May 24th, 2021

Intra-helical salt bridge contribution to membrane protein insertion

Gerard Duart [1 #], John Lamb [2 #], 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 (alphabetical order)

[*] 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 [1]. 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 [2]. This fact is sometimes not contemplated in modern membrane topology prediction tools [3,4], 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 [5], 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 [6]. Many studies have shown that pairs of charged residues that form potential salt-bridges stabilize soluble α-helices [7]. 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 [8]. Apart from that, it has been suggested that potential salt bridges could help in the insertion of TM α-helices [5,9], 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 [10]. 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.

90 **RESULTS**

91 **Charge pair interactions in a model transmembrane helix.**

92 To test the contribution of potential salt bridges to the translocon-mediated
93 membrane insertion, we used the vehicle protein leader peptidase (Lep) from
94 *Escherichia coli* (Fig.1a). The Lep protein consists of two TM segments (H1 and H2)
95 connected by a cytoplasmic loop (P1) and a large C-terminal (P2) domain, which
96 inserts into endoplasmic reticulum (ER)-derived rough microsomes with both
97 termini located in the microsomes lumen. The designed TM segments were inserted
98 into the luminal P2 domain and flanked by two acceptor sites (G1 and G2) for N-
99 linked glycosylation. Glycosylation occurs exclusively in the lumen of the ER (or
100 microsomes) because of the location of the oligosaccharyltransferase (OST) active
101 site (a translocon-associated enzyme responsible for the oligosaccharide transfer)
102 [11]. In this case, the engineered glycosylation sites can be used as membrane
103 insertion reporters because G1 will always be glycosylated due to its native luminal
104 localization, but G2 will be glycosylated only upon translocation of the analyzed
105 sequence across the microsomal membrane. A singly glycosylated construct in
106 which a tested sequence is inserted into the membrane has a molecular mass ~2.5
107 kDa higher than the molecular mass of Lep molecule expressed in the absence of
108 microsomes; the molecular mass shifts by ~5 kDa upon double glycosylation which
109 facilitates its identification by gel electrophoresis when expressed in the presence
110 of [^{35}S-labeled] amino acids. Then, *in vitro* transcription/translation of these chimeric
111 proteins in the presence of rough microsomal membranes (RMs) allows for accurate
112 and quantitative description of membrane insertion of designed sequences [12-15].
113 The degree of membrane insertion is quantified by analyzing the fractions of singly

114 glycosylated (i.e., membrane inserted) and doubly glycosylated (i.e., non-inserted)
115 molecules, which can be expressed as an experimental apparent free energy of
116 membrane insertion, ΔG_{exp} (see Materials and Methods) [12].

117 We first compared the effects of oppositely charged Lys and Asp residues
118 on the insertion of a 19-residue-long hydrophobic stretch (L4/A15 scaffold, 4
119 leucines and 15 alanines), which was designed to insert stably into the RM
120 membranes [12], including charges centered in the TM segment at different
121 positions (Fig. 1b) and "insulated" from the surrounding sequence by N- and C-
122 terminal GGPG- and -GPGG tetrapeptides. Single Lys and Asp residues were
123 placed in positions 8 and 12 respectively, and pairs of Lys-Asp residues were
124 designed to cover positions 7-12 (that is, more than one helical turn). When pairs of
125 charged residues are present, our results showed a tendency to insert better when
126 pair charges were placed in positions (*i*, *i*+1; *i*, *i*+3; *i*, *i*+4) that are permissive with
127 salt bridge formation (Fig. 1c), actually an effect not observed in the predictions (Fig.
128 1b). Being a thermodynamically inconvenient phenomenon within the membrane
129 environment, it is expected that the sequence context and the amino acid
130 composition of the TM helix would be determinant for salt bridge formation.
131 Accordingly, we scrutinized a large dataset of membrane proteins of known three-
132 dimensional structures in order to focus on natural salt bridges present within TM
133 segments.

134

135 **Charged pairs in transmembrane helices**

136 Alpha helices are a common secondary structure in both globular and membrane
137 proteins. We created two main datasets, TM dataset with helical membrane

138 proteins of known structure from the PDBTM-dataset [16], and GLOB dataset with
139 globular alpha helical proteins selected from the SCOP-database [17,18], see
140 methods for the full creation steps. Table 1 shows a breakdown of alpha-helices,
141 charges and salt bridges in the two datasets. What is clear is that long alpha helices
142 (≥ 17 residues) form a larger proportion in the TM dataset than in the globular helices
143 (GLOB) dataset, which is logical as most TM helices need to span through the
144 hydrophobic core of the lipid bilayer of thickness ~30 Å.

145 Our TM dataset showed the same distribution of polar charges as previous
146 studies [19], with about 10 % of polar residues in the core membrane regions (see
147 Table 1), and about one-third of these are charged. Over half of the charged
148 residues could form pairs with other charged residues at intervals of *i*, *i*+1; *i*, *i*+3 and
149 *i*, *i*+4. In contrast, the GLOB dataset had a much higher proportion of charged polar
150 residues. The GLOB dataset also contained about 15 times as many charged pairs
151 relative to its size compared to the TM dataset, again indicating that charged
152 residues are more common in globular than in TM helices.

153 The two datasets, TM and GLOB, were extended with homologous
154 sequences identified by searching with jackhammer against UniProt. All sequences
155 with an E-value lower than 10^{-3} were included. These datasets are named TM-MSA
156 and GLOB-MSA. Using data in the TM-MSA dataset to produce the log odds ratios,
157 we identified periodicity patterns of charged residue pairs that are more common
158 than what would have been expected from the underlying amino acid composition
159 (see Figure 2). We found that polar residues at pairs *i*, *i*+3; *i*, *i*+4 and *i*, *i*+7 are
160 significantly enhanced (Fig. 2). This feature is strengthened when we examine the
161 same plot for the GLOB-MSA dataset, where these patterns were not observed

(Figure S1). This was also clear when statistical significance was taken into account, see Figures S2 and S3.

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 [20] at *i*, *i*+6 and contacting with prosthetic groups by His-His pair at *i*, *i*+7 [19].

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 all 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).

Selection of natural salt bridges from membrane protein structures

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

210 When we now make a selection for potential candidates, we also look at the
211 estimated ΔG_{pred} values and choose helices with a positive value above one, as per
212 our hypothesis, a salt bridge helps TM insertion of sequences for which the
213 hydrophobic force would not be enough to insert. This stricter definition results in a
214 set of 426 candidates of a total of 431 salt bridges and we selected TM7 (helix G)
215 from halorhodopsin protein (PDB ID: 3QBG) with an estimated ΔG_{pred} value above
216 +1.7 kcal/mol, and helix A from calcium ATPase (PDB ID: 1SU4) with a higher
217 estimated ΔG_{pred} value (above +4.1 kcal/mol), as candidates for systematic studies
218 to cover a wide range of insertion penalties. See the github repository
219 (https://github.com/ElofssonLab/salt_bridges) for the full lists.
220
221 **Intra-helical salt bridge stabilizes the insertion of helix G from halorhodopsin.**
222 Halorhodopsin (hR) from *Natronomonas pharaonis* (3QBG) is a protein made up of
223 seven TM helices (helix A through G) and a retinal chromophore that is bound via a
224 protonated Schiff base to the ε-amino group of a lysine (K258) residue located
225 roughly in the middle of helix G [21]. *In silico* analysis of 3QBG structure, the anion-
226 free form of the protein, showed a charged pair of amino acids (Asp-Lys), involving
227 the functional K258 and the D254 in the center of helix G (Fig. 4). The position (*i*,
228 *i*+4) and the distance between the anionic carboxylate ($RCOO^-$) from the Asp residue
229 and the cationic ammonium (RNH_3^+) from the lysine residue, in the crystal structure,
230 was about 3.5 Å, a permissive distance for a salt bridge formation (Fig. 4g), which
231 has been established as being lower than 4 Å [22]. To get insights into this
232 interaction, we designed three mutants that were supposed to perturb the salt
233 bridge interaction by different ways: K258D mutant by placing two charged residues

234 with the same polarity at positions *i*, *i*+4; K258A mutant by replacing one of the
235 charged residues by a non-polar amino acid, and K258Y/Y259K double mutant by
236 placing the charged pair at a non-permissive salt bridge position (*i*, *i*+5; Fig. 4g),
237 while keeping the same amino acid composition (Fig. 4a).

238 Halorhodopsin is a trimeric protein in which the helix G is neither exposed at
239 the monomer-monomer interface nor oriented to the inner part of the trimeric
240 structure (Fig. 4, panels b-d). In fact, the charged pair found in helix G is oriented
241 toward the core of the 'globular' structure in each monomer. Then, the insertion of
242 helix G was studied using the Lep-based assay. When constructs harboring helix G
243 wild type sequence were translated *in vitro* in the presence of RMs singly-
244 glycosylated (reporting insertion) forms were found (Fig. 4e, lane 2), despite its
245 positive (suggesting non-insertion) ΔG_{pred} value (Fig. 4a). The nature of the higher
246 molecular weight polypeptide species was analysed by endoglycosidase H (EndoH)
247 treatment, a highly specific enzyme that cleaves N-linked oligosaccharides.
248 Treatment with EndoH of the samples eliminated higher molecular mass bands (Fig.
249 4a, lane 1), confirming the sugar source of the retarded electrophoretic mobility
250 bands and suggesting helix G insertion into the microsomal membrane. However,
251 locating the Asp-Lys pair at *i*, *i*+5, which is non-compatible with salt bridge
252 interaction (Fig. 4g), strongly reduced the experimental insertion efficiency (Fig. 4e,
253 lane 5). Interestingly, replacing the positively charged lysine residue by a negatively
254 charged aspartic acid residue (K258D) rendered similarly low levels of insertion
255 efficiency (Fig. 4e, lane 4). As expected, replacement of the ionizable lysine residue
256 by the aliphatic alanine increased the insertion efficiency (Fig. 4e, lane 3). In this
257 later mutant, the most likely cause is the absence of the positively charged lysine

258 amino side chain and by the presence of the methyl side chain group of the mutant
259 alanine residue. The results of the Lep-based glycosylation assay indicated that wild
260 type (wt) and DA sequences (two stabilized charges or only one charge,
261 respectively) are inserted properly into the microsomal membrane (ΔG_{exp} values -
262 0.24 and -0.88 kcal/mol, respectively), but when the salt bridge is disrupted, either
263 by having two charged amino acids with the same polarity (DD) or by placing
264 oppositely charged residues at a non-permissive distance (*i*, *i*+5) in the center of the
265 helix, the translocation of the segment increases substantially. It should be
266 mentioned that the K258Y/Y259K double mutant has the same amino acid
267 composition than the original helix G, but insertion efficiency is remarkably
268 decreased (ΔG_{exp} = +0.31 kcal/mol). Together these results show that the interaction
269 (salt bridge) between Asp and Lys residues in the center of the helix G from 3QBG
270 is essential for its proper insertion into the microsomal membrane. The salt bridge
271 contributes approximately ~0,5 kcal/mol to the apparent experimental free energy
272 of microsomal membrane insertion, as this is the difference found between the
273 ΔG_{exp} values for the wt and *i*, *i*+5 mutant.
274 Next, to ensure that the *in vitro* results are relevant to the *in vivo* situation, wt
275 and *i*+5 constructs were also expressed *in vivo* in HEK-293T cells. To this end, a *c*-
276 myc tag was engineered at the C-terminus of the Lep chimera to allow immune-
277 detection of our constructs in cell extracts. As shown in Fig. 4f, transfected cells
278 with the chimera containing helix G wt sequence rendered singly-glycosylated
279 molecules, indicating *in vivo* membrane insertion. In contrast, cells transfected with
280 the construct harboring *i*, *i*+5 sequence rendered almost exclusively doubly-
281 glycosylated forms, as proved by EndoH treatment (Fig. 4f, lane 2), suggesting

282 membrane translocation. These results emphasized the relevance of salt bridge
283 interactions in translocon-mediated TM insertion, especially in the *in vivo* (cellular)
284 environment.
285
286 **Salt bridge contribution to the insertion of a heavily charged helix.**
287 In order to challenge salt bridge interactions in a more hydrophilic TM helix we focus
288 on helix A from the sarcoplasmic/ER calcium ATPase 1 [23]. Calcium ATPase (PDB
289 ID: 1SU4, *Oryctolagus cuniculus*) is a member of the P-type ATPases that transport
290 ions across the membrane against a concentration gradient involving 10 α-helices
291 (helix A through J) in the membrane-embedded region [23]. *In silico* analysis of 1SU4
292 structure, the crystal structure of the calcium ATPase with two bound calcium ions,
293 showed Asp-Arg (DR) pair (D59 and R63) in the center of the helix A. This helix
294 extends beyond the membrane (Fig. 5a, from Leu49 to Phe78) and shows some
295 particularities. On the one hand, in the structure the membrane-embedded stretch
296 (the N-terminal region of helix A) encompasses from Leu49 to Ala69 residues, and
297 includes several charged amino acids, probably involved in the Ca^{2+} transport
298 across the membrane (Glu51, Glu55, Glu58, Asp59 and Arg63). Therefore, the ΔG_{pred}
299 value for this segment (L49-A69) is remarkably higher (and positive) than expected
300 for a TM helix (Fig. 5b). On the other hand, the C-terminal region of this helix
301 contains a high prevalence of non-polar amino acids that is more compatible with
302 the hydrophobicity of the membrane core. Nevertheless, the presence of the
303 functional amino acids (e.g. Glu58) in the membrane-embedded region reinforce
304 the idea that the more hydrophilic N-terminal region must ultimately be embedded
305 within the membrane [23]. It has been previously shown that the position in the

306 membrane of TM helices in protein folded structures does not always correspond
307 to the thermodynamically favored positions in the membrane of the isolated helices
308 [24]. Instead, after translocon-mediated insertion of the more hydrophobic region,
309 repositioning of TM helices relative to the lipid bilayer provides a convenient way
310 for non-hydrophobic polypeptide segments to become buried within the
311 membrane. Then, the nature of helix A suggested the possibility that initial insertion
312 of the hydrophobic region can be followed by subsequent repositioning of the
313 charged region into the membrane hydrophobic core, which will be the final
314 segment embedded in the lipid bilayer (Fig. 5a). Interestingly, the ΔG Predictor
315 server (https://dgpred.cbr.su.se/index.php?p=home) selected the adjacent (C-
316 terminus, L60-F78) hydrophobic region as TM (Fig. 5b), instead of the charged
317 region found at the high-resolution structure (Uniprot code: P04191).

318 Focusing on the potential salt bridge residues (D59 and R63 pair), the
319 distance between the anionic carboxylate ($RCOO^-$) from the D59 and the cationic
320 guanidinium ($RC(NH_2)_2^+$) from the R63 was about 3.0 Å in the crystal structure
321 (Figure 5a), clearly within the permissive range for salt bridge interaction. To
322 investigate the contribution of this potential salt bridge interaction in the translocon-
323 mediated membrane insertion of this region, we worked with two different scaffold
324 sequences: the full helix A involving the residues 49-78 (Long, L); and a shorter
325 membrane-embedded version including residues 49-69 (Short, S) as found in the
326 solved structure. We also challenged the D59-R63 charge pair interaction in both
327 sequences by increasing the separation between the ionizable residues from the
328 native *i*, *i*+4 to non-permissive *i*, *i*+5, while maintaining amino acid composition
329 (R63I/I64R double mutant). *In vitro* transcription/translation of these sequences in

330 the presence of microsomes rendered singly-glycosylated molecules for the
331 construct containing full-length helix A (Fig. 5c, lane 2). In contrast, when only the
332 membrane-embedded sequence was included, the Lep chimera was mainly
333 doubly-glycosylated (Fig. 5c, lane 3), suggesting that in the full-length protein helix
334 A inserts initially through the more hydrophobic L60-F78 region and then, after
335 protein rearrangements, repositions the more hydrophilic L49-A69 region at the
336 membrane core, as found in the solved structure. Accordingly, the translocon
337 inefficiently inserted the isolated membrane-embedded (L49-A69) region (Fig. 5c,
338 lane 3), properly inserting helix A only when the full helical sequence is present (Fig.
339 5c, lane 2). When the charge paired residues were placed at a non-permissive
340 distance in terms of salt bridge interaction (*i*, *i*+5; Fig. 5e) the insertion efficiency
341 was reduced (Fig. 5c, lane 4), with a ΔG_{app} decrease (absolute values) of ~0,7
342 kcal/mol relative to the wild type sequence (Fig. 5b). As expected, this effect was
343 not observed when the same mutations were grafted on the membrane-embedded
344 (S) sequence (Fig. 5c, lane 5).

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

352

353 **DISCUSSION**

Charged residues found in alpha-helices can both give a stabilizing effect [25] as well as being important for interaction and function such as in zinc finger motifs [26]. Charged residues have also been found to be more common in globular (soluble) helices compared to TM helices [27]. This can be explained by the fact that globular helices reside in a more hydrophilic environment compared to the hydrophobic environment of the lipid bilayer. An intermediate between these is the existence of amphipathic alpha-helix in contact with the surface of a bilayer, where one face contains mainly polar residues facing an aqueous environment and the opposite face with mostly nonpolar residues facing a hydrophobic environment [28].

Whereas pairs of charged residues of the same charge facing the inside pores in TM regions fill an essential functional role, pairs of oppositely charged residues can form salt bridges that can stabilize the helix and play an important role in function such as for transportation [29]. Salt bridges might also be important for hiding the charges during translocon-mediated TM helix insertion, as the charged residues are hidden from the hydrophobic environment. Previous work has shown that charged and polar residues are conserved within TM segments [19], indicating they are crucial for function and/or stability.

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

378 An interesting observation is that positively charged Arg-Arg pairs at position
379 *i*, *i*+3 is numerically the most common pair at this position. Positively charged pairs
380 at *i*, *i*+6 are also more prevalent than expected and although Arg-Arg pairs
381 numerically only show up half as often at *i*, *i*+6 as in *i*, *i*+3, 80% of the Arg-Arg pairs
382 at position *i*, *i*+6 also contain an arginine at *i*+3, as found in helices from voltage-
383 gated ion channels that contain three or more periodically aligned Arg residues with
384 two intervening hydrophobic residues [30].

385 The fraction of salt bridges at positions *i*, *i*+5 (Fig. 3b) is an anomaly due to
386 bend alpha-helices, as found in the helix E from bacterial translocon (PDB ID: 5MG3)
387 and in the helix A from a lipid flippase (PDB ID: 6CC4) due to the presence of glycine
388 residues (Supplementary Fig. S4). Without the observed helix bending, these salt
389 bridges would not be formed.

390 By analyzing two native helices containing intra-helical salt bridges we now
391 find that the free energy of insertion (ΔG_{app}) is significantly reduced if both oppositely
392 charged residues are spaced at a permissive distance. These results indicate that
393 intra-helix salt bridge can form during translocon-assisted insertion or even earlier,
394 since in contrast to globular (soluble) helices, TM helices can be compacted inside
395 the ribosome exit tunnel [31]. The maximal reduction in ΔG_{app} seen with Asp-Lys
396 and Asp-Arg pairs in both hR and Ca ATPase helices is 0.5-0.7 kcal/mol, which is
397 in good agreement with the 0-1 kcal/mol estimated for these two pairs from
398 thermodynamic peptide partition into octanol experiments [32]. As found in the case
399 of hR helix G (Fig. 4f), this reduction might be even higher in the cell context, since
400 some auxiliary components of the membrane insertion machinery [15,33,34] can be
401 in suboptimal conditions in the microsomal vesicles.

402 As mentioned above, in the case of hR helix G the lysine residue involved in
403 the salt bridge (K258) is bound to the retinal chromophore via a protonated Schiff
404 base as found in the crystal structure of a close homologue [35]. Then, the lysine
405 residue plays a fundamental role for protein function but at the same time introduces
406 a penalty for membrane insertion. Interestingly, our data suggest that helix G
407 translocon-mediated insertion efficiency could be increased by salt bridge
408 formation between K258 and D254, and once in the membrane, retinol binding to
409 the apoprotein could occur by covalently binding the retinal as a protonated Schiff
410 base to K258 and perturbing D254 salt bridge interaction.

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

423

424 **MATERIALS AND METHODS**

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

434 **DNA Manipulation.** The sequences of interest were introduced into the modified
435 Lep sequence from the pGEM1 plasmid [12] between the *Spe*I and *Kpn*I sites using
436 two double-stranded oligonucleotides with overlapping overhangs at the ends. The
437 complementary oligonucleotides pairs were first annealed at 85 °C for 10 min
438 followed by gradual cooling to 30 °C and ligated into the vector (a kind gift from G.
439 von Heijne's lab). Mutations were obtained by site-directed mutagenesis using the
440 QuikChange kit (Stratagene, La Jolla, California). Lep system including the
441 sequences of interest in the P2 region were subcloned into *Kpn*I linearized pCAGGS
442 in-house version using In-Fusion HD cloning Kit (Takara) according to the
443 manufacturer's instructions. A *c*-myc tag (EQKLISEEDL) at the C-terminus of the
444 Lep-derived sequence was added by PCR before cloning. For *in vitro* assays, DNA
445 was amplified by PCR adding the T7 promoter during the process. All sequences
446 were confirmed by sequencing the plasmid DNA at Macrogen Company (Seoul,
447 South Korea).

448 **Translocon-mediated insertion into microsomal membranes.** Lep constructs in
449 pGEM with L4/A15, 3QBG and 1SU4 segments and its variations were transcribed
450 and translated using the TNT T7 Quick Coupled System (#L1170, Promega). Each
451 reaction containing 1 µL of PCR product, 0.5 of EasyTag™ EXPRESS 35S Protein
452 Labeling Mix (Perkin Elmer) (5.5 µCi) and 0.3 µL of microsomes (tRNA Probes) was
453 incubated at 30ºC for 90 min. Endo H treatment was done following the
454 manufacturer's instructions. Samples were analysed by SDS-PAGE (12-14%
455 polyacrylamide). The bands were quantified using a Fuji FLA-3000 phosphoimager
456 and the Image Reader 8.1j software. Free energy was calculated using: ΔG_{app}=-RT
457 $\ln K_{app}$, where $K_{app}=f_{2g}/f_{1g}$ being f_{1g} and f_{2g} the fraction of singly glycosylated and
458 double glycosylated protein, respectively.
459 **Free apparent insertion energy, ΔG.** The free insertion energy of a TM region, ΔG,
460 is calculated as per the experimentally defined Biological hydrophobicity scale [12].
461 This scoring is amphiphilic with hydrophobic residues contributing a lower
462 (negative) ΔG while hydrophilic contributes a higher (positive) ΔG. The total ΔG of a
463 region is the sum of individual position specific scores. This scoring can give an
464 indication of how favorable the amino acid composition of a TM region is to be
465 inserted in a lipid bilayer membrane. To note is that hydrophobicity alone is not the
466 only driving force and that the positive inside rule [36,37] and help from proceeding
467 TM regions [38,39] can assist insertion in polytopic TM proteins especially.
468 **Expression in mammalian cells.** Lep constructs containing 3QBG or 1SU4
469 segments and its variations were tagged with *c*-myc epitope at their Ct
470 (EQKLISEEDL) and inserted in a pCAGGS-ampicillin plasmid. Once the sequence
471 was verified, plasmids were transfected into HEK293-T cells using Lipofectamine

472 2000 (Life Technologies) according to the manufacturer's protocol. Approximately
473 24 h post-transfection cells were harvested and washed with PBS buffer. After a
474 short centrifugation (1000 rpm for 5 min on a table-top centrifuge) cells were lysed
475 by adding 100 µL of lysis buffer (30 mM Tris-HCl, 150 mM NaCl, 0.5% Nonidet P-
476 40) were sonicated in an ice bath in a bioruptor (Diagenode) during 10min and
477 centrifuged. After protein quantification, equal amounts of protein were submitted
478 to Endo H treatment or mock-treated followed by SDS-PAGE analysis and
479 transferred into a PVDF transfer membrane (ThermoFisher Scientific) as previously
480 described [40]. Protein glycosylation status was analysed by Western Blot using an
481 anti-c-myc antibody (Sigma), anti-rabbit IgG-peroxidase conjugated (Sigma), and
482 with ECL developing reagent (GE Healthcare). Chemiluminescence was visualized
483 using an ImageQuantTM LAS 4000mini Biomolecular Imager (GE Healthcare).
484 **Core segment definition.** We define core segments as a transmembrane region
485 minus the first and last 5 residues. This is to ignore the interface regions which are
486 known to contain polar residues.
487 **Salt bridge definition.** Salt bridges are defined as per [41], where a salt bridge is
488 defined if a side chain carbonyl oxygen atom in Asp-Glu is within 4.0 Å from the
489 nitrogen atom in Arg-Lys. This conforms to other works [7,42] with the definition
490 that the atoms are within hydrogen bond distance. We also define local salt bridges
491 as being bridges that are separated by at most 7 residues in the sequence. This is
492 to separate long salt bridge interactions, which can occur between spatially close
493 residues that are separated in sequence, such as coiled coils where salt bridges
494 can be between separate alpha-helices.

495 **Transmembrane helices dataset (TM dataset).** The full pipeline is available as a
496 Makefile together with supporting scripts in the github repository. The full PDBTM
497 database [16] was downloaded together with their list of non-redundant protein pdb
498 ids. This list is used to generate both sequence and topology of the proteins by
499 extracting both from the PDBTM-xml. For each protein in the non-redundant list,
500 the membrane regions are extracted as per the PDBTM database annotation. All
501 non-membrane regions are annotated 'i' for convenience. To support future
502 analysis, membrane regions longer than 10 were run through DSSP and annotated
503 with 'M' if all residues in the core segment were defined as alpha-helix ('H'),
504 otherwise, the full membrane region is annotated 'm'. Observe that this creates
505 fasta-like 3line files, that only contain topological annotation with ambiguous TM
506 regions annotated as 'm' instead of the normal 'M'. These proteins are then cluster
507 using cd-hit [43,44] at 40% identity using the parameter -c 0.4 -n 2 -T 0 -M 0 -d 0.
508 During the extraction of TM regions, the corresponding structure file from
509 RCSB was used to calculate all salt bridges within the current protein and any salt
510 bridge that has at least one residue within any TM region was saved. Additionally,
511 all salt bridges whose both residues were within the same segment and within 7
512 residues of each other were annotated as local as per the salt bridge definition
513 above.
514 **Extraction of charged residues.** From all annotated TM regions of length 17 or
515 longer [2], the core segment was extracted. All these core regions were then
516 scanned and when a charged residue was encountered, we recorded any other
517 charged residue from 7 residues before the current one to 7 residues after. This
518 results in charged residues that can contain a charged pairing partner outside of

519 the TM segment and will therefore differ slightly from charged pairs which are
520 defined next.

521 **Extraction of charged pairs.** From all annotated TM regions of length 17 or longer
522 [2], the core segment was extracted. All these core regions were then scanned and
523 when a charged residue was encountered, we look at up to 7 residues in front of it
524 or to the end of that core region, whichever came first. All occurrences of charged
525 pairs were recorded, resulting in charged pairs where both residues were fully within
526 the core segment of a TM helix.

527 **MSA-dataset extension (TM-MSA).** Using the TM dataset, we extended it by
528 creating an MSA alignment of each protein using jackhmmer [45] against Uniref90
529 with one iteration and an E-value cut off of 10^{-3} with the following parameters:
530 -N 1 -E 1e-3 --incE 1e-3 --cpu 14

531 From each alignment, we then sampled up to 200 hits, including the initial
532 seed sequence. If there were fewer than 200 hits, we used them all. We then used
533 the original topology for each alignment to extract all TM regions, only to include
534 parts where the sequence covers the full TM region and where the sequence did
535 not contain any insertions or deletions.

536 **Dataset of helices from globular proteins (GLOB and GLOB-MSA).** To create a
537 reference dataset of globular α-helical proteins we extracted all globular all-alpha
538 protein domains from SCOP. As SCOP classifies domains of proteins resulting in
539 that one domain of a protein can be annotated as globular whereas another domain
540 as TM (see 1PPJ chain D as an example) we reduced the SCOP list against the
541 redundant list of all PDBTM [16] chains to clear any overlap. This results in 4,500
542 proteins in total.

543 Topological files with sequence and membrane topology are created with the
544 help of the RCSB secondary structure file and only membrane segments whose
545 core region (*i.e.*, central 15 residues) is annotated as pure (canonical) α-helices were
546 retained, *i.e.*, those TM segments containing any residues within the core annotated
547 as loops or other types of secondary structures were removed. This file was further
548 homology reduced and alignments prepared in the same manner and using the
549 same parameters as the TM dataset described above.

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

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

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

562 $logOddsRatio = log((A/B)/(C/D))$

563 Where for two amino acids p_1 and p_2:

564 A = number of pairs of p_1 to p_2

565 B = number of total pairs

566 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 based on the number of hypothesis, 20 * 20 * 10, number of amino acids square times the number of steps.

Acknowledgments

We thank Pilar Selvi for excellent technical assistance. This work was supported by grants BFU2016-79487-P form the Spanish Ministry of Economy and Competitiveness 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 (FPU) from the Spanish Ministry of Education.

Conflict of Interest Statement

None declared.

588 **Table 1**. About 10 % of the core residues in the TM dataset are polar and just under
589 4% are charged residues. This is significantly less than 37.5% and 25.5%
590 respectively in the globular dataset. In the transmembrane dataset, just over half of
591 the charged core residues form a charged pair at distance 1, 3 or 4. *In the case of
592 the globular set, there are more potential charged pairs than charged core residues
593 as multiple residues are counted more than once as they form more than one
594 potential pairing with separation of 1, 3 or 4 steps. It is clearly seen that although
595 the TM dataset contains more helices per protein and has a higher proportion of
596 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*	-

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*	*$1.97e^{-10}$*

611 **FIGURE LEGENDS**

612

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

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

639

640 **Figure 2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8'**
641 **for the TM-MSA dataset.** The rows on the y-axis indicate the first amino acid in
642 the pair and the columns on the x-axis the second. The residues are ordered by
643 hydrophobicity according to the Engelman order [46]. See figure S1 for the
644 equivalent of the globular dataset. S2 and S3 show the same plots masked for
645 statistical significance.

646

647 **Figure 3. Charge pairs in TM helical sequences and structures.** Helical wheel
648 projection and lateral views of an α-helix are shown on top. The initial position *i* and
649 the following 8 residues are numbered. Residues in positions *i*+3 (orange), *i*+4
650 (yellow), and *i*+7 (light brown) are mainly on the same face of the helix, but *i*+7 is
651 placed too far for a salt bridge interaction. **a)** The log odds ratios of charged pairs
652 for '*i*, *i*+1' through '*i*, *i*+8' in the TM dataset. Plain filled bars refer to oppositely
653 charged pairs and the forward slash are the and same charged pairs, all with error
654 bars. **b)** Fraction of oppositely charged pairs that form local salt bridges. The small
655 bump at *i* +5 are the two proteins 6CC4 (helix A) and 5MG3 (helix E), which both
656 exhibit a bend in the alpha helix due to the presence of glycine residues (see Fig.
657 S4).

658

659 **Figure 4. Insertion of halorhodopsin helix G from *Natronomonas pharaonis***
660 **(3QBG) into microsomal and cellular membranes. (a)** Tested sequences from
661 3QBG including the gap distance, and the predicted (ΔG_{pred}) and experimental (*in*
662 *vitro* $\Delta G_{exp}^{in\ vitro}$ and in vivo $\Delta G_{exp}^{in\ vivo}$, respectively) ΔG values in kcal/mol are shown.
663 Amino acids with a positive or negative charge are highlighted in blue (K) and red
664 (D), respectively. Green numbers indicate negative ΔG (insertion) values, while red
665 numbers denote ΔG values above 0 (translocation). **(b)** Frontal view of 3QBG
666 monomer structure. The helix G is highlighted in orange with the D254 and K258
667 shown in sticks colored red and blue respectively. The membrane position is
668 indicated by a red (outer) and blue (inner) discontinuous line, according to OPM
669 dataset [22]. Lateral **(c)** and upper **(d)** views of the 3QBG trimeric structure. The
670 helix G is highlighted in orange with the D and K shown in sticks colored red and
671 blue, respectively. The different monomers are shown in transparent blue, pink and
672 green. Representative examples (*n*=3) of *in vitro* protein translations in the presence
673 of ER-derived microsomes **(e)** and Western blots (*n*=3) of *in vivo* protein translations
674 in HEK-293T cells **(f)** in the presence (+) or absence (-) of Endoglycosidase H
675 (EndoH), a glycan-removing enzyme. The absence of glycosylation of G1 and G2
676 acceptor sites is indicated by two white dots, single glycosylation by one white and
677 one black dot, and double glycosylation by two black dots. **(g)** Zoom view centered
678 on the salt bridge between D254 and K257 at *i*, *i*+4 (left) and *i*, *i*+5 (right) gaps. D
679 and K residues are shown in sticks colored red and blue, respectively, while the
680 dashed line indicates the $RCOO^-$ to RNH_3^+ distance.
681
682 **Figure 5. Insertion of Calcium ATPase (1SU4) helix A into microsomal and**

683 **cellular membranes. (a)** Lateral view of 1SU4 structure. Zoom view of the A helix
684 (right panel). The membrane-embedded region of helix A is highlighted in yellow.
685 Charged amino acids are shown as sticks in blue (R), red (D) and pink (E),
686 respectively. L49, A69 and F78 are also shown as sticks to define helix's
687 subdomains. The membrane location is indicated by red (outer) and blue (inner)
688 discontinuous lines according to OPM dataset [22] and the distance between the R
689 and D charges is indicated in Å. **(b)** Helix A-derived sequences from 1SU4 including
690 the gap between charged residues, and the predicted (ΔG_{pred}) and experimental (*in*
691 *vitro* $\Delta G_{exp}^{in\ vitro}$ and in vivo $\Delta G_{exp}^{in\ vivo}$, respectively) ΔG values in kcal/mol are shown.
692 Amino acids with a positive charge are highlighted in blue (K) while negatively
693 charged are marked in red (D) and pink (E). Green numbers indicate negative ΔG
694 (insertion) values while red numbers denote ΔG values above 0 (translocation).
695 Representative examples (n=3) of *in vitro* protein translations in the presence of ER-
696 derived microsomes **(c)** and Western blots (n=3) of *in vivo* protein translation in
697 HEK-293T cells **(d)** in the presence (+) or absence (-) of Endoglycosidase H (EndoH),
698 a glycan-removing enzyme. The absence of glycosylation of G1 and G2 acceptor
699 sites is indicated by two white dots, single glycosylation by one white and one black
700 dot, and double glycosylation by two black dots. **(e)** 1SU4 helix A *i*, *i*+5 mutant. The
701 membrane-embedded region of helix A is highlighted in yellow. Charged residues
702 are shown as sticks in blue (R), red (D) and pink (E). L49, A69 and F78 are also
703 shown as sticks to define helix's subdomains. The membrane is indicated by red
704 (outer) and blue (inner) discontinuous lines as in (a), and the dashed line indicates
705 the $RCOO^-$ and $RC(NH_2)_2^+$ distance.
706

Figure S1. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the GLOB-MSA dataset. Log odds ratios for the middle core residues of α-helices of at least 17 residues in length in the GLOB-MSA dataset. The rows on the y-axis indicate the first amino acid in the pair and the columns on the x-axis the second. The residues are ordered as in Fig. 2.

Figure S2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the GLOB-MSA dataset. Log odds ratios as in S1 but all pairs with P-value >0.05 have been masked.

Figure S3. Masked log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the TM-MSA dataset. Log odds ratios as in Fig. 1 but all pairs with P-value >0.05 have been masked.

Figure S4. Salt bridges at pair *i*, *i*+5. Left: helix E from bacterial translocon (PDB ID: 5MG3). Glycine (orange) residue causes a kink facilitating salt bridge interaction between Asp50 (red) and Arg55 (blue). Right: helix A from a lipid flippase (PDB ID: 6CC4). Glycine (orange) residue causes a kink facilitating salt bridge interaction between Arg153 (blue) and Glu158 (red). Distances are shown in Ångström.

REFERENCES

1. Martínez-Gil L, Saurí A, Marti-Renom MA, Mingarro I (2011) Membrane protein integration into the endoplasmic reticulum. *FEBS J* **278**: 3846–3858.
2. Baeza-Delgado C, Marti-Renom MA, Mingarro I (2012) Structure-based statistical analysis of transmembrane helices. *Eur Biophys J* **42**: 199–207.

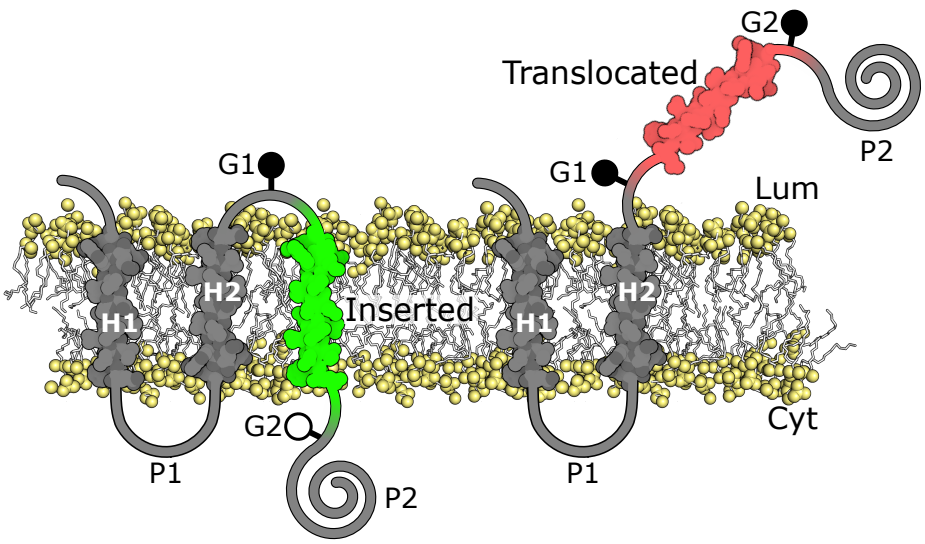
3. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* **43**: W401–W407.
4. Tsirigos KD, Govindarajan S, Bassot C, Västermark Å, Lamb J, Shu N, Elofsson A (2018) ScienceDirect Topology of membrane proteins — predictions, limitations and variations. *Curr Opin Struct Biol* **50**: 9–17.
5. Bañó-Polo M, Baeza-Delgado C, Orzáez M, Marti-Renom MA, Abad C, Mingarro I (2012) Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing. *PLoS ONE* **7**: e44263.
6. Marqusee S, Baldwin RL (1987) Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci USA* **84**: 8898–8902.
7. Donald JE, Kulp DW, DeGrado WF (2011) Salt bridges: Geometrically specific, designable interactions. *Proteins* **79**: 898–915.
8. Mbaye MN, Hou Q, Basu S, Teheux F, Pucci F, Rooman M (2019) A comprehensive computational study of amino acid interactions in membrane proteins. *Sci Rep* **9**: 12043–14.
9. Baeza-Delgado C, Heijne von G, Marti-Renom MA, Mingarro I (2016) Biological insertion ofcomputationally designed shorttransmembrane segments. *Sci Rep* 1–9.
10. Chin CN, Heijne von G (2000) Charge pair interactions in a model transmembrane helix in the ER membrane. *J Mol Biol* **303**: 1–5.
11. Braunger K, Pfeffer S, Shrimal S, Gilmore R, Berninghausen O, Mandon EC, Becker T, Förster F, Beckmann R (2018) Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. *Science* **360**: 215–219.
12. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, Heijne von G (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**: 377–381.
13. Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, Heijne von G (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450**: 1026–1030.
14. Bañó-Polo M, Martínez-Gil L, Barrera FN, Mingarro I (2019) Insertion of Bacteriorhodopsin Helix C Variants into Biological Membranes. *ACS Omega* **5**: 556–560.
15. Tamborero S, Vilar M, Martínez-Gil L, Johnson AE, Mingarro I (2011) Membrane insertion and topology of the translocating chain-associating membrane protein (TRAM). *J Mol Biol* **406**: 571–582.
16. Kozma D, Simon I, Tusnády GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res* **41**: D524–D529.
17. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2013) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* **42**: D310–D314.
18. Andreeva A, Kulesha E, Gough J, Murzin AG (2019) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res* **48**: D376–D382.
19. Illergård K, Kauko A, Elofsson A (2011) Why are polar residues within the membrane core evolutionary conserved? *Proteins* **79**: 79–91.

20. Samanta U, Pal D, Chakrabarti P (1999) Packing of aromatic rings against tryptophan residues in proteins. *Acta Cryst (1999) D55, 1421-1427 [doi:101107/S090744499900726X]* 1–7.
21. Kanada S, Takeguchi Y, Murakami M, Ihara K, 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.
22. Kumar S, Nussinov R (2002) Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. *Biophys J* **83**: 1595–1612.
23. Toyoshima C, Nakasako M, Nomura H, Ogawa H (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**: 647–655.
24. Kauko A, Hedin LE, Thebaud E, Cristobal S, Elofsson A, Heijne von G (2010) Repositioning of transmembrane alpha-helices during membrane protein folding. *J Mol Biol* **397**: 190–201.
25. Armstrong KM, Baldwin RL (1993) Charged histidine affects alpha-helix stability at all positions in the helix by interacting with the backbone charges. *Proc Natl Acad Sci USA* **90**: 11337–11340.
26. Lin C-Y, 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.
27. Kauko A, Illergård K, Elofsson A (2008) Coils in the Membrane Core Are Conserved and Functionally Important. *J Mol Biol* **380**: 170–180.
28. Giménez-Andrés M, Čopič A, Antonny B (2018) The Many Faces of Amphipathic Helices. *Biomolecules* **8**: 45–14.
29. Walther TH, Ulrich AS (2014) ScienceDirectTransmembrane helix assembly and the role of salt bridges. *Curr Opin Struct Biol* **27**: 63–68.
30. Okamura Y, Fujiwara Y, Sakata S (2015) Gating Mechanisms of Voltage-Gated Proton Channels. *Annu Rev Biochem* **84**: 685–709.
31. Bañó-Polo M, Baeza-Delgado C, Tamborero S, Hazel A, Grau B, Nilsson I, Whitley P, Gumbart JC, Heijne von G, Mingarro I (2018) Transmembrane but not soluble helices fold inside the ribosome tunnel. *Nat Commun* **9**: 5246.
32. Jayasinghe S, Hristova K, White SH (2001) Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* **312**: 927–934.
33. Shurtleff MJ, Itzhak DN, Hussmann JA, Schirle Oakdale NT, Costa EA, Jonikas M, Weibezahn J, Popova KD, Jan CH, Sinitcyn P, et al. (2018) The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *eLife* **7**: 382.
34. Chitwood PJ, Hegde RS (2020) An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature* **584**: 1–22.
35. Kolbe M, Besir H, Essen LO, Oesterhelt D (2000) Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. *Science* **288**: 1390–1396.
36. Heijne von G (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341**: 456–458.
37. Lerch-Bader M, Lundin C, Kim H, Nilsson I, Heijne von G (2008) Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. **105**: 4127–4132.

38. Hedin LE, 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.
39. Bañó-Polo M, Martínez-Gil L, Wallner B, Nieva JL, Elofsson A, 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.
40. Duart G, García-Murria MJ, Grau B, Acosta-Cáceres JM, Martínez-Gil L, Mingarro I (2020) SARS-CoV-2 envelope protein topology in eukaryotic membranes. *Open Biol* **10**: 200209–6.
41. Kumar S, Ma B, Tsai CJ, Nussinov R (2000) Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins* **38**: 368–383.
42. Bosshard HR, Marti DN, Jelesarov I (2004) Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J Mol Recognit* **17**: 1–16.
43. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.
44. Li W, Godzik A (2006) *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*.
45. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.
46. Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15**: 321–353.

Figure 1

a)



b)

AA	Gap	Sequence	$\Delta G_{pred.}$	$\Delta G_{exp.}$
L4/A15	-	AAAALALAAAAAALALAAAA (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)

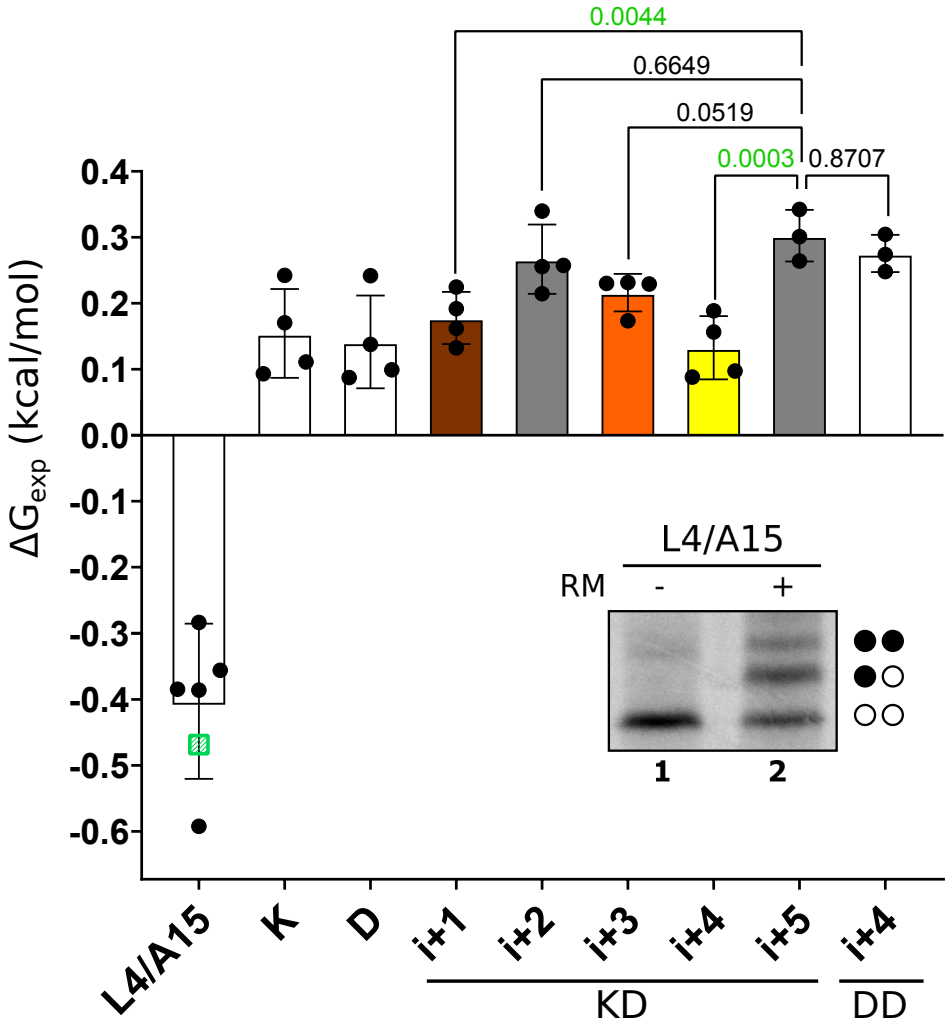


Figure 2

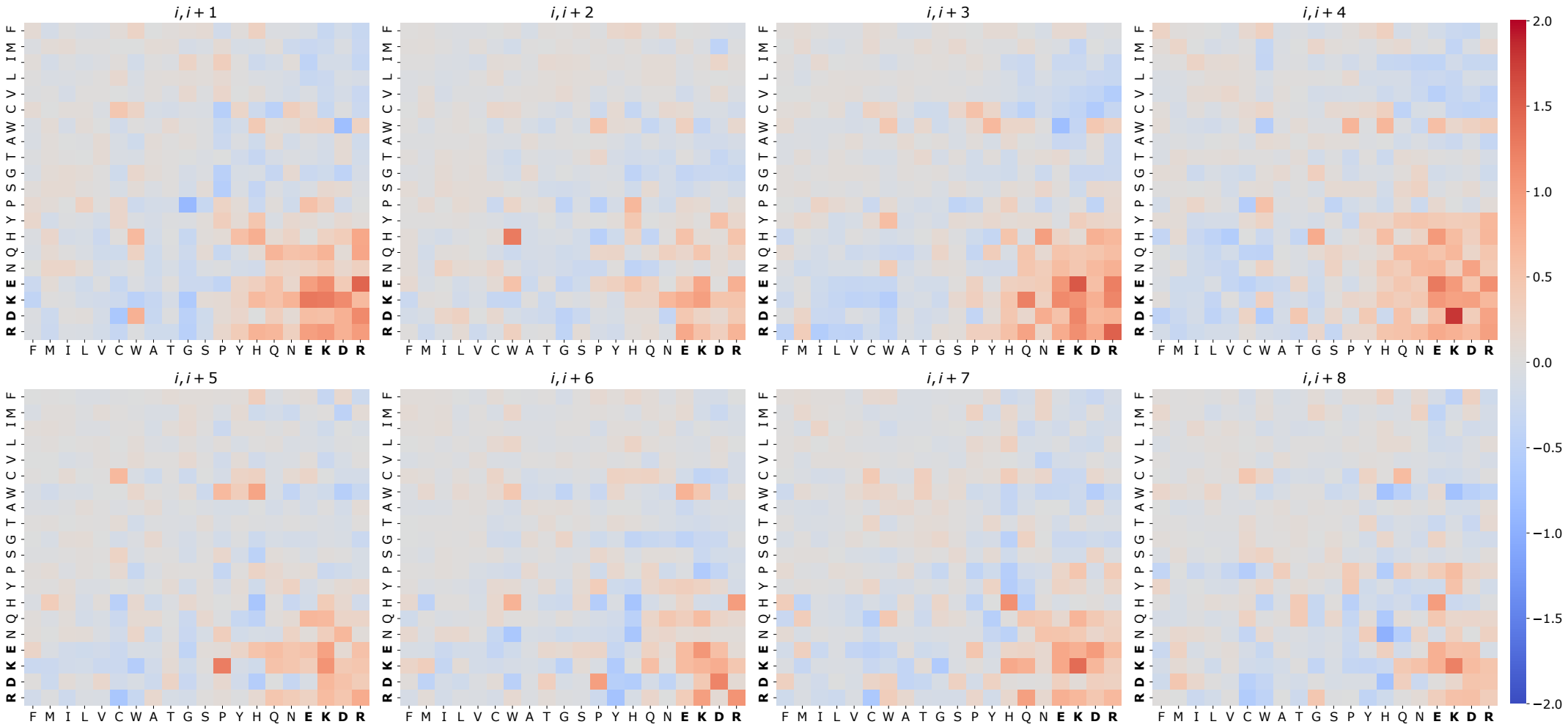

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure 3

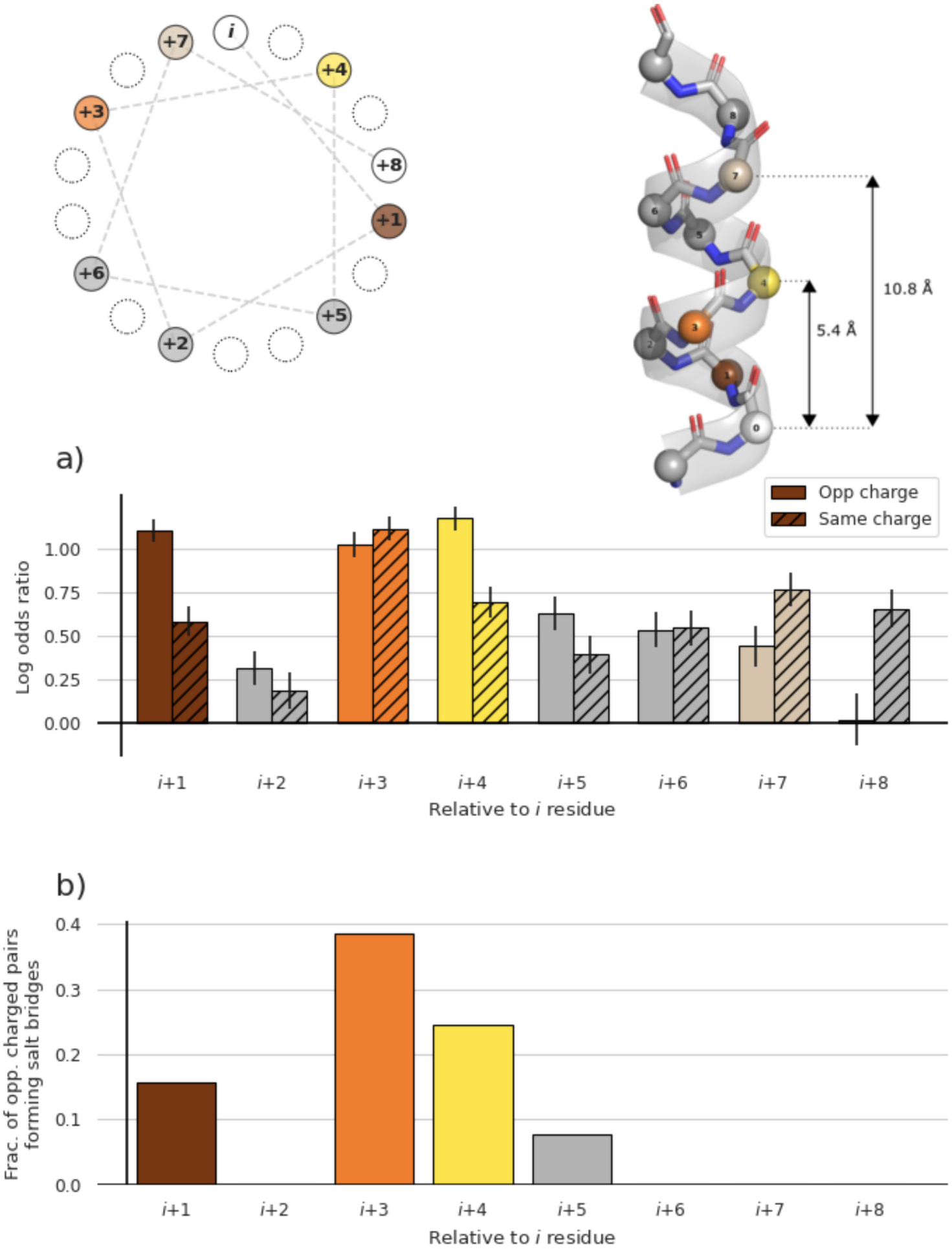
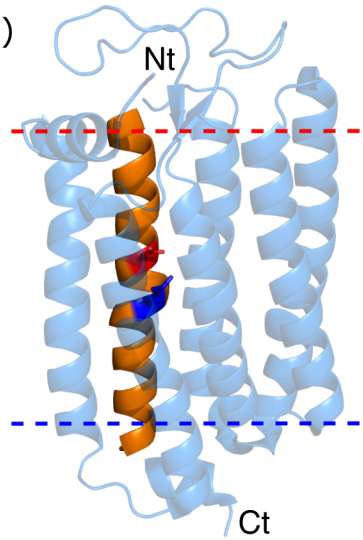

i
+1
+2
+3
+4
+5
+6
+7
+8
10.8 Å
5.4 Å
a)
Opp charge
Same charge
Log odds ratio
1.00
0.75
0.50
0.25
0.00
i+1
i+2
i+3
i+4
i+5
i+6
i+7
i+8
Relative to i residue
b)
Frac. of opp. charged pairs forming salt bridges
0.4
0.3
0.2
0.1
0.0
i+1
i+2
i+3
i+4
i+5
i+6
i+7
i+8
Relative to i residue

Figure 4

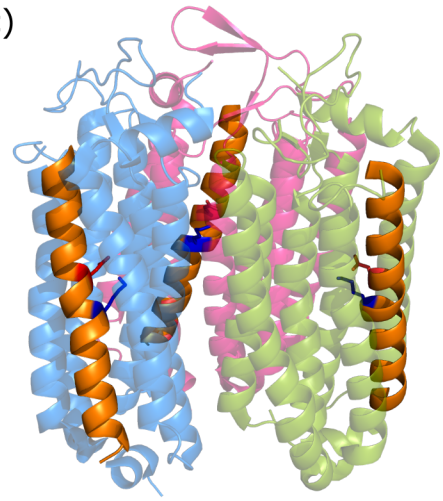
a)

Name	Sequence	Gap	$\Delta G_{pred.}$	$\Delta G^{in\ vitro}_{exp.}$	$\Delta G^{in\ vivo}_{exp.}$
wt	243 254 258 271 VGVTSWGYSFL**D**IVA**K**YIFAFLLLNYLTS	4	+1.73	-0.24	-1.54
DD	VGVTSWGYSFL**D**IVA**D**YIFAFLLLNYLTS	4	+1.66	+0.11	-
DA	VGVTSWGYSFL**D**IVA**A**YIFAFLLLNYLTS	4	-0.38	-0.88	-
DK *i*, *i*+5	VGVTSWGYSFL**D**IVAY**K**IFAFLLLNYLTS	5	+1.25	+0.31	+1.25

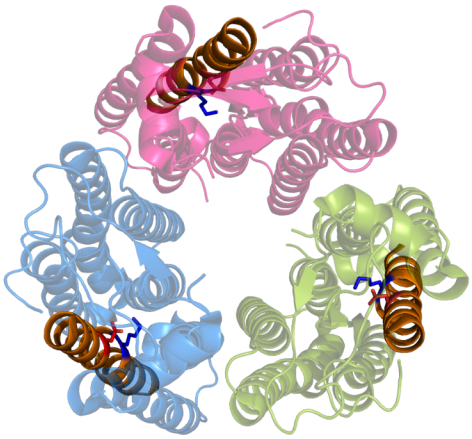
b)



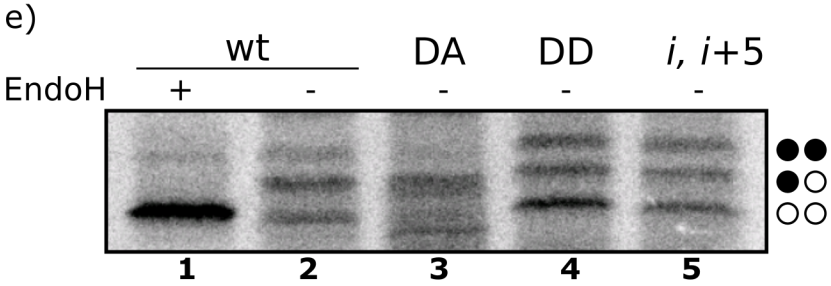
c)



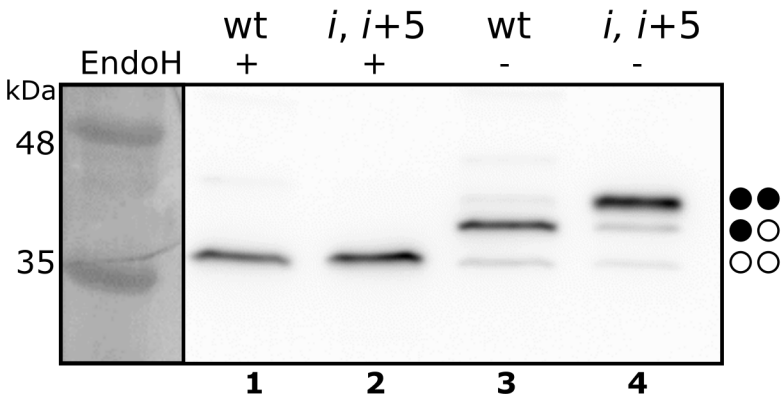
d)



e)



f)



g)

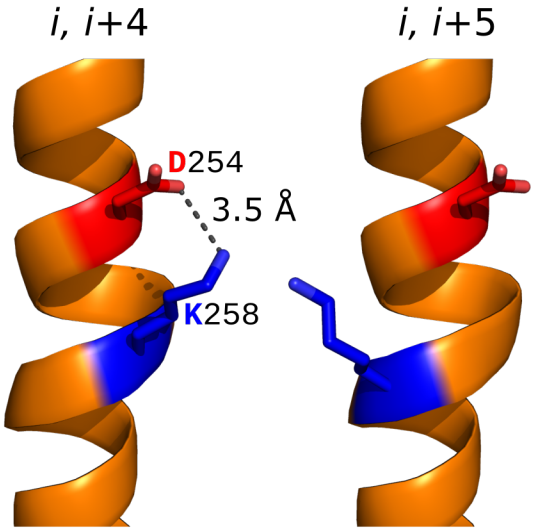
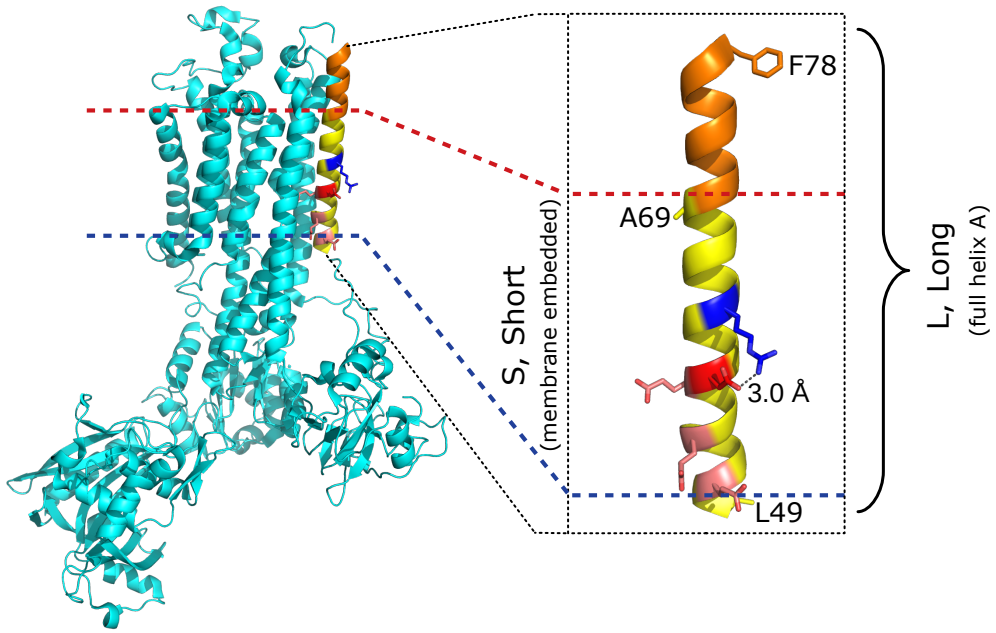


Figure 5

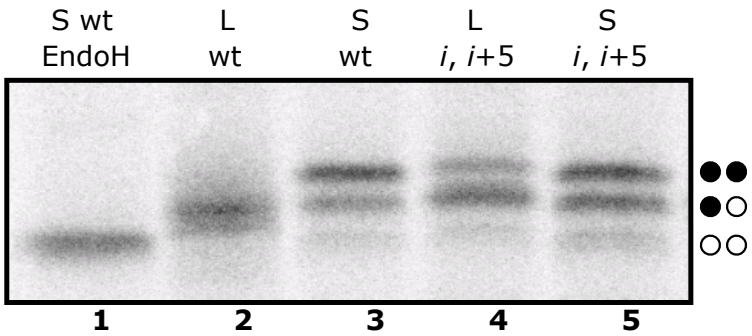
a)



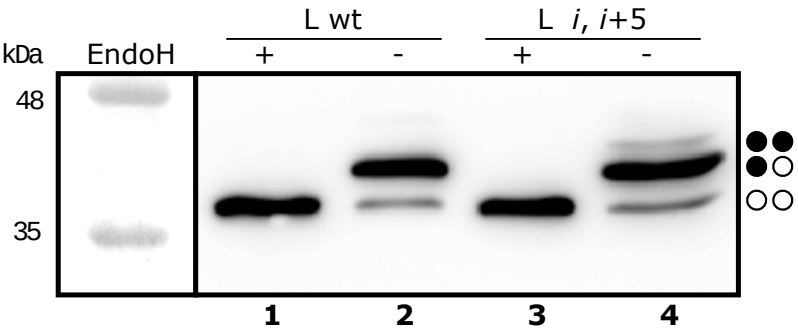
b)

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)

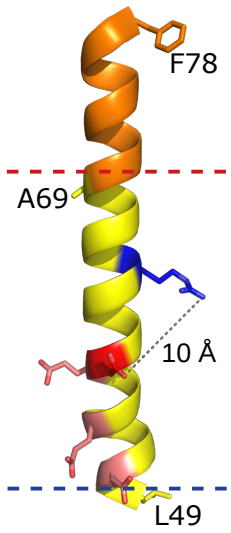


Figure S1

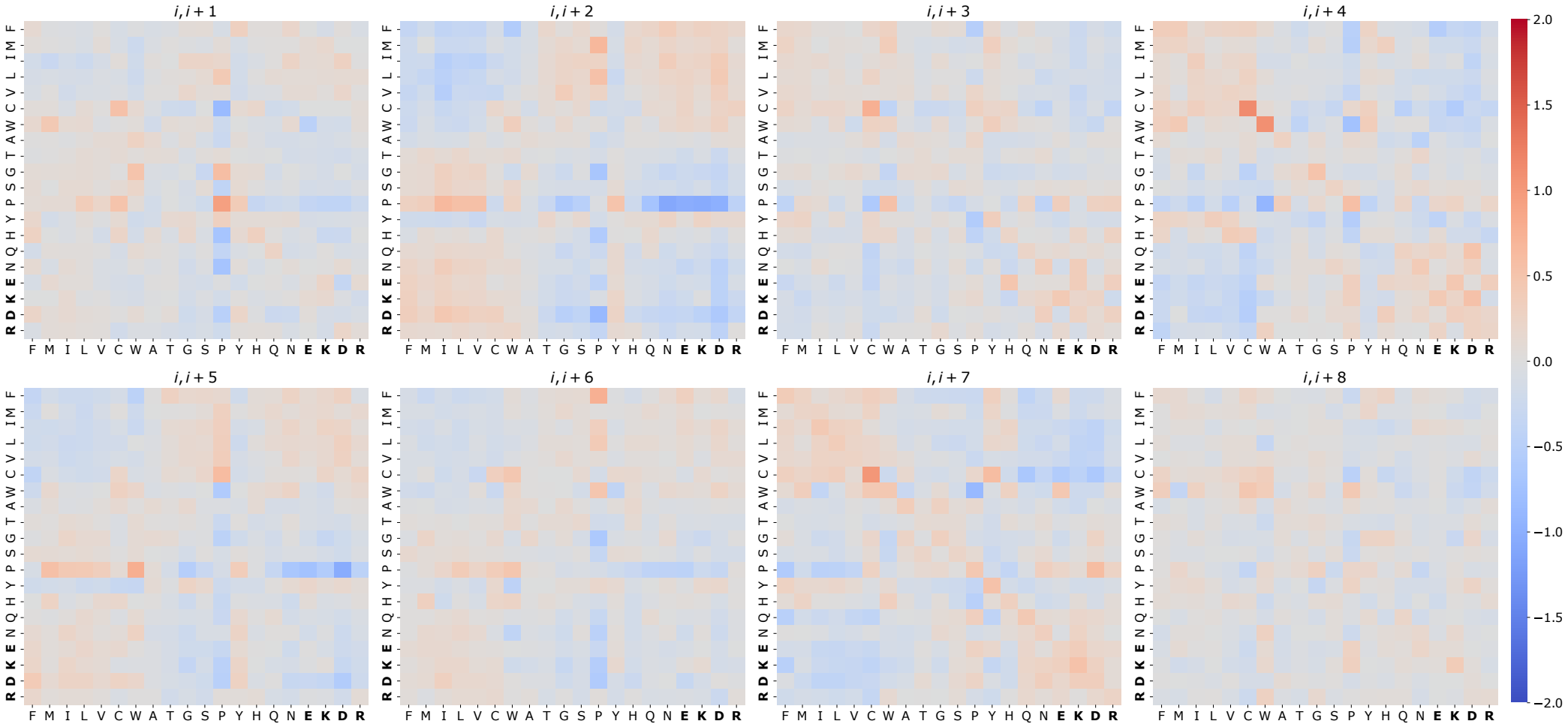

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure S2

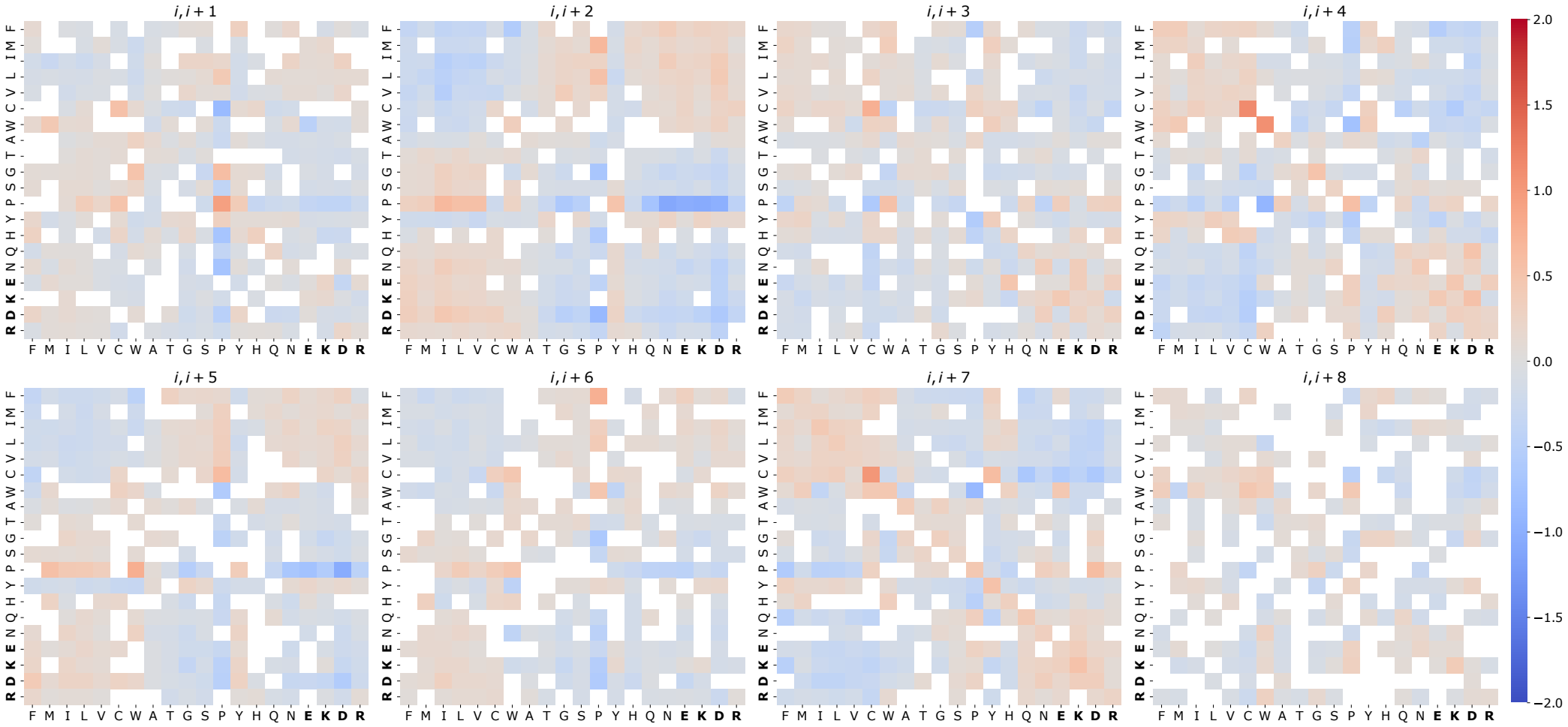
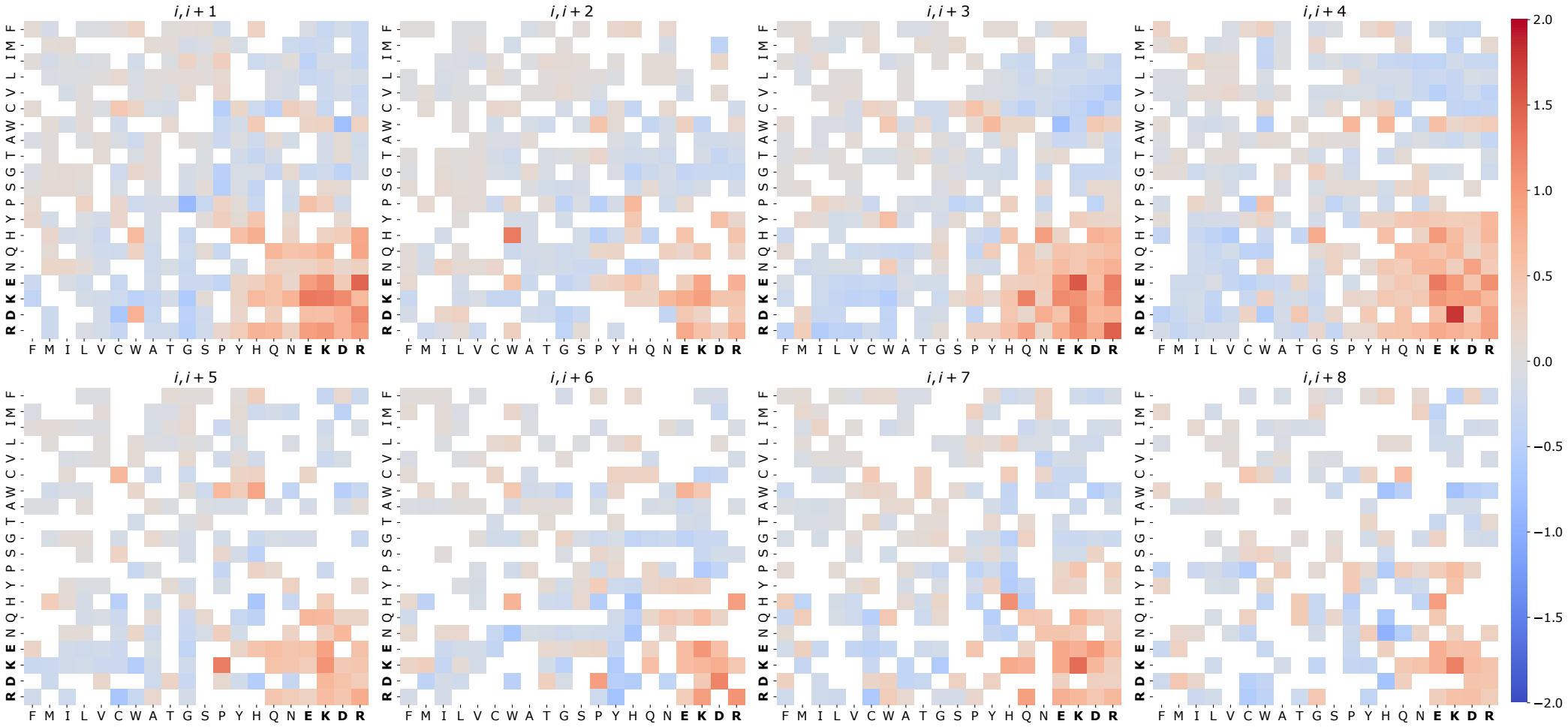

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

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

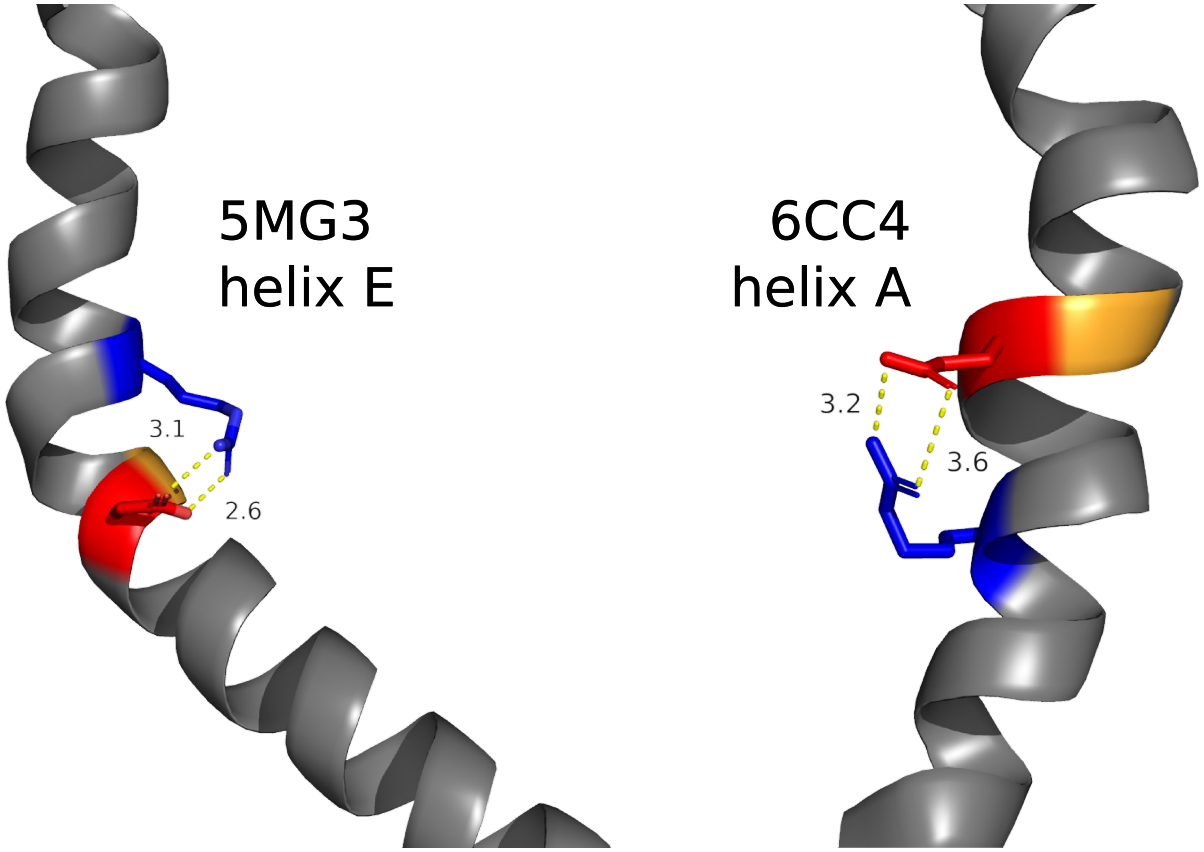



Figure S4