Aug 31st, 2021

Intra-helical salt bridge contribution to membrane protein insertion

Gerard Duart [1#], John Lamb [2#], Juan Ortiz-Mateu [1], Arne Elofsson [2*] and Ismael Mingarro [1*]

[1] Departament de Bioquímica i Biologia Molecular, Institut Universitari de Biotecnologia i Biomedicina (BioTecMed), Universitat de València. E-46100 Burjassot, Spain.

[2] Science for Life Laboratory and Department of Biochemistry and biophysics, Stockholm University, 171 21 Solna, Sweden

[#] Equal contribution

[*] Corresponding authors

ABSTRACT

Salt bridges between negatively (D, E) and positively charged (K, R, H) amino acids play an important role in protein stabilization. This has a more prevalent effect in membrane proteins where polar amino acids are exposed to a very hydrophobic environment. In transmembrane (TM) helices the presence of charged residues can hinder the insertion of the helices into the membrane. This can sometimes be avoided by TM region rearrangements after insertion, but it is also possible that the formation of salt bridges could decrease the cost of membrane integration. However, the presence of intra-helical salt bridges in TM domains and their effect on insertion has not been properly studied yet. In this work, we use an analytical pipeline to study the prevalence of charged pairs of amino acid residues in TM α-helices, which shows that potentially salt-bridge forming pairs are statistically over-represented. We then selected some candidates to experimentally determine the contribution of these electrostatic interactions to the translocon-assisted membrane insertion process. Using both *in vitro* and *in vivo* systems, we confirm the presence of intra-helical salt bridges in TM segments during biogenesis and determined that they contribute between 0.5-0.7 kcal/mol to the apparent free energy of membrane insertion (ΔG_{app}). Our observations suggest that salt bridge interactions can be stabilized during translocon-mediated insertion and thus could be relevant to consider for the future development of membrane protein prediction software.

KEYWORDS

electrostatic interactions; membrane insertion; salt bridge; translocon; transmembrane helix;

INTRODUCTION

Most integral membrane proteins have to insert their transmembrane (TM) segments into the lipid bilayer in a helical conformation and then acquire a defined three-dimensional structure by packaging their helices (Martínez-Gil et al., 2011). α-Helical TM segments are largely composed of apolar residues because of the hydrophobic nature of the membrane environment. Nevertheless, in some cases, it is necessary for the protein activity to include polar amino acids in a TM region in order to develop a functional or structural role (Baeza-Delgado et al., 2012). This fact is sometimes not contemplated in modern membrane topology prediction tools (Tsirigos et al., 2018; 2015), in which the presence of charged amino acids in a sequence automatically suppose a penalty increase in the predicted free energy (ΔG_{pred}) of insertion. The presence of polar amino acids in TM regions is more frequent than what would be expected (Bañó-Polo et al., 2012), especially when these are in pairs on the same face of an α-helix.

Salt bridges are electrostatic interactions between negatively (D, E) and positively charged (K, R, H) amino acids that play an important role in protein stabilization (Marqusee and Baldwin, 1987). Many studies have shown that pairs of charged residues that form potential salt-bridges stabilize soluble α-helices (Donald et al., 2011). Salt bridges play an important role in the folding of globular proteins and, despite their low occurrence in TM domains, it seems that the contribution in membrane protein stability could be even more determinant. This contribution is especially important in membrane protein biogenesis, as salt bridges help to bury the polarity of charged residues in a hydrophobic environment (Mbaye et al., 2019). Apart from that, it has been suggested that potential salt bridges could help in the

insertion of TM α-helices (Baeza-Delgado et al., 2016; Bañó-Polo et al., 2012), even though their predicted ΔG penalty is well above what is usually seen for TM segments.

To investigate the potential formation of intra-helical salt bridges in TM α-helices, we analyzed the composition of the TM domains from membrane proteins of known structures looking for preferences in the pairing of charged amino acids. This analysis showed that charged residue pairing is more prevalent than expected for pairs located on the same face of α-helices within membranes. Likely, salt bridge formation on the same face of α-helices reduces the unfavorable energetics of inserting ionizable residues into the hydrophobic membrane core (Chin and Heijne, 2000). We use this knowledge together with the known membrane protein structures to generate a list of potential candidates for further *in vitro* and *in vivo* experiments.

In this work, we have studied the presence of intra-helical salt bridges in TM domains using *in silico*, *in vitro* and *in vivo* systems, showing that despite this being an unexpected phenomenon in nature, salt bridges can be crucial for membrane insertion. Also, we determined in a quantitative manner that the apparent free energy (ΔG_{app}) of membrane insertion through the translocon machinery can be decreased between 0.5-0.7 kcal/mol by position-specific charge pair interaction, which is not contemplated in the commonly used ΔG predictors. These findings will lead to a better understanding of the insertion mechanism of TM helices and to improve prediction tools that would more accurately be able to model the presence of charged residues in these helices.

93 **RESULTS**

94 **Charge pair interactions in model transmembrane helices.**

95 To test the contribution of potential salt bridges to the translocon-mediated
96 membrane insertion, we used the vehicle protein leader peptidase (Lep) from
97 *Escherichia coli* (Fig.1a). The Lep protein consists of two TM segments (H1 and H2)
98 connected by a cytoplasmic loop (P1) and a large C-terminal (P2) domain, which
99 inserts into endoplasmic reticulum (ER)-derived rough microsomes with both
100 termini located in the microsomes lumen. The designed TM segments were inserted
101 into the luminal P2 domain and flanked by two acceptor sites (G1 and G2) for N-
102 linked glycosylation. Glycosylation occurs exclusively in the lumen of the ER (or
103 microsomes) because of the location of the oligosaccharyltransferase (OST) active
104 site (a translocon-associated enzyme responsible for the oligosaccharide transfer)
105 (Braunger et al., 2018). In this case, the engineered glycosylation sites can be used
106 as membrane insertion reporters because G1 will always be glycosylated due to its
107 native luminal localization, but G2 will be glycosylated only upon translocation of
108 the analyzed sequence across the microsomal membrane. A singly glycosylated
109 construct in which a tested sequence is inserted into the membrane has a molecular
110 mass ~2.5 kDa higher than the molecular mass of Lep molecule expressed in the
111 absence of microsomes; the molecular mass shifts by ~5 kDa upon double
112 glycosylation, which facilitates its identification by gel electrophoresis when
113 expressed in the presence of [^{35}S-labeled] amino acids. Then, *in vitro*
114 transcription/translation of these chimeric proteins in the presence of rough
115 microsomal membranes (RMs) allows for accurate and quantitative description of
116 membrane insertion of designed sequences (Bañó-Polo et al., 2019; Hessa et al.,

117 2005; 2007; Tamborero et al., 2011). The degree of membrane insertion is quantified
118 by analyzing the fractions of singly glycosylated (i.e., membrane inserted) and
119 doubly glycosylated (i.e., non-inserted) molecules, which can be expressed as an
120 experimental apparent free energy of membrane insertion, ΔG_{exp} (see Materials and
121 Methods) (Hessa et al., 2005).

122 We first compared the effects of oppositely charged Lys and Asp residues
123 on the insertion of a 19-residue-long hydrophobic stretch (L4/A15 scaffold, 4
124 leucines and 15 alanines), which was designed to insert stably into the RM
125 membranes (Hessa et al., 2005), including charges centered in the TM segment at
126 different positions (Fig. 1b) and "insulated" from the surrounding sequence by N-
127 and C-terminal GGPG- and -GPGG tetrapeptides. Single Lys and Asp residues
128 were placed in positions 8 and 12 respectively, and pairs of Lys-Asp residues were
129 designed to cover positions 7-12 (that is, more than one helical turn). When pairs of
130 charged residues are present, our results showed a tendency to insert more
131 efficiently when pair charges were placed in positions (*i*, *i*+1; *i*, *i*+3; *i*, *i*+4) that are
132 permissive with salt bridge formation (Fig. 1c), actually an effect not observed in the
133 predictions (Fig. 1b). Similar results were obtained on a different Leu/Ala
134 background with a slightly higher insertion efficiency (L5/A14, 5 leucines and 14
135 alanines), those mutants harbouring charged pairs compatible with salt bridge
136 formation (i.e. *i*, *i*+3; *i*, *i*+4) insert more efficiently than the non-compatible one *i*, *i*+5
137 (Figure S1). Being the insertion of charged residues a thermodynamically
138 inconvenient phenomenon within the membrane environment, it is expected that
139 the sequence context and the amino acid composition of the TM helix would be
140 determinant for salt bridge formation. Accordingly, we scrutinized a large dataset of

141 membrane proteins of known three-dimensional structures in order to focus on
142 natural salt bridges present within TM segments.
143
144 **Charged pairs in transmembrane helices.**
145 Alpha helices are a common secondary structure in both globular and membrane
146 proteins. We created two main datasets, TM dataset with helical membrane
147 proteins of known structure from the PDBTM-dataset (Kozma et al., 2013), and
148 GLOB dataset with globular alpha helical proteins selected from the SCOP-
149 database (Andreeva et al., 2013; 2019), see methods for the full creation steps.
150 Table 1 shows a breakdown of alpha-helices, charged residues and salt bridges in
151 the two datasets. What is clear is that long alpha helices (≥ 17 residues) form a
152 larger proportion in the TM dataset than in the globular helices (GLOB) dataset,
153 which is logical as most TM helices need to span through the hydrophobic core of
154 the lipid bilayer of thickness ~30 Å.
155 Our TM dataset showed the same distribution of polar charges as previous
156 studies (Illergård et al., 2011), with about 10 % of polar residues in the core
157 membrane regions (see Table 1), and about one-third of these are charged. Over
158 half of the charged residues could form pairs with other charged residues at
159 intervals of *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4. In contrast, the GLOB dataset had a much higher
160 proportion of charged polar residues. The GLOB dataset also contained about 15
161 times as many charged pairs relative to its size compared to the TM dataset, again
162 indicating that charged residues are more common in globular than in TM helices.
163 The two datasets, TM and GLOB, were extended with homologous
164 sequences identified by searching with jackhammer against UniProt. All sequences

with an E-value lower than 10^{-3} were included. These datasets are named TM-MSA and GLOB-MSA. Using data in the TM-MSA dataset to produce the log odds ratios, we identified periodicity patterns of charged residue pairs that are more common than what would have been expected from the underlying amino acid composition (see Figure 2). We found that polar residues at pairs *i*, *i*+3; *i*, *i*+4 and *i*, *i*+7 are significantly enhanced (Fig. 2). This feature is strengthened when we examine the same plot for the GLOB-MSA dataset, where these patterns were not observed (Figure S2). This was also clear when statistical significance was taken into account, see Figures S3 and S4.

Table 2 lists the log odds ratios together with errors and familywise error-corrected *p*-values for all pairs at separation up to seven residues. It is clear that pair residues placed at *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4 positions are most significant, especially in the case of oppositely charged pairs.

Structural analysis of charged residues in transmembrane helices.

The abundance of salt bridges also shows a remarkable difference between TM and globular proteins. In the GLOB dataset over 13% of the proteins contains at least one local salt bridge in an alpha helix whereas just over 5% of the membrane proteins contains a local salt bridge, see Table 1. This does conform to our current understanding of soluble versus TM alpha-helices and their different environments, and suggests that salt bridges in TM regions perform a vital functional and/or structural role.

As seen in Figure 2, charged pairs of amino acids are especially prevalent at positions *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4. Oppositely charged residues stand out, especially

Glu-Arg at *i*, *i*+1, Glu-Lys at *i*, *i*+3 and Asp-Lys at *i*, *i*+4. Also, several same charged pairings at *i*, *i*+3 and *i*, *i*+6 are more frequent than expected. Other known structural features can also be hinted at, including aromatic ring stacking by His-Trp pair (Samanta et al., 1999) at *i*, *i*+6, and contacting with prosthetic groups by His-His pair at *i*, *i*+7 (Illergård et al., 2011).

Charged pairs placed at *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4 could potentially form salt bridges as they are all on the same relative face of the alpha helix and are close enough in vertical separation on the helix (see Figure 3, top). Although oppositely charged pairs at *i*, *i*+7 are also on the same face of the alpha helix, unless the alpha helix has a bend, both residues are too far separated to form a salt bridge. This was clearly seen in Figure 3 where the TM dataset was used. In both absolute count and log odds ratios (Fig. 3a) it is clear that residues at *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4 are by far the most common and overrepresented pairings. Figure 3a also shows that oppositely and same charged pairs have about the same overrepresentation at *i*, *i*+3, whereas oppositely charged pairs are stronger than same charged pairs at *i*, *i*+1and *i*, *i*+4, both within salt bridge range, and same charged pairs are stronger at positions *i*, *i*+7 and *i*, *i*+8, too far to form salt bridges.

When structure-observed salt bridges in the different positions were compared to the oppositely charged pairs a clear image aroused, see Figure 3b. Even though there are more oppositely charged pairs at position *i*, *i*+1 than in positions *i*, *i*+3 ~~and *i*, *i*+4~~ with both log odds ratios over 1.0 (Fig. 3a), only about 15% of the oppositely charged pairs at *i*, *i*+1 form salt bridge (Fig. 3b). This is in contrast to *i*, *i*+3 where almost 40 % of the pairs form salt bridges, and just under 25% at *i*, *i*+4 (Fig. 3b).

213

214 **Selection of natural salt bridges from membrane protein structures.**

215 To further look for candidate proteins containing membrane-spanning helices with
216 salt bridges we started with the redundant (TM-Red) dataset of known structures
217 (see Materials and Methods). Each of the 8,687 proteins in our dataset were
218 scanned for oppositely charged pairs in positions *i, i*+1*; i, i*+3 and *i, i*+4 within any
219 TM segments core region. For each of these, we only kept proteins with at least one
220 TM segment that contains such a pair and where this pair was within salt bridging
221 distance.

222 To select for potential candidates, we also look at the estimated ΔG_{pred}
223 values and choose helices with a positive value above one, as per our hypothesis,
224 a salt bridge helps TM insertion of sequences for which the hydrophobic force
225 would not be enough to insert. This stricter definition results in a set of 426
226 candidates of a total of 431 salt bridges with a wide range of estimated ΔG_{pred}
227 penalty values (Figure S5). As shown in Figure S5, most TM segments with a salt
228 bridge exhibit a surprisingly high ΔG_{pred} value above 0 that in normal circumstances
229 are not expected to insert into a membrane. Then, we selected TM7 (helix G) from
230 halorhodopsin protein (PDB ID: 3QBG) with an estimated ΔG_{pred} value above +1.7
231 kcal/mol, and helix A from calcium ATPase (PDB ID: 1SU4) with a higher estimated
232 ΔG_{pred} value (above +4.1 kcal/mol), as candidates for systematic studies to cover a
233 wide range of insertion penalties. See the github repository
234 (https://github.com/ElofssonLab/salt_bridges) for the full lists.

235

236 **Intra-helical salt bridge stabilizes the insertion of helix G from halorhodopsin.**

237 Halorhodopsin (hR) from *Natronomonas pharaonis* (3QBG) is a protein made up of
238 seven TM helices (helix A through G) and a retinal chromophore that is bound via a
239 protonated Schiff base to the ε-amino group of a lysine (K258) residue located
240 roughly in the middle of helix G (Kanada et al., 2011). *In silico* analysis of 3QBG
241 structure, the anion-free form of the protein, showed a charged pair of amino acids
242 (Asp-Lys), involving the functional K258 and the D254 in the center of helix G (Fig.
243 4). The position (*i*, *i*+4) and the distance between the anionic carboxylate ($RCOO^-$)
244 from the Asp residue and the cationic ammonium (RNH_3^+) from the lysine residue,
245 in the crystal structure, was about 3.5 Å, a permissive distance for a salt bridge
246 formation (Fig. 4g), which has been established as being lower than 4 Å (Kumar and
247 Nussinov, 2002). To get insights into this interaction, we designed three mutants
248 that were supposed to perturb the salt bridge interaction by different ways: K258D
249 mutant by placing two charged residues with the same polarity at positions *i*, *i*+4;
250 K258A mutant by replacing one of the charged residues by a non-polar amino acid,
251 and K258Y/Y259K double mutant by placing the charged pair at a non-permissive
252 salt bridge position (*i*, *i*+5; Fig. 4g), while keeping the same amino acid composition
253 (Fig. 4a).

254 Halorhodopsin is a trimeric protein in which the helix G is neither exposed at
255 the monomer-monomer interface nor oriented to the inner part of the trimeric
256 structure (Fig. 4, panels b-d). In fact, the charged pair found in helix G is oriented
257 toward the core of the 'globular' structure in each monomer. Then, the insertion of
258 helix G was studied using the Lep-based assay. When constructs harboring helix G
259 wild type sequence were translated *in vitro* in the presence of RMs singly-
260 glycosylated (reporting insertion) forms were found (Fig. 4e, lane 2), despite its

261 positive (suggesting non-insertion) ΔG_{pred} value (Fig. 4a). The nature of the higher
262 molecular weight polypeptide species was analysed by endoglycosidase H (EndoH)
263 treatment, a highly specific enzyme that cleaves N-linked oligosaccharides.
264 Treatment with EndoH of the samples eliminated higher molecular mass bands (Fig.
265 4E, lane 1), confirming the sugar source of the retarded electrophoretic mobility
266 bands and suggesting helix G insertion into the microsomal membrane. However,
267 locating the Asp-Lys pair at *i*, *i*+5, which is non-compatible with salt bridge
268 interaction (Fig. 4g), strongly reduced the experimental insertion efficiency (Fig. 4e,
269 lane 5). Interestingly, replacing the positively charged lysine residue by a negatively
270 charged aspartic acid residue (K258D) rendered similarly low levels of insertion
271 efficiency (Fig. 4e, lane 4). As expected, replacement of the ionizable lysine residue
272 by the aliphatic alanine increased the insertion efficiency (Fig. 4e, lane 3). In this
273 later mutant, the most likely cause for the increased insertion is the absence of the
274 positively charged lysine amino side chain and by the presence of the methyl side
275 chain group of the mutant alanine residue. The results of the Lep-based
276 glycosylation assay indicated that wild type (wt) and DA sequences (two stabilized
277 charges or only one charge, respectively) are inserted properly into the microsomal
278 membrane (ΔG_{exp} values -0.24 and -0.88 kcal/mol, respectively), but when the salt
279 bridge is disrupted, either by having two charged amino acids with the same polarity
280 (DD) or by placing oppositely charged residues at a non-permissive distance (*i*, *i*+5)
281 in the center of the helix, the translocation of the segment increases substantially.
282 It should be mentioned that the K258Y/Y259K double mutant has the same amino
283 acid composition than the original helix G, but insertion efficiency is remarkably
284 decreased (ΔG_{exp} = +0.31 kcal/mol). Together these results show that the interaction

285 (salt bridge) between Asp and Lys residues in the center of the helix G from 3QBG
286 is essential for its proper insertion into the microsomal membrane. The salt bridge
287 contributes approximately ~0,5 kcal/mol to the apparent experimental free energy
288 of microsomal membrane insertion, as this is the difference found between the
289 ΔG_{exp} values for the wt and *i*, *i*+5 mutant.

290 Next, to ensure that the *in vitro* results are relevant to the *in vivo* situation, wt
291 and *i, i*+5 constructs were also expressed *in vivo* in HEK-293T cells. To this end, a
292 *c*-myc tag was engineered at the C-terminus of the Lep chimera to allow immune-
293 detection of our constructs in cell extracts. As shown in Fig. 4f, transfected cells
294 with the chimera containing helix G wt sequence rendered singly glycosylated
295 molecules, indicating *in vivo* membrane insertion. In contrast, cells transfected with
296 the construct harboring *i*, *i*+5 sequence rendered almost exclusively doubly
297 glycosylated forms, as proved by EndoH treatment (Fig. 4f, lane 2), suggesting
298 membrane translocation. These results emphasized the relevance of salt bridge
299 interactions in translocon-mediated TM insertion, especially in the *in vivo* (cellular)
300 environment.

301

302 **Salt bridge contribution to the insertion of a heavily charged helix.**

303 In order to challenge salt bridge interactions in a more hydrophilic TM helix (Figure
304 S5), we focus on helix A from the sarcoplasmic/ER calcium ATPase 1 (Toyoshima
305 et al., 2000). Calcium ATPase (PDB ID: 1SU4, *Oryctolagus cuniculus*) is a member
306 of the P-type ATPases that transport ions across the membrane against a
307 concentration gradient involving 10 α-helices (helix A through J) in the membrane-
308 embedded region (Toyoshima et al., 2000). *In silico* analysis of 1SU4 structure, the

309 crystal structure of the calcium ATPase with two bound calcium ions, showed Asp-
310 Arg (DR) pair (D59 and R63) in the center of the helix A. This helix extends beyond
311 the membrane (Fig. 5a, from Leu49 to Phe78) and shows some particularities. On
312 the one hand, in the structure the membrane-embedded stretch (the N-terminal
313 region of helix A) encompasses from Leu49 to Ala69 residues, and includes several
314 charged amino acids, probably involved in the Ca^{2+} transport across the membrane
315 (Glu51, Glu55, Glu58, Asp59 and Arg63). Therefore, the ΔG_{pred} value for this
316 segment (L49-A69) is remarkably higher (and positive, +4.12 kcal/mol) than
317 expected for a TM helix (Fig. 5b). On the other hand, the C-terminal region of this
318 helix contains a high prevalence of non-polar amino acids that is more compatible
319 with the hydrophobicity of the membrane core. Nevertheless, the presence of the
320 functional amino acids (e.g. Glu58) in the membrane-embedded region reinforce
321 the idea that the more hydrophilic N-terminal region must ultimately be embedded
322 within the membrane (Toyoshima et al., 2000). It has been previously shown that
323 the position in the membrane of TM helices in protein folded structures does not
324 always correspond to the thermodynamically favored positions in the membrane of
325 the isolated helices (Kauko et al., 2010). Instead, after translocon-mediated insertion
326 of the more hydrophobic region, repositioning of TM helices relative to the lipid
327 bilayer provides a convenient way for non-hydrophobic polypeptide segments to
328 become buried within the membrane. Then, the nature of helix A suggested the
329 possibility that initial insertion of the hydrophobic region can be followed by
330 subsequent repositioning of the charged region into the membrane hydrophobic
331 core, which will be the final segment embedded in the lipid bilayer (Fig. 5a).
332 Interestingly, the ΔG Predictor server (https://dgpred.cbr.su.se/index.php?p=home)

333 selected the adjacent (C-terminus, L60-F78) hydrophobic region as TM (Fig. 5b),
334 instead of the charged region found at the high-resolution structure (Uniprot code:
335 P04191).

336 Focusing on the potential salt bridge residues (D59 and R63 pair), the
337 distance between the anionic carboxylate ($RCOO^-$) from the D59 and the cationic
338 guanidinium ($RC(NH_2)_2^+$) from the R63 was about 3.0 Å in the crystal structure
339 (Figure 5a), clearly within the permissive range for salt bridge formation. To
340 investigate the contribution of this potential salt bridge interaction in the translocon-
341 mediated membrane insertion of this region, we worked with two different scaffold
342 sequences: the full helix A involving the residues 49-78 (Long, L); and a shorter
343 membrane-embedded version including residues 49-69 (Short, S) as found in the
344 solved structure. We also challenged the D59-R63 charge pair interaction in both
345 sequences by increasing the separation between the ionizable residues from the
346 native *i*, *i*+4 to non-permissive *i*, *i*+5, while maintaining amino acid composition
347 (R63I/I64R double mutant). *In vitro* transcription/translation of these sequences in
348 the presence of microsomes rendered singly glycosylated molecules for the
349 construct containing full-length helix A (Fig. 5c, lane 2). In contrast, when only the
350 membrane-embedded sequence was included, the Lep chimera was mainly doubly
351 glycosylated (Fig. 5c, lane 3), suggesting that in the full-length protein helix A inserts
352 initially through the more hydrophobic L60-F78 region and then, after protein
353 rearrangements, repositions the more hydrophilic L49-A69 region at the membrane
354 core, as found in the solved structure. Accordingly, the translocon inefficiently
355 inserted the isolated membrane-embedded (L49-A69) region (Fig. 5c, lane 3),
356 properly inserting helix A only when the full helical sequence is present (Fig. 5c, lane

357 2). When the charge paired residues were placed at a non-permissive distance in
358 terms of salt bridge interaction (*i*, *i*+5; Fig. 5e) the insertion efficiency was reduced
359 (Fig. 5c, lane 4), with a ΔG_{app} decrease (absolute values) of ~0,7 kcal/mol relative to
360 the wild type sequence (Fig. 5b). As expected, this effect was not observed when
361 the same mutations were grafted on the membrane-embedded (S) sequence (Fig.
362 5c, lane 5).

363 Next, we analysed the salt bridge interaction in HEK-293T cells to study the
364 translocon performance *in vivo*. When cell cultures were transfected with a Lep-
365 derived chimera containing the helix A wild type sequence only singly glycosylated
366 molecules were observed (Fig. 5d, lanes 1 and 2). However, a construct harboring
367 double mutant (R63I/I64R; *i*, *i*+5) sequence showed doubly glycosylated molecules
368 to a measurable extent (Fig. 5d, lanes 3 and 4), indicating a lower insertion efficiency
369 that can be attributed to the altered salt bridge interaction.

370

371 **Effect of salt bridge formation in the absence of previous TM regions.**

372 To investigate the effect of salt bridge formation in translocon-mediated membrane
373 insertion in the absence of precedent TM regions we used a different glycosylation-
374 based reporter system in which the TM sequences of interest (bR helix G and
375 ATPase helix A) were connected to the well-folded constant domain of the antibody
376 λ light chain, C_L (Feige and Hendershot, 2013). The C-terminal of the TM sequence
377 carries an Asn-Val-Thr glycosylation site (G2), and we engineered an extra
378 glycosylation site (G1) within the C_L sequence (Figure 6a, top). As in the case of the
379 Lep system, G1 will always be glycosylated due to the native translocation of the λ
380 light chain, but G2 will be glycosylated only upon translocation of the analyzed

381 (tested) sequence across the ER membrane (Figure 6a, bottom). When constructs
382 harboring either hR helix G or ATPase helix A wild type sequences were transfected
383 into Hek293T cells, singly glycosylated (reporting insertion) forms were
384 predominantly found (Fig. 6b, lane 1 and Fig. 6c, lane 2, respectively), discarding
385 any potential contribution to membrane insertion of precedent TM segments. On
386 the contrary, sequences containing non-permissive (*i*, *i*+5) salt-bridge forming pairs
387 were more efficiently doubly glycosylated (Figures 6b, lane 3 and 6c, lane 4,
388 respectively). Thus, similar to what was observed using the Lep system (Figures 4f
389 and 5d), the presence of chair pairs at permissive salt bridge formation distances
390 strongly promotes helix integration into the ER membrane.
391

392 **DISCUSSION**

393 Charged residues found in alpha-helices can both give a stabilizing effect
394 (Armstrong and Baldwin, 1993) as well as being important for interaction and
395 function such as in zinc finger motifs (Lin and Lin, 2018). Charged residues have
396 also been found to be more common in globular (soluble) helices compared to TM
397 helices (Kauko et al., 2008). This can be explained by the fact that globular helices
398 reside in a more hydrophilic environment compared to the hydrophobic
399 environment of the lipid bilayer. An intermediate between these is the existence of
400 amphipathic alpha-helix in contact with the surface of a bilayer, where one face
401 contains mainly polar residues facing an aqueous environment and the opposite
402 face with mostly nonpolar residues facing a hydrophobic environment (Giménez-
403 Andrés et al., 2018).

404 Whereas pairs of charged residues of the same charge facing the inside
405 pores in TM regions fill an essential functional role, pairs of oppositely charged
406 residues can form salt bridges that can stabilize the helix and play an important role
407 in function such as for transportation (Walther and Ulrich, 2014). Salt bridges might
408 also be important for hiding the charges during translocon-mediated TM helix
409 insertion, as the charged residues are hidden from the hydrophobic environment.
410 Previous work has shown that charged and polar residues are conserved within TM
411 segments (Illergård et al., 2011), indicating they are crucial for function and/or
412 stability.

413 Asp-Lys pairs at position *i*, *i*+4 and Glu-Lys pairs at position *i*, *i*+3 are the
414 most prevalent as seen previously in Figure 2. They are both among the most
415 prevalent oppositely charged pairs and the charged pairs that form the highest
416 number of salt bridges in membrane protein structures. This is in stark contrast to
417 Glu-Arg pair at position *i*, *i*+1 that although as frequent in pairs as Asp-Lys and Glu-
418 Lys at positions *i*, *i*+4 and *i*, *i*+3 respectively, only form salt bridges in one-fourth of
419 the cases as found in Fig. 3b.

420 An interesting observation is that positively charged Arg-Arg pairs at position
421 *i*, *i*+3 is numerically the most common pair at this position. Positively charged pairs
422 at *i*, *i*+6 are also more prevalent than expected and although Arg-Arg pairs
423 numerically only show up half as often at *i*, *i*+6 as in *i*, *i*+3, 80% of the Arg-Arg pairs
424 at position *i*, *i*+6 also contain an arginine at *i*+3, as found in helices from voltage-
425 gated ion channels that contain three or more periodically aligned Arg residues with
426 two intervening hydrophobic residues (Okamura et al., 2015).

427 The fraction of salt bridges at positions *i*, *i*+5 (Fig. 3b) is an anomaly due to
428 bend alpha-helices, as found in the helix E from bacterial translocon (PDB ID: 5MG3)
429 and in the helix A from a lipid flippase (PDB ID: 6CC4) due in both cases to the
430 presence of a glycine residue (Figure S6). Without the observed helix bending, these
431 salt bridges would not be formed.

432 By analyzing two native helices containing intra-helical salt bridges we now
433 find that the free energy of insertion (ΔG_{app}) is significantly reduced if both oppositely
434 charged residues are spaced at a permissive distance. These results indicate that
435 intra-helix salt bridge can form during translocon-assisted insertion or even earlier,
436 since in contrast to globular (soluble) helices, TM helices can be compacted inside
437 the ribosome exit tunnel (Bañó-Polo et al., 2018). The maximal reduction in ΔG_{app}
438 seen with Asp-Lys and Asp-Arg pairs in both hR and Ca^{2+} ATPase helices is 0.5-0.7
439 kcal/mol, which is in good agreement with the 0-1 kcal/mol estimated for these two
440 pairs from thermodynamic peptide partition into octanol experiments (Jayasinghe
441 et al., 2001). As found in the case of hR helix G (Figures 4f and 6b), this reduction
442 might be even higher in the cell context, since some auxiliary components of the
443 membrane insertion machinery (Chitwood and Hegde, 2020; Shurtleff et al., 2018;
444 Tamborero et al., 2011) can be in suboptimal conditions in the microsomal vesicles.

445 As mentioned above, in the case of hR helix G the lysine residue involved in
446 the salt bridge (K258) is bound to the retinal chromophore via a protonated Schiff
447 base as found in the crystal structure of a close homologue (Kolbe et al., 2000).
448 Then, the lysine residue plays a fundamental role for protein function but at the
449 same time introduces a penalty for membrane insertion. Interestingly, our data
450 suggest that helix G translocon-mediated insertion efficiency could be increased by

451 salt bridge formation between K258 and D254, and once in the membrane, retinol
452 binding to the apoprotein could occur by covalently binding the retinal as a
453 protonated Schiff base to K258 and perturbing D254 salt bridge interaction.

454 Beyond the conceptual issues involving the membrane insertion process, we
455 note that the availability of quantitative experimental data on the contribution of salt
456 bridge interactions to the free energy of insertion (ΔG_{app}) will make it possible to fine-
457 tune current membrane protein topology-prediction methods based on free energy
458 calculations. Although today's state of the art topology prediction tools uses
459 amphiphatic biologically derived scales, they do not take these types of salt bridge
460 interaction into account. Current algorithms tend to overestimate the free energy
461 of insertion due to the penalty acquainted by charged residues in the TM region.
462 However, distinguishing between charged residues of the same or opposite
463 polarity, i.e., incorporating the effect of potential salt bridges in the reduction of
464 ΔG_{app} during membrane integration should help to make prediction tools even more
465 accurate.
466

467 **MATERIALS AND METHODS**

468 **Enzymes and chemicals.** TNT T7 Quick for PCR DNA was from Promega
469 (Madison, WI, USA). Dog pancreas ER column washed rough microsomes were
470 from tRNA Probes (College Station, TX, USA). EasyTag™ EXPRESS^{35}S Protein
471 Labeling Mix, [^{35}S]-L-methionine and [^{35}S]-L-cysteine, for *in vitro* labeling was
472 purchased from Perkin Elmer (Waltham, MA, USA). Restriction enzymes were from
473 New England Biolabs (Massachusetts, USA) and endoglycosidase H was from
474 Roche Molecular Biochemicals (Basel, Switzerland). PCR and plasmid purification
475 kits were from Thermo Fisher Scientific (Ulm, Germany). All oligonucleotides were
476 purchased from Macrogen (Seoul, South Korea).

477 **DNA Manipulation.** The sequences of interest were introduced into the modified
478 Lep sequence from the pGEM1 plasmid (Hessa et al., 2005) between the *Spe*I and
479 *Kpn*I sites using two double-stranded oligonucleotides with overlapping overhangs
480 at the ends. The complementary oligonucleotides pairs were first annealed at 85 °C
481 for 10 min followed by gradual cooling to 30 °C and ligated into the vector (a kind
482 gift from G. von Heijne's lab). Mutations were obtained by site-directed mutagenesis
483 using the QuikChange kit (Stratagene, La Jolla, California). Lep system including the
484 sequences of interest in the P2 region were subcloned into *Kpn*I linearized pCAGGS
485 in-house version using In-Fusion HD cloning Kit (Takara) according to the
486 manufacturer's instructions. An engineered glycosylation site (Q36N) was added to
487 the C_L-TM plasmid (a kind gift from L. Hendershot's lab), in which the sequences
488 from hR helix G and ATPase helix A were introduced flanked by 'insulating' Gly-Pro
489 tetrapeptides. A *c*-myc tag (EQKLISEEDL) at the C-terminus of the Lep- and C_L-
490 derived sequences was added by PCR before cloning. For *in vitro* assays, DNA was

491 amplified by PCR adding the T7 promoter during the process. All sequences were
492 confirmed by sequencing the plasmid DNA at Macrogen Company (Seoul, South
493 Korea).

494 **Translocon-mediated insertion into microsomal membranes.** Lep constructs in
495 pGEM with L4/A15, L5/A14, 3QBG and 1SU4 segments and its variations were
496 transcribed and translated using the TNT T7 Quick Coupled System (#L1170,
497 Promega). Each reaction containing 1 μL of PCR product, 0.5 of EasyTag™
498 EXPRESS 35S Protein Labeling Mix (Perkin Elmer) (5.5 μCi) and 0.3 μL of
499 microsomes (tRNA Probes) was incubated at 30ºC for 90 min. Endo H treatment
500 was done following the manufacturer's instructions. Samples were analysed by
501 SDS-PAGE (12-14% polyacrylamide). The bands were quantified using a Fuji FLA-
502 3000 phosphoimager and the Image Reader 8.1j software. Free energy was
503 calculated using: ΔG_{app}=-RT lnK_{app}, where $K_{app}=f_{2g}/f_{1g}$ being f_{1g} and f_{2g} the fraction of
504 singly glycosylated and double glycosylated protein, respectively.

505 **Free apparent insertion energy, ΔG.** The free insertion energy of a TM region, ΔG,
506 is calculated as per the experimentally defined Biological hydrophobicity scale
507 (Hessa et al., 2005). This scoring is amphiphilic with hydrophobic residues
508 contributing a lower (negative) ΔG while hydrophilic contributes a higher (positive)
509 ΔG. The total ΔG of a region is the sum of individual position specific scores. This
510 scoring can give an indication of how favorable the amino acid composition of a TM
511 region is to be inserted in a lipid bilayer membrane. To note is that hydrophobicity
512 alone is not the only driving force and that the positive inside rule (Heijne, 1989;
513 Lerch-Bader et al., 2008) and help from proceeding TM regions (Bañó-Polo et al.,
514 2013; Hedin et al., 2010) can assist insertion in polytopic TM proteins especially.

515 **Core segment definition.** We define core segments as a TM region minus the first
516 and last 5 residues. This is to ignore the interface regions which are known to
517 contain polar residues.
518 **Salt bridge definition.** Salt bridges are defined as per (Kumar et al., 2000), where
519 a salt bridge is defined if a side chain carbonyl oxygen atom in Asp-Glu is within 4.0
520 Å from the nitrogen atom in Arg-Lys. This conforms to other works (Bosshard et al.,
521 2004; Donald et al., 2011) with the definition that the atoms are within hydrogen
522 bond distance. We also define local salt bridges as being bridges that are separated
523 by at most 7 residues in the sequence. This is to separate long salt bridge
524 interactions, which can occur between spatially close residues that are separated
525 in sequence, such as coiled coils where salt bridges can be between separate
526 alpha-helices.
527 **Transmembrane helices dataset (TM dataset).** The full pipeline is available as a
528 Makefile together with supporting scripts in the github repository. The full PDBTM
529 database (Kozma et al., 2013) was downloaded together with their list of non-
530 redundant protein pdb ids. This list is used to generate both sequence and topology
531 of the proteins by extracting both from the PDBTM-xml. For each protein in the non-
532 redundant list, the membrane regions are extracted as per the PDBTM database
533 annotation. All non-membrane regions are annotated 'i' for convenience. To support
534 future analysis, membrane regions longer than 10 were run through DSSP and
535 annotated with 'M' if all residues in the core segment were defined as alpha-helix
536 ('H'), otherwise, the full membrane region is annotated 'm'. Observe that this creates
537 fasta-like 3line files, that only contain topological annotation with ambiguous TM
538 regions annotated as 'm' instead of the normal 'M'. These proteins are then cluster

539 using cd-hit (Fu et al., 2012; Li and Godzik, 2006) at 40% identity using the
540 parameter -c 0.4 -n 2 -T 0 -M 0 -d 0.

541 During the extraction of TM regions, the corresponding structure file from
542 RCSB was used to calculate all salt bridges within the current protein and any salt
543 bridge that has at least one residue within any TM region was saved. Additionally,
544 all salt bridges whose both residues were within the same segment and within 7
545 residues of each other were annotated as local as per the salt bridge definition
546 above.

547 **Extraction of charged residues.** From all annotated TM regions of length 17 or
548 longer (Baeza-Delgado et al., 2012), the core segment was extracted. All these core
549 regions were then scanned and when a charged residue was encountered, we
550 recorded any other charged residue from 7 residues before the current one to 7
551 residues after. This results in charged residues that can contain a charged pairing
552 partner outside of the TM segment and will therefore differ slightly from charged
553 pairs which are defined next.

554 **Extraction of charged pairs.** From all annotated TM regions of length 17 or longer
555 (Baeza-Delgado et al., 2012), the core segment was extracted. All these core
556 regions were then scanned and when a charged residue was encountered, we look
557 at up to 7 residues in front of it or to the end of that core region, whichever came
558 first. All occurrences of charged pairs were recorded, resulting in charged pairs
559 where both residues were fully within the core segment of a TM helix.

560 **MSA-dataset extension (TM-MSA).** Using the TM dataset, we extended it by
561 creating an MSA alignment of each protein using jackhmmer (Eddy, 2011) against

562 Uniref90 with one iteration and an E-value cut off of 10^{-3} with the following
563 parameters:
564 -N 1 -E 1e-3 --incE 1e-3 --cpu 14
565 From each alignment, we then sampled up to 200 hits, including the initial
566 seed sequence. If there were fewer than 200 hits, we used them all. We then used
567 the original topology for each alignment to extract all TM regions, only to include
568 parts where the sequence covers the full TM region and where the sequence did
569 not contain any insertions or deletions.
570 **Dataset of helices from globular proteins (GLOB and GLOB-MSA).** To create a
571 reference dataset of globular α-helical proteins we extracted all globular all-alpha
572 protein domains from SCOP. As SCOP classifies domains of proteins resulting in
573 that one domain of a protein can be annotated as globular whereas another domain
574 as TM (see 1PPJ chain D as an example) we reduced the SCOP list against the
575 redundant list of all PDBTM (Kozma et al., 2013) chains to clear any overlap. This
576 results in 4,500 proteins in total.
577 Topological files with sequence and membrane topology are created with the
578 help of the RCSB secondary structure file and only membrane segments whose
579 core region (*i.e.*, central 15 residues) is annotated as pure (canonical) α-helices were
580 retained, *i.e.*, those TM segments containing any residues within the core annotated
581 as loops or other types of secondary structures were removed. This file was further
582 homology reduced and alignments prepared in the same manner and using the
583 same parameters as the TM dataset described above.

The GLOB-MSA dataset was created in the same way as the TM-MSA dataset using jackhmmer to extend the sequences to alignments and then to extract helix sequences.

Redundant (TM-Red) dataset. The full PDBTM database was used as in the preparation of the TM dataset. We skip the clustering step and instead use all redundant proteins to generate their respective topology files. We added in the constraint that each selected TM region must fully contain at least one potential salt bridge. This means a local salt bridge where both residues are within the core segment. This dataset was only used to find potential candidates for further *in vitro* and *in vivo* experiments. See section of natural salt bridges above.

Calculation of log odds ratio. The log odds ratios for each amino acid pair for steps 1 through 10 are calculated as follows:

logOddsRatio = log((A/B)/(C/D))

Where for two amino acids p_1 and p_2:

A = number of pairs of p_1 to p_2

B = number of total pairs

C = number of p_1 times number of p_2

D = number of total pairs squared

The standard error, SE, and z-value is calculated as follows allowing for a two-sided test:

$$SE = \sqrt{(1/A + 1/B + 1/C + 1/D)}$$

z = abs(logOddsValue/SE)

The survival function, sf, from the scipy python packages is used to calculate the p-values. To correct for multiple hypothesis, the Bonferroni Correction is used

608 based on the number of hypothesis, 20 * 20 * 10, number of amino acids square
609 times the number of steps.
610 **Expression in mammalian cells.** Lep or C_L-derived constructs containing 3QBG
611 or 1SU4 segments and its variations were tagged with *c*-myc epitope at their Ct
612 (EQKLISEEDL) and inserted in the appropriate plasmids. Once the sequence was
613 verified, plasmids were transfected into HEK293-T cells using Lipofectamine 2000
614 (Life Technologies) according to the manufacturer's protocol. Approximately 24 h
615 post-transfection cells were harvested and washed with PBS buffer. After a short
616 centrifugation (1000 rpm for 5 min on a table-top centrifuge) cells were lysed by
617 adding 100 µL of lysis buffer (30 mM Tris-HCl, 150 mM NaCl, 0.5% Nonidet P-40)
618 were sonicated in an ice bath in a bioruptor (Diagenode) during 10min and
619 centrifuged. After protein quantification, equal amounts of protein were submitted
620 to Endo H treatment or mock-treated followed by SDS-PAGE analysis and
621 transferred into a PVDF transfer membrane (ThermoFisher Scientific) as previously
622 described (Duart et al., 2020). Protein glycosylation status was analysed by Western
623 Blot using an anti-*c*-myc antibody (Sigma), anti-rabbit IgG-peroxidase conjugated
624 (Sigma), and with ECL developing reagent (GE Healthcare). Chemiluminescence
625 was visualized using an ImageQuantTM LAS 4000mini Biomolecular Imager (GE
626 Healthcare).
627

628 **Acknowledgments**

629 We thank Pilar Selvi and Beatriz Iborra for excellent technical and administrative
630 assistance, respectively. The C_L-TM plasmid was a kind gift from Prof. Linda M.
631 Hendershot (St. Jude Children's Research Hospital). This work was supported by

grants PID2020-119111GB-100 from the Spanish Ministry of Science and Innovation and PROMETEU/2019/065 from Generalitat Valenciana (to I.M.), and by grant from the Swedish Research Council (VR-NT 2016-03798 to A.E.). G.D. was recipient of a predoctoral contract (FPU18/05771) from the Spanish Ministry of Education.

Conflict of Interest Statement

None declared.

641 **Table 1**. About 10 % of the core residues in the TM dataset are polar and just under
642 4% are charged residues. This is significantly less than 37.5% and 25.5%
643 respectively in the globular dataset. In the transmembrane dataset, just over half of
644 the charged core residues form a charged pair at distance 1, 3 or 4. *In the case of
645 the globular set, there are more potential charged pairs than charged core residues
646 as multiple residues are counted more than once as they form more than one
647 potential pairing with separation of 1, 3 or 4 steps. It is clearly seen that although
648 the TM dataset contains more helices per protein and has a higher proportion of
649 long (≥ 17 residues) helices it contains significantly fewer charged pairs.

Dataset	**TM**	**TM-MSA**	**GLOB**	**GLOB-MSA**
Charge statistics				
Core residues	40116	3990464	74299	7388277
Polar core residues	4080 (10.2%)	490861 (12.3%)	27851 (37.5%)	2696564 (36.5%)
Charged core residues	1467 (3.7%)	204544 (5.1%)	18959 (25.5%)	1819047 (24.6%)
Pairs of charged core residues (+1, +3 or +4)	825	88380	23755*	2271405*
# of same charged pairs	373	45220	10835	1117270
# of oppositely charged pairs	452	43160	12920	1154135
Charge Pairs statistics				
Proteins	925	130934	2475	363610
Total alpha helices	5435	481729	23991	2118158
Alpha helices ≥17	3755	306632	5895	485797
Proteins with charged pairs	204	25187	1780	193534
Helices with charged pairs	265	34029	4251	344142

Proteins with any salt bridge	241	-	*751*	-
Proteins with local salt bridge in helix	56	-	*335*	-
Total number of local salt bridges in helices	59	-	*577*	-

650

651

Table 2. Log odds ratios for all charged pairs, same charged pairs and oppositely charged pairs with calculated errors and multiple hypotheses corrected p-values. It is clear that charged pairs occur more often than predicted, evidenced by the positive log odds ratios in all cases. It is clear that positions +1, +3 and +4 are the most prevalent pairings, with the bolded values highlighting log odds ratio above 0.8. It is also clear that oppositely charged residues which have the potential to form salt bridges are prevalent in all three positions whereas the same charge is mainly prevalent in position 3. Most likely these same charges facing the same face of the helix are involved in functions such as ion transport.

Spacing	***All Log odds***			***Same charged Log odds***			***Oppositely charged Log odds***		
	ratio	***error***	***p-value***	***ratio***	***error***	***p-value***	***ratio***	***error***	***p-value***
+1	***0.879***	*0.025*	*$1.31e^{-257}$*	*0.548*	*0.042*	*$4.03e^{-41}$*	***1.107***	*0.032*	*$1.65e^{-256}$*
+2	*0.252*	*0.037*	*$2.02e^{-08}$*	*0.185*	*0.054*	*$1.88e^{-00}$*	*0.316*	*0.050*	*$1.09e^{-06}$*
+3	***1.073***	*0.026*	*$<0.00e^{-300}$*	***1.118***	*0.035*	*$1.15e^{-214}$*	***1.026***	*0.037*	*$5.78e^{-165}$*
+4	***0.965***	*0.028*	*$1.14e^{-253}$*	*0.696*	*0.046*	*$2.88e^{-49}$*	***1.177***	*0.036*	*$2.74e^{-233}$*
+5	*0.517*	*0.037*	*$4.88e^{-42}$*	*0.390*	*0.055*	*$4.15e^{-09}$*	*0.630*	*0.049*	*$1.21e^{-34}$*
+6	*0.540*	*0.038*	*$1.66e^{-43}$*	*0.545*	*0.053*	*$2.33e^{-21}$*	*0.536*	*0.053*	*$2.10e^{-20}$*
+7	*0.615*	*0.038*	*$1.21e^{-55}$*	*0.764*	*0.050*	*$7.78e^{-50}$*	*0.439*	*0.058*	*$1.97e^{-10}$*

664 **FIGURE LEGENDS**

665

666 **Figure 1. Effects on membrane insertion of single or pairs of Asp and Lys**
667 **residues in a model TM segment. (a)** Schematic representation of the leader
668 peptidase (Lep) model protein. G1 and G2 denote artificial glycosylation acceptor
669 sites. The sequence under investigation was introduced in the P2 region after H2.
670 Recognition of the tested sequence as a TM by the translocon machinery
671 (highlighted in green) results in the modification of the G1 site but not G2. The Lep
672 chimera will be double glycosylated if the sequence being tested is not recognized
673 as a TMD and thus translocated into the microsomes lumen (shown in red). **(b)** The
674 tested sequences from L4/A15 model TM (including the charged residues, bold),
675 the gap distance, and the predicted ΔG (ΔG_{pred}) values in kcal/mol are shown. Amino
676 acids with positive and negative charge are highlighted in blue (K) and red (D)
677 respectively. **(c)** Experimental ΔG (ΔG_{exp}) in kcal/mol of each tested sequence in the
678 Lep-based microsomal assay. The mean and standard deviation of at least 3
679 independent experiments is represented (*n* values: 4 [from 1 to 7] and 3 [8 and 9]).
680 The individual value of each experiment is represented by a solid dot, *p*-values
681 (ordinary one-way ANOVA test with Dunnett correction) are indicated above the
682 corresponding bars with values <0.005 highlighted in green. In addition, a green
683 square represents the experimental ΔG value for the L4/A15 sequence from an
684 earlier study [12]. The wt and single mutants are shown in white bars. Charges at
685 compatible distances with salt bridge formation (*i*, *i*+1; *i*, *i*+3; and *i*, *i*+4) are shown
686 in brown, orange and yellow, respectively. Not compatible distances with salt bridge
687 formation (*i*, *i*+2; and *i*, *i*+5) are shown in dark gray. The inset shows a representative

688 SDS-PAGE gel for L4/A15 construct. The construct was expressed in rabbit
689 reticulocyte lysed in the presence (+RM) or absence (–RM) of rough microsomes.
690 Bands of non-glycosylated proteins are indicated by a white dot; mono and double
691 glycosylated proteins are indicated by one and two black dots, respectively.

692

693 **Figure 2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8'**
694 **for the TM-MSA dataset.** The rows on the y-axis indicate the first amino acid in
695 the pair and the columns on the x-axis the second. The residues are ordered by
696 hydrophobicity according to the Engelman order (Engelman et al., 1986). See figure
697 S1 for the equivalent of the globular dataset. S2 and S3 show the same plots
698 masked for statistical significance.

699

700 **Figure 3. Charge pairs in TM helical sequences and structures.** Helical wheel
701 projection and lateral views of an α-helix are shown on top. The initial position *i* and
702 the following 8 residues are numbered. Residues in positions *i*+3 (orange), *i*+4
703 (yellow), and *i*+7 (light brown) are mainly on the same face of the helix, but *i*+7 is
704 placed too far for a salt bridge interaction. **a)** The log odds ratios of charged pairs
705 for '*i*, *i*+1' through '*i*, *i*+8' in the TM dataset. Plain filled bars refer to oppositely
706 charged pairs and the forward slash are the and same charged pairs, all with error
707 bars. **b)** Fraction of oppositely charged pairs that form local salt bridges. The small
708 bump at *i* +5 are the two proteins 6CC4 (helix A) and 5MG3 (helix E), which both
709 exhibit a bend in the alpha helix due to the presence of glycine residues (see Fig.
710 S6).

711

712 **Figure 4. Insertion of halorhodopsin helix G from *Natronomonas pharaonis***
713 **(3QBG) into microsomal and cellular membranes. (a)** Tested sequences from
714 3QBG including the gap distance, and the predicted (ΔG_{pred}) and experimental (*in*
715 *vitro* $\Delta G_{exp}^{in\ vitro}$ and in vivo $\Delta G_{exp}^{in\ vivo}$, respectively) ΔG values in kcal/mol are shown.
716 Amino acids with a positive or negative charge are highlighted in blue (K) and red
717 (D), respectively. Green numbers indicate negative ΔG (insertion) values, while red
718 numbers denote ΔG values above 0 (translocation). **(b)** Frontal view of 3QBG
719 monomer structure. The helix G is highlighted in orange with the D254 and K258
720 shown in sticks colored red and blue respectively. The membrane position is
721 indicated by a red (outer) and blue (inner) discontinuous line, according to OPM
722 dataset [22]. Lateral **(c)** and upper **(d)** views of the 3QBG trimeric structure. The
723 helix G is highlighted in orange with the D and K shown in sticks colored red and
724 blue, respectively. The different monomers are shown in transparent blue, pink and
725 green. Representative examples (*n*=3) of *in vitro* protein translations in the presence
726 of ER-derived microsomes **(e)** and Western blots (*n*=3) of *in vivo* protein translations
727 in HEK-293T cells **(f)** in the presence (+) or absence (-) of Endoglycosidase H
728 (EndoH), a glycan-removing enzyme. The absence of glycosylation of G1 and G2
729 acceptor sites is indicated by two white dots, single glycosylation by one white and
730 one black dot, and double glycosylation by two black dots. **(g)** Zoom view centered
731 on the salt bridge between D254 and K257 at *i*, *i*+4 (left) and *i*, *i*+5 (right) gaps. D
732 and K residues are shown in sticks colored red and blue, respectively, while the
733 dashed line indicates the $RCOO^-$ to RNH_3^+ distance.
734
735 **Figure 5. Insertion of Calcium ATPase (1SU4) helix A into microsomal and**

736 **cellular membranes. (a)** Lateral view of 1SU4 structure. Zoom view of the A helix
737 (right panel). The membrane-embedded region of helix A is highlighted in yellow.
738 Charged amino acids are shown as sticks in blue (R), red (D) and pink (E),
739 respectively. L49, A69 and F78 are also shown as sticks to define helix's
740 subdomains. The membrane location is indicated by red (outer) and blue (inner)
741 discontinuous lines according to OPM dataset [22] and the distance between the R
742 and D charges is indicated in Å. **(b)** Helix A-derived sequences from 1SU4 including
743 the gap between charged residues, and the predicted (ΔG_{pred}) and experimental (*in*
744 *vitro* $\Delta G_{exp}^{in\ vitro}$ and *in vivo* $\Delta G_{exp}^{in\ vivo}$, respectively) ΔG values in kcal/mol are shown.
745 Amino acids with a positive charge are highlighted in blue (K) while negatively
746 charged are marked in red (D) and pink (E). The residues predicted as TM by the ΔG
747 Prediction server are underlined. Green numbers indicate negative ΔG (insertion)
748 values while red numbers denote ΔG values above 0 (translocation). Representative
749 examples (n=3) of *in vitro* protein translations in the presence of ER-derived
750 microsomes **(c)** and Western blots (n=3) of *in vivo* protein translation in HEK-293T
751 cells **(d)** in the presence (+) or absence (-) of Endoglycosidase H (EndoH), a glycan-
752 removing enzyme. The absence of glycosylation of G1 and G2 acceptor sites is
753 indicated by two white dots, single glycosylation by one white and one black dot,
754 and double glycosylation by two black dots. **(e)** 1SU4 helix A *i*, *i*+5 mutant. The
755 membrane-embedded region of helix A is highlighted in yellow. Charged residues
756 are shown as sticks in blue (R), red (D) and pink (E). L49, A69 and F78 are also
757 shown as sticks to define helix's subdomains. The membrane is indicated by red
758 (outer) and blue (inner) discontinuous lines as in (a), and the dashed line indicates
759 the $RCOO^-$ and $RC(NH_2)_2^+$ distance.

760

761 **Figure 6. Salt bridge effect on the insertion in the absence of preceding TM**
762 **segments. (a)** (Top) Schematic of the C_LTM construct used. It is composed of the
763 domain of the antibody λ light chain (C_L) containing a glycosylation site (G1)
764 connected by flexible linkers to the sequence of analysis followed by a C-terminal
765 glycosylation site (G2) and a *c*-myc tag. (Bottom) Scheme depicting the main
766 features of the C_LTM insertion assay. Black dots represent glycosylated sites while
767 white dots represent non-glycosylated sites. **(b)** Representative halorhodopsin
768 (3QBG) helix G western blot (*n*=3) of *in vivo* protein translations in HEK-293T cells
769 in the presence (+) or absence (-) of Endoglycosidase H (EndoH), a glycan-removing
770 enzyme. **(c)** Representative Ca^{2+} ATPase (1SU4) helix A western blot (*n*=3) of *in vivo*
771 protein translations in HEK-293T cells in the presence (+) or absence (-) of EndoH.
772 Non-glycosylated proteins are indicated by two white dots, singly-glycosylated
773 proteins are indicated by one white and one black dot, and doubly-glycosylated
774 proteins are indicated by two black dots. Experimental ΔG values (kcal/mol) are
775 shown above each sample (*n*=3).

776

777 **Figure S1. Effects on membrane insertion of single or pairs of Asp and Lys**
778 **residues in L5/A14. a)** The tested sequences from L5/A14 model TM (including the
779 charged residues, bold), the gap distance, and the predicted ΔG (ΔG_{pred}) and
780 experimental (ΔG_{exp}) values in kcal/mol are shown. Amino acids with positive and
781 negative charge are highlighted in blue (K) and red (D) respectively. **b)** Experimental
782 ΔG (ΔG_{exp}) in kcal/mol of each tested sequence in the Lep-based microsomal assay.
783 The mean and standard deviation of 3 independent experiments are represented.

784 The individual value of each experiment is represented by a solid dot, *p*-values are
785 indicated above. In addition, a green dot represents the ΔG_{pred} value for the L5/A14
786 sequence. The wt and single mutants are shown in white bars. Charges at
787 compatible distances with salt bridge formation (*i*, *i*+3; and *i*, *i*+4) are shown in
788 orange and yellow, respectively. Not compatible distances with salt bridge
789 formation (*i*, *i*+5) is shown in gray. The inset shows a representative SDS-PAGE gel
790 for L4/A15 and L5/A14 constructs. The construct was expressed in rabbit
791 reticulocyte lysed in the presence (+RM) or absence (–RM) of column washed rough
792 microsomes. Bands of non-glycosylated proteins are indicated by a white dot;
793 mono and double glycosylated proteins are indicated by one and two black dots,
794 respectively.
795
796 **Figure S2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8'**
797 **for the GLOB-MSA dataset.** Log odds ratios for the middle core residues of α-
798 helices of at least 17 residues in length in the GLOB-MSA dataset. The rows on the
799 y-axis indicate the first amino acid in the pair and the columns on the x-axis the
800 second. The residues are ordered as in Fig. 2.
801
802 **Figure S3. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8'**
803 **for the GLOB-MSA dataset.** Log odds ratios as in S2 but all pairs with P-
804 value >0.05 have been masked.
805
806 **Figure S4. Masked log odds ratios of each pair of amino acids for '*i*, *i*+1'**
807 **through '*i*, *i*+8' for the TM-MSA dataset.** Log odds ratios as in Fig. 2 but all pairs

808 with P-value >0.05 have been masked.

809

810 **Figure S5.** Histogram of ΔG_{pred} values for the salt bridge containing TM segments.
811 ΔG values represented with a bin size of 0.8 kcal/mol, with negative (indicative of
812 insertion) and positive (indicative of non-inserted) values highlighted with a green
813 and red background, respectively. Halorhodopsin 3QBG helix G (ΔG_{pred} of 1.73
814 kcal/mol) and Ca2+ ATPase helix A (ΔG_{pred} of 4.15 kcal/mol) are part of the left and
815 right blue highlighted bars, respectively.

816

817 **Figure S6. Salt bridges at pair *i*, *i*+5.** Left: helix E from bacterial translocon (PDB
818 ID: 5MG3). Glycine (orange) residue causes a kink facilitating salt bridge interaction
819 between Asp50 (red) and Arg55 (blue). Right: helix A from a lipid flippase (PDB ID:
820 6CC4). Glycine (orange) residue causes a kink facilitating salt bridge interaction
821 between Arg153 (blue) and Glu158 (red). Distances are shown in Ångström.

822

823 **REFERENCES**

824 Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2013).
825 SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res.
826 *42*, D310–D314.

827 Andreeva, A., Kulesha, E., Gough, J., and Murzin, A.G. (2019). The SCOP
828 database in 2020: expanded classification of representative family and
829 superfamily domains of known protein structures. Nucleic Acids Res. *48*, D376–
830 D382.

831 Armstrong, K.M., and Baldwin, R.L. (1993). Charged histidine affects alpha-helix
832 stability at all positions in the helix by interacting with the backbone charges.
833 Proc. Natl. Acad. Sci. U.S.a. *90*, 11337–11340.

834 Baeza-Delgado, C., Heijne, von, G., Marti-Renom, M.A., and Mingarro, I. (2016).
835 Biological insertion ofcomputationally designed shorttransmembrane segments.
836 Sci. Rep. 1–9.

Baeza-Delgado, C., Marti-Renom, M.A., and Mingarro, I. (2012). Structure-based statistical analysis of transmembrane helices. Eur Biophys J *42*, 199–207.

Bañó-Polo, M., Baeza-Delgado, C., Orzáez, M., Marti-Renom, M.A., Abad, C., and Mingarro, I. (2012). Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing. PLoS ONE *7*, e44263.

Bañó-Polo, M., Baeza-Delgado, C., Tamborero, S., Hazel, A., Grau, B., Nilsson, I., Whitley, P., Gumbart, J.C., Heijne, von, G., and Mingarro, I. (2018). Transmembrane but not soluble helices fold inside the ribosome tunnel. Nat Commun *9*, 5246.

Bañó-Polo, M., Martínez-Gil, L., Barrera, F.N., and Mingarro, I. (2019). Insertion of Bacteriorhodopsin Helix C Variants into Biological Membranes. ACS Omega *5*, 556–560.

Bañó-Polo, M., Martínez-Gil, L., Wallner, B., Nieva, J.L., Elofsson, A., and Mingarro, I. (2013). Charge Pair Interactions in Transmembrane Helices and Turn Propensity of the Connecting Sequence Promote Helical Hairpin Insertion. J Mol Biol *425*, 830–840.

Bosshard, H.R., Marti, D.N., and Jelesarov, I. (2004). Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. J. Mol. Recognit. *17*, 1–16.

Braunger, K., Pfeffer, S., Shrimal, S., Gilmore, R., Berninghausen, O., Mandon, E.C., Becker, T., Förster, F., and Beckmann, R. (2018). Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. Science *360*, 215–219.

Chin, C.N., and Heijne, von, G. (2000). Charge pair interactions in a model transmembrane helix in the ER membrane. J Mol Biol *303*, 1–5.

Chitwood, P.J., and Hegde, R.S. (2020). An intramembrane chaperone complex facilitates membrane protein biogenesis. Nature *584*, 1–22.

Donald, J.E., Kulp, D.W., and DeGrado, W.F. (2011). Salt bridges: Geometrically specific, designable interactions. Proteins *79*, 898–915.

Duart, G., García-Murria, M.J., Grau, B., Acosta-Cáceres, J.M., Martínez-Gil, L., and Mingarro, I. (2020). SARS-CoV-2 envelope protein topology in eukaryotic membranes. Open Biol. *10*, 200209–6.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

Engelman, D.M., Steitz, T.A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem *15*, 321–353.

Feige, M.J., and Hendershot, L.M. (2013). Quality Control of Integral Membrane Proteins by Assembly-Dependent Membrane Integration. Mol Cell *51*, 297–309.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Giménez-Andrés, M., Čopič, A., and Antonny, B. (2018). The Many Faces of Amphipathic Helices. Biomolecules *8*, 45–14.

Hedin, L.E., Ojemalm, K., Bernsel, A., Hennerdal, A., Illergård, K., Enquist, K., Kauko, A., Cristobal, S., Heijne, von, G., Lerch-Bader, M., et al. (2010). Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. J Mol Biol *396*, 221–229.

Heijne, von, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. Nature *341*, 456–458.

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and Heijne, von, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature *433*, 377–381.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H., and Heijne, von, G. (2007). Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature *450*, 1026–1030.

Illergård, K., Kauko, A., and Elofsson, A. (2011). Why are polar residues within the membrane core evolutionary conserved? Proteins *79*, 79–91.

Jayasinghe, S., Hristova, K., and White, S.H. (2001). Energetics, stability, and prediction of transmembrane helices. J Mol Biol *312*, 927–934.

Kanada, S., Takeguchi, Y., Murakami, M., Ihara, K., and Kouyama, T. (2011). Crystal Structures of an O-Like Blue Form and an Anion-Free Yellow Form of pharaonis Halorhodopsin. J Mol Biol *413*, 162–176.

Kauko, A., Hedin, L.E., Thebaud, E., Cristobal, S., Elofsson, A., and Heijne, von, G. (2010). Repositioning of transmembrane alpha-helices during membrane protein folding. J Mol Biol *397*, 190–201.

Kauko, A., Illergård, K., and Elofsson, A. (2008). Coils in the Membrane Core Are Conserved and Functionally Important. J Mol Biol *380*, 170–180.

Kolbe, M., Besir, H., Essen, L.O., and Oesterhelt, D. (2000). Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. Science *288*, 1390–1396.

Kozma, D., Simon, I., and Tusnády, G.E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res. *41*, D524–D529.

Kumar, S., Ma, B., Tsai, C.J., and Nussinov, R. (2000). Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. Proteins *38*, 368–383.

Kumar, S., and Nussinov, R. (2002). Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. Biophys J *83*, 1595–1612.

Lerch-Bader, M., Lundin, C., Kim, H., Nilsson, I., and Heijne, von, G. (2008). Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. *105*, 4127–4132.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.

Lin, C.-Y., and Lin, L.-Y. (2018). The conserved basic residues and the charged amino acid residues at the α-helix of the zinc finger motif regulate the nuclear transport activity of triple C2H2 zinc finger proteins. PLoS ONE *13*, e0191971–20.

Marqusee, S., and Baldwin, R.L. (1987). Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design. Proc. Natl. Acad. Sci. U.S.a. *84*, 8898–8902.

Martínez-Gil, L., Saurí, A., Marti-Renom, M.A., and Mingarro, I. (2011). Membrane protein integration into the endoplasmic reticulum. Febs J *278*, 3846–3858.

Mbaye, M.N., Hou, Q., Basu, S., Teheux, F., Pucci, F., and Rooman, M. (2019). A comprehensive computational study of amino acid interactions in membrane proteins. Sci. Rep. *9*, 12043–14.

Okamura, Y., Fujiwara, Y., and Sakata, S. (2015). Gating Mechanisms of Voltage-Gated Proton Channels. Annu Rev Biochem *84*, 685–709.

Samanta, U., Pal, D., and Chakrabarti, P. (1999). Packing of aromatic rings against tryptophan residues in proteins. Acta Cryst (1999). D55, 1421-1427 [Doi:10.1107/S090744499900726X] 1–7.

Shurtleff, M.J., Itzhak, D.N., Hussmann, J.A., Schirle Oakdale, N.T., Costa, E.A., Jonikas, M., Weibezahn, J., Popova, K.D., Jan, C.H., Sinitcyn, P., et al. (2018). The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. eLife *7*, 382.

Tamborero, S., Vilar, M., Martínez-Gil, L., Johnson, A.E., and Mingarro, I. (2011). Membrane insertion and topology of the translocating chain-associating membrane protein (TRAM). J Mol Biol *406*, 571–582.

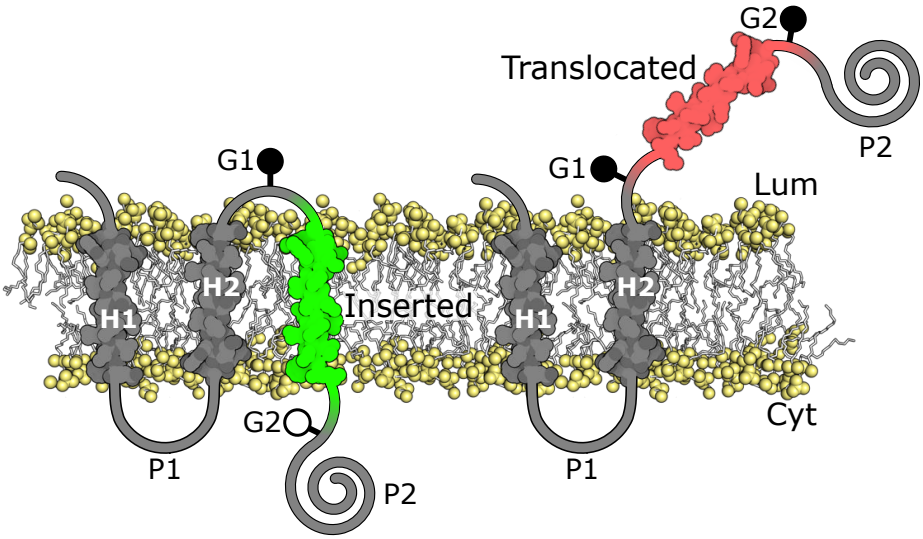
Toyoshima, C., Nakasako, M., Nomura, H., and Ogawa, H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. Nature *405*, 647–655.

Tsirigos, K.D., Govindarajan, S., Bassot, C., Västermark, Å., Lamb, J., Shu, N., and Elofsson, A. (2018). ScienceDirect Topology of membrane proteins — predictions, limitations and variations. Curr Opin Struct Biol *50*, 9–17.

Tsirigos, K.D., Peters, C., Shu, N., Käll, L., and Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. *43*, W401–W407.

Walther, T.H., and Ulrich, A.S. (2014). Transmembrane helix assembly and the role of salt bridges. Curr Opin Struct Biol *27*, 63–68.

a)



b)

AA	Gap	Sequence	$\Delta G_{pred.}$	$\Delta G_{exp.}$
L4/A15	-	AAAALALAAAAALALAAAA (8, 12)	-0.49	-0.40 ± 0.11
K	-	AAAALALKAAAALALAAAA	+1.41	+0.15 ± 0.07
D	-	AAAALALAAAADLALAAAA	+1.39	+0.14 ± 0.07
KD	1	AAAALALAAAKDLALAAAA	+3.22	+0.18 ± 0.04
KD	2	AAAALALAAKADLALAAAA	+2.93	+0.27 ± 0.05
KD	3	AAAALALAKAADLALAAAA	+3.35	+0.22 ± 0.03
KD	4	AAAALALKAAADLALAAAA	+3.34	+0.13 ± 0.05
KD	5	AAAALAKLAAADLALAAAA	+2.98	+0.31 ± 0.02
DD	4	AAAALALDAAADLALAAAA	+3.55	+0.23 ± 0.01

c)

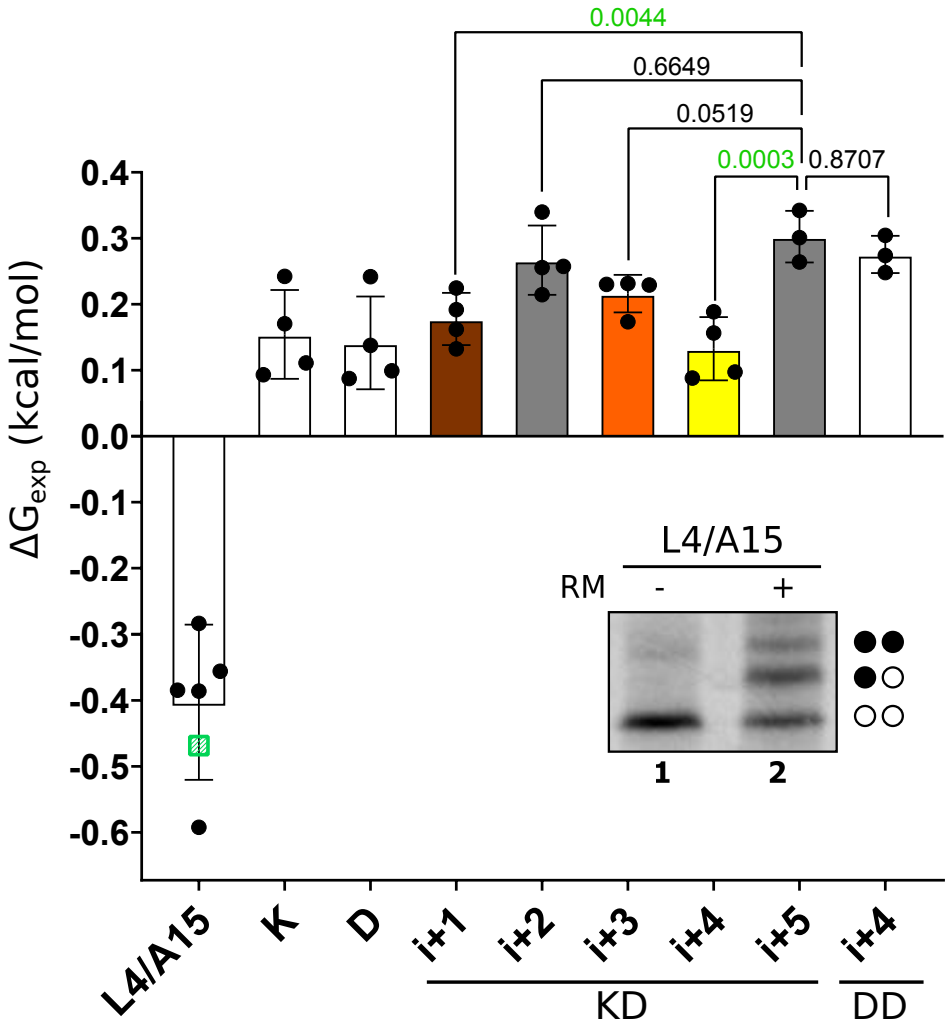


Figure 2

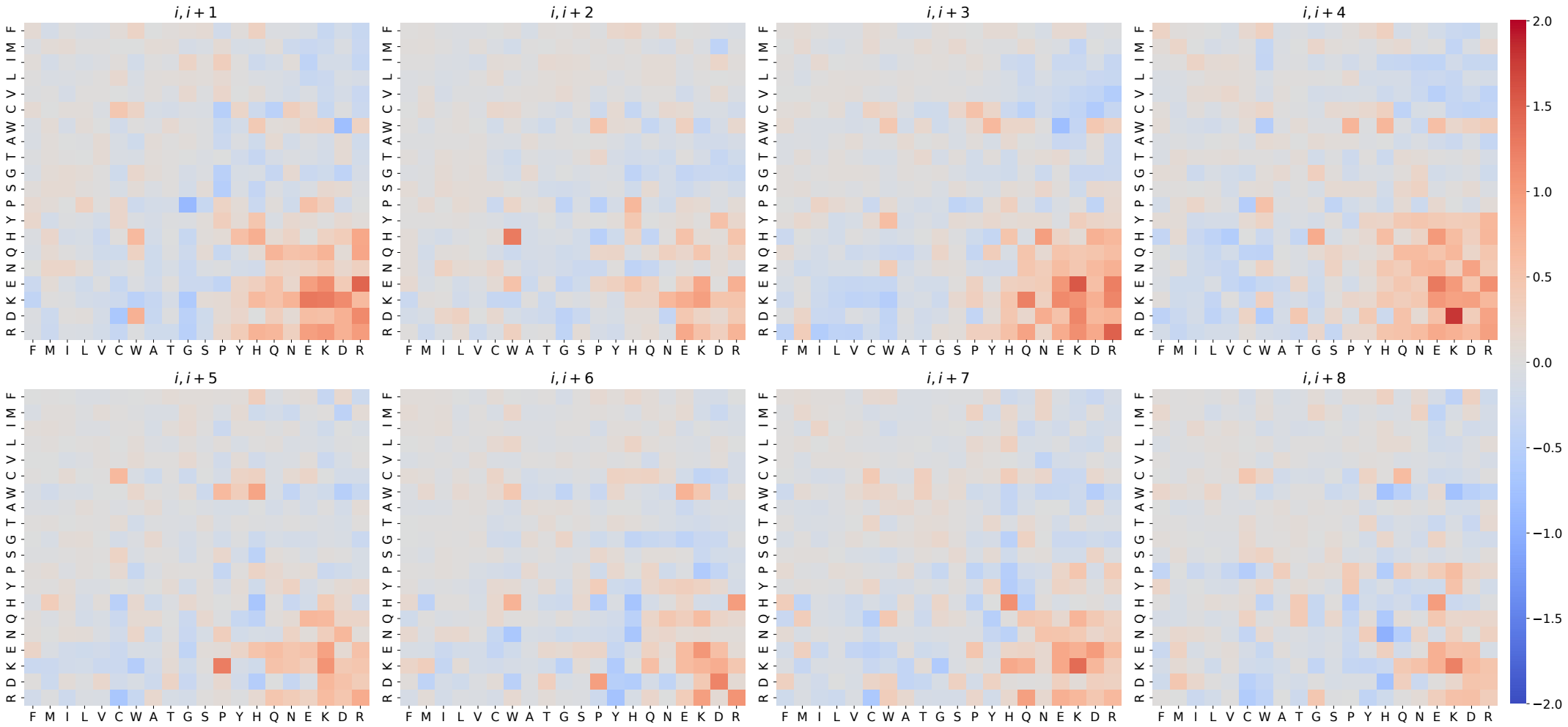

$i, i+1$
$i, i+2$
$i, i+3$
$i, i+4$
$i, i+5$
$i, i+6$
$i, i+7$
$i, i+8$
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure 3

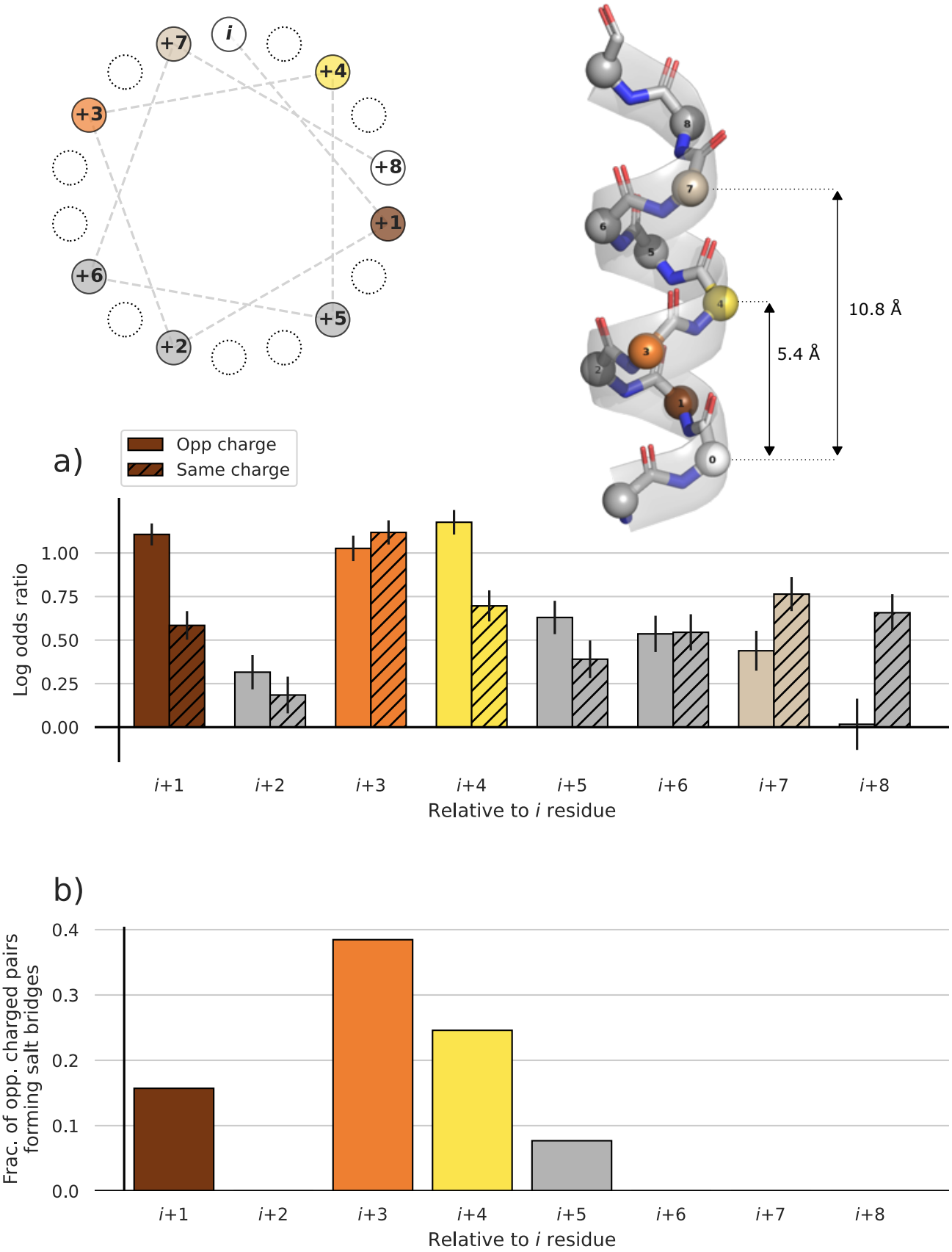
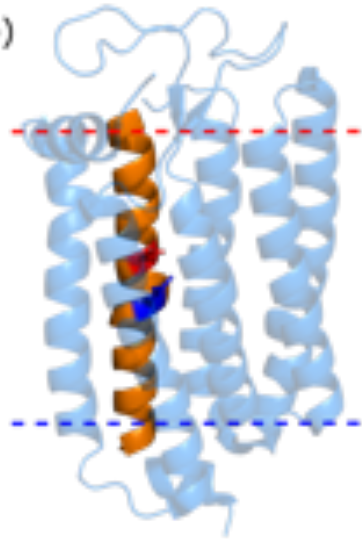

i
+1
+2
+3
+4
+5
+6
+7
+8
10.8 Å
5.4 Å
a)
Opp charge
Same charge
Log odds ratio
1.00
0.75
0.50
0.25
0.00
i+1
i+2
i+3
i+4
i+5
i+6
i+7
i+8
Relative to i residue
b)
Frac. of opp. charged pairs forming salt bridges
0.4
0.3
0.2
0.1
0.0
i+1
i+2
i+3
i+4
i+5
i+6
i+7
i+8
Relative to i residue

Figure 4

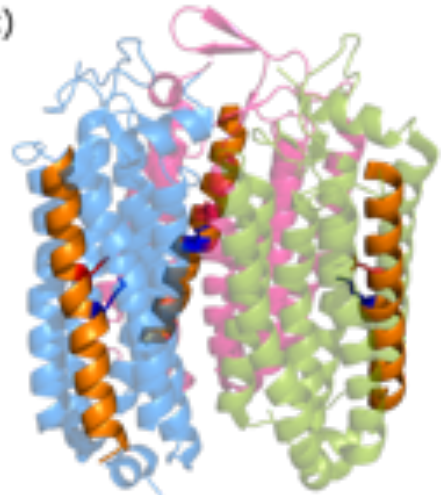
a)

Name	Sequence	Gap	$\Delta G_{pred.}$	$\Delta G^{in\ vitro}_{exp.}$	$\Delta G^{in\ vivo}_{exp.}$
wt	243 254 258 271 VGVTSWGYSFLDIVAKYIFAFLLLNYLTS	4	+1.73	-0.24	-1.54
DD	VGVTSWGYSFLDIVADYIFAFLLLNYLTS	4	+1.66	+0.11	-
DA	VGVTSWGYSFLDIVAAYIFAFLLLNYLTS	4	-0.38	-0.88	-
DK *i*, *i*+5	VGVTSWGYSFLDIVAYKIFAFLLLNYLTS	5	+1.25	+0.31	+1.25

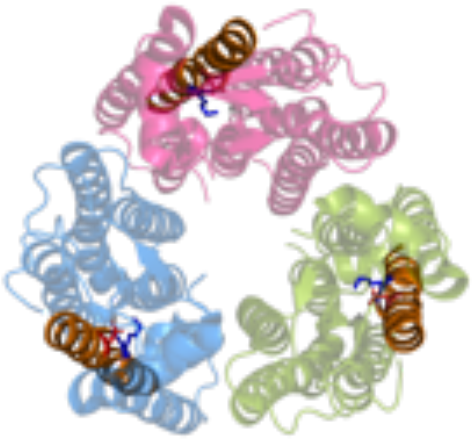
b)



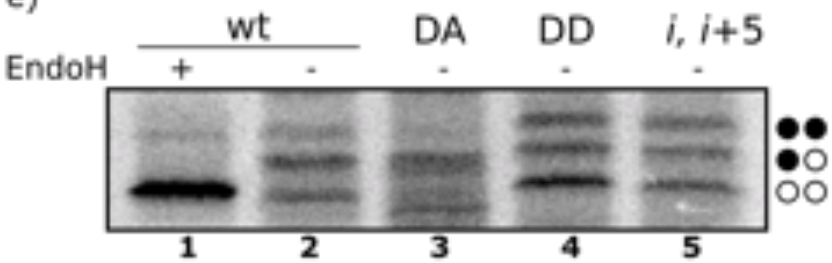
c)



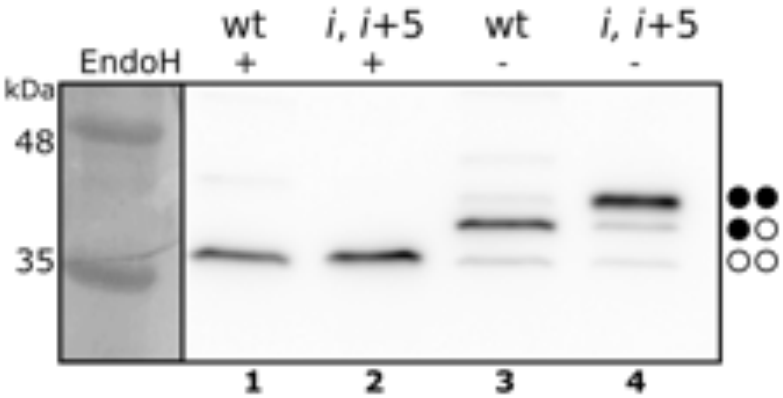
d)



e)



f)



g)

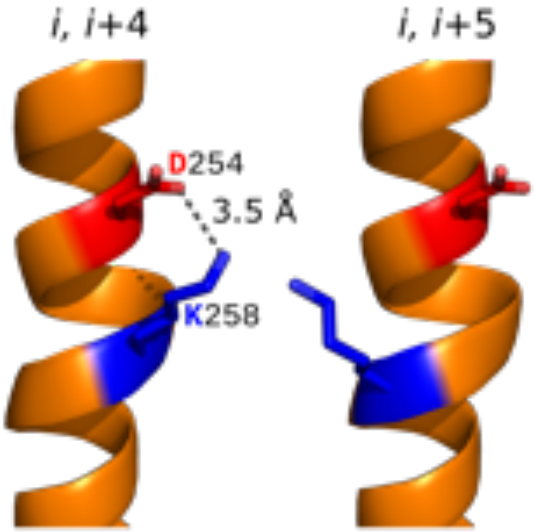
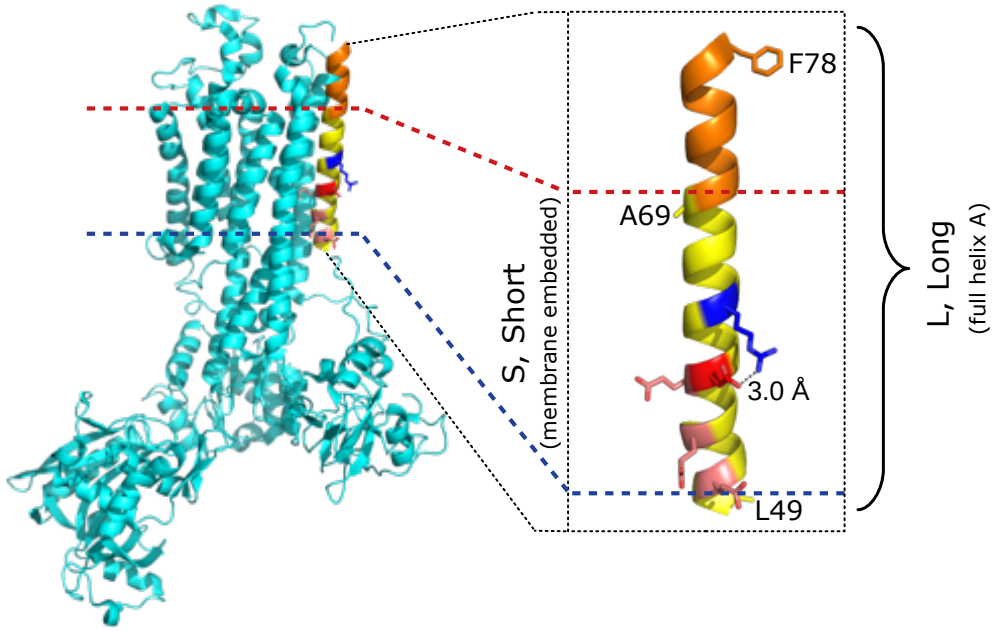


Figure 5

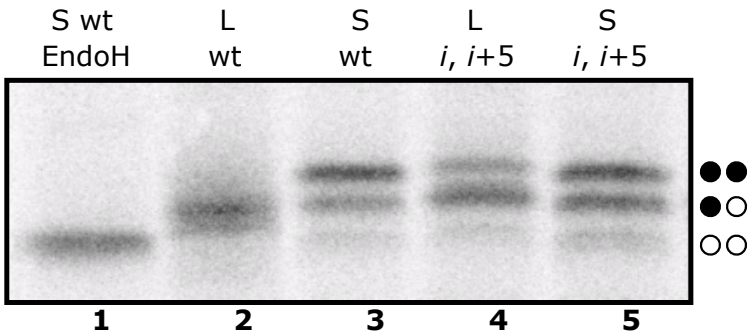
a)



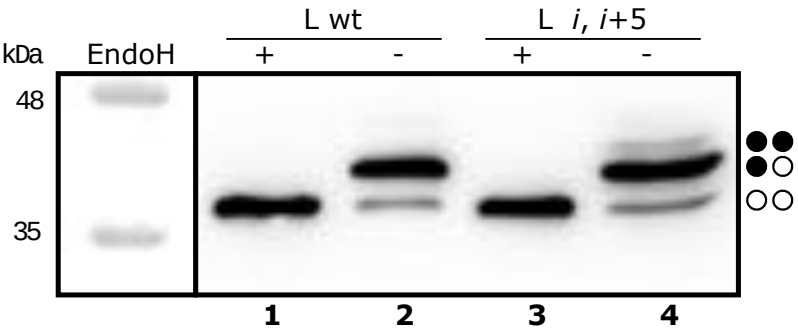
b)

Name	Sequence	Gap	$\Delta G_{pred.}$	$\Delta G^{in\ vitro}_{exp.}$	$\Delta G^{in\ vivo}_{exp.}$
L (L49-F78)	49 59 63 78 LWELVIEQFEDLLVRILLLAACISFVLAWF	4	+1.06	-1.05	-2.06
L *i*, *i*+5 (L49-F78)	LWELVIEQFEDLLVIRLLLAACISFVLAWF	5	+0.44	-0.34	-1.05
S (L49-A69)	LWELVIEQFEDLLVRILLLAA	4	+4.12	+0.22	-
S *i*, *i*+5 (L49-A69)	LWELVIEQFEDLLVIRLLLAA	5	+3.46	+0.06	-

c)



d)



e)

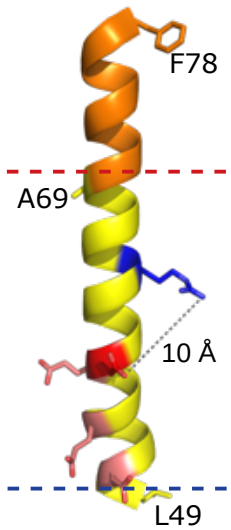
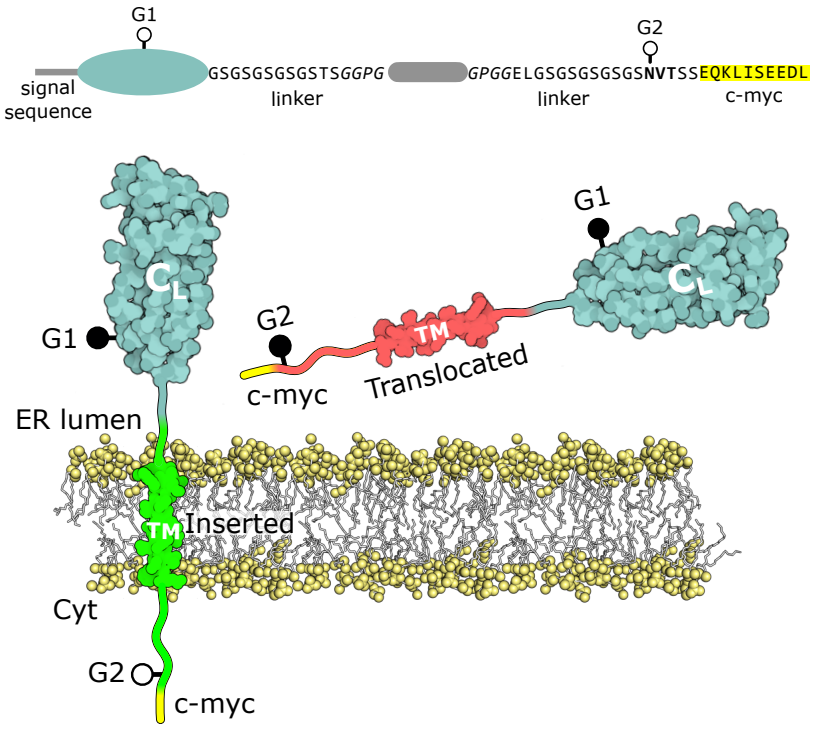
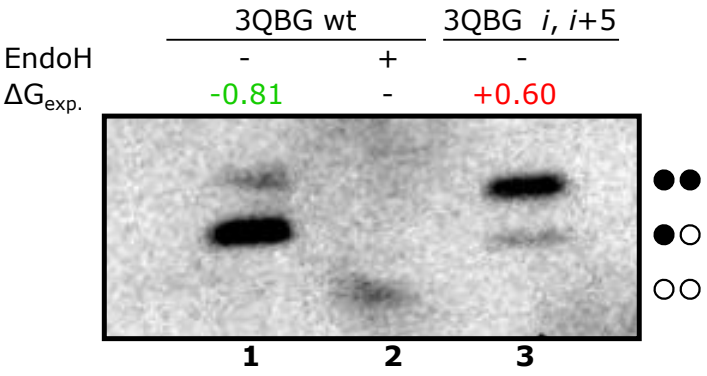


Figure 6

a)


G1
signal sequence
GSGSGSGSGSTSGGPG
linker
GPGGELGSGSGSGSGSNVTSSEQKLISEEDL
linker
G2
c-myc
G1
CL
G2
TM
Translocated
c-myc
G1
CL
ER lumen
TM
Inserted
Cyt
G2
c-myc

b)


3QBG wt
3QBG i, i+5
EndoH
-
+
-
ΔGexp.
-0.81
-
+0.60
1
2
3

c)

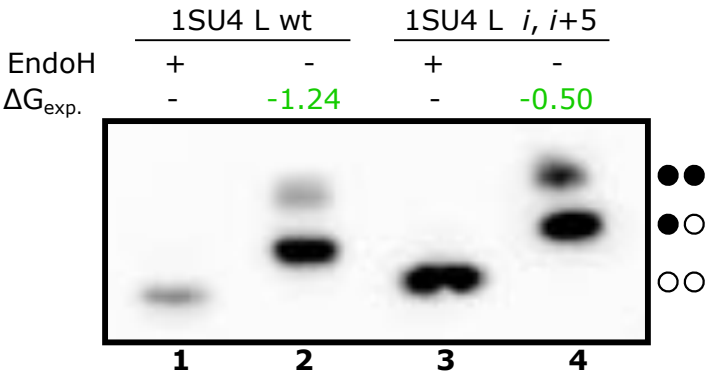
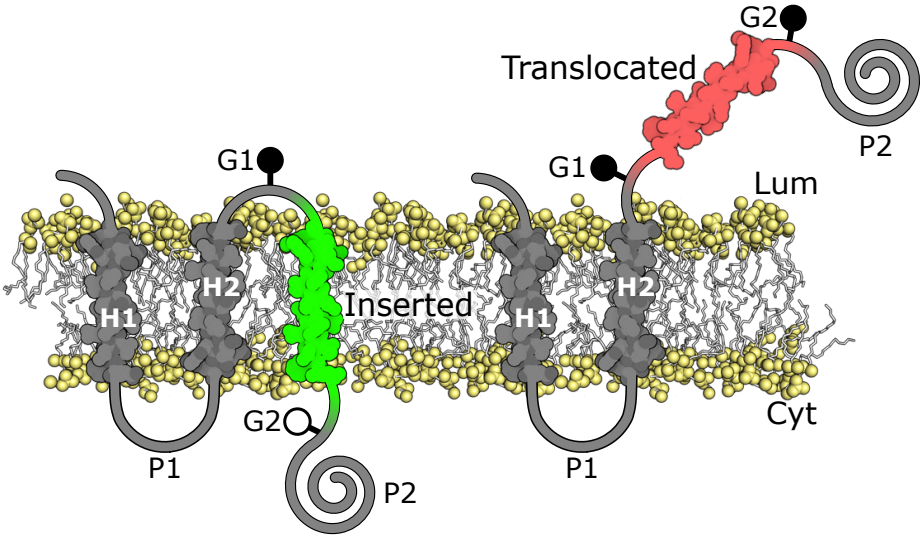

1SU4 L wt
1SU4 L i, i+5
EndoH
+
-
+
-
ΔGexp.
-
-1.24
-
-0.50
1
2
3
4

Figure S1

a)



a)

AA	Gap	Sequence	$\Delta G_{pred.}$	$\Delta G_{exp.}$
L5/A14	-	AAAALALAALAALALAAAA (8, 12)	-1.00	-1.17 ± 0.12
K	-	AAAALALKALAALALAAAA	1.00	0.31 ± 0.08
D	-	AAAALALAALADLALAAAA	0.98	0.57 ± 0.08
KD	3	AAAALALAKLADLALAAAA	2.97	0.69 ± 0.03
KD	4	AAAALALKAAADLALAAAA	2.93	0.58 ± 0.05
KD	5	AAAALAKLAAADLALAAAA	2.14	0.88 ± 0.04

b)

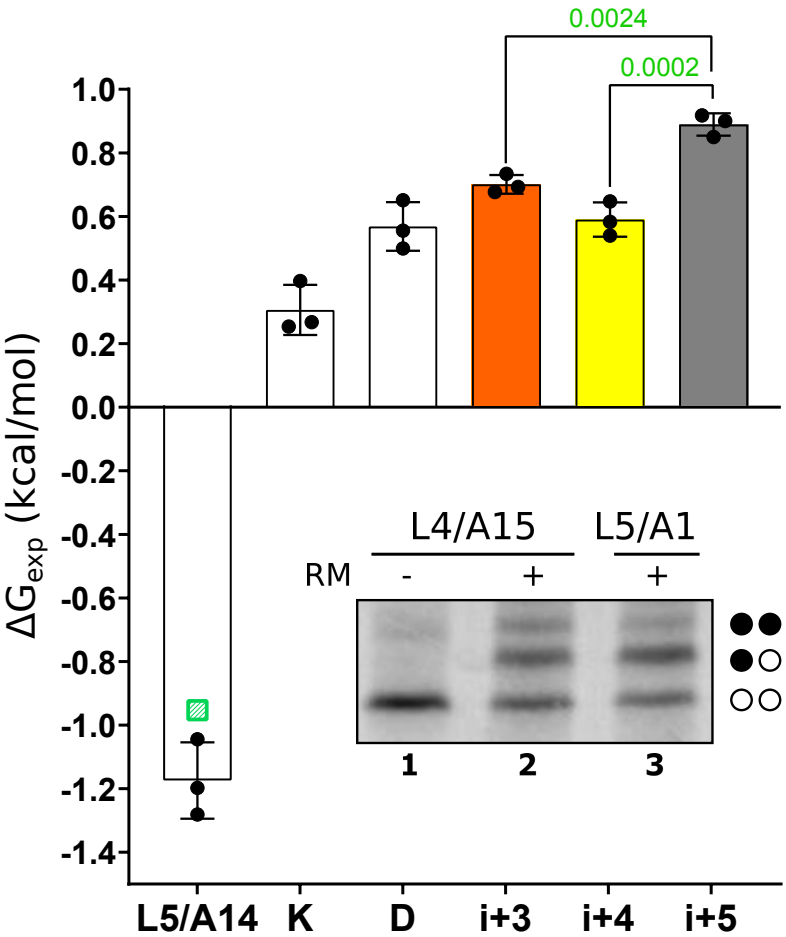


Figure S2

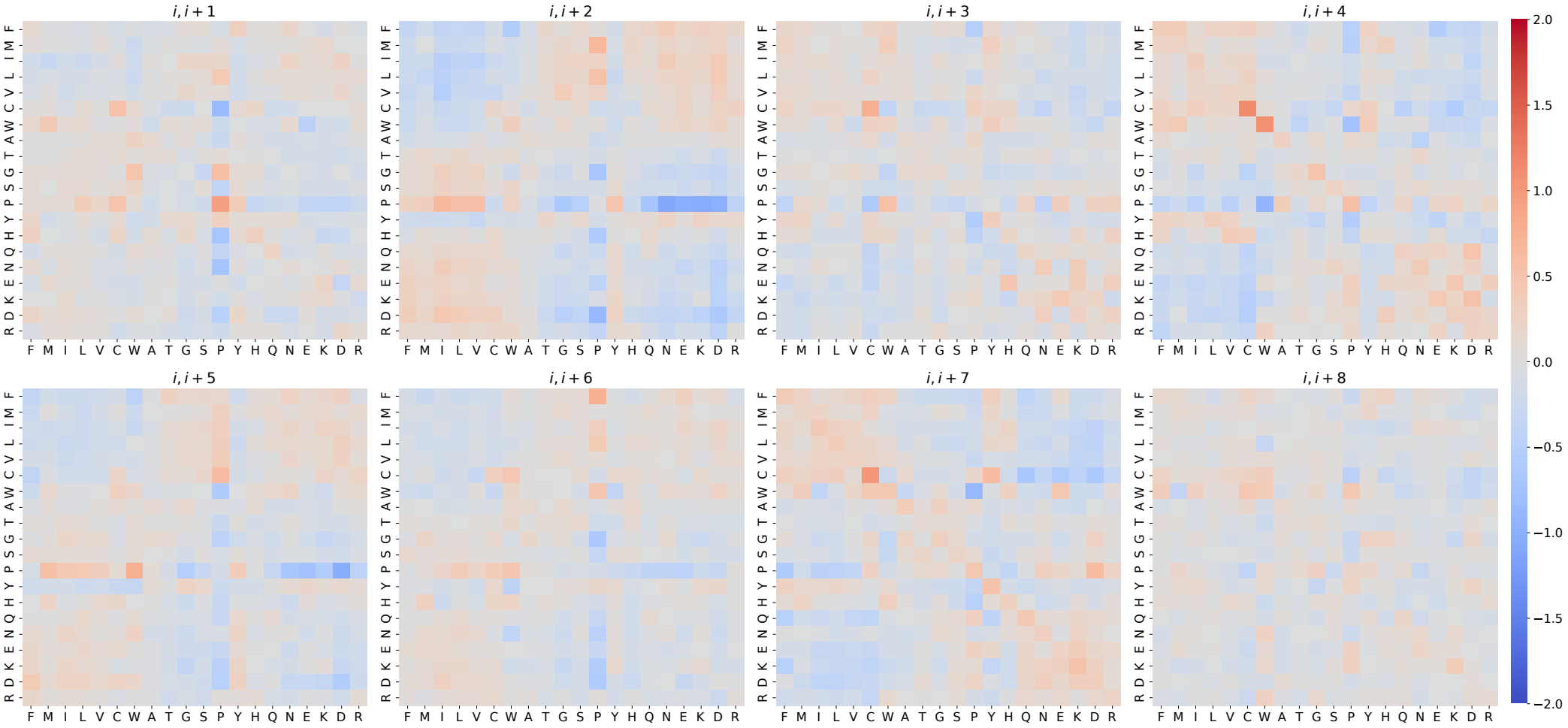

$i, i+1$
$i, i+2$
$i, i+3$
$i, i+4$
$i, i+5$
$i, i+6$
$i, i+7$
$i, i+8$
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure S3

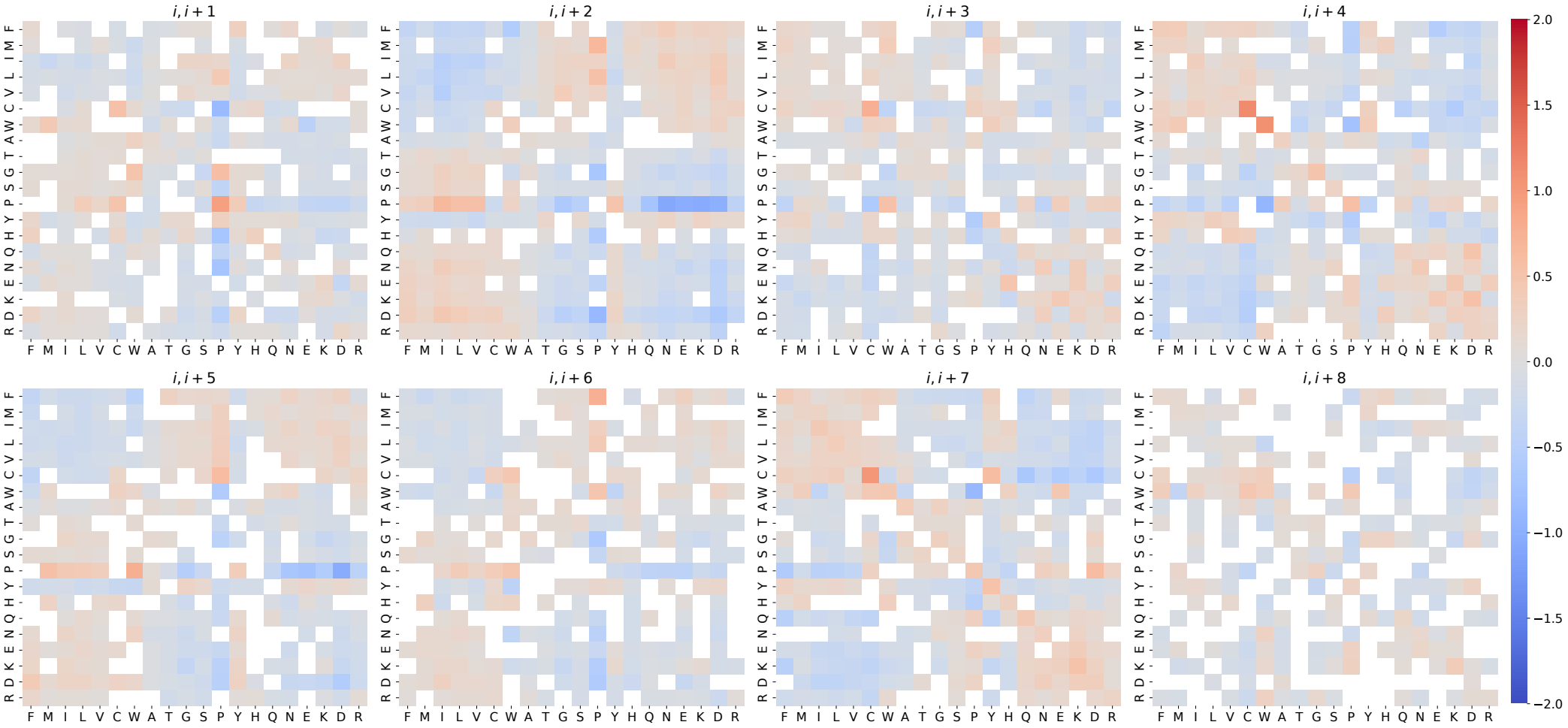

$i, i+1$
$i, i+2$
$i, i+3$
$i, i+4$
$i, i+5$
$i, i+6$
$i, i+7$
$i, i+8$
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure S4

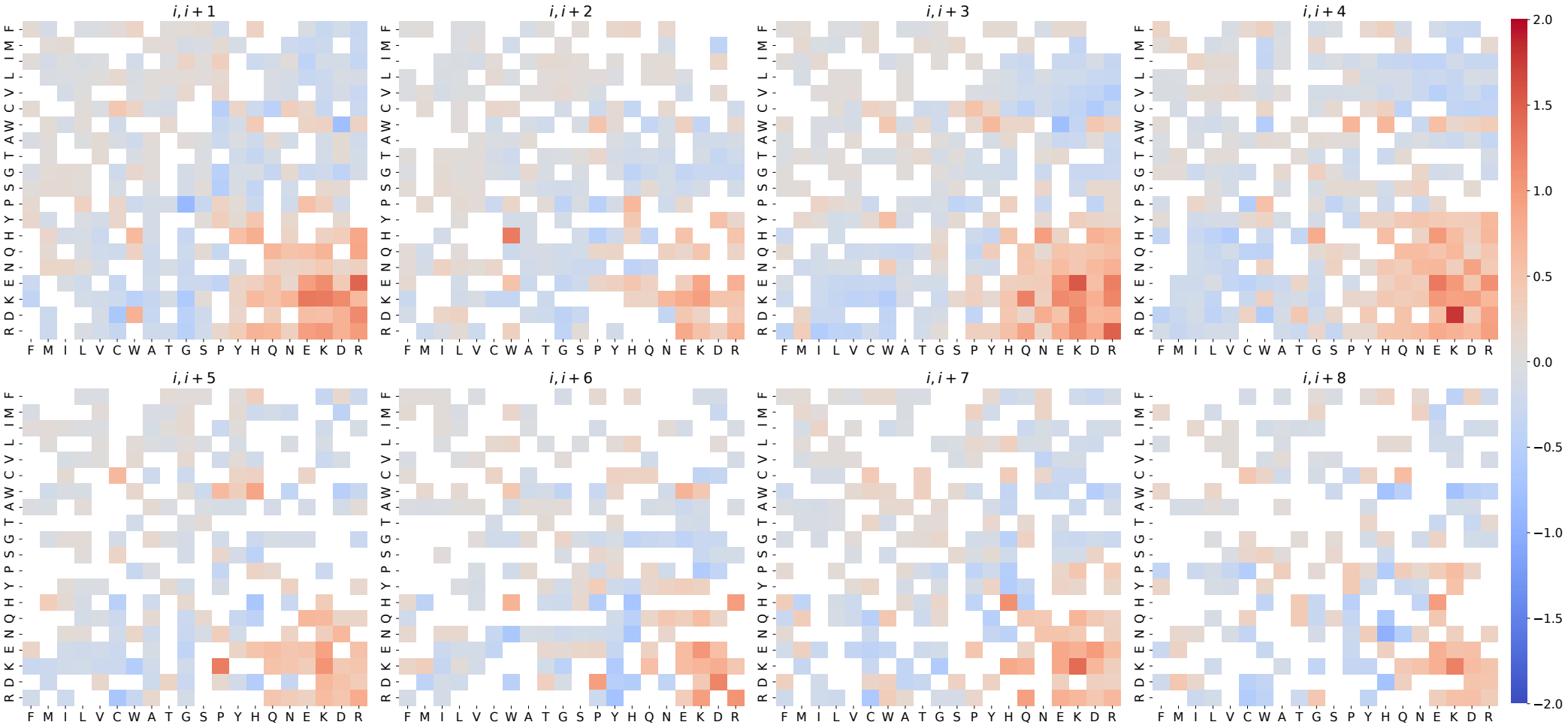
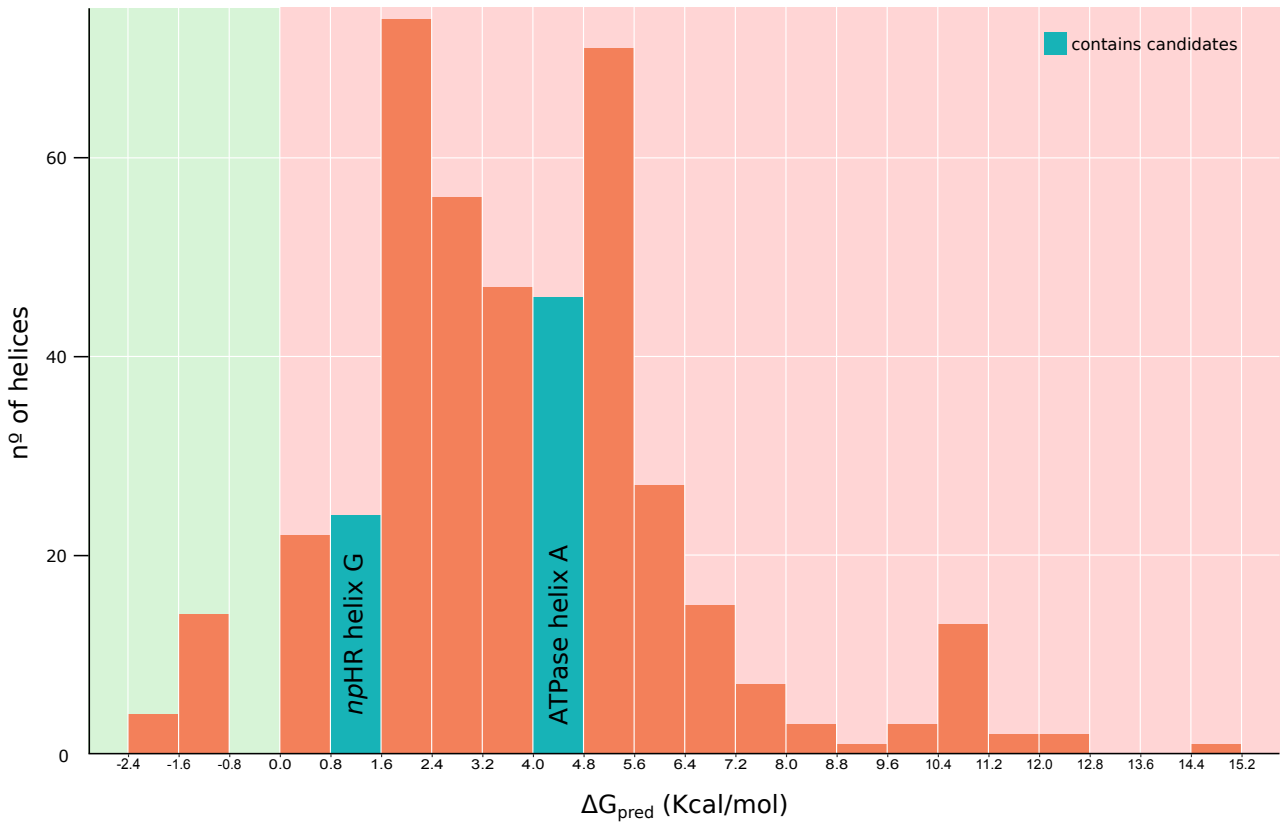

$i, i+1$
$i, i+2$
$i, i+3$
$i, i+4$
$i, i+5$
$i, i+6$
$i, i+7$
$i, i+8$
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure S5


contains candidates
nº of helices
0
20
40
60
npHR helix G
ATPase helix A
-2.4
-1.6
-0.8
0.0
0.8
1.6
2.4
3.2
4.0
4.8
5.6
6.4
7.2
8.0
8.8
9.6
10.4
11.2
12.0
12.8
13.6
14.4
15.2
ΔG_pred (Kcal/mol)

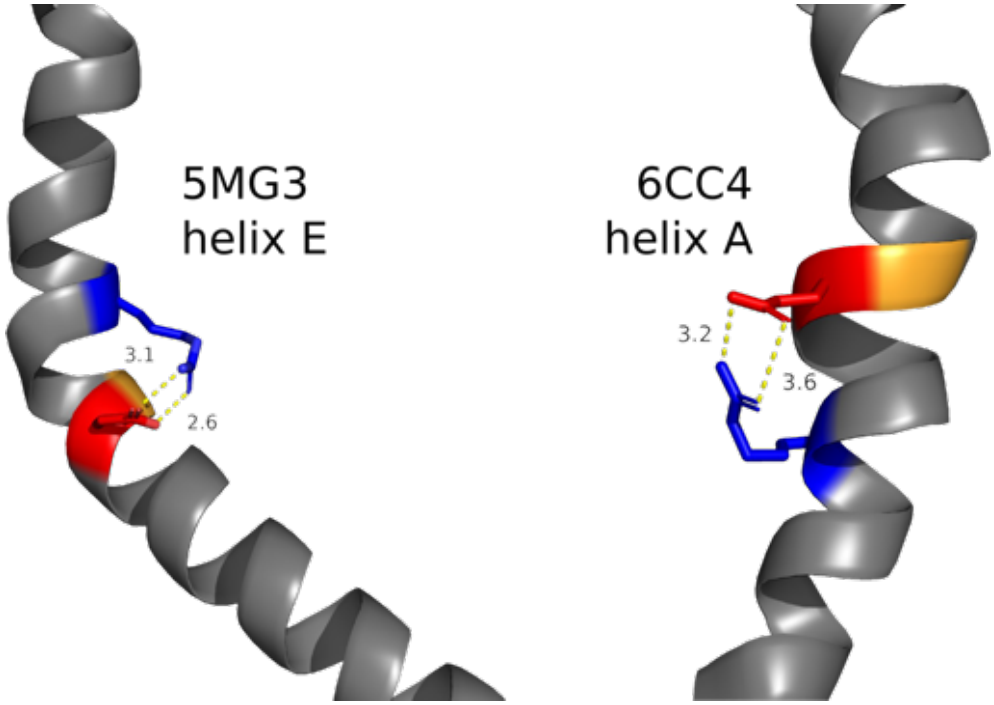



Figure S6