February 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, Insitut 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;

42 **INTRODUCTION**

43 Most integral membrane proteins have to insert their transmembrane (TM) segments into
44 the lipid bilayer in a helical conformation and then acquire a defined three-dimensional
45 structure by packaging their helices [1]. α-Helical TM segments are largely composed of
46 apolar residues because of the hydrophobic nature of the membrane environment.
47 Nevertheless, in some cases, it is necessary for the protein activity to include polar amino
48 acids in a TM region in order to develop a functional or structural role [2]. This fact is
49 sometimes not contemplated in modern membrane topology prediction tools [3,4], in
50 which the presence of charged amino acids in a sequence automatically suppose a penalty
51 increase in the predicted free energy (ΔG_{pred}) of insertion. The presence of polar amino
52 acids in TM regions is more frequent than what would be expected [5], especially when
53 these are in pairs on the same face of an α-helix.

54 Salt bridges are electrostatic interactions between negatively (D, E) and positively
55 charged (K, R, H) amino acids that play an important role in protein stabilization [6].
56 Many studies have shown that pairs of charged residues that form potential salt-bridges
57 stabilize soluble α-helices [7]. Salt bridges play an important role in the folding of
58 globular proteins and, despite their low occurrence in TM domains, it seems that the
59 contribution in membrane protein stability could be even more determinant. This
60 contribution is especially important in membrane protein biogenesis, as salt bridges help
61 to bury the polarity of charged residues in a hydrophobic environment [8]. Apart from
62 that, it has been suggested that potential salt bridges could help in the insertion of TM α-

63 helices [5,9], even though their predicted ΔG penalty is well above what is usually seen
64 for TM segments.

65 To investigate the potential formation of intra-helical salt bridges in TM α-helices,
66 we analyzed the composition of the TM domains from membrane proteins of known
67 structures looking for preferences in the pairing of charged amino acids. This analysis
68 showed that charged residue pairing is more prevalent than expected for pairs located on
69 the same face of α-helices within membranes. Likely, salt bridge formation on the same
70 face of α-helices reduces the unfavorable energetics of inserting ionizable residues into
71 the hydrophobic membrane core [10]. We use this knowledge together with the known
72 membrane protein structures to generate a list of potential candidates for further *in vitro*
73 and *in vivo* experiments.

74 In this work, we have studied the presence of intra-helical salt bridges in TM
75 domains using *in silico*, *in vitro* and *in vivo* systems, showing that despite this being an
76 unexpected phenomenon in nature, salt bridges can be crucial for membrane insertion.
77 Also, we determined in a quantitative manner that the apparent free energy (ΔG_{app}) of
78 membrane insertion through the translocon machinery can be decreased between 0.5-
79 0.7 kcal/mol by position-specific charge pair interaction, which is not contemplated in
80 the commonly used ΔG predictors. These findings will lead to a better understanding of
81 the insertion mechanism of TM helices and to improve prediction tools that would more
82 accurately be able to model the presence of charged residues in these helices.

83

84

85

86 **RESULTS**

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

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

108 membrane insertion is quantified by analyzing the fractions of singly glycosylated (i.e.,
109 membrane inserted) and doubly glycosylated (i.e., non-inserted) molecules, which can
110 be expressed as an experimental apparent free energy of membrane insertion, ΔG_{exp} (see
111 Materials and Methods) [12].

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

127

128 **Charged pairs in transmembrane helices**

129 Alpha helices are a common secondary structure in both globular and membrane
130 proteins. We created two main datasets, TM dataset with helical membrane proteins of
131 known structure from the PDBTM-dataset [16], and GLOB dataset with globular alpha
132 helical proteins selected from the SCOP-database [17,18], see methods for the full
133 creation steps. Table 1 shows a breakdown of alpha-helices, charges and salt bridges in
134 the two datasets. What is clear is that long alpha helices (≥ 17 residues) form a larger
135 proportion in the TM dataset than in the globular helices (GLOB) dataset, which is logical
136 as most TM helices need to span through the hydrophobic core of the lipid bilayer of
137 thickness ~30 Å.

138 Our TM dataset showed the same distribution of polar charges as previous studies
139 [19], with about 10 % of polar residues in the core membrane regions (see Table 1), and
140 about one-third of these are charged. Over half of the charged residues could form pairs
141 with other charged residues at intervals of *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4. In contrast, the GLOB
142 dataset had a much higher proportion of charged polar residues. The GLOB dataset also
143 contained about 15 times as many charged pairs relative to its size compared to the TM
144 dataset, again indicating that charged residues are more common in globular than in TM
145 helices.

146 The two datasets, TM and GLOB, were extended with homologous sequences
147 identified by searching with jackhammer against UniProt. All sequences with an E-value
148 lower than 10^{-3} were included. These datasets are named TM-MSA and GLOB-MSA.
149 Using data in the TM-MSA dataset to produce the log odds ratios, we identified
150 periodicity patterns of charged residue pairs that are more common than what would

151 have been expected from the underlying amino acid composition, see Figure 2). We found
152 that polar residues at pairs *i*, *i*+3; *i*, *i*+4 and *i*, *i*+7 are significantly enhanced (Fig. 2). This
153 feature is strengthened when we examine the same plot for the GLOB-MSA dataset,
154 where these patterns were not observed (Figure S1). This was also clear when statistical
155 significance was taken into account, see Figures S2 and S3.

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

160

161 **Structural analysis of charged residues in transmembrane helices**

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

169 As seen in Figure 2, charged pairs of amino acids are especially prevalent at
170 positions *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4. Oppositely charged residues stand out, especially Glu-Arg
171 at *i*, *i*+1, Glu-Lys at *i*, *i*+3 and Asp-Lys at *i*, *i*+4. Also, several same charged pairings at *i*,

172 *i*+3 and *i*, *i*+6 are more frequent than expected. Other known structural features can also
173 be hinted at, including aromatic ring stacking by His-Trp pair [20] at *i*, *i*+6.
174 Charged pairs placed at *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4 could potentially form salt bridges
175 as they are all on the same relative face of the alpha helix and are close enough in vertical
176 separation on the helix (see Figure 3, top). Although oppositely charged pairs at *i*, *i*+7 are
177 also on the same face of the alpha helix, unless the alpha helix has a bend, both residues
178 are too far separated to form a salt bridge. This was clearly seen in Figure 3 where the TM
179 dataset was used. In both absolute count (Fig. 3 panel a) and log odds ratios (Fig. 3 panel
180 b) it is clear that residues at *i*, *i*+1; *i*, *i*+3 and *i*, *i*+4 are by far the most common and
181 overrepresented pairings. Figure 3b also shows that oppositely and same charged pairs
182 have about the same overrepresentation at *i*, *i*+3, whereas oppositely charged pairs are
183 stronger than same charged pairs at *i*, *i*+1and *i*, *i*+4, both within salt bridge range, and
184 same charged pairs are stronger at positions *i*, *i*+7 and *i*, *i*+8, too far to form salt bridges.
185 When structure-observed salt bridges in the different positions were compared to
186 the oppositely charged pairs a clear image aroused, see Figure 3c. Even though there are
187 more oppositely charged pairs at position *i*, *i*+1 than in positions *i*, *i*+3 and *i*, *i*+4 with all
188 log odds ratios over 1.0 (Fig. 3b), only about 15% of the oppositely charged pairs at *i*, *i*+1
189 form salt bridge (Fig. 3c). This is in contrast to *i*, *i*+3 where almost 40 % of the pairs form
190 salt bridges, and just under 25% at *i*, *i*+4 (Fig. 3c).
191
192 **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 (183 protein chains with a total of 188 salt bridges).

When we now make a selection for potential candidates, we also look at the estimated ΔG_{pred} values and choose helices with a positive value above one, as per our hypothesis, a salt bridge helps TM insertion of sequences for which the hydrophobic force would not be enough to insert. This stricter definition results in a set of 136 candidates of a total of 138 salt bridges and we selected TM7 (helix G) from Halorhodopsin protein (PDB ID: 3QBG) with an estimated ΔG_{pred} value above +1.7 kcal/mol, and helix A from calcium ATPase (PDB ID: 1SU4) with a higher estimated ΔG_{pred} value (above 4.1 kcal/mol), as candidates for systematic studies. See the github repository (https://github.com/ElofssonLab/salt_bridges) for the full lists.

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

Halorhodopsin (hR) from *Natronomonas pharaonis* (3QBG) is a protein made up of seven TM helices (helix A through G) and a retinal chromophore that is bound via a protonated Schiff base to the ε-amino group of a lysine (K258) residue located roughly in the middle of helix G [21]. *In silico* analysis of 3QBG structure, the anion-free form of the protein, showed a charged pair of amino acids (Asp-Lys), involving the

215 functional K258 and the D254 in the center of helix G (Fig. 4). The position (*i*, *i*+4) and
216 the distance between the anionic carboxylate ($RCOO^-$) from the Asp residue and the
217 cationic ammonium (RNH_3^+) from the lysine residue, in the crystal structure, was about
218 3.5 Å, a permissive distance for a salt bridge formation (Fig. 4g), which has been
219 established as being lower than 4 Å [22]. To get insights into this interaction, we designed
220 three mutants that were supposed to perturb the salt bridge interaction by different ways:
221 K258D mutant by placing two charged residues with the same polarity at positions *i*, *i*+4;
222 K258A mutant by replacing one of the charged residues by a non-polar amino acid, and
223 K258Y/Y259K double mutant by placing the charged pair at a non-permissive salt bridge
224 position (*i*, *i*+5; Fig. 4g), while keeping the same amino acid composition (Fig. 4a).

225 hR is a trimeric protein in which the helix G is neither exposed at the monomer-
226 monomer interface nor oriented to the inner part of the trimeric structure (Fig. 4, panels
227 b-d). In fact, the charged pair found in helix G is oriented toward the core of the 'globular'
228 structure in each monomer. Then, the insertion of helix G was studied using the Lep-
229 based assay. When constructs harboring helix G wild type sequence were translated *in*
230 *vitro* in the presence of RMs singly-glycosylated (reporting insertion) forms were found
231 (Fig. 4e, lane 2), despite its positive (suggesting non-insertion) ΔG_{pred} value (Fig. 4a). The
232 nature of the higher molecular weight polypeptide species was analysed by
233 endoglycosidase H (EndoH) treatment, a highly specific enzyme that cleaves N-linked
234 oligosaccharides. Treatment with EndoH of the samples eliminated higher molecular
235 mass bands (Fig. 4a, lane 1), confirming the sugar source of the retarded electrophoretic
236 mobility bands and suggesting helix G insertion into the microsomal membrane.

237 However, locating the Asp-Lys pair at *i*, *i*+5, which is non-compatible with salt bridge
238 interaction (Fig. 4g), strongly reduced the experimental insertion efficiency (Fig. 4e, lane
239 5). Interestingly, replacing the positively charged lysine residue by a negatively charged
240 aspartic acid residue (K258D) rendered similarly low levels of insertion efficiency (Fig.
241 4e, lane 4). As expected, replacement of the ionizable lysine residue by the aliphatic
242 alanine increased the insertion efficiency (Fig. 4e, lane 3). In this later mutant, the most
243 likely cause is the absence of the positively charged lysine amino side chain and by the
244 presence of the methyl side chain group of the mutant alanine residue. The results of the
245 Lep-based glycosylation assay indicated that wild type (wt) and DA sequences (two
246 stabilized charges or only one charge, respectively) are inserted properly into the
247 microsomal membrane (ΔG_{exp} values -0.24 and -0.88 kcal/mol, respectively), but when
248 the salt bridge is disrupted, either by having two charged amino acids with the same
249 polarity (DD) or by placing oppositely charged residues at a non-permissive distance (*i*,
250 *i*+5) in the center of the helix, the translocation of the segment increases substantially. It
251 should be mentioned that the K258Y/Y259K double mutant has the same amino acid
252 composition than the original helix G, but insertion efficiency is remarkably decreased
253 (ΔG_{exp} = +0.31 kcal/mol). Together these results show that the interaction (salt bridge)
254 between Asp and Lys residues in the center of the helix G from 3QBG is essential for its
255 proper insertion into the microsomal membrane. The salt bridge contributes
256 approximately ~0,5 kcal/mol to the apparent experimental free energy of membrane
257 insertion.

258 Next, to ensure that the *in vitro* results are relevant to the *in vivo* situation, wt
259 and *i*+5 constructs were also expressed *in vivo* in HEK-293T cells. To this end, a *c*-myc
260 tag was engineered at the C-terminus of the Lep chimera to allow immune-detection of
261 our constructs in cell extracts. As shown in Fig. 4f, transfected cells with the chimera
262 containing helix G wt sequence rendered singly-glycosylated molecules, indicating *in*
263 *vivo* membrane insertion. In contrast, cells transfected with the construct harboring *i*,
264 *i*+5 sequence rendered almost exclusively doubly-glycosylated forms, as proved by
265 EndoH treatment (Fig. 4f, lane 2), suggesting membrane translocation. These results
266 emphasized the relevance of salt bridge interactions in translocon-mediated TM
267 insertion, especially in the *in vivo* (cellular) environment.
268
269 **Salt bridge contribution to the insertion of a heavily charged helix.**
270 In order to challenge salt bridge interactions in a more hydrophilic TM helix we focus on
271 helix A from the sarcoplasmic/ER calcium ATPase 1 [23]. Calcium ATPase (PDB ID:
272 1SU4, *Oryctolagus cuniculus*) is a member of the P-type ATPases that transport ions
273 across the membrane against a concentration gradient involving 10 α-helices (helix A
274 through J) in the membrane-embedded region [23]. *In silico* analysis of 1SU4 structure,
275 the crystal structure of the calcium ATPase with two bound calcium ions, showed Asp-
276 Arg (DR) pair (D59 and R63) in the center of the helix A. This helix extends beyond the
277 membrane (Fig. 5a, from Leu49 to Phe78) and shows some particularities. On the one
278 hand, in the structure the membrane-embedded stretch (the N-terminal region of helix
279 A) encompasses from Leu49 to Ala69 residues, and includes several charged amino acids,

280 probably involved in the Ca^{2+} transport across the membrane (Glu51, Glu55, Glu58,
281 Asp59 and Arg63). Therefore, the ΔG_{pred} value for this segment (L49-A69) is remarkably
282 higher (and positive) than expected for a TM helix (Fig. 5b). On the other hand, the C-
283 terminal region of this helix contains a high prevalence of non-polar amino acids that is
284 more compatible with the hydrophobicity of the membrane core (Fig. 5b). Nevertheless,
285 the presence of the functional amino acids (e.g. Glu58) in the membrane-embedded
286 region reinforce the idea that the more hydrophilic N-terminal region must ultimately
287 be embedded within the membrane [23]. It has been previously shown that the position
288 in the membrane of TM helices in protein folded structures does not always correspond
289 to the thermodynamically favored positions in the membrane of the isolated helices [24].
290 Instead, after translocon-mediated insertion of the more hydrophobic region,
291 repositioning of TM helices relative to the lipid bilayer provides a convenient way for
292 non-hydrophobic polypeptide segments to become buried within the membrane. Then,
293 the nature of helix A suggested the possibility that initial insertion of the hydrophobic
294 region can be followed by subsequent repositioning of the charged region into the
295 membrane hydrophobic core, which will be the final segment embedded in the lipid
296 bilayer (Fig. 5a). Interestingly, the ΔG Predictor server
297 (https://dgpred.cbr.su.se/index.php?p=home) selected the adjacent (C-terminus, L60-
298 F78) hydrophobic region as TM (Fig. 5b), instead of the charged region found at the high-
299 resolution structure (Uniprot code: P04191).

300 Focusing on the potential salt bridge residues (D59 and R63 pair), the distance
301 between the anionic carboxylate ($RCOO^-$) from the D59 and the cationic guanidinium

302 ($RC(NH_2)_2^+$) from the R63 was about 3.0 Å in the crystal structure (Figure 5a), clearly
303 within the permissive range for salt bridge interaction. To investigate the contribution of
304 this potential salt bridge interaction in the translocon-mediated membrane insertion of
305 this region, we worked with two different scaffold sequences: the full helix A involving
306 the residues 49-78 (Long, L); and a shorter membrane-embedded version including
307 residues 49-69 (Short, S) as found in the solved structure. We also challenged the D59-
308 R63 charge pair interaction in both sequences by increasing the separation between the
309 ionizable residues from the native *i*, *i*+4 to non-permissive *i*, *i*+5, while maintaining amino
310 acid composition (R63I/I64R double mutant). *In vitro* transcription/translation of these
311 sequences in the presence of microsomes rendered singly-glycosylated molecules for the
312 construct containing full-length helix A (Fig. 5c, lane 2). In contrast, when only the
313 membrane-embedded sequence was included, the Lep chimera was mainly doubly-
314 glycosylated (Fig. 5c, lane 3), suggesting that helix A is properly inserted when the full
315 helical sequence is present. Still, the translocon inefficiently inserted the isolated
316 membrane-embedded region. When the charge paired residues were placed at a non-
317 permissive distance in terms of salt bridge interaction (*i*, *i*+5; Fig. 5e) the insertion
318 efficiency was reduced (Fig. 5c, lane 4), with a ΔG_{app} decrease (absolute values) of ~0,7
319 kcal/mol relative to the wild type sequence (Fig. 5b). As expected, this effect was not
320 observed when the same mutations were grafted on the membrane-embedded (S)
321 sequence (Fig. 5c, lane 5).

322 Finally, we analysed the salt bridge interaction in HEK-293T cells to study the
323 translocon performance *in vivo*. When cell cultures were transfected with a Lep-derived

324 chimera containing the helix A wild type sequence only singly-glycosylated molecules
325 were observed (Fig. 5d, lanes 1 and 2). However, a construct harboring double mutant
326 (R63I/I64R; *i*, *i*+5) sequence showed doubly-glycosylated molecules to a measurable
327 extent (Fig. 5d, lanes 3 and 4), indicating a lower insertion efficiency that can be
328 attributed to the altered salt bridge interaction.
329

330 **DISCUSSION**

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

340 Whereas pairs of charged residues of the same charge facing the inside pores in
341 TM regions fill an essential functional role, pairs of oppositely charged residues can form
342 salt bridges that can stabilize the helix and play an important role in function such as for
343 transportation [29]. Salt bridges might also be important for hiding the charges during
344 translocon-mediated TM helix insertion, as the charged residues are hidden from the
345 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.

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

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

By analyzing two native helices containing intra-helical salt bridges we now find that the free energy of insertion (ΔG_{app}) is significantly reduced if both oppositely charged

368 residues are spaced at a permissive distance. These results indicate that intra-helix salt
369 bridge can form during translocon-assisted insertion or even earlier, since in contrast to
370 globular (soluble) helices, TM helices can be compacted inside the ribosome exit tunnel
371 [31]. The maximal reduction in ΔG_{app} seen with Asp-Lys and Asp-Arg pairs in both hR
372 and Ca ATPase helices is 0.5-0.7 kcal/mol, which is in good agreement with the 0-1
373 kcal/mol estimated for these two pairs from thermodynamic peptide partition into
374 octanol experiments [32]. As found in the case of hR helix G (Fig. 4f), this reduction might
375 be even higher in the cell context, since some auxiliary components of the membrane
376 insertion machinery [15,33,34] can be in suboptimal conditions in the microsomal
377 vesicles.

378 Beyond the conceptual issues involving the membrane insertion process, we note
379 that the availability of quantitative experimental data on the contribution of salt bridge
380 interactions to the free energy of insertion (ΔG_{app}) will make it possible to fine-tune
381 current membrane protein topology-prediction methods based on free energy
382 calculations.

383

MATERIALS AND METHODS

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

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

407 **Translocon-mediated insertion into microsomal membranes.** Lep constructs in pGEM
408 with L4/A15, 3QBG and 1SU4 segments and its variations were transcribed and translated
409 using the TNT T7 Quick Coupled System (#L1170, Promega). Each reaction containing 1
410 μL of PCR product, 0.5 of EasyTag™ EXPRESS 35S Protein Labeling Mix (Perkin Elmer)
411 (5.5 μCi) and 0.3 μL of microsomes (tRNA Probes) was incubated at 30ºC for 90 min. Endo
412 H treatment was done following the manufacturer's instructions. Samples were analysed
413 by SDS-PAGE. The bands were quantified using a Fuji FLA-3000 phosphoimager and the
414 Image Reader 8.1j software. Free energy was calculated using: $\Delta G_{app}=-RT \ln K_{app}$, where
415 $K_{app}=f_{2g}/f_{1g}$ being f_{1g} and f_{2g} the fraction of singly glycosylated and double glycosylated
416 protein, respectively.
417 **Free apparent insertion energy, ΔG.** The free insertion energy of a TM region, ΔG, is
418 calculated as per the experimentally defined Biological hydrophobicity scale [12]. This
419 scoring is amphiphilic with hydrophobic residues contributing a lower (negative) ΔG
420 while hydrophilic contributes a higher (positive) ΔG. A region's total ΔG is the sum of
421 these individual scores weighted on where in the region the residue, a residue in the
422 middle of the helix has a higher weight than residues at the ends. This scoring can give
423 an indication of how favourable the amino acid composition of a TM region is to be
424 inserted in a lipid bilayer membrane. To note is that hydrophobicity alone is not the only
425 driving force and that the positive inside rule [35,36] and help from proceeding TM
426 regions [37,38] can help insertion in polytopic TM proteins especially.
427 **Expression in mammalian cells.** Lep constructs containing 3QBG or 1SU4 segments and
428 its variations were tagged with c-myc epitope at their Ct (Glu-Gln-Lys-Leu-Ile-Ser-Glu-

429 Glu-Asp-Leu, EQKLISEEDL) and inserted in a pCAGGS-ampicillin plasmid. Once the
430 sequence was verified, plasmids were transfected into HEK293-T cells using
431 Lipofectamine 2000 (Life Technologies) according to the manufacturer's protocol.
432 Approximately 24 h post-transfection cells were harvested and washed with PBS buffer.
433 After a short centrifugation (1000 rpm for 5 min on a table-top centrifuge) cells were
434 lysed by adding 100 μL of lysis buffer (30 mM Tris-HCl, 150 mM NaCl, 0.5% Nonidet P-
435 40) were sonicated in an ice bath in a bioruptor (Diagenode) during 10min and
436 centrifugated. Total protein was quantified and equal amounts of protein submitted to
437 Endo H treatment or mock-treated, followed by SDS-PAGE analysis and transferred into
438 a PVDF transfer membrane (ThermoFisher Scientific). Protein glycosylation status was
439 analysed by Western Blot using an anti-c-myc antibody (Sigma), anti-rabbit IgG-
440 peroxidase conjugated (Sigma), and with ECL developing reagent (GE Healthcare).
441 Chemiluminescence was visualized using an ImageQuantTM LAS 4000mini
442 Biomolecular Imager (GE Healthcare).

443 **Core segment definition.** We define core segments as a transmembrane region minus
444 the first and last 5 residues. This is to ignore the interface regions which are known to
445 contain polar residues.

446 **Salt bridge definition.** Salt bridges are defined as per [39], where a salt bridge is defined
447 if a side chain carbonyl oxygen atom in Asp-Glu is within 4.0 Å from the nitrogen atom
448 in Arg-Lys. This conforms to other works [7,40] with the definition that the atoms are
449 within hydrogen bond distance. We also define local salt bridges as being bridges that
450 are separated by at most 7 residues in the sequence. This is to separate long salt bridge

451 interactions, which can occur between spatially close residues that are separated in
452 sequence, such as coiled coils where salt bridges can be between separate alpha-helices.
453 **Transmembrane helices dataset (TM dataset).** The full pipeline is available as a Makefile
454 together with supporting scripts in the github repository. The full PDBTM database [16]
455 was downloaded together with their list of non-redundant protein pdb ids. This list is
456 used to generate both sequence and topology of the proteins by extracting both from the
457 PDBTM-xml. For each protein in the non-redundant list, the membrane regions are
458 extracted as per the PDBTM database annotation. All non-membrane regions are
459 annotated 'i' for convenience. To support future analysis, membrane regions longer than
460 10 were run through DSSP and annotated with 'M' if all residues in the core segment
461 were defined as alpha-helix ('H'), otherwise, the full membrane region is annotated 'm'.
462 Observe that this creates fasta-like 3line files, that only contain topological annotation
463 with ambiguous TM regions annotated as 'm' instead of the normal 'M'. These proteins
464 are then clustered using cd-hit [41,42] at 40% identity using the parameter -c 0.4 -n 2 -T
465 0 -M 0 -d 0.
466 During the extraction of TM regions, the corresponding structure file from RCSB
467 was used to calculate all salt bridges within the current protein and any salt bridge that
468 has at least one residue within any TM region was saved. Additionally, all salt bridges
469 whose both residues were within the same segment and within 7 residues of each other
470 were annotated as local as per the salt bridge definition above.
471 **Extraction of charged residues.** From all annotated TM regions of length 17 or longer [2],
472 the core segment was extracted. All these core regions were then scanned and when a

473 charged residue was encountered, we recorded any other charged residue from 7 residues
474 before the current one to 7 residues after. This results in charged residues that can contain
475 a charged pairing partner outside of the TM segment and will therefore differ slightly
476 from charged pairs which are defined next.

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

483 **MSA-dataset extension (TM-MSA).** Using the TM dataset, we extended it by creating an
484 MSA alignment of each protein using jackhmmer [43] against Uniref90 with one iteration
485 and an E-value cut off of 10^{-3} with the following parameters:
486 -N 1 -E 1e-3 --incE 1e-3 --cpu 14

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

492 **Dataset of helices from globular proteins (GLOB and GLOB-MSA).** To create a reference
493 dataset of globular α-helical proteins we extracted all globular all-alpha protein domains
494 from SCOP. As SCOP classifies domains of proteins resulting in that one domain of a

495 protein can be annotated as globular whereas another domain as TM (see 1PPJ chain D
496 as an example) we reduced the SCOP list against the redundant list of all PDBTM [16]
497 chains to clear any overlap. This results in 4,500 proteins in total.

498 Topological files with sequence and membrane topology are created with the help
499 of the RCSB secondary structure file and only membranes annotated as pure α-helices
500 were retained. This file was further homology reduced and alignments prepared in the
501 same manner and using the same parameters as the TM dataset described above.

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

505 **Redundant (TM-Red) dataset**. The full PDBTM database was used as in the preparation
506 of the TM dataset. In this case we used all redundant sequences and generate topology
507 files, but we did not clustered the proteins. We added in the constraint that each selected
508 TM region must fully contain at least one potential salt bridge. This means a local salt
509 bridge where both residues are within the core segment. This dataset was only used to
510 find potential candidates for further *in vitro* and *in vivo* experiments.

511

512

513 **Acknowledgments**

514 We thank Pilar Selvi for excellent technical assistance. This work was supported by grants
515 BFU2016-79487-P form the Spanish Ministry of Economy and Competitiveness and
516 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.

523 **Table 1**. About 10 % of the core residues in the TM dataset are polar and just under 4%
524 are charged residues. This is significantly less than 37.5% and 25.5% respectively in the
525 globular dataset. In the transmembrane dataset, just over half of the charged core residues
526 form a charged pair at distance 1, 3 or 4. *In the case of the globular set, there are more
527 potential charged pairs than charged core residues as multiple residues are counted more
528 than once as they form more than one potential pairing with separation of 1, 3 or 4 steps.
529 It is clearly seen that although the TM dataset contains more helices per protein and has
530 a higher proportion of long (≥ 17 residues) helices it contains significantly fewer charged
531 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	265	34029	4251	344142

pairs				
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*	-

532

533

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*	*$0.00e^{+00}$*	***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*	*$1.97e^{-10}$*

546 **FIGURE LEGENDS**

547

548 **Figure 1. Effects on membrane insertion of single or pairs of Asp and Lys residues in a**
549 **model TM segment. (a)** Schematic representation of the leader peptidase (Lep) model
550 protein. G1 and G2 denote artificial glycosylation acceptor sites. The sequence under
551 investigation was introduced in the P2 region after H2. Recognition of the tested
552 sequence as a TM by the translocon machinery (highlighted in green) results in the
553 modification of the G1 site but not G2. The Lep chimera will be double glycosylated if
554 the sequence being tested is not recognized as a TMD and thus translocated into the
555 microsomes lumen (shown in red). **(b)** The tested sequences from L4/A15 model TM
556 (including the charged residues, bold), the gap distance, and the predicted ΔG (ΔG_{pred})
557 values in kcal/mol are shown. Amino acids with positive and negative charge are
558 highlighted in blue (K) and red (D) respectively. **(c)** Experimental ΔG (ΔG_{exp}) in kcal/mol
559 of each tested sequence in the Lep assay. The mean and standard deviation of at least 3
560 independent experiments are represented (*n* values: 4 [from 1 to 7] and 3 [8 and 9]). The
561 individual value of each experiment in represented by a solid dot being represented as a
562 green square the experimental ΔG value for the L4/A15 sequence from [2]. The wt and
563 simple mutants are shown in white bars. Charges at compatible distances with salt bridge
564 formation (*i*, *i*+1; *i*, *i*+3; and *i*, *i*+4) are shown in yellow. Not compatible distances with
565 salt bridge formation (*i*, *i*+2; and *i*, *i*+5) are shown in dark gray. Compatible distances but
566 not compatible amino acid pair (*i*, *i*+4 DD pair) is shown in clear gray.

567

568 **Figure 2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the TM-**
569 **MSA dataset.** The rows on the y-axis indicate the first amino acid in the pair and the
570 columns on the x-axis the second. The residues are ordered by hydrophobicity according
571 to the Engelman order [44]. See figure S1 for the equivalent of the globular dataset. S2
572 and S3 show the same plots masked for statistical significance.
573
574 **Figure 3. Charge pairs in TM helical sequences and structures.** Helical wheel projection
575 and lateral views of an α-helix are shown on top. The initial position *i* and the following
576 8 residues are numbered. Residues in positions *i*+3 (orange), *i*+4 (yellow), and *i*+7 (light
577 brown) are mainly on the same face of the helix, but *i*+7 is placed too far for a salt bridge
578 interaction. **a)** The absolute count of charged pairs for '*i*, *i*+1' through '*i*, *i*+8' in the TM
579 dataset. The forward slash are the oppositely charged pairs and backward slash are the
580 same charged pairs. **b)** The log odds ratios of the same data for all charges, oppositely
581 charged and same charged pairs, all with error bars. **c)** Fraction of oppositely charged
582 pairs that form local salt bridges. The small bump at *i* +5 are the two proteins 6CC4 (helix
583 A) and 5MG3 (helix E), which both exhibit a bend in the alpha helix due to the presence
584 of glycine residues (see Fig. S4).
585
586 **Figure 4. Insertion of halorhodopsin helix G from *Natronomonas pharaonis* (3QBG) into**
587 **microsomal and cellular membranes. (a)** Tested sequences from 3QBG including the gap
588 distance, the distance between charges in Å, and the predicted and experimental ΔG
589 values (ΔG_{pred} and Δg_{exp} respectively) in kcal/mol are shown. Amino acids with a positive

590 or negative charge are highlighted in blue (K) and red (D), respectively. Green numbers
591 indicate negative ΔG (insertion) values, while red numbers denote ΔG values above 0
592 (translocation). **(b)** Frontal view of 3QBG monomer structure. The helix G is highlighted
593 in orange with the D254 and K258 shown in sticks colored red and blue respectively. The
594 membrane position is indicated by a red (outer) and blue (inner) discontinuous line,
595 according to OPM dataset [22]. Lateral **(c)** and upper **(d)** views of the 3QBG trimeric
596 structure. The helix G is highlighted in orange with the D and K shown in sticks colored
597 red and blue, respectively. The different monomers are shown in transparent blue, purple
598 and indigo. Representative examples (*n*=3) of *in vitro* protein translations in the presence
599 of ER-derived microsomes **(e)** and Western blots (*n*=3) of *in vivo* protein translations in
600 HEK-293T cells **(f)** in the presence (+) or absence (-) of Endoglycosidase H (EndoH), a
601 glycan-removing enzyme. The absence of glycosylation of G1 and G2 acceptor sites is
602 indicated by two white dots, single glycosylation by one white and one black dot, and
603 double glycosylation by two black dots. **(g)** Zoom view centered on the salt bridge
604 between D254 and K257 at *i*, *i*+4 (left) and *i*, *i*+5 (right) gaps. D and K residues are shown
605 in sticks colored red and blue, respectively, while the dashed line indicates the $RCOO^-$ to
606 RNH_3^+ distance.
607
608 **Figure 5. Insertion of Calcium ATPase (1SU4) helix A into microsomal and cellular**
609 **membranes. (a)** Lateral view of 1SU4 structure. Zoom view of the A helix (right panel).
610 The membrane-embedded region of helix A is highlighted in yellow. Charged amino
611 acids are shown as sticks in blue (R), red (D) and pink (E), respectively. L49, A69 and F78

612 are also shown as sticks to define helix's subdomains. The membrane location is indicated
613 by red (outer) and blue (inner) discontinuous lines according to OPM dataset [22] and
614 the distance between the R and D charges is indicated in Å. **(b)** Helix A-derived sequences
615 from 1SU4 including the gap between charged residues, the distance between charges in
616 Å, and the predicted and experimental ΔG values (ΔG_{pred} and Δg_{exp} respectively) in
617 kcal/mol are shown. Amino acids with a positive charge are highlighted in blue (K) while
618 negatively charged are marked in red (D) and pink (E). Green numbers indicate negative
619 ΔG (insertion) values while red numbers denote ΔG values above 0 (translocation).
620 Representative examples (n=3) of *in vitro* protein translations in the presence of ER-
621 derived microsomes **(c)** and Western blots (n=3) of *in vivo* protein translation in HEK-
622 293T cells **(d)** in the presence (+) or absence (-) of Endoglycosidase H (EndoH), a glycan-
623 removing enzyme. The absence of glycosylation of G1 and G2 acceptor sites is indicated
624 by two white dots, single glycosylation by one white and one black dot, and double
625 glycosylation by two black dots. **(e)** 1SU4 helix A *i*, *i*+5 mutant. The membrane-embedded
626 region of helix A is highlighted in yellow. Charged residues are shown as sticks in blue
627 (R), red (D) and pink (E). L49, A69 and F78 are also shown as sticks to define helix's
628 subdomains. The membrane is indicated by red (outer) and blue (inner) discontinuous
629 lines as in (a), and the dashed line indicates the $RCOO^-$ and $RC(NH_2)_2^+$ distance.
630
631 **Figure S1. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the**
632 **GLOB-MSA dataset.** Log odds ratios for the middle core residues of α-helices of at least
633 17 residues in length in the GLOB-MSA dataset. The rows on the y-axis indicate the first

634 amino acid in the pair and the columns on the x-axis the second. The residues are ordered

635 as in Fig. 2.

636

637 **Figure S2. Log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8' for the**

638 **GLOB-MSA dataset.** Log odds ratios as in S1 but all pairs with P-value >0.05 have been

639 masked.

640

641 **Figure S3. Masked log odds ratios of each pair of amino acids for '*i*, *i*+1' through '*i*, *i*+8'**

642 **for the TM-MSA dataset.** Log odds ratios as in Fig. 1 but all pairs with P-value >0.05 have

643 been masked.

644

645 **Figure S4. Salt bridges at pair *i*, *i*+5.** Left: helix E from bacterial translocon (PDB ID:

646 5MG3). Glycine (orange) residue causes a kink facilitating salt bridge interaction between

647 Asp50 (red) and Arg55 (blue). Right: helix A from a lipid flippase (PDB ID: 6CC4).

648 Glycine (orange) residue causes a kink facilitating salt bridge interaction between Arg153

649 (blue) and Glu158 (red). Distances are shown in Ångström.

650

651 **REFERENCES**

652

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

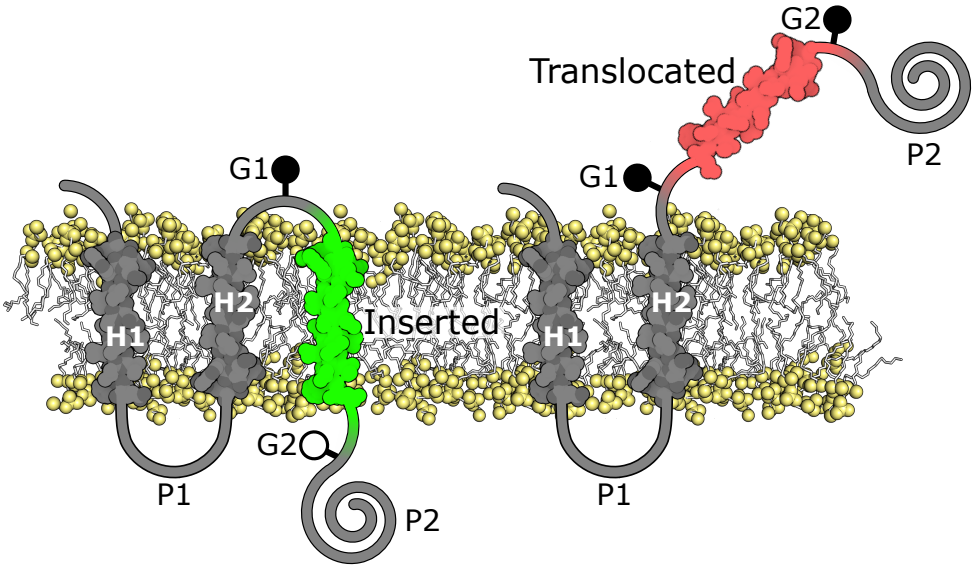
3. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* **43**: W401–W407.
4. Tsirigos KD, Govindarajan S, Bassot C, Västermark Å, Lamb J, Shu N, Elofsson A (2018) ScienceDirect Topology 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. Heijne von G (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341**: 456–458.

36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. Li W, Godzik A (2006) *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*.
43. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.
44. 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	$\Delta G_{pred.}$
L4/A15	-	AAAALALAAAAALALAAAA (8, 12)	-0.490
K	-	AAAALALKAAAALALAAAA	1.414
D	-	AAAALALAAAADLALAAAA	1.389
KD	1	AAAALALAAAKDLALAAAA	3.222
KD	2	AAAALALAAKADLALAAAA	2.927
KD	3	AAAALALAKAADLALAAAA	3.349
KD	4	AAAALALKAAADLALAAAA	3.339
KD	5	AAAALAKLAAADLALAAAA	2.980
DD	4	AAAALALDAAADLALAAAA	3.548

c)

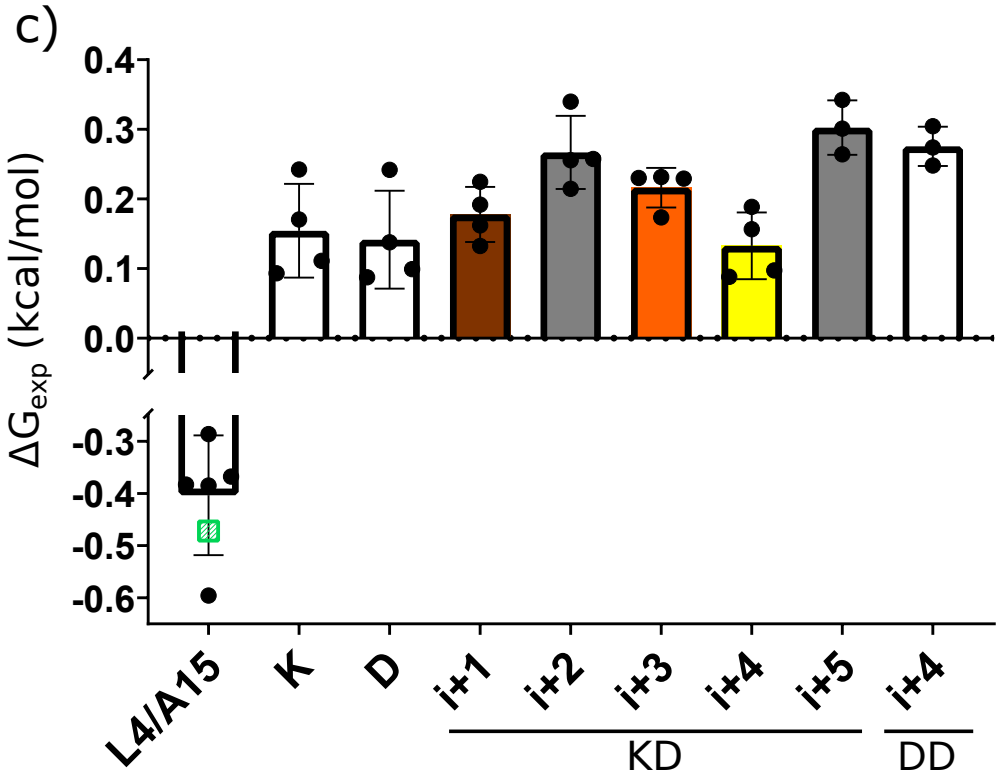


Figure 2

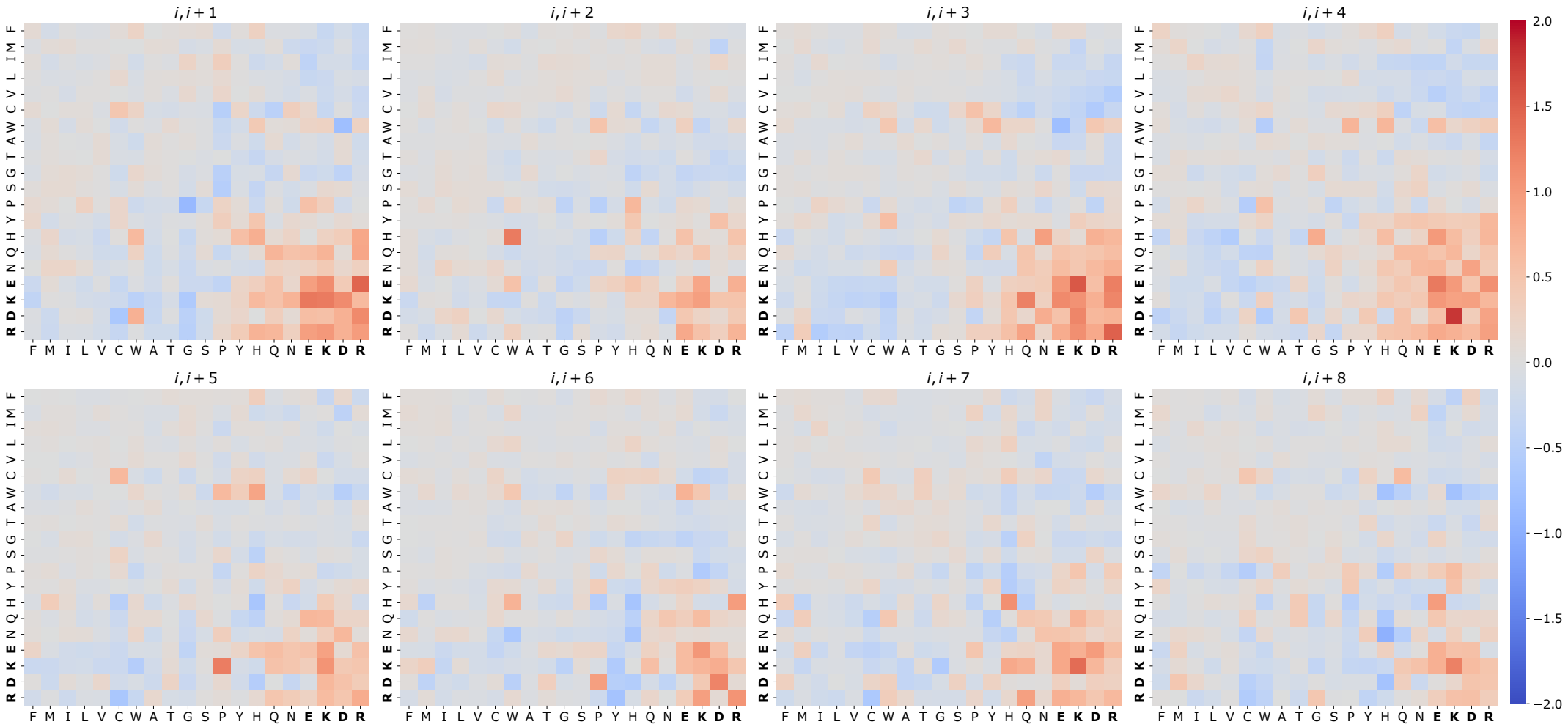

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figura 3

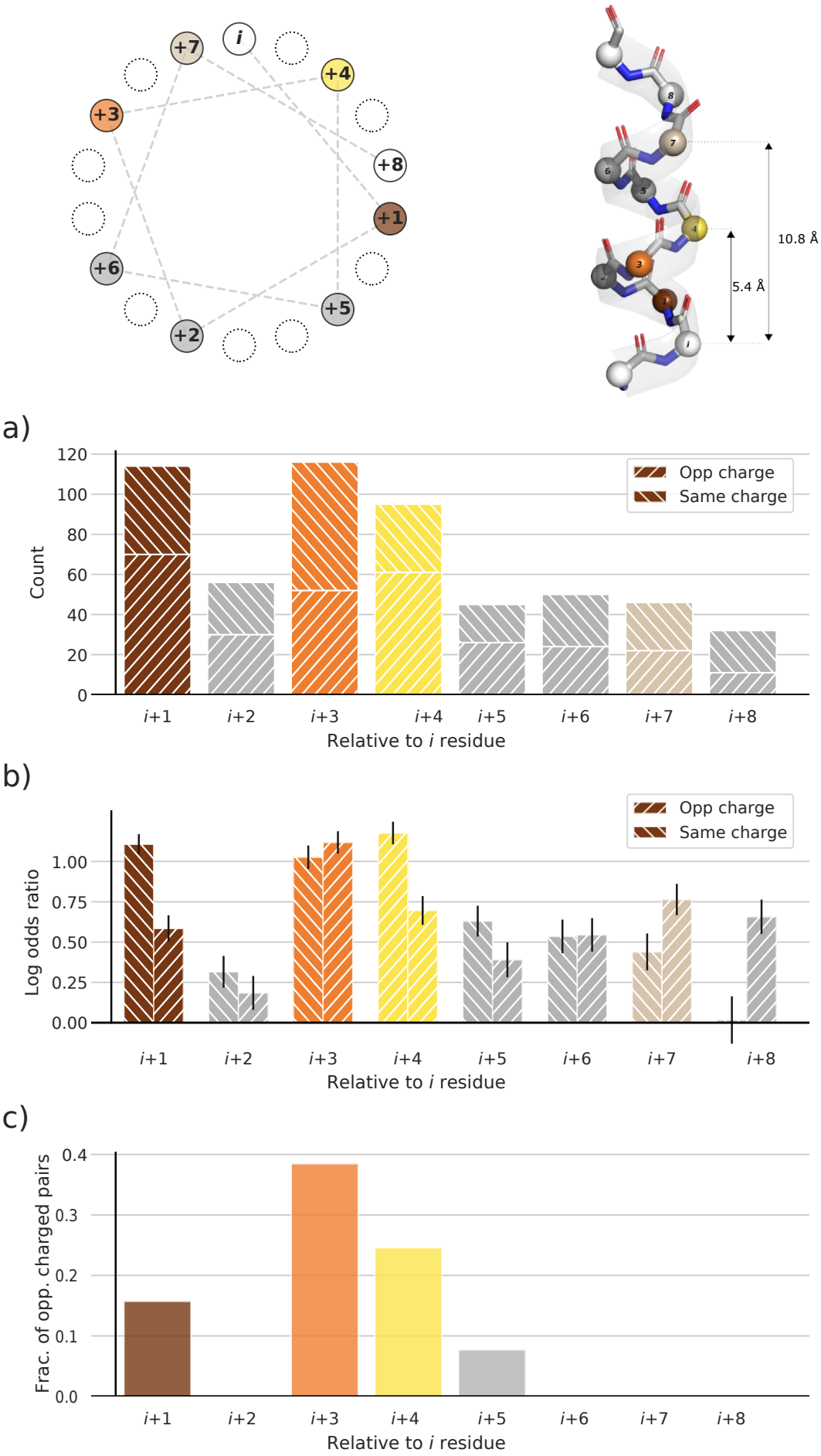
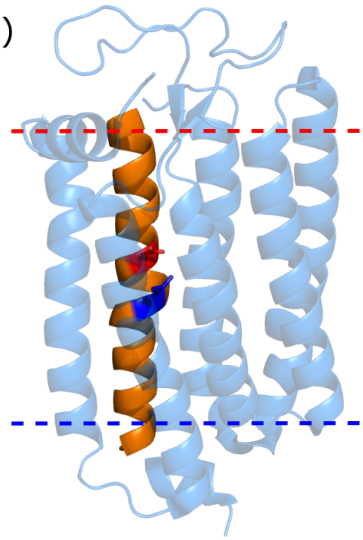

i
+1
+2
+3
+4
+5
+6
+7
+8
10.8 Å
5.4 Å
a)
Count
Opp charge
Same charge
i+1
i+2
i+3
i+4
i+5
i+6
i+7
i+8
Relative to i residue
b)
Log odds ratio
Opp charge
Same charge
Relative to i residue
c)
Frac. of opp. charged pairs
Relative to i residue

Figure 4

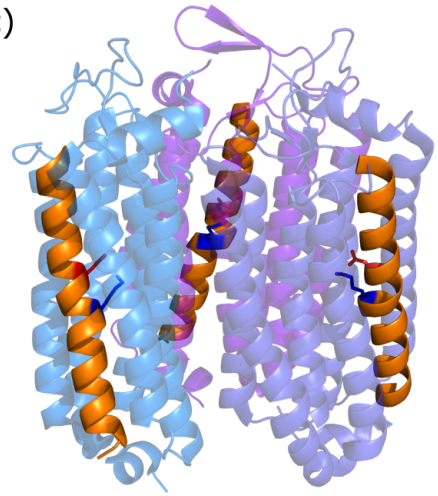
a)

Name	Sequence	Charges	Gap	Å	$\Delta G_{pred.}$	$\Delta G_{exp.}$
wt	243 254 258 271 VGVTSWGYSFL**D**IVA**K**YIFAFLLLNYLTS	DK	4	3.5	+1.728	-0.243
DD	VGVTSWGYSFL**D**IVA**D**YIFAFLLLNYLTS	DD	4	-	+1.664	+0.112
DA	VGVTSWGYSFL**D**IVA**A**YIFAFLLLNYLTS	DA	4	-	-0.376	-0.880
DK *i*, *i*+5	VGVTSWGYSFL**D**IVAY**K**IFAFLLLNYLTS	DK	5	-	+1.252	+0.307

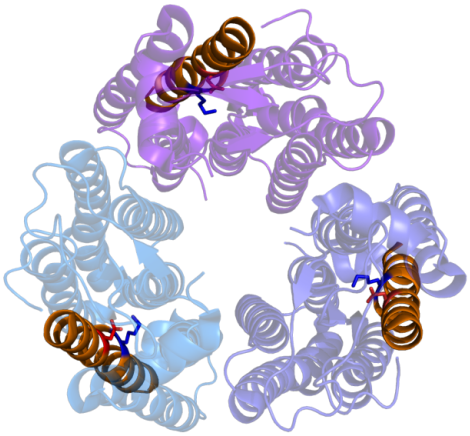
b)



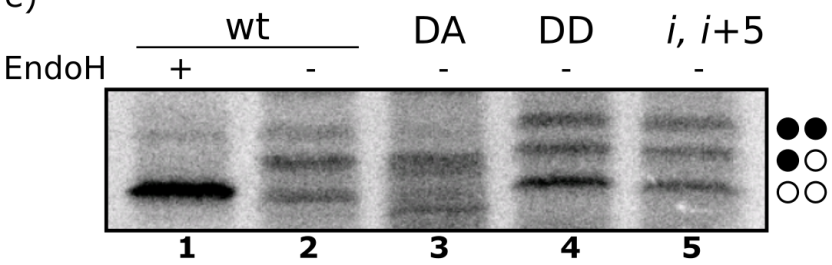
c)



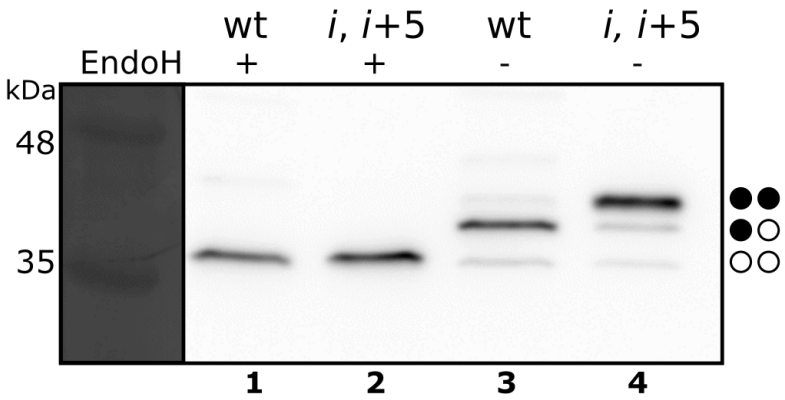
d)



e)



f)



g)

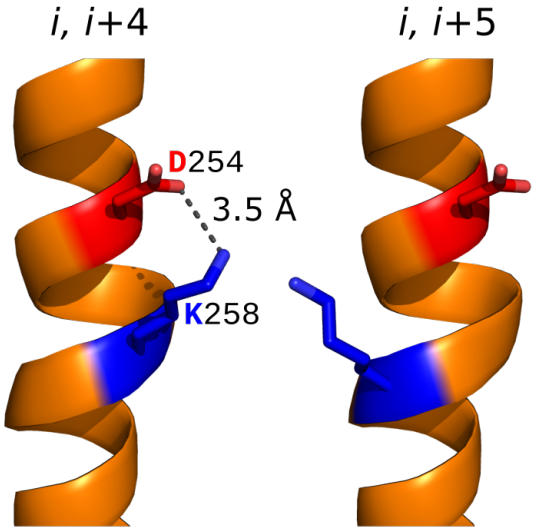
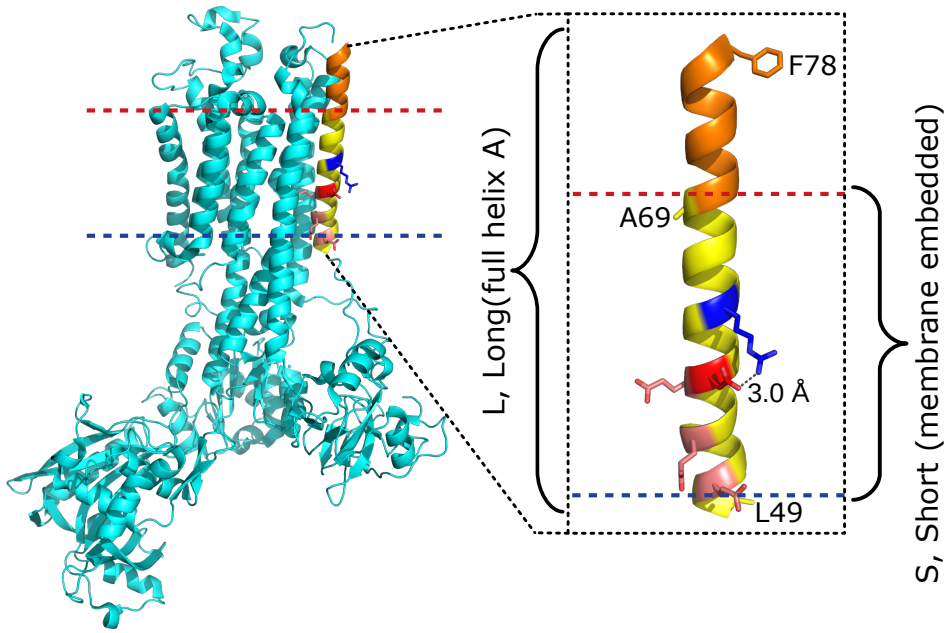


Figura 5

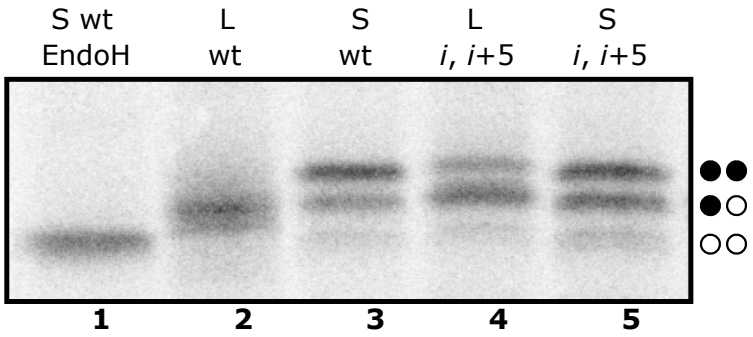
a)



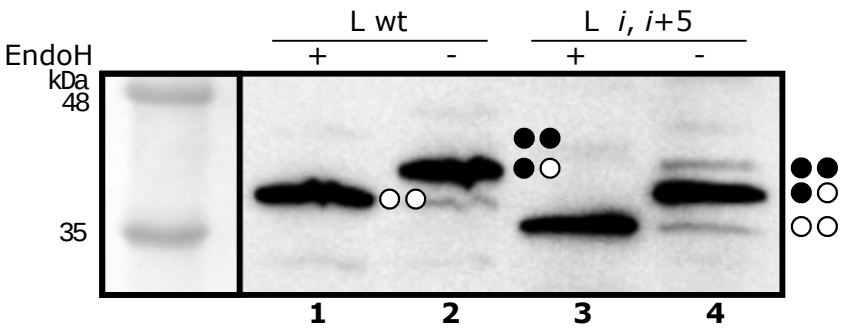
b)

Name	Sequence	Gap	(Å)	$\Delta G_{pred.}$	$\Delta G_{exp.}$
L (L49-F78)	49 59 63 78 LWELVIEQFEDLLVRILLLAACISFVLAWF	4	3.0	+1.063	-1.050
S (L49-A69)	LWELVIEQFEDLLVRILLLAA	4	3.0	+4.122	+0.216
ΔG_{pred} TM (L60-F78)	LLVRILLLAACISFVLAWF	-	-	-2.478	-
S *i*, *i*+5 (L49-A69)	LWELVIEQFEDLLVIRLLLAA	5	-	+3.459	+0.060
L *i*, *i*+5 (L49-F78)	LWELVIEQFEDLLVIRLLLAACISFVLAWF	5	-	+0.442	-0.336

c)



d)



e)

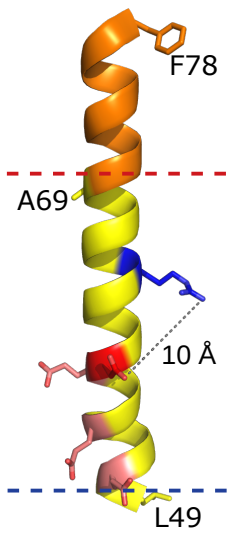


Figure S1

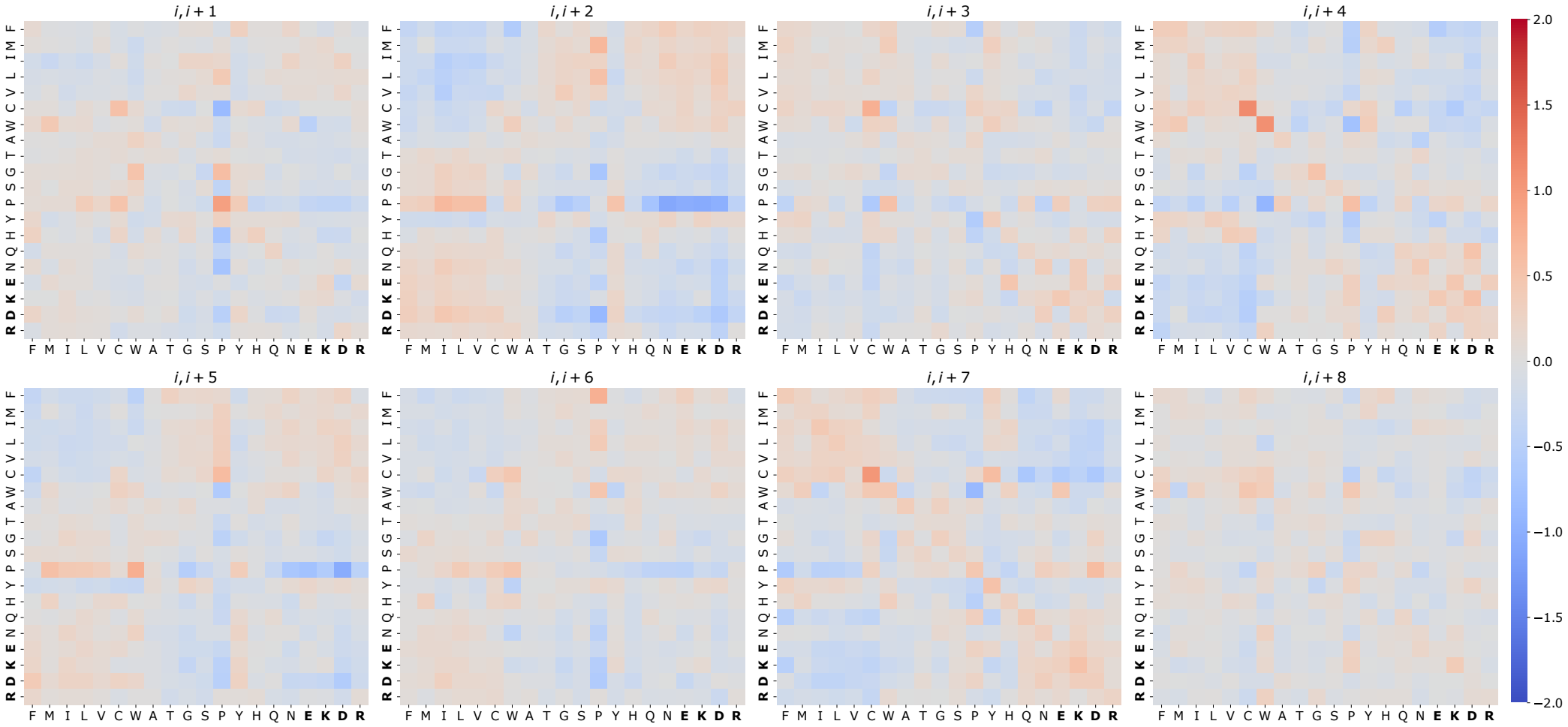

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

Figure S2

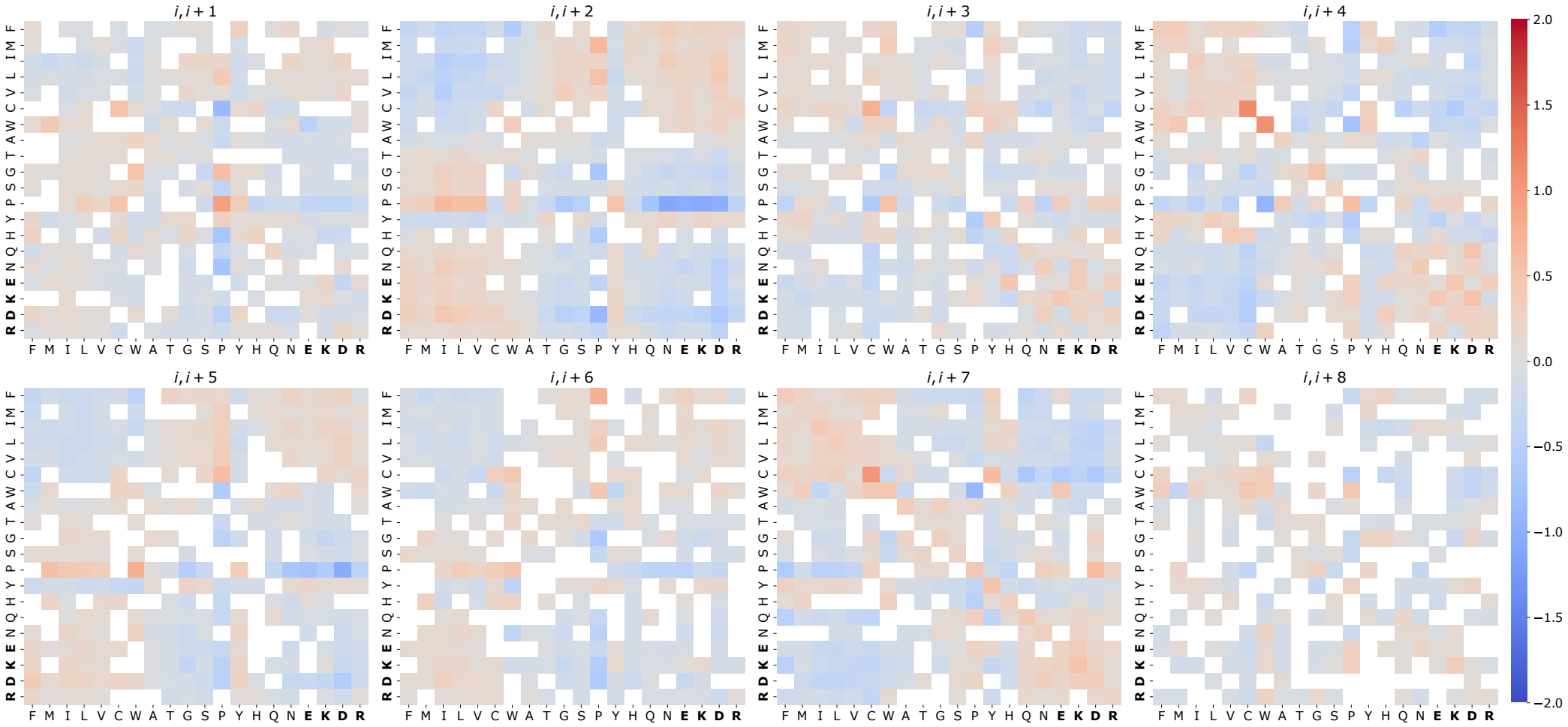
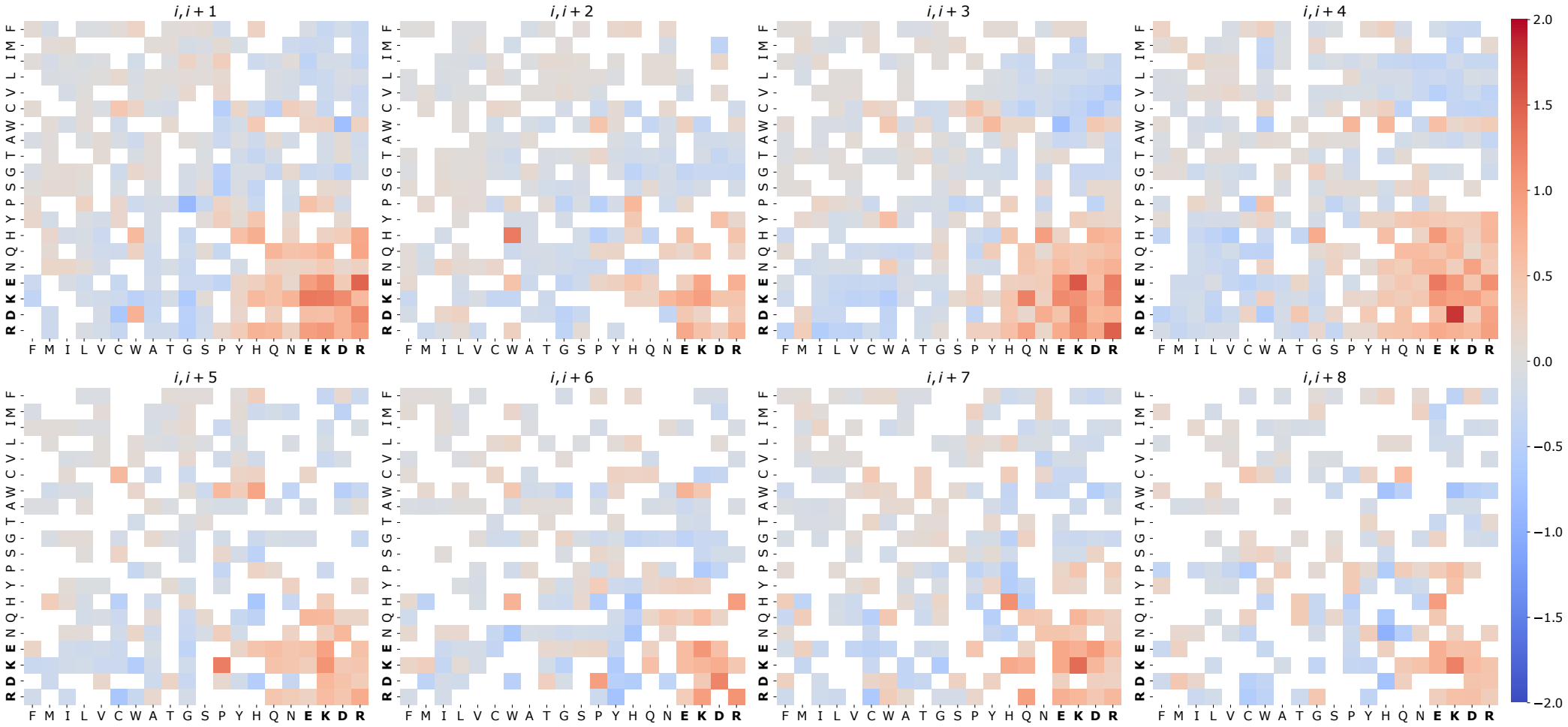

i, i + 1
i, i + 2
i, i + 3
i, i + 4
i, i + 5
i, i + 6
i, i + 7
i, i + 8
F M I L V C W A T G S P Y H Q N E K D R
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

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

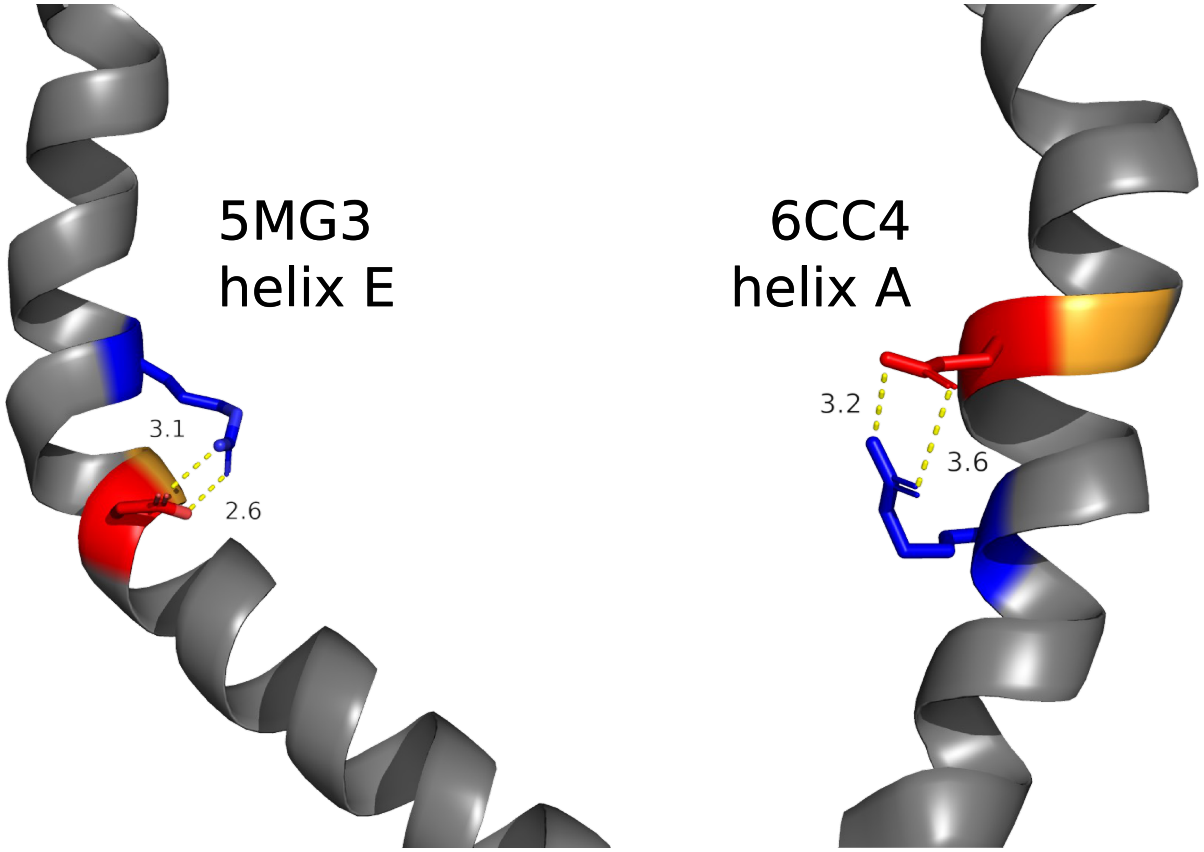



Figure S4