Rad23 (169 sequences) within the STI1-domains (blue) and the rest of the protein (orange). Amino acids that were overrepresented across all five STI1-domain containing protein groups are shown.

co-chaperones or AUPS – with several common features noted. Overall, this work presents the first in-depth examination of STI1-domains and the essential proteins for where they are found.

Previously, STI1-domains were identified by the SMART database, which lead to the identification of erroneous hits and omissions for the family. Our revised HMM that encompasses a minimal five helix region helps uncover and clarify the full breadth of STI1-domains. As a result, this new definition revealed new STI1-domains and corrections to previous identifications. For example, the annotations of four abutting

STI1-domains in UBQLNs are actually a set of two STI1-domains and the annotated TPR domain in the co-chaperone Tic40 is more likely a STI1-domain. Furthermore, it is now clear that Rad23 does not contain a STI1-domain, but has a distinct helical hand formed of only four helices.

This revised list was evaluated to determine common structural features. The overall five amphipathic helices forming a flexible helical hand are seen in the structures of DP2 and Tic40-STI1-II (Schmid et al., 2012). Differing from the structure of the DP2, the DP1 structure contains a sixth helix (H0) that resides in the groove (Fig. 4.1 *B* & 4.5). With our new HMM, an H0 was identified preceding most STI1-domains and its role is yet to be determined (Fig. 4.3). Due to the flexibility of STI1-domains, one possibility is that H0 fills the groove as seen in the DP1 structure (Guna and Hegde, 2018). In this model the hydrophobic residues in H0 would dock in the groove of the helical hand stabilizing the domain in the absence of client. H0 would then be displaced by an appropriate client.

Outside of the STI1-domains, several common features surface in these STI1-domain containing proteins. A distinct feature of the co-chaperone family are the TPR domains. We previously discussed the various TPR domains identified in Sgt2, HOP, and HIP, noting sequence features that define specificity in interacting partners (Chartron, Gonzalez, and Clemons, 2011). TPR domains consist of multiple repeats of 34 amino acids in a helix-turn-helix fold, with anti-parallel alpha-helices. Differences in the binding pocket of TPR domains allow for selectivity of a diverse set of chaperone partners. In the case of SGTA, the TPR-domain works together with the STI1-domain to coordinate client binding. Ssa1 carrying STI1-domain clients interact directly with the TPR domains allowing for client capture by the STI1-domains. Here we demonstrate that the suggested TPR domain of Tic40 is more likely a STI1-domain based on its higher score against the STI1 HMM vs the TPR HMM. This adjustment to the domain structure within Tic40 may suggest that Tic40 does not interact with HSPs to capture clients as seen for the other co-chaperones.

Pairs of STI1-domains are found in both co-chaperones and the AUPS family, either encoded in a single monomer or joined through a dimerization region. While the co-chaperones identified in this paper contain STI1-domains in pairs, for the AUPS family this is only true for UBQLNs. Of the AUPS family members, UBQLNs are the only ones so far shown to play a direct role in preventing client aggregation in the cytosol and facilitating the degradation of mitochondrial membrane proteins that

fail to insert into the mitochondrial membrane (Itakura et al., 2016). These roles in both protein targeting and degradation are similar to those of SGTA – handing off TA proteins to chaperones in the GET pathway for insertion and handing off mislocalized proteins to Bag6 for degradation (Pawel Leznicki and High, 2012). The pair of STI1-domains lie in the identified client binding domain of UBQLN-1. Perhaps due to similar roles, these domains in UBQLNs function similarly to the STI1-domains in SGTA. The details of how STI1-domain pairs affect function and if they interact with one-another are important areas for future study.

How might pairs of STI1-domains cooperate for client specificity and selection? Conceptually, a pair of STI1-domains may simultaneously bind the same client TMD in the case of UBQLNs and SGTA. We have shown previously that a single STI1-domain from Sgt2 can bind to a minimum of 11 amino acids in a client (Lin et al., 2021). Conceivably, one model is that the two STI1-domains in the Sgt2 dimer simultaneously bind a single client — binding side-by-side on a client TMD that averages 20aa. This would require that these domains come close together altering the overall architecture. A related model is that the pair of STI1-domains could cooperate to increase the apparent affinity for a client TMD by increasing the local concentration of the binding domain. The simplest model is that each STI1-domain binds a separate client either to increase client load (two clients per monomer/dimer instead of one) or there is a necessity to bind two clients at once. On the other hand, it is also possible that two STI1-domains are necessary for different functions, as proposed previously in the case of HOP by Schmid and colleagues where glucocorticoid receptor activation cannot by rescued by replacing DP2 with DP1. It is worth noting that low resolution structural studies have suggested that dimeric Sgt2 and HIP position their STI1-domains on opposite ends of a dimer molecule in the absence of client (Coto et al., 2018; Chartron, Gonzalez, and Clemons, 2011). Still, given the noted flexibility in these proteins (Lin et al., 2021; Schmid et al., 2012), the possibility of cooperation remains, with the molecular details an open question.

Flexibility is a common motif seen within STI1-domains and the proteins where they are identified. NMR studies of STI1-domains have suggested that these domains are flexible. We consider this flexibility a feature of these helical-hands for reversible and specific binding of a variety of clients. But what is the benefit of the flexible helical-hand structure for hydrophobic helix binding? While it remains an open question, it is notable that evolution has settled on similar simple solutions to the

complex problem of specific but temporary binding of hydrophobic helices. For all of the domains with experimentally determined structures, the flexible helical-hands provide an extensive hydrophobic surface to capture the client-helix. Required to only engage temporarily, the flexibility of the helical hand could offset the favorability of the domain to bind a hydrophobic client, allowing the client to be released. This would account for the favorable transfer seen from Sgt2 (Cho and Shan, 2018) and SGTA (Shao, Rodrigo-Brenni, et al., 2017) to downstream components.

## 4.4 Conclusion

This work provides a comprehensive HMM to define and identify STI1-domains in proteins and recognizes common features observed within the domains themselves and across STI1-domain containing proteins. These patterns leave open questions that have yet to be determined. What is the role of the flexibility within STI1-domains and STI1-domain containing proteins? Does helix zero act as a stand in for clients in their absence or does it have a role as a lid? What is the benefit of having two STI1-domains in the co-chaperones? Do UBQLNs have a larger role than other AUPS family members and do their two STI1-domains contribute to this different role? There is much still to be understood about the underlying mechanisms that result in specificity and client hand-off of STI1-domains. This comprehensive list of STI1-domains provides a coherent starting point.

## 4.5 Material and Methods

*Molecular visualization*

All STI1-domains with experimentally determined structures were retrieved from the RCSB. *Ab initio* structure prediction was employed for other STI1-domains using RobettaTR (transform restrained) (Yang, Anishchenko, et al., 2020a) with the full-length protein sequences of each protein, with the specific STI1 region of interest visualized. Images were rendered using PyMOL 2.4 (www.pymol.org) with a viridis coloring scheme (Saladi et al., 2020). Helical wheel diagrams were rendered in R using a fork (https://github.com/smsaladi/heliquest) of the HELIQUEST source code (Gautier et al., 2008).

*Sequence analyses*

Alignments of Sti1 (DP1/DP2) and STI1-domains were created by pulling all unique domain structures with annotated STI1-domains from Uniprot. Sequences were clustered at 50% similarity present, with the human, yeast, and *A. thaliana* preferred

and then aligned with PROMALS3D (Pei, Kim, and Grishin, 2008) along with all experimentally determined structures of STI1-domains. PROMALS3D provides a way of integrating a variety of costs into the alignment procedure, including 3D structure, secondary structure predictions, and known homologous positions. The human, yeast, and A. thaliana homologs were selected from this alignment for display. An HMM for the STI1-domain was generated using HMMER v3.3.1. Alignments were visualized using Jalview (Waterhouse et al., 2009). Secondary structure where indicated is calculated using DSSP (Kabsch and Sander, 1983) on experimentally determined or predicted structures (Yang, Anishchenko, et al., 2020a).

*Amino acid composition of STI1-domains*

Homologs of *Sc*Sti1, *Hs*HOP, *Sc*Sgt2, *Hs*SGTA, *Hs*HIP, *Sc*Dsk2, *Sc*KPC2, *Hs*UBQLN-1, and *Hs*Rad23A were compiled from EnsemblGenomes Howe:2019jk and filtered for redundancy at 70% sequence identity using the CD-HIT Suite Li:2006hr. Sequences were then aligned using MAFTT (Katoh and Standley, 2013). The STI1-domain(s) in each sequence were identified by alignment to the known STI1-domains of HsHOP and HsSGTA to yield two segments: the STI1-domain and the non-STI1-domain region, i.e. the "rest" of the protein. For each segment of each protein the percentage of all individual amino acids was calculated. Significance was determined by permutation testing, comparing the difference between the first quartile of an amino acid's percentage between each segment, i.e. within vs outside of the STI1-domain.

*Data Availability*

Our refined HMM for the STI1-domain (HMMER format), alignment shown in Fig. 4.2 (FASTA format), and each of the structural models (PDB format) are included as supplementary data. Structural models should be used with caution.

## 4.6 Acknowledgements

This work was supported by the National Institutes of Health (NIH) Pioneer Award GM105385 and Grant GM097572 (to WMC), NIH/National Research Service Award Training Grant T32 GM07616 (to SMS and MYF), and a National Science Foundation Graduate Research fellowship under Grant 1144469 (to SMS). The authors declare that they have no conflict of interest.

*Chapter 5*

# A COMPREHENSIVE STRUCTURAL VIEW OF THE TAIL-ANCHOR TARGETING CHAPERONE GET3 CATALYTIC CYCLE BASED ON STRUCTURES FROM A HUMAN PATHOGEN

*Adapted from:*

Fry, Michelle Y et al. (2021). "The conformational changes during the catalytic cycle of the tail-anchor targeting chaperone Get3 based on structures from a human pathogen". *Submitted*.

M.Y. Fry collected and refined X-ray diffraction data sets and collected and processed the single particle cryo-electron microscopy data sets. M.Y. Fry participated in the design and execution of biochemical experiments.

## Abstract

The correct targeting and insertion of membrane proteins is crucial for the maintenance of cellular homeostasis. The tail-anchored (TA) protein class of membrane proteins that account for 2% of eukaryotic genomes are targeted post-translationally. The pathway responsible for delivering and inserting the most hydrophobic of these proteins to the ER membrane is the Guided Entry of TA proteins (GET) pathway. The central targeting factor is the ATPase Get3 and its nucleotide state has been shown to regulate the GET pathway and drive TA protein delivery to the ER membrane. Homologs of Get3 have been predicted across all kingdoms and functionally demonstrated in several eukaryotes, including metazoans, fungi, and plants. Here we present a series of structures of Get3 from the human pathogen, *Giardia intestinalis* in different nucleotide states using single particle cryo-electron microscopy (cryo-EM) and x-ray crystallography in six conformational states. These structures of Get3, the first from a protozoan, reflect key states in the ATPase cycle in the nucleotide free (apo), pre-hydrolysis ATP bound, and post-hydrolysis tail-anchor and ADP bound conformations. This provides the first comprehensive characterization of Get3 from a single organism and a structural picture that illustrates the molecular changes in Get3 required for successful targeting of TA proteins to the ER membrane.

## 5.1 Introduction

The delivery and insertion of integral membrane proteins (IMPs) into designated membranes is a complex process for cells. Most IMPs are targeted co-translationally to the ER membrane via the signal recognition particle (SRP) pathway (Guna and Hegde, 2018). One class of IMPs, Tail-anchored (TA) proteins, contain a single transmembrane domain (TMD) near their C-terminus and must be targeted to cellular membranes post-translationally (Kutay, Hartmann, and Rapoport, 1993; Denic, 2012; Wattenberg and Lithgow, 2001; Chartron, Clemons, and Suloway, 2012). TA proteins account for ~2% of all genomes and targeting of these proteins have been studied in both fungi and mammals (Borgese, Colombo, and Pedrazzini, 2003; Guna and Hegde, 2018; Rabu et al., 2009b).

In the Opisthokont clade, which includes fungi and mammals (Cavalier-Smith and Chao, 2003), the most well-studied pathway for targeting TA proteins to the ER membrane is the conserved Guided Entry of TA protein (GET) pathway, described here for fungi. The central targeting component is the homodimeric ATPase Get3, whose ATP-dependent conformational changes drives targeting of TA proteins (Chartron, Clemons, and Suloway, 2012; F. Wang, Brown, et al., 2010; Simpson et al., 2010; Rome, Rao, et al., 2013). Upstream from Get3 is the co-chaperone Sgt2 (SGTA in mammals) that binds specifically to ER TA proteins (Chartron, Gonzalez, and Clemons, 2011; Borgese, Colombo, and Pedrazzini, 2003). Upon binding to ATP, Get3 transitions from a 'open' to a 'closed' form that is recognized by a Get4/Get5 complex (analogous to the Bag6 complex in mammals), which facilitates the transfer of the TA protein to Get3 through an interaction with Sgt2 (Chio, Chung, et al., 2017; Chartron, Suloway, et al., 2010; Shan, 2019). Once a TA protein binds, ATP hydrolysis occurs driving Get3 into an 'intermediate' state causing Get3 to disassociate from the Get4/Get5 loading complex (Chio, Chung, et al., 2017). A membrane bound Get1/Get2 complex (Get1/CAML in mammals) localizes the TA bound Get3 to the ER, drives the disassociation of the TA protein from Get3 by favoring Get3 in an 'open' state, and then facilitates insertion of the TA protein into the ER membrane (McDowell et al., 2020; Mariappan et al., 2011).

Considerable effort has been dedicated to structurally characterizing the various Get3 states to understand the conformational changes that drive TA protein targeting. Currently, fungal Get3 structures have been solved of the 'open' form as apo (Mateja, Szlachcic, et al., 2009), ADP bound (Suloway, Chartron, et al., 2009), or bound to cytosolic regions of Get1 (Mariappan et al., 2011; Stefer et al., 2011) and of the

'closed' form with AMPPNP (Bozkurt et al., 2009), ADP•$AlF_4$ (Mateja, Szlachcic, et al., 2009), ADP (Bozkurt et al., 2009), with ATP in complex with Get4 (Gristick et al., 2014), ATP and TA-protein ('pre-hydrolysis') (Mateja, Paduch, et al., 2015), or bound to the either the soluble domains or full ER receptors (Kubota et al., 2012; Mariappan et al., 2011; Stefer et al., 2011; McDowell et al., 2020). In all structures, Get3 is a homodimer hinged through a coordinated $Zn^{2+}$ ion liganded by four conserved cysteines. Two distinct structural regions are observed, a well ordered nucleotide binding domain (NBD) and a flexible $\alpha$-helical region that forms the client binding domain (CBD). In the NBD, parts that are characteristic of G-type hydrolases are present (Simpson et al., 2010), a P-loop (formed by a deviant Walker A motif (Suloway, Chartron, et al., 2009)), A-loop, and Switch I & II. In the current model, binding of ATP drives the transition from the 'open' to the 'closed' state causing the $\alpha$-helices in the CBD to rearrange to form a hydrophobic groove that is believed to be the site of TA protein (client) binding (Mateja, Szlachcic, et al., 2009).

Despite this success, a number of additional conformational states of Get3 remain to be determined. A key missing structural state is the Get3/TA protein complex after nucleotide hydrolysis and release from the Get4/Get5 transfer complex for delivery to the ER receptors. This post-hydrolysis structural state would likely require a significant conformational change as evidenced by single-molecule Foster resonance energy transfer (smFRET) studies where Get3 shifted to a significantly lower FRET efficiency state when bound to a TA protein and ADP relative to the 'closed' AMPPNP bound state (Chio, Chung, et al., 2017). It is not possible to infer this key 'post-hydrolysis' state with current structures, leaving the conformational changes that drive Get4/5 release a mystery.

Outside of the Opisthokont supergroup, there has been little characterization of the GET pathway. Some GET components have been identified in *Arabidopsis thaliana* (Xing et al., 2017) and most recently in the apicomplexan *Plasmodium falciparum* (Kumar et al., 2021). In plants, there are four copies of Get3, with the ER targeting Get3 named Get3a and the genomic deletion of this gene resulted in a stunted root growth phenotype. Whereas in the Apicomplexan, *P. falciparum*, knocking out Get3 lead to a sensitivity to $CuSO_4$ (Kumar et al., 2021). These works begin to highlight the conservation of Get3 throughout eukaryotes, but the conservation of mechanism has yet to be demonstrated.

Here, we identify key GET pathway components in the Excavatan *Giardia intesti-*

*nalis* – homologs of Get3, Get4, Sgt2, and Get2. *G. intestinalis* is a single cell protist that causes giardiasis (Cernikova, Faso, and Hehl, 2018) which affects ~30% of the population in developing countries (Feng and Xiao, 2011). We present the first structures of Get3 in this branch of the eukaryotic tree in three functional states and six conformations: *Gi*Get3 in the apo (nucleotide-free), ATP-bound (pre-hydrolysis), and ADP/client-bound (post-hydrolysis) states using either x-ray crystallography or single particle analysis (SPA) cryo-electron microscopy (cryo-EM). This comprehensive structural characterization of Get3 provides the first structural characterization of client and hydrolysis induced conformational changes in Get3, filling in a critical missing piece of the GET pathway story. Together these structures produce the most complete picture to date of the Get3 catalytic cycle, explaining the conformational changes in Get3 necessary for driving the successful targeting of TA proteins.

## 5.2 Results

### Identifying GET pathway components in *Giardia intestinalis*

The predicted Get3 homolog (GL50803_7953) in the parasite *Giardia intestinalis* was captured by an antibody and the protein binding partners were determined by mass spectrometery indentifying additional potential GET components, homologs of Get4 and Get2 (Fig. 5.1*A-C* & S5.2 & S5.4). The sequence identity of *Gi*Get3 to human and yeast Get3 are 42.18% and 44.67%, respectively. Important features of Get3 are conserved in the nucleotide binding site (P-loop, Switch I&II, and A-loop) and the TRC40-insert (Fig. 5.1*G*). Residues that form the Get4 interface are also conserved including critical binding residues (Fig. S5.1) (Gristick et al., 2014). Hydrolase activity of *Gi*Get3 was verified through an ATPase assay and a D54N point mutation corresponding to the inactivating mutation in yeast (Mateja, Szlachcic, et al., 2009; Suloway, Chartron, et al., 2009) has the same effect in *G. intestinalis* (D53N in giardia) (Fig. 5.1*F*). The identified Get4 homolog (GL50803_112893) has a 33.3% identity to yeast Get3 with residues critical for Get3 binding conserved (Fig. S5.2) supporting the capture by mass spectrometry and suggesting a similar mechanism of interaction. The identified Get2 homolog has a 15.1% identity to yeast Get2 with a similar predicted architecture, three transmembrane domains and extended N-terminal tether that contains potential Get3 binding residues (Fig. S5.4) (McDowell et al., 2020). An Sgt2 homolog was identified a structure-based sequence search of the giardia genome. This identified homolog (GL50803_7287) has a 27.2% identity to yeast and contains the three distinct domains for this protein:

dimerization, TPR, and client binding (Fig. S5.3) (Lin et al., 2021). An *in vitro* capture assay, optimized for yeast GET components, demonstrates that *Gi*Sgt2 can capture a yeast TA protein from the yeast Hsp70, Ssa1 (Lin et al., 2021; Cho and Shan, 2018) (Fig. 5.1*E*).

Lastly, we sought to identify potential substrates for this newly identified Get3. The giardia genome was screened for putative TA proteins and several were selected based on hydrophobicity to test for binding to *Gi*Get3. The predicted TMDs of these putative TA proteins were cloned with an N-terminal SUMO-tag and co-expressed in *E. coli* with wild-type Get3. Complexes were purified using affinity chromatography and detected by western blots using anti-*Gi*Get3 and anti-SUMO antibodies (Fig. S5.5). Of the selected TA proteins, clear bands for Get3 were present for two – GL50803_9489 and GL50803_24512, two hypothetical proteins in the GiardiaDB (Aurrecoechea et al., 2008). Calculated hydrophobicities using the transmembrane (TM) tendency scale of these TMDs are 25.76 and 33.05, respectively (Zhao and London, 2006)correlate with the expected hydrophobicity for Get3 binding and ER targeting in yeast (Fry, Saladi, et al., 2021). The identification of these proteins as binders for Get3 suggests these proteins localize to the ER which may aid in identifying their functions.

**Structures of *Gi*Get3 in the 'open' and 'closed' states**

To gain mechanistic insight into Get3, we aimed to structurally characterize *Gi*Get3 in various nucleotide states during the catalytic cycle (Fig. 5.2 & 5.3). The first states we characterized are similar to those previously seen for fungal Get3 homologs in the presence and absence of nucleotide (Suloway, Chartron, et al., 2009; Mateja, Szlachcic, et al., 2009; Bozkurt et al., 2009). *Gi*Get3 was cloned and expressed with an N-terminal 6x His-tag followed by a small ubiquitin-like modifier (SUMO) fusion. *Gi*Get3 was purified by chromatography as a dimer and the tag was removed (Fig. S5.6). We began structural characterization using single particle analysis (SPA) cryo-electron microscopy (cryo-EM) (Fig. S5.7). The apo, or nucleotide-free, *Gi*Get3 was placed on grids and 9,730 movies were collected. Despite starting with $\geq$10 million particles, only 51,340 particles were selected to obtain a 8.43Å map (Fig. S5.7, Table S5.2). The resulting reconstruction gave a model that most resembled a hybrid of the nucleotide-bound closed and 'apo' open states of fungal Get3s (Fig. 5.2*A* & Fig. S5.7). SmFRET data demonstrated that the 'apo' dimer adopts a spectrum of 'open' to 'closed' conformations and the refinement presented here supports this as the model reflects a small subset of particles that were able to

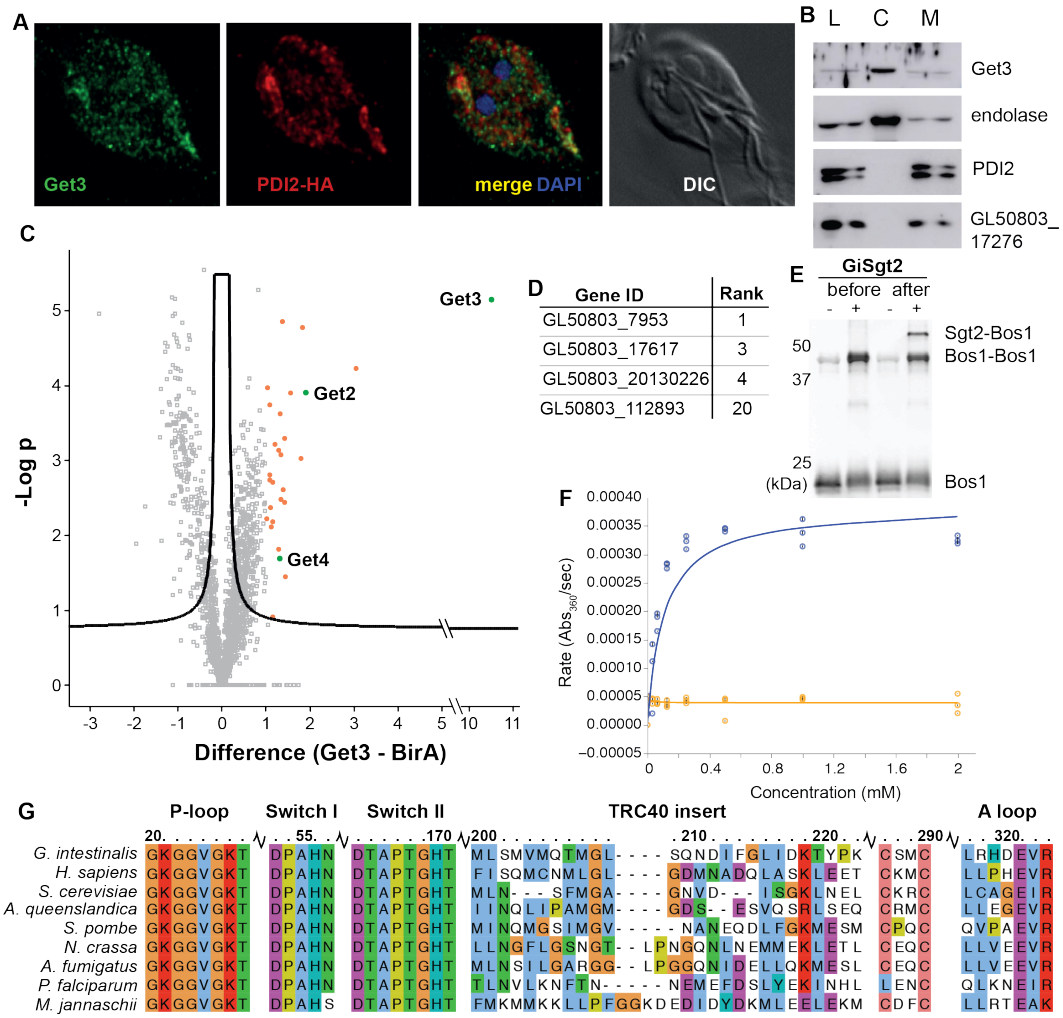



Figure 5.1: **Identification of the GET pathway in *Giardia intestinalis***
*A*) Immunofluorescent images of *G. inestinalis* trophozites to detect the localization of Get3. Get3 was detected by an anti-*Gi*Get3 antibody *green*, the ER membrane is detected by an anti-PDI2 antibody *red*, and the nucleus is stained with DAPI *blue*. Images are merged in the far right panel. *B*) Western-blots of *Gi*Get3 trophozite lysate, cytosol, and membrane fractions for anti-*Gi*Get3, anti-endolase, anti-PDI2, and anti-GL50803_17276 antibodies. *C*) A volcano plot of the mass spectrometry analysis of elution from the native Get3 pull-downs, with the fold-change on the x-axis and significance on the y-axis, both on the log scale. GET components identified are highlighted with green dots. *D*) A table of the proteins identified through mass spectrometry with their ranking sorted by enrichment. *E*) Anti-Strep western blot of UV treated and untreated samples from an *in vitro* transfer assay both before and after transfer, demonstrating *Gi*Sgt2 captures *Sc*Bos1. *F*) ATPase assays with *Gi*Get3 & *Gi*Get3·D53N at nucleotide concentrations of 0mM, 0.031, 0.0625, 0.125, 0.250, 0.5, 1, & 2mM. *G*) Sequence alignments of the conserved regions in Get3 across several eukaryotes (*G. intestinalis*, *H. sapiens*, *S. cerevisiae*, *A. queenslandica*, *S. pombe*, *N. crassa*, *A. fumigatus*, *P. falciparum*, and *M. jannaschii*). Residues are colored by the ClustalX color scheme and numbered by sequence position in *G. intestinalis*.

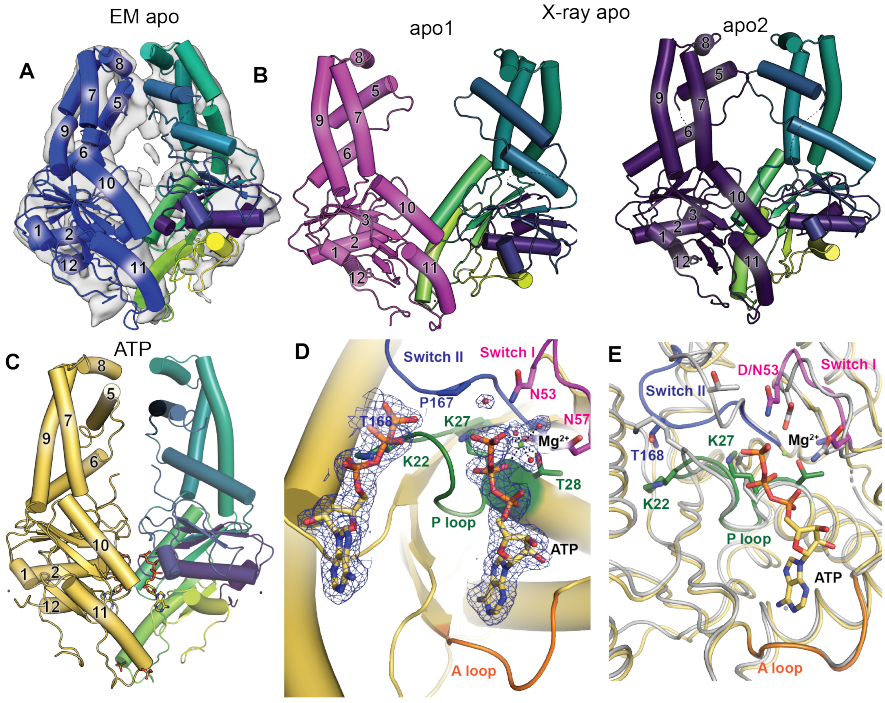



Figure 5.2: **Structures of *Gi*Get3 in the 'open' and 'closed' states.**
*A*) The cryo-EM map of apo *Gi*Get3 with two monomers from the apo1 crystal structure in (*B*) fitted in. *B*) The crystal structures of the two conformations of dimeric apo *Gi*Get3 named apo1 and apo2. *C*) The crystal structure of *Gi*Get3$_{D53N}$ bound to ATP. For all structures, helices are numbered based on the secondary structure prediction order. *D)* A $Mg^{2+}$ ion in the active site of ATP-bound *Gi*Get3 is coordinated by the $\gamma$- and $\beta$-phosphates in the ATP molecule, three waters, and Thr$_2$8. *E*) A comparison of the active site from apo2 (*grey*) and *Gi*Get3 $_{D53N}$·ATP with the P-loop (*green*), A-loop (*orange*), and Switch I (*magenta*) & II (*blue*) highlighted.

be successfully classified in a single conformation.

For a high resolution 'open' state structure, apo *Gi*Get3 was crystallized and the resulting crystals diffracted to 3.0Å in space group P2$_1$2$_1$2. Remarkably, the asymmetric unit contained two separate monomers that generated two different symmetry related dimer conformations, which we refer to as apo1 and apo2 (Fig. 5.2 *B*). The two *Gi*Get3 monomers have 0.66Å rmsd to each other and overall are structurally similar to fungal Get3s. Both apo1 and apo2 form symmetric homodimers that contain the well-ordered nucleotide binding domain (NBD) and the $\alpha$-helical TA client

binding domain (CBD), which consists of helices 4-9 (H4-9). Both are consistent with an 'open' conformation, with apo1 more closely resembling the open conformation seen in fungal structures (PDB:3IBG, 2WOO, 3A36) (Suloway, Chartron, et al., 2009; Mateja, Szlachcic, et al., 2009; Yamagata et al., 2010) (Fig. 5.1*A*, S5.1*A*, & S5.9). The apo2 conformation has not yet been seen and adopts a state that is in between the fungal 'open' and 'closed' forms which we consider to be an intermediate open conformation(Fig. 5.4*A*, *purple*). In this context, the low resolution SPA cryo-EM model can be classified as a closed apo form where the monomer has a similar structure to the yeast apo structures. This range of conformational states is consistent with smFRET data where the apo yeast Get3 sampled a broad range of conformations (Chio, Chung, et al., 2017).

We were able to confidently build the dynamic Helix 5 (H5) which had been seen in various fungal apo (PDIB:2WOO & 3A36) and 'closed' structures in different conformations (Mateja, Szlachcic, et al., 2009; Yamagata et al., 2010; Mateja, Paduch, et al., 2015; Gristick et al., 2014) (Fig. S5.9 & S5.11). Here, the amphipathic H5 in giardia packs under H8 and against H7 & 9 to shield the hydrophobic residues residing in these helices and resulting in an overall hydrophilic surface that masks the putative hydrophobic groove for TA protein binding. This positioning of H5 most closely reflects that of the structure of apo *S. cerevisiae* Get3 (PDBID:3A36), where H5 is shown to pack against H7 & H9, but H8 could not be built (Fig. S5.9). While H8 has been modeled in in two fungal 'open' structures, their positioning differs from what is seen here; in one case H8 acts as a linker between H7 & H9 (PDBID:3IBG) and in the other, H8 covers the hydrophobic residues in H9 of *Sc*Get3 (PDBID: 2WOO) (Fig. S5.9). Here, the giardia apo structure reveals H8 forms a cap that links H7 and H9 and then stabilizes the buried hydrophobic interface with H5, completing the exposed hydrophilic surface. The active site resembles that of the yeast structures where the P loop, A loop, Switch I, and Switch II all adopt similar conformations (Fig. S5.8*B*) (Suloway, Chartron, et al., 2009).

To to obtain the ATP-bound *Gi*Get3 structure, we purified a non-hydrolyzing mutant, D53N, which we demonstrated to be catalytically inactive (Fig. S5.6*B* & 5.1*F*). Crystal trays were set in the presence of ATP and $MgCl_2$, resulting in crystals that diffracted to 2.23Å in the space group $P3_221$ with a monomer in the asymmetric unit that formed a symmetry related dimer in a 'closed' conformation (Fig. 5.2*C*). In the active site, a $Mg^{2+}$ ion is coordinated by the $\gamma$- and $\beta$-phosphates in ATP, three waters, and $Thr_{28}$. A water molecule is coordinated by the asparagine residue that

replaced the catalytic aspartate above the $\gamma$-phosphate of the ATP molecule, primed for nucleophilic attack. This structure is the first of Get3 bound to ATP alone, *i.e.* with or without any client or binding partners, and adopts a 'closed' form with the two monomers rotating closer together relative to the apo-hybrid state. Excitingly, all but one stretch (residues 88-114) can be modeled in this structure, providing the most complete structure of a 'closed' Get3. Previous prediction that H8 forms a lid in the TA-bound 'closed' state to cover the hydrophobic groove to protect TMDs of TA proteins is not supported by our structure (Mateja, Szlachcic, et al., 2009; Mateja, Paduch, et al., 2015). Instead, in the 'closed' ATP alone structure, the H8 participate in stabilizing the chamber in the absence of substrate (Fig. 5.2). This is consistent with biochemical work that demonstrated the deletion of H8 did not affect the targeting of TA proteins by Get3 to the ER and only decreased loading efficiency of TA proteins onto Get3. Likely, H8 regulates opening of the hydrophobic groove to allow TA protein capture. Contrary to yeast 'closed' Get3 structures where H5 is partially modeled to be parallel to H7 & 9 , this structure reveals that H5 remains sitting against these two helices as in the apo structures, occluding the binding pocket.

The rest of the helices that make up the CBD are in similar orientations, in particular H6 is in an upward position, away from the ATP molecule. In fungal structures of Get3 such as the the transition state bound to ADP·$AlF_4$ or to ATP in complex with other GET components, H6 is seen shifted down towards the nucleotide binding pocket, forming the bottom of the hydrophobic groove (Mateja, Szlachcic, et al., 2009; Gristick et al., 2014; Mateja, Paduch, et al., 2015). This shift is not observed in our ATP-bound form nor in the AMPPNP-bound *Ct*Get3 crystal structure. It is likely that ATP binding alone is not sufficient to cause H6 to shift likely required for H5 to rearrange from being nearly perpendicular to parallel to H7 & 9. In the state here, ATP primes Get3 for Get4 binding and likely stabilizes the groove in preparation for client capture. The previous fungal structures where H6 has shifted down likely represent the next step which requires client or Get4 binding to generate the transition state for hydrolysis. Combined, these structures reveal that *Gi*Get3 adopts conformational states similar to the Opisthokont Get3s and have identical mechanisms.

**Single particle cryo-EM analysis of *Gi*Get3/TA protein complexes**

We next sought to fill in the key missing piece of the Get3 ATPase cycle by structurally characterizing Get3 bound to TA protein in the post-hydrolysis state using

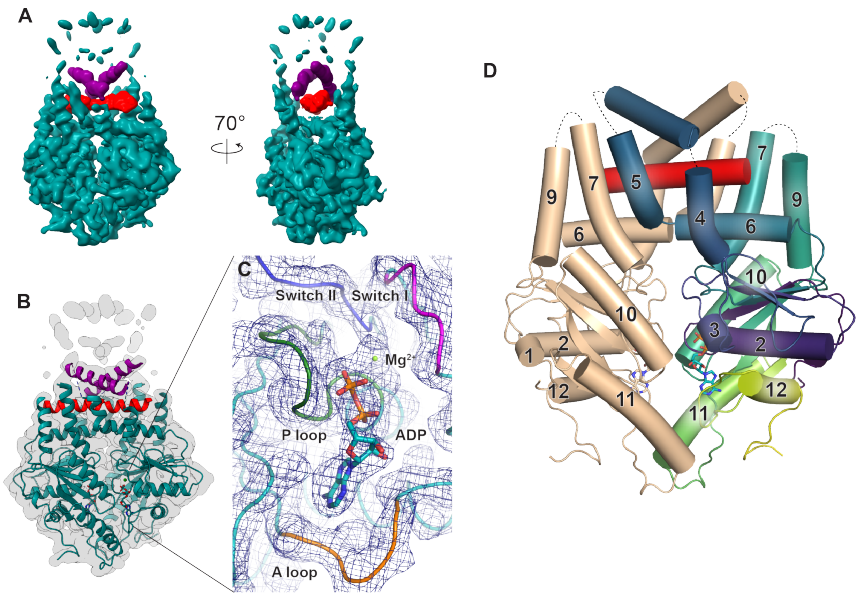



Figure 5.3: **Cryo-EM structure of *Gi*Get3/TA complex bound to ADP**
*A)* Two views of the unsharpened overall map of *Gi*Get3 in the post-hydrolysis state shown in Fig. S5.14. Assigned density in the hydrophobic groove corresponding to the TMD (*red*) and H4/5 (*purple*) are colored. *B)* The sharpened map after a focused refinement shown in Fig. S5.14 with the molecular model of *Gi*Get3 fitted in. A close up of the density in the nucleotide binding pocket, with the model fitted it and key features are labeled and an ADP and $Mg^{2+}$ ion are fitted into the density. *D)* Molecular model of *Gi*Get3 in the 'intermediate' state, represented as a cartoon colored from N- to C-terminus using the viridis color map (*purple to yellow*). The TMD from the TA protein is colored in red.

SPA cryo-EM. To accomplish this, wild-type *Gi*Get3 was recombinantly expressed with a His-tagged client, (BRIL·Bos1$_{TMD}$), that had previously been used to investigate the binding mechanism of Sgt2 (Lin et al., 2021) (Fig. S5.13). The complex was purified using nickle affinity and size exclusion chromatography and the elution fraction correlating to 150kDa was frozen on grids and used to determined a reconstruction with an overall resolution of 4.1 Å and a focused refined resolution of the NBD and partial CBD at an average resolution of 3.7 Å (Fig. 5.3*A&B* & Fig. S5.14). In the focused refinement map there is clear density for all of Get3 except for the C-terminus of H7, all of H8, and the N-terminus of H9 which were masked. At this resolution (ranging from 3.5-7.5 Å), we were able to fully build the NBD as well as the sections of the CBD that form the hydrophobic groove as seen in the pre-hydrolysis structure (Mateja, Paduch, et al., 2015). Three new helical densities were observed in the CBD that did not correspond to helices previously seen in

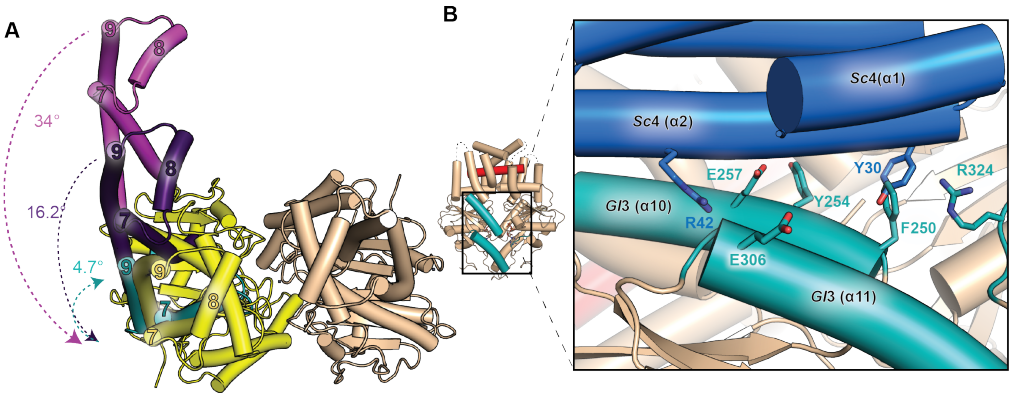



Figure 5.4: **Comparison of *Gi*Get3 in the 'closed', 'intermediate', and 'open' states**
*A)* A top view of the twist opening of Get3. Structures of nucleotide-free, ATP- and post-hydrolysis *Gi*Get3 are aligned by the P-loop in Chain A and only Chain A from the 'closed' state of Get3 is shown (*wheat*). For clarity only helices 7-9 are shown and colored as they are in 5.2 & 5.4. Rotations of chain B are annotation by arrows and degrees measured from the P-loop. *B*) *Gi*Get3/TA·ADP (*wheat*) with the binding interface of Get4 colored in teal. A zoomed in image of the post-hydrolysis complex docked into the *Sc*Get3/4/5 structure (*blue*), demonstrating a clash due to the movement in H10.

other Get3 structures highlighted in *red* and *purple* in Fig. 5.3). These helices were placed as poly-alanine C$\alpha$s as they could not be unambiguously assigned in the density. Density colored in red coincides with the positioning of the TMD of the TA protein Pep12 in the pre-hydrolysis complex form, suggesting this density represents the Bos1 TMD. For the other two helices, only the region between H4 and H5, loop H4/5, was able to account for the density colored in purple. Clear density for an ADP molecule and $Mg^{+2}$ ion are visible in the nucleotide binding pocket revealing that this map reflects the post-hydrolysis form of the Get3/TA complex (Fig. 5.3*C*). No nucleotide was added during the purification process indicating that the *Gi*Get3·TA complex binds tightly to ADP until disruption by the receptor.

The post-hydrolysis structure reveals a number of new general features. It adopts a slightly more open conformation in agreement with the intermediate state observed from smFRET data (Chio, Chung, et al., 2017) (Fig. 5.4 *A*). The sequence that makes up H4 as part of the TA binding groove is disordered in both the apo and ATP state and then becomes structured in the post-hydrolysis state. H6 is seen in the lowered position in the post-hydrolysis structure pulling H5 down and away from its position

protecting the sides of the groove. H5 then joins H4, H7, and H9 in forming the walls of the hydrophobic groove. Relative to the transition state structure, H4 shifts away from the center of the groove to accommodate the change in H6, resulting in an expansion of the CBD (Fig. 5.5*A*). The most striking new feature is that loop H4/5 becomes helical and docks on top of the client TMD completing the hydrophobic chamber and protecting the entire hydrophobic client TMD from solvent (Fig. 5.3 & 5.5*B & C*).

**Conformational changes induced by nucleotide and client binding and hydrolysis**

Together these crystal and cryo-EM structures provide a comprehensive view of the catalytic cycle of Get3, supplying the conformational changes regulated by nucleotide binding and hydrolysis (Fig. 5.4 & 5.5). Apo Get3 adopts a range of conformations swinging between fully open and closed. Upon ATP binding, the 'closed' Get3 state is stabilized generating the binding interface for Get4 (Fig. 5.4*A & B*), as the 'closed' apo Get3 is different from this ATP bound conformation (Fig. 5.2 & S5.16). ATP binding stabilizes Get3 into a fully closed interaction. A possible contributing factor of this stabilization is a conserved cation-pi interaction across the dimer interface, a $Phe_{250}$ and $Arg_{324}$ (Fig. S5.12*C* & S5.17). These interactions are known to stabilize protein interfaces (Salonen, Ellermann, and Diederich, 2011). Also observed in fungal 'closed' Get3 structures (PDBID:2WOJ, 4PWX, 4XTR), the cation-pi is absent in 'open' structures (both fungal and giardia) (Fig. S5.12*C & D*) (Mateja, Szlachcic, et al., 2009; Gristick et al., 2014; Mateja, Paduch, et al., 2015). $Phe_{250}$ and $Arg_{324}$ are conserved across Get3s and mutating these to Ala in yeast resulted in a phenotypic growth defect and a decrease in Get4 binding (Gristick et al., 2014; Suloway, Chartron, et al., 2009).

Nucleotide binding drives significant remodeling within each monomer. When transitioning from the apo to ATP-bound state, Switch I moves towards the nucleotide, placing the catalytic $Asp_{53}$ above the $\gamma$-phosphate in the ATP molecule to position the water for nucleophilic attack (Fig. 5.2*D* & 5.5*B*). The backbone of the P-loop interacts with the $\alpha$- and $\beta$-phosphates of the ATP molecule, wrapping around the nucleotide and $Lys_{27}$ rotates towards the nucleotide, interacting with the $\beta$-phosphate. Switch II moves towards H10 and the 'deviant Walker A' $Lys_{22}$ in the P-loop shifts out to interact with the $\beta$-phosphate of the ATP molecule across the dimer interface. These stabilized interactions generate the Get4 binding site, priming Get3 for the transfer complex. These changes in the active site are also

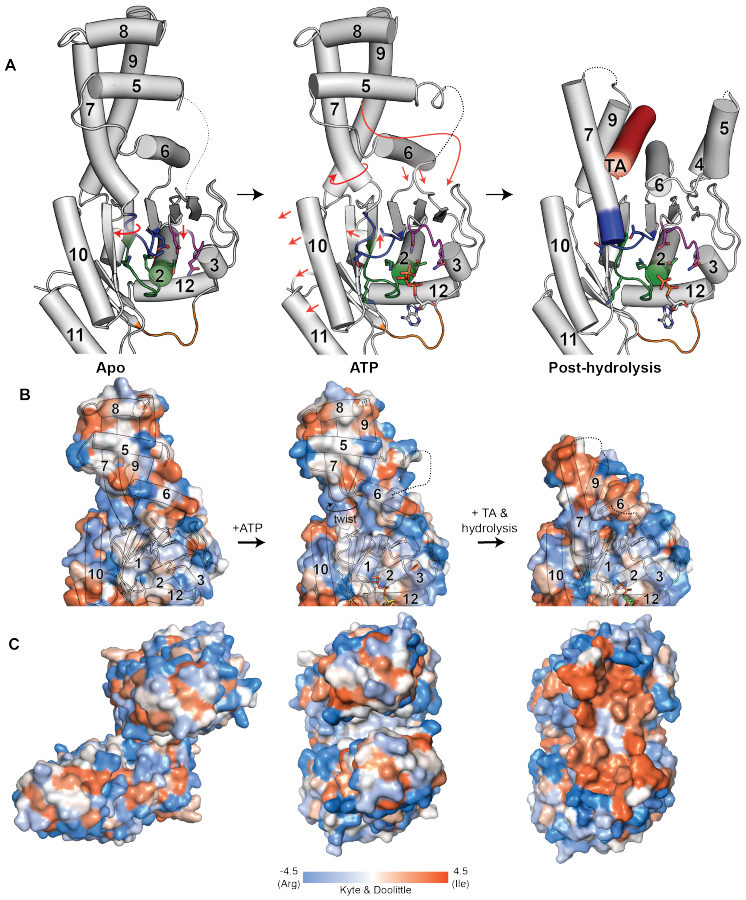



Figure 5.5: **Conformational changes induced by TA protein binding and hydrolysis stabilize a hydrophobic groove**
The nucleotide binding cycle of *Gi*Get3 in the apo (*left*), ATP-bound (*center*), and post-hydrolysis (*right*) states. *A*) a cartoon representation of these states with the P-loop, A-loop, and Switch I & II colored as they are in Fig. 5.2. Conformational changes between states are highlighted with red arrows. Get3 monomers are colored in grey and the TMD of the TA protein is colored in red. *B)* A surface representation of the CBD of apo1, ATP-bound, and post-hydrolysis TA bound Get3. For clarity, a cartoon representation is outlined and helices are numbered appropriately. Changes in H7 between the pre- and post-hydrolysis form are indicated by a black arrow. *C)* A bird's eye view of the hydrophilic (apo1, ATP-bound) and hydrophobic (post-hydrolysis) states. Hydrophobicity is colored from blue (*hydrophilic*) to orange (*hydrophobic*) based on the Kyte & Doolittle hydrophobicity scale (Kyte and Doolittle, 1982).

observed in yeast Get3, suggesting a conservation in the mechanistic function of Get3 in yeast and giardia (Mateja, Szlachcic, et al., 2009).

Additional specific changes are observed in the post-hydrolysis state. First, the Get3 dimer rotates to a more open conformation in between the apo2 and the ATP-bound state (Fig. 5.4*A*) and in this intermediate state the cation-pi interaction between $Phe_{250}$ and $Arg_{324}$ is preserved (Fig. 5.4*B*, & S5.17*C*). In the active site, the most significant change is in Switch II where $Pro_{166}$ moves away from the nucleotide allowing the C-terminus of this sequence to transition from a loop into a helix resulting in a twist in H7 and the formation of a new N-terminus for the helix (Fig. 5.5*A*). This twist in H7 organizes the hydrophobic residues in the helix to turn inwards towards the center of the TA protein-binding groove, creating the hydrophobic interior (Fig. 5.5). This helical transition of Switch II results in H10 moving away from the nucleotide and accommodates the downward movement of H6. H4 shifts away from the active site avoiding a clash with H6 and, together with H10 movements, results in a swelling of the Get3 monomer. As the cation-pi and other interactions across the dimer interface are preserved, the majority of the changes in the intermediate conformation are a consequence of the rearrangements to Switch II and the resulting shift in H10 (Fig. 5.4*B*). While the Get3/Get4 binding interface has been explored biochemically and structurally, how this interface is disrupted by client recognition was unknown. Here, the observed shift in H10 and slight opening of the dimer explains the disruption of the Get4 binding interface. Aligning a Get3 monomer between the post-hydrolysis Get3/TA complex and the yeast Get3/4/5 crystal structure shows that the Get3 H10 movement generates clashes with residues on H2 of Get4 (Fig. 5.4*B*). These changes do not disrupt the hydrophobic groove seen in the 'closed' and post-hydrolysis form, clarifying how TA proteins remain bound to Get3 after Get4 disassociation (Fig. 5.5).

These structures allow a more complete model of how Get3 successfully targets TA proteins to emerge. In the apo state, Get3 fluctuates between the closed and open form as highlighted by our three apo structures. The hydrophobic surfaces of the CBD are hidden by H5 with H4 in a disordered loop. Nucleotide binding results in a stabilized closed complex that includes two symmetrical cation-pi interactions that bridge the dimers. This 'closed' Get3 can bind to Get4 in preparation for client capture. Get4 binding likely induces the conformational changes in Get3 to prime it for TA protein capture including the movement of H5 and H4 to become parallel

with H7 and H9 forming the hydrophobic walls of the client binding pocket and H8 becoming disordered. Evidence suggests that H8 plays an important role in client loading (Chio, Chung, et al., 2019) and movement of H8 in the Get3/Get4 complex may drive the release of H5 leading to the other CBD conformational changes. H6 must shift down to drive catalysis and the lower orientation is most likely stabilized by a correct client. This binding would result in ATP hydrolysis and then phosphate release. Get3 disassociates from Get4 through conformational changes induced by changes in the active site that are transmitted to the Get3 surface. In this form, Get3 remains bounds to TA proteins completely shielding the hydrophobic TMD with the helix formed from loop H4/5.

## 5.3 Conclusion

The work presented here is the most completed picture to date of Get3 from a single organism. Despite the evolutionary distance, the GET pathway shows remarkable conservation across eukaryotes highlighting the importance of this essential process. *In vitro*, alone Get3 cannot capture a TA substrate and requires Get4 (F. Wang, Brown, et al., 2010). The fully masked TA-binding groove in the apo and ATP state would prevent inadvertent association with non-specific targets. After release from Get4, the Get3/TA protein complex would still be capable of binding to Get2 at the ER as the membrane bound receptor does not bind across the dimer interface (Mariappan et al., 2011). Once localized, the released H8 may then bind the membrane destabilizing the complex so that Get3 can be opened and release the client which is favored due to Get1 binding an open Get3. These structures demonstrate that Get3 is a dynamic protein, sampling many different conformational states during TA protein targeting. All together, this work provides detailed mechanistic insight into nucleotide regulation of the GET pathway. Remaining questions for the pathway are how Get4 inhibits nucleotide hydrolysis (Rome, Rao, et al., 2013), how Sgt2 hands off to Get3 in a privileged interaction (Shao, Rodrigo-Brenni, et al., 2017), and how the Get1/Get2 complex facilitates TA insertion. The human pathogen *Giardia intestinalis* has proven to be a beneficial model system for TA targeting. As the GET pathway has proven to be a drug target (Morgens et al., 2019), further studies in this organism will lead to even deeper insight.

## 5.4 Materials and Methods

*Sequence alignments* Alignments of Get3, Get4, and Sgt2 were created by downloading genes from *G. intestinalis*, *H. sapiens*, *S. cerevisiae*, *A. queenslandica*, *S.*

*pombe*, *N. crassa*, *A. fumigata*, *M. jannaschii*, and *P. falciparum* from Uniprot. Sequences were aligned with PROMALS3D (Pei, Kim, and Grishin, 2008) along with all experimentally determined structures of of Get3, Get4, and Sgt2 homologs. PROMALS3D provides a way of integrating a variety of costs into the alignment procedure, including 3D structure, secondary structure predictions, and known homologous positions. Alignments were visualized using Jalview (Waterhouse et al., 2009).

*Immunofluorescence microscopy*

*G. intestinalis* trophozoites were fixed in 1% paraformaldehyde for 30min at 37°C, collected by centrifugation at 1000×g for 5 min, washed in PEM buffer (100mM PIPES pH 6.9, 1mM EGTA, and 0.1mM MgSO4) and placed on cover slips. The cells were permeabilized by 0.2% Triton X-100 for 20 min, washed three times with PEM buffer, and incubated with primary antibodies in PEMBALG (100mM PIPES pH 6.9, 1mM EGTA, 0.1mM MgSO4, 1% BSA, 0.1% NaN3, 100mM lysine, and 0.5% cold-water fish skin gelatin) for 1 hr. Cells were probed with the following primary antibodies: rat anti-*Gi*Get3 polyclonal antibody (1:100 dilution) and rabbit anti-HA tag antibody for PDI2 (1:1000 dilution). The cover slips were washed three times with 1mL of PEM buffer and then incubated with appropriate secondary antibodies: Alexa Fluor™ 488-conjugated goat anti-rat IgG and Alexa FLuor™ 594-conjugated donkey anti-rabbit IgG (Invitrogen, Waltham, MA). After three 5 min washes in PEM buffer, slides were mounted with Vectashield containing DAPI (4‘, 6-diamidino-2-phenylindole; Vector Laboratories, Burlingame, CA).

*Cell culture, cloning, and transfection in G. intestinalis*

The *Giardia intestinalis* strain WB (ATCC 3095) was grown in TYI-S-33 medium supplemented with 10% heat-inactivated bovine serum, 0.1% bovine bile and antibiotics at 37°C (Keister, 1983). Gene encoding Get3 homolog (GL50803_7953) were amplified from genomic DNA and inserted to the pOndra plasmid (Dolezal et al., 2005) with C-terminal BAP-tag. $1 \times 10^7$ cells expressing cytosolic BirA (biotin ligase)(Martincová et al., 2015) were electroporated with a Bio-Rad Gene Pulser (Hercules, CA) using an exponential protocol (U=30V; C=1,000$\mu$F; R = 750$\Omega$). The transfected cells were grown in medium supplemented with antibiotics (57 $\mu$g/ml puromycin and 600 $\mu$g/ml G418).

*Fractionation of G. intestinalis cells*

*G. intestinalis* cells were harvested in cold Phosphate-Buffered Saline (PBS), pH 7.4

by centrifugation 1000×g at 4°C for 10 min, washed with 20mM MOPS, 250mM sucrose, pH 7.4, and again collected by centrifugation. The pellet was resuspended in 20mM MOPS, 250mM sucrose, pH 7.4 supplemented with protease inhibitors (cOmplete™ Protease Inhibitor Cocktail, Roche, Switzerland). The cells were lysed on ice by sonication for 2 min (1 sec pulses, 40% amplitude). The lysate was subjected to centrifugation at 2680×g and 4°C, for 20 min to sediment nuclei, cytoskeleton, and remaining unbroken cells. The supernatant was subjected to centrifugation at 180,000×g, for 30 min at 4°C. The resulting supernatant corresponded to cytosolic fraction, and the high-speed pellet (HSP) contained organelles including mitosomes and the endoplasmic reticulum.

*Native pull-down assays of Get3*

*G. intestinalis* cells were grown in TYI-S-33 medium with 50$\mu$M biotin for 24 hours before harvesting. The cell lysate was diluted to final concentration of 1 mg/ml in PBS (pH7.4) supplemented with protease inhibitors and incubated with 50$\mu$L of streptavidin-coupled magnetic beads (Dynabeads MyOne Streptavidin C1, Invitrogen, Waltham, MA) for 1 hr at 4°C with gently rotation. Isolation was made in quadruplicates, where each sample contain 5 mg of proteins. The magnetic beads were washed 3 times in 50mM HEPES pH 7.4, 150mM potassium acetate, 5mM magnesium acetate, 1mM DTT, and 10% glycerol and 3 times in PBS. Beads with bound proteins were submitted to mass spectrometry analysis.

*Protein cloning, expression, and purification in Opisthokonts*

Full-length *Gi*Get3 used to form Get3/TA complexes was cloned into a pET28a vector. For experiments using only *Gi*Get3 and single point mutants, the gene was cloned in a pET28a vector that was modified to have an N-terminal His-tag and SUMO fusion protein. To create the TA protein variant used for cryo-EM experiments, a BRIL fusion protein followed by the TMD of *Sc*Bos1 was cloned into a pACYC-Duet vector modified to contain a TEV cleavage site between an N-terminal His-tag and the fusion protein, a thermostabilized apocytochrome b562 (BRIL) (Chun et al., 2012). The TMD of *Gi*TA proteins were cloned into a pET21b vector and flanked by an N-terminal 3xStrep-tag followed by a modified non-cleavable SUMO fusion protein and a C-terminal opsin tag. *Gi*Get3 and mutants were expressed in *E. coli* nicoDE3 cells in 2xYT at 37 °C and induced by 0.3mM IPTG at an $OD_{600}$ ~0.6. Cells were harvested 4 hours after induction. Cells were disrupted in lysis buffer (50mM Tris pH 7.5, 300mM NaCl, 20mM imidazole, and 10mM $\beta$-ME) supplemented with protease inhibitors, 1mM PMSF and 1mM

benzamidine using a M-110 Microfluidizer Processor (Microfluidics). Lysate was centrifuged to separate the soluble and membrane fractions. Protein was purified by batch incubation with NiNTA resin at 4C for 1.5 hours. Resin was washed with lysis buffer and *Gi*Get3 and mutants were eluted in 50mM Tris, 300mM NaCl, 300mM imidazole, and 10mM $\beta$-ME, pH 7.5. The affinity tag was removed from the elution collected after nickel chromatography by an overnight ULP1 digestion in dialysis buffer (20mM Tris, 150mM NaCl, and 10mM $\beta$-ME). The dialyzed fraction flowed over a nickel column again to remove the His-SUMO particles and *Gi*Get3 was collected in the flow-through. This pool was further purified using size-exclusion chromatography (SEC) through a HiLoad 16/600 Superdex 200 (GE). Protein purified for crystallography purposes were purified in 10mM Tris, 75mM NaCl and 10mM $\beta$-ME while protein used in ATPase assays was purified in 50mM HEPES, 150mM KAc, and 10mM $\beta$-ME.

Get3/TA complexes were formed by co-expressing TA proteins with tag-less *Gi*Get3 in *E. coli* nicoDE3 cells in 2xYT at 37C and induced by 0.5mM IPTG at an $OD_600$ ~0.7. Cells were lysed in 50mM Tris, 300mM NaCl, and 10mM $\beta$-ME, pH 7.5 supplemented with 1mM PMSF and benzamidine using a M-110 Microfluidizer Processor (Microfluidics). The lysis was separated by centrifugation. Complexes used for structural determination via cryo-EM were purified by incubating the soluble fraction with NiNTA resin at 4C for 1.5 hours. Resin was washed with 50mM Tris, 300mM NaCl, 35mM imidazole, and 10mM $\beta$-ME, pH 7.5 and protein was eluted in 50mM Tris, 300mM imidazole, and 10mM $\beta$-ME, pH 7.5. The elutate was further purified via SEC using a HiLoad 16/600 Superdex 200 (GE) column.

*Gi*Get3/TA complexes used for pull-down experiments were purified by affinity chromatography using Strep-tactin resin. The resin was washed with lysis buffer and complexes were eluted in lysis buffer plus 2.5 mM destihiobiotin.

Full-length *Gi*Sgt2 was cloned into a pET33b vector that was modified to have an N-terminal His-tag and TEV cleavage site. *Gi*Sgt2 was expressed in *E. coli* nicoDE3 cells in 2xYT at 37°C and induced by 0.3mM IPTG at and $OD_{600}$ ~0.6. Cells were harvested 4 hours after inductions and disrupted in lysis buffer (50mM Tris pH 7.5, 300mM NaCl, 10mM imidazole, and 10mM $\beta$-ME) supplemented with protease inhibitors, 1mM PMSF and 1mM benzamidine using a M-110 Microfluidizer Processor (Microfluidics). Lysis was centrifuged to separate the soluble and membrane fractions. Protein was purified by batch incubation with NiNTA resin at 4C for 1.5 hours. Resin was washed with wash buffer (20mM Tris pH7.5, 150mM NaCl,

20mM imidazole, and 10mM $\beta$ME and protein was eluted in 20mM Tris, 150mM NaCl, 300mM imidazole, and 10mM $\beta$-ME, pH 7.5.

*Crystallization*

Purified *Gi*Get3 or *Gi*Get3-D57N were concentrated to 10-12mg/mL and crystal trays were set using the hanging-drop vapor-diffusion method by equilibriation of equal volumes of protein and well liquor in VDX plates with sealant (Hamptons). Wild-type *Gi*Get3 crystals formed in 0.1M MES pH 5.3, 0.1M $MgCl_2$, and 21% PEG3350 at 4°C. *Gi*Get3-D57N was incubated with 5mM ATP and 2mM $MgCl_2$ on ice for 1 hour prior to setting trays and crystals formed in 0.1M Tris pH 7.5, 0.2M ammonium sulfate, and 15% PEG3350 at room temperature. Crystals were cryo-protected by transfer into 30$\mu$L of well liquor supplemented with 2mM $MgCl_2$ and, in the case of *Gi*Get3-D57N, 5mM ATP and increasing amounts of glycerol (10%, 15%, and 20%). Crystals were incubated in each cryoprotectant drop for <5 minutes before flash freezing in liquid nitrogen.

*Data collection, structure determination, and refinement*

Structures were solved from data collected on the 12-2 beamline at SSRL at 12.6keV. Structures were solved using a single data set and the *Gi*Get3 structure was integrated and scaled using XDS and the *Gi*Get3-D57N dataset was integrated and scaled using HKL3000. The wild-type crystal diffracted to 3.0Åand the mutant crystal diffracted to 2.23Å. Both structures were solved using molecular replacement in PHENIX, using the monomer of yeast Get3 in the open state (PDBID: 3IBG). Sequences were adjusted using Sculptor. Refinement was performed using Refmac5. Manual building was done in COOT (residues 105-126 and 193-210) and the final refinement was done in PHENIX to an Rfactor of 0.27 (Rfree 0.34) and 0.17 (Rfree 0.21) for wild-type and mutant respectively. Final refinements resulted in 97% (wild-type) and 98% (mutant) Ramachandrian favored (see Table S5.1).

*Cryo-EM grid preparation and data collection*

*Gi*Get3/His-BRIL-Bos1$_{TMD}$ complexes taken immediately after elution from SEC at a concentration of ~0.73mg/mL. 3$\mu$L of sample was placed on Holey carbon grids (Quantifoil R1.2/1.3, 300 mesh) that were glow discharged in air with a 20A plasma current for 2 minutes using a Pelco easiGlow, Emeritech K100X. Grids were blotted at a force of 10 for 3.5 seconds and frozen in liquid ethane with the chamber at 4°C and 100% humidity using a FEI Vitrobot Mark v4 x2. Data was collected using an automated data collection program, SerialEM, on a FEI Titan Krios equipped with

an energy filter (20eV slit width) at 300keV and a Gatan K3 direct detector. Beam illumination was adjusted to a fluence of 13 $e^-$/upix/Å. Images were collected using a defocus range of -0.7 to -3.0$\mu$m using super resolution mode at a calibrated pixel size of 0.433Å/pixel. Using counting mode, 1.82 second images were collected with a frame rate of 45.5ms and dosage of 1.58$e^-$/Å/frame.

A purified apo *Gi*Get3 sample was taken immediately after elution and diluted to ~0.55mg/mL. 3$\mu$L of sample was placed on Holey carbon grids (Quantifoil R2/2, 100 mesh NH2 Finders), which were treated in the same manner as the grid prep for the complex. Data was collected using SerialEM on a FEI titan Krios equipped with an energy filter (GIF) at 300keV and a Gatan K3 direct electron detector. Images were collected using a defocus range of -0.5 to -2.5$\mu$m using super resolution mode at a calibrate pixel size of 0.5295Å/pix. In counting mode a total of 50 frames were collected for a total dosage of 50$e^-$/$AA^2$.

*Image processing*

For the *Gi*Get3/TA complex dataset, 2,732 movies were initially processed using cryosparc v.3.2.0 to produce aligned dose-weighted micrographs. During motion correction, movies were down-sampled to a corrected pixel size of 0.866Å/pix and all downstream processing is done at this pixel size. Of the 2732 movies, 2356 were manually selected for further processing. A small set of particles were manually picked and used for template based picking and manually filtered to remove obvious debris, resulting in 1,790,962 particles. An initial round of 2D classification was used for further particle filtration, resulting in 555,998 particles. Four *ab initio* models were generated using cryosparc and two classes were consistent with the expected shape and size of Get3 (a total of 362,614 particles). Several rounds of 3D heterogeneous refinement were carried out to produce a class of 156,446 particles. 2D templates were generated using these class of particles and these templates were used for template picking. The 1,561,353 picked particles were extracted at a 4x bin, followed by 3D heterogeneous refinement to filter out bad particles and good particles were re-extracted at a 2x bin. These 803,265 particles were subjected to a 3D heterogeneous refinement again. Two models had similar levels of detail and shape. The 568,836 particles that result in these two models were re-extracted with no binning and underwent a round of 3D heterogeneous refinement again. This resulted in two classes with high resemblance (338,011 particles).

These particles were exported using the cryosparc2star.py program and imported into RELION 3.1.2. Particles underwent a round of 3D homogeneous refinement

and local CTF refinement. It became clear that there was weaker density above the NBD, which could suggest lower order partially due to the expected flexible BRIL. We applied a soft mask of 6 pixels extended by 4 to the particles, isolating the NBD using particle subtraction. These particles underwent a round of 3D classification into 4 different classes. One class, 70,330 particles, with the most detail refined to 3.86Å. C2 symmetry was imposed and the map refined to 3.72Å. Post-processing was performed with a soft mask of 6 pixels extended by 4 and the B-factor was estimated by RELION. Local resolution was estimated using RELION's own implementation. The disordered region could not be refined.

For the *Gi*Get3 apo dataset, 9,300 movies were first processed in cryosparc v.3.2.0, producing aligned dose-weighted micrographs. Movies were down-sampled to a corrected pixel size of 1.059Å/pix and all future process was done at this pixel size. A subset of 7,607 movies were manually selected for particle picking. A small set of particles were manually picked and used to create 2D templates for template based picking. Particles picked were filtered to remove debris, resulting in 11,596,225 particles that were then filtered using 2D classification resulting in 552,716 particles. Four *ab initio* models were generated and one class was consistent with the expected shape and size of Get3 (174,3012 particles). Several rounds of 3D heterogeneous refinement were carried out to produce a class of 74,013 particles. 2D templates were then generated uses these particles and the resulting templates were used for template picking, resulting in 17,238,072 pick particles. These particles were then extracted at 4 times bin, followed by 3D heterogeneous refinement to filter out bad particles, particles belonging to classes that resembled Get3 were then re-extracted at 2x bin. These 7,599,636 particles were filtered by 3D heterogeneous refinement and particles in 3D classes that resembled Get3 were again re-extracted, this time without any binning. These particles were filtered using several rounds of 3D heterogeneous refinement and the resulting 580,912 particles underwent homogeneous refinement.

These particles were exported using the cryosparc2star.py program and imported into RELION 3.1.2. Particles underwent several rounds of 3D classification into six different classes. Five classes resembled Get3 and the combined 51,340 particles refined to 8.46Å. Post-processing was performed with a soft mask of 6 pixels extended by 4 and the B-factor was again estimated by RELION. Local resolution was estimated using RELION's own implementation.

*Model building into the cryo-EM map*

For the *Gi*Get3/TA complexes, using *phenix.dock_in_map* two molecules were searched for in the map using the monomer of *Sc*Get3 from PDIB:5BW8 as a model. The *G. intestinalis* sequence was then imposed using *phenix.sculptor*. Manual model building was conducted in COOT and the final model was ran through *phenix.real_space_refinement*. Poly-alanine sequences were built into the three helical densities in the CBD, but could not be ambiguously assigned are denoted as UNK the deposited structure.

Chain A from the apo *Gi*Get3 crystal structure was used as a search model in the apo *Gi*Get3 map using *phenix.dock_in_map*. Two molecules were found to fit. The resulting model was then refined to fit the map using the FLEX-EM function in CCPEM (Topf et al., 2008).

*ATPase assays*

ATPase assays were carried out using EnzChek®Phosphate Assay Kit (Thermo Fisher, Waltham, MA). Assays were carried out with 5.03$\mu$M of *Gi*Get3 or 4.51$\mu$M of *Gi*Get3·D53N in a buffer of 50mM HEPES, 150mM Potassium Acetate, 5mM Magnesium Acetate, and 10mM $\beta$ME, pH 7.5 at 37°C. The reaction mixture were incubated in 96 well plates (Corning Costar Assay Plate) at 37°C prior to initiating the reaction with ATP at concentrations of 0$\mu$M, 37.25$\mu$M, 62.5$\mu$M, 125$\mu$M, 250$\mu$M, 500$\mu$M, 1mM, and 2mM. Measurements were taken by a Tecan Infinite M Nano+ at an Abs=360nm every 20 sections for a total of 10min. This method was programmed using Magellan 7.2 software. Data was analyzed using IceKat.

*In vitro capture assays*

The in vitro transfer assays were performed as in previous reports (Chio, Chung, et al., 2019; Shao, Rodrigo-Brenni, et al., 2017). Specifically, 39$\mu$M Bos1·BPA (50mM HEPES, 300mM NaCl, 0.05% LDAO, 20% glycerol) was diluted to a final concentration of 0.1$\mu$M and added to 4$\mu$M Ssa1 supplemented with 2mM ATP (25mM HEPES pH7.5, 150mM KOAc). After one minute, 0.3µM of full-length *Gi*Sgt2 or mutant was added to the reaction. Samples were flash frozen after one minute and placed under a 365nm UV lamp for 2 hours on dry ice to allow for BPA crosslinking.

*Visualization of GiGet3 and GiTA proteins using western blots*

*Gi*Get3/*Gi*TA protein complexes were run on a 12.5% SDS-PAGE gel. The gels were blotted onto 0.45$\mu$m nitrocellulose membranes (BioRad, USA) which were then cut in half and blocked for 1 hour with a 5% dry milk in TTBS buffer. Then the

higher mw half was incubated with an anti-Get3 anti-body and the lower mw half was incubated with an anti-SUMO anti-body at 4°C for 4 hours. Blots were rinsed with TTBS and then incubated with a secondary anti-body (anti-rat or anti-rabbit) conjugated to an $IR_{680}$ fluorophore. The presence of *Gi*Get3 and *Gi*TA proteins were visualized by imaging the blots at a wavelength of 680nm.

**Acknowledgements**

We thank Andrey Malyutin, Songye Chen, Harry Scott, and Jens Kaiser for technical assistance and Shu-ou Shan, Rebecca Voorhees, and Gabriel Lander for discussion and comments. We are grateful to Ailiena Maggiolo for her expertise for crystallography data refinement and Vanessa Mechem for her help with protein purification. Crystallography data was collected at the Stanford Synchrotron Radiation Lightsource (SSRL) beamline 12-2. We are grateful to G. Moore and B. Moore for support of the Molecular Observatory at the California Institute of Technology. SSRL operations are supported by the US Department of Energy and US National Institutes of Health (NIH). Cryo-EM data sets were collected at the Caltech Cryo-EM facility and Portland National Cryo-EM Center. Data was processed using the Extreme Science and Engineering Discovery Environment (XSEDE) resources, which is supported by the National Science Foundation Grant Acl-1052574(108).

## 5.5 Supplementary Data

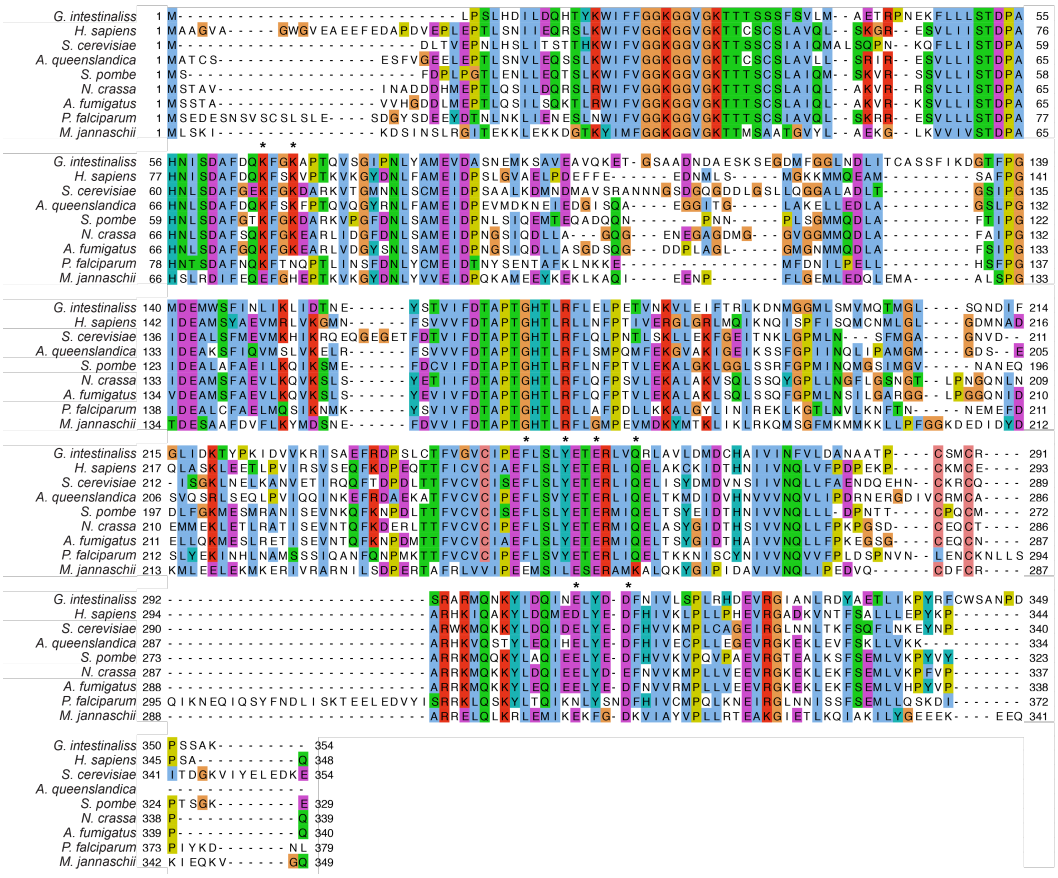



Figure S5.1: **Alignment of Get3**
The full alignment of Get3 partially shown in Fig. 5.1. Conserved residues that were demonstrated to play a role in Get4 binding are highlighted with asterisks above and residues are colored using the ClustalX color scheme.

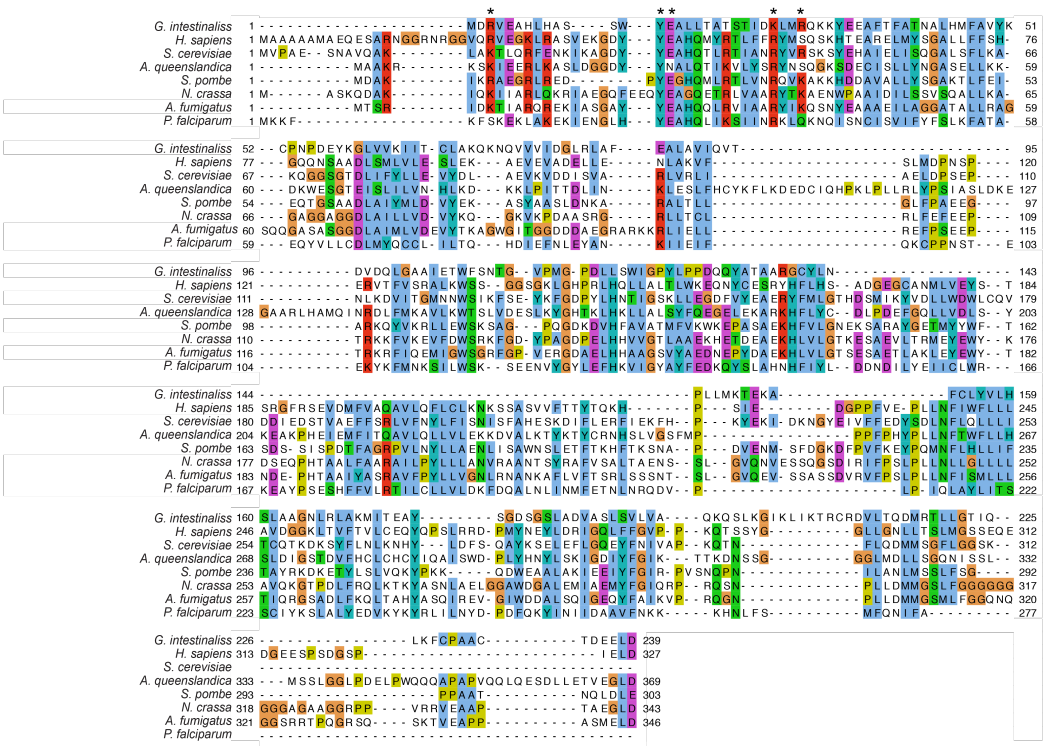



Figure S5.2: **Alignment of Get4**
An alignment of Get4 from *G. intestinalis*, *H. sapiens*, *S. cerevisiae*, *A. queenslandica*, *S. pombe*, *N. crassa*, *A. fumigatus*, and *P. falciparum*. Conserved residues that were demonstrated to play a role in Get3 binding are highlighted with asterisks above and residues are colored the same as in Fig. S5.1

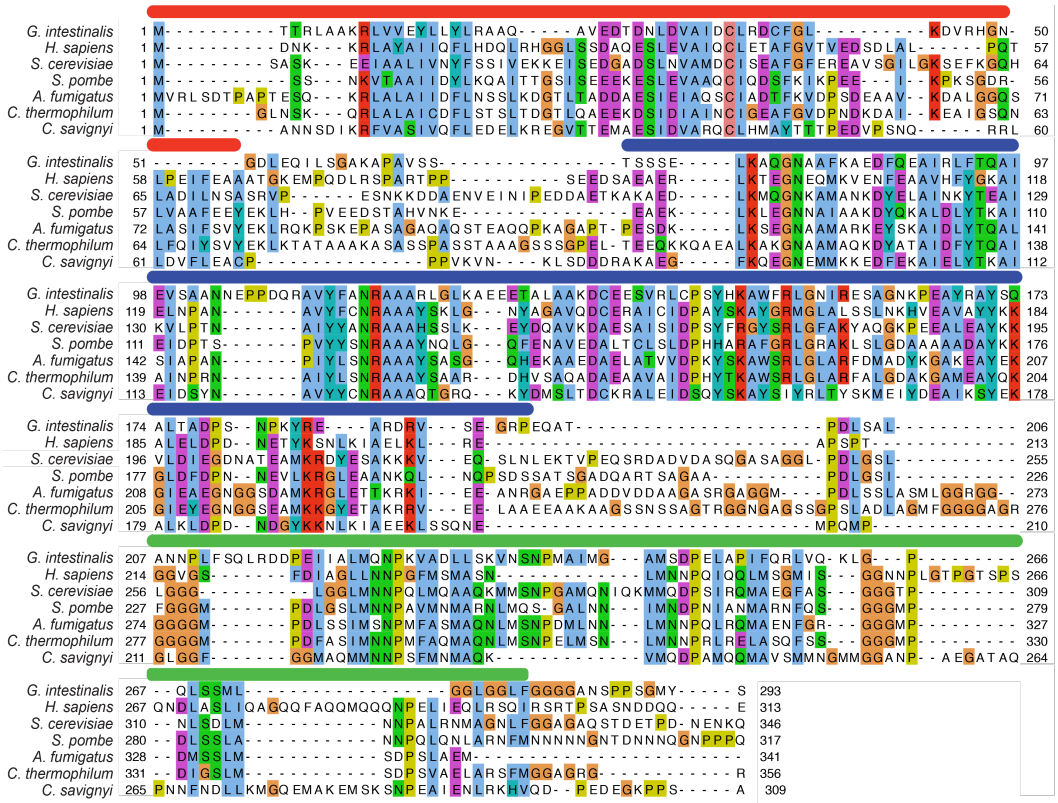



Figure S5.3: **Alignment of identified Sgt2**
An alignment of the identified Sgt2 from *G. intestinalis*, *H. sapiens*, *S. cerevisiae*, *S. pombe*, *A. fumigatus*, *C. thermophium*, and *C. savignyi*. The three domains, N-terminal dimerization (*red*), TPR-domain (*blue*), and substrate binding C-domain (*green*), in Sgt2 are highlighted by cylinders above the alignment. Residues are colored using the ClustalX color scheme.

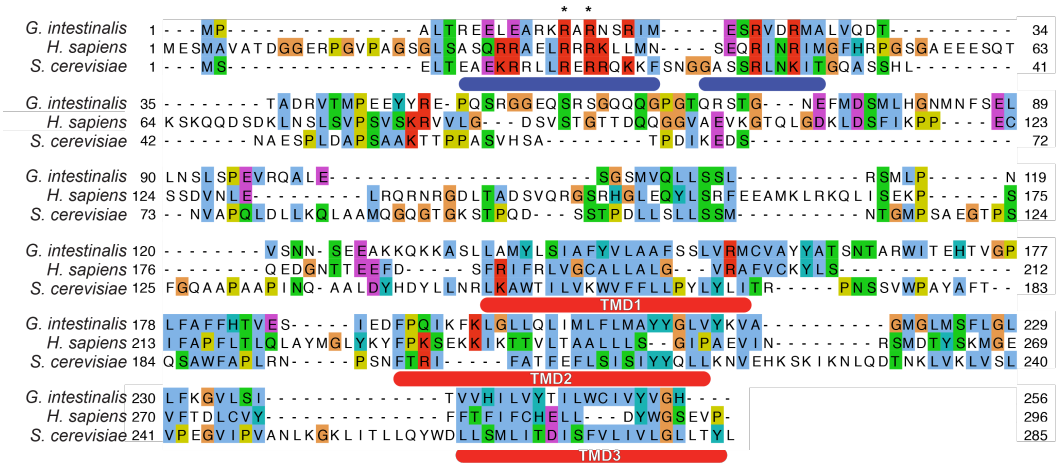



Figure S5.4: **Alignment of identified Get2**
An alignment of the identified Get2 from *G. intestinalis*, *H. sapiens*, and *S. cerevisiae*. The predicted TMDs are highlighted with red cylinders below the alignment and the conserved N-terminal tethers are highlighted with blue. Conserved residues involved in Get3 binding are marked by asterisks above the alignment and residues are colored using the ClustalX color scheme and predicted TMDs are annotated with red bars below.

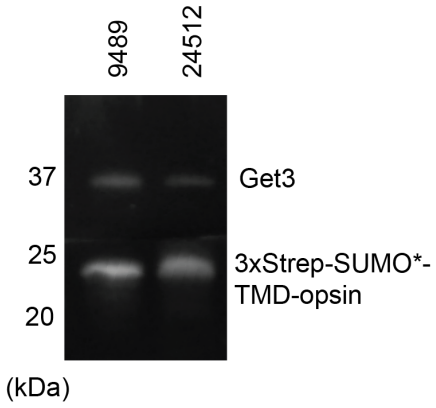



Figure S5.5: **Identification of several TA proteins from *G. intestinalis***
*A)* A schematic of experimental set up. *Gi*Get3-TA complexes were recombinantly purified from *E. coli* through a nickel pull-down on the His-tagged TA protein. *B)* A anti-*Gi*Get3 (top) and an anti-strep (bottom) western blots of the eluate in *(A)*.

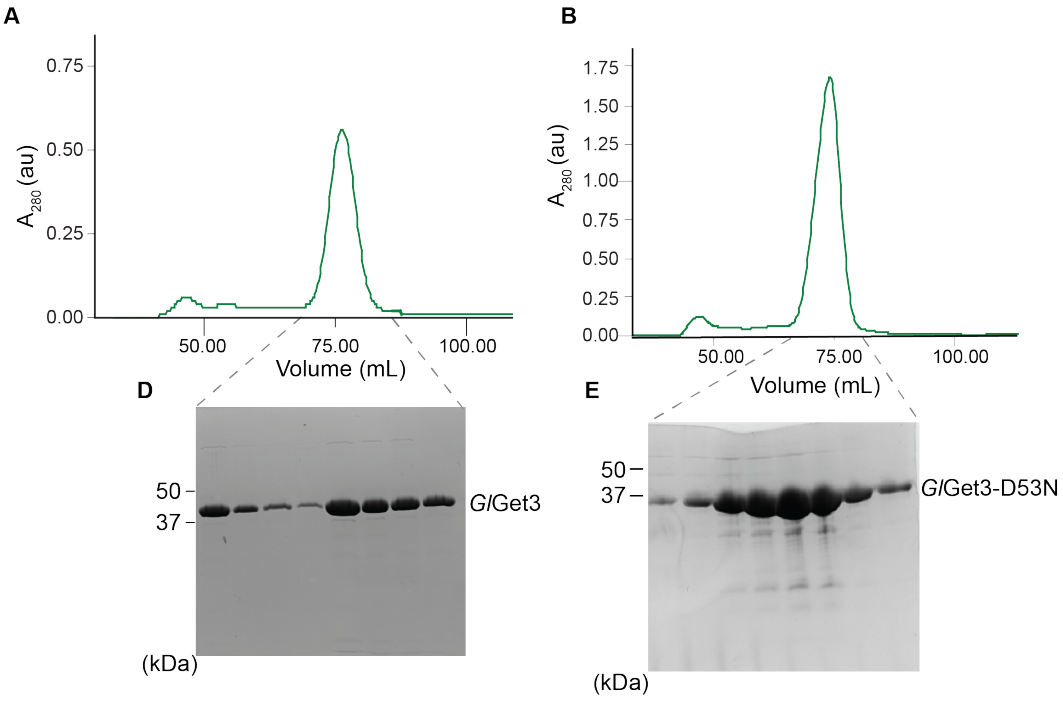



Figure S5.6: **Purification of *Gi*Get3 and *Gi*Get3$_{D53N}$**
Size exclusion chromatograms of nickel eluate of *A)* Get3 and *B)* Get3$_{D53N}$. SDS-PAGE gels of the respective peaks highlighted in *A) & B)* for the *(C)Gi*Get3 and *D)* the non-hydrolyzing mutant.

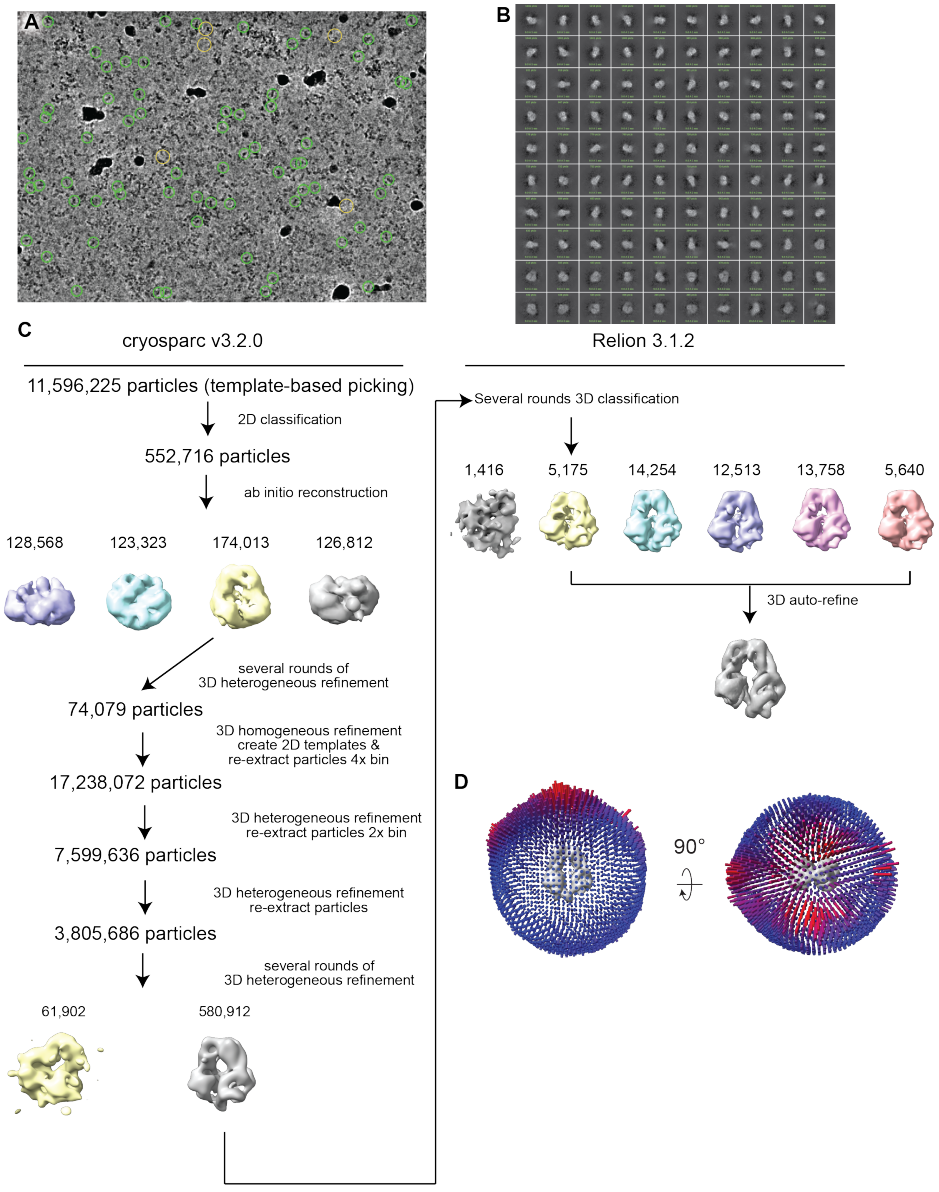



Figure S5.7: **Data processing of apo *Gi*Get3**
*A)* An aligned micrograph from the data collection with sample particles selected with yellow circles. *B)* 2D class averages of particles used for the reconstruction. *C)* Processing of data through cryosparc v3.2.0 and RELION 3.1.2. *D*) Two views of the angular distribution of particles. Particle concentration is displayed by color and length (blue to red).

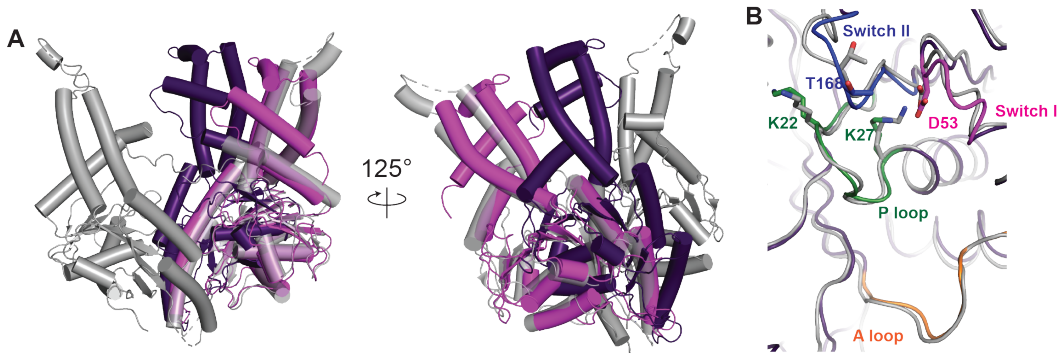



Figure S5.8: **Comparison of Apo *Gi*Get3 and *Af*Get3**
*A)* Two views of *Gi*Get3 apo$_1$ (*magenta*) and apo$_2$ (*purple*) aligned with chain A of *Af*Get3 (*grey*, PDBID:3IBG). *B)* A comparison of the active sites of *Gi*Get3 apo$_2$ (*purple*) and *Af*Get3 (*grey*). The P loop, A loop, and Switch I & II are colored as in Fig. 5.2.

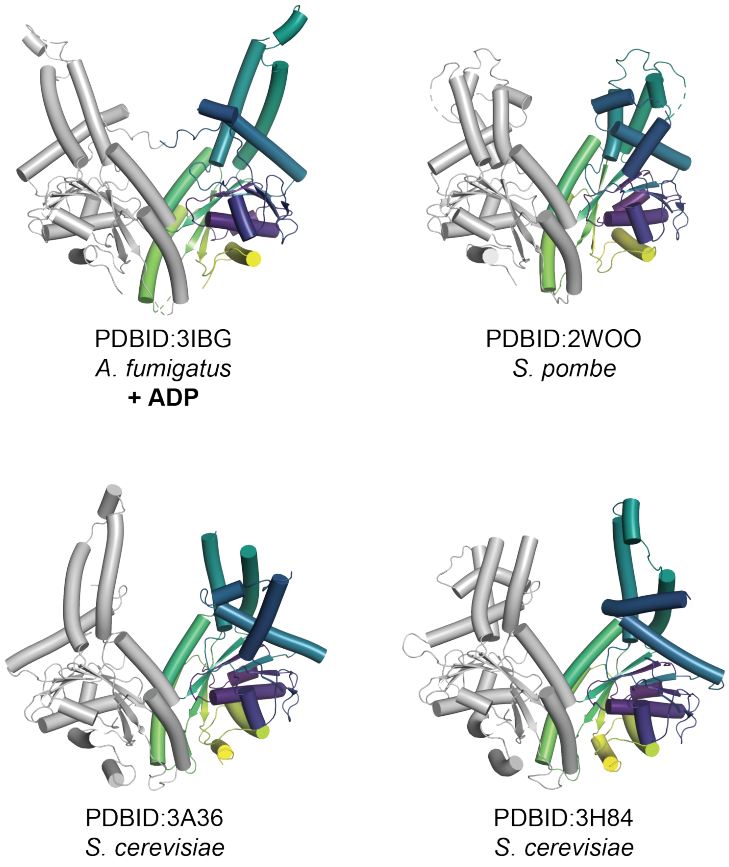



Figure S5.9: **Published structures of Opisthokont Get3s in the 'open' state.**
Cartoon representation of fungal (PDBIDs:3IBG 2WOO, 3A36, & 3H84) Get3 in the open state. Chain A is colored in grey and Chain B is colored from N- to C-terminus using the viridis color map (*purple to yellow*). Species and ligands are specified below the PDBID numbers.

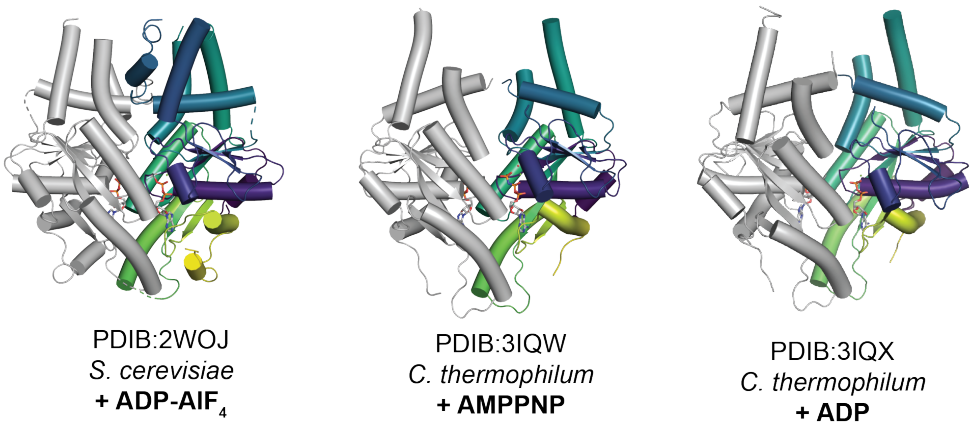



Figure S5.10: **Published structures of Opisthokont Get3s in the 'closed' state.** Cartoon representation of fungal (PDBIDs:2WOJ, 3IQW, 3IQX, & 3VLC) Get3s in the open state. Chain A is colored in grey and Chain B is colored from N- to C-terminus using the viridis color map (*purple to yellow*). Species and ligands are specified below the PDBID numbers.

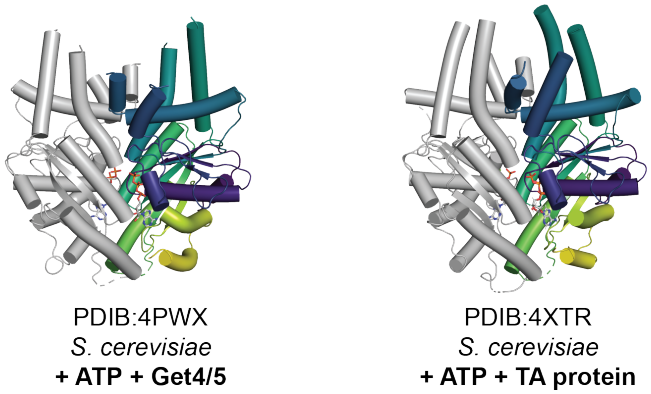



Figure S5.11: **Published structures of Opisthokont Get3s in complexes in the 'closed' state.**
Cartoon representation of fungal Get3 in complex with Get4/5 (PDBIDs:4PWX) and Get3·ATP/TA protein complex the 'closed' state (Gristick et al., 2014; Mateja, Paduch, et al., 2015). Chain A is colored in grey and Chain B is colored from N- to C-terminus using the viridis color map (*purple to yellow*). Species, binding partners, and ligands are specified below the PDBID numbers.

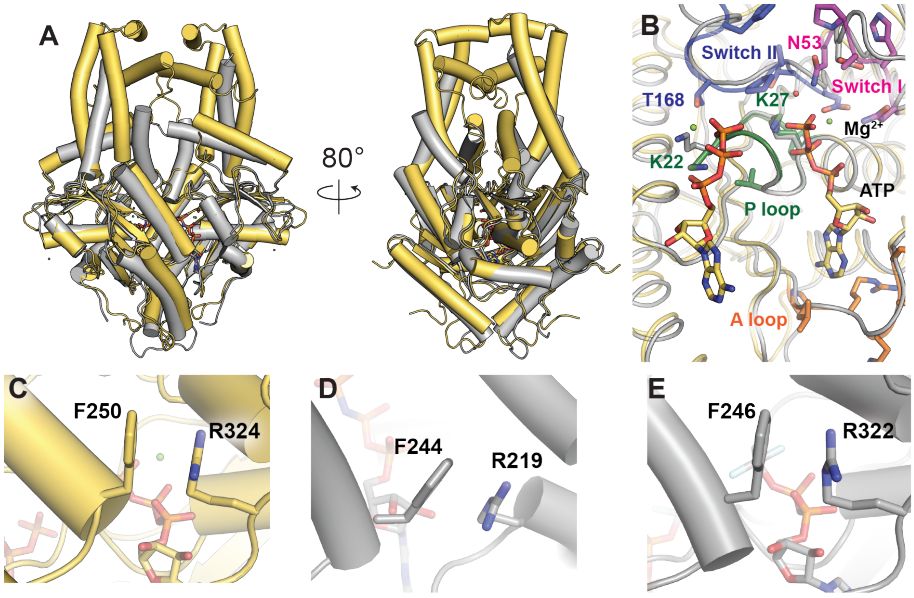



Figure S5.12: **Comparison of ATP-bound *Gi*Get3 to AMPPNP-bound *Ct*Get3**
*A)* Two views of *Gi*Get3$_{D53N}$·ATP (*yellow*) overlayed on structure of *Ct*Get3·AMPPNP (*grey*), aligned by chain A. *B)* A comparison of the active sites of *Gi*Get3 (*yellow*) and *Ct*Get3 (*grey*). The P loop, A loop, and Switch I & II are colored as in Fig. 5.2. The cation-pi stacking of the Phe in H10 and Arg across the dimer interface in structures of *C)* *Gi*Get3·D53N with ATP, *D)* *Ct*Get3·AMPPNP (PDBID:3IQW), and *E)* *Sc*Get3·ADP-AlF$^4$ (PDBID:2WOJ).

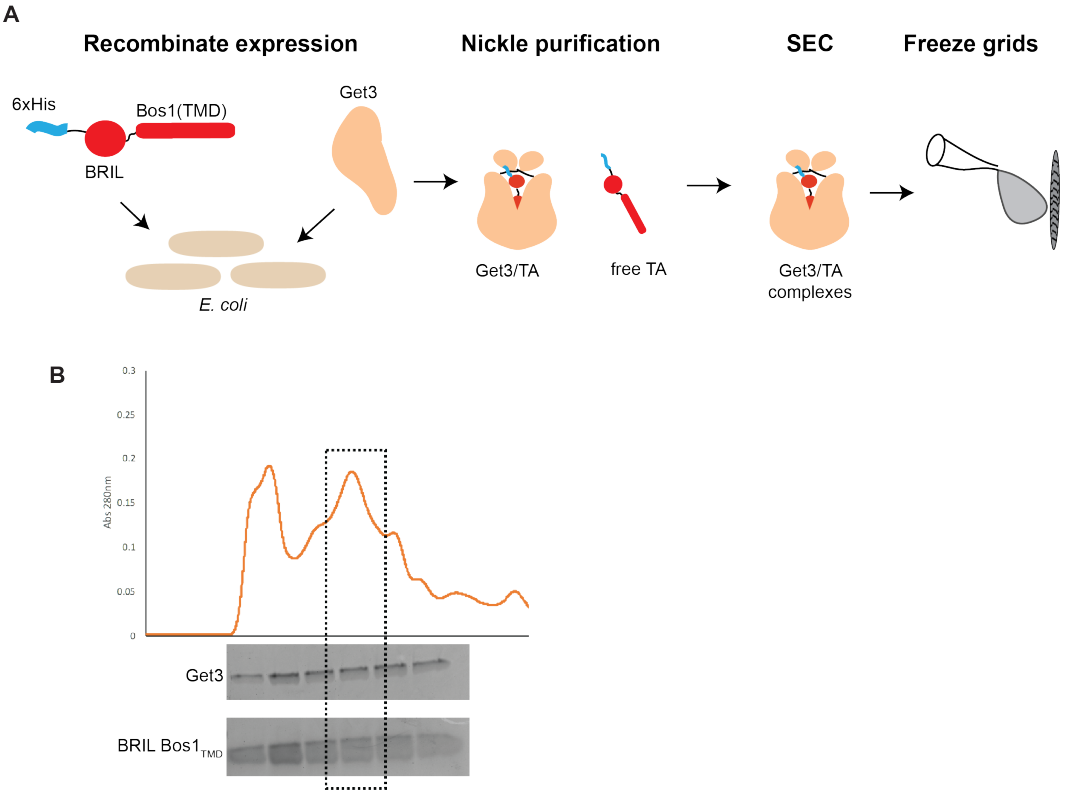



Figure S5.13: **Purification of *Gi*Get3·TA complexes**
*A)* Schematic of the purification of *Gi*Get3 and BRIL·Bos1$_{TMD}$. *B)* Size exclusion chromatogram of nickel eluate of Get3/TA complexes.

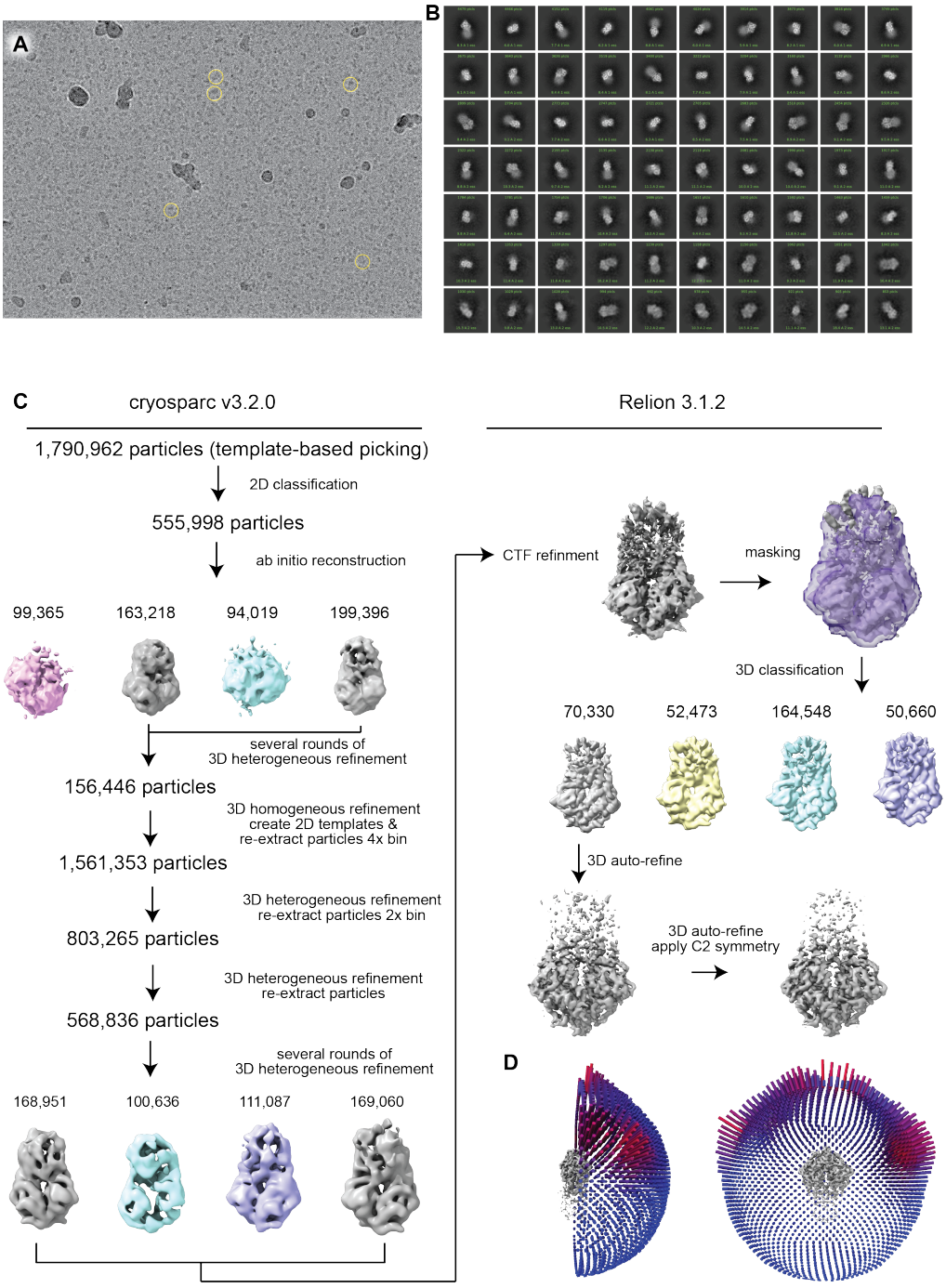



Figure S5.14: **Data processing of *Gi*Get3/TA complexes**
*A)* An aligned micrograph from the data collection with sample particles selected with yellow circles. *B)* 2D class averages of particles used for the reconstruction. *C)* Processing of data through cryosparc v3.2.0 and RELION 3.1.2. *D*) Two views of the angular distribution of particles. Particle concentration is displayed by color and length (*blue to red*).

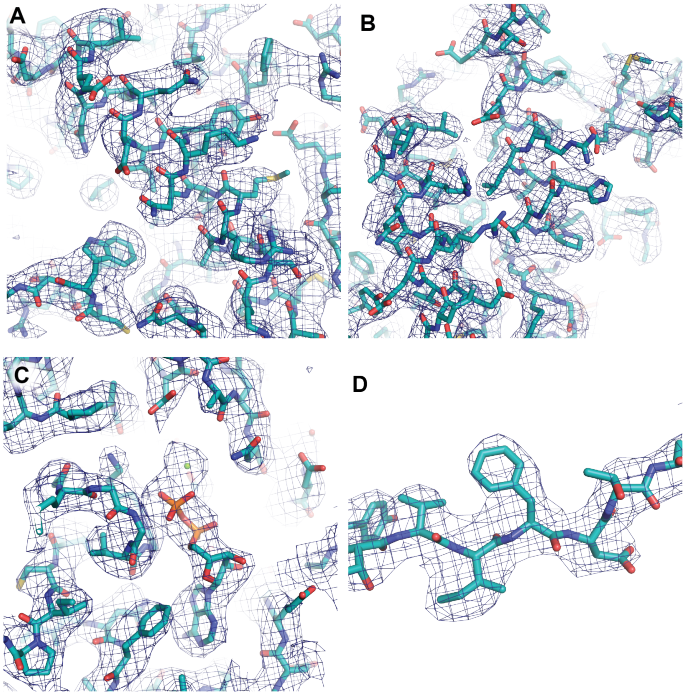



Figure S5.15: **Representative density of *Gi*Get3/TA**
Representative density of *A)* H11, *B)* H7 & H9, *C* active site with ADP molecule and $Mg^{2+}$ ion, and *D* $\beta$-sheet 1. The loops in the active site are colored as in Fig. 5.2

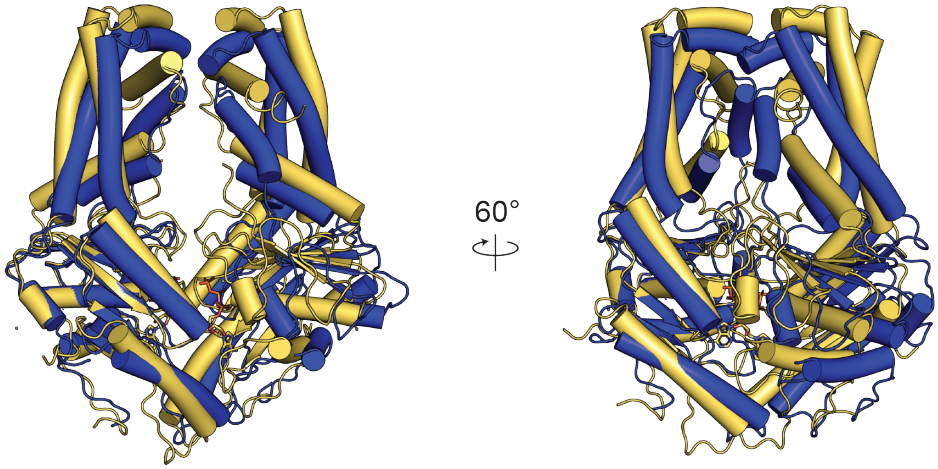



Figure S5.16: **Comparison of ATP-bound and the closed apo conformations of *Gi*Get3**
Two views of an overlay of the *Gi*Get3$_{D53N}$·ATP (*yellow*) and 'closed' apo *Gi*Get3*blue*.

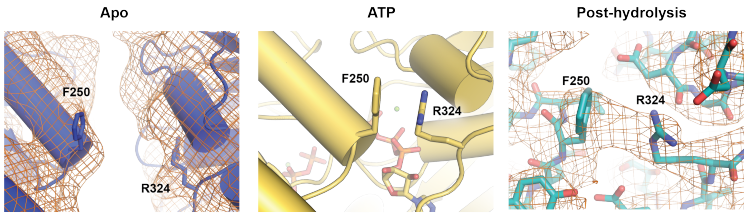



Figure S5.17: **A cation-pi interaction stabilizes the 'closed' conformation**
A close up view of the $Phe_{250}$ and $Arg_{324}$ (shown as sticks) involved in this interaction for the *A*) closed apo form where the two residues are too far apart to form this stabilizing interaction, *B*) ATP-bound and *C*) post-hydrolysis forms where the two residues are close enough to form a cation-pi interaction.

Table S5.1: Crystallography statistics.

| | Apo | ATP |
|---|---|---|
| **Data collection** | | |
| Space group | $P2_12_12$ | $P3_221$ |
| Cell dimensions | | |
| *a, b, c (Å)* | 54.1, 101.9, 137.8 | 81.0, 81.0, 130.1 |
| $\alpha, \beta, \gamma$ (°) | 90.0, 90.0, 90.0 | 90.0, 90.0, 120.0 |
| Resolution (Å) | 39.3-3.0 (3.18-3.0) | 50.0-2.23 (2.32-2.23) |
| Wavelength (Å) | 0.97946 | 0.97946 |
| $R_{merge}$ | 0.112 (0.106) | 0.108 (1.132) |
| $R_{pim}$ | 0.036 (0.405) | 0.035 (0.453) |
| I/$\sigma$ | 12.77 (1.69) | 22.55 (1.75) |
| Completeness (%) | 90.9 (72.2) | 97.5 (79.6) |
| Redundancy (%) | 8.7 (6.6) | 8.9 (5.5) |
| **Refinement** | | |
| Resolution (Å) | 39.3-3.0 | 30.5-2.23 |
| No. reflections | 14,501 | 23,290 |
| $R_{work}/R_{free}$ (%) | 27.5/34.4 | 17.7/20.9 |
| No. atoms | | |
| Protein | 2893 | 2557 |
| Ligand/ion | 2 | 39 |
| Solvent | | 190 |
| B-factors | | |
| Protein | 102.12 | 41.20 |
| Ligand/ion | 127.72 | 25.03 |
| Bond RMSD | | |
| Lengths (Å) | 0.01 | 0.01 |
| Angles (°) | 1.63 | 1.92 |
| **Validation** | | |
| MolProbity score | 1.37 | 1.61 |
| Clashscore | 2.56 | 3.71 |
| rotamer outliers (%) | 0 | 3.48 |
| C$\beta$ outliers (%) | 0 | 1 |
| Ramachandran outliers (%) | 0 | 0 |

*Values in parentheses are for the highest-resolution shell

Table S5.2: Cryo-EM statistics.

| | Apo | Get3/TA (client & H45) | Get3/TA (Get3 only) |
|---|---|---|---|
| **Data collection & processing** | | | |
| Microscope | FEI Titan Krios | FEI Titan Krios | |
| Voltage (kV) | 300 | 300 | |
| Camera | Gatan K3 | Gatan K3 | |
| Energy filter | Gatan Imaging Filter(GIF) Quantum | BioQuantum | |
| Energy filter slit width (eV) | 20 | 20 | |
| Magnification (nominal) | | 105,000 | |
| Defocus range ($\mu$m) | -0.5 to -2.5 | -0.7 to -3.0 | |
| Calibrated pixel size (Å/pix) | 0.5295 | 0.433 | |
| Electron exposure ($e^-/Å^2$) | 50 | 63.2 | |
| Exposure rate ($e^-/Å^2$/frame) | 1 | 1.58 | |
| Number of frames per movie | 50 | 40 | |
| Automation software | SerialEM | SerialEM | |
| Number of micrographs | 9,300 | 2,732 | |
| Initial particle images (no.) | 11,596,225 | 1,790,962 | |
| Final particle images (no.) | 51,340 | 70,330 | 70,330 |
| Estimated accuracy of translations (pix) (RELION) | 1.076 | 2.143 | 1.066 |
| Estimated accuracy of rotations (°) (RELION) | 3.355 | 3.461 | 3.302 |
| Local resolution range | 6.0-8.45 | 4.15-7.36 | 3.61-7.06 |
| Map resolution (Å, FSC=0.143) | 8.46 | 6.26 | 3.72 |
| Model fitting | | | |
| Software (*FLEX-EM*) | CCPEM 1.5.0 | | |
| **Refinement** | | | |
| Software (*phenix.real_space_refine*) | ccpem 1.5.0 | PHENIX 1.16-3549 | PHENIX 1.16-3549 |
| Initial model used (PDB code) | | 5bw8 | |
| Resolution of unmasked reconstructions (Å, FSC=0.5) | | 3.89 | 3.95 |
| Resolution of masked reconstructions (Å, FSC=0.5) | | 3.95 | 3.64 |
| Correlation coefficient ($CC_{mask}$) | | 0.80 | 0.83 |
| Map sharpening *B* factor ($Å^2$) | | -145 | -121 |
| Model composition | | | |
| Non-hydrogen atoms | | 4322 | 4145 |
| Protein residues | | 630 | 528 |
| Ligand | | 5 | 5 |
| *B* factors ($Å^2$) | | min/max/mean | min/max/mean |
| Protein | | 30.00/140.27/76.28 | 59.71/131.16/80.59 |
| Ligand | | 6.90/118.10/10.34 | 6.78/123.57/10.35 |
| Bond RMSD | | | |
| Bond lengths (Å) (# $4\sigma$) | | 0.056 (25) | 0.003 (4) |
| Bond angles (°) (# $4\sigma$) | | 2.44 (48) | 1.073 (14) |
| **Validation** | | | |
| MolProbity score | | 1.72 | 1.02 |
| Clashscore | | 8.37 | 1.19 |
| rotamer outliers (%) | | 0.99 | 0 |
| C$\beta$ outliers (%) | | 0.50 | 0 |
| CaBLAM outliers (%) | | 0.84 | 0.40 |
| EMRinger score | | 3.08 | 4.24 |
| Ramachandran plot | | | |
| Favored (%) | | 96.08 | 0 |
| Allowed (%) | | 3.43 | 3.10 |
| Disallowed (%) | | 0.49 | 96.90 |

*UNK is the code for the unknown amino acids of H4/5 and the TMD as a poly-Ala

*Chapter 6*

# TA PROTEIN TARGETING: A COMPLEX MULTIFARIOUS PROCESS

## 6.1 Concluding Remarks

Recent reports of alternative pathways for targeting ER TA proteins highlight the complexity of TA protein biogenesis. Two of these pathways, the GET pathway and the EMC, are conserved throughout the eukaryotic tree. Despite this, the majority of the literature on the proteins involved in these processes are limited to the Opisthokonts supergroup. This thesis aimed to answer questions surrounding how components in TA protein targeting pathways select and shield clients as well as the conservation of these mechanisms.

A combination of computational analysis and experimental localization clarified how the multiple ER targeting pathways select for ER-bound TA proteins rather than mitochondria-bound TA proteins. In fungi and humans, a hydrophobic face is sufficient for selective targeting of TA proteins to the ER membrane. A hydrophobic face distinguishes clients of the EMC insertase from mitochondria-bound TA proteins with TMDs of similar overall hydrophobicities and explains the observed Sgt2 and SGTA preferences for clients with clustered hydrophobics. It is intriguing that a segment of the TMD is sufficient for classifying TA proteins, suggesting that targeting factors only interact with a section of the TMD. This is reflected in the computational model and minimal client binding requirements of the Sgt2 TA protein binding domain – the model only facilitates ~11 amino acids and clients $\geq$11 residues bind to Sgt2 and SGTA. While we expect slight changes to the leading metric as more TA localizations are determined, the high performance of metrics focusing on a subset of the TMD will aid in predicting protein folds in binding partners for TA proteins. The newly identified STI1-domain is in a number of client binding co-chaperones and may also be present in other TA protein targeting factors. Future application of this method of organelle classification by a minimal segment will be helpful for identifying ER and mitochondria-bound TA proteins in humans and other organisms as more targeting pathways are identified. Accurate and sensitive localization of known TA proteins to specific organelles besides just the ER and mitochondria will allow for a more extensive search for information encoded in TA proteins that are organelle specific, i.e. Golgi Apparatus, nuclear, lysosomal.

The identification and structural characterization of a Get3 homolog from Excavata, *Giardia intestinalis*, demonstrates conservation of the mechanisms for TA protein targeting in two distant eukaryotic organisms. This work not only presented the first structure of a protozoan GET pathway component, but it also presents the first comprehensive structural characterization of the catalytic cycle of Get3 from

a single organism. For the first time, apo Get3 is structurally characterized in different conformations spanning the 'open' and 'closed' conformations, supporting biophysical data demonstrating that apo Get3 is not static. This work presents structural evidence that a swelling of the Get3 monomer paired with the opening of the dimer induced by TA protein binding and ATP hydrolysis results in the disassociation from Get4. A high resolution structure of *Gi*Get3 bound to *Gi*Get4 will verify the observations made in Chapter 5 as well as provide mechanistic insight into how Get4 catalytically inhibits Get3.

These works answer several long standing questions, while opening more. Most importantly, how conserved is the GET pathway in other distant relatives of humans? Are these components different enough where drugs can be developed to specifically target GET components in pathogens? Future work investigating this pathway in *G. intestinalis* and other organisms will answer these questions as well as help develop tools that specifically target key proteins in pathogens.

# BIBLIOGRAPHY

Almagro Armenteros, José Juan et al. (Nov. 2017). "DeepLoc: prediction of protein subcellular localization using deep learning." In: *Bioinformatics* 33.21, pp. 3387–3395. DOI: 10.1093/bioinformatics/btx431.

Aurora, Rajeev and George D Rosee (1998). "Helix capping". In: *Protein Science* 7.1, pp. 21–38. DOI: 10.1002/pro.5560070103.

Aurrecoechea, Cristina et al. (Sept. 2008). "GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis". In: *Nucleic Acids Research* 37.suppl$_1$, pp. D526–D530. DOI: 10.1093/nar/gkn631.

Aviram, Naama, Elizabeth A Costa, et al. (Dec. 2016). "The SND proteins constitute an alternative targeting route to the endoplasmic reticulum". In: *Nature* 540.7631, pp. 134–138. DOI: 10.1038/nature20169.

Aviram, Naama and Maya Schuldiner (Dec. 2017). "Targeting and translocation of proteins to the endoplasmic reticulum at a glance". In: *Journal of Cell Science* 130.24, pp. 4079–4085. DOI: 10.1242/jcs.204396.

Bédard, Jocelyn, Sybille Kubis, et al. (July 2007). "Functional similarity between the chloroplast translocon component, Tic40, and the human co-chaperone, Hsp70-interacting protein (Hip)." In: *Journal of Biological Chemistry* 282.29, pp. 21404–21414. DOI: 10.1074/jbc.M611545200.

Bédard, Jocelyn, Raphael Trösch, et al. (Aug. 2017). "Suppressors of the Chloroplast Protein Import Mutant tic40Reveal a Genetic Link between Protein Import and Thylakoid Biogenesis". In: *The Plant Cell* 29.7, pp. 1726–1747. DOI: 10.1105/tpc.16.00962.

Beilharz, Traude et al. (Feb. 2003). "Bipartite Signals Mediate Subcellular Targeting of Tail-anchored Membrane Proteins in Saccharomyces cerevisiae". In: *Journal of Biological Chemistry* 278.10, pp. 8219–8223. DOI: 10.1074/jbc.M212725200.

Bernstein, Harris D et al. (Aug. 1989). "Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle". In: *Nature* 340.6233, pp. 482–486. DOI: 10.1038/340482a0.

Biebl, Maximilian M and Johannes Buchner (Sept. 2019). "Structure, Function, and Regulation of the Hsp90 Machinery." In: *Cold Spring Harbor Perspectives in Biology* 11.9, a034017–33. DOI: 10.1101/cshperspect.a034017.

Bois, Justin S (2020). "justinbois/bebi103: Version 0.1.0". In: DOI: 10.22002/D1.1615.

Borgese, Nica, Sara Colombo, and Emanuela Pedrazzini (June 2003). "The tale of tail-anchored proteins". In: *The Journal of Cell Biology* 161.6, pp. 1013–1019. DOI: 10.1083/jcb.200303069.

Bozkurt, Gunes et al. (2009). "Structural insights into tail-anchored protein binding and membrane insertion by Get3". In: *Proceedings of the National Academy of Sciences* 106.50, pp. 21131–21136. DOI: 10.1073/pnas.0910223106.

Bradley, Philip, Kira M Misura, and David Baker (Sept. 2005). "Towards high-resolution de novo structure prediction for small proteins". In: *Science* 309.5742, pp. 1868–1871. DOI: 10.1126/science.1113801.

Buchanan, Grant et al. (2007). "Control of Androgen Receptor Signaling in Prostate Cancer by the Cochaperone Small Glutamine–Rich Tetratricopeptide Repeat Containing Protein ". In: 67.20, pp. 10087–10096. DOI: 10.1158/0008-5472.CAN-07-1646.

Burri, Lena and Trevor Lithgow (Nov. 2003). "A Complete Set of SNAREs in Yeast". In: *Traffic* 5.1, pp. 45–52. DOI: 10.1046/j.1600-0854.2003.00151.x.

Callahan, Michael A et al. (June 1998). "Functional interaction of human immunodeficiency virus type 1 Vpu and Gag with a novel member of the tetratricopeptide repeat protein family." In: *Journal of Virology* 72.6, pp. 5189–5197.

Caplan, Avrom J (2003). "What is a co-chaperone?" In: *Cell stress & chaperones* 8.2, pp. 105–107. DOI: 10.1379/1466-1268(2003)008$<$0105:wiac$>$2.0.co;2.

Cavalier-Smith, Thomas and Ema E-Y Chao (2003). "Phylogeny and Classification of Phylum Cercozoa (Protozoa)". In: *Protist* 154.3, pp. 341–358. DOI: 10.1078/143446103322454112.

Cernikova, Lenka, Carmen Faso, and Adrian B Hehl (Sept. 2018). "Five facts about Giardia lamblia". In: *PLOS Pathogens* 14.9, pp. 1–5. DOI: 10.1371/journal.ppat.1007250.

Chang, Hui-Chen J, Debra F Nathan, and Susan Lindquist (1997). "In vivo analysis of the Hsp90 cochaperone Sti1 (p60)". In: *Molecular and Cellular Biology* 17.1, pp. 318–325. DOI: 10.1128/MCB.17.1.318.

Chartron, Justin W, William M Clemons Jr, and Christian JM Suloway (Apr. 2012). "The complex process of GETting tail-anchored membrane proteins to the ER." In: *Current Opinion in Structural Biology* 22.2, pp. 217–224. DOI: 10.1016/j.sbi.2012.03.001.

Chartron, Justin W, Grecia M Gonzalez, and William M Clemons Jr (Sept. 2011). "A Structural Model of the Sgt2 Protein and Its Interactions with Chaperones and the Get4/Get5 Complex". In: *Journal of Biological Chemistry* 286.39, pp. 34325–34334. DOI: 10.1074/jbc.M111.277798.

Chartron, Justin W, Christian JM Suloway, et al. (2010). "Structural characterization of the Get4/Get5 complex and its interaction with Get3". In: *Proceedings of the National Academy of Sciences* 107.27, pp. 12127–12132. DOI: 10.1073/pnas.1006036107.

Chartron, Justin W, David G VanderVelde, and William M Clemons Jr (Dec. 2012). "Structures of the Sgt2/SGTA Dimerization Domain with the Get5/UBL4A UBL Domain Reveal an Interaction that Forms a Conserved Dynamic Interface". In: *Cell Reports* 2.6, pp. 1620–1632. DOI: 10.1016/j.celrep.2012.10.010.

Chen, Shiying and David F Smith (1998). "Hop as an Adaptor in the Heat Shock Protein 70 (Hsp70) and Hsp90 Chaperone Machinery". In: *Journal of Biological Chemistry* 273.52, pp. 35194–35200. DOI: 10.1074/jbc.273.52.35194.

Chio, Un Seng, Hyunju Cho, and Shu-ou Shan (Oct. 2017). "Mechanisms of Tail-Anchored Membrane Protein Targeting and Insertion". In: *Annual Review of Cell and Developmental Biology* 33.1, pp. 417–438. DOI: 10.1146/annurev-cellbio-100616-060839.

Chio, Un Seng, SangYoon Chung, et al. (Sept. 2017). "A protean clamp guides membrane targeting of tail-anchored proteins". In: *Proceedings of the National Academy of Sciences* 41, pp. 201708731–10. DOI: 10.1073/pnas.1708731114.

– (Jan. 2019). "A Chaperone Lid Ensures Efficient and Privileged Client Transfer during Tail-Anchored Protein Targeting". In: *Cell Reports* 26.1, 37–44.e7. DOI: 10.1016/j.celrep.2018.12.035.

Chitwood, Patrick J et al. (Nov. 2018). "EMC Is Required to Initiate Accurate Membrane Protein Topogenesis". In: *Cell* 175.6, 1507–1519.e16. DOI: 10.1016/j.cell.2018.10.009.

Cho, Hyunju and Shu-ou Shan (Aug. 2018). "Substrate relay in an Hsp70-cochaperone cascade safeguards tail-anchored membrane protein targeting." In: *EMBO Journal* 37.16. DOI: 10.15252/embj.201899264.

Chou, Ming-Lun, Chiung-Chih Chu, et al. (Dec. 2006). "Stimulation of transit-peptide release and ATP hydrolysis by a cochaperone during protein import into chloroplasts". In: *Journal of Cell Biology* 175.6, pp. 893–900. DOI: 10.1083/jcb.200609172.

Chou, Ming-Lun, Lynda M Fitzpatrick, et al. (June 2003). "Tic40, a membrane-anchored co-chaperone homolog in the chloroplast protein translocon." In: *EMBO Journal* 22.12, pp. 2970–2980. DOI: 10.1093/emboj/cdg281.

Chu, Ruiai et al. (Oct. 2002). "Redesign of a Four-helix Bundle Protein by Phage Display Coupled with Proteolysis and Structural Characterization by NMR and X-ray Crystallography". In: *Journal of Molecular Biology* 323.2, pp. 253–262. DOI: 10.1016/s0022-2836(02)00884-7.

Chun, Eugene et al. (June 2012). "Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors". In: *Cell Structure and Function* 20.6, pp. 967–976. DOI: 10.1016/j.str.2012.04.010.

Clemons Jr, William M et al. (Sept. 1999). "Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 A resolution: evidence for the mechanism of signal peptide binding." In: *Journal of Molecular Biology* 292.3, pp. 697–705. DOI: 10.1006/jmbi.1999.3090.

Conklin, Darrell et al. (May 2000). "Molecular cloning, chromosome mapping and characterization of UBQLN3 a testis-specific gene that contains an ubiquitin-like domain." In: *Gene* 249.1-2, pp. 91–98. DOI: 10.1016/s0378-1119(00)00122-0.

Consortium, The UniProt (Nov. 2020). "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic Acids Research* 49.D1, pp. D480–D489. DOI: 10.1093/nar/gkaa1100.

Costello, Joseph L et al. (May 2017). "Predicting the targeting of tail-anchored proteins to subcellular compartments in mammalian cells." In: *Journal of Cell Science* 130.9, pp. 1675–1687. DOI: 10.1242/jcs.200204.

Coto, Amanda L S et al. (Oct. 2018). "Structural and functional studies of the Leishmania braziliensis SGT co-chaperone indicate that it shares structural features with HIP and can interact with both Hsp90 and Hsp70 with similar affinities." In: *International Journal of Biological Macromolecules* 118.Pt A, pp. 693–706. DOI: 10.1016/j.ijbiomac.2018.06.123.

Cziepluch, Celina et al. (May 1998). "Identification of a novel cellular TPR-containing protein, SGT, that interacts with the nonstructural protein NS1 of parvovirus H-1." In: *Journal of Virology* 72.5, pp. 4149–4156. DOI: 10.1128/JVI.72.5.4149-4156.1998.

Dantuma, Nico P, Christian Heinen, and Deborah Hoogstraten (Apr. 2009). "The ubiquitin receptor Rad23: At the crossroads of nucleotide excision repair and proteasomal degradation". In: *DNA Repair* 8.4, pp. 449–460. DOI: 10.1016/j.dnarep.2009.01.005.

Darby, John F et al. (Nov. 2014). "Solution Structure of the SGTA Dimerisation Domain and Investigation of Its Interactions with the Ubiquitin-Like Domains of BAG6 and UBL4A". In: *PLoS ONE* 9.11, e113281–19. DOI: 10.1371/journal.pone.0113281.

Deng, Han-Xiang et al. (Sept. 2011). "Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia". In: *Nature* 477.7363, pp. 211–215. DOI: 10.1038/nature10353.

Denic, Vladimir (Oct. 2012). "A portrait of the GET pathway as a surprisingly complicated young man". In: *Trends in Biochemical Sciences* 37.10, pp. 411–417. DOI: 10.1016/j.tibs.2012.07.004.

Dolezal, Pavel et al. (2005). "Giardia mitosomes and trichomonad hydrogenosomes share a common mode of protein targeting". In: *Proceedings of the National Academy of Sciences* 102.31, pp. 10924–10929. DOI: 10.1073/pnas.0500349102.

Drozdetskiy, Alexey et al. (June 2015). "JPred4: a protein secondary structure prediction server". In: *Nucleic Acids Research* 43.W1, W389–W394. DOI: 10.1093/nar/gkv332.

Dupzyk, Allison et al. (June 2017). "SGTA-Dependent Regulation of Hsc70 Promotes Cytosol Entry of Simian Virus 40 from the Endoplasmic Reticulum." In: *Journal of Virology* 91.12, p. 23. DOI: 10.1128/JVI.00232-17.

Dutta, Sujit and Yee-Joo Tan (Sept. 2008). "Structural and functional characterization of human SGT and its interaction with Vpu of the human immunodeficiency virus type 1." In: *Biochemistry* 47.38, pp. 10123–10131. DOI: 10.1021/bi800758a.

Dyson, H Jane and Peter E Wright (Aug. 2004). "Unfolded Proteins and Protein Folding Studied by NMR". In: *Chemical Reviews* 104.8, pp. 3607–3622. DOI: 10.1021/cr030403s.

Eisenberg, David et al. (1984). "Analysis of membrane and surface protein sequences with the hydrohobic moment plot". In: *Journal of Molecular Biology* 179, pp. 125–142. DOI: 10.1016/0022-2836(84)90309-7.

El Ayadi, Amina et al. (Mar. 2013). "Ubiquilin-1 and protein quality control in Alzheimer disease." In: *Prion* 7.2, pp. 164–169. DOI: 10.4161/pri.23711.

Elsasser, Suzanne et al. (Sept. 2002). "Proteasome subunit Rpn1 binds ubiquitin-like protein domains". In: *Nature Cell Biology* 4.9, pp. 725–730. DOI: 10.1038/ncb845.

Fan, Guo-Huang et al. (Feb. 2002). "Hsc/Hsp70 Interacting Protein (Hip) Associates with CXCR2 and Regulates the Receptor Signaling and Trafficking". In: *Journal of Biological Chemistry* 277.8, pp. 6590–6597. DOI: 10.1074/jbc.M110588200.

Fauchere, J L and Vladimir Pliska (1983). "Hydrophobic parameters $\pi$ of amino acid side chains from partitioning of N-acetyl amino acid amides". In: *European Journal of Medical Chemistry* 18, pp. 369–375.

Feng, Yaoyu and Lihua Xiao (2011). "Zoonotic Potential and Molecular Epidemiology of *Giardia* Species and Giardiasis". In: *Clinical Microbiology Reviews* 24.1, pp. 110–140. DOI: 10.1128/CMR.00033-10.

Fernandez-Patron, C et al. (Jan. 1995). "Double staining of coomassie blue-stained polyacrylamide gels by imidazole-sodium dodecyl sulfate-zinc reverse staining: sensitive detection of coomassie blue-undetected proteins." In: *Analytical Biochemistry* 224.1, pp. 263–269. DOI: 10.1006/abio.1995.1039.

Figueiredo Costa, Bruna et al. (Feb. 2018). "Discrimination between the endoplasmic reticulum and mitochondria by spontaneously inserting tail-anchored proteins". In: *Traffic* 19.3, pp. 182–197. DOI: 10.1111/tra.12550.

Fishbain, Susan et al. (Feb. 2011). "Rad23 escapes degradation because it lacks a proteasome initiation region". In: *Nature Communications* 2.1, p. 192. DOI: 10.1038/ncomms1194.

Fonseca, Anna Carolina Carvalho da et al. (Jan. 2021). “The multiple functions of the co-chaperone stress inducible protein 1”. In: *Cytokine and Growth Factor Reviews*, pp. 1–11. DOI: 10.1016/j.cytogfr.2020.06.003.

Fowler, Douglas M and Stanley Fields (Aug. 2014). “Deep mutational scanning: a new style of protein science.” In: *Nature Methods* 11.8, pp. 801–807. DOI: 10.1038/nmeth.3027.

Fry, Michelle Y and William M Clemons Jr (2018). “Complexity in targeting membrane proteins”. In: *Science* 359.6374, pp. 390–391. DOI: 10.1126/science.aar5992.

Fry, Michelle Y, Shyam M Saladi, et al. (2021). “Sequence-based features that are determinant for tail-anchored membrane protein sorting in eukaryotes”. In: *Traffic* 22.9, pp. 306–318. DOI: 10.1111/tra.12809.

Gautier, Romain et al. (Sept. 2008). “HELIQUEST: a web server to screen sequences with specific  -helical properties”. In: *Bioinformatics* 24.18, pp. 2101–2102. DOI: 10.1093/bioinformatics/btn392.

El-Gebali, Sara et al. (Oct. 2018). “The Pfam protein families database in 2019”. In: *Nucleic Acids Research* 47.D1, pp. D427–D432. DOI: 10.1093/nar/gky995.

Gene Ontology Consortium (Jan. 2021). “The Gene Ontology resource: enriching a GOld mine.” In: *Nucleic Acids Research* 49.D1, pp. D325–D334.

Graether, Steffen P (Jan. 2019). “Troubleshooting Guide to Expressing Intrinsically Disordered Proteins for Use in NMR Experiments”. In: *Frontiers in Molecular Biosciences* 5, pp. 49–9. DOI: 10.3389/fmolb.2018.00118.

Gristick, Harry B et al. (May 2014). “Crystal structure of ATP-bound Get3–Get4–Get5 complex reveals regulation of Get3 by Get4”. In: *Nature Structural & Molecular Biology* 21.5, pp. 437–442. DOI: 10.1038/nsmb.2813.

Guna, Alina and Ramanujan S Hegde (Apr. 2018). “Transmembrane Domain Recognition during Membrane Protein Biogenesis and Quality Control”. In: *Current Biology* 28.8, R498–R511. DOI: 10.1016/j.cub.2018.02.004.

Guna, Alina, Norbert Volkmar, et al. (Jan. 2018). “The ER membrane protein complex is a transmembrane domain insertase”. In: *Science* 359.6374, pp. 470–473. DOI: 10.1126/science.aao3099.

Hara, Taichi et al. (Oct. 2005). “Role of the UBL-UBA Protein KPC2 in Degradation of p27 at G1 Phase of the Cell Cycle”. In: *Molecular and Cellular Biology* 25.21, pp. 9292–9303. DOI: 10.1128/MCB.25.21.9292-9303.2005.

Hartley, Andrew M et al. (Dec. 2018). “Structure of yeast cytochrome c oxidase in a supercomplex with cytochrome bc1”. In: *Nature Structural & Molecular Biology*, pp. 1–10. DOI: 10.1038/s41594-018-0172-z.

Hegde, Ramanujan S and Robert J Keenan (Dec. 2011). “Tail-anchored membrane protein insertion into the endoplasmic reticulum”. In: *Nature Reviews Molecular Cell Biology* 12.12, pp. 787–798. DOI: 10.1038/nrm3226.

Hessa, Tara, Nadja M Meindl-Beinker, et al. (Dec. 2007). “Molecular code for transmembrane-helix recognition by the Sec61 translocon.” In: *Nature* 450.7172, pp. 1026–1030. DOI: 10.1038/nature06387.

Hessa, Tara, Ajay Sharma, et al. (July 2011). “Protein targeting and degradation are coupled for elimination of mislocalized proteins”. In: *Nature* 475.7356, pp. 394–397. DOI: 10.1038/nature10181.

Howe, Kevin L et al. (Oct. 2019). “Ensembl Genomes 2020—enabling non-vertebrate genomic research”. In: *Nucleic Acids Research* 48.D1, pp. D689–D695. DOI: 10.1093/nar/gkz890.

Itakura, Eisuke et al. (July 2016). “Ubiquilins Chaperone and Triage Mitochondrial Membrane Proteins for Degradation”. In: *Molecular Cell* 63.1, pp. 21–33. DOI: 10.1016/j.molcel.2016.05.020.

Jonikas, Martin C et al. (Mar. 2009). “Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum.” In: *Science* 323.5922, pp. 1693–1697. DOI: 10.1126/science.1167983.

Kabsch, Wolfgang and Christian Sander (1983). “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22.12, pp. 2577–2637. DOI: 10.1002/bip.360221211.

Kajander, Tommi et al. (Sept. 2009). “Electrostatic interactions of Hsp-organizing protein tetratricopeptide domains with Hsp70 and Hsp90: computational analysis and protein engineering.” In: *Journal of Biological Chemistry* 284.37, pp. 25364–25374. DOI: 10.1074/jbc.M109.033894.

Kalbfleisch, Ted, Alex Cambon, and Binks W Wattenberg (Sept. 2007). “A Bioinformatics Approach to Identifying Tail-Anchored Proteins in the Human Genome”. In: *Traffic* 8.12, pp. 1687–1694. DOI: 10.1111/j.1600-0854.2007.00661.x.

Katoh, Kazutaka and Daron M Standley (Mar. 2013). “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. In: *Molecular Biology and Evolution* 30.4, pp. 772–780. DOI: 10.1093/molbev/mst010.

Keenan, Robert J et al. (July 1998). “Crystal Structure of the Signal Sequence Binding Subunit of the Signal Recognition Particle”. In: *Cell* 94.2, pp. 181–191. DOI: 10.7554/eLife.07975.

Keister, David B (Jan. 1983). “Axenic culture of Giardia lamblia in TYI-S-33 medium supplemented with bile”. In: *Transactions of The Royal Society of Tropical Medicine and Hygiene* 77.4, pp. 487–488. DOI: 10.1016/0035-9203(83)90120-7.

Kiktev, Denis A et al. (Dec. 2012). "Regulation of chaperone effects on a yeast prion by cochaperone Sgt2." In: *Molecular and Cellular Biology* 32.24, pp. 4960–4970. DOI: 10.1128/MCB.00875-12.

Kirschke, Elaine et al. (June 2014). "Glucocorticoid Receptor Function Regulated by Coordinated Action of the Hsp90 and Hsp70 Chaperone Cycles". In: *Cell* 157.7, pp. 1685–1697. DOI: 10.1016/j.cell.2014.04.038.

Ko, Kenton et al. (2004). "The Tic40 translocon components exhibit preferential interactions with different forms of the Oee1 plastid protein precursor". In: *Functional Plant Biology* 31.3, pp. 285–294. DOI: 10.1071/FP03195.

Kotoshiba, Shuhei et al. (May 2005). "Molecular dissection of the interaction between p27 and Kip1 ubiquitylation-promoting complex, the ubiquitin ligase that regulates proteolysis of p27 in G1 phase." In: *Journal of Biological Chemistry* 280.18, pp. 17694–17700. DOI: 10.1074/jbc.M500866200.

Kovacheva, Sabina et al. (Dec. 2004). "In vivo studies on the roles of Tic110, Tic40 and Hsp93 during chloroplast protein import". In: *The Plant Journal* 41.3, pp. 412–428. DOI: 10.1111/j.1365-313X.2004.02307.x.

Krogh, Anders et al. (Jan. 2001). "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen". In: *Journal of Molecular Biology* 305.3, pp. 567–580. DOI: 10.1006/jmbi.2000.4315.

Kubota, Keiko et al. (2012). "Get1 Stabilizes an Open Dimer Conformation of Get3 ATPase by Binding Two Distinct Interfaces". In: *Journal of Molecular Biology* 422.3, pp. 366–375. DOI: 10.1016/j.jmb.2012.05.045.

Kumar, Tarkeshwar et al. (2021). "A conserved Guided Entry of Tail-anchored pathway is involved in the trafficking of tail-anchored membrane proteins in Plasmodium falciparum". In: *bioRxiv*.

Kutay, Ulrike, Enno Hartmann, and Tom A Rapoport (Mar. 1993). "A class of membrane proteins with a C-terminal anchor". In: *Trends in Cell Biology* 3.3, pp. 72–75. DOI: 10.1016/0962-8924(93)90066-a.

Kyte, Jack and Russell F Doolittle (May 1982). "A simple method for displaying the hydropathic character of a protein." In: *Journal of Molecular Biology* 157.1, pp. 105–132. DOI: 10.1016/0022-2836(82)90515-0.

Lässle, Michael et al. (Jan. 1997). "Stress-inducible, murine protein mSTI1. Characterization of binding domains for heat shock proteins and in vitro phosphorylation by different kinases." In: *Journal of Biological Chemistry* 272.3, pp. 1876–1884. DOI: 10.1074/jbc.272.3.1876.

Letunic, Ivica and Peer Bork (Oct. 2017). "20 years of the SMART protein domain annotation resource". In: *Nucleic Acids Research* 46.D1, pp. D493–D496. DOI: 10.1093/nar/gkx922.

Levchenko, Maria et al. (2016). "Cox26 is a novel stoichiometric subunit of the yeast cytochrome c oxidase". In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863.7, Part A, pp. 1624–1632. DOI: 10.1016/j.bbamcr.2016.04.007.

Leznicki, P et al. (Sept. 2015). "Binding of SGTA to Rpn13 selectively modulates protein quality control". In: *Journal of Cell Science* 128.17, pp. 3187–3196. DOI: 10.1242/jcs.165209.

Leznicki, Pawel and Stephen High (Nov. 2012). "SGTA antagonizes BAG6-mediated protein triage". In: *Proceedings of the National Academy of Sciences* 109.47, pp. 19214–19219. DOI: 10.1073/pnas.1209997109.

Li, Weizhong and Adam Godzik (June 2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22.13, pp. 1658–1659. DOI: 10.1093/bioinformatics/btl158.

Li, Zhuo, F Ulrich Hartl, and Andreas Bracher (Aug. 2013). "Structure and function of Hip, an attenuator of the Hsp70 chaperone cycle". In: *Nature Structural & Molecular Biology* 20.8, pp. 929–935. DOI: 10.1038/nsmb.2608.

Lim, Precious J et al. (Oct. 2009). "Ubiquilin and p97/VCP bind erasin, forming a complex involved in ERAD." In: *Journal of Cell Biology* 187.2, pp. 201–217. DOI: 10.1083/jcb.200903024.

Lin, Ku-Feng et al. (Jan. 2021). "Molecular basis of tail-anchored integral membrane protein recognition by the cochaperone Sgt2". In: *Journal of Biological Chemistry* 296. DOI: 10.1016/j.jbc.2021.100441.

Liou, Shen-Ting and Chung Wang (2005). "Small glutamine-rich tetratricopeptide repeat-containing protein is composed of three structural units with distinct functions". In: *Archives of Biochemistry and Biophysics* 435.2, pp. 253–263. DOI: 10.1016/j.abb.2004.12.020.

Liu, Fu-Hwa et al. (Nov. 1999). "Specific interaction of the 70-kDa heat shock cognate protein with the tetratricopeptide repeats." In: *Journal of Biological Chemistry* 274.48, pp. 34425–34432. DOI: 10.1074/jbc.274.48.34425.

Long, Philip et al. (Sept. 2012). "A yeast two-hybrid screen reveals that osteopontin associates with MAP1A and MAP1B in addition to other proteins linked to microtubule stability, apoptosis and protein degradation in the human brain." In: *European Journal of Neuroscience* 36.6, pp. 2733–2742. DOI: 10.1111/j.1460-9568.2012.08189.x.

Lott, Antonia, Javier Oroz, and Markus Zweckstetter (Oct. 2020). "Molecular basis of the interaction of Hsp90 with its co-chaperone Hop". In: *Protein Science* 29.12, pp. 2422–2432. DOI: 10.1002/pro.3969.

Lu, Alex X et al. (Nov. 2019). "YeastSpotter: accurate and parameter-free web segmentation for microscopy images of yeast cells." In: *Bioinformatics* 35.21, pp. 4525–4527. DOI: 10.1093/bioinformatics/btz402.

Luo, Peizhi and Robert L Baldwin (July 1997). "Mechanism of helix induction by trifluoroethanol: a framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water." In: *Biochemistry* 36.27, pp. 8413–8421. DOI: 10.1021/bi9707133.

Mariappan, Malaiyalam et al. (Sept. 2011). "The mechanism of membrane-associated steps in tail-anchored protein insertion". In: *Nature* 477.7362, pp. 61–66. DOI: 10.1038/nature10362.

Marín, Ignacio (Mar. 2014). "The ubiquilin gene family: evolutionary patterns and functional insights". In: *BMC Evolutionary Biology* 14.1, p. 63. DOI: 10.1186/1471-2148-14-63.

Martincová, Eva et al. (2015). "Probing the Biology of Giardia intestinalis Mitosomes Using *in vivo* Enzymatic Tagging". In: *Molecular and Cellular Biology* 35.16, pp. 2864–2874. DOI: 10.1128/MCB.00448-15.

Martínez-Lumbreras, Santiago et al. (July 2018). "Structural complexity of the co-chaperone SGTA: a conserved C-terminal region is implicated in dimerization and substrate quality control". In: *BMC Biology* 16.1, p. 76. DOI: 10.1186/s12915-018-0542-3.

Martins, Vilma R et al. (Dec. 1997). "Complementary hydropathy identifies a cellular prion protein receptor". In: *Nature Medicine* 3.12, pp. 1376–1382. DOI: 10.1038/nm1297-1376.

Mateja, Agnieszka, Marcin Paduch, et al. (2015). "Structure of the Get3 targeting factor in complex with its membrane protein cargo". In: *Science* 347.6226, pp. 1152–1155. DOI: 10.1126/science.1261671.

Mateja, Agnieszka, Anna Szlachcic, et al. (Sept. 2009). "The structural basis of tail-anchored membrane protein recognition by Get3". In: *Nature* 461.7262, pp. 361–366. DOI: 10.1038/nature08319.

McDowell, Melanie A et al. (2020). "Structural Basis of Tail-Anchored Membrane Protein Biogenesis by the GET Insertase Complex". In: *Molecular Cell* 80.1, 72–86.e7. DOI: 0.1016/j.molcel.2020.08.012.

McGibbon, Robert T et al. (Oct. 2015). "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories". In: *Biophysical Journal* 109.8, pp. 1528–1532. DOI: 10.1016/j.bpj.2015.08.015.

Micsonai, András et al. (June 2015). "Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy". In: *Proceedings of the National Academy of Sciences* 112.24, E3095–E3103. DOI: 10.1073/pnas.1500851112.

Mock, Jee-Young et al. (Jan. 2015). "Bag6 complex contains a minimal tail-anchor–targeting module and a mock BAG domain". In: *Proceedings of the National Academy of Sciences* 112.1, pp. 106–111. DOI: 10.1073/pnas.1402745112.

Morgens, David W et al. (Nov. 2019). "Retro-2 protects cells from ricin toxicity by inhibiting ASNA1-mediated ER targeting and insertion of tail-anchored proteins". In: *eLife* 8, e48434. DOI: 10.7554/eLife.48434.

Nelson, Gregory M, Holly Huffman, and David F Smith (Apr. 2003). "Comparison of the carboxy-terminal DP-repeat region in teh co-chaperones Hop and Hip". In: *EMBO Journal* 8.2, pp. 125–133. DOI: 10.1379/1466-1268(2003)008$<$0125:cotcdr$>$2.0.co;2.

Nielsen, Henrik (2017). "Predicting Secretory Proteins with SignalP". In: *Protein Function Prediction: Methods and Protocols*. Ed. by Daisuke Kihara. New York, NY: Springer New York, pp. 59–73. DOI: 10.1007/978-1-4939-7015-5_6.

Okreglak, Voytek and Peter Walter (June 2014). "The conserved AAA-ATPase Msp1 confers organelle specificity to tail-anchored proteins". In: *Proceedings of the National Academy of Sciences* 111.22, pp. 8019–8024. DOI: 10.1073/pnas.1405755111.

Onuoha, S C et al. (June 2008). "Structural Studies on the Co-chaperone Hop and Its Complexes with Hsp90". In: *Journal of Molecular Biology* 379.4, pp. 732–744. DOI: 10.1016/j.jmb.2008.02.013.

Pace, C Nick and J Martin Scholtz (July 1998). "A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins". In: *Biophysical Journal* 75.1, pp. 422–427. DOI: 10.1016/s0006-3495(98)77529-0.

Pei, Jimin, Bong-Hyun Kim, and Nick V Grishin (Feb. 2008). "PROMALS3D: a tool for multiple protein sequence and structure alignments". In: *Nucleic Acids Research* 36.7, pp. 2295–2300. DOI: 10.1093/nar/gkn072.

Pieper, Ursula et al. (Feb. 2013). "Coordinating the impact of structural genomics on the human $\alpha$-helical transmembrane proteome". In: *Nature Structural & Molecular Biology* 20.2, pp. 135–138. DOI: 10.1038/nsmb.2508.

Pierce, Brian G et al. (June 2014). "ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers". In: *Bioinformatics* 30.12, pp. 1771–1773. DOI: 10.1093/bioinformatics/btu097.

Prapapanich, Viravan, Shiying Chen, and David F Smith (Feb. 1998). "Mutation of Hip's Carboxy-Terminal Region Inhibits a Transitional Stage of Progesterone Receptor Assembly". In: *Molecular and cellular biology* 18.2, pp. 944–952. DOI: 10.1128/MCB.18.2.944.

Prapapanich, Viravan, Shiying Chen, Eric J Toran, et al. (Nov. 1996). "Mutational analysis of the hsp70-interacting protein Hip." In: *Molecular and Cellular Biology* 16.11, pp. 6200–6207. DOI: 10.1128/MCB.16.11.6200.

Prodromou, Chrisostomos (Aug. 2016). "Mechanisms of Hsp90 regulation". In: *Biochemical Journal* 473.16, pp. 2439–2452. DOI: 10.1042/BCJ20160005.

Rabu, Catherine et al. (Oct. 2009a). "Biogenesis of tail-anchored proteins: the beginning for the end?" In: *Journal of Cell Science* 122.Pt 20, pp. 3605–3612. DOI: 10.1242/jcs.041210.

– (Oct. 2009b). "Biogenesis of tail-anchored proteins: the beginning for the end?" In: *Journal of Cell Science* 122.20, pp. 3605–3612. DOI: 10.1242/jcs.041210.

Rao, Meera et al. (Dec. 2016). "Multiple selection filters ensure accurate tail-anchored membrane protein targeting". In: *eLife* 5, e21301–24. DOI: 10.7554/eLife.21301.

Reidy, Michael et al. (Aug. 2018). "Dual Roles for Yeast Sti1/Hop in Regulating the Hsp90 Chaperone Cycle." In: *Genetics* 209.4, pp. 1139–1154. DOI: 10.1534/genetics.118.301178.

Reißer, Sabine et al. (June 2014). "3D Hydrophobic Moment Vectors as a Tool to Characterize the Surface Polarity of Amphiphilic Peptides". In: *Biophysical Journal* 106.11, pp. 2385–2394. DOI: 10.1016/j.bpj.2014.04.020.

Rivera-Monroy, Jhon et al. (Dec. 2016). "Mice lacking WRB reveal differential biogenesis requirements of tail-anchored proteins in vivo". In: *Scientific Reports* 6.1, p. 39464. DOI: 10.1038/srep39464.

Rodrigo-Brenni, Monica C, Erik Gutierrez, and Ramanujan S Hegde (July 2014). "Cytosolic Quality Control of Mislocalized Proteins Requires RNF126 Recruitment to Bag6". In: *Molecular Cell* 55.2, pp. 227–237. DOI: 10.1016/j.molcel.2014.05.025.

Röhl, Alina, Julia Rohrberg, and Johannes Buchner (2013). "The chaperone Hsp90: changing partners for demanding clients". In: *Trends in Biochemical Sciences* 38.5, pp. 253–262. DOI: 10.1016/j.tibs.2013.02.003.

Röhl, Alina et al. (Apr. 2015). "Hsp90 regulates the dynamics of its cochaperone Sti1 and the transfer of Hsp70 between modules." In: *Nature Communications* 6.1, p. 6655. DOI: 10.1038/ncomms7655.

Rome, Michael E, Un Seng Chio, et al. (Nov. 2014). "Differential gradients of interaction affinities drive efficient targeting and recycling in the GET pathway." In: *Proceedings of the National Academy of Sciences* 111.46, E4929–35. DOI: 10.1073/pnas.1411284111.

Rome, Michael E, Meera Rao, et al. (2013). "Precise timing of ATPase activation drives targeting of tail-anchored proteins". In: *Proceedings of the National Academy of Sciences* 110.19, pp. 7666–7671. DOI: 10.1073/pnas.1222054110.

Roseman, Mark A (Apr. 1988). "Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds." In: *Journal of Molecular Biology* 200.3, pp. 513–522. DOI: 10.1016/0022-2836(88)90540-2.

Saeki, Yasushi et al. (Aug. 2002). "Identification of ubiquitin-like protein-binding subunits of the 26S proteasome." In: *Biochemical and Biophysical Research Communications* 296.4, pp. 813–819. DOI: 10.1016/s0006-291x(02)02002-8.

Saladi, Shyam M et al. (Sept. 2020). "Structural biologists, let's mind our colors". In: *bioRxiv* 27, pp. 14–16.

Salonen, Laura M, Manuel Ellermann, and François Diederich (2011). "Aromatic Rings in Chemical and Biological Recognition: Energetics and Structures". In: *Angewandte Chemie International Edition* 50.21, pp. 4808–4842. DOI: 10.1002/anie.201007560.

Schägger, Hermann (June 2006). "Tricine–SDS-PAGE". In: *Nature Protocols* 1.1, pp. 16–22. DOI: 0.1038/nprot.2006.4.

Scheufler, Clemens et al. (Apr. 2000). "Structure of TPR Domain–Peptide Complexes". In: *Cell* 101.2, pp. 199–210. DOI: 10.1016/S0092-8674(00)80830-2.

Schiffer, Marianne and Allen B Edmundson (Mar. 1967). "Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential". In: *Biophysical Journal* 7.2, pp. 121–135. DOI: 10.1016/S0006-3495(67)86579-2.

Schmid, Andreas B et al. (Mar. 2012). "The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop." In: *EMBO Journal* 31.6, pp. 1506–1517. DOI: 10.1038/emboj.2011.472.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri (July 2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9.7, pp. 671–675. DOI: 10.1038/nmeth.2089.

Schopf, Florian H, Maximilian M Biebl, and Johannes Buchner (June 2017). "The HSP90 chaperone machinery". In: *Nature Reviews Molecular Cell Biology* 18.6, pp. 345–360. DOI: 10.1038/nrm.2017.20.

Schrum, Jason P, Ting F Zhu, and Jack W Szostak (Sept. 2010). "The origins of cellular life." In: *Cold Spring Harbor Perspectives in Biology* 2.9, a002212. DOI: 10.1101/cshperspect.a002212.

Schuldiner, Maya, Sean R Collins, et al. (Nov. 2005). "Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile". In: *Cell* 123.3, pp. 507–519. DOI: 10.1016/j.cell.2005.08.031.

Schuldiner, Maya, Jutta Metz, et al. (Aug. 2008). "The GET Complex Mediates Insertion of Tail-Anchored Proteins into the ER Membrane". In: *Cell* 134.4, pp. 634–645. DOI: 10.1016/j.cell.2008.06.025.

Şentürk, Mümine et al. (Feb. 2019). "Ubiquilins regulate autophagic flux through mTOR signalling and lysosomal acidification". In: *Nature Cell Biology*, pp. 1–19. DOI: 10.1038/s41556-019-0281-x.

Seok Ko, Han et al. (2004). “Ubiquilin interacts with ubiquitylated proteins and proteasome through its ubiquitin-associated and ubiquitin-like domains”. In: *FEBS Letters* 566.1-3, pp. 110–114. DOI: 10.1016/j.febslet.2004.04.031.

Shan, Shu-ou (2019). “Guiding tail-anchored membrane proteins to the endoplasmic reticulum in a chaperone cascade”. In: *Journal of Biological Chemistry* 294.45, pp. 16577–16586. DOI: 10.1074/jbc.REV119.006197.

Shao, Sichen and Ramanujan S Hegde (Dec. 2011a). “A Calmodulin-Dependent Translocation Pathway for Small Secretory Proteins”. In: *Cell* 147.7, pp. 1576–1588. DOI: 10.1016/j.cell.2011.11.048.

– (Nov. 2011b). “Membrane Protein Insertion at the Endoplasmic Reticulum”. In: *Annual Review of Cell and Developmental Biology* 27.1, pp. 25–56. DOI: 10.1146/annurev-cellbio-092910-154125.

Shao, Sichen, Monica C Rodrigo-Brenni, et al. (Jan. 2017). “Mechanistic basis for a molecular triage reaction”. In: *Science* 355.6322, pp. 298–302. DOI: 10.1126/science.aah6130.

Shindyalov, Ilya N and Philip E Bourne (Sept. 1998). “Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.” In: *Protein Engineering* 11.9, pp. 739–747. DOI: 10.1093/protein/11.9.739.

Simon, Aline C et al. (Jan. 2013). “Structure of the Sgt2/Get5 complex provides insights into GET-mediated targeting of tail-anchored membrane proteins”. In: *Proceedings of the National Academy of Sciences* 110.4, pp. 1327–1332. DOI: 10.1073/pnas.1207518110.

Simpson, Peter J et al. (Aug. 2010). “Structures of Get3, Get4, and Get5 Provide New Models for TA Membrane Protein Targeting”. In: *Structure* 18.8, pp. 897–902. DOI: 10.1016/j.str.2010.07.003.

Stefanovic, Sandra and Ramanujan S Hegde (Mar. 2007). “Identification of a Targeting Factor for Posttranslational Membrane Protein Insertion into the ER”. In: *Cell* 128.6, pp. 1147–1159. DOI: 10.1016/j.cell.2007.01.036.

Stefer, Susanne et al. (Aug. 2011). “Structural basis for tail-anchored membrane protein biogenesis by the Get3-receptor complex.” In: *Science* 333.6043, pp. 758–762. DOI: 10.1126/science.1207125.

Stringer, Carsen et al. (Dec. 2020). “Cellpose: a generalist algorithm for cellular segmentation”. In: *Nature Methods*, pp. 1–21. DOI: 10.1038/s41592-020-01018-x.

Studier, F William (May 2005). “Protein production by auto-induction in high-density shaking cultures”. In: *Protein Expression and Purification* 41.1, pp. 207–234. DOI: 10.1016/j.pep.2005.01.016.

Subudhi, Ipsita and James Shorter (Mar. 2018). “Ubiquilin 2: Shuttling Clients Out of Phase?” In: *Molecular Cell* 69.6, pp. 919–921. DOI: 10.1016/j.molcel.2018.02.030.

Suloway, Christian JM, Justin W Chartron, et al. (Sept. 2009). "Model for eukaryotic tail-anchored protein binding based on the structure of Get3". In: *Proceedings of the National Academy of Sciences* 106.35, pp. 14849–14854. DOI: 10.1073/pnas.0907522106.

Suloway, Christian JM, Michael E Rome, and William M Clemons Jr (Nov. 2011). "Tail-anchor targeting by a Get3 tetramer: the structure of an archaeal homologue". In: *EMBO Journal* 31.3, pp. 707–719. DOI: 10.1038/emboj.2011.433.

Swets, John A, Robyn M Dawes, and John Monahan (Oct. 2000). "Better decisions through science." In: *Scientific American* 283.4, pp. 82–87. DOI: 10.1038/scientificamerican1000-82.

Thompson, Julie D et al. (Dec. 1997). "The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools". In: *Nucleic Acids Research* 25.24, pp. 4876–4882. DOI: 10.1093/nar/25.24.4876.

Tidow, Henning and Poul Nissen (May 2013). "Structural diversity of calmodulin binding to its target sites". In: *FEBS Journal* 280.21, pp. 5551–5565. DOI: 10.1111/febs.12296.

Topf, Maya et al. (2008). "Protein Structure Fitting and Refinement Guided by Cryo-EM Density". In: *Structure* 16.2, pp. 295–307. DOI: 10.1016/j.str.2007.11.016.

Trempe, Jean-François et al. (Sept. 2016). "Structural studies of the yeast DNA damage-inducible protein Ddi1 reveal domain architecture of this eukaryotic protein family". In: *Scientific Reports*, pp. 1–13. DOI: 10.1038/srep33671.

Trotta, Andrew P et al. (Dec. 2013). "Knockdown of the cochaperone SGTA results in the suppression of androgen and PI3K/Akt signaling and inhibition of prostate cancer cell proliferation." In: *International Journal of Cancer* 133.12, pp. 2812–2823. DOI: 10.1002/ijc.28310.

Tsirigos, Konstantinos D et al. (June 2015). "The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides". In: *Nucleic Acids Research* 43.W1, W401–W407. DOI: 10.1093/nar/gkv485.

Vida, Thomas A and Scott D Emr (Mar. 1995). "A new vital stain for visualizing vacuolar membrane dynamics and endocytosis in yeast." In: *Journal of Cell Biology* 128.5, pp. 779–792. DOI: 10.1083/jcb.128.5.779.

Vilardi, F, H Lorenz, and B Dobberstein (Mar. 2011). "WRB is the receptor for TRC40/Asna1-mediated insertion of tail-anchored proteins into the ER membrane". In: *Journal of Cell Science* 124.8, pp. 1301–1307. DOI: 10.1242/jcs.084277.

Voorhees, Rebecca M and Ramanujan S Hegde (July 2015). "Structures of the scanning and engaged states of the mammalian SRP-ribosome complex". In: *eLife* 4, pp. 1485–21. DOI: 10.7554/eLife.07975.

Wade, Staton L and David T Auble (Oct. 2014). "The Rad23 ubiquitin receptor, the proteasome and functional specificity in transcriptional control". In: *Transcription* 1.1, pp. 22–26. DOI: 10.4161/trns.1.1.12201.

Waheed, Abdul A et al. (Apr. 2016). "The Vpu-interacting Protein SGTA Regulates Expression of a Non-glycosylated Tetherin Species". In: *Scientific Reports* 6.1, p. 24934. DOI: 10.1038/srep24934.

Wallner, Bjorn and Arne Elofsson (Mar. 2006). "Identification of correct regions in protein models using structural, alignment, and consensus information". In: *Protein Science* 15.4, pp. 900–913. DOI: 10.1110/ps.051799606.

Wang, Fei, Emily C Brown, et al. (Oct. 2010). "A Chaperone Cascade Sorts Proteins for Posttranslational Membrane Insertion into the Endoplasmic Reticulum". In: *Molecular Cell* 40.1, pp. 159–171. DOI: 10.1016/j.molcel.2010.08.038.

Wang, Fei, Andrew Whynot, et al. (Sept. 2011). "The Mechanism of Tail-Anchored Protein Insertion into the ER Membrane". In: *Molecular Cell* 43.5, pp. 738–750. DOI: 10.1016/j.molcel.2011.07.020.

Wang, Haixia, Qiang Zhang, and Dahai Zhu (Nov. 2003). "hSGT interacts with the N-terminal region of myostatin". In: *Biochemical and Biophysical Research Communications* 311.4, pp. 877–883. DOI: 10.1016/j.bbrc.2003.10.080.

Ward, Natalie and Gabriel Moreno-Hagelsieb (July 2014). "Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss?" In: *PLoS ONE* 9.7, e101850–6. DOI: 10.1371/journal.pone.0101850.

Waterhouse, Andrew M et al. (May 2009). "Jalview Version 2–a multiple sequence alignment editor and analysis workbench." In: *Bioinformatics* 25.9, pp. 1189–1191. DOI: 10.1093/bioinformatics/btp033.

Watkins, John F et al. (Dec. 1993). "The Saccharomyces cerevisiae DNA repair gene RAD23 encodes a nuclear protein containing a ubiquitin-like domain required for biological function." In: *Molecular and Cellular Biology* 13.12, pp. 7757–7765. DOI: 10.1128/mcb.13.12.7757-7765.1993.

Wattenberg, Binks W and Trevor Lithgow (Jan. 2001). "Targeting of C-Terminal (Tail)-Anchored Proteins: Understanding how Cytoplasmic Activities are Anchored to Intracellular Membranes". In: *Traffic* 2.1, pp. 66–71. DOI: 10.1034/j.1600-0854.2001.20108.x.

Webb, Benjamin and Andrej Sali (2016). *Comparative Protein Structure Modeling Using MODELLER*. Vol. 54. 1, pp. 5.6.1–5.6.37. DOI: 10.1002/cpbi.3.

Weill, Uri et al. (Dec. 2018). "Genome-wide SWAp-Tag yeast libraries for proteome exploration". In: *Nature Methods*, pp. 1–13. DOI: 10.1038/s41592-018-0044-9.

Weisman, Caroline M, Andrew W Murray, and Sean R Eddy (Nov. 2020). "Many, but not all, lineage-specific genes can be explained by homology detection failure." In: *PLoS Biology* 18.11, e3000862. DOI: 10.1371/journal.pbio.3000862.

Wideman, Jeremy G (2015). "The ubiquitous and ancient ER membrane protein complex (EMC): tether or not?" In: *F1000Research* 4, p. 624. DOI: 10.12688/f1000research.6944.2.

Wimley, William C, Trevor P Creamer, and Stephen H White (Apr. 1996). "Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides." In: *Biochemistry* 35.16, pp. 5109–5124. DOI: 10.1021/bi9600153.

Winnefeld, Marc et al. (Aug. 2006). "Human SGT interacts with Bag-6/Bat-3/Scythe and cells with reduced levels of either protein display persistence of few misaligned chromosomes and mitotic arrest". In: *Experimental Cell Research* 312.13, pp. 2500–2514. DOI: 10.1016/j.yexcr.2006.04.020.

Wunderley, Lydia et al. (Nov. 2014). "SGTA regulates the cytosolic quality control of hydrophobic substrates." In: *Journal of Cell Science* 127.Pt 21, pp. 4728–4739. DOI: 10.1242/jcs.155648.

Xing, Shuping et al. (2017). "Loss of GET pathway orthologs in Arabidopsis thaliana causes root hair growth defects and affects SNARE abundance". In: *Proceedings of the National Academy of Sciences* 114.8, E1544–E1553. DOI: 10.1073/pnas.1619525114.

Xu, Dong and Yang Zhang (July 2012). "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field." In: *Proteins* 80.7, pp. 1715–1735. DOI: 10.1002/prot.24065.

Xu, Yue, Mengli Cai, et al. (Dec. 2012). "SGTA Recognizes a Noncanonical Ubiquitin-like Domain in the Bag6-Ubl4A-Trc35 Complex to Promote Endoplasmic Reticulum-Associated Degradation". In: *Cell Reports* 2.6, pp. 1633–1644. DOI: 10.1016/j.celrep.2012.11.010.

Xu, Yue, Yanfen Liu, et al. (June 2013). "A Ubiquitin-like Domain Recruits an Oligomeric Chaperone to a Retrotranslocation Complex in Endoplasmic Reticulum-associated Degradation". In: *Journal of Biological Chemistry* 288.25, pp. 18068–18076. DOI: 10.1074/jbc.M112.449199.

Yamagata, Atsushi et al. (2010). "Structural insight into the membrane insertion of tail-anchored proteins by Get3". In: *Genes to Cells* 15.1, pp. 29–41. DOI: 10.1111/j.1365-2443.2009.01362.x.

Yamamoto, Yasunori and Toshiaki Sakisaka (Nov. 2012). "Molecular Machinery for Insertion of Tail-Anchored Membrane Proteins into the Endoplasmic Reticulum Membrane in Mammalian Cells". In: *Molecular Cell* 48.3, pp. 387–397. DOI: 10.1016/j.molcel.2012.08.028.

Yang, Jianyi, Ivan Anishchenko, et al. (2020a). “Improved protein structure prediction using predicted interresidue orientations”. In: *Proceedings of the National Academy of Sciences* 117.3, pp. 1496–1503. DOI: 10.1073/pnas.1914677117.

– (Jan. 2020b). “Improved protein structure prediction using predicted interresidue orientations.” In: *Proceedings of the National Academy of Sciences* 117.3, pp. 1496–1503. DOI: 10.1073/pnas.1914677117.

Yang, Jianyi, Renxiang Yan, et al. (Jan. 2015). “The I-TASSER Suite: protein structure and function prediction”. In: *Nature Methods* 12.1, pp. 7–8. DOI: 10.1038/nmeth.3213.

Yuan, Shuiqiao et al. (Apr. 2015). “Ubqln3, a testis-specific gene, is dispensable for embryonic development and spermatogenesis in mice.” In: *Molecular Reproduction and Development* 82.4, pp. 266–267. DOI: 10.1002/mrd.22475.

Zanata, Silvio M et al. (July 2002). “Stress-inducible protein 1 is a cell surface ligand for cellular prion that triggers neuroprotection.” In: *EMBO Journal* 21.13, pp. 3307–3316. DOI: 10.1093/emboj/cdf325.

Zeytuni, Natalie and Raz Zarivach (Mar. 2012). “Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module”. In: *Cell Structure and Function* 20.3, pp. 397–405. DOI: 10.1016/j.str.2012.01.006.

Zhang, Daoning, Shahri Raasi, and David Fushman (Mar. 2008). “Affinity Makes the Difference: Nonselective Interaction of the UBA Domain of Ubiquilin-1 with Monomeric Ubiquitin and Polyubiquitin Chains”. In: *Journal of Molecular Biology* 377.1, pp. 162–180. DOI: 10.1016/j.jmb.2007.12.029.

Zhao, Gang and Erwin London (Aug. 2006). “An amino acid “transmembrane tendency” scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity”. In: *Protein Science* 15.8, pp. 1987–2001. DOI: 10.1110/ps.062286306.

Zientara-Rytter, Katarzyna and Suresh Subramani (Jan. 2019). “The Roles of Ubiquitin-Binding Protein Shuttles in the Degradative Fate of Ubiquitinated Proteins in the Ubiquitin-Proteasome System and Autophagy”. In: *Cells* 8.1, pp. 40–32. DOI: 10.3390/cells8010040.

Zopf, Dieter et al. (1990). “The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be crosslinked to a signal sequence.” In: *The EMBO Journal* 9.13, pp. 4511–4517. DOI: 10.1002/j.1460-2075.1990.tb07902.x.