Title: Combinatorial mutagenesis of N-terminal sequences reveals unexpected and expanded stability determinants of the *Escherichia coli* N-degron pathway

Authors: Sabyasachi Sen, Nastassja Corrado[#], Alexander Tiso[#], Khai Khee Kho[#], Aditya M. Kunjapur*

Affiliations:

Department of Chemical & Biomolecular Engineering, University of Delaware, Newark, DE 19716

#: Authors contributed equally

*Correspondence: Correspondence to A.M.K. at kunjapur@udel.edu

Abstract:

Although it is known that penultimate N-terminal residues can influence protein stability by the N-degron pathway, there has not been a comprehensive effort to document nor predict these effects for natural and synthetic sequences in prokaryotes. Here, we present the deepest sequence coverage screen of the bacterial N-degron pathway using combinatorial site-saturation mutagenesis, fluorescence-activated cell sorting (FACS), and next-generation sequencing (NGS). Our results reveal new categories of exceptions and nuance to the N-degron pathway observed in bulk and context-specific analyses. For example, we find cases where penultimate residues can reinforce or shift the stability expected based only on the identity of position 1 (P1). We observe lowered stability for P2-P5 motifs rich in bulky residues and Gln as well as heightened stability for P2-P5 motifs rich in negatively charged residues, Gly, and Pro, particularly when these residues are at P2. We find that P1 Cys can be a component of an N-degron in a sequence-specific manner. Furthermore, we employ machine learning to predict protein stability, identifying several motifs with unexpected fates. Our work expands the sequence determinants of the bacterial N-degron pathway with unprecedented granularity, providing a set of sequence guidelines for proteomic investigation and protein engineering.

Main Text:

Introduction

Protein degradation plays a key role in supporting natural and engineered functions of bacteria. Bacterial protein degradation pertains to human health in myriad ways, such as in the development of tools such as BacPROTACs[1] that can guide antibiotic target discovery through inducible degradation of essential proteins. Furthermore, bacterial proteins directly involved in protein degradation, such as the Clp protease system in the pathogen *Mycobacterium tuberculosis*[2,3], are attractive antibiotic targets. In applied biotechnology efforts, protein degradation tags are frequently used to attenuate hysteresis in genetic circuits that are expressed in industrial workhorse strains such as *Escherichia coli*[4–6]. Additionally, bacterial proteins that recognize and recruit N-terminal degrons (N-recognins) have also been repurposed for next-generation protein sequencing methods[7–10].

Protein degradation by the N-degron pathway, which is conserved across both prokaryotes and eukaryotes[11], can modulate protein half-life by orders of magnitude in bacteria and is therefore of significant relevance to fundamental and applied researchers. In *E. coli,* bulky (Leu, Phe, Trp, Tyr) or positively charged (Arg, Lys) amino acids at the N-terminus serve as the primary signal for N-degradation based on recognition by one or both of two N-recognins[12,13]. The adaptor protein ClpS, a prokaryotic homolog of the eukaryotic N-recognin Ubr1[14], recognizes and binds to N-degron-tagged proteins that carry bulky amino acids at the N-terminus and deposits them at the ClpAP protease, accelerating substrate turnover [15–18]. A second N-recognin, LFTR (leucyl-phenylalanyl transferase), recognizes positively charged amino acids at the N-terminus and will append one or more Leu/Phe onto the N-terminus to generate a new substrate for ClpS[19–21]. However, the identity of the N-terminal residue is not a definitive determinant of protein fate. Recent work has shown that residues downstream of the N-terminus through at least the first 5 positions of a protein can impact stability through the pathway, including situations where the fate of a protein appears to defy its N-terminal residue[18,20,22–27]. These important yet relatively limited observations strongly motivate a comprehensive analysis of N-degron sequence stability determinants in bacteria (Figure 1a). Additionally, synthetic and chemical biologists have increasingly designed alternative

55 protein N-termini for non-degradative purposes, such as for bio-orthogonal conjugation reactions
56 specific to certain N-terminal amino acids[28–31], live cell cleavage of N-terminal tags or signal
57 peptides, or protein ligation or excision via inteins, sortases, or subtiligases[32–36]. For these
58 applications and more, it is valuable to ensure that any generated neo-N-termini do not form an N-
59 degron.

60 To date, the characterization of the bacterial N-degron pathway has been performed using
61 low-throughput protein half-life assays, *in vitro* pathway reconstitution, and affinity purification to
62 screen dozens to hundreds of putative N-degrons at a time [12,19,20,24,26,37,38]. While these methods
63 have pioneered our understanding of N-degrons, they have their limitations. Pulldown assays are
64 limited to sequences found and enriched within the bacterial proteome. Low-throughput reporter
65 screens require extrapolation from smaller data sets, and *in vitro* experimentation can sometimes
66 present artifacts or other deviations from *in vivo* pathway behavior. To address these limitations,
67 recent advances within the protein degradation field, including exemplary work by Timms et al.[39],
68 have leveraged high-throughput screening and next-generation of screening N-terminal libraries
69 derived from the human proteome to analyze roughly 10^4 sequences within human cells. Similar
70 approaches have been taken to study protein degradation[40–42], but no studies to date have
71 screened large DNA libraries within bacteria to characterize the N-degron pathway.

72 To address this gap, we evaluated the role of the N-terminal 5 amino acids (P1-P5) using
73 degenerate oligonucleotides for combinatorial mutagenesis, reporter protein expression in wild-
74 type and N-recognin-deficient bacterial hosts, fluorescence-activated cell sorting (FACS), and
75 next-generation sequencing (NGS). From over 2 million unique sequences screened and collected,
76 we have mapped N-degron pathway with increased granularity, as visualized through position-by-
77 position substrate specificity maps and motif analysis. We additionally share a series of sequence-
78 based stability determinants. Specifically, at P2 Gln (destabilizing), Pro (stabilizing), and Gly
79 (stabilizing) show the largest single residue stability shifts. Furthermore, additive stability effects
80 were observed for motifs rich in bulky residues (destabilizing), negatively charged residues
81 (stabilizing), and flexible Gly/Ser residues (stabilizing). In an analysis of low stability sequences
82 without the canonically destabilizing P1 FLYWRK, we found motifs rich in P1 Cys and bulky and
83 negatively charged residues in P2-P5. We then trained a machine learning model, N-FIVE, on our

84 dataset and harnessed it to predict sequences that break canonicalstability rules. We conclude by
85 discussing the implications of this work on pathway specificity, N-degron design, and future
86 applications of our platform.

87 Results

88 **Overview and development of an *in vivo* N-degron pathway fluorescent reporter**

89 To monitor *in vivo* protein degradation, we adapted the ubiquitin fusion technique[43] into a
90 dual fluorescent reporter, creating an mCherry-ubiquitin-degron-sfGFP-His6x genetic fusion
91 which we cloned into a plasmid (p15a ori, CmR, P_{araBAD}). Ubiquitin (Ub), which is generally
92 considered inert within *E. coli*, is scarlessly cleaved at its C-terminus upon expression of Ubp1, a
93 Ub protease from *Saccharomyces cerevisiae*, using a second plasmid (ColE1 ori, KanR, P_{tet}). This
94 cleavage reveals a neo-N-terminus (neo-Nt) that is attached to sfGFP while simultaneously
95 producing a stable fluorescent mCherry-ubiquitin motif. The subsequent degradation of a
96 destabilizing motif attached to sfGFP should result in lower sfGFP:mCherry ratios for pathway
97 substrates in comparison to stabilizing motifs that are attached to sfGFP (Figure 1b). This dual
98 fluorescent reporter assay avoids various confounding effects, including variable protein
99 expression due to 5' ORF mRNA sequence and the amino acid sequence near the translational
100 start site[44–52]. We selected *E. coli* BL21 as the expression host due to its deficiency of the Lon and
101 OmpT proteases, helping to minimize proteolytic crosstalk and to isolate the effect of the ClpSAP
102 system. To further isolate the impact of each key N-recognin and control for possible changes in
103 reporter expression, we generated two separate variants of BL21 where either ClpS or LFTR were
104 inactivated through the insertion of in-frame stop codons using multiplex automatable genome
105 engineering[53,54]. This configuration paired with mutagenesis of neo-Nt regions serves as the basis
106 for our high-throughput screening of the pathway (Figure 1c).

107 To verify assay functionality, we screened neo-Nt sequences previously reported to be
108 strong and weak ClpS/N-degron pathway substrates. For positive controls, we selected reported
109 ClpS ligand FLFVQEL[26,55] and LVKTKASNLL, the latter derived from ClpSAP substrate Dps[20]. For
110 negative controls, we selected SLFVQEL, a known ClpS non-interacting sequence, and
111 DHGSGAWLLP[12], a reported stable motif derived from the first 10 amino acids of β-galactosidase.

112 Using flow cytometry, we observed distinct unimodal distributions of sfGFP/mCherry ratio across
113 all tested degrons and selected median sfGFP/mCherry ratio as a comparative metric. Ratiometric
114 comparisons revealed an average 6.0 and 5.9-fold dynamic range between SLFVQEL/FLFVQEL
115 and DHGSGAWLLP/LVKTKASNLL pairs, respectively (Figure 1d). We additionally screened
116 FLFVQEL in the ClpS^{-} host and observed a 5.2-fold ratio increase over the wild-type strain, further
117 evidencing that observed ratiometric changes from our assay matched expectations for N-degron
118 pathway mediated turnover (Figure 1e).

119 **Deep sequence profiling reveals penultimate residue stability determinants**

120 We next validated our N-degron screening platform on a 60-member library and profiled
121 the P1-P3 stability of eukaryotic-derived N-degrons containing Nt-Arg-Cys, which led us to identify
122 a strongly stabilizing influence of acidic residues at P3, which we verified by Western blotting[56].
123 After this validation on a small library, we sought to expand library size to the million-scale to
124 determine the stability contributions for all combinations of the first five amino acids through the
125 N-degron pathway. As the starting point for mutagenesis, we selected an N-degron derived from
126 the eukaryotic protein RAP2.2 (CGGAIISDFI)[57] due to its consistent behavior and high dynamic
127 range between stable and unstable arginylated variants. Using 5 consecutive NNK codons located
128 at the neo-N-terminus of the reporter (XXXXXISDFI, theoretical size: 3.2E6), we transformed the
129 library into WT and knockout strains. We then induced expression of both reporter and Ubp1 and
130 subsequently collected 10M gated events using fluorescence activated cell sorting (FACS). We
131 sorted mCherry-expressing cells into four bins that covered over three orders of GFP fluorescence
132 magnitude (Figure 2a). Upon analyzing 100-150M reads of an amplicon containing the
133 mutagenized region from the WT, LFTR^{-}, and ClpS^{-} sorts, we obtained 2.29M, 2.34M, and 2.19M
134 unique sequences per host, respectively. To evaluate the stability of an analyzed sequence, we
135 then calculated the Protein Stability Index (PSI), a weighted average recently pioneered in high-
136 throughput protein degradation experimentation[39,58–60] . In line with expectations that most P1-P5
137 sequences are stable, PSI distributions were skewed towards high stability values, with over 82%
138 of wild-type sequences being in the top half of the PSI range (Figure 2b).

139 To study the sequence-specific substrate preferences of each N-recognin, we next
140 generated average PSI heatmaps for all amino acid and position combinations for the datasets
141 resulting from expression in WT (Supplemental Figure 1), LFTR$^-$(Supplemental Figure 2), and ClpS$^-$
142 hosts (Supplemental Figure 3). We observed high consistency between datasets; in a comparison
143 of the wild-type dataset and LFTR$^-$ dataset, a linear correlation of PSI achieved an r-squared of 1.0
144 for the 17 amino acids that are expected to behave similarly when at P1 (Supplemental Figure 4).
145 As expected, Arg and Lys at P1 exhibited differential stability between LFTR$^-$ and WT hosts. In
146 analyses of P1, Pro was excluded due to confounding effects where it is not uniformly cleaved by
147 Ulp1. We also examined how a protease capable of modifying some neo-Nt that contain Met at P1,
148 the native *E. coli* methionine aminopeptidase (MetAP), might affect assay results when the residue
149 at P2 is expected to be destabilizing if at P1. We compared the PSI for motifs carrying Nt-Met
150 (MXYYY) to those same motifs when they appear in the first four positions (XYYYZ) within the wild-
151 type dataset. We observed that residues that are canonically destabilizing at P1 (L,F,W,Y,R,K) had
152 increased mean PSI at the X-position in MXYYY motifs when compared to XYYYZ motifs
153 (Supplemental Figure 5). In line with the published specificity of MetAP[61,62], it did not appear that
154 Met-LFWYRK motifs produced low PSI N-degrons in bulk, though the aminopeptidase may be
155 functional on other motifs.

156 To study how individual amino acids impact stability as a function of position in the P1-P5
157 region, we computed the average PSI for each AA-position combination and plotted the differences
158 between WT and ClpS$^-$ hosts in a heatmap (Figure 2c). In line with the reported role of ClpS as a
159 sequence-specific degradation initiator, the ClpS$^-$ data shows sequence-agnostic parity in the
160 absence of the N-recognin. All cells in the corresponding P1-P5 heatmap lied within 0.17 PSI units
161 of each other, demonstrably smaller than the greater than 1 PSI differences observed between
162 canonically stable and unstable P1 residues in WT BL21 and LFTR$^-$ BL21. Furthermore, P1 stability
163 trends match literature expectations. Notably, the PSI order for destabilizing residues is roughly
164 tiered as Phe = Arg < Leu < Trp = Lys = Tyr from lowest to highest average PSI. An analysis of the
165 100,000 lowest PSI sequences revealed an enrichment for five of these six residues (excluding
166 Lys) at and near P1 (Supplemental Figure 6).

167 Given the dominant role of P1 binding in stability determination, we next turned our focus
168 towards the stability contribution of P2. As P2 is positioned on the periphery of the ClpS binding
169 pocket in crystallized bound peptide ligands, docking studies have not clearly elucidated a role for
170 this position (Supplemental Figure 7). First, we observed that select residues in P2 can generate
171 highly destabilizing sequences in tandem with LFWYRK at P1. To this end, in an analysis of the 100
172 most enriched P1-P2 motifs at low PSI, 90/100 contained at least one FLWYRK residue, and 34/36
173 possible combinations containing two of FLWYRK residues were present, showing a clear
174 overrepresentation of these six residues within destabilizing motifs relative to other sequence
175 combinations clustered at P1-P2 (Figure 3d-3e).

176 We further observed that sequences containing Gln at or near P1 were amongst the most
177 common sequences enriched at low PSI, which has not been previously reported to the best of our
178 knowledge. More distinctly, Gln at position 2 (and to a lesser extent at P3-5) is amongst the most
179 destabilizing P2 amino acid in both WT and LFTR$^-$ BL21, providing a -0.51 PSI unit decrease at P2
180 in the WT dataset. This effect was amplified and significant when Gln was paired with P1 LFWYRK
181 ($p < 1E-99$, ES = 0.26), showing a 3.2x decrease in the difference in mean PSI compared to Gln
182 paired with an alternative P1 residue (Figure 2f). The Gln trend was visible across WT and LFTR$^-$
183 data and absent in the ClpS$^-$ dataset, suggesting a ClpS-mediated preference (Supplemental
184 Figures 1-3). Furthermore, amongst P1-P2 combinations that were enriched at low PSI but did not
185 carry a canonically destabilizing P1 Gln was overwhelmingly present (Supplemental Figure 8). We
186 sought to validate this finding using clonal examples. Upon pairing Gln with a destabilizing P1
187 residue we identified sequences that resulted in sfGFP:mCherry ratios that were 3.4 and 4.0 times
188 lower below our previous degron control (RQWYHISDFI : FLFVQEL and FQAQWISDFI : FLFVQEL,
189 respectively), and FQAQWISDFI showed no visible degron-sfGFP band when visualized via
190 Western blot (Figure 2g).

191 Several amino acids showed stabilizing effects at P2. We observed this most distinctly
192 from Gly and Pro, waning with distance from P1 with a maximal impact at position 2. Commonly
193 enriched in stable, PSI >3 motifs (Supplemental Figure 9), these two residues had the highest PSI
194 change in the (WT – ClpS$^-$) heatmap. Notably, P2 Pro paired with P1 FLWYRK led to a significant
195 increase 1.05-unit increase in PSI ($p < 1E-99$, ES = 0.71) (Figure 2H). A similar, but smaller in

196 magnitude 0.63-unit increase was observed for P2 Gly ($p<1E-99$, ES = 0.47) (Figure 2I). Upon
197 testing motifs with P2 Gly, we found clonal isolates that can stabilize screened destabilizing N-
198 terminal amino acids in Leu and Trp, while N-terminal Arg remained destabilizing, potentially due
199 to Gly having an extended distance from the neo-P1 upon LFTR activity (Figure 2J).

200

201 **Similar residue side chains can cumulatively amplify or dampen N-degrons in an additive**
202 **manner**

203 We next investigated the combinatorial effect of multiple amino acids, hypothesizing that
204 certain stability determinants could be a function of aggregate properties of amino acid sidechains,
205 such as hydrophobicity or charge, or otherwise influenced by local sequence context. To select
206 phenomena to target, we grouped amino acids based on R-group characteristics and evaluated
207 their enrichment in low stability sequences within the WT dataset (Figure 3a). Unstable sequences
208 showed an enrichment for aromatic, hydrophobic, and positively charged residues, and a
209 deficiency in flexible, negatively charged, and small amino acids at various P2-P5 position, but
210 most distinctly at P2. First, recognizing the differential enrichment in charged residues as well as
211 the results presented in several papers that show that a penultimate charged residue can impact
212 ClpSAP recognition and turnover[16,25,55], we investigated whether there could be a relationship
213 between the net charge of the N-terminal region and protein stability. We observed that lower PSI
214 sequences are more commonly found in neutral to positively charged motifs, whereas negatively
215 charged motifs, even with P1 FLYWRK, demonstrated large jumps in mean stability (Figure 3b). To
216 further investigate, we tested clonal isolates carrying P1 Phe with motifs of various charge. We
217 found FDEDE (-4) to be highly stabilizing and FDEAA (-2) to be lightly stabilized relative to a control
218 degron or to a positively charged motif in FRKAA (+2) (Figure 3c). This charge-based phenomenon
219 was absent in the $ClpS^{-}$ dataset, suggesting that neutral to positive charges with P1 FLWYRK are
220 critical to their recognition by ClpS (Supplemental Figure 10). Additionally, in the WT dataset both
221 bulky (FLWY) and positively charged (RK) P1 residues showed increases in mean PSI for negatively
222 charged P2-P5 motifs, though P1 RK sequences demonstrated a higher increase in mean PSI
223 between neutral and negatively charged P2-P5 datasets (Supplemental Figure 11).

224 We next analyzed the impact of clustered FLWY residues in P2-P5 to inquire about the role
225 of multiple bulky residues in this region. For sequences carrying multiple FLWY residues at and
226 near the N-terminus, the average PSI was 0.56 units lower ($p < 1E-99$, ES = 0.45) for 2 residues
227 compared to a singular P1 FLWY. Comparable distributions were observed for motifs carrying 2+
228 FLYW residues (Figure 3d). On average, sequences carrying a single positively charged residue at
229 P1 exhibited a lower mean PSI than sequences with a bulky P1, though, notably, these sequences
230 may have multiple bulky residues appended onto P1 Arg or P1 Lys via LFTR polyaddition[19,20]
231 (Supplemental Figure 12). We screened several clonal isolates with multiple bulky and positively
232 charged residues at P1 + at least one of P2-P5 and found them all to be destabilizing (Supplemental
233 Figure 13).

234 We additionally analyzed the additive impact of multiple smaller flexible amino acids,
235 focusing on Gly and Ser, which frequently compose flexible N-terminal linkers. In line with the
236 depletion of small amino acids in destabilizing motifs, we observed an increase in PSI with
237 increasing Gly/Ser count downstream of a P1 FLWYRK residue (PSI increases for Gly/Ser counts
238 0-1: 0.09, 0-2: 0.32, 0-3: 0.64, 0-4: 0.99, all $p<1E-33$) (Figure 3e). As such, the presence of these
239 flexible motifs may aid in the functionalization and stability of N-terminal motifs with canonically
240 unstable residues.

241 **N-terminal cysteine can be destabilizing under specific sequence constraints**

242 We next investigated whether there were generalizable sequence patterns for an N-degron
243 with a canonically stable N-terminus (P1 not FLYWRK). To this end, we first analyzed amino acid
244 enrichment in sequences with moderate to low stability (PSI < 3) that did not have LFWYRK at P1.
245 Within these sequences, Cys and His were enriched at P1, and bulky residues (LFWY), acidic
246 residues (DE), and other residues (MVC) were enriched at P2-P5 (Figure 4a). In line with the
247 enrichment of acidic residues, we observed that canonically stable, low PSI motifs often
248 maintained a neutral to negative charge (Figure 4b). Interestingly, low PSI sequences with
249 noncanonically destabilizing P1 follow a charge preference that is opposite to that of the canonical
250 six destabilizing P1 amino acids, with negatively charged motifs appearing more frequently at lower
251 PSI. Furthermore, as previous studies have reported that bulky residues downstream of the N-

252 terminus can be functionally destabilizing components of an N-degron[22], we investigated
253 sequence patterns with multiple bulky residues paired with a canonically stable P1 residue. We
254 had observed that for sequences with bulky residues in P2/P3 that there was a bimodal distribution
255 of sequences with high and low PSI (Figure 4c). In an analysis of low PSI sequences with bulky
256 residues at P2 and P3, we observed that sequences with Gln, Cys, or His at P1 had the lowest
257 average PSIs (Supplemental Figure 14). The largest decreases in PSI were observed for canonically
258 stable P1 sequences with both bulky and negatively charged residues (Supplemental Figure 15).

259 We next turned our attention to P1 Cys to understand what downstream residues could
260 make this synthetically useful P1 residue an N-degron. Interestingly, P1 Cys had a lower mean PSI
261 (3.42) compared to other stable P1 residues (PSI 3.80), suggesting that there was a subset of
262 unstable sequences (Supplemental Figure 1). Upon visualizing the sequence composition of the
263 20,000 lowest PSI P1 Cys motifs, we observed an enrichment of bulky and negatively charged
264 residues, whereas positively charged and flexible residues as well as P2 Pro were depleted in these
265 motifs (Figure 4d). These enrichment trends aligned with residue-by-position PSI changes (Figure
266 4e). We generated clonal isolates of five sequences that carried candidate destabilizing motifs with
267 P1 Cys that had bulky and negatively charged residues in P2-P5. Two of the five candidates
268 presented sfGFP:mCherry ratios on par with the FLFVQEL degron, while the remaining three
269 offered intermediate ratio values (Supplemental Figure 16). When we performed Western blotting
270 analysis, at least one candidate sequence (CFYEAISDFI) was fully degraded, producing a non-
271 visible band on par with the strongest identified degrons (Figure 5f).

272 We additionally investigated other commonly functionalized P1 residues in Gly and Ser.
273 Both residues were notably stable at P1, though an analysis of P2-P5 residues for the 20,000
274 lowest PSI sequences for P1 Gly and P1 Ser revealed an enrichment in Gln, Glu, and Lys at P2-P5.
275 (Supplemental Figure 17).

276 **Training and validation of N-FIVE, an N-terminal stability predictive model**

277 To interpret the full sequence space, our next objective was to develop a stability-
278 predictive machine learning model. During training, we observed increased statistical performance
279 of our models upon applying a minimum read count threshold of 20 (RC>20). As such, we applied

280 this filter to trim our WT dataset to approximately 798,000 sequences prior to final model training.
281 To generate a base matrix, we one-hot encoded each enriched motif by amino acid and position.
282 We then trained an extreme gradient boosted model using XGBoost. We supplied these encodings
283 and the PSI from 80% of the WT dataset to generate the input for training. After hyperparameter
284 tuning, our model obtained an RMSE of 0.34 and an r-squared of 0.82 when validated using the
285 remaining 20% of data. Satisfied with the performance of our model, we dubbed our model N-FIVE
286 and set out to test its predictive capabilities (Figure 4a-b). We first simulated the PSI of 100,000
287 random N-terminal sequences and observed the characteristic decrease in mean PSI for
288 sequences carrying destabilizing N-termini (Figure 4c). Upon simulating the full 3.2M sequences
289 using N-FIVE, we observed a high correlation compared to our actual dataset (Supplementary
290 Figure 18). We generated SHAP plots to understand the underlying architecture, observing trends
291 on par with known N-degron pathway rules such as the dominant role of P1 identity
292 (Supplementary Figure 19).

293 To validate the predictive capabilities of N-FIVE, we sought to generate N-terminal motifs
294 with unexpected stability profiles. To this end, we sought to identify five sequences with a
295 canonically destabilizing P1 Leu that would have stable profiles. Excitingly, all five of the
296 sequences (sequence LXXXXISDFI) suggested by N-FIVE displayed stable profiles (Figure 4d). We
297 additionally investigated the challenging task to identify sequences that are unexpectedly unstable
298 with P1 residues that are not one of FLWYRK. From our five-sequence screen, one motif,
299 SYMMDISDFI, generated a moderately destabilizing profile (Figure 4f). SHAP values were
300 calculated for candidate sequences LAASYISDFI (Figure 4e) and SYMMDISDFI (Figure 4g),
301 revealing the cumulative effect of P2-P5 residues in overriding the P1 residue as a stability
302 determinant.

303

304 Discussion

305 This work presents the most comprehensive mapping of the *Escherichia coli* N-degron
306 pathway to date. Having collected and analyzed over 2.2M screened sequences in the P1-P5 space
307 within wild-type BL21, we both reinforce and provide insight into the pathway with increased

308 granularity. Our findings corroborate the consensus understanding of the pathway within *E. coli*,
309 namely that six amino acids (Leu, Phe, Tyr, Trp, Arg, Lys) when present at P1 by and large lead to
310 protein degradation through the ClpSAP proteolytic cascade. We have expanded this to
311 demonstrate how residues in the P2-P5 positions can augment or dampen P1 effects, in rare cases
312 even fundamentally altering the expected level of protein stability. This is most evident for the
313 destabilizing P2 Gln and the stabilizing P2 Pro and Gly. Notably, we observe that most destabilizing
314 N-terminal motifs frequently have multiple bulky and positively charged residues in P1-P5.
315 Exceptions to this rule frequently carry P1 Cys and His and have bulky or negatively charged
316 residues in P2-P5. It is unclear how P2-P5 bulky amino acids may be causing instability as there is
317 conflicting evidence about whether bulky residue downstream of P1 can be bound directly in the
318 hydrophobic pocket of ClpS[22]. It is possible that this instability could be caused by strengthened
319 interactions in the substrate channel or surface of the protein. Furthermore, in an analysis of N-
320 degrons with canonically stable N-termini, we observed an enrichment in Cys, Gln, and His at P1.

321 Amongst the most destabilizing motifs that we observed were those that could be generated
322 by LFTR. Multiple previous studies have identified that polyaddition of Leu and Phe, rather than
323 addition of a singular Leu or Phe, is a possibility.[19,20] The highly destabilizing nature of multiple
324 Leu/Phe-Arg residues in tandem coincide with previous observations that the turnover of LR-
325 substrates is greater than just L-substrates.[16] To this end, others have suggested that the ClpS
326 binding site has been optimized to accommodate LFTR-modified substrates.[63] On balance, the
327 generation of a highly destabilizing motif in front of Nt-Arg and Lys may routinely contribute to the
328 generation of some of the most highly destabilizing degrons, demonstrating the rapid turnover of
329 LFTR-substrates despite the requirement of an additional modification.

330 These findings bear relevance to the design of synthetic N-termini for applications in small
331 molecule or protein ligation in live cells. For example, bioconjugation efforts that rely on a P1 Cys
332 should avoid bulky and negatively charged residues near the N-terminus. Gly-Ser motifs are likely
333 to provide a stabilizing effect that would offset any degradation-promoting elements. The
334 overrepresentation of negatively charged residues and bulky residues in low PSI sequences with
335 stable N-termini and small residues in high PSI sequences is worth considering when designing
336 protein neo-N-termini with targeted stability levels in mind. For such designs of synthetic neo-N-

termini, the stability-predictive N-FIVE model can be consulted, including for design with sequence constraints. Our confidence in the predictive power of N-FIVE is evidenced by its ability to solve a complex sequence-stability problem – in this instance successfully suggesting unexpectedly stable sequences. The identification of unexpectedly unstable sequences remains a challenge, in part due to the small number of candidate sequences to form the dataset. Further development of more complex ML models may improve stability prediction capabilities.

Code availability

NGS analysis code, sequence databases, and the N-FIVE model have been made available at GitHub (https://github.com/KunjapurLab/N-terminal-cluster-stability).

Acknowledgements

The authors thank the following people: Dr. Arit Ghosh of the University of Delaware Flow Cytometry Core for training and advice on cell sorting and analysis. Bruce Kingham and Mark Shaw of the University of Delaware Sequencing & Genotyping Center for consultations and performing NGS sequencing. Ishika Govil and Zhao Wei Wang for discussions and code for the analysis of NGS data. Prof. Karl Schmitz for_discussions regarding ClpS binding and the N-degron pathway. Prof. Kevin Solomon for discussions regarding bioinformatic analysis. We further thank members of the Kunjapur Lab for their helpful discussions and review of the manuscript.

Author Contributions

S.S. designed and performed most experiments, wrote the bioinformatics and machine learning code, analyzed data, and wrote the manuscript. N.C. assisted in optimizing DNA library generation. A.T. assisted in developing NGS data analysis code. K.K. assisted with cloning and assay development. A.M.K. oversaw the project, designed experiments, and assisted in writing and editing the manuscript.

Conflict of Interest Statement

A.M.K. has a financial interest in a commercial entity, Nitro Biosciences Inc. The remaining authors declare no competing interests.

AI Statement

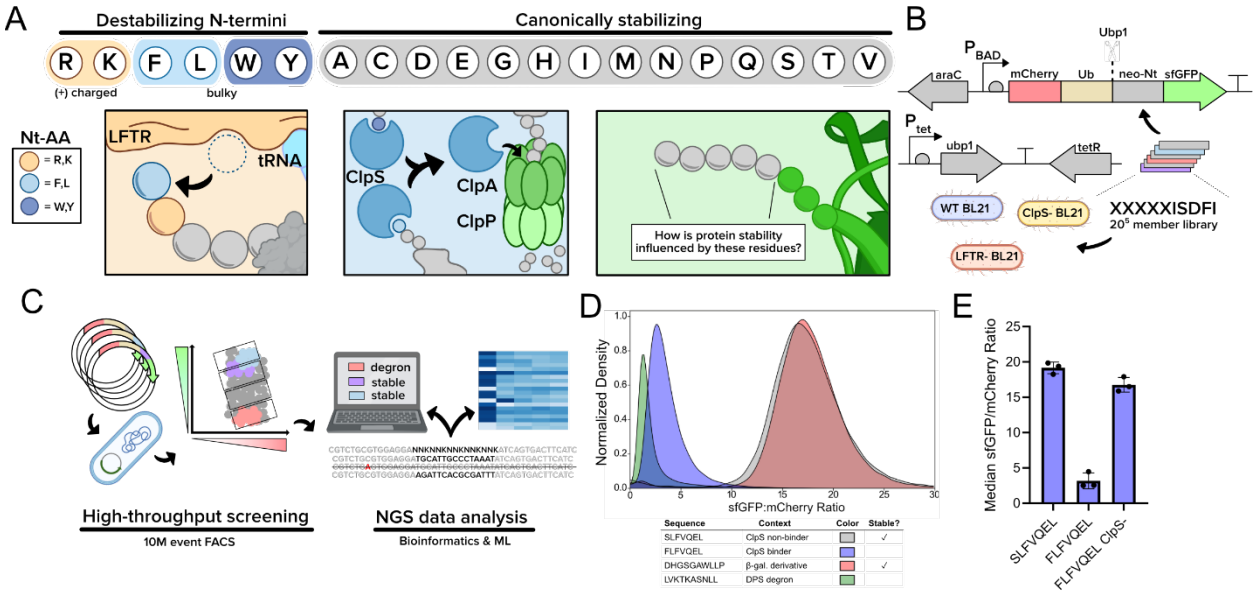
Bioinformatics analysis code was developed and debugged with the assistance of ChatGPT.

Funding

This work was funded by the following: the U.S. Department of Agriculture's National Institute of Food and Agriculture (project award no. 2021-33522-35807), grant 1DP2AI176137-01 provided

369 by the Office of the Director and the National Institute of Allergy and Infectious Diseases (NIAID) at
370 the National Institutes of Health (NIH), grant P20GM104316B from the National Institutes of
371 Health (NIH)-funded COBRE Center for Discovery of Chemical Probes and Therapeutic Leads, as
372 well as the University of Delaware Research Foundation. Additional support for Sabyasachi Sen
373 was provided by the Mort & Donna Collins Fellowship. Access to FACS and NGS was supported
374 by the Institutional Development Award (IDeA) from the National Institute of Health's National
375 Institute of General Medical Sciences under grant number P20GM103446 through DE-INBRE.

376 **Figures:**



377

378 **Figure 1.**

379 **A.** An overview of the *Escherichia coli* Leu/N-degron pathway. The top row shows the expected
380 fate for proteins with the listed amino acids at the N-terminus, where the orange and blue
381 shaded residues are canonically destabilizing. The left-most box illustrates the
382 leucyl/phenylalanyl transferase LFTR, appending an F or L onto an N-terminal R or K. The
383 middle box illustrates the ClpS adaptor protein that binds N-terminal F, L, W, and Y and recruits
384 the protein to the ClpAP protease for degradation. The right-most box shows the dual reporter
385 after the action of a ubiquitin (Ub) cleavase, revealing a neo-N-terminus (neo-Nt) on an sfGFP
386 reporter. **B.** Genetic constructs for the protein stability assay. Wild-type or knockout *E. coli*
387 strains were transformed to harbor two plasmids. The first plasmid harbors a synthetic gene
388 fusion that encodes the dual fluorescent reporter under control of an arabinose-inducible
389 promoter. The second plasmid harbors a gene encoding ubiquitin cleavase from
390 *Saccharomyces cerevisiae* under control of a tetracycline-inducible promoter. Constructs
391 were tested in wild-type BL21 as well as BL21 derivatives deficient in *clpS* and *aat* (LFTR). **C.**
392 The HTS protein stability workflow used in this study. It includes 10^6 sequence DNA library
393 generation, high-efficiency transformation, 10^7 event cell sorting, next-generation sequencing,
394 and data analysis that includes machine learning, quality control, stability evaluation, and motif

395 enrichment. **D.** The protein stability assay differentiates literature controls. sfGFP:mCherry
396 ratio density histograms for over 100,000 events collected for cells expressing Ubp1 as well as
397 literature control N-terminal sequences embedded in a dual fluorescent reporter construct.
398 Stable sequences were SLFVQEL and DHGSGAWLLP and unstable sequences were FLFVQEL
399 and LVKTKASNLL. Dual reporter ratiometric comparison of literature controls using flow
400 cytometry. **E.** Known ClpS-binding degron FLFVQEL returns stabilizing sfGFP:mCherry ratio
401 values in the absence of ClpS. Data shown is median fluorescence intensity ± standard
402 deviation for at least 100,000 events captured across clonal biological triplicates using flow
403 cytometry. Figure generated using Biorender content.

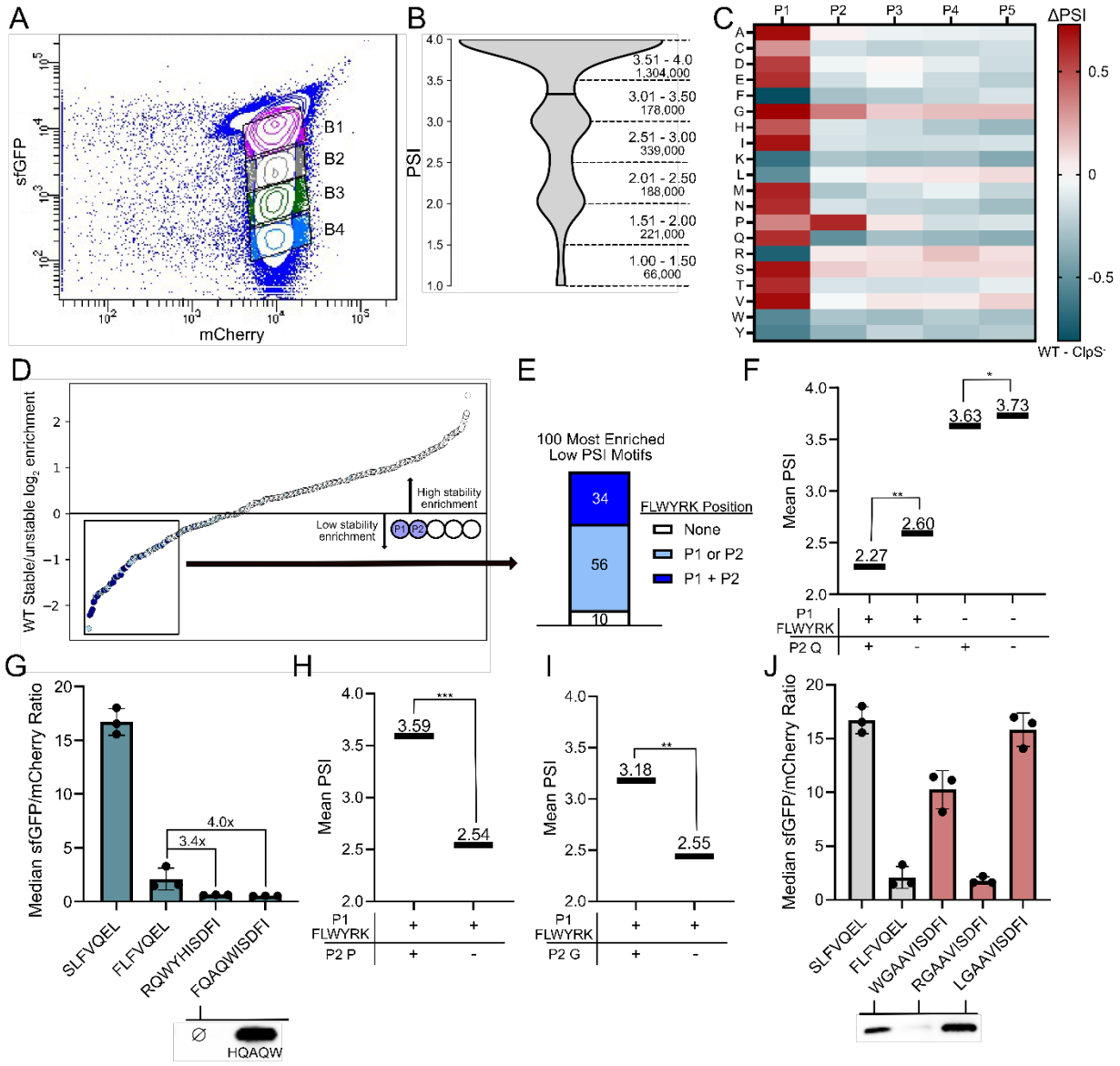


Figure 2

A. FACS binning of the XXXXXISDFI library within wild-type BL21 **B.** Violin plot showing the PSI distribution of 2.30M collected sequences from the wild-type BL21 dataset. The number of sequences rounded to the nearest thousand can be found below the corresponding PSI sub-range. **C.** Mean change in PSI heatmap between wild-type BL21 and ClpS- BL21. **D.** Log_2 enrichment ratio plot of the 400 possible stable (PSI >3)/unstable

412 (PSI<2) P1-P2 dipeptide motifs. Sequences containing LFWYRK in either P1 or P2 are
413 highlighted in light blue and sequences containing LFWYRK in both P1 and P2 are
414 highlighted in dark blue. **E.** LFWYRK composition of the 100 lowest PSI P1/P2 motifs. **F.**
415 Mean PSIs for wild-type sequence subsets, revealing an amplified PSI lowering effect by P2
416 Gln when paired with a canonically unstable N-terminus. The number of sequences per
417 condition are (++): 29,982, (+-): 754, 580, (-+): 52,021, (--): 1,459,940; (++/+-) ES: 0.26. **G.**
418 Clonal P2 Gln isolates show decreased stability measurements relative to a literature
419 control. Flow cytometry data for various P2 glutamine motifs (top). Western blots of
420 hardcoded sequences (bottom) No band was visible for the FQAQWISDFI sequence. **H.**
421 Mean PSIs for wild-type sequence subsets, revealing an amplified PSI raising effect by P2
422 Pro when paired with a canonically unstable N-terminus. The number of sequences per
423 condition are (++): 33,674, (+-): 750, 888; (++/+-) ES: 0.71. **I.** Mean PSIs for wild-type
424 sequence subsets, revealing an amplified PSI lowering effect by P2 Gly when paired with a
425 canonically unstable N-terminus. The number of sequences per condition are (++): 41,975,
426 (+-): 742,587; (++/+-) ES: 0.47. **J.** Clonal P2 Gly isolates show stabilizing effects for two
427 non-LFTR modified N-termini. Flow cytometry data for various P2 glycine motifs (top).
428 Western blots of hardcoded sequences (bottom). Comparisons are abbreviated as (P1
429 FLYWRK presence, P2 presence). p-values evaluated using a Mann-Whitney U test. Effect
430 size magnitude (ES) ranges from 0 to 1, as evaluated using Cliff's Delta. * = p <0.05, ES 0-
431 0.25; ** = p <0.05, ES 0.25-0.50, *** = p <0.05, ES 0.50-0.75, **** = p <0.05, ES 0.75-1.0

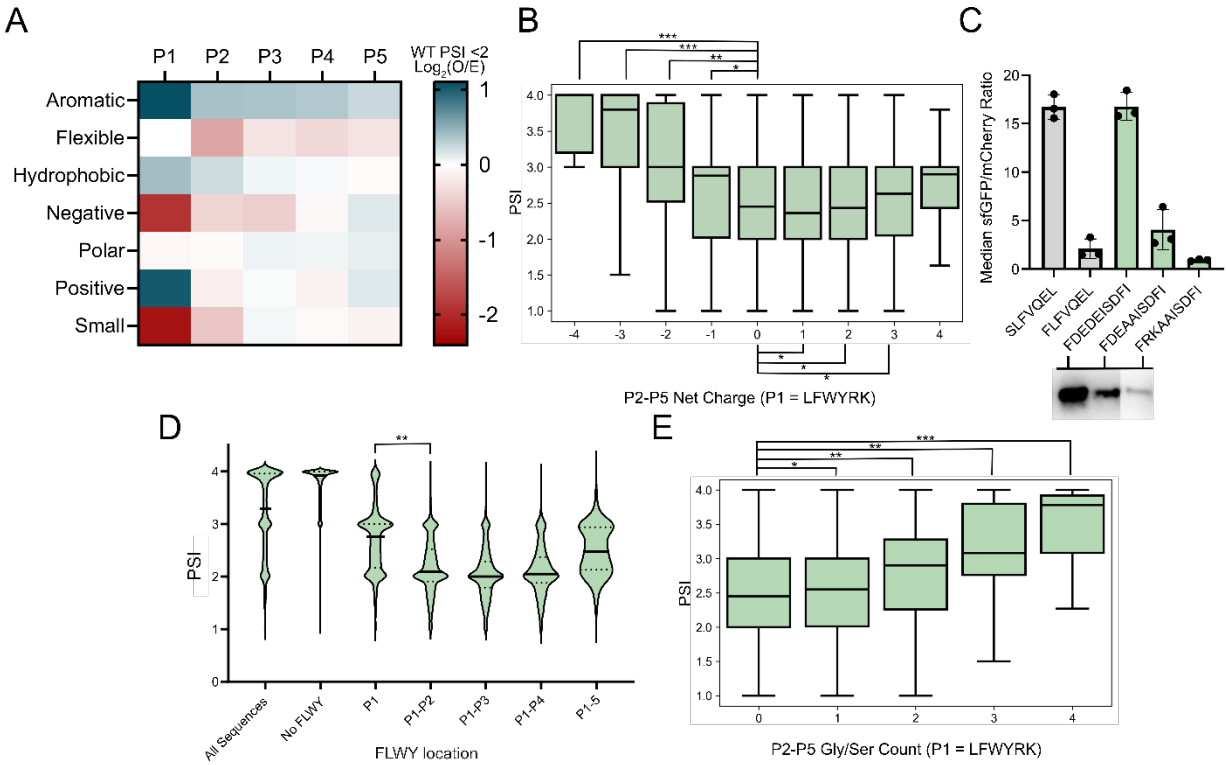


Figure 3.

A. Low stability motif heatmap depicting $\log_2$ enrichment (observed/entire dataset) of grouped amino acids. Negative = DE, Positive = RK, Hydrophobic = AVILMFWY, Polar = RNDQEKHSTYC, Aromatic = FWYH, Small = GASC, Flexible = GSDN. **B.** Boxplot showing that sequences with P1 FLWYRK and negatively charged P2-P5 show increased bulk stability. (Net charge, n, ES relative to 0 net P2-P5 charge): (-4, 35, 0.62), (-3, 1,341, 0.61), (-2, 22,647, 0.39), (-1,155,936,0.15), (0,399,998, N/A), (+1, 171, 933, 0.04), (+2, 30,185, 0.01), (+3, 2,422, 0.07), (+4, 75, 0.24). All p-values in comparison to 0 net charge data show $p < 0.05$. **c.** Negative charge can stabilize a canonically destabilizing P1 residue. (Top) Flow cytometry data for P1-P5 sequences with varying P2-P5 charge (FDEDE: -4, FDEAA: -2, FRKAA: +2). (Bottom) Western blots of corresponding sequences. **d.** Boxplot depicting that clustered FLWY at and near the N-terminus leads to decreased stability. x-axis lists exclusive locations for FLWY in the P1-P5 motif. P1(n=206,557) compared to P1+P2 (n=54,022) shows significance ($p<0.05$, ES = 0.45). **e.** Boxplot showing increased PSI means and distributions for sequences with P1 FLWYRK and

increasing Gly/Ser residue counts in P2-P5. (Gly/Ser count, n, ES relative to 0 P2-P5 Gly/Ser): (0, 477,006, N/A), (1, 257,126, 0.07), (2, 46,829, 0.25), (3, 3,505, 0.50), (4,96, 72). All p-values in comparison to 0 P2-P5 Gly/Ser show $p < 0.05$. p-values evaluated using a Mann-Whitney U test, effect size magnitude (ES) ranges from 0 to 1, as evaluated using Cliff's delta. * = $p < 0.05$, ES 0-0.25; ** = $p < 0.05$, ES 0.25-0.50, *** = $p < 0.05$, ES 0.50-0.75, **** = $p < 0.05$, ES 0.75-1.0

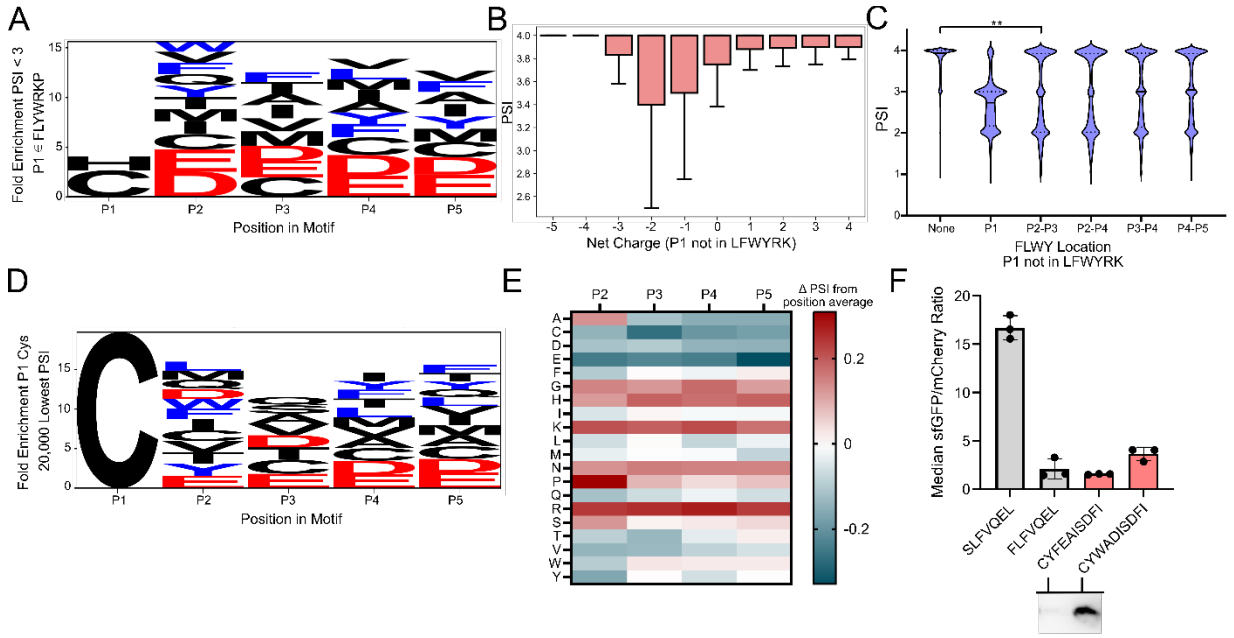


Figure 4.

A. Low stability sequences with canonically stable N-termini are rich in negatively charged and bulky amino acids in P2-P5. Weblogo depicting fold enrichment amino acids in low PSI (PSI <3) motifs that do not contain FLWYRKP at P1. Bulky residues (FLWY) depicted in blue and negatively charged residues (DE) depicted in red. **B.** PSI boxplot for sequences with varying net charge and canonically stable P1 residues shows the opposite charge trend to low stability P1 FLWYRK sequences. P1 Pro is excluded. **C.** Multiple bulky residues lower mean PSI. Violin plot of wild-type BL21 data showing PSI distributions for sequences with bulky residues at various positions. A comparison between motifs with no bulky residues (n = 1,511,962) and motifs with bulky residues only in P2 and P3 (n=56,341) shows a significant difference between mean PSI values ($p < 0.05$, ES = 0.25) **D.** Weblogo depicting enriched amino acids for the 20,000 lowest PSI motifs with P1 Cys. Bulky residues (FLWY) depicted in blue and negatively charged residues (DE) depicted in red. **E.** Amino acid enrichment heatmap for motifs with P1 Cys. Change in PSI is calculated for each amino acid relative to the position average. **F.** Identification of an Nt-Cys degron. (Top) flow cytometry data, (bottom) Western blots of the

471 degron-sfGFP product for candidate degrons with N-terminal cysteine. p-values evaluated
472 using a Mann-Whitney U test, effect size magnitude (ES) ranges from 0 to 1, as evaluated using
473 Cliff's delta. * = $p < 0.05$, ES 0-0.25; ** = $p < 0.05$, ES 0.25-0.50, *** = $p < 0.05$, ES 0.50-0.75, ****
474 = $p < 0.05$, ES 0.75-1.0
475
476

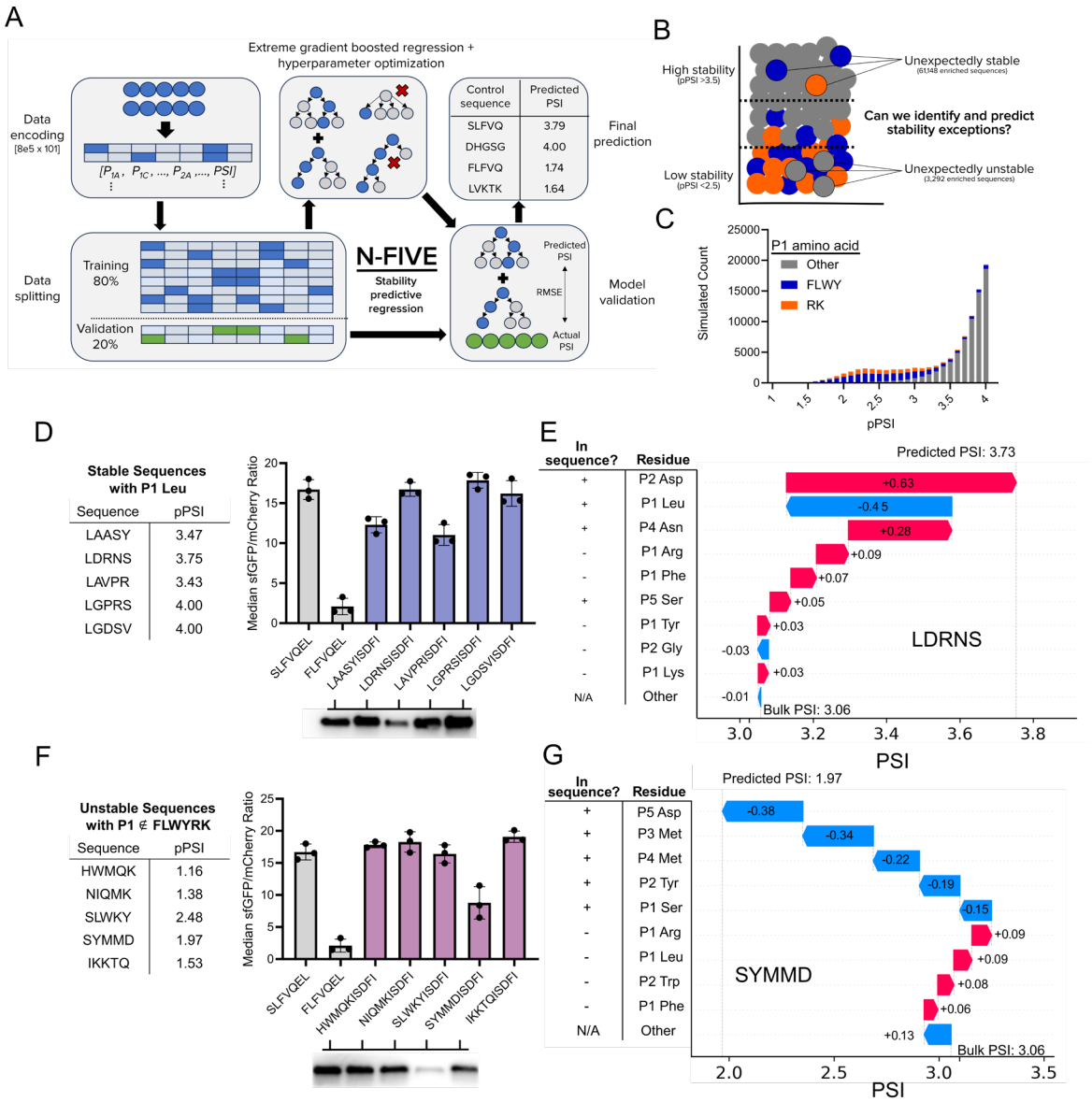


Figure 5.

A. Architecture of the N-FIVE machine learning model. Enriched sequences were one-hot encoded into a matrix. 80% of sequences were used to train and optimize an extreme gradient boosted machine learning model. Model validation was performed on the remaining 20% of sequences. The first 5 amino acids of control sequences (Figure 1d) were predicted using N-FIVE and found to match the expected stability profile. **B.** High stability sequences (predicted PSI >3.5) are under-enriched in P1 LFWYRK and low stability sequences (predicted PSI <2.5) are over-enriched in P1 LFYWRK. Stability exception predictions would be a challenging task.

486 **C.** PSI prediction of 100,000 randomly selected 5-amino acid sequences using N-FIVE.
487 Sequences with a canonically destabilizing P1 are highlighted (bulky residues in blue, positively
488 charged residues in orange). **D.** (Left) Predicted PSI values for five unexpectedly stable (pPSI
489 >3.5) candidate sequences selected with a P1 Leu constraint. (Top right) Candidate sequence
490 flow cytometry data, (right left) Western blots of the degron-sfGFP product. **E.** (right) SHAP
491 values for individual matrix contributions to the predicted PSI value of 3.73 for the candidate
492 sequence LDRNS. **F.** Predicted PSI values for five unexpectedly unstable (pPSI < 2.5) candidate
493 sequences selected with a P1 ∉ FLWYRK constraint. (Top right) Candidate sequence flow
494 cytometry data, (bottom right) Western blots of the degron-sfGFP product. **G.** (right) SHAP
495 values for individual matrix contributions to the predicted PSI value of 1.97 for the candidate
496 sequence SYMMD.

497

498 Methods

499 **Strains, plasmid cloning, and transformations**

500 *Escherichia coli* strains and plasmids used are listed in Table S1 and Table S2, respectively.
501 *clpS* and *aat* inactivations were performed as described in the Genetic knockout of N-recognins
502 section. Genes were purchased as G-Blocks or gene fragments from Integrated DNA Technologies
503 (IDT) or Twist Bioscience and were optimized for E. coli K12 using the IDT Codon Optimization Tool.
504 A version of scUbp1 without the N-terminal region was used for improved expression in *E. coli*[64].
505 Essential genetic sequences and primers can be found in Table S3 and Table S4, respectively.
506 Cloning amplicons were generated using KOD XTREME Hot Start polymerase and corresponding
507 buffering reagents. Amplicons were verified and purified by running through a 1% agarose gel for
508 200 V for 20 minutes followed by gel excision and extraction. Samples were assembled in
509 homemade Gibson assembly aliquots run at 50 °C for 30 minutes. Assemblies were transformed
510 into *E. coli* cloning strains DH5α or DH10β using either electroporation or chemical heat shock.
511 Following a one-hour outgrowth after transformation, cells were plated on permissive antibiotic
512 plates and were incubated at 37 °C overnight. The following afternoon, multiple colonies were
513 picked and incubated overnight in separate culture tubes containing LB media (10 g/L tryptone, 5
514 g/L sodium chloride, 5 g/L yeast extract) and the permissive antibiotic. The next day, cells were
515 stocked 1:1 with 30% glycerol at negative 80 °C for future usage. The remaining cells were
516 miniprepped for sequencing. Plasmid sequences were verified through a mixture of Sanger
517 sequencing through Eurofins, Genewiz, and Azenta and full plasmid sequencing through
518 Plasmidsaurus. Plasmids were transformed into BL21 and BL21 derivatives using either
519 electroporation or chemical heat shock. Two plasmid systems were transformed in series.

520 **Materials and chemicals**

521 The following compounds were purchased from MilliporeSigma (Burlington, MA, USA):
522 phosphate-buffered saline (PBS), kanamycin sulfate, glycerol, 25 nm membranes (VSWP02500)
523 and KOD XTREME Hot Start polymerase (Millipore 71975-3). D-glucose was purchased from TCI
524 America (Portland, OR, USA). Agarose, Laemmli SDS sample reducing buffer, and ethanol were
525 purchased from Alfa Aesar (Ward Hill, MA, USA). Anhydrotetracycline (aTc) was purchased from

526 Cayman Chemical (Ann Arbor, MI, USA Methanol, sodium chloride, LB Broth powder (Lennox), LB
527 Agar powder (Lennox), Amersham ECL Prime chemiluminescent detection reagent, and Thermo
528 Scientific™ Spectra™ Multicolor Broad Range Protein Ladder were purchased from Fisher Chemical
529 (Hampton, NH, USA). Taq DNA ligase was purchased from GoldBio (St. Louis, MO, USA). Phusion
530 DNA polymerase and T5 exonuclease were purchased from New England BioLabs (NEB) (Ipswich,
531 MA, USA). SybrSafe DNA gel stain was purchased from Invitrogen (Waltham, MA, USA). E. cloni
532 10G Supreme electrocompetent cells were purchased from Biosearch Technologies. Miniprep kits
533 were purchased from Zymo. 4-20% precast protein gels were purchased from Bio-Rad. Flow
534 cyometry filter caps (Chemglass Life Sciences CLS4380009), arabinose, DpnI enzyme
535 (FERFD1704) and Immobilon-E Western blot membranes were obtained from Fisher Scientific.
536 HRP-conjugated anti-6*His antibody (Proteintech HRP-66005) was obtained from Proteintech
537 (Rosemont, IL, USA).

538 **Genetic knockout of N-recognins**

539 Multiplex automatable genome engineering (MAGE) was used to inactivate the endogenous
540 *aat* (LFTR) and *clpS* genes[65]. MAGE oligonucleotides were designed using MODEST[66] to insert three
541 in-frame stop codons into the gene of interest. Freshly made electrocompetent cells were
542 resuspended in 5 μM oligonucleotide and subsequently electroporated to enable cell permeation
543 of the oligonucleotide. Cells were outgrown in LB media for an hour prior to plating on permissive
544 plates. To verify the desired genetic knockouts, allele-specific colony PCR was performed using
545 KAPA 2G Fast HotStart polymerase (Roche KK5801). Additional Sanger sequencing was
546 performed to verify asPCR hits.

547 **Fluorescence analysis of degrons**

548 Overnights made from biological triplicate colonies of BL21 and recognin-deficient BL21
549 derivatives carrying both the dual reporter and Ubp1 expressing plasmids were inoculated at a
550 1:100 ratio in fresh LB media containing 0.2% arabinose, 100 ng/mL aTc, 20 μg/mL
551 chloramphenicol, and 15 μg/mL kanamycin. Cells were grown at 37 °C for 18 hours prior to
552 analysis. Cells were then filtered using 35 μm filter caps into cytometry tubes containing fresh PBS
553 at a roughly 1:250 ratio of cell culture: PBS. Cells were analyzed on a NovoCyte Flow Cytometer

554 (Agilent Technologies). Prior to analysis, scatter gates were used to isolate singlet bacterial cells
555 of the appropriate morphology. sfGFP fluorescence was analyzed using a 488 nm laser and a
556 530/30 nm bandpass filter and mCherry fluorescence was analyzed using a 488 nm laser and a
557 660/20 nm bandpass filter. For all analyses, at least 100,000 events that passed through the
558 singlet gate were collected.

559 **Western blotting of hexahistidine-tagged reporter proteins**

560 1 mL of overnighted culture containing expressed dual reporter and Ubp1 protease were
561 lysed using glass beads for 15 minutes. The sample was transferred to a fresh microcentrifuge tube
562 and centrifuged at 12,000 RPM for 10 minutes. Subsequently, the supernatant was collected.
563 Sample concentration was evaluated using a Bradford assay on a SpectraMax i3x with a BSA
564 calibration curve. Sample concentration was normalized to 0.1 mg/mL in water with 1x SDS PAGE
565 loading dye and denatured at 95 °C for 10 minutes. To separate proteins by size, 10 uL of sample
566 is loaded into a 4–20% Mini-PROTEAN TGX gel and run at 180 V for 35 minutes. Subsequently, the
567 gel is transferred onto an Immobilon membrane through wet transfer overnight (25V for 15 hours
568 on ice). The following day, the membrane is blocked for at least one hour with 5% milk in TBST
569 buffer. An HRP-conjugated anti-His6x antibody is added to the solution at a 1:10,000 dilution and
570 is allowed to incubate at room temperature for 1 hour. Following three washes in TBST, ECL Prime
571 chemiluminescent reagent is added to the membrane. Following a 10 second incubation,
572 chemiluminescent images at varying intensities are captured using an Azure c280.

573 **DNA library preparation**

574 DNA library plasmids were designed to be assembled from two separate amplicons.
575 Primers containing 1-5 consecutive NNK codons were ordered from IDT, using a
576 25A/25C/25G/25T (N) or 50G/50T (K) hand-mixed nucleotide split to maximize diversity in the
577 mutagenized region. PCR amplicons were generated using KOD XTREME polymerase using roughly
578 200 ng of plasmid template per reaction with an annealing temperature of 59 °C and extension time
579 of 2 minutes for inserts and 3 minutes for the backbone. To maximize sequence diversity and avoid
580 overamplification, 5 separate PCR amplicons were generated at a low cycle count (20 cycles) and
581 high volume (50 µL per reaction). Amplicons were then digested using DpnI to remove the

582 template. The amplicons were then isolated on a 1% agarose gel and subsequently purified using
583 gel extraction. Amplicons were pooled and assembled using an isothermal Gibson assembly
584 method with a total of 300 ng of total DNA at a 1:1 molar ratio of insert: backbone. Assemblies
585 were then dialyzed on a 25 nm membrane placed over deionized water for 60 minutes. Five
586 separate assemblies were pooled and then electroporated into E. cloni 10G Supreme
587 electrocompetent cells. Immediately following electroporation, cells were mixed with 500 µL of
588 SOC media and were placed in a culture tube in a 250 RPM shaking incubator set to 37 °C.
589 Following a 30-minute outgrowth, cells were plated at serial dilutions to verify transformation
590 efficiency. For 20^5 member libraries, a transformation efficiency of 10^8 colonies/transformation
591 was the minimum threshold for further experimentation. Following a second 30-minute outgrowth,
592 5 mL of SOC media was added to the tube and cultures were allowed to grow overnight.

593 The following morning, library stocks were made by mixing 500 µL of undiluted culture at a
594 1:1 ratio with 30% glycerol, which were then stored at -80 °C. The remaining 4-4.5 mL of culture
595 was miniprepped using a Zymo ZR Plasmid Miniprep-Classic kit. BL21 and recognin-deficient BL21
596 derivatives carrying a Ubp1-expressing plasmid were made electrocompetent using standard
597 molecular biology procedures, concentrating 10 mL cultures to 200 µL competent cell aliquots.
598 Miniprepped libraries were electroporated into freshly made competent cells and followed parallel
599 outgrowth, efficiency measurements, overnight growth, and freezer stocking to original library
600 transformations. A transformation efficiency of 10^8 was obtained for all conditions. Library diversity
601 was confirmed prior to sorting using 2.5 Mbp next-generation sequencing through Azenta's
602 AmpliconEZ platform.

603 **Cell culturing and fluorescence activated cell sorting**

604 Overnights of BL21 and recognin-deficient BL21 derivatives carrying both library plasmids
605 and Ubp1-expressing plasmids were inoculated at a 1:100 ratio in fresh LB media containing 0.2%
606 arabinose, 100 ng/mL aTc, 20 µg/mL chloramphenicol, and 15 µg/mL kanamycin. Cells were
607 grown at 37 °C for 18 hours prior to analysis. Cells were then filtered using 35 µm filter caps into
608 cytometry tubes containing fresh PBS at a roughly 1:250 ratio of cell culture: PBS. Cells were
609 sorted on a BD FACS Aria Fusion using a 70 µm or 85 µm nozzle and a neutral density 1.0 filter.
610 Prior to analysis, scatter gates were used to isolate singlet bacterial cells of the appropriate

611 morphology. sfGFP fluorescence was analyzed using a 488 nm laser and a 530/30 nm bandpass
612 filter and mCherry fluorescence was analyzed using a 561 nm laser and a 610/20 nm bandpass
613 filter. Sorting conditions and sample concentrations were optimized to obtain event rates on the
614 order of 1,000 – 5,000 events per second and a sort efficiency >95%. 10 million gated cells were
615 sorted into four equally sized bins covering the sfGFP range, excluding a small gap at the bottom of
616 the distribution that was observed to have a higher concentration of debris and GFP-inactivating
617 mutants. Cells were recovered overnight in a mixture of 1:1 recovered cells: LB media containing
618 1% glucose. The subsequent day, cells were stocked and miniprepped using methods described
619 in the DNA library preparation section.

620 **Next generation sequencing sample preparation**

621 250 base pair (AVITI) and 405 base pair (Amplicon EZ) DNA amplicons for next-generation
622 sequencing were designed to contain a 5' and 3' sequencing adapter, a 3 base pair barcode for
623 demultiplexing, and a centralized mutagenized region to ensure coverage by both paired-end
624 reads. The full sequence can be found in our corresponding methods publication[56]. Amplicons
625 were generated using KAPA 2G polymerase, minimizing cycle counts (20 was default) to decrease
626 over-enrichment of select sequences. The amplicons were then isolated on a 1% agarose gel and
627 subsequently purified using gel extraction. Samples were then submitted to Azenta's Amplicon EZ
628 service (~100,000-300,000 2x250 paired end reads) or to the UD Sequencing & Genotyping Center
629 (~100,000,000 - 150,000,000 2x150 paired end reads on an Element AVITI).

630 **Next generation sequencing analysis**

631 NGS data was first converted from FASTQ.gz format to FASTA for ease of analysis. When
632 necessary, reads were demultiplexed using the three base pair barcodes that corresponded to
633 unique sorted bins. Reads were compared to the consensus sequence for the submitted amplicon.
634 Any reads that varied from the consensus sequence (excluding the barcode and mutagenized
635 region) or those that contained stop codons (TAG, TAA, TGA) in the mutagenized region were then
636 excluded from subsequent analysis. The mutagenized region was identified from each read based
637 on a 15 base pair consensus sequence located directly after the mutagenized region and translated
638 into an amino acid sequence. Read counts for each 5 amino acid motifs across the four bins of any

given strain. From this data, the protein stability index (PSI) was calculated using the formula $PSI = \sum_{i=1}^{4} R_i * i$, where R_i is the fraction of reads present in bin i[58]. Each unique sequence, its read counts per bin, and PSI were stored in a Pandas database for subsequent analysis.

For general enrichment analysis, the counts of any given residue or motif were first measured. To calculate enrichment differences between two populations, the average occurrence of an amino acid or motif in the appropriate population was calculated and subsequently subtracted to obtain a difference. To obtain a ratio, groupings were normalized to the same sample size and subsequently divided for ratiometric comparison. Log_2 analysis was performed to simplify data visualization.

Heatmaps, boxplots, and violin plots were visualized using the Seaborn and matplotlib Python packages[67,68] as well as GraphPad Prism. Weblogos were generated using Logomaker[69]. All code was developed in Python. Code, sequence databases, and N-FIVE are accessible on GitHub: https://github.com/KunjapurLab/N-terminal-cluster-stability. Full NGS data will be made available prior to publication.

Development of N-FIVE machine learning model

During model training, we observed that applying a 20 read minimum filter minimized RMSE and improved R^2 for the model. Utilizing the sequence-PSI database, wild-type BL21 data were one-hot encoded into an 798,837 x 101 matrix. The first 100 columns represented a binary encoding of each of the 20 amino acids across 5 positions, while the final column corresponded to the PSI value. The dataset was randomly split, with 80% of the data being used for model training and 20% being used for model validation and evaluation. An extreme gradient boosted nonlinear regression model was then trained using the Python XGBoost[70] package for 20 rounds. Hyperparameter optimization was performed using Optuna[71] over 30 trials, aiming to minimize RMSE. The following parameters were set: 'n_estimators': 986, 'max_depth': 11, 'learning_rate': 0.085, 'subsample': 0.904, 'colsample_bytree': 0.810, providing an RMSE of 0.34. The final model was then saved as a .pkl file for future analysis. To evaluate the contribution of each matrix component, SHAP (Shapley Additive exPlanations) values were calculated[72] for the whole model as well as for individual sequences.

667 **Statistical analysis**

668 To quantify differences in PSI distributions between defined groups, we applied two non-
669 parametric statistical methods: the Mann-Whitney U test and Cliff's Delta for effect size
670 calculations. The Mann-Whitney U test was selected due to the skewed nature of PSI distribution
671 as well as its robustness to unequal sample sizes. We used a two-sided Mann–Whitney U test to
672 assess whether the PSI distributions between two groups differed significantly. Cliff's Delta was
673 computed to quantify the magnitude and direction of the distributional shift between compared
674 groups. Unlike common effect size measurements such as Cohen's d, Cliff's Delta makes no
675 assumptions about the underlying distribution of PSI values and remains valid for skewed,
676 bounded, or ordinal data. Here, Cliff's Delta represents the probability that a randomly selected
677 value from one group is greater than a randomly selected value from the other, minus the reverse
678 probability to yield an effect size bounded between −1 and +1. For clonal isolate testing,
679 fluorescence measurements from biological triplicates were taken for at least 100,000 gated cells
680 using flow cytometry. In line with common practices in the field, median fluorescence ratios +/-
681 standard deviation were reported.

682

683

References:

(1) Morreale, F. E.; Kleine, S.; Leodolter, J.; Junker, S.; Hoi, D. M.; Ovchinnikov, S.; Okun, A.; Kley, J.; Kurzbauer, R.; Junk, L.; Guha, S.; Podlesainski, D.; Kazmaier, U.; Boehmelt, G.; Weinstabl, H.; Rumpel, K.; Schmiedel, V. M.; Hartl, M.; Haselbach, D.; Meinhart, A.; Kaiser, M.; Clausen, T. BacPROTACs Mediate Targeted Protein Degradation in Bacteria. *Cell* **2022**, *185* (13), 2338-2353.e18. https://doi.org/10.1016/j.cell.2022.05.009.

(2) Lupoli, T. J.; Vaubourgeix, J.; Burns-Huang, K.; Gold, B. Targeting the Proteostasis Network for Mycobacterial Drug Discovery. *ACS Infect. Dis.* **2018**, *4* (4), 478–498. https://doi.org/10.1021/acsinfecdis.7b00231.

(3) d'Andrea, F. B.; Poulton, N. C.; Froom, R.; Tam, K.; Campbell, E. A.; Rock, J. M. The Essential M. Tuberculosis Clp Protease Is Functionally Asymmetric in Vivo. *Sci. Adv.* **2022**, *8* (18), eabn7943. https://doi.org/10.1126/sciadv.abn7943.

(4) Chassin, H.; Müller, M.; Tigges, M.; Scheller, L.; Lang, M.; Fussenegger, M. A Modular Degron Library for Synthetic Circuits in Mammalian Cells. *Nat. Commun.* **2019**, *10* (1), 2013. https://doi.org/10.1038/s41467-019-09974-5.

(5) Kunjapur, A. M.; Stork, D. A.; Kuru, E.; Vargas-Rodriguez, O.; Landon, M.; Söll, D.; Church, G. M. Engineering Posttranslational Proofreading to Discriminate Nonstandard Amino Acids. *Proc. Natl. Acad. Sci.* **2018**, *115* (3), 619–624. https://doi.org/10.1073/pnas.1715137115.

(6) Prindle, A.; Selimkhanov, J.; Li, H.; Razinkov, I.; Tsimring, L. S.; Hasty, J. Rapid and Tunable Post-Translational Coupling of Genetic Circuits. *Nature* **2014**, *508* (7496), 387–391. https://doi.org/10.1038/nature13238.

(7) Reed, B. D.; Meyer, M. J.; Abramzon, V.; Ad, O.; Ad, O.; Adcock, P.; Ahmad, F. R.; Alppay, G.; Ball, J. A.; Beach, J.; Belhachemi, D.; Bellofiore, A.; Bellos, M.; Beltrán, J. F.; Betts, A.; Bhuiya, M. W.; Blacklock, K.; Boer, R.; Boisvert, D.; Brault, N. D.; Buxbaum, A.; Caprio, S.; Choi, C.; Christian, T. D.; Clancy, R.; Clark, J.; Connolly, T.; Croce, K. F.; Cullen, R.; Davey, M.; Davidson, J.; Elshenawy, M. M.; Ferrigno, M.; Frier, D.; Gudipati, S.; Hamill, S.; He, Z.; Hosali, S.; Huang, H.; Huang, L.; Kabiri, A.; Kriger, G.; Lathrop, B.; Li, A.; Lim, P.; Liu, S.; Luo, F.; Lv, C.; Ma, X.; McCormack, E.; Millham, M.; Nani, R.; Pandey, M.; Parillo, J.; Patel, G.; Pike, D. H.; Preston, K.; Pichard-Kostuch, A.; Rearick, K.; Rearick, T.; Ribezzi-Crivellari, M.; Schmid, G.; Schultz, J.; Shi, X.; Singh, B.; Srivastava, N.; Stewman, S. F.; Thurston, T.; Thurston, T. R.; Trioli, P.; Tullman, J.; Wang, X.; Wang, Y.-C.; Webster, E. A. G.; Zhang, Z.; Zuniga, J.; Patel, S. S.; Griffiths, A. D.; van Oijen, A. M.; McKenna, M.; Dyer, M. D.; Rothberg, J. M. Real-Time Dynamic Single-Molecule Protein Sequencing on an Integrated Semiconductor Device. *Science* **2022**, *378* (6616), 186–192. https://doi.org/10.1126/science.abo7651.

(8) Ikonomova, S. P.; Yan, B.; Sun, Z.; Lyon, R. B.; Zatopek, K. M.; Marino, J. P.; Kelman, Z. Engineering GID4 for Use as an N-Terminal Proline Binder via Directed Evolution. *Biotechnol. Bioeng.* **2025**, *122* (1), 179–188. https://doi.org/10.1002/bit.28868.

(9) Alfaro, J. A.; Bohländer, P.; Dai, M.; Filius, M.; Howard, C. J.; van Kooten, X. F.; Ohayon, S.; Pomorski, A.; Schmid, S.; Aksimentiev, A.; Anslyn, E. V.; Bedran, G.; Cao, C.; Chinappi, M.; Coyaud, E.; Dekker, C.; Dittmar, G.; Drachman, N.; Eelkema, R.; Goodlett, D.; Hentz, S.; Kalathiya, U.; Kelleher, N. L.; Kelly, R. T.; Kelman, Z.; Kim, S. H.; Kuster, B.; Rodriguez-Larrea, D.; Lindsay, S.; Maglia, G.; Marcotte, E. M.; Marino, J. P.; Masselon, C.; Mayer, M.; Samaras, P.; Sarthak, K.; Sepiashvili, L.; Stein, D.; Wanunu, M.; Wilhelm, M.; Yin, P.; Meller, A.; Joo, C. The Emerging Landscape of Single-Molecule Protein Sequencing Technologies. *Nat. Methods* **2021**, *18* (6), 604–617. https://doi.org/10.1038/s41592-021-01143-1.

(10) Chinnaraj, M.; Lin, J.; Blacklock, K.; Hermes, E.; Meyer, M.; Pike, D.; Vieceli, J.; Chorny, I. Detecting Amino Acid Variants Using Next-Generation Protein Sequencing (NGPS). bioRxiv December 20, 2024, p 2024.12.17.629036. https://doi.org/10.1101/2024.12.17.629036.

(11) Varshavsky, A. N-Degron Pathways. *Proc. Natl. Acad. Sci.* **2024**, *121* (39), e2408697121. https://doi.org/10.1073/pnas.2408697121.
(12) Tobias, J. W.; Shrader, T. E.; Rocap, G.; Varshavsky, A. The N-End Rule in Bacteria. *Science* **1991**, *254* (5036), 1374–1377. https://doi.org/10.1126/science.1962196.
(13) Dougan, D. A.; Truscott, K. N.; Zeth, K. The Bacterial N-End Rule Pathway: Expect the Unexpected. *Mol. Microbiol.* **2010**, *76* (3), 545–558. https://doi.org/10.1111/j.1365-2958.2010.07120.x.
(14) Kim, J. G.; Shin, H.-C.; Seo, T.; Nawale, L.; Han, G.; Kim, B. Y.; Kim, S. J.; Cha-Molstad, H. Signaling Pathways Regulated by UBR Box-Containing E3 Ligases. *Int. J. Mol. Sci.* **2021**, *22* (15), 8323. https://doi.org/10.3390/ijms22158323.
(15) Dougan, D. A.; Reid, B. G.; Horwich, A. L.; Bukau, B. ClpS, a Substrate Modulator of the ClpAP Machine. *Mol. Cell* **2002**, *9*, 673–683. https://doi.org/10.1016/S1097-2765(02)00485-9.
(16) Erbse, A.; Schmidt, R.; Bornemann, T.; Schneider-Mergener, J.; Mogk, A.; Zahn, R.; Dougan, D. A.; Bukau, B. ClpS Is an Essential Component of the N-End Rule Pathway in Escherichia Coli. *Nature* **2006**, *439* (7077), 753–756. https://doi.org/10.1038/nature04412.
(17) Román-Hernández, G.; Hou, J. Y.; Grant, R. A.; Sauer, R. T.; Baker, T. A. The ClpS Adaptor Mediates Staged Delivery of N-End-Rule Substrates to the AAA+ ClpAP Protease. *Mol. Cell* **2011**, *43* (2), 217–228. https://doi.org/10.1016/j.molcel.2011.06.009.
(18) Yeom, J.; Gao, X.; Groisman, E. A. Reduction in Adaptor Amounts Establishes Degradation Hierarchy among Protease Substrates. *Proc. Natl. Acad. Sci.* **2018**, *115* (19), E4483–E4492. https://doi.org/10.1073/pnas.1722246115.
(19) Humbard, M. A.; Surkov, S.; De Donatis, G. M.; Jenkins, L. M.; Maurizi, M. R. The N-Degradome of Escherichia Coli: Limited Proteolysis in Vivo Generates a Large Pool of Proteins Bearing N-Degrons. *J. Biol. Chem.* **2013**, *288* (40), 28913–28924. https://doi.org/10.1074/jbc.M113.492108.
(20) Ninnis, R. L.; Spall, S. K.; Talbo, G. H.; Truscott, K. N.; Dougan, D. A. Modification of PATase by L/F-transferase Generates a ClpS-dependent N-end Rule Substrate in Escherichia Coli. *EMBO J.* **2009**, *28* (12), 1732–1744. https://doi.org/10.1038/emboj.2009.134.
(21) Watanabe, K.; Toh, Y.; Suto, K.; Shimizu, Y.; Oka, N.; Wada, T.; Tomita, K. Protein-Based Peptide-Bond Formation by Aminoacyl-tRNA Protein Transferase. *Nature* **2007**, *449* (7164), 867–871. https://doi.org/10.1038/nature06167.
(22) Gao, X.; Yeom, J.; Groisman, E. A. The Expanded Specificity and Physiological Role of a Widespread N-Degron Recognin. *Proc. Natl. Acad. Sci.* **2019**, *116* (37), 18629–18637. https://doi.org/10.1073/pnas.1821060116.
(23) Wang, K. H.; Roman-Hernandez, G.; Grant, R. A.; Sauer, R. T.; Baker, T. A. The Molecular Basis of N-End Rule Recognition. *Mol. Cell* **2008**, *32* (3), 406–414. https://doi.org/10.1016/j.molcel.2008.08.032.
(24) Ottofuelling, R. D.; Ninnis, R. L.; Truscott, K. N.; Dougan, D. A. Novel Modification by L/F-tRNA-Protein Transferase (LFTR) Generates a Leu/N-Degron Ligand in Escherichia Coli. bioRxiv July 27, 2021, p 2021.07.26.453911. https://doi.org/10.1101/2021.07.26.453911.
(25) Stein, B. J.; Grant, R. A.; Sauer, R. T.; Baker, T. A. Structural Basis of an N-Degron Adaptor with More Stringent Specificity. *Struct. Lond. Engl. 1993* **2016**, *24* (2), 232–242. https://doi.org/10.1016/j.str.2015.12.008.
(26) Wang, K. H.; Oakes, E. S. C.; Sauer, R. T.; Baker, T. A. Tuning the Strength of a Bacterial N-End Rule Degradation Signal. *J. Biol. Chem.* **2008**, *283* (36), 24600–24607. https://doi.org/10.1074/jbc.M802213200.
(27) Yeom, J.; Wayne, K. J.; Groisman, E. A. Sequestration from Protease Adaptor Confers Differential Stability to Protease Substrate. *Mol. Cell* **2017**, *66* (2), 234-246.e5. https://doi.org/10.1016/j.molcel.2017.03.009.

(28) Hempfling, J. P.; Sekera, E. R.; Sarkar, A.; Hummon, A. B.; Pei, D. Generation of Proteins with Free N-Terminal Cysteine by Aminopeptidases. *J. Am. Chem. Soc.* **2022**, *144* (47), 21763–21771. https://doi.org/10.1021/jacs.2c10194.
(29) Tolbert, T. J.; Wong, C.-H. New Methods for Proteomic Research: Preparation of Proteins with N-Terminal Cysteines for Labeling and Conjugation. *Angew. Chem. Int. Ed.* **2002**, *41* (12), 2171–2174. https://doi.org/10.1002/1521-3773(20020617)41:12<2171::AID-ANIE2171>3.0.CO;2-Q.
(30) Rosen, C. B.; Francis, M. B. Targeting the N Terminus for Site-Selective Protein Modification. *Nat. Chem. Biol.* **2017**, *13* (7), 697–705. https://doi.org/10.1038/nchembio.2416.
(31) Paupelin-Vaucelle, H.; Boschiero, C.; Lazennec-Schurdevin, C.; Schmitt, E.; Mechulam, Y.; Marlière, P.; Pezo, V. Cys-tRNAj as a Second Translation Initiator for Priming Proteins with Cysteine in Bacteria. *ACS Omega* **2025**, *10* (5), 4548–4560. https://doi.org/10.1021/acsomega.4c08326.
(32) Wang, H.; Wang, L.; Zhong, B.; Dai, Z. Protein Splicing of Inteins: A Powerful Tool in Synthetic Biology. *Front. Bioeng. Biotechnol.* **2022**, *10*. https://doi.org/10.3389/fbioe.2022.810180.
(33) Theile, C. S.; Witte, M. D.; Blom, A. E. M.; Kundrat, L.; Ploegh, H. L.; Guimaraes, C. P. Site-Specific N-Terminal Labeling of Proteins Using Sortase-Mediated Reactions. *Nat. Protoc.* **2013**, *8* (9), 1800–1807. https://doi.org/10.1038/nprot.2013.102.
(34) Williamson, D. J.; Fascione, M. A.; Webb, M. E.; Turnbull, W. B. Efficient N-Terminal Labeling of Proteins by Use of Sortase. *Angew. Chem. Int. Ed.* **2012**, *51* (37), 9377–9380. https://doi.org/10.1002/anie.201204538.
(35) Yoshihara, H. A. I.; Mahrus, S.; Wells, J. A. Tags for Labeling Protein N-Termini with Subtiligase for Proteomics. *Bioorg. Med. Chem. Lett.* **2008**, *18* (22), 6000–6003. https://doi.org/10.1016/j.bmcl.2008.08.044.
(36) Weeks, A. M.; Wells, J. A. N-Terminal Modification of Proteins with Subtiligase Specificity Variants. *Curr. Protoc. Chem. Biol.* **2020**, *12* (1), e79. https://doi.org/10.1002/cpch.79.
(37) Wang, K. H.; Roman-Hernandez, G.; Grant, R. A.; Sauer, R. T.; Baker, T. A. The Molecular Basis of N-End Rule Recognition. *Mol. Cell* **2008**, *32* (3), 406–414. https://doi.org/10.1016/j.molcel.2008.08.032.
(38) Presloid, C. J.; Jiang, J.; Kandel, P.; Anderson, H. R.; Beardslee, P. C.; Swayne, T. M.; Schmitz, K. R. ClpS Directs Degradation of N-Degron Substrates With Primary Destabilizing Residues in Mycolicibacterium Smegmatis. *Mol. Microbiol.* **2025**, *123* (1), 16–30. https://doi.org/10.1111/mmi.15334.
(39) Timms, R. T.; Zhang, Z.; Rhee, D. Y.; Harper, J. W.; Koren, I.; Elledge, S. J. A Glycine-Specific N-Degron Pathway Mediates the Quality Control of Protein N-Myristoylation. *Science* **2019**, *365* (6448), eaaw4912. https://doi.org/10.1126/SCIENCE.AAW4912.
(40) Khmelinskii, A.; Keller, P. J.; Bartosik, A.; Meurer, M.; Barry, J. D.; Mardin, B. R.; Kaufmann, A.; Trautmann, S.; Wachsmuth, M.; Pereira, G.; Huber, W.; Schiebel, E.; Knop, M. Tandem Fluorescent Protein Timers for in Vivo Analysis of Protein Dynamics. *Nat. Biotechnol.* **2012**, *30* (7), 708–714. https://doi.org/10.1038/nbt.2281.
(41) Kats, I.; Khmelinskii, A.; Kschonsak, M.; Huber, F.; Knieß, R. A.; Bartosik, A.; Knop, M. Mapping Degradation Signals and Pathways in a Eukaryotic N-Terminome. *Mol. Cell* **2018**, *70* (3), 488-501.e5. https://doi.org/10.1016/j.molcel.2018.03.033.
(42) Beardslee, P. C.; Schmitz, K. R. Toxin-Based Screening of C-Terminal Tags in Escherichia Coli Reveals the Exceptional Potency of ssrA-like Degrons. *eLife* **2024**, *13*. https://doi.org/10.7554/eLife.98528.1.
(43) Varshavsky, A. Ubiquitin Fusion Technique and Related Methods. In *Methods in Enzymology*; Ubiquitin and Protein Degradation, Part B; Academic Press, 2005; Vol. 399, pp 777–799. https://doi.org/10.1016/S0076-6879(05)99051-4.

(44) Stenström, C. M.; Holmgren, E.; Isaksson, L. A. Cooperative Effects by the Initiation Codon and Its Flanking Regions on Translation Initiation. *Gene* **2001**, *273* (2), 259–265. https://doi.org/10.1016/S0378-1119(01)00584-4.
(45) Stenström, C. M.; Jin, H.; Major, L. L.; Tate, W. P.; Isaksson, L. A. Codon Bias at the 3'-Side of the Initiation Codon Is Correlated with Translation Initiation Efficiency in *Escherichia Coli*. *Gene* **2001**, *263* (1), 273–284. https://doi.org/10.1016/S0378-1119(00)00550-3.
(46) Zhang, W.; Xiao, W.; Wei, H.; Zhang, J.; Tian, Z. mRNA Secondary Structure at Start AUG Codon Is a Key Limiting Factor for Human Protein Expression in *Escherichia Coli*. *Biochem. Biophys. Res. Commun.* **2006**, *349* (1), 69–78. https://doi.org/10.1016/j.bbrc.2006.07.209.
(47) Bivona, L.; Zou, Z.; Stutzman, N.; Sun, P. D. Influence of the Second Amino Acid on Recombinant Protein Expression. *Protein Expr. Purif.* **2010**, *74* (2), 248–256. https://doi.org/10.1016/j.pep.2010.06.005.
(48) Allert, M.; Cox, J. C.; Hellinga, H. W. Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. *J. Mol. Biol.* **2010**, *402* (5), 905–918. https://doi.org/10.1016/j.jmb.2010.08.010.
(49) Goodman, D. B.; Church, G. M.; Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **2013**, *342* (6157), 475–479. https://doi.org/10.1126/science.1241934.
(50) Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M. Composability of Regulatory Sequences Controlling Transcription and Translation in Escherichia Coli. *Proc. Natl. Acad. Sci.* **2013**, *110* (34), 14024–14029. https://doi.org/10.1073/pnas.1301301110.
(51) Ojima-Kato, T.; Nagai, S.; Nakano, H. N-Terminal SKIK Peptide Tag Markedly Improves Expression of Difficult-to-Express Proteins in *Escherichia Coli* and *Saccharomyces Cerevisiae*. *J. Biosci. Bioeng.* **2017**, *123* (5), 540–546. https://doi.org/10.1016/j.jbiosc.2016.12.004.
(52) Ojima-Kato, T.; Nishikawa, Y.; Furukawa, Y.; Kojima, T.; Nakano, H. Nascent MSKIK Peptide Cancels Ribosomal Stalling by Arrest Peptides in Escherichia Coli. *J. Biol. Chem.* **2023**, *299* (5). https://doi.org/10.1016/j.jbc.2023.104676.
(53) Wang, H. H.; Isaacs, F. J.; Carr, P. A.; Sun, Z. Z.; Xu, G.; Forest, C. R.; Church, G. M. Programming Cells by Multiplex Genome Engineering and Accelerated Evolution. *Nature* **2009**, *460* (7257), 894–898. https://doi.org/10.1038/nature08187.
(54) Wannier, T. M.; Ciaccia, P. N.; Ellington, A. D.; Filsinger, G. T.; Isaacs, F. J.; Javanmardi, K.; Jones, M. A.; Kunjapur, A. M.; Nyerges, A.; Pal, C.; Schubert, M. G.; Church, G. M. Recombineering and MAGE. *Nat. Rev. Methods Primer* **2021**, *1* (1), 7. https://doi.org/10.1038/s43586-020-00006-x.
(55) Wang, K. H.; Sauer, R. T.; Baker, T. A. ClpS Modulates but Is Not Essential for Bacterial N-End Rule Degradation. *Genes Dev.* **2007**, *21* (4), 403–408. https://doi.org/10.1101/gad.1511907.
(56) Sen, S.; Kho, K. K.; Kunjapur, A. Developing a Flow Cytometric Method to Evaluate the Stability of Protein N-Termini. 2025.
(57) Licausi, F.; Kosmacz, M.; Weits, D. A.; Giuntoli, B.; Giorgi, F. M.; Voesenek, L. A. C. J.; Perata, P.; van Dongen, J. T. Oxygen Sensing in Plants Is Mediated by an N-End Rule Pathway for Protein Destabilization. *Nature* **2011**, *479* (7373), 419–422. https://doi.org/10.1038/nature10536.
(58) Yen, H.-C. S.; Xu, Q.; Chou, D. M.; Zhao, Z.; Elledge, S. J. Global Protein Stability Profiling in Mammalian Cells. *Science* **2008**, *322* (5903), 918–923. https://doi.org/10.1126/science.1160489.
(59) Makaros, Y.; Raiff, A.; Timms, R. T.; Wagh, A. R.; Gueta, M. I.; Bekturova, A.; Guez-Haddad, J.; Brodsky, S.; Opatowsky, Y.; Glickman, M. H.; Elledge, S. J.; Koren, I. Ubiquitin-Independent Proteasomal Degradation Driven by C-Degron Pathways. *Mol. Cell* **2023**, *83* (11), 1921-1935.e7. https://doi.org/10.1016/j.molcel.2023.04.023.
(60) Zhang, Z.; Sie, B.; Chang, A.; Leng, Y.; Nardone, C.; Timms, R. T.; Elledge, S. J. Elucidation of E3 Ubiquitin Ligase Specificity Through Proteome-Wide Internal Degron Mapping. *Mol. Cell* **2023**, *83* (18), 3377-3392.e6. https://doi.org/10.1016/j.molcel.2023.08.022.

(61) Ben-Bassat, A.; Bauer, K.; Chang, S. Y.; Myambo, K.; Boosman, A.; Chang, S. Processing of the Initiation Methionine from Proteins: Properties of the Escherichia Coli Methionine Aminopeptidase and Its Gene Structure. *J. Bacteriol.* **1987**, *169* (2), 751–757. https://doi.org/10.1128/jb.169.2.751-757.1987.

(62) Xiao, Q.; Zhang, F.; Nacev, B. A.; Liu, J. O.; Pei, D. Protein N-Terminal Processing: Substrate Specificity of Escherichia Coli and Human Methionine Aminopeptidases. *Biochemistry* **2010**, *49* (26), 5588–5599. https://doi.org/10.1021/bi1005464.

(63) Schuenemann, V. J.; Kralik, S. M.; Albrecht, R.; Spall, S. K.; Truscott, K. N.; Dougan, D. A.; Zeth, K. Structural Basis of N-end Rule Substrate Recognition in *Escherichia Coli* by the ClpAP Adaptor Protein ClpS. *EMBO Rep.* **2009**, *10* (5), 508–514. https://doi.org/10.1038/embor.2009.62.

(64) Wojtowicz, A.; Mazurkiewicz-Pisarek, A.; Plucienniczak, G.; Mikiewicz-Sygula, D.; Chojnacka, L.; Lukasiewicz, N.; Plucienniczak, A. Expression of Yeast Deubiquitination Enzyme UBP1 Analogues in E. Coli. *Microb. Cell Factories* **2005**, *4*, 17. https://doi.org/10.1186/1475-2859-4-17.

(65) Wannier, T. M.; Ciaccia, P. N.; Ellington, A. D.; Filsinger, G. T.; Isaacs, F. J.; Javanmardi, K.; Jones, M. A.; Kunjapur, A. M.; Nyerges, A.; Pal, C.; Schubert, M. G.; Church, G. M. Recombineering and MAGE. *Nat. Rev. Methods Primer* **2021**, *1* (1), 1–24. https://doi.org/10.1038/s43586-020-00006-x.

(66) Bonde, M. T.; Klausen, M. S.; Anderson, M. V.; Wallin, A. I. N.; Wang, H. H.; Sommer, M. O. A. MODEST: A Web-Based Design Tool for Oligonucleotide-Mediated Genome Engineering and Recombineering. *Nucleic Acids Res.* **2014**, *42* (W1), W408–W415. https://doi.org/10.1093/nar/gku428.

(67) Waskom, M. L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.

(68) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.

(69) Tareen, A.; Kinney, J. B. Logomaker: Beautiful Sequence Logos in Python. *Bioinformatics* **2020**, *36* (7), 2272–2274. https://doi.org/10.1093/bioinformatics/btz921.

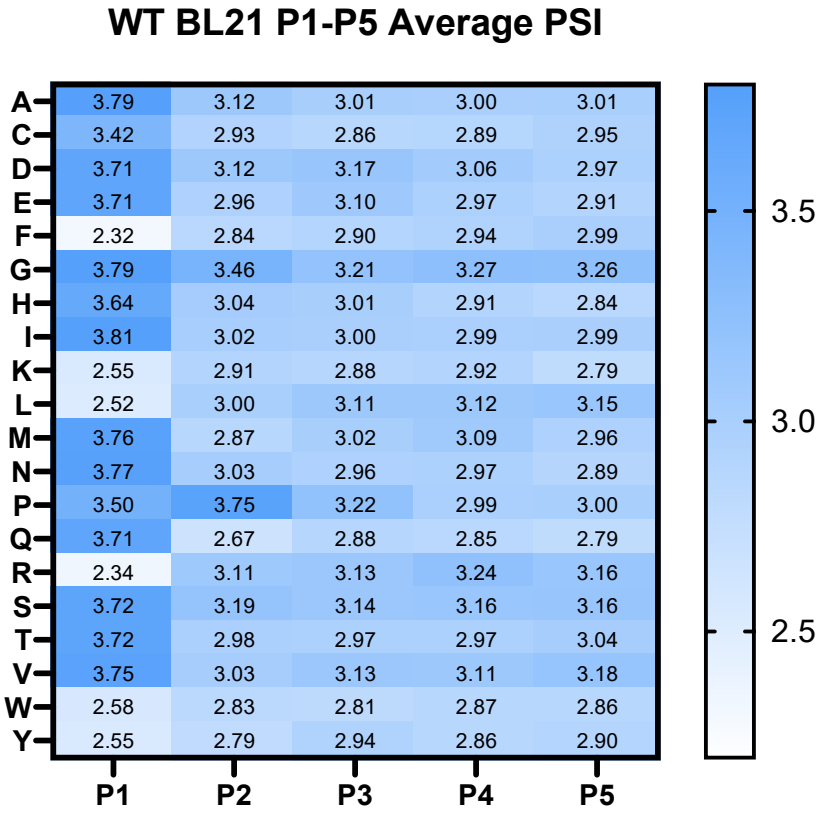
(70) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD ’16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(71) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; KDD ’19; Association for Computing Machinery: New York, NY, USA, 2019; pp 2623–2631. https://doi.org/10.1145/3292500.3330701.

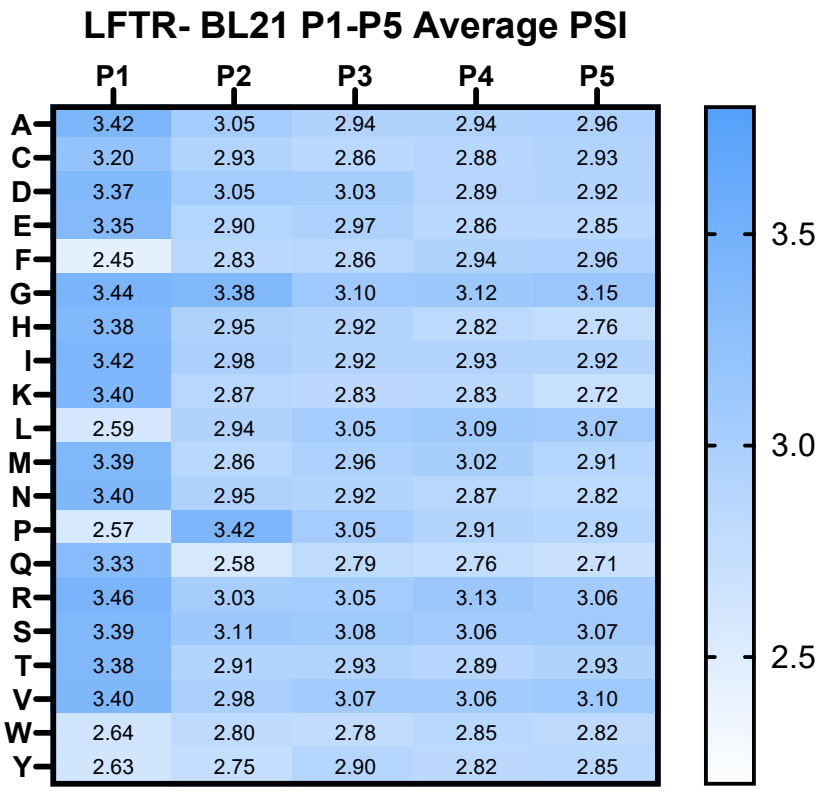
(72) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; NIPS’17; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp 4768–4777.

Supplementary Materials

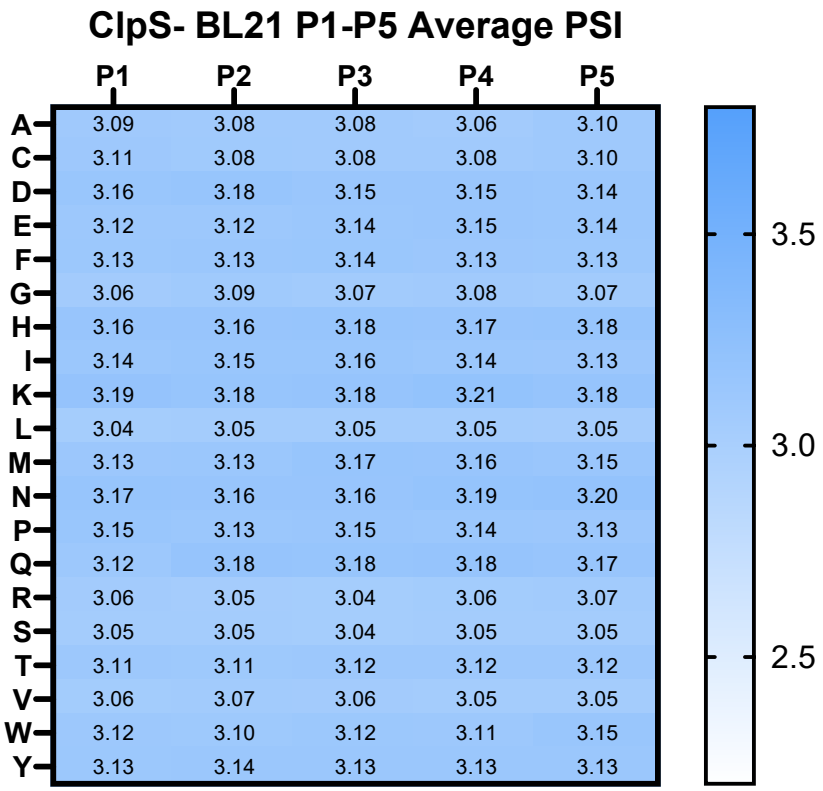
Supplemental Figures



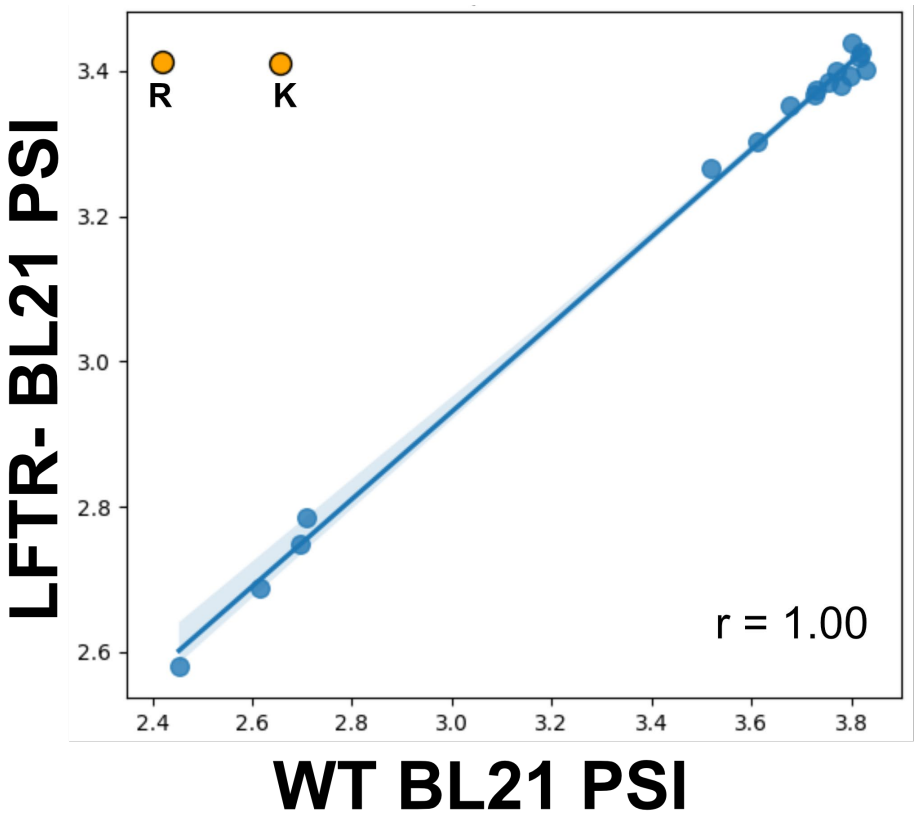
Supplemental Figure 1. P1-P5 WT BL21 Heatmap A mean PSI heatmap generated from 2.29 million sequences for the first five positions of a protein in wild-type BL21.



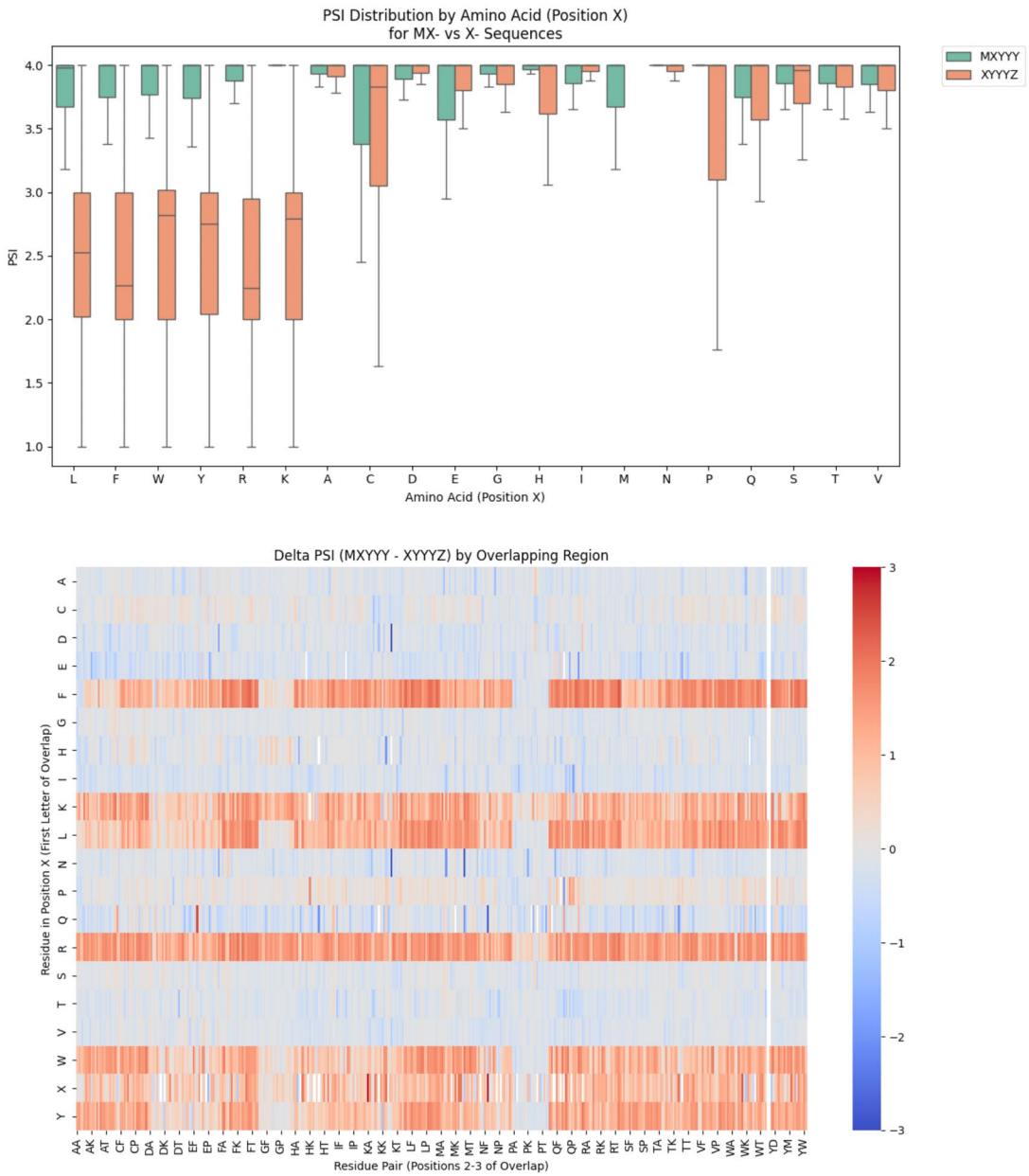
Supplemental Figure 2. P1-P5 LFTR- BL21 Heatmap A mean PSI heatmap generated from 2.34 million sequences for the first five positions of a protein in BL21 deficient in *aat* (LFTR).



Supplemental Figure 3. P1-P5 ClpS- BL21 Heatmap A mean PSI heatmap generated from 2.19 million sequences for the first five positions of a protein in BL21 deficient in *clpS*.



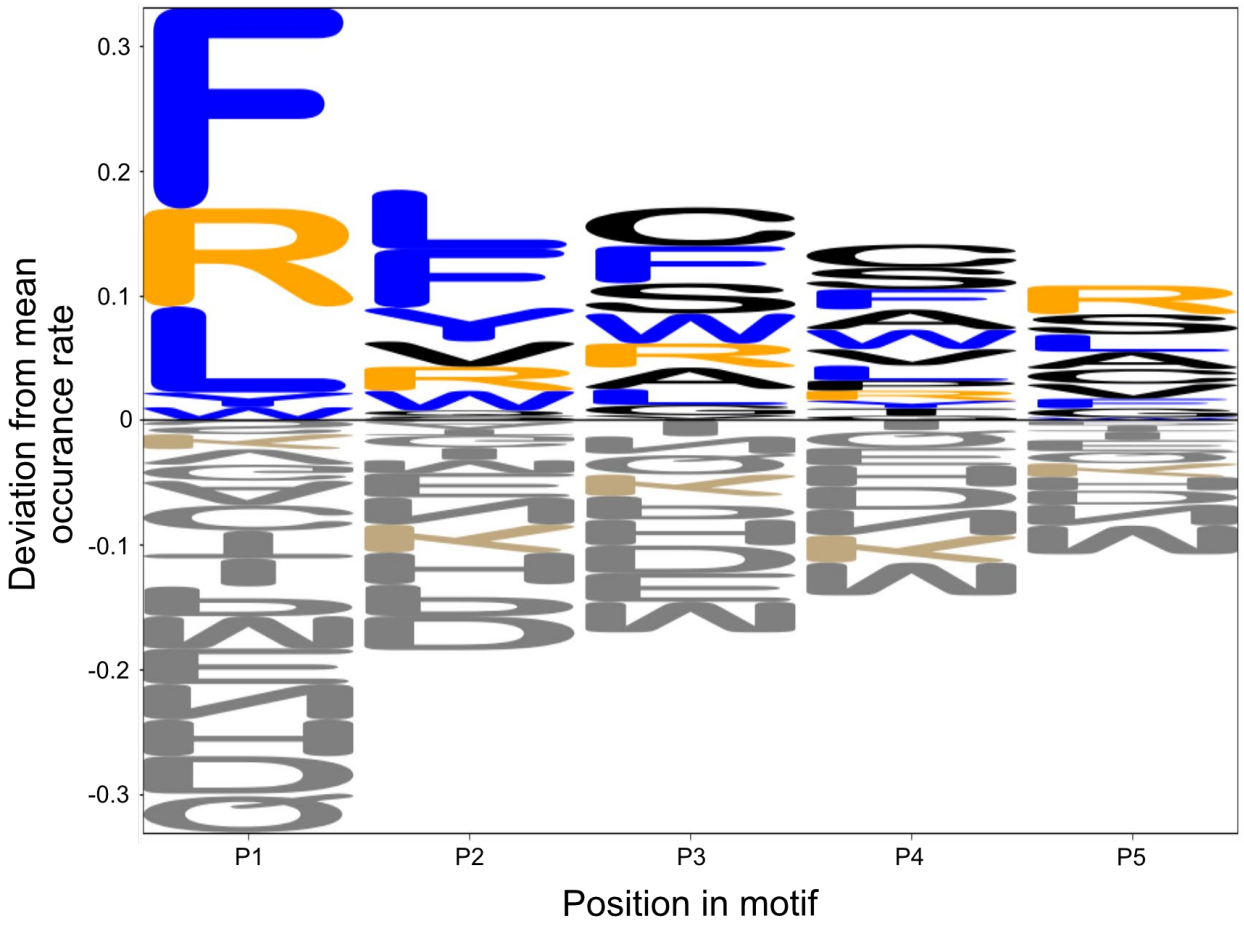
Supplemental Figure 4. P1 amino acid correlation between WT BL21 and LFTR- BL21 sorts Correlation between PSI scores for N-terminal residues between WT BL21 and LFTR- BL21 (r = 1.00) in the 10 million event sort. Arg/Lys/Pro were excluded from regression analysis. Arg and Lys separately highlighted in orange. Ubiquitin upstream of P1 Pro is not fully cleaved, complicating analysis.



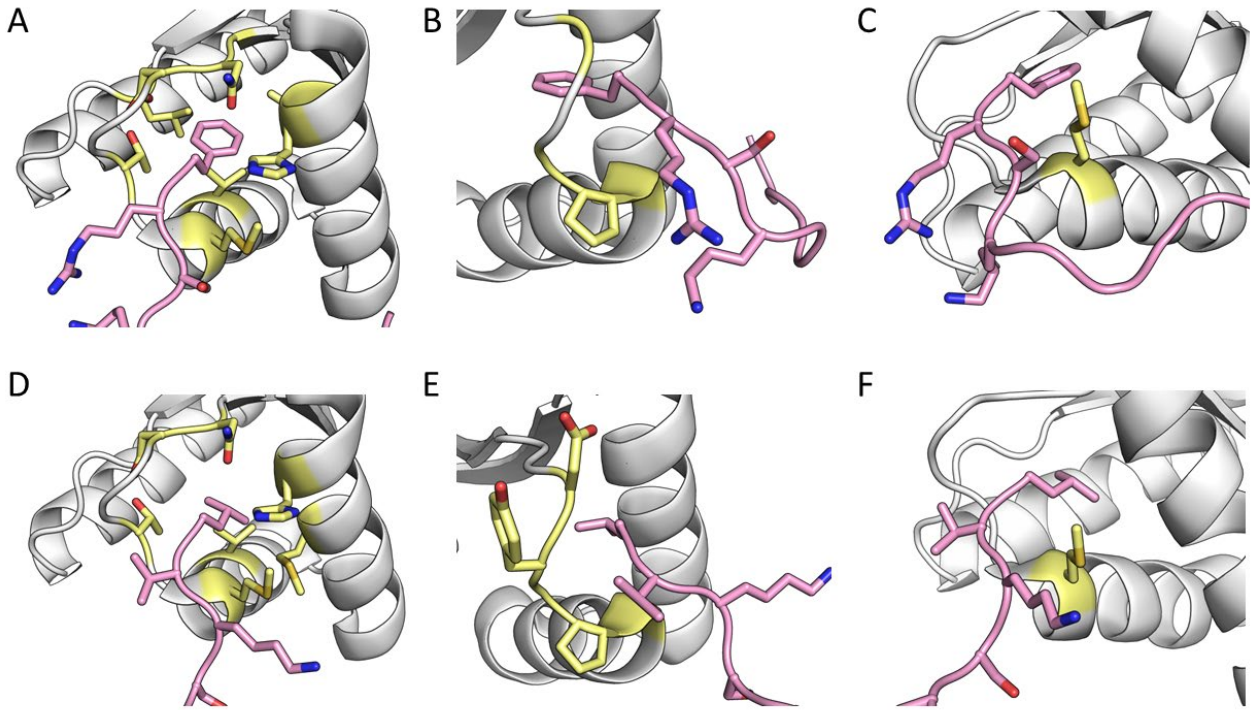
931

932 **Supplemental Figure 5. MX- Motifs do not show similar PSI distributions for P1 LFWYRK**
933 **motifs.** (Top) Mean PSI boxplots for residues in position X from the 10M event sort. MXYYY is in
934 green and XYYYZ is in orange. (Bottom) Change in PSI between MXYYY and XYYYZ datasets, with
935 the X-position on the y-axis and the first two Y positions in the x-axis.

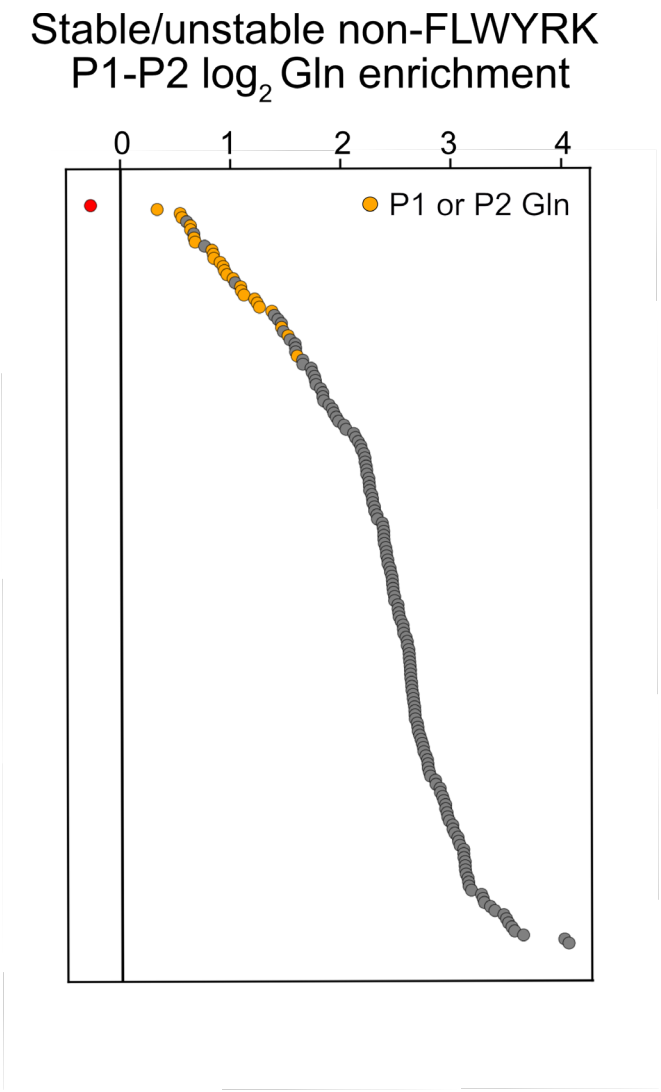
936



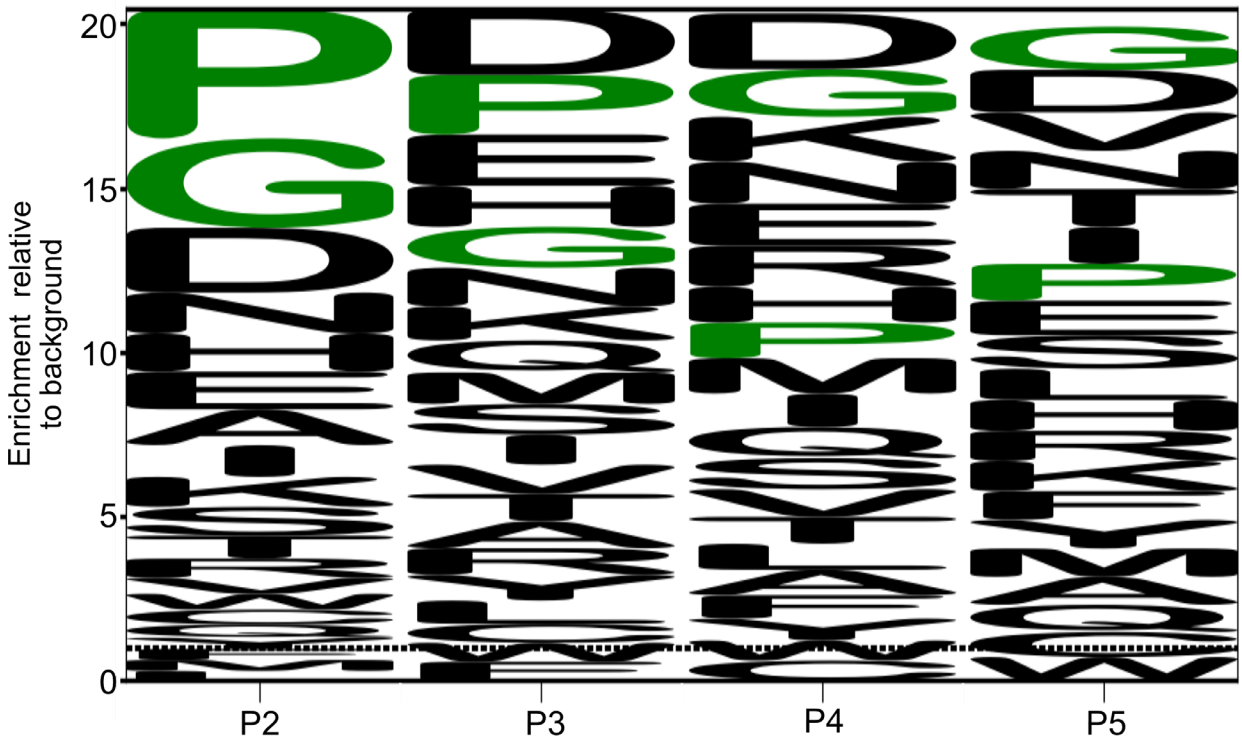
Supplemental Figure 6. Bulky and positively charged amino acids are enriched in low stability sequences. Weblogo depicting the difference in enrichment for each amino acid in each of the five positions in low PSI (PSI <2) samples compared to the entire dataset.



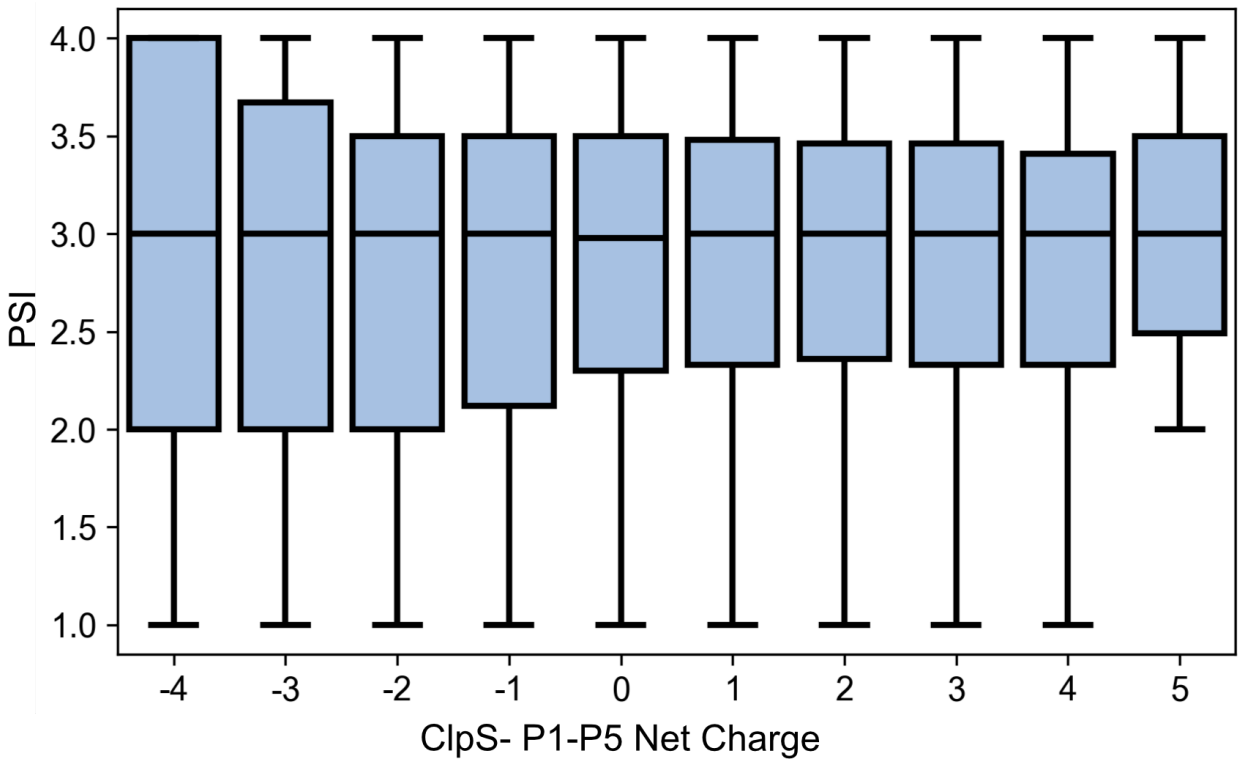
Supplemental Figure 7. Structures of ClpS bound to peptide substrates provide limited insight into the molecular basis of substrate specificity downstream of P1. A-C: Structures of ClpS bound to Fpep with nearby ClpS residues highlighted in yellow for (A) P1, (B) P2, and (C) P3. (PDB ID: 2WA8) D-F: Structures of ClpS bound to Lpep with nearby ClpS residues highlighted in yellow for (D) P1, (E) P2, and (F) P3. (PDB ID: 2W9R). No interacting residues within 4 A were obtained for P4 and P5 of the peptides.



Supplemental Figure 8. P1/P2 Gln is noticeably enriched in low stability P1-P2 motifs that do not have P1 FLWYRK. Log_2 enrichment of P1-P2 motifs in low stability (PSI <2) motifs relative to high stability (PSI >3) motifs where FLWYRK are excluded from P1. Motifs with Gln in P1 or P2 are in orange, QQ is in red.



Supplemental Figure 9. Glycine and proline are highly enriched in high stability P1 FLWYRK sequences. A P2-P5 WebLogo showing high stability P1 FLWYRK (PSI >3) sequence enrichment relative to the entire wild-type dataset for each of the five positions studied. Glycine and proline highlighted in green.

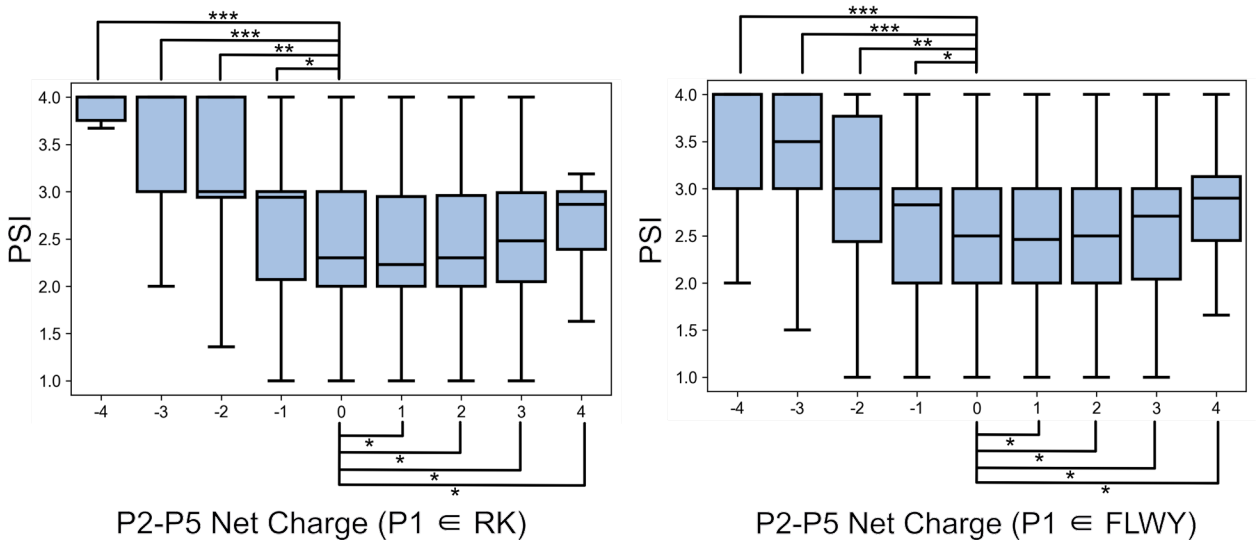


963

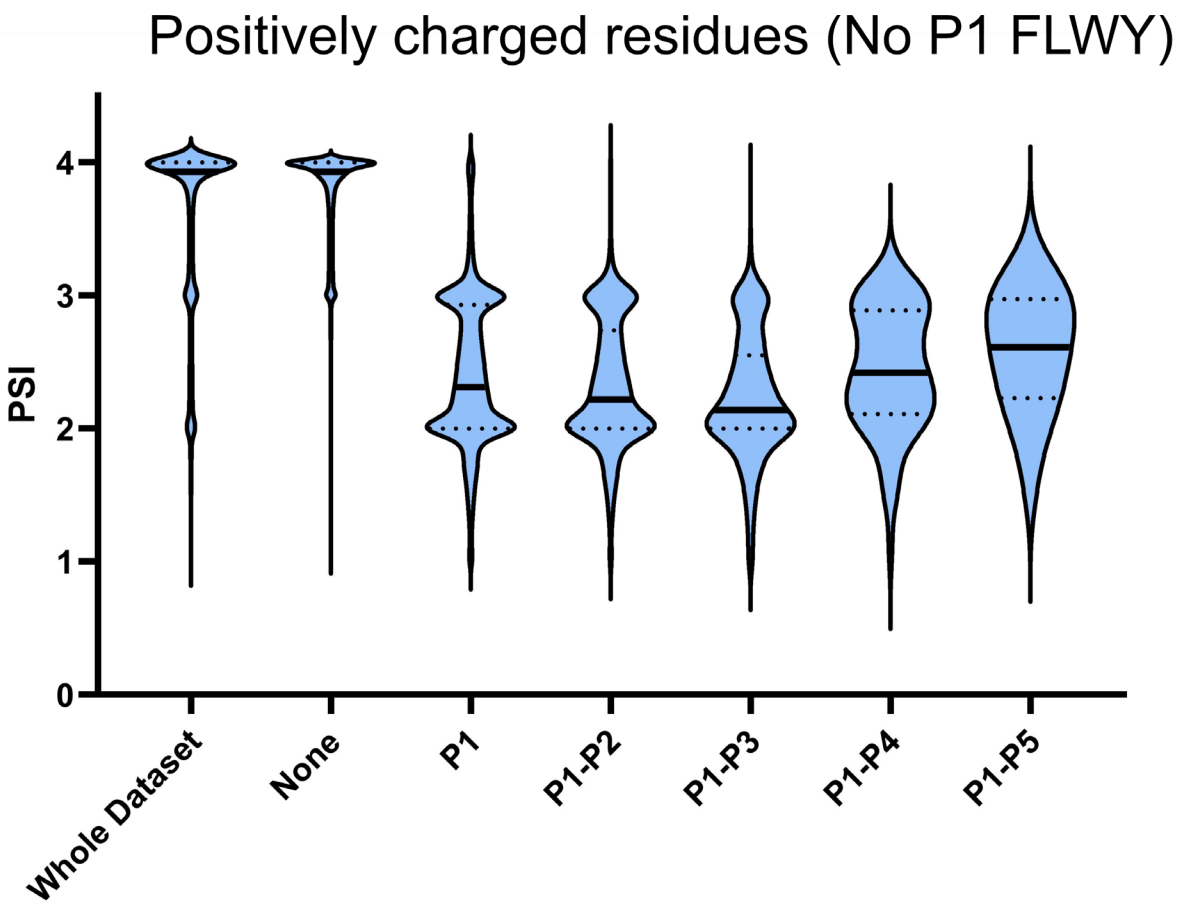
964 **Supplemental Figure 10. ClpS- sequences show charge parity.** P1-P5 net charge heatmap for

965 data collected in ClpS- Bl21.

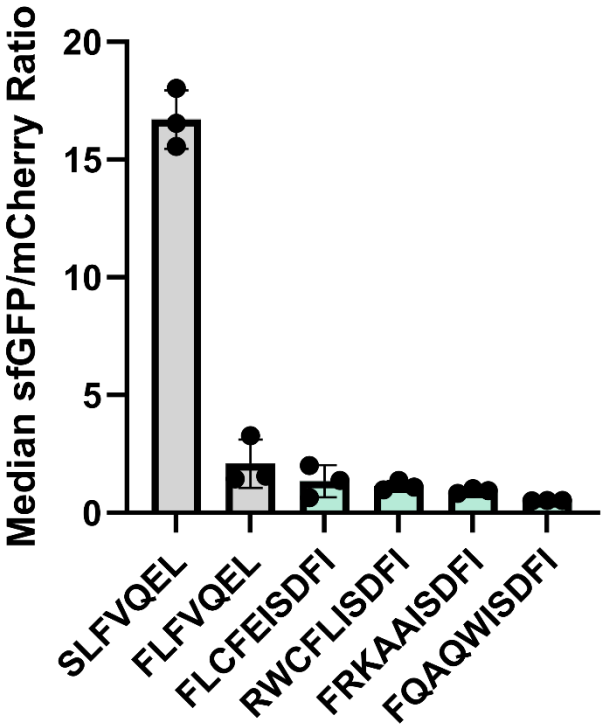
966



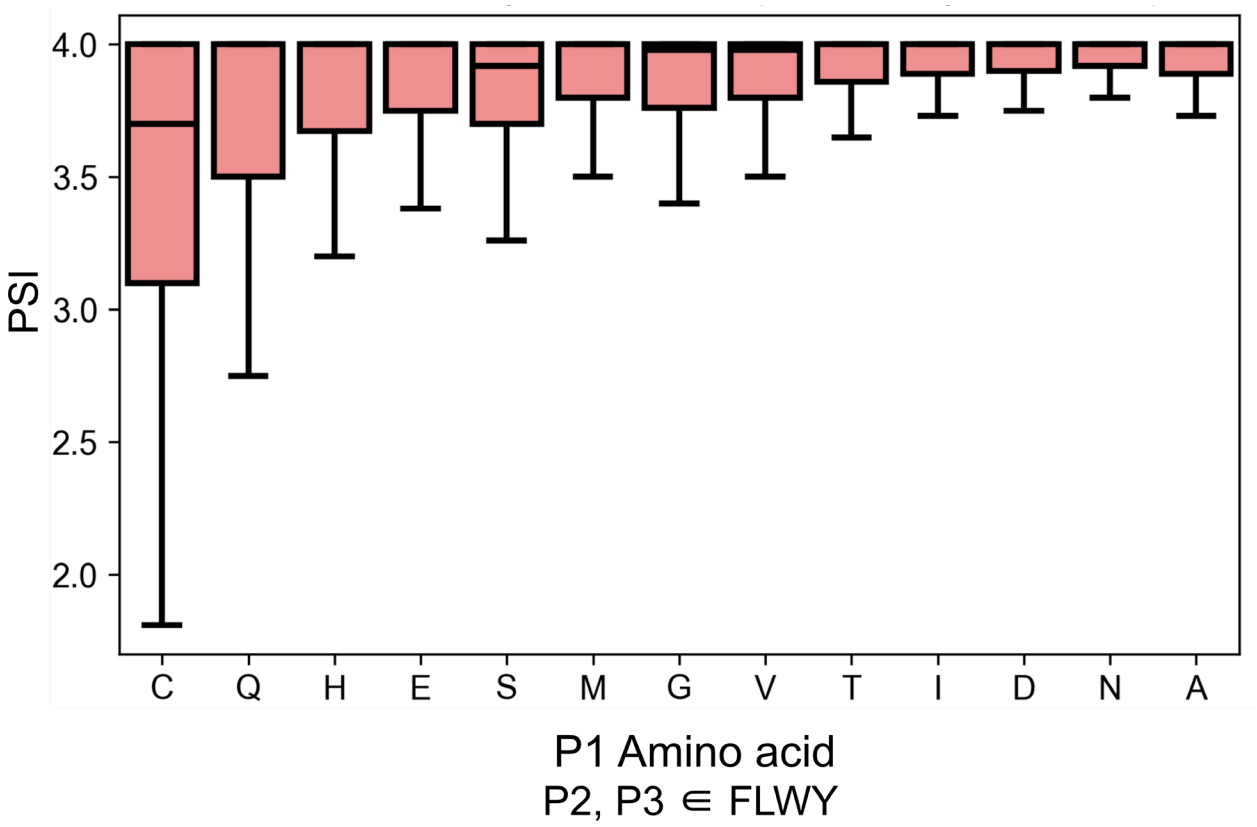
Supplemental Figure 11. Separation of charged and bulky P1 residue net charge heatmaps. All p-values in comparison to 0 P2-P5 Gly/Ser show p <0.05. p-values evaluated using a Mann-Whitney U test, effect size magnitude (ES) ranges from 0 to 1, as evaluated using Cliff's delta. * = p <0.05, ES 0-0.25; ** = p <0.05, ES 0.25-0.50, *** = p <0.05, ES 0.50-0.75, **** = p <0.05, ES 0.75-1.0



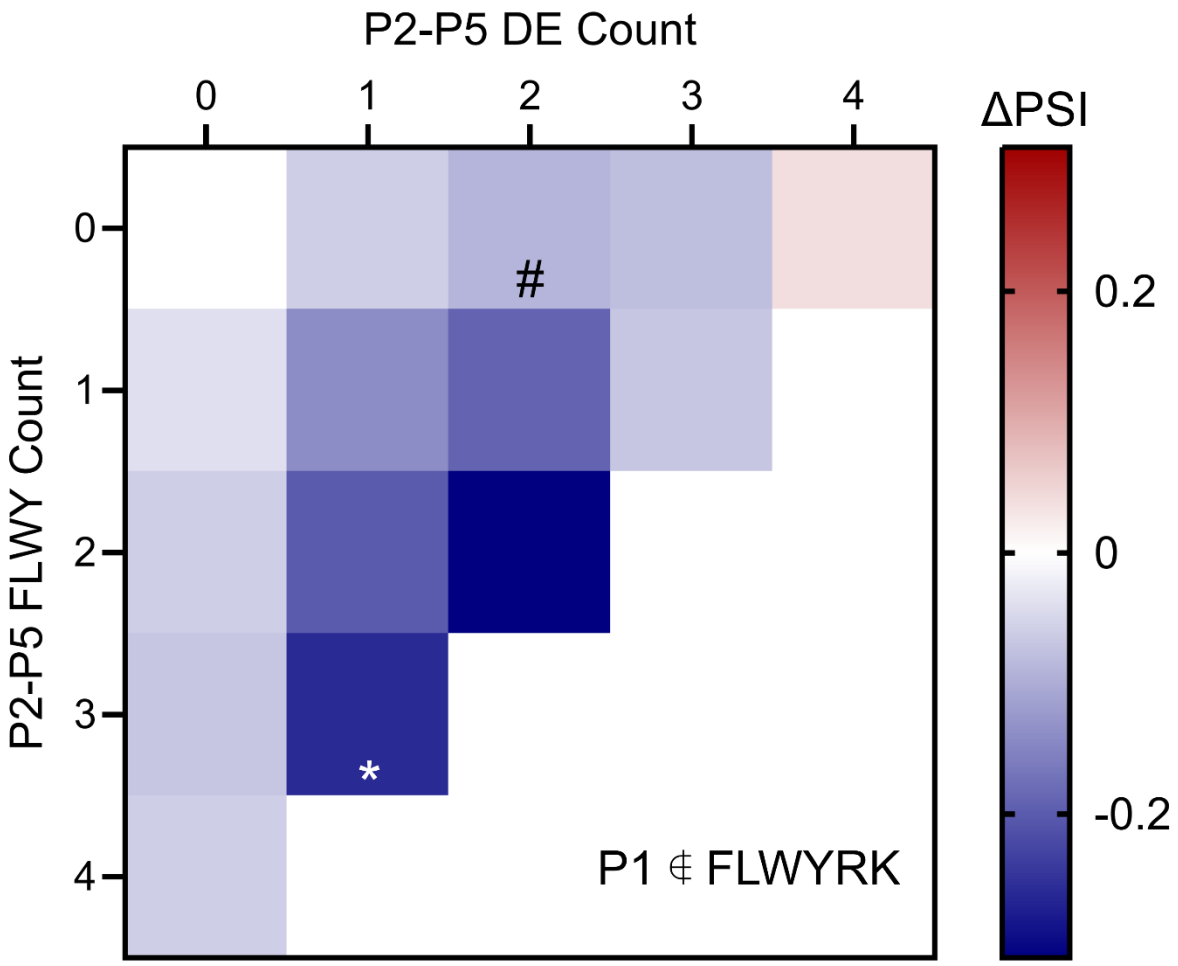
Supplemental Figure 12. Multiple positively charged residues at and near P1 lead to low PSI sequences. Violin plot depicting the PSI distribution in wild-type BL21 for sequences with one or more RK residues in series at and adjacent to the N-terminus.



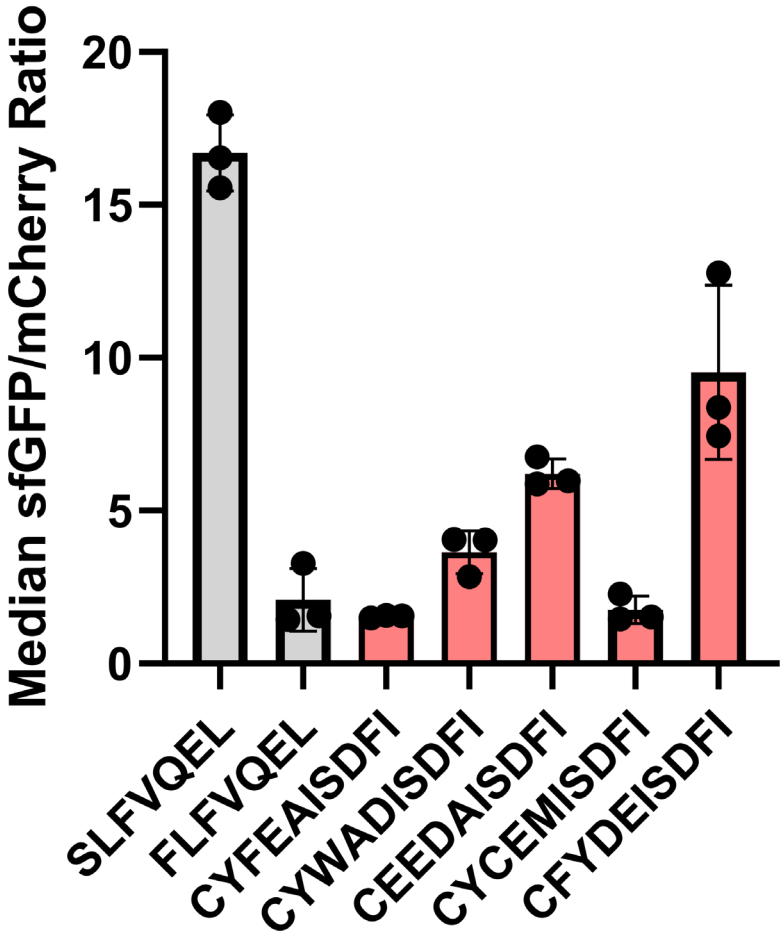
Supplemental Figure 13. Sequences with multiple FLWYRK(Q) residues show low stability. Flow cytometry data from the protein stability assay depicting sequences carrying multiple FLWYRK or P2 Gln residues exhibited low sfGFP:mCherry ratios relative to literature controls.



Supplemental Figure 14. P1 Cys is the most destabilizing non FLWYRK residue when P2 and P3 are bulky. Boxplot showing PSI distributions for non-P1 FLWYRKP motifs that have P2 and P3 residues in FLWY.



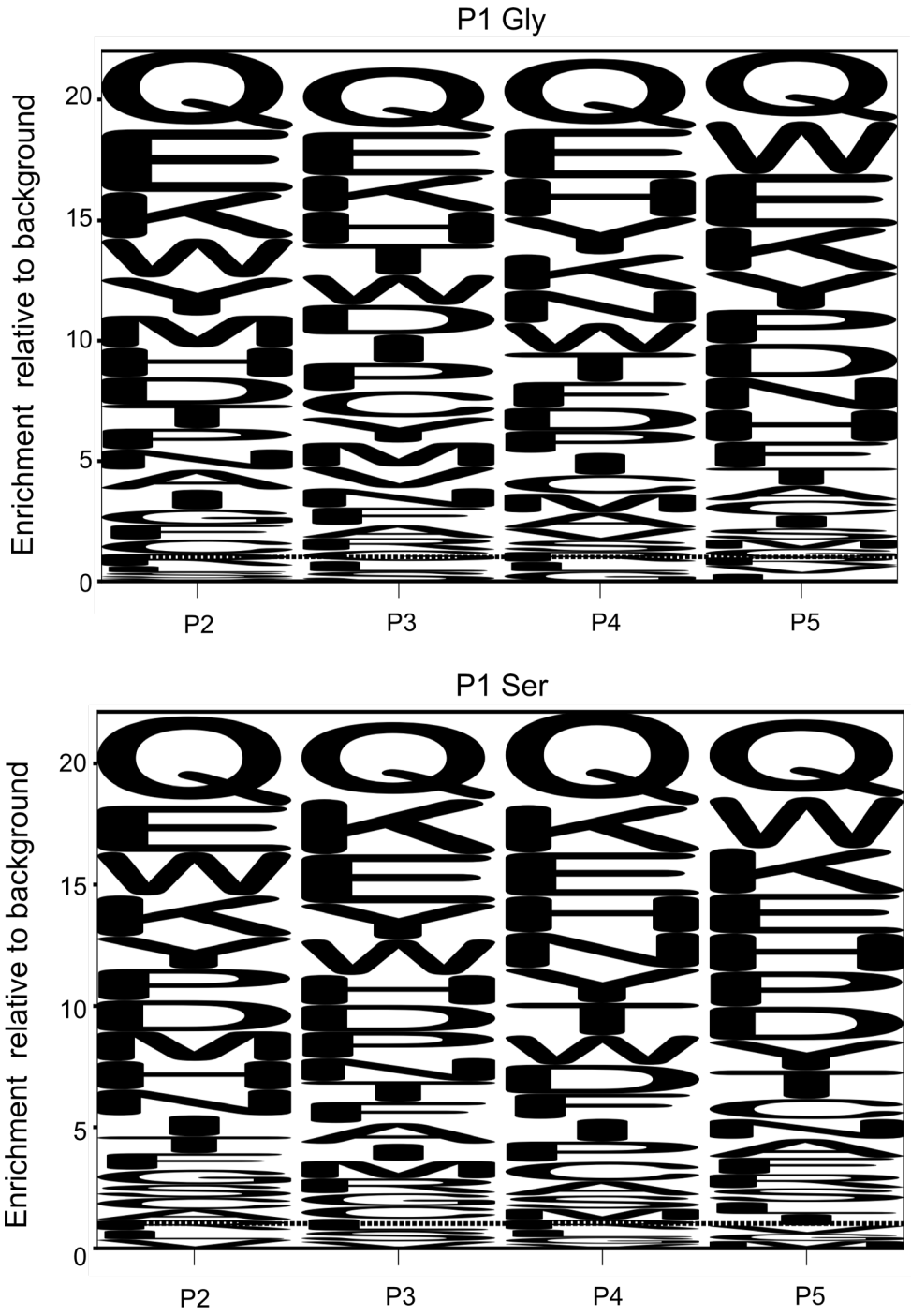
Supplemental Figure 15. For WT sequences with canonically stable P1, the largest PSI drops are observed for motifs with both bulky and negatively charged P2-P5 residues. Change in PSI is relative to the 0 P2-P5 FLWY, 0 P2-P5 DE dataset (mean PSI = 3.78). All p-values relative to (0,0) are p <0.5 except cells marked with # which are p >0.05. * = ES > 0.25.



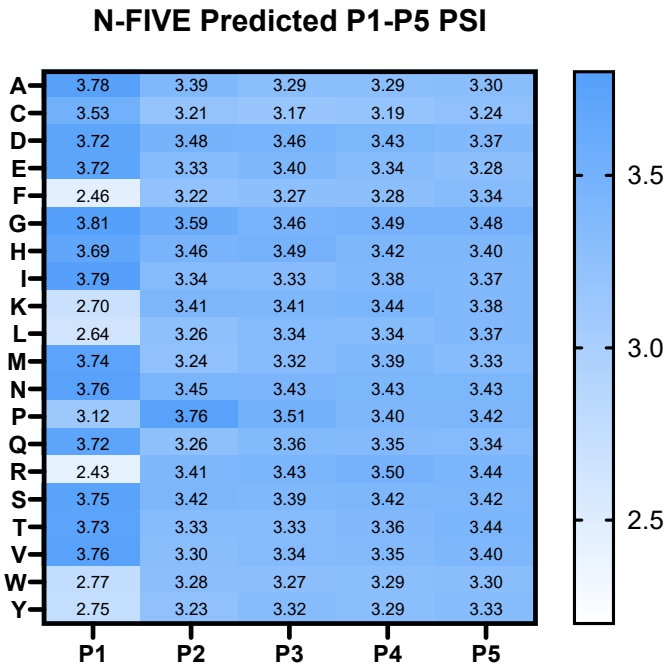
995

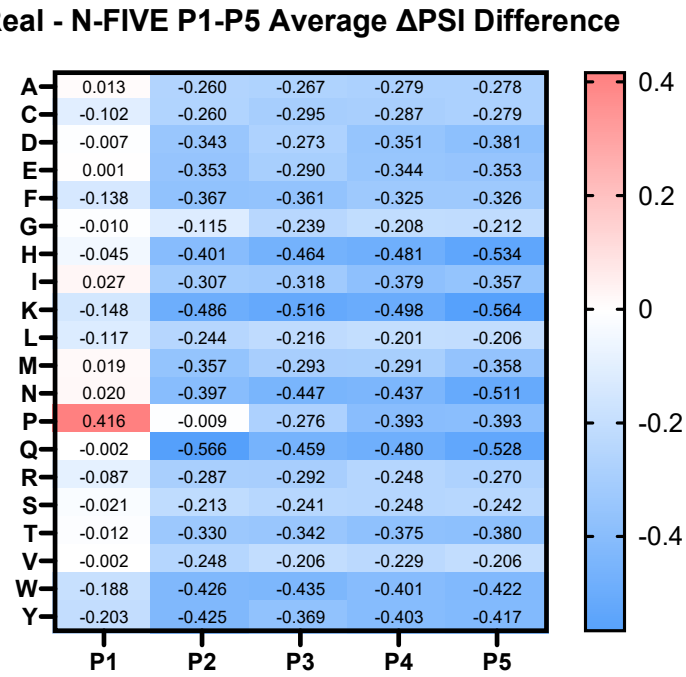
996 **Supplemental Figure 16. Full panel P1 Cys screen reveals putative N-degrons.** Flow
997 cytometry data collected from biological triplicates for the full panel of Nt-Cys screened
998 sequences.

999

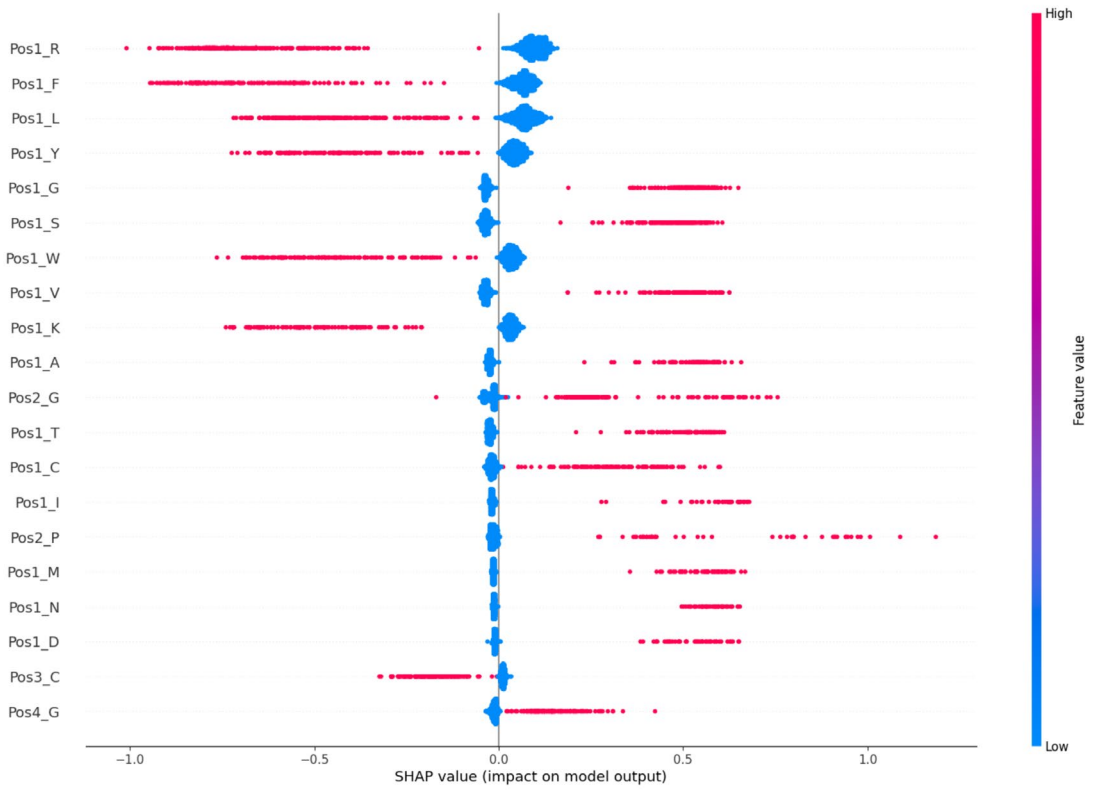


Supplemental Figure 17. Enriched residues in low stability sequences for P1 Gly and P1 Ser. Weblogo depicting the enrichment of amino acids relative to the whole population for the lowest 20,000 PSI sequences with P1 Gly or Ser.





Supplemental Figure 18. N-FIVE PSI predictions. (top) Wild-type BL21 mean PSI heatmap for all 3.2 million sequences as predicted by N-FIVE. (bottom) Change in PSI for each cell between the collected WT BL21 dataset and the N-FIVE predicted dataset.



Supplemental Figure 19. Individual matrix components for N-FIVE predictions. SHAP values showing the contribution of the 20 most impactful matrix features onto the N-FIVE model.

1017 **Supplemental Tables**

1018 **Supplemental Table 1. Strains**

1019

Strain	Description	Key mutations
E. coli BL21 DE3	Standard *E. coli* BL21 D3	N/A
E. coli BL21 DE3 *clpS-*	BL21 with inserted in-frame stop codons for *clpS*	V46*, L47*, Q48*
E. coli BL21 DE3 *aat-*	BL21 with inserted in-frame stop codons for *aat* (LFTR)	L20*, R21*, E22*

1020

1021 **Supplemental Table 2. Plasmids**

1022 pEVOL constructs are arabinose inducible, chloramphenicol resistant and pZE constructs are

1023 aTc inducible, kanamycin resistant unless listed otherwise.

Plasmid ID	Description	Description
pSS052	pEVOL mCherry-Ub-SLFVQEL-sfGFP-His6x	Dual reporter construct
pSS054	pEVOL mCherry-Ub-FLFVQEL-sfGFP-His6x	Dual reporter construct
pSS060	pEVOL mCherry-ArgRAP2.2_nt10-sfGFP-His6x	Dual reporter construct
pSS129	pEVOL mCherry-DPSnt6-16-sfGFP-His6x	Dual reporter construct
pSS142	pEVOL mCherry-Asp-Bgalnt10-sfGFP-His6x	Dual reporter construct
pSS168	pZE Ubp1 w TetR	scUlp1 protease construct
pSS303	pEVOL mCherry-Ub-(P1-P5NNK)-RAP2.2nt10-sfGFP	Dual reporter library with mutagenized P1-P5 residues
pSS0489	pEVOL mCherry-Ub-CFYEAISDFI-sfGFP	Dual reporter construct
pSS0490	pEVOL mCherry-Ub-CYWADISDFI-sfGFP	Dual reporter construct
pSS0491	pEVOL mCherry-Ub-CEEDAISDFI-sfGFP	Dual reporter construct
pSS0492	pEVOL mCherry-Ub-CYCEMISDFI-sfGFP	Dual reporter construct
pSS0493	pEVOL mCherry-Ub-CFYDEISDFI-sfGFP	Dual reporter construct
pSS0494	pEVOL mCherry-Ub-PYFDEISDFI-sfGFP	Dual reporter construct
pSS0495	pEVOL mCherry-Ub-IFFAEISDFI-sfGFP	Dual reporter construct
pSS0496	pEVOL mCherry-Ub-IWYKHISDFI-sfGFP	Dual reporter construct
pSS0497	pEVOL mCherry-Ub-FQKQKISDFI-sfGFP	Dual reporter construct
pSS0498	pEVOL mCherry-Ub-RQWYHISDFI-sfGFP	Dual reporter construct
pSS0499	pEVOL mCherry-Ub-FDEDEISDFI-sfGFP	Dual reporter construct
pSS0500	pEVOL mCherry-Ub-FDEAAISDFI-sfGFP	Dual reporter construct
pSS0501	pEVOL mCherry-Ub-FAAAAISDFI-sfGFP	Dual reporter construct

pSS0502	pEVOL mCherry-Ub-FRKAAISDFI-sfGFP	Dual reporter construct
pSS0504	pEVOL mCherry-Ub-GLVRGISDFI-sfGFP	Dual reporter construct
pSS0505	pEVOL mCherry-Ub-GLRLRISDFI-sfGFP	Dual reporter construct
pSS0506	pEVOL mCherry-Ub-LVRLGISDFI-sfGFP	Dual reporter construct
pSS0507	pEVOL mCherry-Ub-GSVLRISDFI-sfGFP	Dual reporter construct
pSS0508	pEVOL mCherry-Ub-HWMQKISDFI-sfGFP	Dual reporter construct
pSS0509	pEVOL mCherry-Ub-NIQMKISDFI-sfGFP	Dual reporter construct
pSS0510	pEVOL mCherry-Ub-SLWKYISDFI-sfGFP	Dual reporter construct
pSS0511	pEVOL mCherry-Ub-SYMMDISDFI-sfGFP	Dual reporter construct
pSS0512	pEVOL mCherry-Ub-IKKTQISDFI-sfGFP	Dual reporter construct
pSS0513	pEVOL mCherry-Ub-LAASYISDFI-sfGFP	Dual reporter construct
pSS0514	pEVOL mCherry-Ub-LDRNSISDFI-sfGFP	Dual reporter construct
pSS0515	pEVOL mCherry-Ub-LAVPRISDFI-sfGFP	Dual reporter construct
pSS0516	pEVOL mCherry-Ub-LGPRSISDFI-sfGFP	Dual reporter construct
pSS0517	pEVOL mCherry-Ub-LGDSVISDFI-sfGFP	Dual reporter construct
pSS0518	pEVOL mCherry-Ub-QQCKWISDFI-sfGFP	Dual reporter construct
pSS0519	pEVOL mCherry-Ub-HQAQWISDFI-sfGFP	Dual reporter construct
pSS0520	pEVOL mCherry-Ub-FQAQWISDFI-sfGFP	Dual reporter construct
pSS0521	pEVOL mCherry-Ub-WGAAVISDFI-sfGFP	Dual reporter construct
pSS0522	pEVOL mCherry-Ub-RGAAVISDFI-sfGFP	Dual reporter construct
pSS0523	pEVOL mCherry-Ub-LGAAVISDFI-sfGFP	Dual reporter construct
pSS0524	pEVOL mCherry-Ub-FLCFEISDFI-sfGFP	Dual reporter construct
pSS0525	pEVOL mCherry-Ub-RWCFLISDFI-sfGFP	Dual reporter construct

1025 **Supplemental Table 3. Important gene sequences**

Genetic construct	DNA Sequence
Dual fluorescent reporter (mCherry-Ub-RCKGLAALPHS-sfGFP-His6x	ATGGTGAGCAAGGGCGAGGaggataacatggccatcatcaaggagt tcatgcgcttcaaggtgcacatggagggctccgtgaacggccacgagttcgagat cgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctg aaggtgaccaagggtggccccctgcccttcgcctgggacatcctgtcccctcag ttcatgtacggctccaaggcctacgtgaagcaccccgccgacatccccgacta cttgaagctgtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgagg acggcggcgtggtgaccgtgacccaggactcctcactgcaggacggcgagttca tctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtaatg cagaagaagaccatgggctgggaggcctcctccgagcggatgtaccccgagga cggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggc cactacgacgctgaggtcaagaccacctacaaggccaagaagcccgtgcagc tgcccggcgcctacaacgtcaacatcaagttggacatcacctcccacaacgag gactacaccatcgtggaacagtacgaacgcgctgagggccgccactccaccg gcggcATGGACGAGCTGTACAAGCAGATTTTTGTGAAGACTT TAACAGGTAAGACGATTACCCTGGAGGTGGAGTCCTCGG ACACCATCGATAATGTAAAATCAAAAATCCAAGATAAGGAA GGAATCCCTCCAGACCAGCAACGTCTGATTTTCGCAGGTA AACAACTGGAGGATGGTCGCACGCTTTCGGACTACAACAT CCAGAAAGAATCTACCCTTCATTTGGTTCTGCGTCTGCGTG GAGGACGCTGCAAAGGACTTGCAGCACTGCCCCATTCCg catccaagggcgaggagctctttactggcgtagtaccaattctcgtagagctcgat ggcgatgtaaatggccataagttttccgtacgcggcgagggcgagggcgatgcaa ctaacggcaagctcactctcaagtttatttgtactactggcaagctcccagtacc atggccaactctcgtaactactctgacctatggcgtacaatgtttttcccgctatc cagatcacatgaagcaacatgatttttttaagtccgcaatgccagagggctatgta caagagcgcactattagctttaaggatgatggcacctataagactcgcgcagag gtaaagtttgagggcgatactctcgtaaatcgcattgagctcaagggcattgatttt aaggaggatggcaatattctcggccataagctggagtataatttcaattcccataa tgtatacattaccgcagataagcaaaagaatggcattaaggcgaattttaagattc gccataatgtggaggatggctccgtacaactcgcagatcattatcaacaaaata ctccaattggcgatggcccagtactcctcccagataatcattatctctccactca atccgtgctctccaaagatccaaatgagaagcgcgatcacatggtactcctgga gtttgtaactgcagcaggcattactcatggcatggatgagctctataagctcgagc accaccaccaccaccactaa

scUbp1, truncated	ATGGGGAGTGGGTCTTTCATTGCTGGGCTTGTCAACGATG GTAATACGTGTTTTATGAACTCGGTTCTTCAGTCCCTTGCTA GTAGCCGTGAACTTATGGAGTTTTTGGATAATAATGTAATCC GTACATATGAAGAAATTGAACAGAACGAGCACAATGAGGA AGGTAATGGCCAAGAGAGCGCACAAGATGAGGCAACTCA CAAAAAAAACACTCGCAAGGGAGGTAAGGTCTATGGGAA GCATAAAAAGAAATTAAACCGCAAATCTTCTAGCAAGGAA GACGAAGAAAAGTCGCAAGAACCAGACATTACGTTTTCGG TGGCGTTGCGTGATCTGCTGAGCGCATTAAATGCTAAGTAT TATCGCGACAAACCCTACTTTAAGACTAACTCTTTATTAAAA GCGATGAGCAAGTCCCCGCGCAAAAATATCTTGCTTGGGT ACGATCAAGAAGACGCTCAGGAATTTTTTCAAAACATTCTT GCGGAGTTAGAATCTAATGTCAAGTCGTTAAACACAGAAA AGCTTGATACTACACCGGTAGCCAAGTCCGAACTTCCAGA CGATGCTCTGGTTGGCCAATTAAACCTTGGTGAGGTAGGC ACCGTGTACATTCCCACAGAACAAATTGACCCCAATTCGA TTTTACATGACAAATCGATTCAAAACTTTACCCCCTTTAAAC TGATGACCCCGTTGGATGGGATCACGGCTGAGCGCATCG GCTGCCTGCAATGCGGAGAGAACGGGGGAATTCGCTACA GTGTTTTCAGCGGATTAAGTTTGAACCTGCCGAATGAAAAT ATTGGAAGCACTCTTAAACTGTCCCAGTTACTGTCCGATTG GTCGAAACCCGAGATTATCGAGGGTGTTGAATGCAACCGT TGCGCTTTAACAGCTGCGCACTCACACTTGTTTGGCCAATT AAAGGAGTTTGAGAAGAAACCTGAAGGCTCGATTCCCGAA AAACTTATTAATGCCGTAAAGGACCGCGTGCACCAGATCG AAGAGGTCTTGGCAAAGCCGGTTATCGACGATGAAGATTA TAAAAAATTGCATACTGCGAATATGGTCCGCAAGTGTTCAA AAAGTAAACAAATTCTTATCTCTCGTCCACCACCTTTGTTGT CTATTCATATCAACCGCTCTGTTTTCGACCCGCGCACCTA CATGATTCGCAAGAACAACTCCAAGGTTTTGTTCAAGTCAC GCTTGAACCTGGCACCCTGGTGCTGTGATATCAACGAAAT CAATCTTGACGCACGCCTTCCGATGTCGAAGAAGGAAAAA GCAGCTCAACAAGATTCTTCTGAAGACGAGAACATTGGCG GAGAGTACTATACTAAATTGCATGAACGTTTTGAGCAGGAG TTTGAAGATTCTGAAGAAGAGAAGGAATACGATGATGCAG AGGGTAATTATGCATCGCATTATAACCATACCAAGGACATC TCCAACTACGATCCATTGAATGGAGAAGTCGACGGTGTGA CTTCCGATGATGAGGATGAATACATTGAAGAGACAGACGC GTTGGGGAATACCATCAAAAAACGTATTATTGAACACTCCG ACGTGGAGAACGAAAACGTGAAGGATAATGAAGAACTTCA GGAGATCGATAACGTTAGCTTGGATGAGCCAAAAATTAAT GTCGAGGACCAGCTTGAAACGAGTTCTGATGAGGAAGAC GTTATTCCTGCTCCACCCATCAACTACGCTCGCAGCTTTA GTACGGTCCCAGCGACCCCTTTAACTTACTCTTTGCGCAG

	CGTCATCGTGCACTATGGGACTCACAACTACGGACATTAT ATTGCATTTCGCAAGTATCGTGGATGTTGGTGGCGCATCTC CGATGAGACGGTCTATGTGGTAGATGAGGCCGAAGTACTG TCAACACCGGGGGTATTTATGCTTTTCTACGAGTATGATTT CGACGAGGAGACCGGAAAAATGAAAGACGACTTAGAAGC TATCCAGAGCAATAATGAGGAAGATGACGAGAAAGAACAG GAACAGAAGGGTGTCCAGGAGCCAAAAGAATCCCAGGAG CAAGGCGAAGGCGAAGAACAAGAAGAAGGGCAAGAGCA AATGAAATTTGAGCGTACGGAGGATCATCGCGACATTTCA GGGAAGGATGTGAATTAA

1026