The client-binding domain of the cochaperone SGTA/Sgt2 has a helical-hand structure that binds a short hydrophobic helix

(120 of 120)

Ku-Feng Lin, Michelle Fry*, Shyam Saladi*, and William M. Clemons, Jr.

* contributed equally to this work

Correspondence: W.M.C, clemons@caltech.edu
Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

Keywords: Tail-anchored proteins, Protein targeting, co-chaperones, Ubiquilins, Sti1, Hop

Abstract (150 words – currently 149)

The correct targeting and insertion of tail-anchored (TA) integral membrane proteins (IMP) is critical for cellular homeostasis. The mammalian protein SGTA, and its fungal homolog Sgt2 (Sgt2/A), binds hydrophobic clients and is the entry point for targeting of ER-bound TA IMPs. Here we reveal molecular details that underlie the mechanism of Sgt2/A binding to TA clients. We establish that the Sgt2/A C-terminal region is conserved but flexible, sufficient for client binding, and has functional and structural similarity to the DP domains of Sti1. A molecular model for Sgt2/A-C reveals a helical hand forming a hydrophobic groove, consistent with a higher affinity for TA clients with hydrophobic faces and a minimal length of 11 residues. Finally, we show that a hydrophobic face metric improves the predictions for TA localization *in vivo*. The structure and binding mechanism positions Sgt2/A into a broader class of helical-hand domains that reversibly bind hydrophobic clients.

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 [1], 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 [2-4]. 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 targeted post-translationally to either the ER or mitochondria [5-9]. In the case of the canonical pathway for ER-destined TA IMPs, each is first recognized by homologs of mammalian SGTA (small glutamine tetratricopeptide repeat protein) [4,6,10,11]. 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 the human SGTA and fungal Sgt2 (hereafter referred to as *Hs*SGTA for *Homo sapiens* and *Sc*Sgt2 fo*r Saccharomyces cerevisiae*, collectively Sgt2/A) are involved in a variety of cellular processes regarding the homeostasis of membrane proteins including the targeting of TA IMPs [9,12-14], retrograde transport of membrane proteins for ubiquitination and subsequent proteasomal degradation [15], and regulation of mislocalized membrane proteins (MLPs) [16,17]. Among these, the role of Sgt2/A in the primary pathways responsible for targeting TA clients to the endoplasmic reticulum (ER) are 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 client then transferring the TA to the ATPase chaperone Get3 (human TRC40) with the aid of the heteromeric Get4/Get5 complex (human TRC35/Ubl4A/Bag6 complex) [13,18-20]. In this process, TA binding to Sgt2, after hand-off from Hsp70, is proposed as the first committed step to ensure that ER, but not mitochondrial, TAs are delivered to the ER membrane [3,13,21]. Subsequent transfer of the TA 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 complex to the Get1/2 (mammalian CAML/WRB) receptor at the ER leading to release of the TA into the membrane [22-26]. Deletions of GET genes (*i.e. get1*Δ, *get2*Δ, or *get3*Δ) cause cytosolic aggregation of TAs

dependent on Sgt2 [26,27].

In addition to targeting TA IMPs, SGTA may also promote degradation of IMPs through the proteasome by cooperating with the Bag6 complex, a heterotrimer containing Bag6, TRC35, and Ubl4A, which acts as a central hub for a diverse physiological network related to protein targeting and quality control [19,28-30]. 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 SGTA to ER associated degradation (ERAD) and proteasomal activity. Depletion of SGTA significantly inhibits turnover of ERAD IMP clients and elicits the unfolded protein response[16]. Furthermore, the cellular level of MLPs in the cytoplasm could be maintained by co-expression with SGTA, which possibly antagonize ubiquitination of MLPs to prevent proteasomal degradation [15,17]. These studies demonstrate an active role of SGTA in triaging membrane proteins in the cytoplasm and the breadth of SGTA clients including TAs, ERAD clients, and MLPs all harboring one or more TMD. SGTA roles in disease have been linked to polyomavirus infection [31], neurodegenerative disease [27,32], hormone-regulated carcinogenesis [33,34], and myogenesis [35], although the underlying molecular mechanisms are still unclear.

The architecture of Sgt2/A includes three structurally independent domains that define the three different interactions of Sgt2/A (Fig 1A) [12,36-39]_ENREF_19. 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 targeting [36,40] or interacts with the UBL on the N-terminal region of BAG6 [41] where it is thought to initiate downstream degradation processes [15,28,29]. 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) [12,42]. The highly conserved TPR domain was demonstrated to be critical in modulating propagation of yeast prions by recruiting HSP70 [27] and may associate with the proteasomal factor Rpn13 to regulate MLPs [43]. More recently, it was demonstrated that mutations to residues in the TPR domain which prevent Hsp70 binding impair the loading of TA IMPs onto Sgt2 in yeast [21], consistent with a direct role of Hsp70 in TA IMP targeting via the TPR domain. The C-terminal methionine-rich domain of Sgt2/A is responsible for binding to hydrophobic clients such as TA IMPs [11,37,44]. Other hydrophobic segments have been demonstrated to interact with the domain such as the membrane protein Vpu (viral protein U)

from human immunodeficiency virus type-1 (HIV1), the TMD of tetherin [44], the signal peptide of myostatin [35], and the N-domain of the yeast prion forming protein Sup35 [27]. All of these studies suggest that the C-terminus of Sgt2/A 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/A (Sgt2/A-C) and show that, in the absence of substrate, they are relatively unstructured. We demonstrate 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 crucial for client binding. Combining this with *ab initio* structure prediction and biochemistry in total demonstrates that C_{cons} has structural homology to the client-binding domain of the co-chaperone Sti1/Hop. Artificial TA clients are used to define the properties critical for binding to Sgt2/A-C. We further show that these principles extend to the TA proteome and are sufficient to properly categorize the cellular localization of TA clients. Finally, the combined results lead to a mechanistic model where Sgt2/A-C falls into the broader class of helical-hand containing proteins involved in the binding and release of hydrophobic alpha-helices.

Results

The flexible Sgt2/A-C domain

Based on sequence alignment (Fig. 1*A*), the C-domain of Sgt2/A contains a conserved core of six predicted helices flanked by unstructured loops that vary in length and sequence. Previous experimental work had suggested that this region was particularly flexible, as the domain in the *Aspergillus fumigatus* was sensitive to proteolysis [12]. Similarly, for *Sc*Sgt2-TPR-C, the sites sensitive to limited proteolysis primarily occur within the loops flanking the conserved helices (Fig. 1*A, red arrows* and S1*B*). This flexible nature of the C-domain likely contributes to its anomalous passage through a gel-filtration column where *Sc*Sgt2/A-C elutes much earlier than the similarly-sized, but well-folded Sgt2/A TPR-domain (Fig. 1*B*). The circular dichroism (CD) spectra for both homologs suggests that the C-domain largely assumes a random-coil conformation, with 40-45% not assignable to a defined secondary structure category (Fig. 1*C*) [45]. This lack of stable tertiary structure is further highlighted by the well-resolved, sharp, but narrowly dispersed chemical shifts of the backbone amide protons in ^{1}H-^{15}N HSQC spectra of Sgt2/A-C (Fig. 1*D,E*), indicating a significant degree of backbone mobility, similar to natively unfolded proteins [46] and

consistent with results seen by others [47]. The larger hydrodynamic radius matches previous small-angle X-ray scattering measurement of the TPR-C protein that indicated a partial unfolded characteristic in a Kratky plot analysis [12].

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

Several lines of evidence suggest the conserved region of the C-domain binds substrates. First, during purification the Sgt2-C-domain was cut at several specific sites (Fig. 1*A*). Proteolysis occurred primarily in the poorly-conserved N-terminal region (between Asp_{235}-Gly_{258}) and at Leu_{327}. This suggests that the intervening conserved region, Gly_{258} and Leu_{327} on *Sc*Sgt2 and its corresponding region on *Hs*SGTA, may mediate TA client binding (Fig. 2*A, grey*). To test this, various his-tagged Sgt2/A constructs were co-expressed with MBP-tagged TA client (Sbh1) and binding was detected by the presence of captured TA by various Sgt2/A constructs bound to the affinity resin (Fig 2*B*). As previously seen [13], we confirm that the Sgt2/A-TPR-C alone is sufficient for capturing a TA client (Sbh1) (Fig. 2*B*). As one might expect, the C-domain was also sufficient for binding the TA client. The central region of Sgt2/A-C contains six conserved helixes, hereafter referred to as Sgt2/A-C_{cons}, and is sufficient for binding to the TMD of Sbh1 with an N-terminal MBP-tag. For Sgt2, the minimal conserved region H1-H5 (ΔH0) poorly captures a TA client, while in SGTA this minimal region is sufficient for capturing the client at a similar level as the longer C_{cons} domain (Fig. 2*D*).

The six predicted helices in Sgt2/A-C_{cons} are amphipathic (Fig. 1*A* and Fig. 2*C*) suggesting that they use the hydrophobic faces of the helices for binding to client. To probe this, each of these helices 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/A (Fig. 2*E*). While these mutants expressed at similar levels to the wild-type sequence, one cannot rule out that these changes do not broadly affect the tertiary structure of this domain. In general, these results imply that these amphipathic helices are directly involved in the interaction with client. The overall effect on binding by each helix is different with mutations in the helices 1-3 having the most dramatic reduction in binding suggesting that these are more crucial for TA complex formation. It is also worth noting, as this is a general trend, that SGTA is more resistant to mutations that affect binding than Sgt2 which likely represents different threshold requirements.

155

156 **Molecular modeling of Sgt2/A-C domain**

157 Despite the need for a molecular model, the C-domains have resisted structural studies, likely
158 due to the demonstrated inherent flexibility. With six conserved α-helical amphipathic segments
159 (Fig. 1*A*) containing hydrophobic residues critical for TA-client binding (Fig. 2*C,E),* we expect
160 some folded structure to exist. Therefore, we performed *ab initio* molecular modeling of Sgt2-C
161 using a variety of prediction methods [48-51] resulting in a diversity of putative structures. Of the
162 various models, only the highest scored structures from Quark [48] consistently result in a similar
163 tertiary fold (Fig. 3*A*). The general architecture contained a clear potential TA client binding site,
164 a hydrophobic groove formed by the amphipathic helices. The groove is approximately 15 Å long,
165 12 Å wide, and 10 Å deep, which is sufficient to accommodate three helical turns of an alpha-helix,
166 ~11 amino acids (Fig. 3*B*). For the prediction, while the entire C-domain was used, the N- and C-
167 termini of Sgt2 do not adopt similar structures across the various models consistent with their
168 expected higher flexibility (Fig. 3*C*).

169 To validate the model, we interrogated the accuracy of the predicted spatial location of the
170 helices by experimentally determining distance constraints from crosslinking experiments. Based
171 on the model, four pairs of residues in close spatial proximity and one pair far-apart were selected
172 and mutated to cysteines (Fig. 3*D*). In the experiment, an artificial TA client containing a cMyc-
173 tagged BRIL (small, 4-helix bundle protein [52]) with a C-terminal TMD consisting of eight
174 leucines and three alanines is co-expressed with Sgt2-TPR-C cysteine variants, purified, and then
175 oxidized to form disulfide crosslinks if the residues are near each other [53]. Crosslink formation
176 is identified by comparing the products after protease digestion where bond formation results in a
177 reducing-agent sensitive ~7.7 kDa fragment (Fig. 3*D*). For the wild-type (cysteine-free) sequence,
178 no higher molecular weight bands are observed at ~7.7kDa. For the N285/G329 pair which is too
179 distant for disulfide bond formation, no higher band is observed. For the remaining pairs that are
180 predicted to be close enough for bond formation, the 7.7kDa fragment is observed in each case
181 and is labile in reducing conditions. These results support the structures obtained in the Quark
182 derived C_{cons} model.

183

184 **Structural similarity of Sgt2/A-C domain to STI1 domains**

185 Attempts to glean functional insight for Sgt2/A-C from BLAST searches did not reliably

186 return other families or non Sgt2/A homologs making functional comparisons difficult. A more
187 extensive domain-based search using definitions from the similar modular architecture database
188 (SMART) [54] identified a similarity to domains in the yeast co-chaperone Sti1. First called DP1
189 and DP2 due to their prevalence of aspartates (D) and prolines (P), these domains have been shown
190 to be required for client-binding [55,56] and are termed 'STI1'-domains in bioinformatics
191 databases [54]. In yeast Sti1, and its human homolog Hop, each of the two STI1 domains (DP1
192 and DP2) are preceded by Hsp70/90-binding TPR domains, similar to the domain architecture of
193 Sgt2/A. Deletion of the second, C-terminal STI1-domain (DP2) from Sti1 *in vivo* is detrimental,
194 impairing native activity of the glucocorticoid receptor [55]. *In vitro*, removal of the DP2 domain
195 from Sti1 results in the loss of recruitment of the progesterone receptor to Hsp90 without
196 interfering in Sti1-Hsp90 binding [57]. These results implicate DP2 in binding of Sti1 clients. In
197 addition, others have noted that, broadly, STI1-domains may present a hydrophobic groove for
198 binding hydrophobic segments of a client [55,56]. Furthermore, the similar domain organizations
199 (*i.e.* Sgt2/A TPR-C, Sti1 TPR-STI1) and molecular roles could imply an evolutionary relationship
200 between these co-chaperones. Indeed, a multiple sequence alignment of the Sgt2-C_{cons} with several
201 yeast STI1 domains (Fig. 4*A*) reveals strong conservation of structural features. H1-H5 of the
202 predicted helical regions in the C_{cons} align directly with the structurally determined helices in the
203 DP2 domain of Sti1; this includes complete conservation of helix breaking prolines and close
204 alignment of hydrophobic residues defining amphipathic helices [55].
205 Based on the domain architecture and homology, we believe it is reasonable to make a direct
206 comparison between the STI1 domain and Sgt2/A-C_{cons}. A structure of DP2 solved by solution
207 NMR reveals that the five amphipathic helices assemble to form a flexible helical-hand with a
208 hydrophobic groove [55]. The lengths of the alpha helices in this structure concur with those
209 inferred from the alignment in Fig. 4*A*. Our molecular model of Sgt2-C_{cons} is strikingly similar to
210 this DP2 structure. An overlay of the DP2 structure and our molecular model in Fig. 4*C*
211 demonstrates both Sgt2-C_{cons} and DP2 have similar lengths and arrangements of their amphipathic
212 helices (Fig. 4*B,C* and Fig. S3). Consistent with our observations of flexibility in Sgt2/A-C_{cons},
213 Sti1-DP2 generates few long-range NOEs between its helices indicating that Sti1-DP2 also a
214 flexible architecture [55]. We consider this flexibility a feature of these helical-hands for reversible
215 and specific binding of a variety of clients.
216

Binding mode of TA clients to Sgt2/A

We examined the C_{cons} surface that putatively interacts with TA clients by constructing hydrophobic-to-charge residue mutations that are expected to disrupt capture of TA clients by Sgt2/A. Similar to the helix mutations in Figure 2, the capture assay was employed to establish the relative effects of individual mutations. A baseline was established based on the amount of TA-client Sbh1 captured by wild-type Sgt2/A-C. In each experiment, Sbh1 expresses at the same level; therefore, differences in binding should directly reflect the affinity of Sgt2/A mutants for clients. In all cases, groove mutations from hydrophobic to aspartate led to a reduction in TA client binding (Fig. 5*A* and *B*). The effects are most dramatic in Sgt2 where each mutant significantly reduced binding by 60% or more (Fig. 5*A*). While all SGTA individual mutants saw a significant loss in binding, the results were subtler with the strongest being only ~36% (Fig. 5*B*). Double mutants were stronger with a significant decrease in binding relative to the individual mutants, more reflective of the individual mutants in Sgt2. As seen before (Fig. 2), we observe that mutations toward the N-terminus of Sgt2/A-C have a stronger effect on binding than those later in the sequence.

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

With a molecular model for Sgt2/A-C_{cons} and multiple lines of evidence for a hydrophobic groove, we sought to better understand the specific requirements for TMD binding. To probe this, Sgt2/A-TPR-C complex binding with designed TA clients where a number of variables are tested including the overall (sum) and average (mean) TMD hydrophobicity, length of the TMD, and the distribution of hydrophobic character within a TMD. These artificial TMDs were constructed as C-terminal fusions with the architecture cMyc-tag, cytoplasmic BRIL, a hydrophilic linker (Gly-Ser-Ser), and the TMD (Leu/Ala helical stretch followed by a Trp) (Fig. 6*A*). The total and mean hydrophobicity are controlled by varying the helix-length and Leu/Ala ratio (1.82/0.38 TM tendency hydrophobicity values). For clarity, we define a syntax for the various artificial TA 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 constructs was to define the minimal length for a TMD to bind to the C-domain. As described earlier, capture of His-tagged Sgt2/A-TPR-C with the various

TA clients were performed. We define a relative binding efficiency as the ratio of captured TA client by a Sgt2/A variant normalized to the ratio of a captured WT TA client by the same Sgt2/A variant, in this case the model ER-bound Bos1. The client 18[L13] shows a comparable binding efficiency to Sgt2/A-TPR-C as that of Bos1 (Fig. 6*B*). From the helical wheel diagram of the TMD for Bos1, we noted that the hydrophobic residues align on one side of the helix. Therefore, we optimized our various model clients to contain a 'hydrophobic face' while shortening the length and maintaining the average hydrophobicity of 18[L13] (Fig. 6*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/A (Fig. 6*B*) establishing a minimal length of 11 residues for the helix consistent with the dimensions of the groove predicted for the structural model (Fig. 3).

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/A and artificial TA clients containing an 11 amino acid TMD with increasing number of leucines (11[Lx]). As shown in Fig. 6*C*, increasing the number of leucines monotonically enhances complex formation, echoing previous results [58]. *Hs*SGTA binds to a wider spectrum of hydrophobic clients than *Sc*Sgt2, which could mean it has a more permissive hydrophobic binding groove as reflected by the milder impact of alanine replacement and Asp mutations in SGTA-C to TA binding (Fig. 2*C* and Fig. 5*A*).

Sgt2/A-C preferentially binds to TMDs with a hydrophobic face

Next, we address the properties within the TMD of TA clients responsible for Sgt2/A binding. In the case of Sgt2/A, it has been suggested that the co-chaperone binds to TMDs based on hydrophobicity and helical propensity [58]. For the most part, varying the hydrophobicity of an artificial TA client acts as expected, the more hydrophobic TMDs bind more efficiently to Sgt2/A TPR-C domains (Fig. 6*C*). Our C_{cons} model suggest the hydrophobic groove of Sgt2/A-C protects a TMD with highly hydrophobic residues clustered to one side (see Fig. 3*B*). Helical wheel diagrams demonstrate the distribution of hydrophobic residues along the helix (*e.g.* bottom Fig. 6*D*). Testing various TMD pairs with the same hydrophobicity, but different distributions of hydrophobic residues demonstrates TA clients with clustered leucines have a higher relative binding efficiency than those with a more uniform distribution (Fig. 6*D*). The clustered leucines

on the TMDs create a hydrophobic face which potentially interacts with the hydrophobic groove formed by the Sgt2/A-C_{cons} region, corresponding to the model in Fig. 3*B*.

Organization of hydrophobic residues in probable client TMDs

So far, the interpretation from the structure that Sgt2/A-C binding to clients via a hydrophobic groove is supported by the binding preferences of Sgt2/A-TPR-C. As Sgt2/A is the entry point into TA IMP targeting to the ER, we were interested in whether TMD hydrophobic faces were relevant to sorting of TA clients in the cell. Previous results demonstrate that hydrophobicity is a dominant factor in selection between the ER and mitochondria [59]; therefore, the reference yeast and human genomes from UniProt [60] were screened for putative TA IMPs and filtered for unique genes longer than 50 residues. Uniprot and TOPCONS2 [61] were used to identify genes that encoded an IMP containing a single TMD within 30 amino acids of the C-terminus [62] and lacked a predicted signal sequence (as determined by SignalP4.1 [63]) (Fig. 7*A* and Table S1). Based on their UniProt-annotated localizations [60], TA IMPs are subcategorized as ER, mitochondrial, peroxisomal, and unknown. While our set encompasses proteins previously predicted as TA IMPs [64,65], it is larger and we believe a more accurate representation of the repertoire of TA IMPs found in each organism. For both yeast and humans, the majority of proteins have no annotated cellular localization. Several previously suggested TA IMPs 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 (full list in Table S1) [59,66].

Broadly, hydrophobicity is considered a dominant feature for discriminating TA IMP localization with those that contain more hydrophobic TMDs localizing to the ER [67]. We explore this in Fig. 7*B*, where the hydrophobicity of the entire TMD for each yeast TA IMP was calculated using the TM tendency scale [68] and is plotted along the y-axis. If we only consider proteins known to localize to the ER or mitochondria, this analysis classifies the majority of the proteins correctly at a best threshold of 16.8 (red dashed line, Fig. 7*B*). While all five mitochondrial TA IMPs are correctly classified, a significant number of ER-bound TAs contain a TMD with a hydrophobicity lower than the threshold (Fig. 7*B*). A notable misclassified example is Sss1 (Sec61γ in chordates) of the ER residing Sec translocon.

We next considered whether the hydrophobic face preference of Sgt2/A might be reflected in

the ability to classify TA IMPs. For yeast, we calculated the maximum hydrophobicity of a helical face of six amino acids and plotted this value (x-axis, Fig. 7*B*). ER targeted TA IMPs are best classified by a helical face threshold of 7.7 (Fig. 7*B*, cyan dotted line). While both metrics correctly categorize mitochondria-bound TA IMPs (all in lower left quadrant), the helical-face metric better categorizes low hydrophobicity ER bound TA IMPs, *e.g.* Sbh1 is now correctly classified as ER-localized (Fig 7*A*). More quantitatively (Fig. 7*C*), as a predictor the AUROC value for classification based on the hydrophobicity of a single face (AUROC = 0.99) is higher than that based on hydrophobicity of the entire TMD (AUROC = 0.87), supporting the relevance of a hydrophobic face in TA IMP targeting by Sgt2/A.

We then applied this analysis to the 587 putative human TA IMPs. Again, proteins were plotted based on the hydrophobicity of the entire TMD (y-axis) and the most hydrophobic face (x-axis) and colored based on UniProt-annotated cellular localization (Fig. 7*D*). The best thresholds determined by our analysis (overall 19.8 and face 9.3) again show that Sec61γ continues to only be correctly categorized by the hydrophobicity of its helical face. As with yeast proteins, an increase in AUROC value was observed when clients were classified based on the hydrophobicity of single face (AUROC = 0.82) instead of the entire TMDs (AUROC = 0.79). With human TA IMPs, a metric focusing on a sufficiently hydrophobic face does just as well if not better than a metric focusing on the hydrophobicity of the TMD. The moderate improvement in predictive capacity likely reveals the higher complexity of the human system and the milder effect of mutants to *Hs*SGTA-C on binding to TA clients.

Interestingly, by considering the hydrophobic face, more information can be gleaned about complex clients that localize to both the mitochondria and ER. Notable examples are members of the Bcl-2 family, which play critical roles in the apoptosis pathway [69,70]. Although many have been reported to localize to several organelles in the cell, some have a preferred localization [69,70]. For example, Bcl-xL has been reported to localize predominantly to the mitochondria, though a fraction of its cellular concentration has been observed to be present in the ER. The case is similar for McL1 [71] and Bcl-B [70,72]. Classified by their hydrophobic face, these proteins are predicted to be mitochondria-bound clients (blue, Fig. 7*D*). Unlike Bcl-xL, the majority of cellular Bok, another Bcl-2 family member, is found in the ER or Golgi [73]. The hydrophobic face metric classifies Bok as an ER bound protein whereas a metric based on the hydrophobicity of the entire TMD misclassifies it as a mitochondrial protein (Fig. 7*D*). This suggests our metric can correctly

341 determine the primary localization of members of the Bcl-2 family TA IMPs, important insight for
342 these medically relevant proteins.

343 Another interesting case for the identification and localization of TA IMPs is the apparent
344 lack of the protein squalene synthase (SQS) in our list, previously used as a model TA [66]. Since
345 SQS is predicted to have two TMDs, it is excluded by the criteria above. However, structural
346 studies of SQS have clearly identified the predicted first TMD to instead be a helical component
347 of the folded soluble domain [74]; therefore, the protein only contains a single TMD at the C-
348 terminus which would fit the standard definition of a TA IMP. Once again, if we consider the
349 human protein SQS and where its TMD falls on the localization metrics (Fig. 8*C*, red x), the TM
350 tendency of its entire TMD (12.5) predicts it to be mitochondrial while considering the most
351 hydrophobic face (9.9) accurately captures its ER localization. How this protein fits into our
352 understanding of ER localized TA IMPs is discussed below. Future refinement of our
353 bioinformatics screen to include details such as known or predicted structure may further hone the
354 list of putative TA IMPs (Table S1).
355

356 ## Discussion

357 Sgt2/A, the most upstream component of the GET/TRC pathway, plays a critical role in the
358 correct insertion of TA IMPs into their designated membranes. Its importance as the first selection
359 step of ER versus mitochondrial bound TA clients necessitates a molecular model for TA client
360 binding. Previous work demonstrated a role for the C-domain of Sgt2/A to bind to hydrophobic
361 clients, yet the exact binding domain remained to be determined. Through the combined use of
362 biochemistry, bioinformatics, and computational modeling, we conclusively identify the minimal
363 client-binding domain of Sgt2/A. This allowed us to present a validated structural model of Sgt2/A
364 C-domain as a methionine-rich helical hand for grasping a hydrophobic helix providing a
365 mechanistic explanation for binding a minimum TMD of 11 hydrophobic residues with the most
366 hydrophobic residues organized onto one face of the helix.

367 Based on these results, we can confidently identify that the C-domain of Sgt2/A contains a
368 STI1 domain for client binding. This places the protein into a larger context of both conserved co-
369 chaperones and adaptors of the ubiquitin-proteasome system (AUPS) (Fig. 8*A*). For the co-
370 chaperone family, the STI1 domains predominantly follow HSP-binding TPR domains connected
371 by a flexible linker. As noted above, it was demonstrated that Sti1/Hop domains are critical for

client-processing and coordinated hand-off between Hsp70 and Hsp90 homologs [75]. Additionally, multiple TPR domains of Sti1/Hop are used to coordinate simultaneous binding of two heat shock proteins. Both Sgt2/A and the co-chaperone Hip share the coordination of two TPR and STI1 domains by forming stable dimers via N-terminal dimerization domains [76]. With evidence for a direct role of the carboxylate-clamp in the TPR domain of Sgt2/A for client-binding now clear [21], one can speculate that the two TPR domains may facilitate TA client entry into other pathways using multiple heat shock proteins. The more distant chloroplast Tic40 contains two putative STI1 domains [77,78] (Fig. 8*A*), with the C-terminal one having a structure clearly similar to that of other co-chaperones (Fig. S2*D*). The rest of the protein has a different domain architecture as it lacks a clear TPR domain [78] and has an N-terminal TMD. Found in the inner chloroplast membrane with the STI1 domain(s) in the stroma, the C-terminal domain can be replaced with the STI1 domain from Hip without loss-of-function [79]. How Tic40 fits mechanistically into this group is less clear.

As annotated, STI1 domains broadly share several features including four to five amphipathic helices (Fig. 8*A* and Fig. S2*A,B*). For structurally characterized domains, these organize into helical hands with a hydrophobic groove (Fig. 8*B* and Fig. S2). In the co-chaperones, all of the domains have the same architecture and are characterized by structural flexibility in the absence of client. While there are no structures of client-bound STI1 domains for this group, the H0 helix in the structure of the DP1 domain from Sti1 likely mimics client binding (grey helix in Fig. 8*B* and Fig. S2*A*). This N-terminal amphipathic helix is conserved among co-chaperone STI1 domains (Fig. S2*A,C*) and the additional helix may be a general feature. Structurally, the co-chaperone STI1 domains contain five core amphipathic helices. Bioinformatics databases, like SMART, use this definition, which can lead to erroneous annotations of putative STI1 domains. The clearest case for this is the two pairs of abutting STI1 motifs predicted for UBQLN -1, -2, & -4. Careful analysis reveals a N-terminal sixth amphipathic helix. When this is considered, it is clear that the abutting STI1 domains are instead a single domain (Fig. S2*B*). While the roles of the additional helices are not clear, they are well conserved within each family. A possible speculation is that they perhaps acts as a lid for protecting the empty groove and/or set the hydrophobic threshold for client-binding, as predicted for other TMD binders [4]. For the AUPS proteins, the only known structure of a STI1 domain comes from the DNA damage response protein Rad23. For this domain, the architecture is different with only four helices that form a different hydrophobic groove for recognition of

clients (Fig. 8*B*). In fact, this difference is underscored by the poor alignment between Rad23 with STI1 domains (Fig. S2*B,D*). Nonetheless, several structures of complexes of Rad23-STI1 bound to amphipathic clients show in each that the client-helix binds via a hydrophobic face (Fig. 8*B* and Fig. S2*D*). Perhaps this represents a second class of STI1-like domains that could include proteins such as Ddi1 [80,81].

The concept of TMD binding by a helical hand is reminiscent of other proteins involved in membrane protein targeting. Like Sgt2/A, 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 [82]. Here once more, the domain has been observed to be flexible in the absence of client [83,84] and, in the resting state, occupied by a region that includes a helix that must be displaced [82]. Another helical-hand example recently shown to be involved in TA-protein targeting is calmodulin where two helical hands coordinate to clasp a TMD from either side (Fig. 9B). Considering an average TMD of 18-20 amino acids (to span a ~40Å bilayer), each half of calmodulin interacts with about 10 amino acids. The close correspondence of this value with the minimal binding length for Sgt2/A C-domain leads one to speculate that the two copies of the Sgt2/A C-domain in the dimer may work together to bind to a full TMD. Cooperation of the two Sgt2/A C-domains in client-binding could elicit conformational changes in the complex that would be recognized by downstream factors, such as increasing the affinity for Get5/Ubl4A. Paired STI1 domains in UBQLN-1, -2, & -4 may cooperate as well. Recently, others noted the ability of the SGTA C-domain to independently dimerize in certain conditions, also hinting at a model of cooperation between across the dimers for client binding [47]. While we see no evidence for dimerization of the C-domain in our constructs, it is clear that interactions between C-domains are likely important.

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, the flexible helical-hands provide an extensive hydrophobic surface to capture the client-helix—driven by the hydrophobic effect. Typically, such extensive interfaces are between pairs of pre-ordered surfaces resulting in very stable binding. Required to only engage temporarily, the flexibility of the helical hand offsets the favorable free energy of binding by charging an additional entropic cost from the need to transition from a flexible unbound form to

that in the client-bound complex. This would account for the favorable transfer seen from Sgt2 [21] and SGTA [85] to downstream components.

While SGTA and Sgt2 share many properties, there are a number of differences between the two proteins that may explain the different biochemical behavior. For the C_{cons}-domains, SGTA appears to be more ordered in the absence of client as the peaks in its NMR spectra are broader (Fig. 1*E*). Comparing the domains at the sequence level, while the high glutamine content in the C-domain is conserved it is higher in SGTA (8.8% versus 15.2%). The additional glutamines are concentrated in the predicted longer H4 helix (Fig. 1*A*). The linker to the TPR domain is shorter compared to Sgt2 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 SGTA than Sgt2 (Fig. 1*E*, 5, and 6) implying that SGTA 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 SGTA C-domain more ordered in the absence of client.

An interesting exception is SQS, which is a client of the EMC, rather than the TRC pathway [66]. The EMC pathway is characterized as targeting ER TA clients of lower hydrophobicity due to a higher affinity of its chaperone calmodulin for these clients over SGTA. Based on experimental results, a threshold for EMC dependence lies approximately at 21.6 [66], slightly higher than the overall hydrophobicity cut-off noted here for ER prediction (Fig. 7*C*). By this metric, mitochondrial and EMC dependent TA clients are indistinguishable. Putative EMC client localization is more accurately predicted by the hydrophobic face metric (ER proteins in the lower right quadrant of Fig. 7*D*). The increased hydrophobicity of TRC/GET pathway clients results in more hydrophobic residues in their TMDs leading to consistently higher values in the hydrophobic face metric. Yet, our analysis reveals the importance of a hydrophobic face for discriminating ER versus mitochondria targeted TAs with low hydrophobicity. As current evidence favors a dependence on the EMC pathway for the ER proteins, one might speculate either a continued role for SGTA for these clients or that the helical-hands of calmodulin also favor hydrophobic face binding. The latter seems unlikely as a discriminatory step as calmodulin is a generalist in client binding [86]. In the absence of calmodulin, SGTA is sufficient for delivering TA clients to the EMC [66] and perhaps acts upstream of calmodulin to discriminate between ER and mitochondrial targets.

The targeting of TA clients 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 focus on a central player, Sgt2/A and its client-binding domain. Through biochemistry and computational analysis, we provide more clarity to client discrimination. A major outcome of this is the clear preference for a hydrophobic face on ER TA IMPs of low hydrophobicity. In yeast, this alone is sufficient to predict the destination of a TA IMP. In mammals, and likely more broadly in metazoans, while clearly an important component, alone the hydrophobic face cannot fully discriminate targets. For a full understanding, we expect other factors to contribute reflective of the increased complexity of higher eukaryotes, perhaps involving more players [87]. Suffice to say, this study highlights the important role of Sgt2/A in TA IMP biogenesis.

Material and Methods

Plasmid constructs

MBP-Sbh1, *Sc*Sgt2$_{95\text{-}346}$ (*Sc*Sgt2-TPR-C), *Sc*Sgt2$_{222\text{-}346}$ (*Sc*Sgt2-C), *Sc*Sgt2$_{260\text{-}327}$ (*Sc*Sgt2-C$_{cons}$), *Sc*Sgt2$_{266\text{-}327}$ (*Sc*Sgt2-ΔH0), *Hs*SGTA$_{87\text{-}313}$ (*Hs*SGTA-TPR-C), *Hs*SGTA$_{213\text{-}313}$ (*Hs*SGTA-C), *Hs*SGTA$_{219\text{-}300}$ (*Hs*SGTA-C$_{cons}$), and *Hs*SGTA$_{228\text{-}300}$ (*Hs*SGTA-ΔH0) were prepared as previously described [12,88]. Genes of *ScSgt2* or *Hs*SGTA 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/A were constructed by site-direct mutagenesis. Artificial TAs were constructed in a pACYC-Duet plasmid with a N-terminal cMyc tag, BRIL protein [89], GSS linker, and a hydrophobic C-terminal tail.

Protein expression and purification

All proteins were expressed in *Escherichia coli* NiCo21 (DE3) (New England BioLabs). To co-express multiple proteins, constructed plasmids were co-transformed as described [88]. Protein expression was induced by 0.3 mM IPTG at OD_{600} ~ 0.7 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, PMSF, and 10 mM β-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 *Sc*Sgt2-TPR-C for MS analysis, PMSF and benzamidine were excluded from the lysis buffer. His-Sgt2/A and their 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 β-ME, pH 7.5. The complexes of interest were eluted in 20 mM Tris, 150 mM NaCl, 300 mM imidazole, 10 mM β-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/A protein concentration was carried out using the bicinchoninic acid (BCA) assay with bovine serum albumin as standard (Pierce Chemical Co.). Samples for NMR and CD analyses were concentrated to 10-15 mg/mL for storage at −80°C before experiments.

NMR Spectroscopy

^{15}N-labeled proteins were generated from cells grown in auto-induction minimal media as described [90] and purified in 20 mM phosphate buffer, pH 6.0 (for ScSgt2-C, 10mM Tris, 100mM NaCl, pH 7.5). The NMR measurements of ^{15}N-labeled Sgt2/A-C proteins (~0.3-0.5 mM) were collected using a Varian INOVA 600 MHz spectrometer at either 25°C (*Sc*Sgt2-C) or 35°C (*Hs*SGTA-C) with a triple resonance probe and processed with TopSpin™ 3.2 (Bruker Co.).

CD Spectroscopy

The CD spectrum was recorded at 24°C with an Aviv 202 spectropolarimeter using a 1 mm path length cuvette with 10 μM protein in 20 mM phosphate buffer, pH 7.0. The CD spectrum of each sample was recorded as the average over three scans from 190 to 250 nm in 1 nm steps. Each spectrum was then decomposed into its most probable secondary structure elements using BeStSel [91].

Glu-C digestion of the double Cys mutants on *Sc*Sgt2-C

Complexes of the co-expressed wild type or double Cys mutated His-*Sc*Sgt2-TPR-C and the artificial TA, 11[L8], were purified as the other His-Sgt2/A complexes described above. The protein solutions were mixed with 0.2 mM $CuSO_4$ and 0.4 mM 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.

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 (50 mM Tris-HCl pH 7.4, 150 mM 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) and 1:5000 anti-cMyc mouse monoclonal (Sigma). A secondary antibody conjugated to alkaline phosphatase (Rockland, 1:8000) was employed, and the blotting signals were chemically visualized with NBT/BCIP (Sigma). All blots were photographed and quantified by image

534 densitometry using ImageJ [92] or ImageStudioLite (LI-COR Biosciences).

535 **Quantification of Sgt2/A—TA complex formation**

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

541 In artificial TA experiments, both his-tagged Sgt2/A and cMyc-tagged artificial TAs were
542 quantified via immunoblotting signals. The complex efficiency of Sgt2/A with various TAs was
543 obtained by

544 $$E_{complex} = \frac{E_{TA}}{T_{TA}} \times \frac{1}{E_{capture}} \quad (1)$$

545 where E_{TA} is the signal intensity of an eluted TA representing the amount of TA co-purified with
546 Sgt2/A. T_{TA} is the signal intensity of a TA in total lysate that corresponds to the expression yield
547 of that TA. Identical volumes of elution and total lysate from different TAs experiments were
548 analyzed and quantified. In order to correct for possible variation in Ni-NTA capture efficiencies,
549 $E_{capture}$ is applied and were obtained by

550

551 $$E_{capture} = \frac{E_{Sgt2}}{E_{purified,\ Sgt2}}, \quad (2)$$

552

553 where E_{Sgt2} is the signal intensity of eluted Sgt2/A, and $E_{purified,\ Sgt2}$ is purified His-tagged Sgt2-
554 TPR-C as an external control. Each E_{TA} and T_{TA} was obtained by blotting both simultaneously, *i.e.*
555 adjacently on the same blotting paper. To facilitate comparison between TAs, the TA complex
556 efficiency $E_{complex,\ TA}$ is normalized by Bos1 complex efficiency $E_{complex,\ Bos1}$.

557 $$\%\ \text{Complex} = \frac{E_{complex,\ TA}}{E_{complex,\ Bos1}} \times 100 \quad (3)$$

558 **Molecular modeling**

559 Putative models for *Sc*Sgt2-C were generated with I-TASSER, PCONS, Quark, and Rosetta
560 via their respective web servers [48-51]. Residue proximity probed by disulfide bond formation
561 suggests that the models put forth by Quark are most plausible. These structures were the only
562 ones with a potential binding groove. The highest scoring model was then chosen to identify

563 putative TA binding sites. To generate complexes, various transmembrane domains were modelled
564 as alpha helices (using 3D-HM [93]) and rigid-body docked into the Sgt2-C_{cons} through the Zdock
565 web server [94]. Images were rendered using PyMOL 2.2 (www.pymol.org).
566 Using the same set of structure prediction servers, we were unable to produce a clear structural
567 model for SGTA-C. We were also unable to get a convincing model by threading the SGTA-C
568 sequence onto the Sgt2-C model [95].

569 **Structure Relaxation**

570 The highest scoring model of Sgt2-C from Quark was relaxed by all-atom molecular
571 dynamics to better account for molecular details not explicitly accounted for by structure
572 prediction methods, *i.e.* to understand an energetic local minimum near the prediction. The protein
573 and solvent system (TIP3P, ~12k atoms, CHARMM36 [96]) once built was minimized (500 steps)
574 and slowly heated to 298K (0.01K/fs) twice: first with a 10 kJ/mol/Å^2 harmonic restraint on each
575 protein atom and then without restraints. The resulting system was equilibrated for 2 ns at constant
576 volume and then for 100 ns at constant pressure (1 atm). All manipulation and calculations were
577 performed using VMD 1.9.2 [97] and NAMD 2.11 [98]. Further details about the simulation
578 protocols and results can be found within the configuration or output files (details below).

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

580 Proteins identified from UniProt [60] containing a single transmembrane domain within 30
581 residues of the C-terminus were separated into groups based on their localization reported in
582 UniProt. The topology of all proteins with 3 TMs or fewer was further analyzed using TOPCONS
583 [61] to avoid missed single-pass TM proteins. Proteins with a predicted signal peptide [63], an
584 annotated transit peptide, problematic cautions, or with a length less than 50 or greater than 1000
585 residues were excluded. Proteins localized to the ER, golgi apparatus, nucleus, endosome,
586 lysosome, and cell membrane were classified as ER-bound, those localized to the outer
587 mitochondrial membrane were classified as mitochondria-bound, those localized to the
588 peroxisome were classified as peroxisomal proteins, and those with unknown localization were
589 classified as unknown. Proteins with a compositional bias overlapping with the predicted TMD
590 were also excluded. A handful of proteins and their inferred localizations were manually corrected
591 or removed (see notebook and Table S1).
592

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 〈 μ_H 〉 [99] was a poor predictor of localization, *e.g.* although a Leu_{18} helix is extremely hydrophobic, it has 〈 μ_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 helical wheel diagram. While a range of hydrophobicity scales were predictive using this metric, we selected the TM Tendency scale [68] to characterize the TMDs of putative TA IMPs 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 [100]). Due to many fewer mitochondrial proteins (*i.e.* a class imbalance), we also confirmed that the AUROC values were consistent with the more robust, but less common, Average Precision (see notebook).

Sequence analyses

An alignment of Sgt2-C domains was carried out as follows: all sequences with an annotated N-terminal Sgt2/A dimerization domain (PF16546 [101]), 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 [102], and then aligned using MAFFT G-INS-i [103]. 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 [104]. 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 [105]. See code repository for additional details.

Data and Code Availability

All configuration, analysis, and figure generation code employed is available openly at github.com/clemlab/sgt2a-modeling with analysis done in Jupyter Lab/Notebooks using Python 3.6 enabled by Numpy, Pandas, Scikit-Learn, BioPython, and Bokeh [106-111]. The system topology and output files (including trajectory sampled at 0.5 ns intervals) can be permanently found here: 10.22002/D1.1100

Acknowledgements

We thank D. G. VanderVelde for assistance with NMR data collection; S. Mayo for providing computing resources; S.-o. Shan, H. J. Cho, and members of the Clemons lab for support and discussion. We thank J.-Y. Mock and A. M. Thinn for comments on the manuscript. This work was supported by the National Institutes of Health (NIH) Pioneer Award 5DP1GM105385 and 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).

References

1. Pieper U, Schlessinger A, Kloppmann E, Chang GA, Chou JJ, Dumont ME, Fox BG, Fromme P, Hendrickson WA, Malkowski MG, et al. (2013) Coordinating the impact of structural genomics on the human α-helical transmembrane proteome. *Nat Struct Mol Biol* **20**: 135–138.
2. Aviram N, Schuldiner M (2017) Targeting and translocation of proteins to the endoplasmic reticulum at a glance. *J Cell Sci* **130**: 4079–4085.
3. Shao S, Hegde RS (2011) Membrane Protein Insertion at the Endoplasmic Reticulum. *Annu Rev Cell Dev Biol* **27**: 25–56.
4. Guna A, Hegde RS (2018) Transmembrane Domain Recognition during Membrane Protein Biogenesis and Quality Control. *Curr Biol* **28**: R498–R511.
5. Kutay U, Hartmann E, Rapoport TA (1993) A class of membrane proteins with a C-terminal anchor. *Trends Cell Biol* **3**: 72–75.
6. Hegde RS, Keenan RJ (2011) Tail-anchored membrane protein insertion into the endoplasmic reticulum. *Nat Rev Mol Cell Biol* **12**: 787–798.
7. Denic V (2012) A portrait of the GET pathway as a surprisingly complicated young man. *Trends Biochem Sci* **37**: 411–417.
8. Wattenberg B, Lithgow T (2001) Targeting of C-Terminal (Tail)-Anchored Proteins: Understanding how Cytoplasmic Activities are Anchored to Intracellular Membranes. *Traffic* **2**: 66–71.
9. Chartron JW, Clemons WM, Suloway CJM (2012) The complex process of GETting tail-anchored membrane proteins to the ER. *Curr Opin Struct Biol* **22**: 217–224.
10. Chio US, Cho H, Shan S-O (2017) Mechanisms of Tail-Anchored Membrane Protein Targeting and Insertion. *Annu Rev Cell Dev Biol* **33**: 417–438.
11. Shao S, Hegde RS (2011) A Calmodulin-Dependent Translocation Pathway for Small Secretory Proteins. *Cell* **147**: 1576–1588.
12. Chartron JW, Gonzalez GM, Clemons WM Jr. (2011) A Structural Model of the Sgt2 Protein and Its Interactions with Chaperones and the Get4/Get5 Complex. *J Biol Chem* **286**: 34325–34334.
13. Wang F, Brown EC, Mak G, Zhuang J, Denic V (2010) A Chaperone Cascade Sorts Proteins for Posttranslational Membrane Insertion into the Endoplasmic Reticulum. *Mol Cell* **40**: 159–171.
14. Simon AC, Simpson PJ, Goldstone RM, Krysztofinska EM, Murray JW, High S, Isaacson RL (2013) Structure of the Sgt2/Get5 complex provides insights into GET-mediated targeting of tail-anchored membrane proteins. *Proc Natl Acad Sci USA* **110**: 1327–1332.
15. Xu Y, Cai M, Yang Y, Huang L, Ye Y (2012) SGTA Recognizes a Noncanonical Ubiquitin-like Domain in the Bag6-Ubl4A-Trc35 Complex to Promote Endoplasmic Reticulum-Associated Degradation. *Cell Rep* **2**: 1633–1644.
16. Wunderley L, Leznicki P, Payapilly A, High S (2014) SGTA regulates the cytosolic quality control of hydrophobic substrates. *J Cell Sci* **127**: 4728–4739.
17. Leznicki P, High S (2012) SGTA antagonizes BAG6-mediated protein triage. *Proc Natl Acad Sci USA* **109**: 19214–19219.
18. Gristick HB, Rao M, Chartron JW, Rome ME, Shan S-O, Clemons WM (2014) Crystal structure of ATP-bound Get3–Get4–Get5 complex reveals regulation of Get3 by Get4. *Nat Struct Mol Biol* **21**: 437–442.

19. Mock J-Y, Chartron JW, Zaslaver M, Xu Y, Ye Y, Clemons WM Jr. (2015) Bag6 complex contains a minimal tail-anchor–targeting module and a mock BAG domain. *Proc Natl Acad Sci USA* **112**: 106–111.
20. Wang F, Whynot A, Tung M, Denic V (2011) The Mechanism of Tail-Anchored Protein Insertion into the ER Membrane. *Mol Cell* **43**: 738–750.
21. Cho H, Shan S-O (2018) Substrate relay in an Hsp70-cochaperone cascade safeguards tail-anchored membrane protein targeting. *EMBO J* **37**:.
22. Stefer S, Reitz S, Wang F, Wild K, Pang Y-Y, Schwarz D, Bomke J, Hein C, Löhr F, Bernhard F, et al. (2011) Structural basis for tail-anchored membrane protein biogenesis by the Get3-receptor complex. *Science* **333**: 758–762.
23. Rome ME, Chio US, Rao M, Gristick H, Shan S-O (2014) Differential gradients of interaction affinities drive efficient targeting and recycling in the GET pathway. *Proc Natl Acad Sci USA* **111**: E4929–E4935.
24. Vilardi F, Lorenz H, Dobberstein B (2011) WRB is the receptor for TRC40/Asna1-mediated insertion of tail-anchored proteins into the ER membrane. *J Cell Sci* **124**: 1301–1307.
25. Yamamoto Y, Sakisaka T (2012) Molecular Machinery for Insertion of Tail-Anchored Membrane Proteins into the Endoplasmic Reticulum Membrane in Mammalian Cells. *Mol Cell* **48**: 387–397.
26. Schuldiner M, Metz J, Schmid V, Denic V, Rakwalska M, Schmitt HD, Schwappach B, Weissman JS (2008) The GET Complex Mediates Insertion of Tail-Anchored Proteins into the ER Membrane. *Cell* **134**: 634–645.
27. Kiktev DA, Patterson JC, Müller S, Bariar B, Pan T, Chernoff YO (2012) Regulation of chaperone effects on a yeast prion by cochaperone Sgt2. *Mol Cell Biol* **32**: 4960–4970.
28. Xu Y, Liu Y, Lee J-G, Ye Y (2013) A Ubiquitin-like Domain Recruits an Oligomeric Chaperone to a Retrotranslocation Complex in Endoplasmic Reticulum-associated Degradation. *J Biol Chem* **288**: 18068–18076.
29. Rodrigo-Brenni MC, Gutierrez E, Hegde RS (2014) Cytosolic Quality Control of Mislocalized Proteins Requires RNF126 Recruitment to Bag6. *Mol Cell* **55**: 227–237.
30. Hessa T, Sharma A, Mariappan M, Eshleman HD, Gutierrez E, Hegde RS (2011) Protein targeting and degradation are coupled for elimination of mislocalized proteins. *Nature* **475**: 394–397.
31. Dupzyk A, Williams JM, Bagchi P, Inoue T, Tsai B (2017) SGTA-Dependent Regulation of Hsc70 Promotes Cytosol Entry of Simian Virus 40 from the Endoplasmic Reticulum. *J Virol* **91**: 23.
32. Long P, Samnakay P, Jenner P, Rose S (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. *Eur J Neurosci* **36**: 2733–2742.
33. Trotta AP, Need EF, Selth LA, Chopra S, Pinnock CB, Leach DA, Coetzee GA, Butler LM, Tilley WD, Buchanan G (2013) Knockdown of the cochaperone SGTA results in the suppression of androgen and PI3K/Akt signaling and inhibition of prostate cancer cell proliferation. *Int J Cancer* **133**:.
34. Buchanan G, Ricciardelli C, Harris JM, Prescott J, Yu ZCL, Jia L, Butler LM, Marshall VR, Scher HI, Gerald WL, et al. (2007) Control of Androgen Receptor Signaling in

Prostate Cancer by the Cochaperone Small Glutamine Rich Tetratricopeptide Repeat Containing Protein. *Cancer Res* **67**: 10087–10096.

35. Wang H, Zhang Q, Zhu D (2003) hSGT interacts with the N-terminal region of myostatin. *Biochem Biophys Res Commun* **311**: 877–883.
36. Chartron JW, VanderVelde DG, Clemons WM Jr. (2012) Structures of the Sgt2/SGTA Dimerization Domain with the Get5/UBL4A UBL Domain Reveal an Interaction that Forms a Conserved Dynamic Interface. *Cell Rep* **2**: 1620–1632.
37. Liou S-T, Wang C (2005) Small glutamine-rich tetratricopeptide repeat-containing protein is composed of three structural units with distinct functions. *Arch Biochem Biophys* **435**: 253–263.
38. Cziepluch C, Kordes E, Poirey R, Grewenig A, Rommelaere J, Jauniaux JC (1998) Identification of a novel cellular TPR-containing protein, SGT, that interacts with the nonstructural protein NS1 of parvovirus H-1. *J Virol* **72**: 4149–4156.
39. Callahan MA, Handley MA, Lee YH, Talbot KJ, Harper JW, Panganiban AT (1998) Functional interaction of human immunodeficiency virus type 1 Vpu and Gag with a novel member of the tetratricopeptide repeat protein family. *J Virol* **72**: 5189–5197.
40. Winnefeld M, Grewenig A, Schn lzer M, Spring H, Knoch TA, Gan EC, Rommelaere J, Cziepluch C (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. *Exp Cell Res* **312**: 2500–2514.
41. Darby JF, Krysztofinska EM, Simpson PJ, Simon AC, Leznicki P, Sriskandarajah N, Bishop DS, Hale LR, Alfano C, Conte MR, et al. (2014) Solution Structure of the SGTA Dimerisation Domain and Investigation of Its Interactions with the Ubiquitin-Like Domains of BAG6 and UBL4A. *PLoS ONE* **9**: e113281–19.
42. Dutta S, Tan Y-J (2008) Structural and functional characterization of human SGT and its interaction with Vpu of the human immunodeficiency virus type 1. *Biochemistry* **47**: 10123–10131.
43. Leznicki P, Korac-Prlic J, Kliza K, Husnjak K, Nyathi Y, Dikic I, High S (2015) Binding of SGTA to Rpn13 selectively modulates protein quality control. *J Cell Sci* **128**: 3187–3196.
44. Waheed AA, MacDonald S, Khan M, Mounts M, Swiderski M, Xu Y, Ye Y, Freed EO (2016) The Vpu-interacting Protein SGTA Regulates Expression of a Non- glycosylated Tetherin Species. *Sci Rep* **6**: 1–14.
45. Luo P, Baldwin RL (1997) Mechanism of helix induction by trifluoroethanol: a framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. *Biochemistry* **36**: 8413–8421.
46. Dyson HJ, Wright PE (2004) Unfolded Proteins and Protein Folding Studied by NMR. *Chem Rev* **104**: 3607–3622.
47. Martínez-Lumbreras S, Krysztofinska EM, Thapaliya A, Spilotros A, Matak-Vinkovic D, Salvadori E, Roboti P, Nyathi Y, Muench JH, Roessler MM, et al. (2018) Structural complexity of the co-chaperone SGTA: a conserved C-terminal region is implicated in dimerization and substrate quality control. *BMC Biol* **16**: 76.
48. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**: 1715–1735.
49. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**: 7–8.

50. Wallner B (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Prot Sci* **15**: 900–913.
51. Bradley P, Misura KM, Baker D (2005) Towards high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.
52. Chun E, Thompson AA, Liu W, Roth CB, Griffith MT, Katritch V, Kunken J, Xu F, Cherezov V, Hanson MA, et al. (2012) Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors. *Cell Struct Funct* **20**: 967–976.
53. Mallick P, Boutz DR, Eisenberg D, Yeates TO (2002) Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc Natl Acad Sci USA* **99**: 9679–9684.
54. Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* **43**: D257–D260.
55. Schmid AB, Lagleder S, Gräwert MA, Röhl A, Hagn F, Wandinger SK, Cox MB, Demmer O, Richter K, Groll M, et al. (2012) The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J* **31**: 1506–1517.
56. Li Z, Hartl FU, Bracher A (2013) Structure and function of Hip, an attenuator of the Hsp70 chaperone cycle. *Nat Struct Mol Biol* **20**: 929–935.
57. Nelson GM, Huffman H, Smith DF (2003) Comparison of the carboxy-terminal DP-repeat region in teh co-chaperones Hop and Hip. *EMBO J* **8**: 125–133.
58. Rao M, Okreglak V, Chio US, Cho H, Walter P, Shan S-O (2016) Multiple selection filters ensure accurate tail-anchored membrane protein targeting. *eLife* **5**: e21301–e21324.
59. Beilharz T, Egan B, Silver PA, Hofmann K, Lithgow T (2003) Bipartite Signals Mediate Subcellular Targeting of Tail-anchored Membrane Proteins in Saccharomyces cerevisiae. *J Biol Chem* **278**: 8219–8223.
60. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, et al. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**: D158–D169.
61. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* **43**: W401–W407.
62. Borgese N, Colombo S, Pedrazzini E (2003) The tale of tail-anchored proteins. *J Cell Biol* **161**: 1013–1019.
63. Nielsen H (2017) *Predicting Secretory Proteins with SignalP*. Methods in molecular biology (Clifton, N.J.).
64. Burri L, Lithgow T (2003) A Complete Set of SNAREs in Yeast. *Traffic* **5**: 45–52.
65. Kalbfleisch T, Cambon A, Wattenberg BW (2007) A Bioinformatics Approach to Identifying Tail-Anchored Proteins in the Human Genome. *Traffic* **8**: 1687–1694.
66. Guna A, Volkmar N, Christianson JC, Hegde RS (2018) The ER membrane protein complex is a transmembrane domain insertase. *Science* **359**: 470–473.
67. Guna A, Hegde RS (2018) Transmembrane Domain Recognition during Membrane Protein Biogenesis and Quality Control. *Curr Biol* **28**: R498–R511.
68. Zhao G, London E (2006) An amino acid 'transmembrane tendency' scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Prot Sci* **15**: 1987–2001.

69. Schinzel A, Kaufmann T, Borner C (2004) Bcl-2 family members: intracellular targeting, membrane-insertion, and changes in subcellular localization. *Biochim Biophys Acta* **1644**: 95–105.
70. Popgeorgiev N, Jabbour L, Gillet G (2018) Subcellular Localization and Dynamics of the Bcl-2 Family of Proteins. *Front Cell Dev Biol* **6**: 2468–11.
71. Pawlikowska P, Leray I, de Laval B, Guihard S, Kumar R, Rosselli F, Porteu F (2010) ATM-dependent expression of IEX-1 controls nuclear accumulation of Mcl-1 and the DNA damage response. *Cell Death Differ* **17**: 1739–1750.
72. Kaufmann T, Schlipf S, Sanz J, Neubert K, Stein R, Borner C (2003) Characterization of the signal that directs Bcl-x L, but not Bcl-2, to the mitochondrial outer membrane. *J Cell Biol* **160**: 53–64.
73. Echeverry N, Bachmann D, Ke F, Strasser A, Simon HU, Kaufmann T (2013) Intracellular localization of the BCL-2 family member BOK and functional implications. *Cell Death Differ* **20**: 785–799.
74. Pandit J, Danley DE, Schulte GK, Mazzalupo S, Pauly TA, Hayward CM, Hamanaka ES, Thompson JF, Harwood HJ Jr. (2000) Crystal Structure of Human Squalene Synthase. *J Biol Chem* **275**: 30610–30617.
75. Röhl A, Wengler D, Madl T, Lagleder S, Tippel F, Herrmann M, Hendrix J, Richter K, Hack G, Schmid AB, et al. (2015) Hsp90 regulates the dynamics of its cochaperone Sti1 and the transfer of Hsp70 between modules. *Nat Commun* **6**: 6655.
76. Coto ALS, Seraphim TV, Batista FAH, Dores-Silva PR, Barranco ABF, Teixeira FR, Gava LM, Borges JC (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. *Int J Biol Macromol* **118**: 693–706.
77. Chou M-L, Fitzpatrick LM, Tu S-L, Budziszewski G, Potter-Lewis S, Akita M, Levin JZ, Keegstra K, Li H-M (2003) Tic40, a membrane-anchored co-chaperone homolog in the chloroplast protein translocon. *EMBO J* **22**: 2970–2980.
78. Balsera M, Soll J, Buchanan BB (2009) *Chapter 10 - Protein Import in Chloroplasts: An Emerging Regulatory Role for Redox*. Elsevier Ltd.
79. Bédard J, Trösch R, Wu F, Ling Q, Flores-Pérez Ú, Töpel M, Nawaz F, Jarvis P (2017) Suppressors of the Chloroplast Protein Import Mutant tic40Reveal a Genetic Link between Protein Import and Thylakoid Biogenesis. *Plant Cell* **29**: 1726–1747.
80. Trempe J-F, Šašková KG, Sivá M, Ratcliffe CDH, Veverka V, Hoegl A, Ménade M, Feng X, Shenker S, Svoboda M, et al. (2016) Structural studies of the yeast DNA damage-inducible protein Ddi1 reveal domain architecture of this eukaryotic protein family. *Sci Rep* **6**: 33671.
81. Sivá M, Svoboda M, Veverka V, Trempe J-F, Hofmann K, Kožíšek M, Hexnerová R, Sedlák F, Belza J, Brynda J, et al. (2016) Human DNA-Damage-Inducible 2 Protein Is Structurally and Functionally Distinct from Its Yeast Ortholog. *Sci Rep* **6**: 30443.
82. Voorhees RM, Hegde RS (2015) Structures of the scanning and engaged states of the mammalian SRP-ribosome complex. *eLife* **4**: 1485–21.
83. Keenan RJ, Freymann DM, Walter P, Stroud RM (1998) Crystal Structure of the Signal Sequence Binding Subunit of the Signal Recognition Particle. *Cell* **94**: 181–191.

84. Clemons WM, Gowda K, Black SD, Zwieb C, Ramakrishnan V (1999) Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 A resolution: evidence for the mechanism of signal peptide binding. *J Mol Biol* **292**: 697–705.
85. Shao S, Rodrigo-Brenni MC, Kivlen MH, Hegde RS (2017) Mechanistic basis for a molecular triage reaction. *Science* **355**: 298–302.
86. Martoglio B (1997) Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *EMBO J* **16**: 6636–6645.
87. Aviram N, Costa EA, Arakel EC, Chuartzman SG, Jan CH, Haßdenteufel S, Dudek J, Jung M, Schorr S, Zimmermann R, et al. (2016) The SND proteins constitute an alternative targeting route to the endoplasmic reticulum. *Nature* **540**: 134–138.
88. Suloway CJ, Rome ME, Clemons WM (2011) Tail-anchor targeting by a Get3 tetramer: the structure of an archaeal homologue. *EMBO J* **31**: 707–719.
89. Chu R, Takei J, Knowlton JR, Andrykovitch M, Pei W, Kajava AV, Steinbach PJ, Ji X, Bai Y (2002) Redesign of a Four-helix Bundle Protein by Phage Display Coupled with Proteolysis and Structural Characterization by NMR and X-ray Crystallography. *J Mol Biol* **323**: 253–262.
90. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. *Protein Express Purif* **41**: 207–234.
91. Micsonai A, Wien F, Kernya L, Lee Y-H, Goto Y, Réfrégiers M, Kardos J (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci USA* **112**: E3095–E3103.
92. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**: 671–675.
93. Reißer S, Strandberg E, Steinbrecher T, Ulrich AS (2014) 3D Hydrophobic Moment Vectors as a Tool to Characterize the Surface Polarity of Amphiphilic Peptides. *Biophys J* **106**: 2385–2394.
94. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**: 1771–1773.
95. Webb B, Sali A (2002) *Comparative Protein Structure Modeling Using MODELLER*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
96. Huang J, MacKerell AD Jr (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* **34**: 2135–2145.
97. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* **14**: 33–38.
98. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* **26**: 1781–1802.
99. Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **299**: 371–374.
100. Swets JA, Dawes RM, Monahan J (2000) Better decisions through science. *Sci Am* **283**: 82–87.
101. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**: D279–D285.

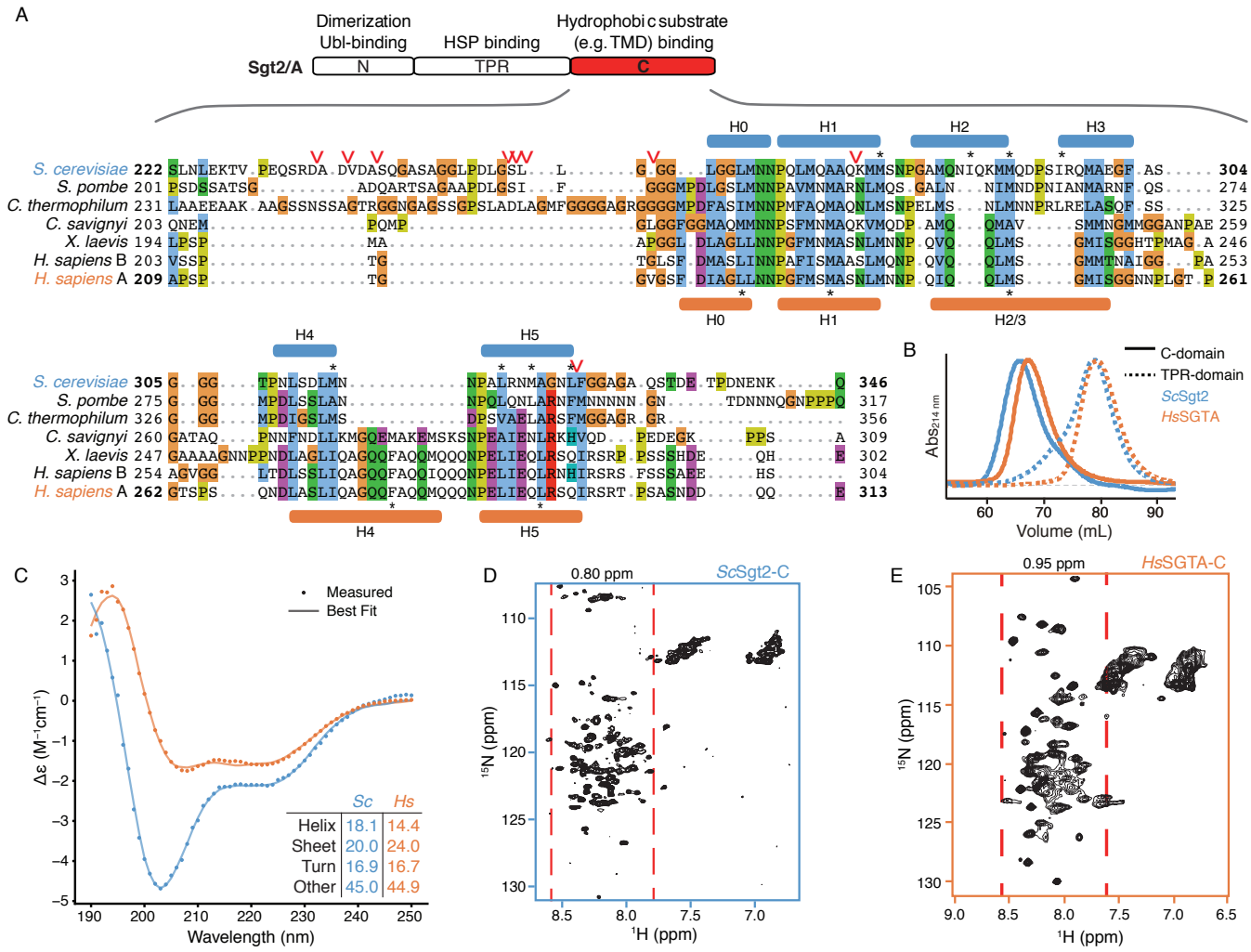
102. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.
103. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772–780.
104. Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**: 2295–2300.
105. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.
106. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, et al. (2016) Jupyter Notebooks -- a publishing format for reproducible computational workflows. In Loizides F, Schmidt B (eds.) pp 87–90.
107. Millman KJ, Aivazis M (2011) Python for Scientists and Engineers. *Comput Sci Eng* **13**: 9–12.
108. McKinney W (2010) Data Structures for Statistical Computing in Python. In pp 51–56.
109. Pedregosa F, Varoquaux GEL, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.
110. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.
111. Bokeh Development Team (2014) *Bokeh: Python library for interactive visualization*.
112. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* **43**: W389–W394.
113. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105–132.
114. Schägger H (2006) Tricine–SDS-PAGE. *Nat Protoc* **1**: 16–22.
115. Fernandez-Patron C, Hardy E, Sosa A, Seoane J, Castellanos L (1995) Double staining of coomassie blue-stained polyacrylamide gels by imidazole-sodium dodecyl sulfate-zinc reverse staining: sensitive detection of coomassie blue-undetected proteins. *Anal Biochem* **224**: 263–269.
116. Gillespie AS, Elliott E (2005) Comparative advantages of imidazole–sodium dodecyl sulfate–zinc reverse staining in polyacrylamide gels. *Anal Biochem* **345**: 158–160.
117. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195–202.
118. Joosten RP, Beek te TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G (2010) A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39**: D411–D419.
119. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.

953 **Figure Legends**

954 **Fig. 1. Structural characteristics of free Sgt2/A C-domain.** (*A*) *Top*, Schematic of the domain
955 organization of Sgt2/A. *Below,* representative sequences from a large-scale multiple sequence
956 alignment of the C domain: fungal Sgt2 from *S. cerevisiae*, *S. pombe*, and *C. thermophilum* and
957 metazoan SGTA from *C. savignyi*, *X. laevis*, and *H. sapiens*. Protease susceptible sites on *Sc*Sgt2-
958 C identified by mass spectrometry are indicated by red arrowheads. Predicted helices of *Sc*Sgt2
959 (blue) and *Hs*SGTA (orange) by Jpred [112] and/or structure prediction are shown. Blue/orange
960 color scheme for *Sc*Sgt2/*Hs*SGTA is used throughout the text. Residues noted in the text are
961 highlighted by an asterisk. (*B*) Overlay of size-exclusion chromatography traces of *Sc*Sgt2-C (blue
962 line), *Hs*SGTA-C (orange line), *Sc*Sgt2-TPR (blue dash) and *Hs*SGTA-TPR (orange dash). Traces
963 are measured at 214 nm, baseline-corrected and normalized to the same peak height. (*C*) Far UV
964 CD spectrum of 10 μM of purified *Sc*Sgt2-C (blue) and *Hs*SGTA-C (orange) at RT with secondary
965 structure decomposition from BestSel [91]. (*D*) ^{1}H-^{15}N HSQC spectrum of *Sc*Sgt2-C at 25°C. The
966 displayed chemical shift window encompasses all N-H resonances from both backbone and side
967 chains. The range of backbone amide protons, excluding possible side-chain NH_2 of Asn/Gln, is
968 indicated by pairs of red dashed lines. (*E*) As in *D* for *Hs*SGTA-C at 25°C.

969

Figure 1



971 **Fig. 2. The minimal binding region of Sgt2/A for TA client binding.** (*A*) Diagram of the protein
972 truncations tested for TA binding that include the TPR-C domain, C-domain (C), C_{cons}, and C_{cons}
973 ΔH0 (ΔH0) from *Sc*Sgt2 and *Hs*SGTA. The residues corresponding to each domain are indicated,
974 and grey blocks highlight the C_{cons} region. *(B)* Schematic of capture experiments of MBP-Sbh1 by
975 Sgt2/A TPR-C variants. After co-expression, cell pellets are lysed and NTA-Ni^{2+} is used to capture
976 His-Sgt2/A TPR-C. (*C*) Helical wheel diagrams of predicted helices (see Fig. 1*A*) in the C_{cons}
977 domain of *Sc*Sgt2 and *Hs*SGTA. Residues are colored by the Kyte and Doolittle hydrophobicity
978 scale [113]. (*D*) Tris-Tricine-SDS-PAGE gel [114] of co-expressed and purified MBP-tagged Sbh1
979 and His-tagged Sgt2/A truncations visualized with Coomassie Blue staining. (*E*) Alanine
980 replacement of hydrophobic residues in the C_{cons}. All of the hydrophobic residues (L, I, F, and M)
981 in a predicted helix (H0, H1, etc.) are replaced by Ala and tested for the ability to capture MBP-
982 Sbh1. Protein levels were quantified by Coomassie staining. Relative binding efficiency of MBP-
983 Sbh1 by Sgt2 C-domain variants was calculated relative to total amount of Sgt2 C-domain captured
984 (MBP-Sbh1/Sgt2 C-domain) then normalized to the wild-type Sgt2-C domain. Experiments were
985 performed 3-4 times and the standard deviations are presented. Total expression levels of the MBP-
986 Sbh1 were similar across experiments as visualized by immunoblotting (IB) of the cell lysate. (*F)*
987 As in *E* but for *Hs*SGTA.

988

Figure 2

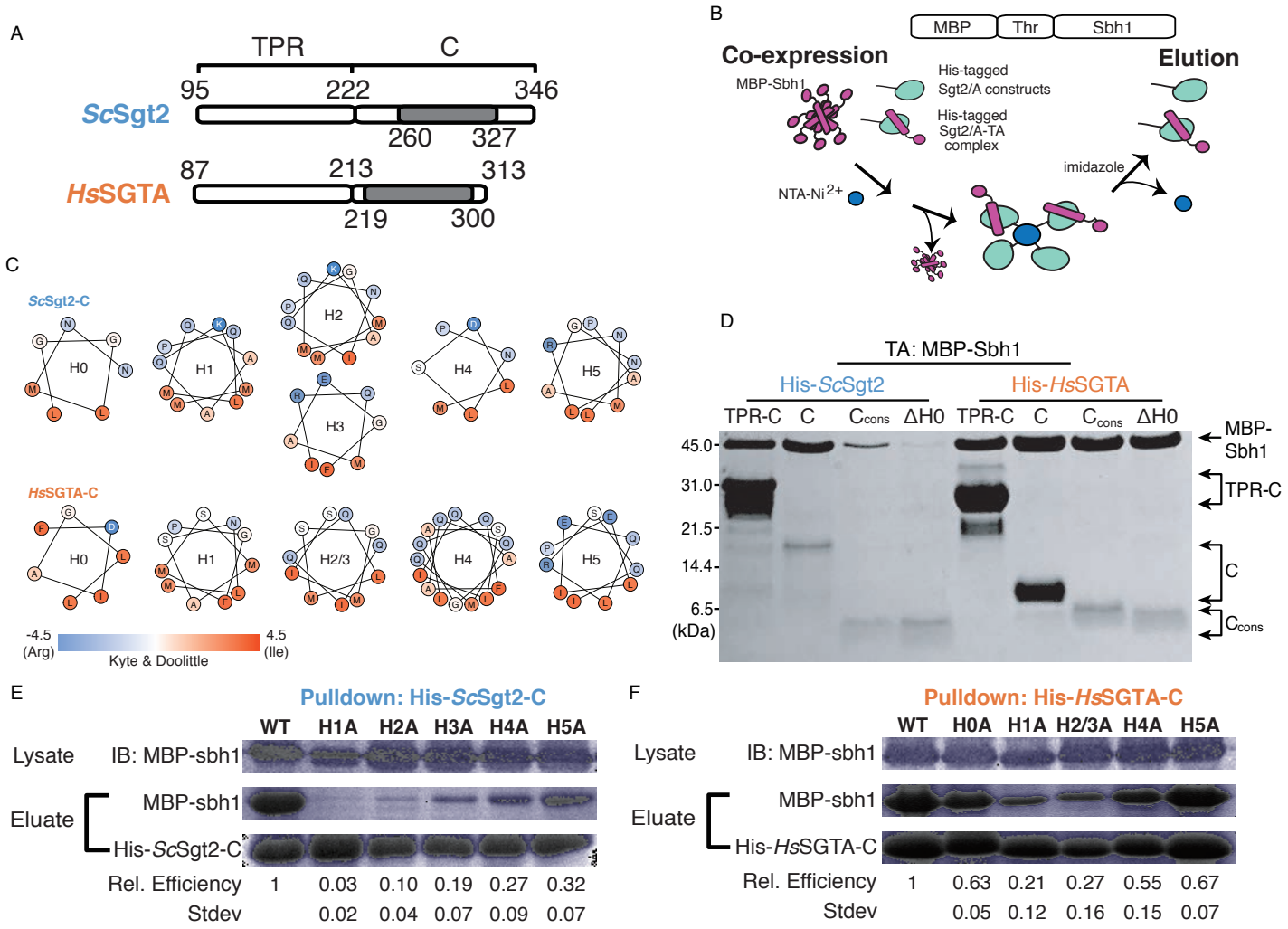


Fig 3. A structural model for Sgt2/A-C_{cons} validated by intramolecular-disulfide bond formation. *(A)* The top 10 models of the Sgt2-C_{cons} generated by the template-free algorithm Quark [48] are overlaid with the highest scoring model in solid. Models are color-ramped from N- (blue) to C-terminus (red). *(B)* A model of Sgt2-C_{cons} (surface colored by Kyte-Doolittle hydrophobicity) bound to a TMD (purple helix) generated by rigid-body docking through Zdock [94]. The darker purple corresponds to an 11 residue stretch. *(C)* The entire Sgt2-C domain 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. *(D* and *E)* Variants of His-*Sc*Sgt2-TPR-C (WT or cysteine double mutants) were co-expressed with the artificial TA client, cMyc-BRIL-11[L8]. After lysis, His-*Sc*Sgt2 proteins were purified, oxidized, then digested by Glu-C protease and analyzed by gel either in non-reducing or reducing buffer. (*E*) Cα ribbon of *Sc*Sgt2-C_{cons} color-ramped with various pairs of Cysteines highlighted. Scissors indicate protease cleavage sites resulting in fragments less than 3 kDa in size. (*F*) Tris-Glycine-SDS-PAGE gel visualized by imidazole-SDS-zinc stain [115,116]. 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 cystine pair (N285/G329) predicted to be far apart by the model does not result in the higher MW species.

Figure 3

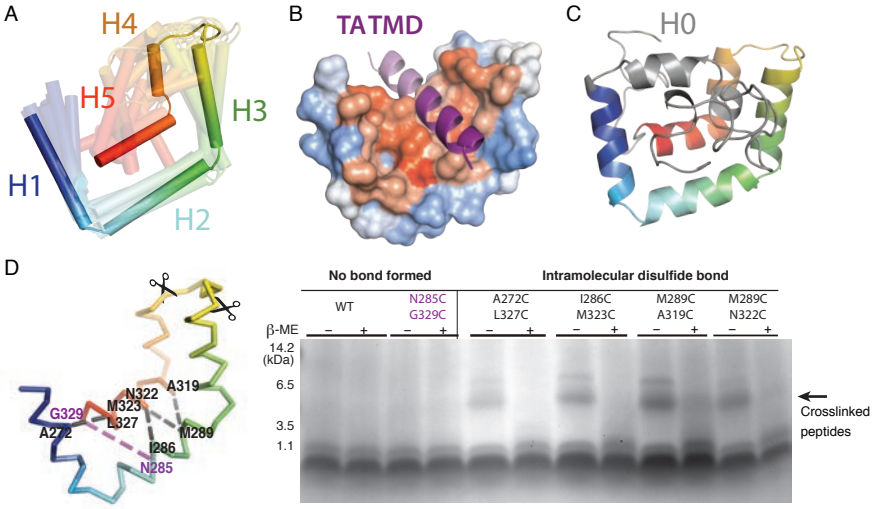


Fig. 4. Comparison of STI1 domains and the Sgt2-C_{cons} model. (*A*) Multiple sequence alignment of Sgt2-C with STI1 domains (DP1, DP2) from STI1/Hop homologs. Helices are shown based on the Sgt2-C_{cons} model and the *Sc*Sti1-DP1/2 structures. Species for representative sequences are from *S. cerevisiae (Scer)*, *S. pombe* (*Spom*), *C. thermophilum* (*Cthe*), *C. savignyi* (*Csav*), and *H. sapiens* (*Hsap*). (*B*) Cα ribbon of *Sc*Sgt2-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-$DP2_{526-582}$ (PDBID: 2LLW) [55]. (*D*) Superposition of the Sgt2-C_{cons} (blue) and Sti1-$DP2_{526-582}$ (red) drawn as cartoons.

Figure 4

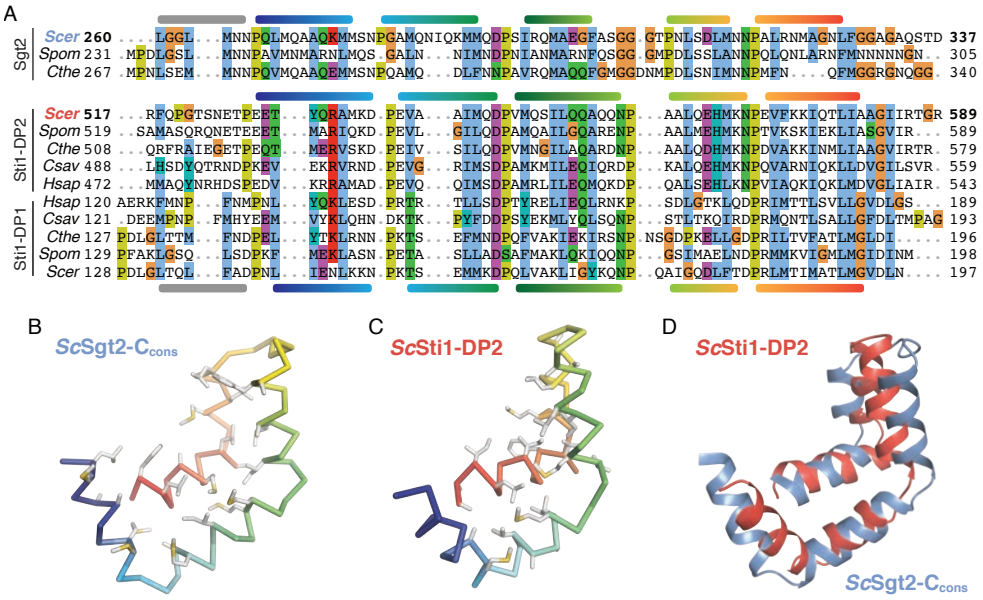


Fig. 5. Effects on TA client binding of charge mutations to the putative hydrophobic groove of Sgt2/A-C_{cons}. For these experiments, individual point mutations are introduced into Sgt2/A-C and tested for their ability to capture Sbh1 quantified as in Figure 2*D*. (*A*) For *Sc*Sgt2-C, a schematic and 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-tagged Sbh1 to His-tagged *Sc*Sgt2-C and mutants were examined as in Figure 2*D*. Lanes for mutated residues are labeled in the same color as the schematic *(B)* Same analysis as in *A* for *Hs*SGTA-C. In addition, double point mutants are included.

Figure 5

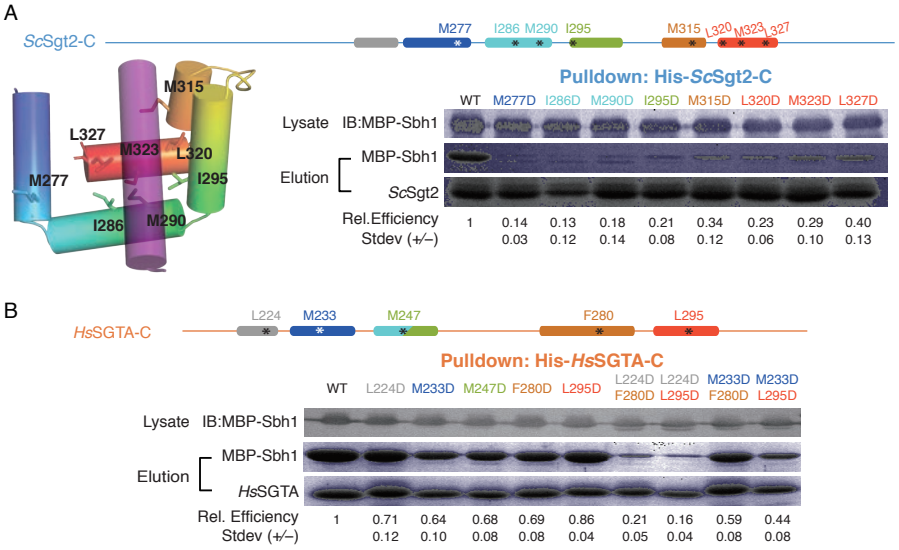


Fig. 6. Minimal requirements for client recognition by Sgt2/A. (*A*) Schematic of model TA clients. Quantification of complex formation is calculated and normalized to that of complexes containing a WT natural TA, here defined as relative binding efficiency. For this figure the WT protein is Bos1. (*B*) Complex formation of *Sc*Sgt2 (blue) and *Hs*SGTA (orange) with the TA Bos1 and several artificial TAs 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 TAs here and for subsequent panels with leucines colored in dark orange, alanines colored in pale orange, and tryptophans colored in grey. (*C*) Complex formation of *Sc*Sgt2 TPR-C and *Hs*SGTA TPR-C with artificial TA IMPs with TMDs of length 11 and increasing numbers of leucine. (*D*) Comparison of complex formation of *Sc*Sgt2 TPR-C and *Hs*SGTA TPR-C with artificial TA IMPs of the same lengths and hydrophobicities but differences in the distribution of leucines, i.e. clustered (solid line) vs distributed (dotted line).

Figure 6

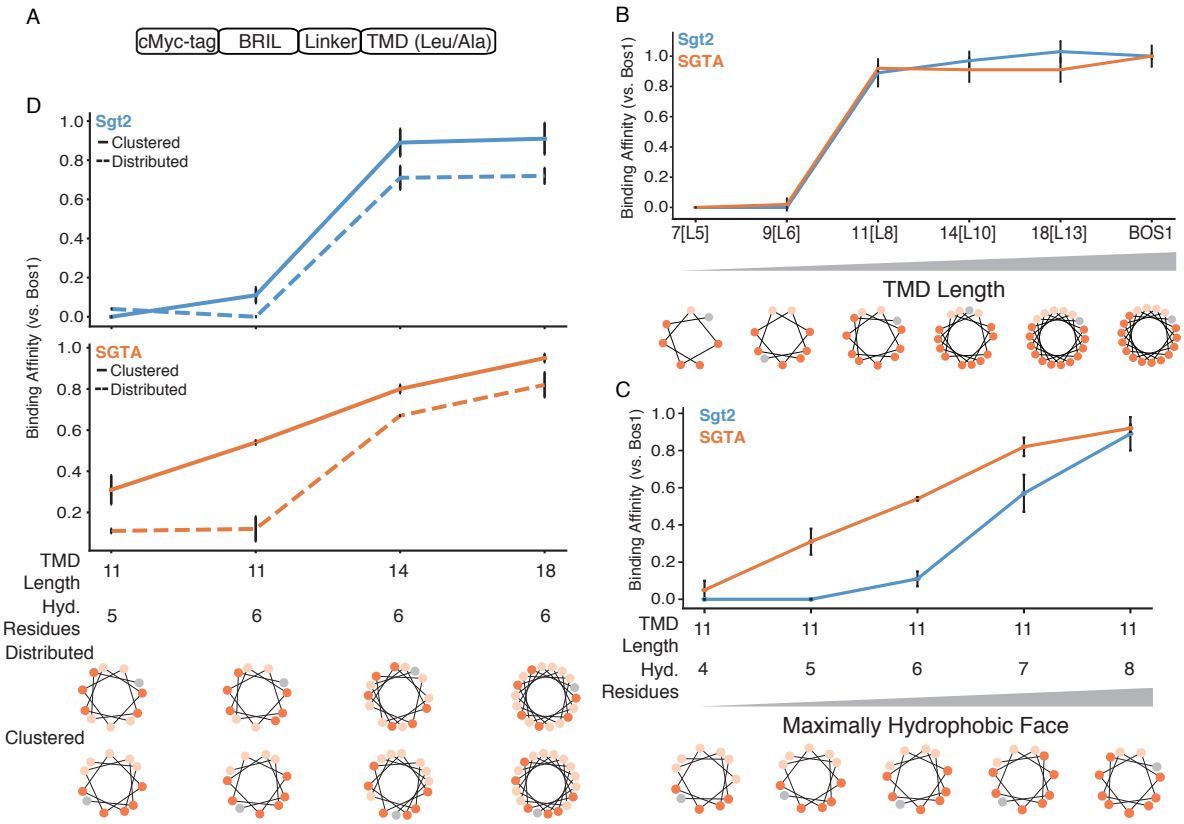


Fig. 7. Hydrophobic properties and localization of natural TA IMPs. (*A*) A summary of putative yeast and human TA IMPs by their experimentally validated localization [60]. Using the definition of a single TMD within 30 residues of their C-terminus gives the final numbers of 90 for yeast and 587 for humans. (*B*) A plot of all predicted yeast TA IMPs comparing two separate metrics for measuring the hydrophobicity of their TMD, either the entire TMD or the most hydrophobic helical face. Each protein is represented by an open circle colored based on localization including those with both mitochondrial and ER localization. Additionally, proteins with ER or mitochondrial localizations are highlighted on each axis. Proteins noted in the text are highlighted. The best cut-offs for predicting mitochondrial versus ER for either metric are represented by dotted lines (dark red, TMD hydrophobicity of entire TMD; light blue: TMD hydrophobicity of the most hydrophobic face). (*C*) As in *B* for putative human TA IMPs. (*D*) Quantitative comparison of the effectiveness of each metric by either the number of correctly predicted ER and mitochondria TA IMPs and the area under a ROC curve (AUROC).

Figure 7

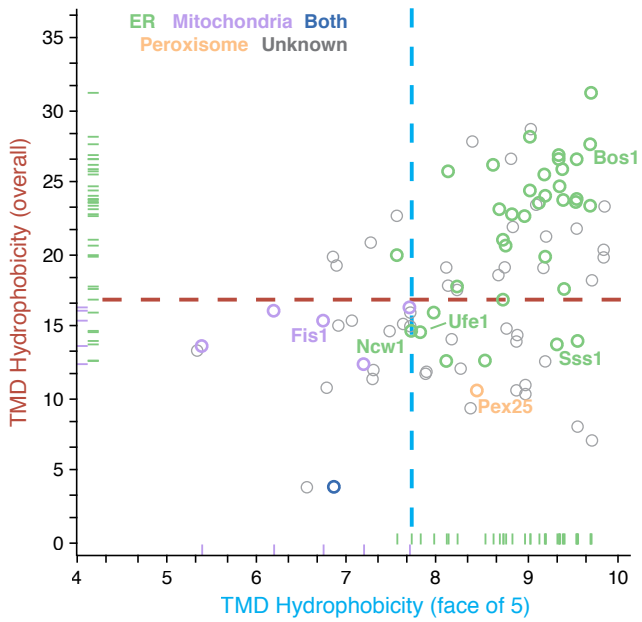
A

	S. cerevisiae	*H. sapiens*
ER	35	117
Mitochondria	5	27
Both	1	10
Peroxisome	1	4
Unknown	46	418
Total	90	576

C

	S. cerevisiae		*H. sapiens*	
	Overall	Face	Overall	Face
Correctly Predicted ER + Mito **(%)**	32 (80)	38 (95)	96 (67)	104 (72)
AUROC	0.87	0.99	0.79	0.82

B



D

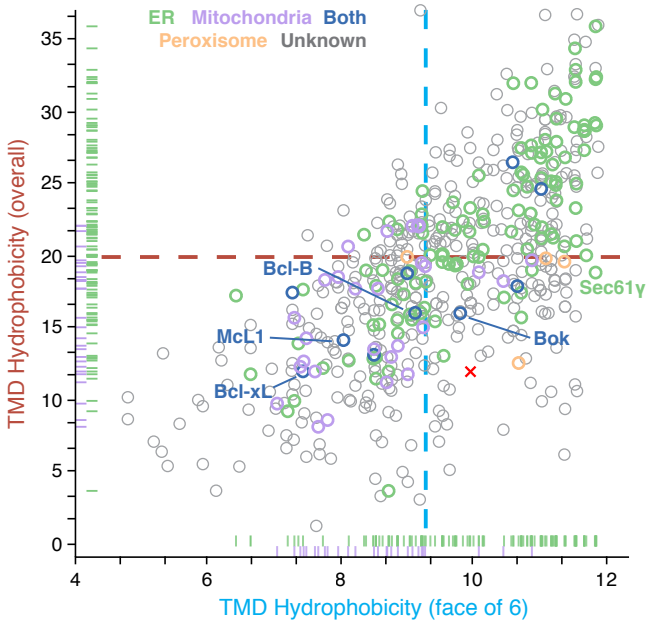


Fig. 8. Various domain structures of STI1 and other helical-hand containing proteins. (*A*) The domain architectures of proteins with a STI1 domain were obtained initially from InterPro [60] and then adjusted as discussed in the text. Each domain within a protein is colored relative to the key. (B) Structural 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 *Sc*Sgt2-C_{cons} and *Sc*Sbh2-TMD, DP1 domain from yeast Sti1 with N-terminus containing H0 in grey (*Sc*Sti1-DP1)(PDBID: 2LLV), STI1 domain from yeast Rad23 (*Sc*Rad23-STI1) bound to the TMD of RAD4 (*Sc*RAD4-TMD) (PDBID: 2QSF), 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).

Figure 8

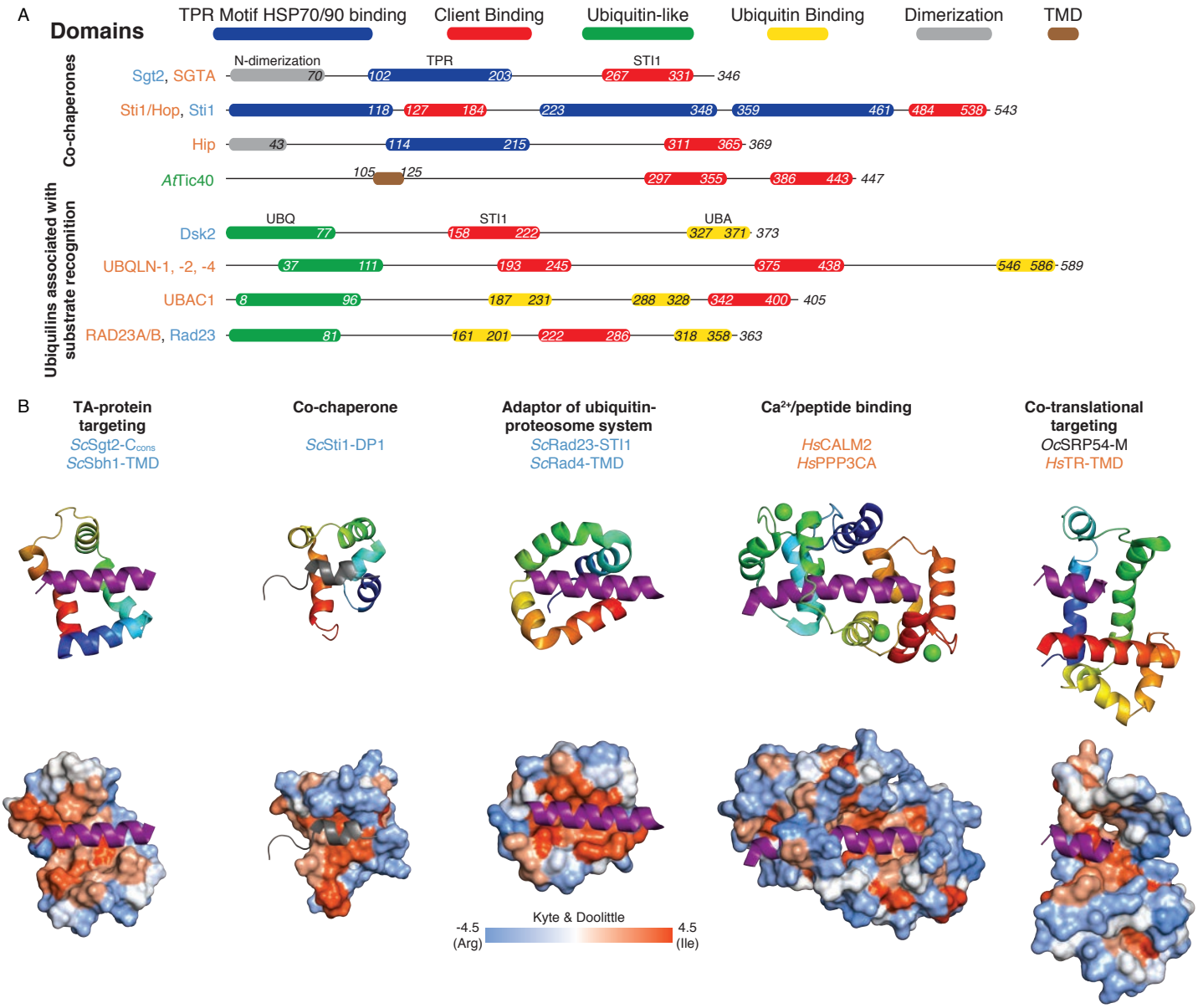


Fig. S1. Biophysical characterization of the Sgt2/A-C_{cons} domain. (*A*) CD spectra as in Fig. 1*C* for the conserved C-terminal domains of Sgt2 (blue) and SGTA (orange). (*B*) NMR spectra as in Fig. 1*D* & *E* for Sgt2-C_{cons} (blue) and SGTA-C_{cons} (orange).

Figure S1

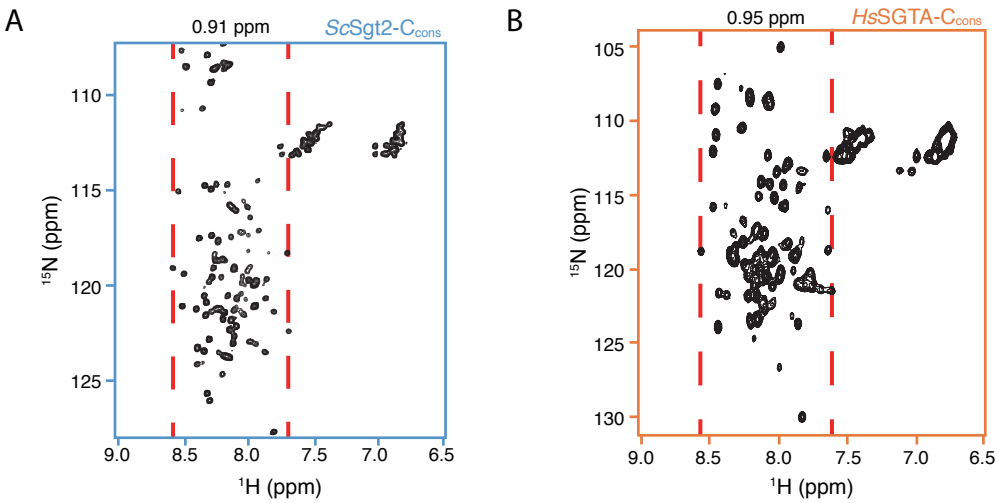


Fig. S2. Characterization of STI1 domains. Predicted [117] and calculated [118] secondary structure elements (*A*) and a structure-based alignment (*B*) of STI1 domains from Fig. 8*A* in the ClustalX color scheme [119]. Dashed lines in *A* depict previous domain boundary annotations. (*C*) Helical wheel diagrams of H0 of STI1 domains. (*D*) Additional STI1 domain structures represented as in Fig. 8*B*. Domains are from the DP2 domain from yeast Sti1 (*Sc*Sti1-DP2) (PDBID:2LLW), the chloroplast import protein Tic40 from *Arabidopsis thaliana* (*At*Tic40-STI1) (PDBID:2LNM), and yeast Rad23 (*Sc*Rad23-STI1) bound to the N-terminus of PNGase (*Sc*PNGase-N-term) (PDBID: 1X3W).

Figure S2

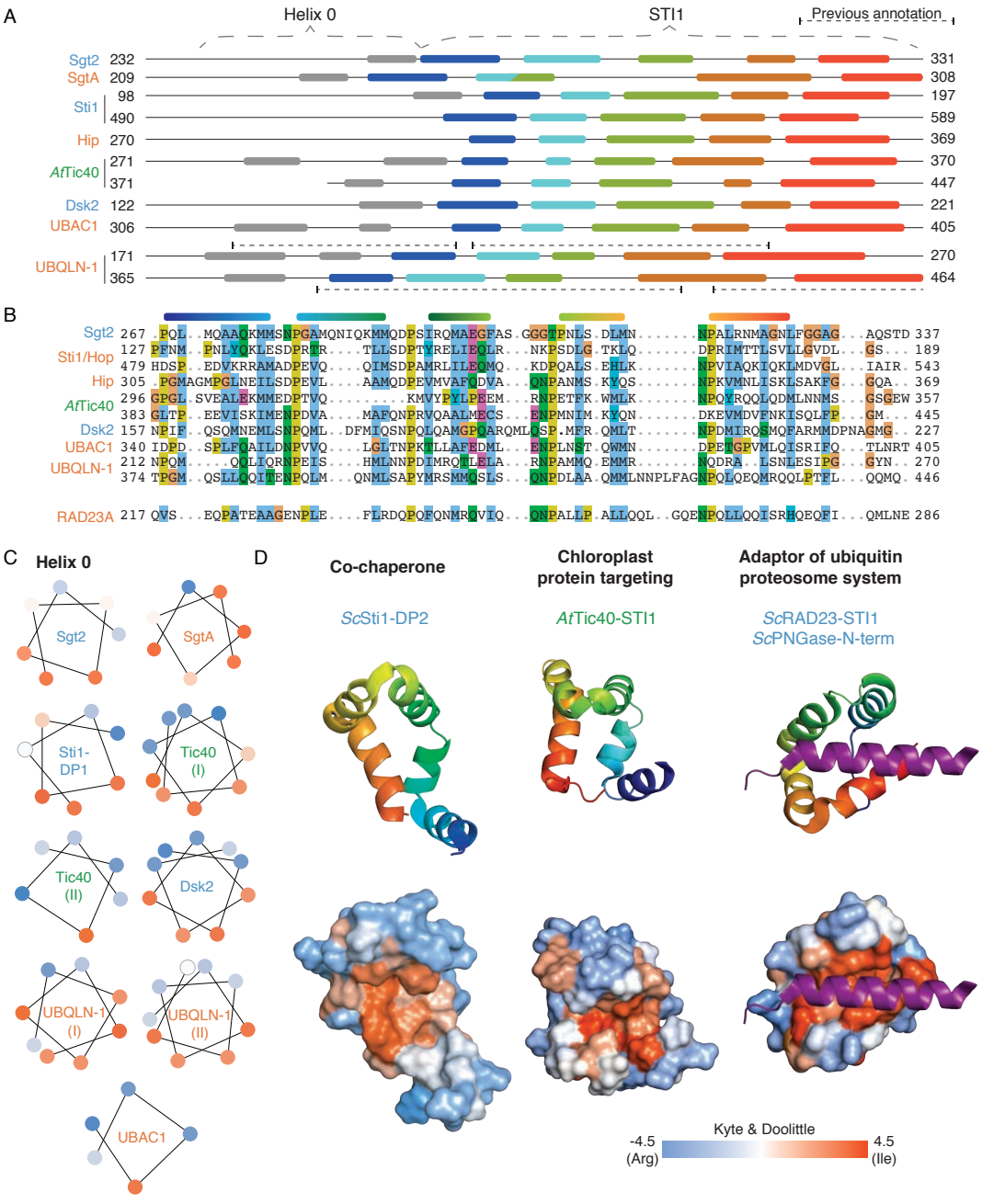


Table S1. TA Database. A compilation of the putative yeast (Sheet 1) and human (Sheet 2) TA proteins shown in Fig. 7*B*,*C*. The Uniprot identifiers, predicted TMD sequence and prediction method, subcellular localization string and resulting inferred target localization, and hydrophobicity metrics (face and overall) are listed for each protein. Those labeled on the plot or mentioned in the text are highlighted along with the abbreviations used. (Sheet 3) A comparison with yeast TA proteins previously compiled by [64] with an explanation of differences, where they exist.