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

Authors: Sabyasachi Sen[1], Nastassja Corrado[1,2], Alexander R. Tiso[1,2], Khai Khee Kho[1,2], Aditya M. Kunjapur[1,3,*]

Affiliations & Notes:

[1] Department of Chemical & Biomolecular Engineering, University of Delaware, Newark, DE 19716, United States of America

[2] These authors contributed equally

[3] Lead contact to A.M.K. at kunjapur@udel.edu

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

Summary:

Although it is known that residues near the N-terminus can influence protein stability, there has not been a comprehensive effort to document how these penultimate residues impact protein stability in prokaryotes. Here, we utilize combinatorial mutagenesis, cell sorting, and next generation sequencing to perform the deepest sequence coverage screen of the bacterial N-degron pathway of proteolysis. We present nuance and exceptions to the N-terminus (P1) functioning as a primary stability determinant. We reveal stability contributions for P2-P5 motifs, including lowered stability for clustered bulky residues and Gln, and heightened stability for negatively charged residues, Pro, and Gly. We find that P1 Cys can be an N-degron component in a sequence-specific manner. Furthermore, we employ stability-predictive machine learning to identify motifs with unexpected fates. Our work expands the stability determinants of the N-degron pathway with unprecedented granularity, serving as a resource for N-degron identification, N-degron design, and future molecular basis elucidation.

Keywords: N-degron, protein degradation, protein stability, *Escherichia coli*, high throughput screening, machine learning, N-terminus, ClpS, LFTR

Main Text:

Introduction

Protein degradation plays a key role in supporting natural and engineered functions of bacteria. Fundamentally, protein degradation enables the recycling of proteins that have either transient, stimuli-responsive lifespans or are aberrantly expressed[1]. This phenomenon intersects with human health in myriad ways. For example, certain bacterial proteins that are directly involved in protein degradation, such as the Clp protease system in the pathogen *Mycobacterium tuberculosis*[2–4], are attractive antibiotic targets. Additionally, inducible degradation of essential proteins using BacPROTACs[5] has been performed to guide antibiotic target discovery. In applied biotechnology efforts, protein degradation tags are frequently used to attenuate hysteresis in genetic circuits within industrial workhorse strains such as *Escherichia coli*[6–8].

One impactful protein degradation system is the N-degron pathway, which modulates protein half-life across minutes to hours. The pathway relies on sequence-specific N-terminal binding interactions to initiate degradation and is conserved across both prokaryotes and eukaryotes[9]. In *E. coli,* bulky (Leu, Phe, Trp, Tyr) or positively charged (Arg, Lys) amino acids at the N-terminus have been identified as a primary signal for N-degradation[10,11]. The N-recognin ClpS, a prokaryotic homolog of the eukaryotic N-recognin Ubr1[12], recognizes and binds to N-degron-tagged proteins that have bulky amino acids at the N-terminus and deposits them at the ClpAP protease to substrate turnover[13–16]. A second key *E. coli* protein, leucyl-phenylalanyl transferase (LFTR), recognizes positively charged amino acids at the N-terminus and will append one or more Leu/Phe to that residue to generate a new ClpS substrate[17–19].

Despite these reported N-terminal patterns, recent work has shown that the identity of the N-terminal (P1) residue is not a definitive determinant of protein fate. Residues through at least the first 5 positions of a protein (P1-P5) have been shown to impact stability through the pathway, including situations where the fate of a protein appears to defy its N-terminal residue[16,18,20–25]. This is consistent with observations that downstream residues influence the dissociation constants of purified N-terminal binding proteins, including ClpS, when they are repurposed for next-generation protein sequencing applications[26–29]. These important yet limited observations strongly motivate a

55 comprehensive analysis of N-degron sequence stability determinants in bacteria (Figure 1A). The
56 benefits of such an analysis would impact various disciplines. For example, synthetic and
57 chemical biologists have increasingly designed alternative protein N-termini for non-degradative
58 purposes such as bio-orthogonal conjugation reactions[30–33], live cell cleavage of N-terminal tags or
59 signal peptides, or protein ligation or excision via inteins, sortases, or subtiligases[34–38]. Additionally,
60 in the fields of *de novo* design and protein engineering there is a need for comprehensive
61 experimental datasets that profile how synthetic sequences impact *in vivo* protein stability[39]. For
62 these applications and more, it is valuable to understand when a generated neo-N-terminus can
63 form an N-degron.

64 To date, the characterization of the bacterial N-degron pathway has been performed using
65 low-throughput protein half-life assays, *in vitro* pathway reconstitution, and affinity purification to
66 screen dozens to hundreds of putative N-degrons at once[10,17,18,22,24,40,41]. While these methods have
67 pioneered our understanding of N-degrons, they have their limitations. Pulldown assays are limited
68 to sequences enriched within the bacterial proteome. Low-throughput reporter screens require
69 extrapolation from smaller datasets. Finally, *in vitro* experimentation can sometimes present
70 artifacts or deviations from *in vivo* pathway behavior. To address these limitations, recent
71 advances within the protein degradation field, including exemplary work by Timms et al.[42], have
72 leveraged high-throughput screening and analysis of N-terminal libraries derived from the human
73 proteome to analyze roughly 10^4 sequences within human cells. Similar approaches have been
74 taken to study protein degradation[43–46], however no reported studies to date have screened large
75 DNA libraries to characterize prokaryotic branches of the N-degron pathway.

76 To investigate this knowledge gap, we evaluated the role of the five N-terminal amino acids
77 using combinatorial mutagenesis, a reporter protein assay in wild-type and N-recognin-deficient
78 bacterial hosts, fluorescence-activated cell sorting (FACS), next-generation sequencing (NGS),
79 and bioinformatic analysis. From over 2.2 million unique sequences screened and collected, we
80 have mapped the N-degron pathway with increased granularity. We share heatmaps that reveal
81 P2-P5 preferences for canonically destabilizing P1 residues. We additionally reveal a series of
82 sequence-based stability determinants. Specifically, at P2, Gln (destabilizing), Pro (stabilizing),
83 and Gly (stabilizing) show the largest single-residue stability shifts, exceeding select destabilizing

84 P1 residues. In an analysis of low stability sequences without the canonically destabilizing P1
85 FLYWRK, we found motifs rich in P1 Gln, Cys, or His with bulky and negatively charged residues in
86 P2-P5. Furthermore, additive stability effects were observed for motifs rich in bulky residues
87 (destabilizing), negatively charged residues (stabilizing), and flexible Gly/Ser residues (stabilizing).
88 We then trained a machine learning model, N-FIVE, on our dataset and harnessed it to predict
89 sequences that break canonical stability rules. In total, our manuscript serves as an in-depth
90 resource for understanding N-degron pathway specificity and designing bacterial degrons.

91 Results

92 **Design, validation, and library sorting using an *in vivo* N-degron pathway fluorescent reporter**

93 To monitor *in vivo* protein degradation, we adapted the ubiquitin reference[47] and global
94 protein stability techniques[48] by creating an mCherry-ubiquitin-degron-sfGFP-His6x genetic fusion
95 that was cloned into a plasmid (p15a ori, CmR, P_{araBAD}). Ubiquitin (Ub), considered inert within *E.*
96 *coli*, is scarlessly cleaved at its C-terminus upon the expression of Ub protease Ubp1 that is
97 housed on a second plasmid (ColE1 ori, KanR, P_{tet}). This cleavage simultaneously reveals a neo-N-
98 terminus (neo-Nt) that is attached to sfGFP while producing a stable fluorescent mCherry-
99 ubiquitin motif. The subsequent degradation of sfGFP that are fused to an N-degron should result
100 in lower sfGFP:mCherry ratios for pathway substrates relative to stable neo-Nt-sfGFP proteins.
101 This dual fluorescent reporter assay avoids various confounding effects, including variable protein
102 expression due to the 5' ORF mRNA sequence or the amino acid sequence near the translational
103 start site[49–57]. We selected *E. coli* BL21 as the expression host due to its deficiency of the Lon and
104 OmpT proteases, helping to minimize proteolytic crosstalk and to isolate the effect of the ClpSAP
105 system. To further isolate the impact of each key degradation adaptor protein, we generated two
106 separate variants of BL21 where either ClpS or LFTR were inactivated through the insertion of in-
107 frame stop codons using multiplex automatable genome engineering[58,59]. These genetic constructs
108 and assay design serve as the experimental basis for our high-throughput screening platform
109 (Figure 1B, Figure 1C).

110 To verify assay functionality, we screened neo-Nt sequences previously reported to be
111 either strong or non-interactive ClpS/N-degron pathway substrates. For positive controls, we

112 selected reported ClpS ligand FLFVQEL[24,60] and LVKTKASNLL, the latter derived from ClpSAP
113 substrate Dps[18]. For negative controls, we selected SLFVQEL, a known ClpS non-interacting
114 sequence, and DHGSGAWLLP[10], a reported stable motif derived from the first 10 amino acids of
115 β-galactosidase. Using flow cytometry, we observed unimodal distributions of sfGFP/mCherry
116 ratio across all tested degrons and we selected median sfGFP/mCherry ratio as a comparative
117 metric. Ratiometric comparisons revealed an average 6.0 and 5.9-fold dynamic range between
118 SLFVQEL/FLFVQEL and DHGSGAWLLP/LVKTKASNLL sets, respectively (Figure 1D). We
119 additionally screened FLFVQEL in the *clpS*$^{-}$ host and observed a 5.2-fold ratio increase over the
120 wild-type strain, further evidencing that observed ratiometric changes from our assay matched
121 expectations (Figure 1E). When we screened our control sequences on a BD FACS Aria II, we
122 observed order of magnitude separation between N-degrons and stable neo-Nt fusions (Figure
123 S1A). Towards deep sequence profiling, we first validated our N-degron screening platform using
124 a 60-member library, where we profiled the P1-P3 stability of eukaryotic-derived N-degrons
125 containing Nt-Arg-Cys[61]. From this experimentation, we identified a strongly stabilizing influence
126 of acidic residues at P3 that we verified *via* Western blot.

127 We next sought to expand our library size to determine the stability contributions for all
128 combinations of the first five amino acids through the N-degron pathway. We selected a sequence
129 template derived from the eukaryotic protein RAP2.2 (CGGAIISDFI)[62] due to its consistent behavior
130 and high dynamic range between the aforementioned sequence (stable) and a variant that contains
131 an N-terminally appended Arg (RCGGAIISDFI, unstable). Using five consecutive NNK codons
132 located at the neo-N-terminus of the reporter (XXXXXISDFI, theoretical library size: 3.2E6), we
133 transformed the library into unmodified, wild-type (WT) and knockout BL21 strains carrying a
134 Ubp1-expressing plasmid. We then induced expression of both the reporter and Ubp1 and
135 subsequently collected 10M gated events using fluorescence activated cell sorting (FACS). We
136 consistently observed most cells within the population in a large, high-sfGFP group with a tail that
137 extended across orders of sfGFP fluorescence magnitude (Figure S1B). We sorted mCherry-
138 expressing cells into four bins that covered the majority of the sfGFP fluorescence range (Figure
139 1F). Upon analyzing 100-150M reads of an amplicon containing the mutagenized region from the
140 WT, LFTR^{-}, and ClpS^{-} sorts, we obtained 2.29M, 2.34M, and 2.19M unique sequences per host,

141 respectively. To evaluate the stability of each collected sequence, we then calculated the Protein
142 Stability Index (PSI), a weighted average utilized in high-throughput protein degradation
143 experimentation[42,48,46,63]. Here, $PSI = \sum_{i=1}^{4} R_i * i$, where R_i is the fraction of reads present in bin *i*.
144 In line with expectations that most P1-P5 sequences are stable, PSI distributions were skewed
145 towards high stability values, with over 82% of wild-type sequences belonging in the top half of the
146 PSI range (Figure 1G).

147 **Deep sequence profiling reveals N-degron specificity through the first five amino acids**

148 To study the substrate-sequence preferences and contributions of various N-degron
149 pathway components, we generated average PSI heatmaps for all amino acid and position
150 combinations for wild-type (Figure S2A), LFTR$^-$ (Figure S2B), and ClpS$^-$ hosts (Figure S2C). Here,
151 the expectation is that amino acid motifs with lower PSI are recognized *in vivo* by LFTR and/or ClpS,
152 which initiate sfGFP degradation. To visualize how individual amino acids downstream of P1
153 impact stability while minimizing expression-based variability, we computed the average PSI for
154 each AA-position combination in the P1-P5 space and plotted the differences between WT and
155 ClpS$^-$ datasets (Figure 2A). The most distinct shifts in PSI were based on the P1 residue, in line with
156 the expected role of P1 as a critical stability determinant. The order for destabilizing P1 residues is
157 roughly tiered from lowest to highest average PSI as Phe = Arg < Leu < Trp = Lys = Tyr. An analysis
158 of the 100,000 lowest PSI sequences revealed an enrichment for five of these six residues
159 (excluding Lys) at and near P1 (Figure S2D). In a comparison of P1 residues between the wild-type
160 and LFTR$^-$ datasets, there was a linear correlation of average PSI values with an r-squared of 1.0
161 for the 17 P1 amino acids that are expected to behave similarly (Figure S2E). As expected, Arg and
162 Lys at P1 exhibited differential stability between LFTR$^-$ and WT hosts with lower average values in
163 the WT dataset. In this analysis, P1 Pro was excluded due to confounding effects when it is not
164 uniformly cleaved by Ulp1; P1 Pro is the only known instance where fused ubiquitin is not fully
165 excised by Ubp1[64]. In line with the reported role of ClpS as a sequence-specific binder and
166 degradation initiator, the ClpS$^-$ dataset shows sequence-agnostic parity in the absence of the N-
167 recognin, with a small bias for Arg, Leu, Pro, and Val at low PSI (Figure S2F). All cells in the
168 corresponding ClpS$^-$ P1-P5 heatmap are within 0.17 PSI units of each other, demonstrably smaller

169 than the >1 PSI differences observed between canonically stable and unstable P1 residues in WT
170 BL21 and LFTR$^-$ BL21 (Figure S2A and S2B).

171 Next, we visualized the P2-P5 preferences for all P1 residues by calculating the average PSI
172 difference between WT and ClpS$^-$ datasets for all amino acid-position combinations (Figure 2B).
173 We observed PSI variability at P2-P5 for sequences with a canonically destabilizing P1.
174 Specifically, while most 2-position sequence combinations led to PSI decreases, there were
175 instances where the drops in PSI varied in magnitude. For example, destabilizing PSI shifts with P1
176 Arg were smaller when negatively charged residues were present particularly in P3 & P4, as well
177 as Pro in P2. In parallel, P1 Phe showed smaller shifts for small residues (Ala, Gly, Ser) in P2-P5
178 and Pro in P2. Frequently, the presence of a bulky residue in P2 lead to greater PSI drop magnitudes.
179 P1 Leu, Trp, Lys and Tyr all exhibit PSI differences between different amino acid combinations,
180 indicating unique P2-P5 preferences for these canonically destable P1 residues. These results
181 further evidence that P1 identity cannot be viewed in as an exclusive stability determinant,
182 particularly when identifying highly destabilizing N-degrons.

183 We next profiled the specificity of LFTR by analyzing the PSI difference between WT and
184 LFTR- datasets. In a comparison of average PSIs values between the WT and LFTR- dataset, a clear
185 PSI difference was observed for P1 Arg and Lys values, suggesting increased stability in the
186 absence of LFTR (Figure S2G). When analyzing the shift in PSI at P2-P5 for P1 Arg and Lys, a clear
187 and uniform increase in PSI is observed (Figure 2C). Arg and Lys residues at P2-P5 did not appear
188 to influence stability, suggesting that internal residues were not contributing to destabilization in
189 bulk (Figure S2H & S2I). One exception was for FRXXX sequences, where P2 Arg led to the largest
190 decrease in mean PSI within the subset of sequences with internal Arg residues. In total, Arg and
191 Lys do not notably impact stability in the absence of LFTR, as expected.

192 **P2 residues including Pro, Gly, Gln, and Ala are critical stability determinants**

193 Having shown that residues adjacent to P1 can have a pivotal role in stability determination,
194 we next focused on key stability determinants at P2. As P2 is positioned on the periphery of the
195 ClpS binding pocket in crystallized bound peptide ligands, docking studies have not clearly
196 elucidated a role for this position (Figure S3A). Within our datasets, several amino acids showed

197 stabilizing effects at this position. We observed this most distinctly from Gly and Pro, with the
198 effect's magnitude having a maximal impact at P2 and waning with distance from P1. Commonly
199 enriched in stable, PSI >3 motifs (Figure S3B), these two residues had the highest PSI change in the
200 WT - ClpS$^-$ heatmap. P2 Pro paired with P1 FLWYRK led to a significant 1.05-unit increase in PSI (p
201 <1E-99, ES = 0.71) (Figure 3A). Notably, P2 Pro often overrode the stability contribution of
202 canonically destabilizing P1 residues, with larger groups being observed in high stability ranges
203 (Figure 3B). Only select P1 Arg or P1 Lys motifs were notably present at lowered PSI values relative
204 to the rest of the P1 dataset. A similar, but smaller in magnitude 0.63-unit PSI increase was
205 observed for P2 Gly (p<1E-99, ES = 0.47) (Figure 3C). Upon testing clonal isolates with P2 Gly, we
206 found motifs that stabilize P1 Leu and Trp. Interestingly, P1 Arg remained destabilizing, potentially
207 due to Gly having an extended distance from the neo-P1 upon LFTR activity (Figure 3D).

208 Gln at position 2 (and to a lesser extent at P3-5) is amongst the most destabilizing P2 amino
209 acids in both WT and LFTR$^-$ BL21, providing a -0.51 PSI unit decrease at P2 in the WT dataset.
210 Fascinatingly, Gln at P2 is just as strong a determinant of protein degradation as L at P1 (-0.52
211 units). The destabilizing effect was amplified when Gln was paired with P1 LFWYRK (p <1E-99, ES
212 = 0.26), showing a 3.2x decrease in the difference in mean PSI compared to Gln paired with an
213 alternative P1 residue (Figure 3E). The stability impact of Gln was visible across WT and LFTR$^-$
214 datasets and absent in the ClpS$^-$ dataset, suggesting a ClpS-mediated preference (Figure S2A and
215 S2C). We sought to validate this observation using clonal examples. Upon pairing Gln with a
216 destabilizing P1 residue, we identified sequences that resulted in sfGFP:mCherry ratios that were
217 3.4 and 4.0 times lower than our degron control (RQWYHISDFI : FLFVQEL and FQAQWISDFI :
218 FLFVQEL, respectively). Furthermore, neither FQAQWISDFI nor RQWYHISDFI showed no visible
219 degron-sfGFP band when visualized via Western blot (Figure 3F, Supplementary File 2), suggesting
220 strong turnover of the construct.

221 A final residue of note was P2 alanine. Specifically for bulky destabilizing P1 residues, P2
222 Ala can have either a stabilizing or destabilizing effect depending on the P1 residue. For P1 Phe or
223 Leu, P2 Ala leads to a significant increase in PSI compared to other P2 residues (p<1E-99, ES =
224 0.43). Furthermore, for P1 Trp or Tyr, P2 Ala provides a significant destabilizing effect (p<1E-99, ES
225 = 0.27). This observation reveals a differential specificity amongst bulky destabilizing P1 residues

226 (Figure 3G). This effect was only observed for Ala at P2 and not P3-P5 for those specific P1 residues
227 (Figure 3H).

228 We additionally examined whether the native *E. coli* methionine aminopeptidase (MetAP)
229 could generate an N-degron by excising Met at P1 to generate a degradation-promoting motif
230 starting at P2. We compared the PSI for motifs carrying Nt-Met (MXYYY) to those same motifs
231 when they appear in the first four positions (XYYYZ) within the wild-type dataset. We observed that
232 residues that are canonically destabilizing at the X-position had increased mean PSI values in
233 MXYYY motifs when compared to XYYYZ motifs (Figure S3C & Figure S3D). In line with the
234 published specificity of MetAP[65,66], it did not appear that Met-L/F/W/Y/R/K motifs produced low
235 PSI N-degrons in bulk.

236 **Destabilizing motifs include unexpected P1 residues Cys, His, and Gln with specific P2-P5**
237 **patterns**

238 Having found that P2 residues can drastically influence stability, we next investigated the
239 stability contributions of all 400 P1/P2 pairs in greater detail. First, we observed that LFWYRK at
240 P1/P2 were highly represented in low stability sequences. In an analysis of the 100 most enriched
241 P1-P2 motifs at low PSI, 90/100 contained at least one FLWYRK residue, and 34/36 possible
242 combinations containing two FLWYRK residues were present, showing an overrepresentation of
243 these six residues within destabilizing motifs relative to other combinations at P1/P2 (Figure 4A).
244 Enriched residues not in FLWYRK within the low stability P1/P2 space of select residues including
245 P1 Cys, P1 and P2 Gln, and P2 Trp (Figure 4B). Furthermore, when visualizing all P1/P2
246 combinations that were enriched at low PSI but did not carry a canonically destabilizing P1 residue,
247 Gln was overwhelmingly present (Figure S4A). From this finding, we next investigated the sequence
248 profiles for low stability sequences with canonically stable P1 residues.

249 We then analyzed moderate to low stability (PSI < 3) sequences without LFWYRK at P1. In
250 this analysis, Cys and His were enriched at P1. Additionally, bulky residues (LFWY), acidic residues
251 (DE), and assorted other residues (MVC) were enriched at P2-P5 (Figure 4C). Next, a series of
252 subsets were analyzed. First, as previous studies have reported that bulky residues downstream
253 of the N-terminus can be destabilizing components of an N-degron[20], we investigated sequence

254 patterns with multiple bulky residues paired with a canonically stable P1 residue. We observed
255 that sequences with bulky residues exclusively in P2/P3 had a significant drop in average PSI
256 relative to the global population (ES = 0.25) (Figure 4D). In a deeper analysis of sequences with
257 bulky P2 & P3 residues, we observed that the P1 residue with the lowest average PSI was Cys,
258 followed by Gln and His (Figure S4B). Second, we observed that low PSI motifs with
259 noncanonically destabilizing P1 often maintained a neutral to negative charge (Figure 4E).
260 Interestingly, these low PSI sequences follow a charge profile that is opposite to that of the
261 canonical six destabilizing P1 amino acids, where positively charged motifs appear more
262 frequently at lower PSI. Third, the largest decreases in PSI were observed for canonically stable P1
263 sequences with both bulky and negatively charged residues (Figure 4F). These sequence
264 dependent trends were not observed in the ClpS- dataset (Figure S4C). In total, these findings
265 reveal a low stability series of sequences with P1 Cys, His, and Gln and bulky and negatively
266 charged P2-P5 motifs that appear to be dependent on ClpS.

267 We next turned our attention to P1 Cys to understand what could make this synthetically
268 useful P1 residue a component of an N-degron. Interestingly, P1 Cys had a lower mean PSI (3.42)
269 than other stable P1 residues (PSI 3.80), suggesting that there are unstable sequences (Figure
270 S2A). Upon visualizing the sequence composition of the 20,000 lowest PSI P1 Cys motifs, we
271 observed an enrichment of bulky and negatively charged residues, whereas positively charged and
272 flexible residues as well as P2 Pro were depleted in these motifs (Figure S4D). In a parallel analysis
273 using ClpS- dataset, there was no enrichment of negatively charged sequences, suggesting that
274 this was unique to ClpS expressing strains (Figure S4E). These enrichment trends aligned with
275 residue-by-position PSI shifts for sequences with P1 Cys (Figure 2B, P1 Cys). We generated clonal
276 isolates of five P1 Cys sequences with candidate destabilizing motifs. Two of the five candidates
277 presented sfGFP:mCherry ratios on par with the FLFVQEL degron control, while the remaining
278 three offered intermediate values (Figure S4F). When we performed a Western blot, at least one
279 candidate sequence (CFYEAISDFI) was fully degraded, producing a non-visible band similar to the
280 strongest identified degrons (Figure 4G).

281 We investigated other commonly functionalized P1 residues in Gly and Ser. Both residues
282 were stable at P1, though an analysis of P2-P5 residues for the 20,000 lowest PSI sequences for

283 P1 Gly and P1 Ser revealed an enrichment in Gln, Glu, Lys, and Trp at P2-P5 (Figure S4G & Figure
284 S4H).

285 **Similar residue sidechains can cumulatively amplify or dampen bacterial N-degrons**

286 We next studied the combinatorial effect of multiple amino acids, as stability can be a
287 function of aggregate properties such as sidechain hydrophobicity or charge, or could otherwise be
288 dictated by local sequence context similarly to other reported degrons[63]. To select which
289 properties to target, we grouped amino acids based on their R-group characteristics and evaluated
290 their enrichment in low stability sequences within the WT dataset. Unstable sequences showed
291 an enrichment for aromatic, hydrophobic, and positively charged residues, and a deficiency in
292 flexible, negatively charged, and small amino acids at various P1-P5 positions, but most distinctly
293 at P1 and P2 (Figure 5A). Recognizing the differential enrichment in charged residues as well as
294 literature precedent that penultimate charged residues can impact ClpSAP recognition and
295 turnover[14,23,60], we investigated whether there could be connection between the net charge of the
296 N-terminal region and protein stability. We observed that lower PSI sequences are more common
297 in neutral to positively charged motifs, whereas negatively charged motifs, even with P1 FLYWRK,
298 had a significant increase in average PSI (Figure 5B). To further investigate, we tested clonal
299 isolates of charged sequences carrying P1 Phe. We found FDEDE (-4) to be highly stabilizing and
300 FDEAA (-2) to be lightly stabilized relative to a control degron, and that positively charged FRKAA
301 (+2) was stabilized (Figure 5C). This finding is supported by the presence of negatively charged
302 residues D36 and E41 flanking the N-degron binding pocket in ClpS crystal structure, which would
303 repulse negatively charged P2-P5 substrates to make binding unfavorable (Figure S5A).
304 Additionally, in the WT dataset both bulky (FLWY) and positively charged (RK) P1 residues showed
305 increases in average PSI when paired with negatively charged P2-P5 motifs. P1 RK sequences
306 demonstrated a higher increase in mean PSI between neutral and negatively charged P2-P5
307 datasets (Figure S5B & Figure S5C). This charge-based phenomenon was absent in the ClpS^{-}
308 dataset (Figure S5D).

309 We next analyzed the impact of clustered FLWY residues to investigate whether multiple
310 bulky residues can additively impact stability. For sequences carrying multiple N-terminal FLWY

311 residues, the average PSI was 0.56 units lower ($p < 1E-99$, ES = 0.45) for two such residues at P1
312 and P2 compared to a singular P1 FLWY. Diminishing PSI shifts were observed for motifs carrying
313 >2 N-terminal FLWY residues (Figure 5D). On average, sequences carrying a single positively
314 charged residue at P1 exhibited a lower average PSI than sequences with a bulky P1, and lower
315 magnitude PSI shifts when multiple charged residues were chained (Figure S5E). However, these
316 motifs may have multiple bulky residues appended onto P1 Arg or P1 Lys via LFTR polyaddition[17,18].
317 We screened several clonal isolates with multiple bulky and positively charged residues at P1-P5
318 and found them all to be destabilizing (Figure S5F).

319 We additionally analyzed the additive impact of multiple smaller flexible amino acids,
320 focusing on Gly and Ser, which frequently compose flexible N-terminal linkers. In line with the
321 depletion of small amino acids in destabilizing motifs, we observed PSI increase with increasing
322 Gly/Ser count downstream of a P1 FLWYRK residue (PSI increases for Gly/Ser counts 0-1: 0.09, 0-
323 2: 0.32, 0-3: 0.64, 0-4: 0.99, all $p<1E-33$) (Figure 5E). As such, the presence of these flexible motifs
324 may aid in the stability of N-terminal motifs with canonically unstable residues.

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

326 Our next objective was to develop a stability-predictive machine learning model to interpret
327 the full sequence space. To generate a base matrix, we one-hot encoded each motif by amino acid
328 and position. During training, we observed improved statistical performance upon applying a
329 minimum read count threshold of 20 for any given amino acid motif. We applied this filter to trim
330 our WT dataset to approximately 798,000 sequences prior to final training. We supplied these
331 encodings and the corresponding PSI from 80% of the WT dataset to generate the training input.
332 We then trained an extreme gradient boosted model using Python's XGBoost library. After
333 hyperparameter tuning, our model obtained an RMSE of 0.34 and an r-squared of 0.82 when
334 validated using the remaining 20% of data. Satisfied with the performance of our model, we dubbed
335 our model N-FIVE and set out to test its predictive capabilities (Figure 6A-6B). We first simulated
336 the PSI of 100,000 random N-terminal sequences and observed the characteristic decrease in
337 mean PSI for sequences carrying destabilizing N-termini (Figure 6C). Upon simulating the full 3.2M
338 sequence space using N-FIVE, we compared its performance to our actual dataset, finding

339 comparable P1 prediction and higher P2-P5 average PSI values (Figures S6A & S6B). We generated
340 SHAP plots to understand the underlying architecture, observing trends on par with known N-
341 degron pathway rules (Figure S6C).

342 To explore the capabilities of N-FIVE, we studied N-terminal motifs with unexpected
343 stability profiles. We sought to identify five stable sequences with a canonically destabilizing P1
344 Leu. Excitingly, all five of the sequences (sequence LXXXXISDFI) suggested by N-FIVE displayed
345 stable profiles (Figure 6D). We additionally attempted the challenging task of identifying sequences
346 that are unexpectedly unstable with P1 residues that are not in FLWYRK. From our five-sequence
347 screen, one motif, SYMMDISDFI, generated a moderately destabilizing profile (Figure 6F). SHAP
348 values were calculated for candidate sequences LAASYISDFI (Figure 6E) and SYMMDISDFI (Figure
349 6G), revealing the cumulative effect of P2-P5 residues in "overriding" the P1 residue.

350 Discussion

351 This work presents the most comprehensive mapping of the *Escherichia coli* N-degron
352 pathway to date. Having collected and analyzed over 2.2M screened sequences that cover the
353 critical P1-P5 space within wild-type and adaptor deficient BL21, we both reinforce and provide
354 insight into the *in vivo* substrate preferences of pathway with unprecedented granularity. Our
355 findings corroborate the consensus understanding within *E. coli* that six amino acids present at P1
356 (Leu, Phe, Tyr, Trp, Arg, Lys) lead to lowered protein stability due to their recognition and
357 degradation through the ClpSAP proteolytic cascade. We have additionally shown that stability
358 determinants beyond P1 appear in our aggregation of screened sequences. Specifically, our
359 analyses show that neutral to positively charged N-terminal motifs with P1 FLWYRK and local
360 bulky and Gln residues are highly probable substrates for ClpS recognition. Frequently, P2-P5
361 bulky amino acids lower protein stability. However, it is unclear how they may be causing instability
362 as there is conflicting evidence about whether bulky residue downstream of P1 can be bound
363 directly in the hydrophobic pocket of ClpS[20]. It is possible that this instability could be caused by
364 strengthened interactions in the substrate channel or surface of the protein. Furthermore, we
365 show cases where the expected level of protein stability based on P1 can be fundamentally
366 altered, such as when multiple small, stabilizing residues flank a canonically destabilizing P1

367 residue. The impact of penultimate residues is evident for destabilizing P2 Gln and P2-P5 bulky
368 residues, the universally stabilizing P2 Pro and Gly, sequence-specific stabilization from internal
369 Asp, Glu, and Ala, and the scattered P2-P5 preferences for P1 FLWY, amongst other examples.

370 These findings bear relevance to various applied settings, including bacterial N-degradation, N-
371 terminal binders, and the design of synthetic N-termini for applications in small molecule/protein
372 ligation in live cells. For example, bioconjugation efforts that rely on a P1 Cys should avoid bulky
373 and negatively charged residues near the N-terminus. Gly-Ser motifs are likely to provide a
374 stabilizing effect that would offset any degradation-promoting elements. For N-terminal
375 identification, the low tolerance for negatively charged motifs may be a challenge. The mutation
376 and screening of variants that have lowered negative electrostatic potential near the binding pocket
377 may address this limitation. When tuning protein expression using degrons, the characteristic
378 structure of an N-degron mentioned in the previous paragraph is worth considering when designing
379 constructs with a target stability in mind.

380 Amongst the most destabilizing motifs were those that could be generated by LFTR. Multiple
381 previous studies have identified that polyaddition of Leu and Phe, rather than addition of a singular
382 Leu or Phe, is a possibility[17,18]. The highly destabilizing nature of multiple Leu/Phe-Arg residues in
383 tandem coincide with previous observations that the turnover of LR-substrates is greater than just
384 Leu-substrates.[14] Furthermore, others have suggested that the ClpS binding site has been
385 optimized to accommodate LFTR-modified substrates[67]. On balance, the generation of Leu/Phe-
386 rich motifs in front of Nt-Arg and Lys may contribute to the generation of highly destabilizing
387 degrons, supporting the rapid turnover of LFTR-substrates despite the requirement of an additional
388 modification.

389 Interestingly, the P2-P5 profile for P1 Cys degrons aligns with that of the eukaryotic Ac/N-
390 degron pathway, where the N-acetyltransferase NatA acetylates N-terminal cysteine with a
391 preference for polar and acidic residues downstream of P1 Cys[68]. Within bacteria, N-acetylation
392 has been observed by GNAT-family enzymes RimI-RimJ, including for Nt-Cys[69], but it has not has
393 not yet been linked to protein degradation. As such, the exact mechanism for the instability of these

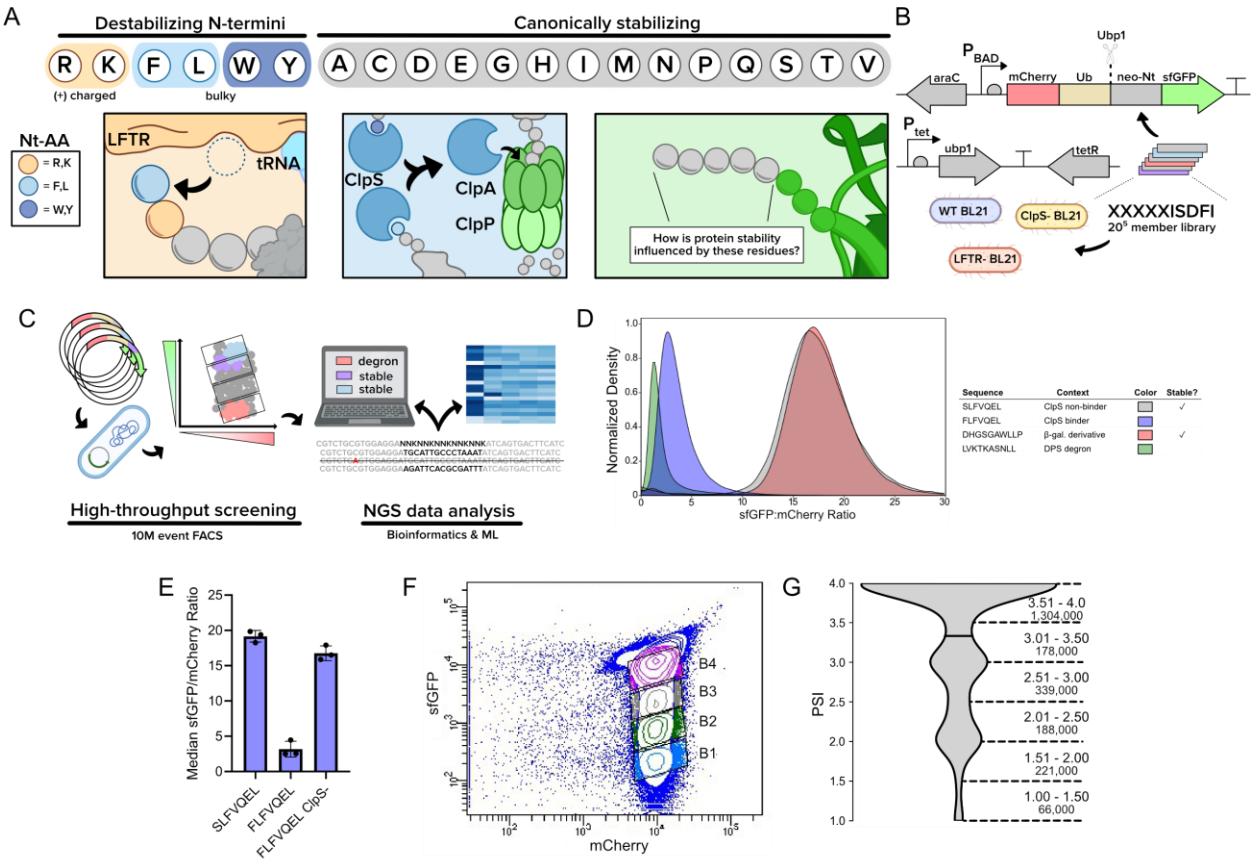
394 sequences is unknown. Our data does suggest a ClpS-mediated phenomenon, yet there may be
395 an uncharacterized protein that modifies these motifs.

396 For the synthetic design of N-terminal motifs, the stability-predictive N-FIVE model can be
397 consulted to evaluate or suggest motifs even when there are sequence constraints. The predictive
398 power of N-FIVE has been evidenced by its ability to solve a complex sequence-stability problem;
399 in this instance successfully suggesting unexpectedly stable sequences. The identification of
400 unexpectedly unstable sequences remains a challenge, in part due to the small number of
401 candidate sequences to form the dataset. Further development of more complex ML models may
402 improve stability prediction capabilities. As such, to aid in such analyses, we have made our full
403 datasets and trimmed, storage friendly sequence-stability databases available for public use.
404 Finally, our application of SHAP values supports a thought process for evaluating the strength of
405 an N-degron; one where a given sequence is measured as accumulation of stabilizing or
406 destabilizing factors that build to a net predicted effect.

407

408 **Figures:**

409



410

411 **Figure 1. Overview and HTS workflow for N-terminal sequence analysis**

412 **A.** An overview of the *Escherichia coli* Leu/N-degron pathway. The top row shows the expected
413 fate for proteins with the listed amino acids at the N-terminus, where the orange and blue shaded
414 residues are canonically destabilizing. The left-most box illustrates the leucyl/phenylalanyl
415 transferase LFTR, appending Phe or Leu onto an N-terminal Arg or Lys. The middle box illustrates
416 the ClpS adaptor protein that binds N-terminal F, L, W, and Y and recruits the protein to the ClpAP
417 protease for degradation. The right-most box shows the dual reporter after the action of a ubiquitin
418 (Ub) cleavase, revealing a neo-N-terminus (neo-Nt) on an sfGFP reporter. **B.** Genetic constructs
419 for the protein stability assay. The first plasmid harbors a synthetic gene fusion that encodes the
420 dual fluorescent reporter under control of an arabinose-inducible promoter. The second plasmid
421 harbors a gene encoding ubiquitin cleavase from *Saccharomyces cerevisiae* under control of a

422 tetracycline-inducible promoter. Constructs were tested in wild-type BL21 as well as BL21
423 derivatives deficient in *clpS* and *aat* (LFTR). **C.** The HTS protein stability workflow used in this study.
424 **D.** sfGFP:mCherry ratio density histograms for over 100,000 events collected for cells expressing
425 Ubp1 as well as literature control N-terminal sequences embedded in a dual fluorescent reporter
426 construct. Stable sequences were SLFVQEL and DHGSGAWLLP and unstable sequences were
427 FLFVQEL and LVKTKASNLL. **E.** Known ClpS-binding degron FLFVQEL returns stabilizing
428 sfGFP:mCherry ratio values in the absence of ClpS. Data shown is median fluorescence intensity
429 ± standard deviation for at least 100,000 events captured across clonal biological triplicates. **F.**
430 FACS binning of the XXXXXISDFI library within wild-type BL21 **G.** Violin plot showing the PSI
431 distribution of 2.30M collected sequences from the wild-type BL21 dataset. The number of
432 sequences is rounded to the nearest thousand. Figures 1A-1C generated using Biorender content.

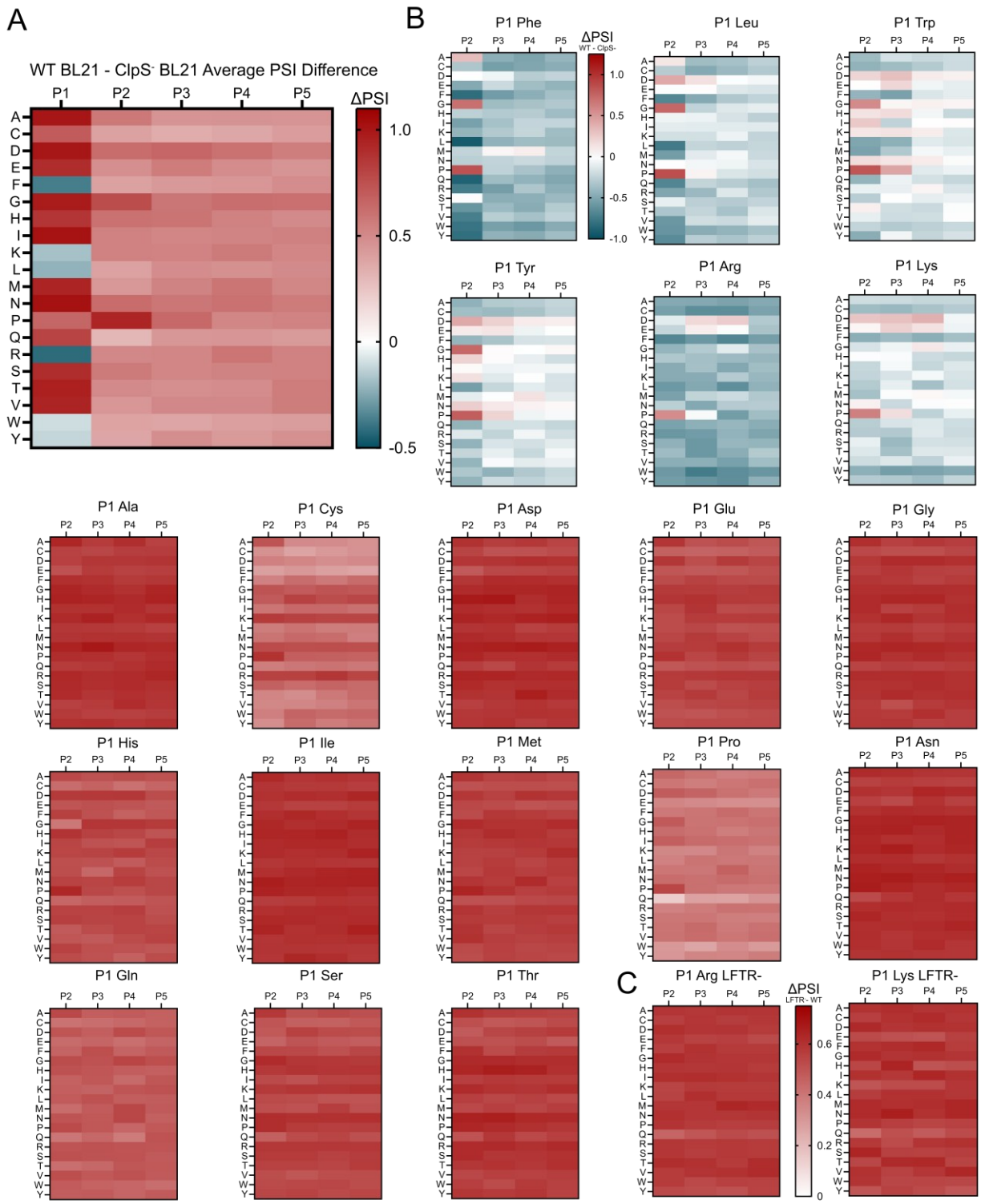


Figure 2. Summary of stability contributions for penultimate amino acids

436 **A.** Heatmap visualizing the bulk P1-P5 preferences of the *Escherichia coli* N-degron
437 pathway visualized as the average change in PSI between wild-type BL21 and ClpS^{-} BL21
438 datasets for P1-P5 amino acid-position combinations. **B.** Heatmaps visualizing P2-P5
439 preferences each standard amino acid at P1 using the average change in PSI between wild-
440 type BL21 and ClpS- BL21 for P2-P5 amino acid-position combinations. **C.** Increased P2-
441 P5 stability profile for P1 Arg and Lys in *Escherichia coli* deficient in LFTR visualized using
442 the average change in PSI between LFTR^{-} BL21 and WT BL21 datasets for P2-P5 amino
443 acid-position combinations.

444

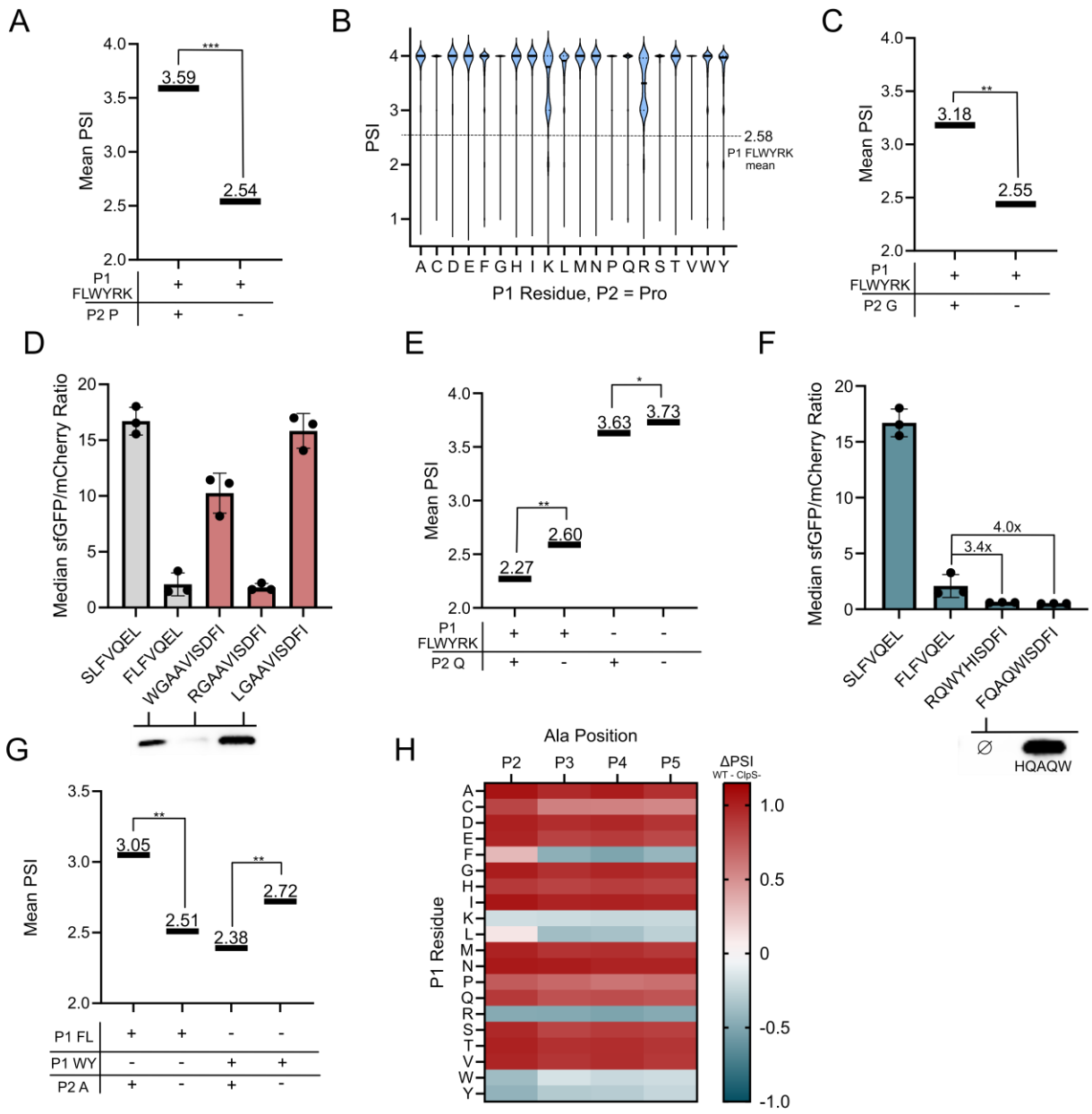


Figure 3. P2 residues including Pro, Gly, Gln, and Ala are critical stability determinants

A. Stability increase from P2 Pro being paired with a canonically unstable N-terminus. Mean PSIs for wild-type sequence subsets where the number of sequences per condition are (++): 33,674, (+-): 750, 888; (++/+-) ES: 0.71. **B.** Violin plot showing PSI distributions for various P1 residues with P2 Pro in the WT dataset. The mean PSI for all P1 FLWYRK residues is shown as a dashed line. **C.** Stability increase from P2 Gly being paired with a canonically unstable N-terminus. Data

453 shown is average PSI for wild-type sequence subsets where the number of sequences per
454 condition are (++): 41,975, (+-): 742,587; (++/+-) ES: 0.47. **D.** Clonal P2 Gly isolates show
455 stabilizing effects for two non-LFTR modified N-terminal motifs. Flow cytometry data for various
456 P2 glycine motifs (top). Western blots of hardcoded sequences (bottom). **E.** An amplified
457 stability decrease when P2 Gln is paired with a canonically unstable N-terminus. Data shown is
458 average PSI for wild-type sequence subsets where the number of sequences per condition are
459 (++): 29,982, (+-): 754, 580, (-+): 52,021, (--): 1,459,940; (++/+-) ES: 0.26. **F.** Clonal P2 Gln
460 isolates show decreased stability measurements relative to a literature control. Flow cytometry
461 data for various P2 glutamine motifs (top). Western blots of hardcoded sequences (bottom). No
462 band was visible for the FQAQWISDFI sequence, indicated by ∅. **G**. P2 Ala stabilizes P1 F or L
463 and further destabilizes P1 W or Y. Data shown is average PSI for wild-type sequence subsets
464 where the number of sequences per condition are (+-+): 14,127, (+--): 272,399, (-++): 12793, (-+-
465): 232,067; (+-+/+--) ES: 0.44, (-++/-+-) ES: 0.27. **H**. Bulky P1 stability shifts from Ala are primarily
466 when Ala is at P2. Heatmap depicting the change in PSI between WT and ClpS$^-$ datasets for all P1
467 residues with Ala at a given P2-P5 position. p-values were evaluated using a Mann-Whitney U
468 test. Effect size magnitude (ES) ranges from 0 to 1, as evaluated using Cliff's Delta. * = $p < 0.05$,
469 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.
470 Comparisons are abbreviated as (P1 FLYWRK presence, P2 presence) or (P1 FL presence, P1 WY
471 presence, P2 presence).

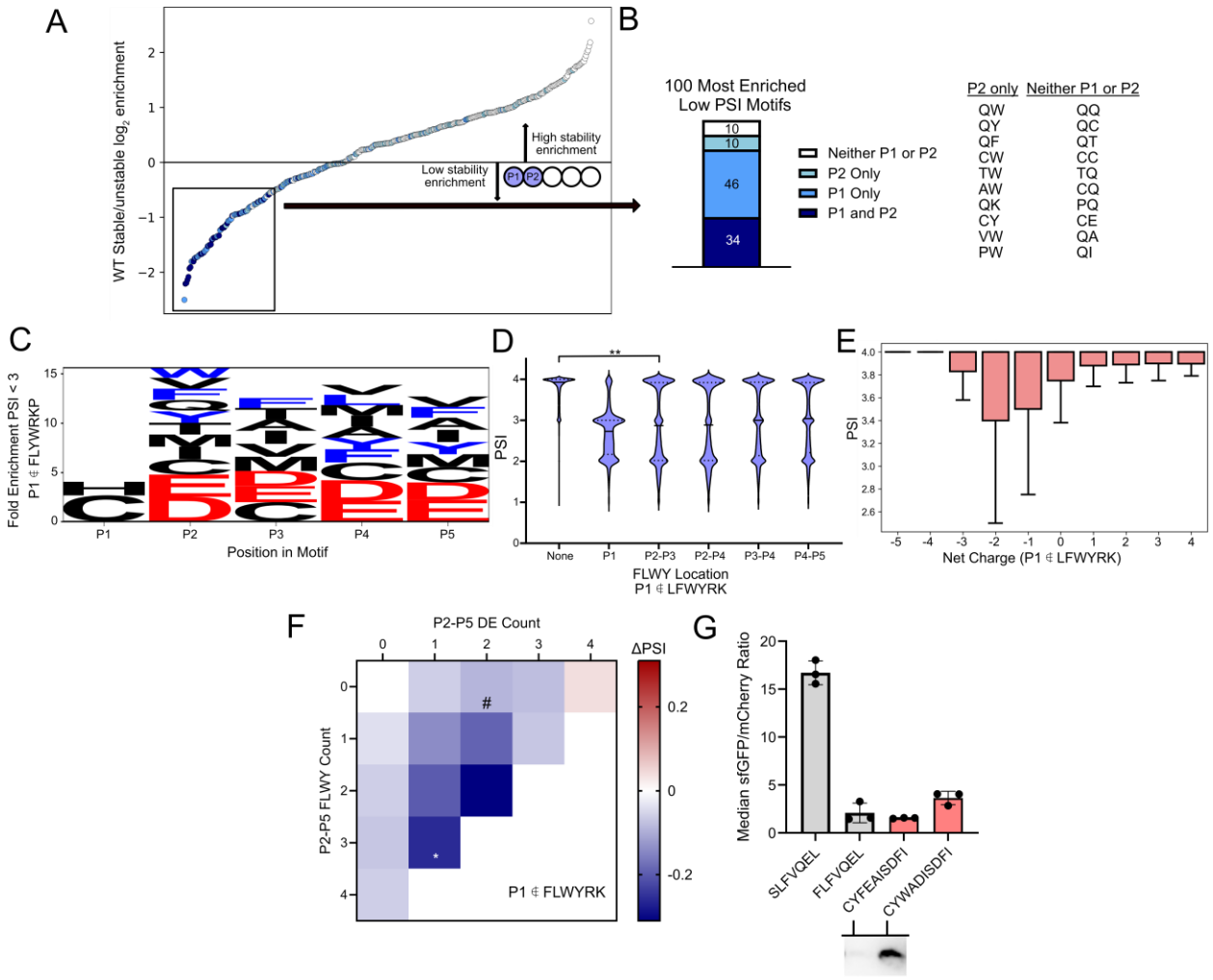


Figure 4. Destabilizing motifs can contain P1 Cys and Gln with specific P2-P5 motifs

A. Log_2 enrichment ratio plot of the 400 possible stable (PSI >3)/unstable (PSI<2) P1-P2 dipeptide motifs. Sequences containing LFWYRK at various positions are listed in the following pairs: (P1 and P2, dark blue), (P1 only, light blue), (P2 only, light green), (neither P1 or P2, white). **B.** LFWYRK composition of the 100 lowest PSI P1/P2 motifs. Motifs with P1 not in FLWYRK are listed below the plot. **C.** 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. **D.** Multiple bulky residues lower mean PSI. Violin

483 plot of wild-type BL21 data showing PSI distributions for sequences with bulky residues at various
484 positions. A comparison between motifs with no bulky residues (n = 1,511,962) and motifs with
485 bulky residues only in P2 and P3 (n=56,341) shows a significant difference between mean PSI
486 values ($p < 0.05$, ES = 0.25) **E.** PSI boxplot for sequences with varying net charge and canonically
487 stable P1 residues shows the opposite charge trend to low stability P1 FLWYRK sequences. P1 Pro
488 is excluded. **F.** For WT sequences with canonically stable P1, the largest PSI drops are observed
489 for motifs with both bulky and negatively charged P2-P5 residues. Change in PSI is relative to the 0
490 P2-P5 FLWY, 0 P2-P5 DE dataset (mean PSI = 3.78). All p-values relative to (0,0) are $p < 0.05$ except
491 cells marked with # which are $p > 0.05$. * = ES > 0.25. **G.** Identification of an Nt-Cys degron. (Top)
492 flow cytometry data, (bottom) Western blots of the degron-sfGFP product for candidate degrons
493 with N-terminal cysteine. p-values evaluated using a Mann-Whitney U test, effect size magnitude
494 (ES) ranges from 0 to 1, as evaluated using Cliff's delta. * = $p < 0.05$, ES 0-0.25; ** = $p < 0.05$, ES
495 0.25-0.50, *** = $p < 0.05$, ES 0.50-0.75, **** = $p < 0.05$, ES 0.75-1.0

496

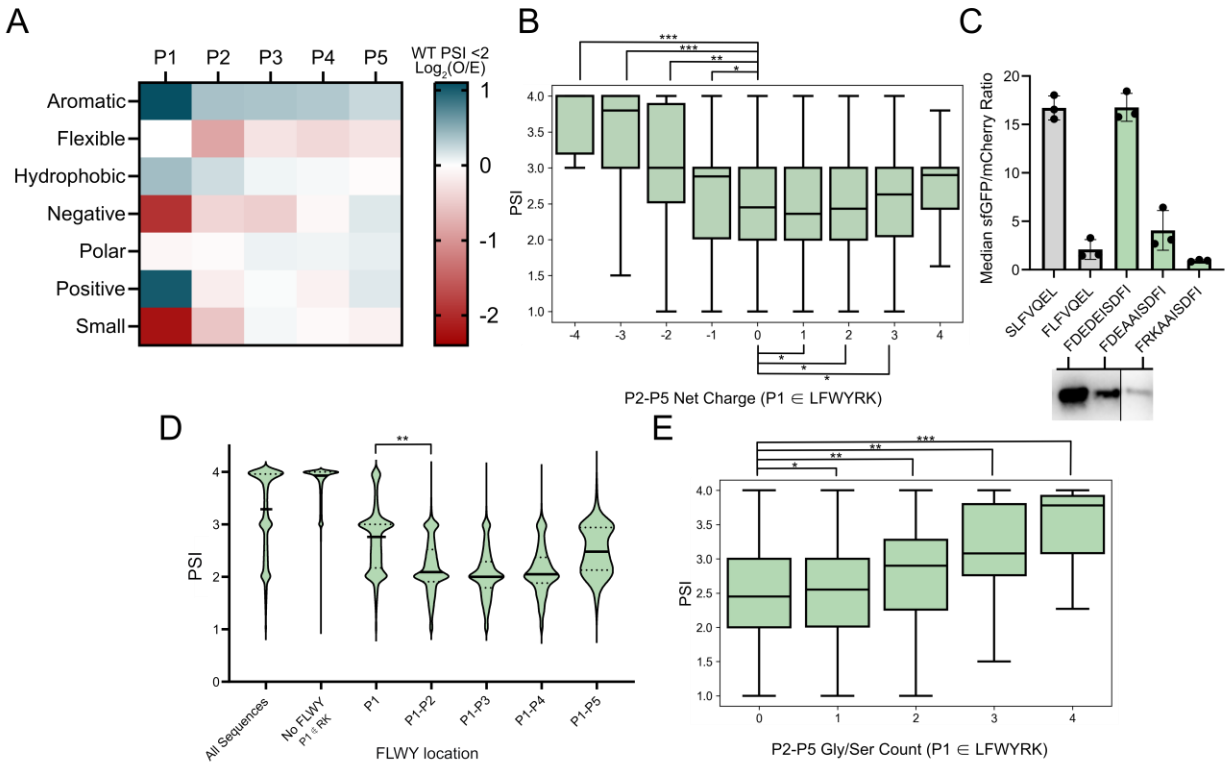


Figure 5. P1-P5 sidechain properties can cumulatively impact stability

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). **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. Non-consecutive lanes from the same blot are separated by a solid line. **D.** Boxplot depicting that clustered FLWY at and near the N-terminus leads to decreased stability. 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

514 p-values in comparison to 0 P2-P5 Gly/Ser show $p < 0.05$. p-values evaluated using a Mann-
515 Whitney U test, effect size magnitude (ES) ranges from 0 to 1, as evaluated using Cliff's delta. * = p
516 <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-
517 1.0

518

519

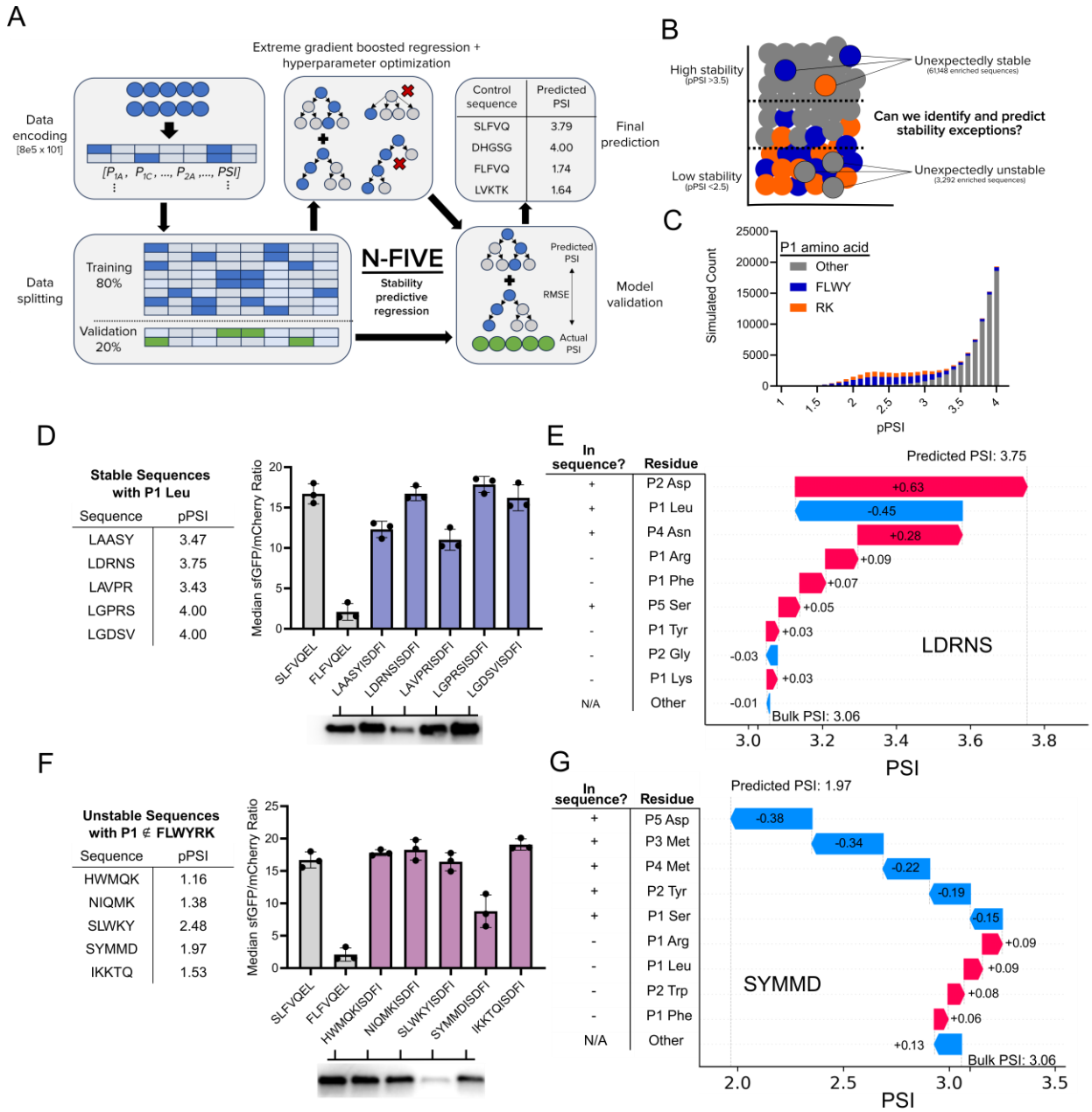


Figure 6. N-FIVE, a machine learning model for stability prediction

A. Architecture of the N-FIVE machine learning model. 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. **C.** PSI prediction of 100,000 randomly selected 5-amino acid sequences using N-FIVE. Sequences with a canonically

528 destabilizing P1 are highlighted (bulky residues in blue, positively charged residues in orange). **D.**
529 (Left) Predicted PSI values for five unexpectedly stable (pPSI >3.5) candidate sequences selected
530 with a P1 Leu constraint. (Top right) Candidate sequence flow cytometry data, (right left) Western
531 blots of the degron-sfGFP product. **E.** (right) SHAP values for individual matrix contributions to the
532 predicted PSI value of 3.75 for the candidate sequence LDRNS. **F.** Predicted PSI values for five
533 unexpectedly unstable (pPSI < 2.5) candidate sequences selected with a P1 ∉ FLWYRK constraint.
534 (Top right) Candidate sequence flow cytometry data, (bottom right) Western blots of the degron-
535 sfGFP product. **G.** (right) SHAP values for individual matrix contributions to the predicted PSI value
536 of 1.97 for the candidate sequence SYMMD.

537

538 **Acknowledgements**

539 The authors thank the following people: Dr. Arit Ghosh of the University of Delaware Flow
540 Cytometry Core for training and advice on cell sorting and analysis. Bruce Kingham and Mark Shaw
541 of the University of Delaware Sequencing & Genotyping Center for consultations and performing
542 NGS sequencing. Ishika Govil and Zhao Wei Wang for discussions and code for the analysis of NGS
543 data. Prof. Karl Schmitz for discussions regarding ClpSAP interactions and manuscript review.
544 Prof. Kevin Solomon for discussions regarding bioinformatic analysis. We further thank members
545 of the Kunjapur Lab for their helpful discussions and review of the manuscript.

546 **Author Contributions**

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

552 **Declaration of Interests Statement**

553 A.M.K. has a financial interest in a commercial entity, Nitro Biosciences Inc. The remaining authors
554 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 by the Office of the Director and the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH), grant P20GM104316B from the National Institutes of Health (NIH)-funded COBRE Center for Discovery of Chemical Probes and Therapeutic Leads, as well as the University of Delaware Research Foundation. Additional support for Sabyasachi Sen was provided by the Mort & Donna Collins Fellowship. Access to FACS and NGS was supported by the Institutional Development Award (IDeA) from the National Institute of Health's National Institute of General Medical Sciences under grant number P20GM103446 through DE-INBRE.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Aditya Kunjapur (kunjapur@udel.edu).

Materials availability

This study did not develop new unique reagents or chemicals.

Plasmids and strains are available from the Lead Contact upon request.

Data and code availability

NGS analysis code, trimmed sequence databases, and the N-FIVE model have been made available at GitHub (https://github.com/KunjapurLab/N-terminal-cluster-stability). The raw FASTQ.gz datasets will be made publicly available by the time of publication. Uncropped blots and raw data can be found in the supplemental files.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

580 *E. coli* DH10β and *E. coli* E. cloni SUPREME were used for plasmid construction and *E. coli* BL21

581 (DE3), *E. coli* BL21*clpS*-, and *E. coli* BL21 *aat*- sequence screening.

582

583 Methods

584 **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**

585 Microbial culture

586 *Escherichia coli* strains, including BL21, ClpS- BL21, and LFTR- BL21, were routinely grown
587 in culture tubes containing LB media at 37 °C while shaking at 250 RPM, unless mentioned
588 otherwise.

589 **METHOD DETAILS**

590 *Escherichia coli* strains and plasmids used are listed in Table S1 and Table S2, respectively.
591 *clpS* and *aat* inactivations were performed as described in the Genetic knockout of N-recognins
592 section. Genes were purchased as G-Blocks or gene fragments from Integrated DNA Technologies
593 (IDT) or Twist Bioscience and were optimized for E. coli K12 using the IDT Codon Optimization Tool.
594 A version of scUbp1 without the N-terminal region was used for improved expression in *E. coli*[70].
595 Essential genetic sequences and primers can be found in Table S3 and Table S4, respectively.
596 Cloning amplicons were generated using KOD XTREME Hot Start polymerase and corresponding
597 buffering reagents. Amplicons were verified and purified by running through a 1% agarose gel for
598 200 V for 20 minutes followed by gel excision and extraction. Samples were assembled in
599 homemade Gibson assembly aliquots run at 50 °C for 30 minutes. Assemblies were transformed
600 into *E. coli* cloning strains DH5α or DH10β using either electroporation or chemical heat shock.
601 Following a one-hour outgrowth after transformation, cells were plated on permissive antibiotic
602 plates and were incubated at 37 °C overnight. The following afternoon, multiple colonies were
603 picked and incubated overnight in separate culture tubes containing LB media (10 g/L tryptone, 5
604 g/L sodium chloride, 5 g/L yeast extract) and the permissive antibiotic. The next day, cells were
605 stocked 1:1 with 30% glycerol at negative 80 °C for future usage. The remaining cells were
606 miniprepped for sequencing. Plasmid sequences were verified through a mixture of Sanger
607 sequencing through Eurofins, Genewiz, and Azenta and full plasmid sequencing through
608 Plasmidsaurus. Plasmids were transformed into BL21 and BL21 derivatives using either
609 electroporation or chemical heat shock. Two plasmid systems were transformed in series.

610 **Materials and chemicals**

611 The following compounds were purchased from MilliporeSigma (Burlington, MA, USA):
612 phosphate-buffered saline (PBS), kanamycin sulfate, glycerol, 25 nm membranes (VSWP02500)
613 and KOD XTREME Hot Start polymerase (Millipore 71975-3). D-glucose was purchased from TCI
614 America (Portland, OR, USA). Agarose, Laemmli SDS sample reducing buffer, and ethanol were
615 purchased from Alfa Aesar (Ward Hill, MA, USA). Anhydrotetracycline (aTc) was purchased from
616 Cayman Chemical (Ann Arbor, MI, USA Methanol, sodium chloride, LB Broth powder (Lennox), LB
617 Agar powder (Lennox), Amersham ECL Prime chemiluminescent detection reagent, and Thermo
618 Scientific™ Spectra™ Multicolor Broad Range Protein Ladder were purchased from Fisher Chemical
619 (Hampton, NH, USA). Taq DNA ligase was purchased from GoldBio (St. Louis, MO, USA). Phusion
620 DNA polymerase and T5 exonuclease were purchased from New England BioLabs (NEB) (Ipswich,
621 MA, USA). SybrSafe DNA gel stain was purchased from Invitrogen (Waltham, MA, USA). E. cloni
622 10G Supreme electrocompetent cells were purchased from Biosearch Technologies. Miniprep kits
623 were purchased from Zymo. 4-20% precast protein gels were purchased from Bio-Rad. Flow
624 cyometry filter caps (Chemglass Life Sciences CLS4380009), arabinose, DpnI enzyme
625 (FERFD1704) and Immobilon-E Western blot membranes were obtained from Fisher Scientific.
626 HRP-conjugated anti-6*His antibody (Proteintech HRP-66005) was obtained from Proteintech
627 (Rosemont, IL, USA).

628 **Genetic knockout of N-recognins**

629 Multiplex automatable genome engineering (MAGE) was used to inactivate the endogenous
630 *aat* (LFTR) and *clpS* genes[59]. MAGE oligonucleotides were designed using MODEST[71] to insert three
631 in-frame stop codons into the gene of interest. Freshly made electrocompetent cells were
632 resuspended in 5 μM oligonucleotide and subsequently electroporated to enable cell permeation
633 of the oligonucleotide. Cells were outgrown in LB media for an hour prior to plating on permissive
634 plates. To verify the desired genetic knockouts, allele-specific colony PCR was performed using
635 KAPA 2G Fast HotStart polymerase (Roche KK5801). Additional Sanger sequencing was
636 performed to verify asPCR hits.

637 **Fluorescence analysis of degrons**

638 Overnights made from biological triplicate colonies of BL21 and recognin-deficient BL21
639 derivatives carrying both the dual reporter and Ubp1 expressing plasmids were inoculated at a
640 1:100 ratio of grown culture : fresh LB media containing 0.2% arabinose, 100 ng/mL aTc, 20 µg/mL
641 chloramphenicol, and 15 µg/mL kanamycin. Cells were grown at 37 °C, 250 RPM for 18 hours prior
642 to analysis. Cells were then filtered using 35 µm filter caps into cytometry tubes containing fresh
643 PBS at a roughly 1:250 ratio of cell culture: PBS. Cells were analyzed on a NovoCyte Flow
644 Cytometer (Agilent Technologies). Prior to analysis, scatter gates were used to isolate singlet
645 bacterial cells of the appropriate morphology. sfGFP fluorescence was analyzed using a 488 nm
646 laser and a 530/30 nm bandpass filter and mCherry fluorescence was analyzed using a 488 nm
647 laser and a 660/20 nm bandpass filter. For all analyses, at least 100,000 events that passed
648 through the singlet gate were collected.

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

650 1 mL of overnighted culture containing expressed dual reporter and Ubp1 protease were
651 lysed using glass beads for 15 minutes. The sample was transferred to a fresh microcentrifuge tube
652 and centrifuged at 12,000 RPM for 10 minutes. Subsequently, the supernatant was collected.
653 Sample concentration was evaluated using a Bradford assay on a SpectraMax i3x with a BSA
654 calibration curve. Sample concentration was normalized to 0.1 mg/mL in water with 1x SDS PAGE
655 loading dye and denatured at 95 °C for 10 minutes. To separate proteins by size, 10 uL of sample
656 is loaded into a 4–20% Mini-PROTEAN TGX gel and run at 180 V for 35 minutes. Subsequently, the
657 proteins were transferred onto an Immobilon-E membrane through an overnight wet transfer (25V
658 for 15 hours on ice). The following day, the membrane was blocked for at least one hour with 5%
659 milk in TBST buffer. Subsequently, an HRP-conjugated anti-His6x antibody was added to the
660 blocking solution at a 1:10,000 dilution, and was allowed to incubate at room temperature for 1
661 hour. Following three washes in TBST to remove the solution, ECL Prime chemiluminescent
662 reagent was added to the membrane. Following a 10 second incubation, chemiluminescent
663 images were captured at varying intensities using an Azure c280.

664 **DNA library preparation**

665 DNA library plasmids were designed to be assembled from two separate PCR amplicons.
666 Primers containing 5 consecutive NNK codons at P1-P5 of the degron template were ordered from
667 IDT, using a 25A/25C/25G/25T (N) or 50G/50T (K) ratio of hand-mixed nucleotides to maximize
668 diversity in the mutagenized region. PCR amplicons were generated using KOD XTREME
669 polymerase using roughly 200 ng of plasmid template per reaction with an annealing temperature
670 of 59 °C and extension time of 2 minutes for inserts and 3 minutes for the backbone. To maximize
671 sequence diversity and avoid overamplification, 5 separate PCR amplicons were generated at a
672 low cycle count (20 cycles) and high volume (50 µL per reaction). Amplicons were then digested
673 using a DpnI enzymatic digestion to remove the template. The amplicons were then isolated on a
674 1% agarose gel and subsequently purified using gel extraction. Amplicons were pooled and
675 assembled using an isothermal Gibson assembly method with a total of 300 ng of total DNA at a
676 1:1 molar ratio of insert: backbone. Assemblies were then dialyzed on a 25 nm membrane placed
677 over deionized water for 60 minutes. Five separate assemblies were pooled and then
678 electroporated into E. cloni 10G Supreme electrocompetent cells. Immediately following
679 electroporation, cells were mixed with 500 µL of SOC media and were placed in a culture tube in a
680 250 RPM shaking incubator set to 37 °C. Following a 30-minute outgrowth, cells were plated at
681 serial dilutions to verify transformation efficiency. For 20^5 member libraries, a transformation
682 efficiency of 10^8 colonies/transformation was the minimum threshold established for further
683 experimentation. Following a second 30-minute outgrowth, 5 mL of SOC media was added to the
684 tube and cultures were allowed to grow overnight.
685 The following morning, library stocks were made by mixing 500 µL of undiluted culture at a
686 1:1 ratio with 30% glycerol, which were then stored at -80 °C indefinitely. The remaining 4-4.5 mL
687 of culture was miniprepped using a Zymo ZR Plasmid Miniprep-Classic kit. BL21 and recognin-
688 deficient BL21 derivatives carrying a Ubp1-expressing plasmid were made electrocompetent using
689 standard molecular biology procedures, concentrating 10 mL cultures to 200 µL competent cell
690 aliquots. Miniprepped libraries were electroporated into freshly made competent cells. Outgrowth,
691 efficiency measurements, overnight growth, and freezer stocking were performed similarly to
692 original library transformations. A transformation efficiency of 10^8 was obtained for all conditions.

693 Library diversity was confirmed prior to sorting using 2.5 Mbp next-generation sequencing through
694 Azenta's AmpliconEZ platform.

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

696 Overnights of BL21 and recognin-deficient BL21 derivatives carrying both library plasmids
697 and Ubp1-expressing plasmids were inoculated at a 1:100 ratio in fresh LB media containing 0.2%
698 arabinose, 100 ng/mL aTc, 20 µg/mL chloramphenicol, and 15 µg/mL kanamycin. Cells were
699 grown at 37 °C for 18 hours prior to analysis. Cells were then filtered using 35 µm filter caps into
700 cytometry tubes containing fresh PBS at a roughly 1:250 ratio of cell culture: PBS. Cells were
701 sorted on a BD FACS Aria Fusion using a 70 µm or 85 µm nozzle and a neutral density 1.0 filter.
702 Prior to analysis, scatter gates were used to isolate singlet bacterial cells of the appropriate
703 morphology. sfGFP fluorescence was analyzed using a 488 nm laser and a 530/30 nm bandpass
704 filter and mCherry fluorescence was analyzed using a 561 nm laser and a 610/20 nm bandpass
705 filter. Bins were set based on controls, where SLFVQEL sorted into B4 and template RCGGAIISDFI
706 sorted into B2 (Supplemental Figure 1). A small gap was placed between bins to maximize sort
707 efficiency, and all bins maintained a sort efficiency ≥ 95%. Sorting conditions and sample
708 concentrations were optimized to obtain event rates on the order of 1,000 – 5,000 events per
709 second. 10 million gated cells were sorted into bins covering the sfGFP range, excluding the bottom
710 of the distribution that was observed to have a higher concentration of debris and GFP-inactivating
711 mutants. The following cell counts were collected in the format [B4; B3; B2; B1]. WT: [5,998,196;
712 1,940,219; 1,130,665; 955,183]. LFTR-: [6,953,470; 1,161,141; 998,851; 913,809]. ClpS-:
713 [7,048,173; 1,651,034; 483,574; 853,250]. Cells were recovered overnight in a mixture of 1:1
714 recovered cells: LB media containing 1% glucose. The subsequent day, cells were stocked and
715 miniprepped using methods described in the DNA library preparation section.

716 **Next generation sequencing preparation and analysis**

717 250 base pair (AVITI) and 405 base pair (Amplicon EZ) DNA amplicons for next-generation
718 sequencing were designed to contain a 5' and 3' sequencing adapter, a 3 base pair barcode for
719 demultiplexing, and a centralized mutagenized region to ensure coverage by both paired-end
720 reads. The full sequence can be found in our corresponding methods paper[61]. Amplicons were

721 generated using KAPA 2G polymerase, minimizing cycle counts (20 was default) to decrease over-
722 enrichment of select sequences. The amplicons were then isolated on a 1% agarose gel and
723 subsequently purified using gel extraction. Samples were then submitted to Azenta's Amplicon EZ
724 service (~100,000-300,000 2x250 paired end reads) or to the UD Sequencing & Genotyping Center
725 (>100 million 2x150 paired end reads on an Element AVITI). The WT, LFTR-, and ClpS- datasets
726 had 146,704,252, 180,209,552, and 188,333,530 reads respectively for B1-B4 and the presorted
727 library. All sequencing datasets had a Q30 >95%.

728 NGS data was first converted from a FASTQ.gz format to FASTA for ease of analysis. When
729 necessary, reads were demultiplexed using the three base pair barcodes that corresponded to
730 unique sorted bins. Reads were compared to the consensus sequence for the submitted amplicon.
731 Any reads that varied from the consensus sequence (excluding the barcode and mutagenized
732 region) or those that contained stop codons (TAG, TAA, TGA) in the mutagenized region were then
733 excluded from subsequent analysis. The mutagenized region was identified in each read based on
734 a 15 base pair consensus sequence located directly after the mutagenized region, and the
735 mutagenized nucleotide sequence was translated into an amino acid sequence. From data
736 collected for each strain, read counts were measured for each five amino acid motif across the
737 four bins. From this data, the protein stability index (PSI) was calculated using the formula $PSI =$
738 $\sum_{i=1}^{4} R_i * i$, where R_i is the fraction of reads present in bin i[48]. Each unique sequence, its read
739 counts per bin, and PSI were stored in a database for subsequent analysis.

740 Heatmaps, boxplots, and violin plots were visualized using the Seaborn and matplotlib
741 Python packages[72,73] as well as GraphPad Prism. Weblogos were generated using Logomaker[74]. All
742 code was developed using Python. Code, sequence databases, and N-FIVE are accessible on
743 GitHub: https://github.com/KunjapurLab/N-terminal-cluster-stability. Full NGS data will be made
744 available prior to publication.

745 **Development of N-FIVE machine learning model**

746 During early rounds of model training, we observed that applying a 20 read minimum filter
747 to included sequences minimized RMSE and improved R^2 for the model. As such, data from the
748 wild-type BL21 sequence-PSI database were one-hot encoded into an 798,837 x 101 matrix. The

749 first 100 columns represented a binary encoding of each of the 20 amino acids across 5 positions,
750 while the final column corresponded to the PSI value. The dataset was randomly split, with 80% of
751 the data being used for model training and 20% being used for model validation and evaluation. An
752 extreme gradient boosted nonlinear regression model was then trained using the Python XGBoost[75]
753 package for 20 rounds. This type of model outperformed random forest regressors and ridge
754 regression models statistically. Hyperparameter optimization was performed using Optuna[76] over
755 30 trials, aiming to minimize RMSE. The following parameters were set: *n_estimators*: 986,
756 *max_depth*: 11, *learning_rate*: 0.085, *subsample*: 0.904, *colsample_bytree*: 0.810, providing an
757 RMSE of 0.34 for predicted PSI. The final model was then saved as a .pkl file for future analysis. To
758 evaluate the contribution of each matrix component, SHAP (Shapley Additive exPlanations) values
759 were calculated[77] for the whole model as well as for individual sequences.

760 **QUANTIFICATION AND STATISTICAL ANALYSIS**

761 To quantify differences in PSI distributions between defined groups, we applied two non-
762 parametric statistical methods: the Mann-Whitney U test and Cliff's Delta for effect size
763 calculations. The Mann-Whitney U test was selected due to the skewed nature of PSI distribution
764 as well as its robustness to unequal sample sizes. We used a two-sided Mann–Whitney U test to
765 assess whether the PSI distributions between two groups differed significantly. Cliff's Delta was
766 computed to quantify the magnitude and direction of the distributional shift between compared
767 groups. Unlike common effect size measurements such as Cohen's d, Cliff's Delta makes no
768 assumptions about the underlying distribution of PSI values and remains valid for skewed,
769 bounded, or ordinal data. Here, Cliff's Delta represents the probability that a randomly selected
770 value from one group is greater than a randomly selected value from the other, minus the reverse
771 probability to yield an effect size bounded between −1 and +1.

772 For general enrichment analysis, the counts of any given residue or motif were first measured.
773 To calculate enrichment differences between two populations, the average occurrence of an
774 amino acid or motif in the appropriate population was calculated and subsequently subtracted to
775 obtain a difference. To obtain a ratio, groupings were normalized to the same sample size and

776 subsequently divided for ratiometric comparison. Log_2 analysis was performed at times to simplify
777 data visualization.

778 For clonal isolate testing, fluorescence measurements were taken for at least 100,000 gated
779 cells using flow cytometry using cultures grown from three separate biological triplicates. Median
780 ratiometric fluorescence ratios +/- standard deviation were reported. Median values were
781 selected over mean values due to large mean shifts observed in select cases due to outliers, to
782 which median values were less sensitive.

783 **Supplemental Content**

784 Document S1. Figures S1–S6, Tables S1 & S2

785 Document S2. Uncropped Western blots presented in Figures 3-6

786

787 **References:**

788 1. Mahmoud, S.A., and Chien, P. (2018). Regulated Proteolysis in Bacteria. Annu. Rev. Biochem.
789 *87*, 677–696. https://doi.org/10.1146/annurev-biochem-062917-012848.

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

793 3. d'Andrea, F.B., Poulton, N.C., Froom, R., Tam, K., Campbell, E.A., and Rock, J.M. (2022). The
794 essential M. tuberculosis Clp protease is functionally asymmetric in vivo. Sci. Adv. *8*,
795 eabn7943. https://doi.org/10.1126/sciadv.abn7943.

796 4. Won, H.I., Zinga, S., Kandror, O., Akopian, T., Wolf, I.D., Schweber, J.T.P., Schmid, E.W., Chao,
797 M.C., Waldor, M., Rubin, E.J., et al. (2024). Targeted protein degradation in mycobacteria
798 uncovers antibacterial effects and potentiates antibiotic efficacy. Nat. Commun. *15*, 4065.
799 https://doi.org/10.1038/s41467-024-48506-8.

800 5. Morreale, F.E., Kleine, S., Leodolter, J., Junker, S., Hoi, D.M., Ovchinnikov, S., Okun, A., Kley,
801 J., Kurzbauer, R., Junk, L., et al. (2022). BacPROTACs mediate targeted protein degradation in
802 bacteria. Cell *185*, 2338-2353.e18. https://doi.org/10.1016/j.cell.2022.05.009.

803 6. Chassin, H., Müller, M., Tigges, M., Scheller, L., Lang, M., and Fussenegger, M. (2019). A
804 modular degron library for synthetic circuits in mammalian cells. Nat. Commun. *10*, 2013.
805 https://doi.org/10.1038/s41467-019-09974-5.

806 7. Kunjapur, A.M., Stork, D.A., Kuru, E., Vargas-Rodriguez, O., Landon, M., Söll, D., and Church,
807 G.M. (2018). Engineering posttranslational proofreading to discriminate nonstandard amino
808 acids. Proc. Natl. Acad. Sci. *115*, 619–624. https://doi.org/10.1073/pnas.1715137115.

809 8. Prindle, A., Selimkhanov, J., Li, H., Razinkov, I., Tsimring, L.S., and Hasty, J. (2014). Rapid and
810 tunable post-translational coupling of genetic circuits. Nature *508*, 387–391.
811 https://doi.org/10.1038/nature13238.

812 9. Varshavsky, A. (2024). N-degron pathways. Proc. Natl. Acad. Sci. *121*, e2408697121.
813 https://doi.org/10.1073/pnas.2408697121.

814 10. Tobias, J.W., Shrader, T.E., Rocap, G., and Varshavsky, A. (1991). The N-End Rule in Bacteria.
815 Science *254*, 1374–1377. https://doi.org/10.1126/science.1962196.

816 11. Dougan, D.A., Truscott, K.N., and Zeth, K. (2010). The bacterial N-end rule pathway: expect
817 the unexpected. Mol. Microbiol. *76*, 545–558. https://doi.org/10.1111/j.1365-
818 2958.2010.07120.x.

819 12. Kim, J.G., Shin, H.-C., Seo, T., Nawale, L., Han, G., Kim, B.Y., Kim, S.J., and Cha-Molstad, H.
820 (2021). Signaling Pathways Regulated by UBR Box-Containing E3 Ligases. Int. J. Mol. Sci. *22*,
821 8323. https://doi.org/10.3390/ijms22158323.

822 13. Dougan, D.A., Reid, B.G., Horwich, A.L., and Bukau, B. (2002). ClpS, a Substrate Modulator of
823 the ClpAP Machine. Mol. Cell *9*, 673–683. https://doi.org/10.1016/S1097-2765(02)00485-9.

824 14. Erbse, A., Schmidt, R., Bornemann, T., Schneider-Mergener, J., Mogk, A., Zahn, R., Dougan,
825 D.A., and Bukau, B. (2006). ClpS is an essential component of the N-end rule pathway in
826 Escherichia coli. Nature *439*, 753–756. https://doi.org/10.1038/nature04412.

827 15. Román-Hernández, G., Hou, J.Y., Grant, R.A., Sauer, R.T., and Baker, T.A. (2011). The ClpS
828 adaptor mediates staged delivery of N-end-rule substrates to the AAA+ ClpAP protease. Mol.
829 Cell *43*, 217–228. https://doi.org/10.1016/j.molcel.2011.06.009.

830 16. Yeom, J., Gao, X., and Groisman, E.A. (2018). Reduction in adaptor amounts establishes
831 degradation hierarchy among protease substrates. Proc. Natl. Acad. Sci. *115*, E4483–E4492.
832 https://doi.org/10.1073/pnas.1722246115.

833 17. Humbard, M.A., Surkov, S., De Donatis, G.M., Jenkins, L.M., and Maurizi, M.R. (2013). The N-
834 degradome of Escherichia coli: limited proteolysis in vivo generates a large pool of proteins
835 bearing N-degrons. J. Biol. Chem. *288*, 28913–28924.
836 https://doi.org/10.1074/jbc.M113.492108.

837 18. Ninnis, R.L., Spall, S.K., Talbo, G.H., Truscott, K.N., and Dougan, D.A. (2009). Modification of
838 PATase by L/F-transferase generates a ClpS-dependent N-end rule substrate in Escherichia
839 coli. EMBO J. *28*, 1732–1744. https://doi.org/10.1038/emboj.2009.134.

19. Watanabe, K., Toh, Y., Suto, K., Shimizu, Y., Oka, N., Wada, T., and Tomita, K. (2007). Protein-based peptide-bond formation by aminoacyl-tRNA protein transferase. Nature *449*, 867–871. https://doi.org/10.1038/nature06167.

20. Gao, X., Yeom, J., and Groisman, E.A. (2019). The expanded specificity and physiological role of a widespread N-degron recognin. Proc. Natl. Acad. Sci. *116*, 18629–18637. https://doi.org/10.1073/pnas.1821060116.

21. Wang, K.H., Roman-Hernandez, G., Grant, R.A., Sauer, R.T., and Baker, T.A. (2008). The Molecular Basis of N-end Rule Recognition. Mol. Cell *32*, 406–414. https://doi.org/10.1016/j.molcel.2008.08.032.

22. Ottofuelling, R.D., Ninnis, R.L., Truscott, K.N., and Dougan, D.A. (2021). Novel modification by L/F-tRNA-protein transferase (LFTR) generates a Leu/N-degron ligand in Escherichia coli. Preprint at bioRxiv, https://doi.org/10.1101/2021.07.26.453911 https://doi.org/10.1101/2021.07.26.453911.

23. Stein, B.J., Grant, R.A., Sauer, R.T., and Baker, T.A. (2016). Structural basis of an N-degron adaptor with more stringent specificity. Struct. Lond. Engl. 1993 *24*, 232–242. https://doi.org/10.1016/j.str.2015.12.008.

24. Wang, K.H., Oakes, E.S.C., Sauer, R.T., and Baker, T.A. (2008). Tuning the Strength of a Bacterial N-end Rule Degradation Signal. J. Biol. Chem. *283*, 24600–24607. https://doi.org/10.1074/jbc.M802213200.

25. Yeom, J., Wayne, K.J., and Groisman, E.A. (2017). Sequestration from Protease Adaptor Confers Differential Stability to Protease Substrate. Mol. Cell *66*, 234-246.e5. https://doi.org/10.1016/j.molcel.2017.03.009.

26. Reed, B.D., Meyer, M.J., Abramzon, V., Ad, O., Ad, O., Adcock, P., Ahmad, F.R., Alppay, G., Ball, J.A., Beach, J., et al. (2022). Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. Science *378*, 186–192. https://doi.org/10.1126/science.abo7651.

27. Ikonomova, S.P., Yan, B., Sun, Z., Lyon, R.B., Zatopek, K.M., Marino, J.P., and Kelman, Z. (2025). Engineering GID4 for use as an N-terminal proline binder via directed evolution. Biotechnol. Bioeng. *122*, 179–188. https://doi.org/10.1002/bit.28868.

28. Alfaro, J.A., Bohländer, P., Dai, M., Filius, M., Howard, C.J., van Kooten, X.F., Ohayon, S., Pomorski, A., Schmid, S., Aksimentiev, A., et al. (2021). The emerging landscape of single-molecule protein sequencing technologies. Nat. Methods *18*, 604–617. https://doi.org/10.1038/s41592-021-01143-1.

29. Chinnaraj, M., Lin, J., Blacklock, K., Hermes, E., Meyer, M., Pike, D., Vieceli, J., and Chorny, I. (2024). Detecting Amino Acid Variants Using Next-Generation Protein Sequencing (NGPS).

875 Preprint at bioRxiv, https://doi.org/10.1101/2024.12.17.629036
876 https://doi.org/10.1101/2024.12.17.629036.

877 30. Hempfling, J.P., Sekera, E.R., Sarkar, A., Hummon, A.B., and Pei, D. (2022). Generation of
878 Proteins with Free N-Terminal Cysteine by Aminopeptidases. J. Am. Chem. Soc. *144*, 21763–
879 21771. https://doi.org/10.1021/jacs.2c10194.

880 31. Tolbert, T.J., and Wong, C.-H. (2002). New Methods for Proteomic Research: Preparation of
881 Proteins with N-Terminal Cysteines for Labeling and Conjugation. Angew. Chem. Int. Ed. *41*,
882 2171–2174. https://doi.org/10.1002/1521-3773(20020617)41:12<2171::AID-
883 ANIE2171>3.0.CO;2-Q.

884 32. Rosen, C.B., and Francis, M.B. (2017). Targeting the N terminus for site-selective protein
885 modification. Nat. Chem. Biol. *13*, 697–705. https://doi.org/10.1038/nchembio.2416.

886 33. Paupelin-Vaucelle, H., Boschiero, C., Lazennec-Schurdevin, C., Schmitt, E., Mechulam, Y.,
887 Marlière, P., and Pezo, V. (2025). Cys-tRNAj as a Second Translation Initiator for Priming
888 Proteins with Cysteine in Bacteria. ACS Omega *10*, 4548–4560.
889 https://doi.org/10.1021/acsomega.4c08326.

890 34. Wang, H., Wang, L., Zhong, B., and Dai, Z. (2022). Protein Splicing of Inteins: A Powerful Tool
891 in Synthetic Biology. Front. Bioeng. Biotechnol. *10*.
892 https://doi.org/10.3389/fbioe.2022.810180.

893 35. Theile, C.S., Witte, M.D., Blom, A.E.M., Kundrat, L., Ploegh, H.L., and Guimaraes, C.P. (2013).
894 Site-specific N-terminal labeling of proteins using sortase-mediated reactions. Nat. Protoc.
895 *8*, 1800–1807. https://doi.org/10.1038/nprot.2013.102.

896 36. Williamson, D.J., Fascione, M.A., Webb, M.E., and Turnbull, W.B. (2012). Efficient N-Terminal
897 Labeling of Proteins by Use of Sortase. Angew. Chem. Int. Ed. *51*, 9377–9380.
898 https://doi.org/10.1002/anie.201204538.

899 37. Yoshihara, H.A.I., Mahrus, S., and Wells, J.A. (2008). Tags for labeling protein N-termini with
900 subtiligase for proteomics. Bioorg. Med. Chem. Lett. *18*, 6000–6003.
901 https://doi.org/10.1016/j.bmcl.2008.08.044.

902 38. Weeks, A.M., and Wells, J.A. (2020). N-Terminal Modification of Proteins with Subtiligase
903 Specificity Variants. Curr. Protoc. Chem. Biol. *12*, e79. https://doi.org/10.1002/cpch.79.

904 39. Kortemme, T. (2024). De novo protein design—From new structures to programmable
905 functions. Cell *187*, 526–544. https://doi.org/10.1016/j.cell.2023.12.028.

906 40. Wang, K.H., Roman-Hernandez, G., Grant, R.A., Sauer, R.T., and Baker, T.A. (2008). The
907 Molecular Basis of N-End Rule Recognition. Mol. Cell *32*, 406–414.
908 https://doi.org/10.1016/j.molcel.2008.08.032.

41. Presloid, C.J., Jiang, J., Kandel, P., Anderson, H.R., Beardslee, P.C., Swayne, T.M., and Schmitz, K.R. (2025). ClpS Directs Degradation of N-Degron Substrates With Primary Destabilizing Residues in Mycolicibacterium smegmatis. Mol. Microbiol. *123*, 16–30. https://doi.org/10.1111/mmi.15334.

42. Timms, R.T., Zhang, Z., Rhee, D.Y., Harper, J.W., Koren, I., and Elledge, S.J. (2019). A glycine-specific N-degron pathway mediates the quality control of protein N-myristoylation. Science *365*, eaaw4912. https://doi.org/10.1126/SCIENCE.AAW4912.

43. Khmelinskii, A., Keller, P.J., Bartosik, A., Meurer, M., Barry, J.D., Mardin, B.R., Kaufmann, A., Trautmann, S., Wachsmuth, M., Pereira, G., et al. (2012). Tandem fluorescent protein timers for in vivo analysis of protein dynamics. Nat. Biotechnol. *30*, 708–714. https://doi.org/10.1038/nbt.2281.

44. Kats, I., Khmelinskii, A., Kschonsak, M., Huber, F., Knieß, R.A., Bartosik, A., and Knop, M. (2018). Mapping Degradation Signals and Pathways in a Eukaryotic N-terminome. Mol. Cell *70*, 488-501.e5. https://doi.org/10.1016/j.molcel.2018.03.033.

45. Beardslee, P.C., and Schmitz, K.R. (2024). Toxin-based screening of C-terminal tags in Escherichia coli reveals the exceptional potency of ssrA-like degrons. eLife *13*. https://doi.org/10.7554/eLife.98528.1.

46. 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., et al. (2023). Ubiquitin-independent proteasomal degradation driven by C-degron pathways. Mol. Cell *83*, 1921-1935.e7. https://doi.org/10.1016/j.molcel.2023.04.023.

47. Varshavsky, A. (2005). Ubiquitin Fusion Technique and Related Methods. In Methods in Enzymology Ubiquitin and Protein Degradation, Part B. (Academic Press), pp. 777–799. https://doi.org/10.1016/S0076-6879(05)99051-4.

48. Yen, H.-C.S., Xu, Q., Chou, D.M., Zhao, Z., and Elledge, S.J. (2008). Global Protein Stability Profiling in Mammalian Cells. Science *322*, 918–923. https://doi.org/10.1126/science.1160489.

49. Stenström, C.M., Holmgren, E., and Isaksson, L.A. (2001). Cooperative effects by the initiation codon and its flanking regions on translation initiation. Gene *273*, 259–265. https://doi.org/10.1016/S0378-1119(01)00584-4.

50. Stenström, C.M., Jin, H., Major, L.L., Tate, W.P., and Isaksson, L.A. (2001). Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. Gene *263*, 273–284. https://doi.org/10.1016/S0378-1119(00)00550-3.

51. Zhang, W., Xiao, W., Wei, H., Zhang, J., and Tian, Z. (2006). mRNA secondary structure at start AUG codon is a key limiting factor for human protein expression in *Escherichia coli*. Biochem. Biophys. Res. Commun. *349*, 69–78. https://doi.org/10.1016/j.bbrc.2006.07.209.

945 52. Bivona, L., Zou, Z., Stutzman, N., and Sun, P.D. (2010). Influence of the second amino acid on
946 recombinant protein expression. Protein Expr. Purif. *74*, 248–256.
947 https://doi.org/10.1016/j.pep.2010.06.005.

948 53. Allert, M., Cox, J.C., and Hellinga, H.W. (2010). Multifactorial Determinants of Protein
949 Expression in Prokaryotic Open Reading Frames. J. Mol. Biol. *402*, 905–918.
950 https://doi.org/10.1016/j.jmb.2010.08.010.

951 54. Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and effects of N-terminal codon
952 bias in bacterial genes. Science *342*, 475–479. https://doi.org/10.1126/science.1241934.

953 55. Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and
954 Church, G.M. (2013). Composability of regulatory sequences controlling transcription and
955 translation in Escherichia coli. Proc. Natl. Acad. Sci. *110*, 14024–14029.
956 https://doi.org/10.1073/pnas.1301301110.

957 56. Ojima-Kato, T., Nagai, S., and Nakano, H. (2017). N-terminal SKIK peptide tag markedly
958 improves expression of difficult-to-express proteins in *Escherichia coli* and *Saccharomyces*
959 *cerevisiae*. J. Biosci. Bioeng. *123*, 540–546. https://doi.org/10.1016/j.jbiosc.2016.12.004.

960 57. Ojima-Kato, T., Nishikawa, Y., Furukawa, Y., Kojima, T., and Nakano, H. (2023). Nascent
961 MSKIK peptide cancels ribosomal stalling by arrest peptides in Escherichia coli. J. Biol. Chem.
962 *299*. https://doi.org/10.1016/j.jbc.2023.104676.

963 58. Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009).
964 Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*,
965 894–898. https://doi.org/10.1038/nature08187.

966 59. Wannier, T.M., Ciaccia, P.N., Ellington, A.D., Filsinger, G.T., Isaacs, F.J., Javanmardi, K., Jones,
967 M.A., Kunjapur, A.M., Nyerges, A., Pal, C., et al. (2021). Recombineering and MAGE. Nat. Rev.
968 Methods Primer *1*, 7. https://doi.org/10.1038/s43586-020-00006-x.

969 60. Wang, K.H., Sauer, R.T., and Baker, T.A. (2007). ClpS modulates but is not essential for
970 bacterial N-end rule degradation. Genes Dev. *21*, 403–408.
971 https://doi.org/10.1101/gad.1511907.

972 61. Sen, S., Kho, K.K., and Kunjapur, A. (2025). Developing a flow cytometric method to evaluate
973 the stability of protein N-termini. Preprint.

974 62. Licausi, F., Kosmacz, M., Weits, D.A., Giuntoli, B., Giorgi, F.M., Voesenek, L.A.C.J., Perata, P.,
975 and van Dongen, J.T. (2011). Oxygen sensing in plants is mediated by an N-end rule pathway
976 for protein destabilization. Nature *479*, 419–422. https://doi.org/10.1038/nature10536.

977 63. Zhang, Z., Sie, B., Chang, A., Leng, Y., Nardone, C., Timms, R.T., and Elledge, S.J. (2023).
978 Elucidation of E3 Ubiquitin Ligase Specificity Through Proteome-wide Internal Degron
979 Mapping. Mol. Cell *83*, 3377-3392.e6. https://doi.org/10.1016/j.molcel.2023.08.022.

980 64. Tobias, J.W., and Varshavsky, A. (1991). Cloning and functional analysis of the ubiquitin-
981 specific protease gene UBP1 of Saccharomyces cerevisiae. J. Biol. Chem. *266*, 12021–
982 12028.

983 65. Ben-Bassat, A., Bauer, K., Chang, S.Y., Myambo, K., Boosman, A., and Chang, S. (1987).
984 Processing of the initiation methionine from proteins: properties of the Escherichia coli
985 methionine aminopeptidase and its gene structure. J. Bacteriol. *169*, 751–757.
986 https://doi.org/10.1128/jb.169.2.751-757.1987.

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

990 67. Schuenemann, V.J., Kralik, S.M., Albrecht, R., Spall, S.K., Truscott, K.N., Dougan, D.A., and
991 Zeth, K. (2009). Structural basis of N-end rule substrate recognition in *Escherichia coli* by the
992 ClpAP adaptor protein ClpS. EMBO Rep. *10*, 508–514.
993 https://doi.org/10.1038/embor.2009.62.

994 68. Heathcote, K.C., Keeley, T.P., Myllykoski, M., Lundekvam, M., McTiernan, N., Akter, S.,
995 Masson, N., Ratcliffe, P.J., Arnesen, T., and Flashman, E. (2024). N-terminal cysteine
996 acetylation and oxidation patterns may define protein stability. Nat. Commun. *15*, 5360.
997 https://doi.org/10.1038/s41467-024-49489-2.

998 69. Perez, L.F.B.-, and Ryu, Y. (2015). RimJ-Catalyzed Sequence-Specific Protein N-Terminal
999 Acetylation in Escherichia coli. Adv. Biosci. Biotechnol. *6*, 182–193.
1000 https://doi.org/10.4236/abb.2015.63018.

1001 70. Wojtowicz, A., Mazurkiewicz-Pisarek, A., Plucienniczak, G., Mikiewicz-Sygula, D., Chojnacka,
1002 L., Lukasiewicz, N., and Plucienniczak, A. (2005). Expression of yeast deubiquitination
1003 enzyme UBP1 analogues in E. coli. Microb. Cell Factories *4*, 17.
1004 https://doi.org/10.1186/1475-2859-4-17.

1005 71. Bonde, M.T., Klausen, M.S., Anderson, M.V., Wallin, A.I.N., Wang, H.H., and Sommer, M.O.A.
1006 (2014). MODEST: a web-based design tool for oligonucleotide-mediated genome engineering
1007 and recombineering. Nucleic Acids Res. *42*, W408–W415.
1008 https://doi.org/10.1093/nar/gku428.

1009 72. Waskom, M.L. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021.
1010 https://doi.org/10.21105/joss.03021.

1011 73. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.
1012 https://doi.org/10.1109/MCSE.2007.55.

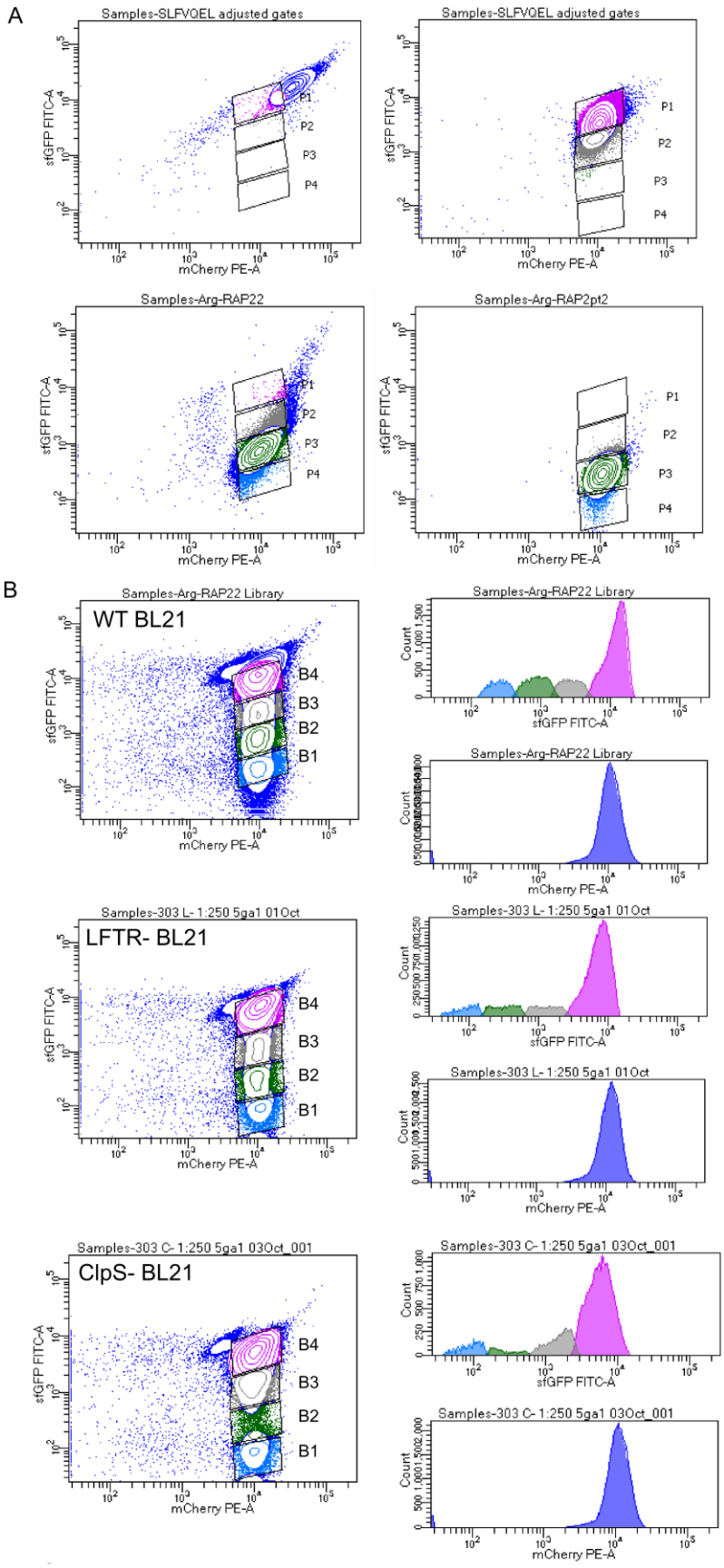
1013 74. Tareen, A., and Kinney, J.B. (2020). Logomaker: beautiful sequence logos in Python.
1014 Bioinformatics *36*, 2272–2274. https://doi.org/10.1093/bioinformatics/btz921.

75. Chen, T., and Guestrin, C. (2016). 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), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

76. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). 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), pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.

77. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS’17. (Curran Associates Inc.), pp. 4768–4777.

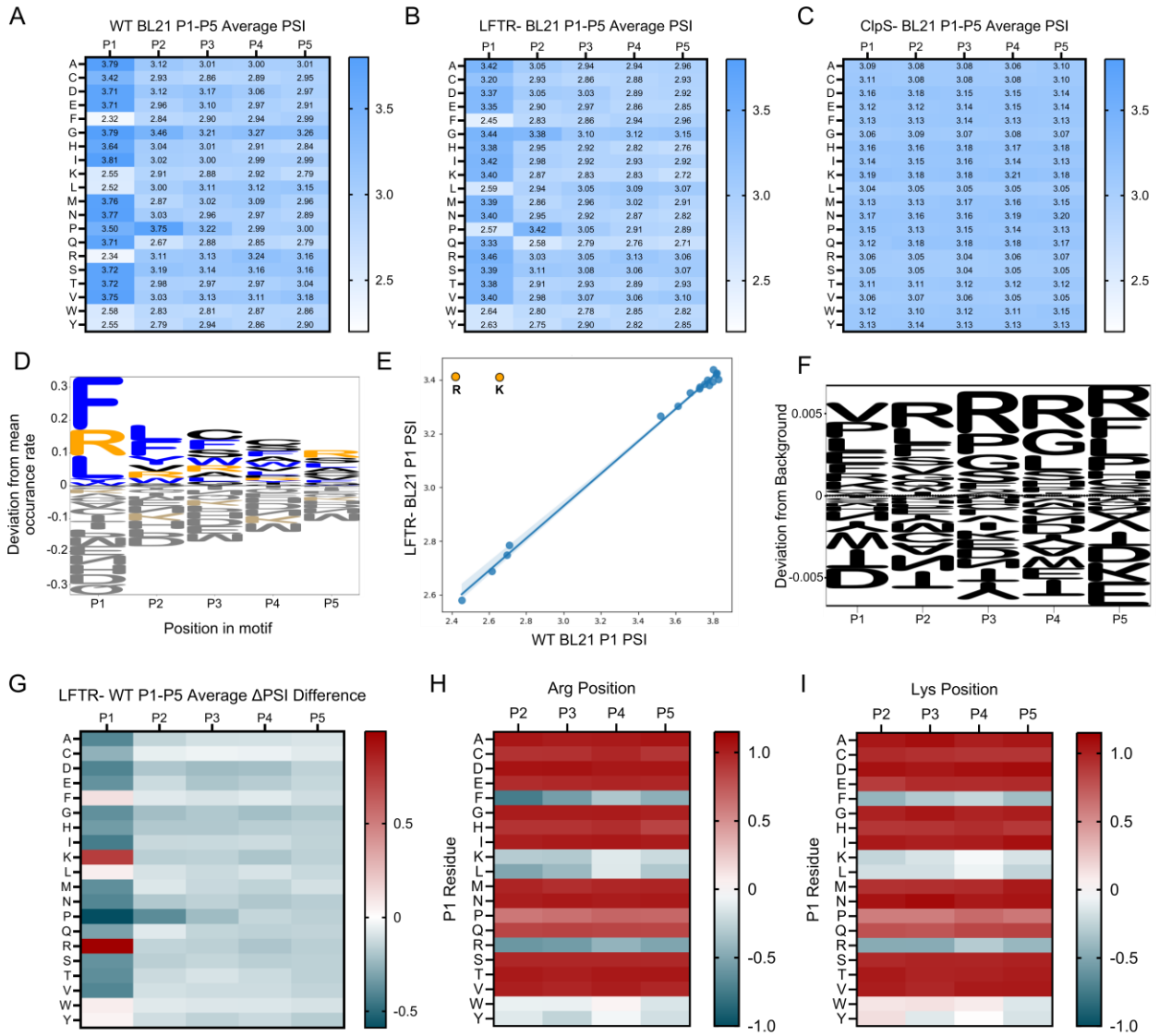
1027 **Supplemental File 1.** Supplemental Figures 1-6 and Supplemental Tables 1-2


A
Samples-SLFVQEL adjusted gates
Samples-SLFVQEL adjusted gates
Samples-Arg-RAP22
Samples-Arg-RAP2pt2
P1
P2
P3
P4
sfGFP FITC-A
mCherry PE-A
B
Samples-Arg-RAP22 Library
WT BL21
Samples-303 L- 1:250 5ga1 01Oct
LFTR- BL21
Samples-303 C- 1:250 5ga1 03Oct_001
ClpS- BL21
B4
B3
B2
B1
Count

1028

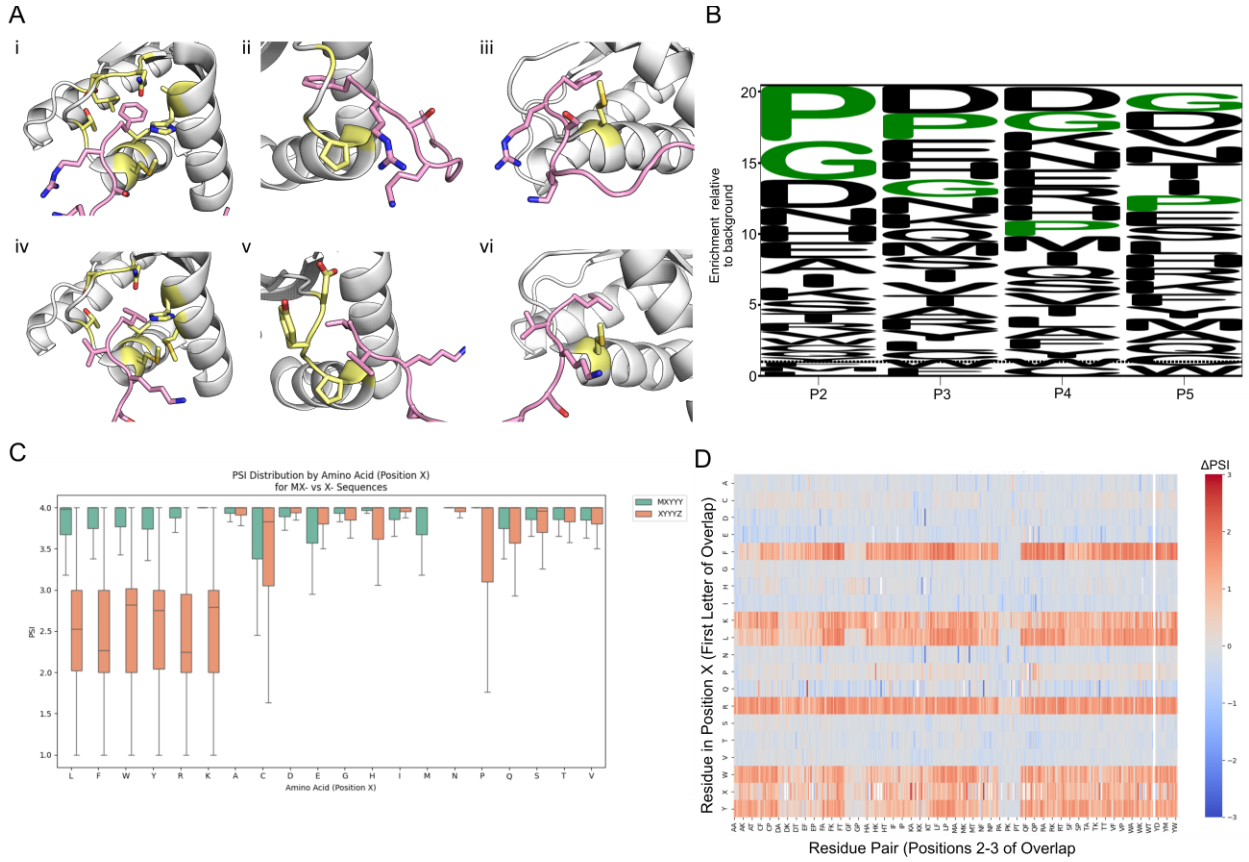
1029 **Supplemental Figure 1.** A.) Differentiation of stable/unstable sort controls. SLFVQEL (stable, top)
1030 and RCGGAIISDFI (unstable, arginylated library template, bottom) sequences hardcoded into the
1031 dual fluorescent reporter are differentiated across two separate examples on a BD FACS Aria
1032 Fusion. Left samples were collected prior to the WT 10M event sort. B.) 10M event sort information.
1033 sfGFP (y-axis) vs. mCherry (x-axis) distributions for the sorted WT (top), $LFTR^-$ (middle), and $ClpS^-$
1034 (bottom) libraries with corresponding bins (left). Histograms for sfGFP & mCherry fluorescence
1035 (right).

1036

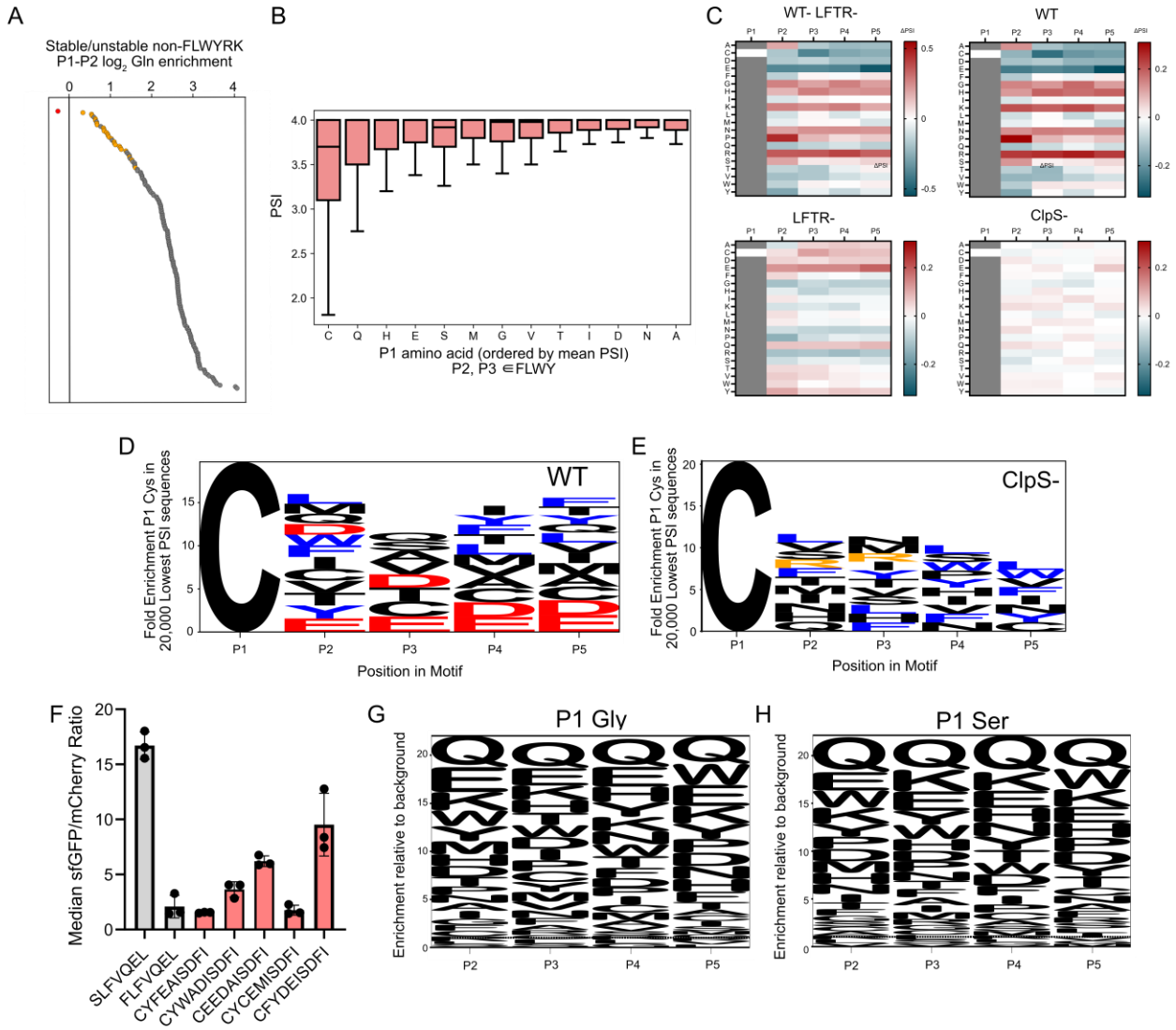


Supplemental Figure 2. A.) P1-P5 WT BL21 Heatmap A mean PSI heatmap generated from 2.29 million sequences for the first five positions of a protein in wild-type BL21. B.) P1-P5 LFTR- BL21 Heatmap A mean PSI heatmap generated from 2.34 million sequences for the first five positions of a protein in BL21 deficient in aat (LFTR). C.) 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. D.) 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. E.) P1 amino acid correlation between WT BL21 and LFTR- BL21 sorts Correlation between PSI scores for N-terminal residues between

1047 WT BL21 and LFTR- BL21 (r = 1.00) in the 10 million event sort. Arg/Lys/Pro were excluded from
1048 regression analysis. Arg and Lys separately highlighted in orange. Ubiquitin upstream of P1 Pro is
1049 not fully cleaved, complicating analysis. F.) Pro, Arg, Leu, and small amino acids are weakly
1050 enriched in low PSI sequences in the ClpS- dataset. Enrichment values of amino acids in
1051 sequences with a PSI <2 in the ClpS- dataset were subtracted from enrichment values from the
1052 whole dataset. Changes in representation cumulatively were <1% in either direction. G. P1 Arg and
1053 P1 Lys stability shifts are observed as a primary difference between LFTR- and WT datasets.
1054 Heatmap depicts the change in PSI for each AA-position combination between LFTR- and WT
1055 datasets. H. Internal Arg heatmaps per P1 residue. Heatmaps depict the average PSI for Arg at each
1056 position for any given P1 residue in the WT dataset. I. Internal Lys heatmaps per P1 residue.
1057 Heatmaps depict the average PSI for Lys at each position for any given P1 residue in the WT
1058 dataset.

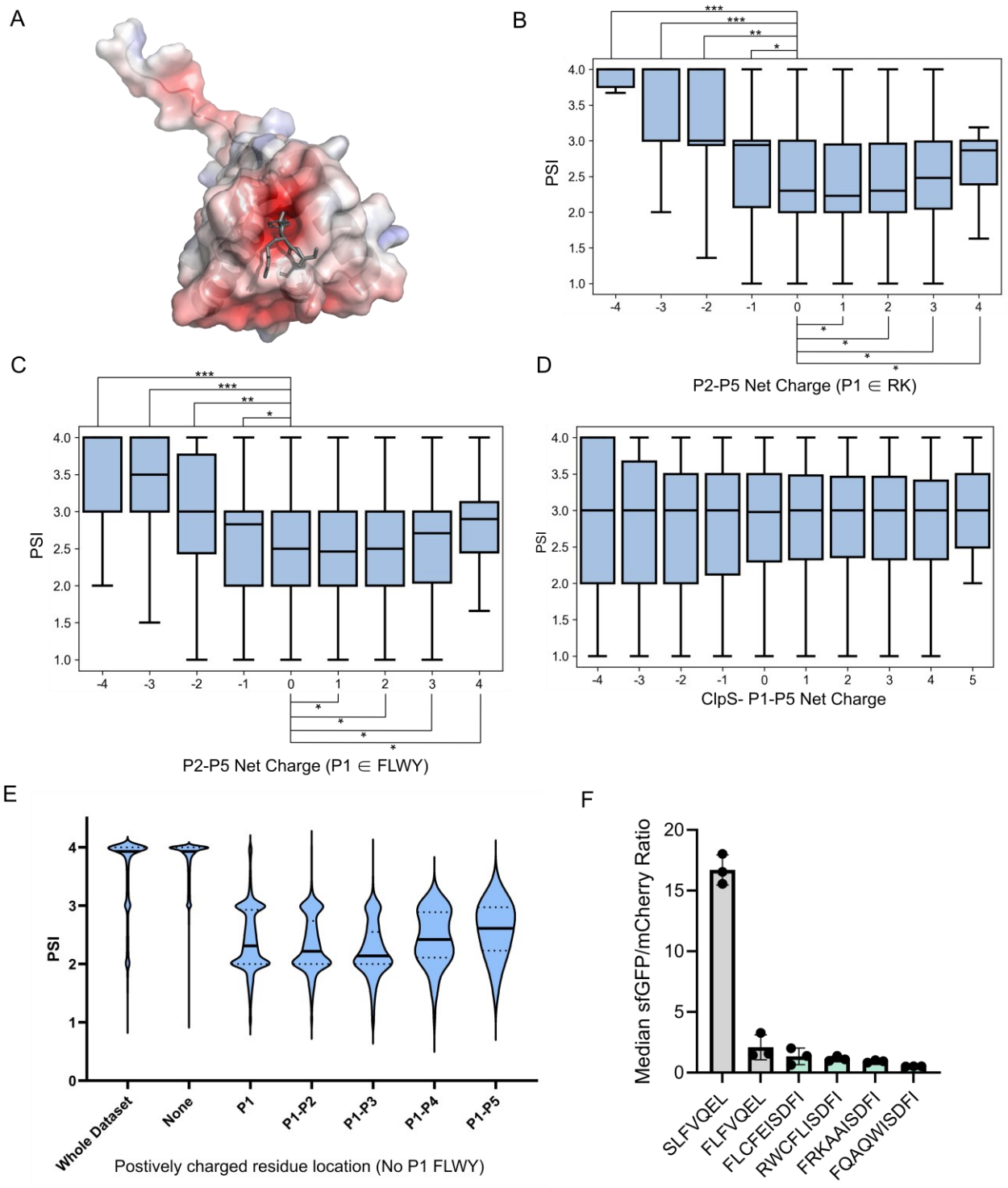


Supplemental Figure 3. A.) Structures of ClpS bound to peptide substrates provide limited insight into the molecular basis of substrate specificity downstream of P1. i-iii: Structures of ClpS bound to Fpep with nearby ClpS residues highlighted in yellow for (i) P1, (ii) P2, and (iii) P3. (PDB ID: 2WA8) iv-vi: Structures of ClpS bound to Lpep with nearby ClpS residues highlighted in yellow for (iv) P1, (v) P2, and (vi) P3. (PDB ID: 2W9R). No interacting residues within 4 A were obtained for P4 and P5 of the peptides. B.) 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. C.) MX- Motifs do not show similar PSI distributions for P1 LFWYRK motifs. Mean PSI boxplots for residues in position X from the 10M event sort. MXYYY is in green and XYYYZ is in orange. D.) Change in PSI between MXYYY and XYYYZ datasets, with the X-position on the y-axis and the first two Y positions in the x-axis.



Supplemental Figure 4. A.) P1/P2 Gln is noticeably enriched in low stability P1-P2 motifs that do not have P1 FLWYRK. Log2 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. B.) 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. C.) P1 Cys mean PSI heatmaps in various backgrounds reveal P2-P5 charge and bulky residue preferences. Sequences with P1 Cys were filtered in each dataset. Data represents differences from the P1 Cys global mean. WT- LFTR- represents the subtraction of difference from the mean values between the WT and LFTR- datasets. D.) Low stability P1 Cys sequences are rich in negatively charged and bulky amino acids at P2-P5. Weblogo depicting enriched amino acids for

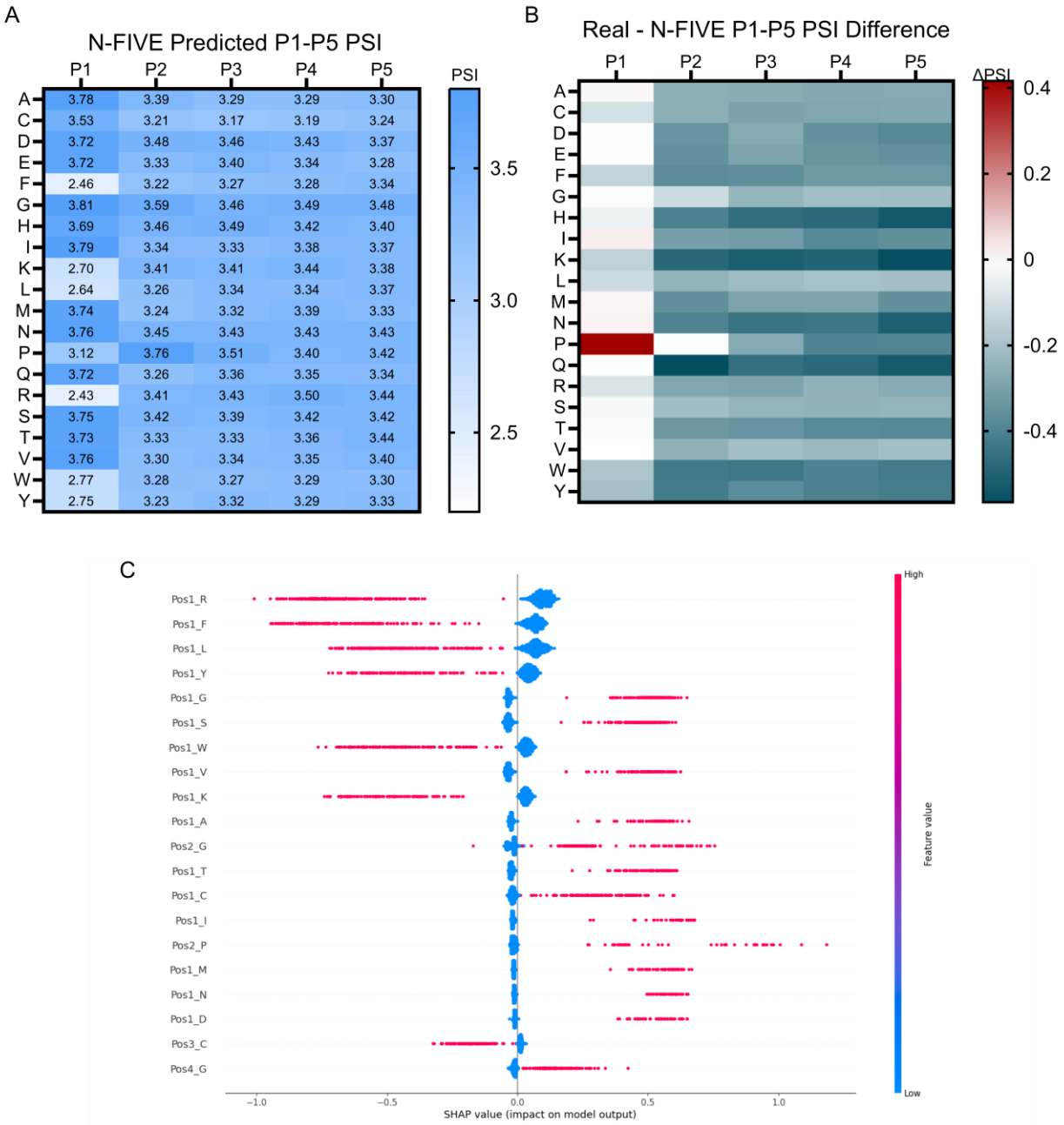
1084 the 20,000 lowest PSI motifs with P1 Cys in the WT dataset. Bulky residues (FLWY) depicted in
1085 blue and negatively charged residues (DE) depicted in red. E.) Negatively charged residues are
1086 absent in low stability P1 Cys sequences. Weblogo depicting enriched amino acids for the 20,000
1087 lowest PSI motifs with P1 Cys in the ClpS- dataset. Bulky residues (FLWY) depicted in blue and
1088 postively charged residues (RK) depicted in orange. F.) Full panel P1 Cys screen reveals putative
1089 N-degrons. Flow cytometry data collected from biological triplicates for the full panel of Nt-Cys
1090 screened sequences. G.) Enriched residues in low stability sequences for P1 Gly. Weblogo
1091 depicting the enrichment of amino acids relative to the whole population for the lowest 20,000 PSI
1092 sequences with P1 Gly. H.) Enriched residues in low stability sequences for P1 Ser. Weblogo
1093 depicting the enrichment of amino acids relative to the whole population for the lowest 20,000 PSI
1094 sequences with P1 Ser.



1095

1096 **Supplemental Figure 5.** A.) The ClpS binding pocket has negative electrostatic potential. ClpS
1097 bound to a substrate (PDB 2WA8) visualized using the APBS Pymol plugin to create a surface map

1098 where regions of negative electrostatic potential are highlighted in red and regions of positive
1099 electrostatic potential are highlighted in blue. B.) Separation of charged and bulky P1 residue net
1100 charge heatmaps. C.) ClpS- sequences show charge parity. P1-P5 net charge heatmap for data
1101 collected in ClpS- Bl21. D.) Multiple positively charged residues at and near P1 lead to low PSI
1102 sequences. Violin plot depicting the PSI distribution in wild-type BL21 for sequences with one or
1103 more RK residues in series at and adjacent to the N-terminus. E.) Sequences with multiple
1104 FLWYRK(Q) residues show low stability. Flow cytometry data from the protein stability assay
1105 depicting sequences carrying multiple FLWYRK or P2 Gln residues exhibited low sfGFP:mCherry
1106 ratios relative to literature controls. All p-values in comparison to 0 P2-P5 Gly/Ser show $p < 0.05$.
1107 p-values evaluated using a Mann-Whitney U test, effect size magnitude (ES) ranges from 0 to 1, as
1108 evaluated using Cliff's delta. * = $p < 0.05$, ES 0-0.25; ** = $p < 0.05$, ES 0.25-0.50, *** = $p < 0.05$, ES
1109 0.50-0.75, **** = $p < 0.05$, ES 0.75-1.0



Supplemental Figure 6. A.) N-FIVE PSI predictions for the mean PSI heatmap for all 3.2 million sequences in wild-type *Escherichia coli* BL21. B.) Change in PSI for each cell between the collected WT BL21 dataset and the N-FIVE predicted dataset. C.) Individual matrix components for N-FIVE predictions. SHAP values showing the contribution of the 20 most impactful matrix features onto the N-FIVE model.

1117 **Supplemental Table 1. 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

Supplemental Table 2. Important oligos

Oligos marked with SEQ were used for Sanger sequencing. Neo-N-terminus (NT) cloning featured a target NT primer + oSS196 as piece 1 and oSS195 + oSS055 as piece 2 used to generate amplicons from a pEVOL dual reporter plasmid.

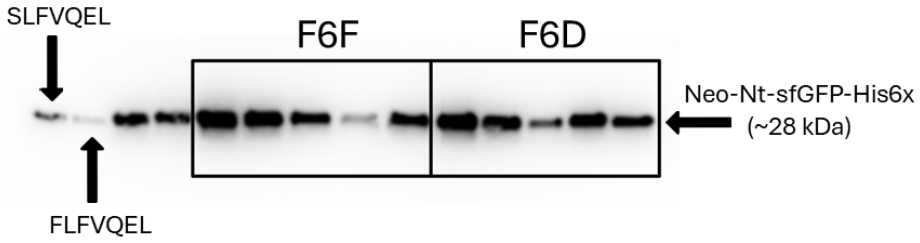
Oligo ID	Sequence	Purpose
Sequencing oligos		
oSS045 SEQ REV	tctagcacgcgtaccatg	sfGFP Sanger sequencing
oSS083 SEQ FWD	Cacggcagaaaagtccac	mCherry Sanger sequencing
oSS084 SEQ FWD	CCAGCAACGTCTGATTTTC	Neo-Nt Sanger sequencing
oSS264 SEQ FWD	caacgaggactacaccatcg	mCherry/Ub Sanger sequencing
Neo-N-terminus insert cloning oligos		
oSS155 FWD	TTCTGCGTCTGCGTGGAGGACGTTGCGG GGGTGCTATCATCAGTGACTTCATCgcatcc aagggcGCGG	ArgRAP2.2 NT cloning
oSS199 FWD	TTCTGCGTCTGCGTGGAGGAAGCCTGTTT GTGCAGGAACTGgcatccaagggcgaggag	SLFVQEL NT cloning
oSS200 FWD	TTCTGCGTCTGCGTGGAGGATTTCTGTTTG TGCAGGAACTGgcatccaagggcgaggag	FLFVQEL NT cloning
oSS262 FWD	TGGTTCTGCGTCTGCGTGGAGGATTGGTG AAGACCAAGGCGTCCAATTTATTAgcatcca agggcgaggag	DPS NT cloning
oSS388 FWD	TTCTGCGTCTGCGTGGAGGAGATCATGGA TCGGGTGCCTGGCTGTTGCCGgcatccaagg gcgaggag	D-Bgal NT cloning

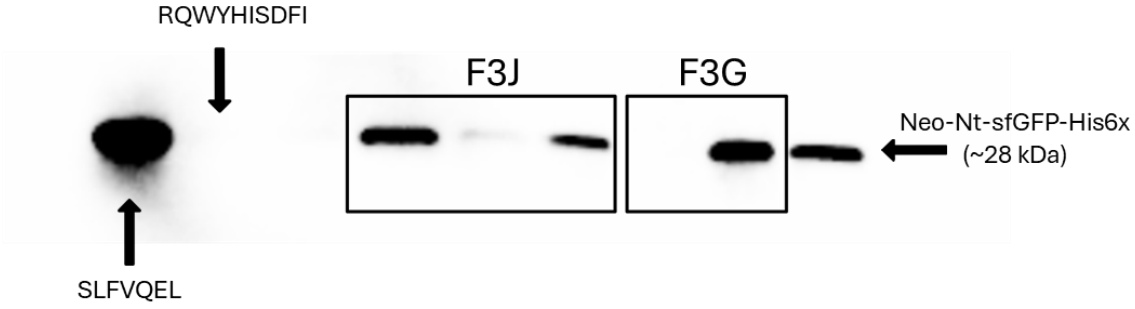
oSS698 FWD	TTCTGCGTCTGCGTGGAGGANNKNNKNNKNNKNNKATCAGTGACTTCATCgcatccaag	P1-P5 NT library cloning
oSS1035 FWD	TTCTGCGTCTGCGTGGAGGATGTTTTTACGAAGCTATCAGTGACTTCATCgcatccaag	CFYEA NT FWD
oSS1036 FWD	TTCTGCGTCTGCGTGGAGGATGCTATTGGGCCGACATCAGTGACTTCATCgcatccaag	CYWAD NT FWD
oSS1037 FWD	TTCTGCGTCTGCGTGGAGGATGCGAGGAGGACGCAATCAGTGACTTCATCgcatccaag	CEEDA NT FWD
oSS1038 FWD	TTCTGCGTCTGCGTGGAGGATGTTACTGTGAGATGATCAGTGACTTCATCgcatccaag	CYCEM NT FWD
oSS1039 FWD	TTCTGCGTCTGCGTGGAGGATGTGACTACGACGAGATCAGTGACTTCATCgcatccaag	CFYDE NT FWD
oSS1040 FWD	TTCTGCGTCTGCGTGGAGGACCGTACTTTGACGAGATCAGTGACTTCATCgcatccaag	PYDFE NT FWD
oSS1041 FWD	TTCTGCGTCTGCGTGGAGGAATTTTTTTTGCAGAGATCAGTGACTTCATCgcatccaag	IFFAE NT FWD
oSS1042 FWD	TTCTGCGTCTGCGTGGAGGAATTTGGTATAAGCATATCAGTGACTTCATCgcatccaag	IWYKH NT FWD
oSS1043 FWD	TTCTGCGTCTGCGTGGAGGATTTCAGAAGCAAAAAATCAGTGACTTCATCgcatccaag	FQKQK NT FWD
oSS1044 FWD	TTCTGCGTCTGCGTGGAGGACGTCAATGGTATCACATCAGTGACTTCATCgcatccaag	RQWYH NT FWD
oSS1045 FWD	TTCTGCGTCTGCGTGGAGGATTTGACGAGGATGAAATCAGTGACTTCATCgcatccaag	FDEDE NT FWD
oSS1046 FWD	TTCTGCGTCTGCGTGGAGGATTCGATGAGGCGGCCATCAGTGACTTCATCgcatccaag	FDEAA NT FWD
oSS1047 FWD	TTCTGCGTCTGCGTGGAGGATTTGCAGCCGCAGCCATCAGTGACTTCATCgcatccaag	FAAAA NT FWD
oSS1048 FWD	TTCTGCGTCTGCGTGGAGGATTTCGCAAGGCAGCTATCAGTGACTTCATCgcatccaag	FRKAA NT FWD
oSS1050 FWD	TTCTGCGTCTGCGTGGAGGAGGATTGGTGCGTGGTATCAGTGACTTCATCgcatccaag	GLVRG NT FWD

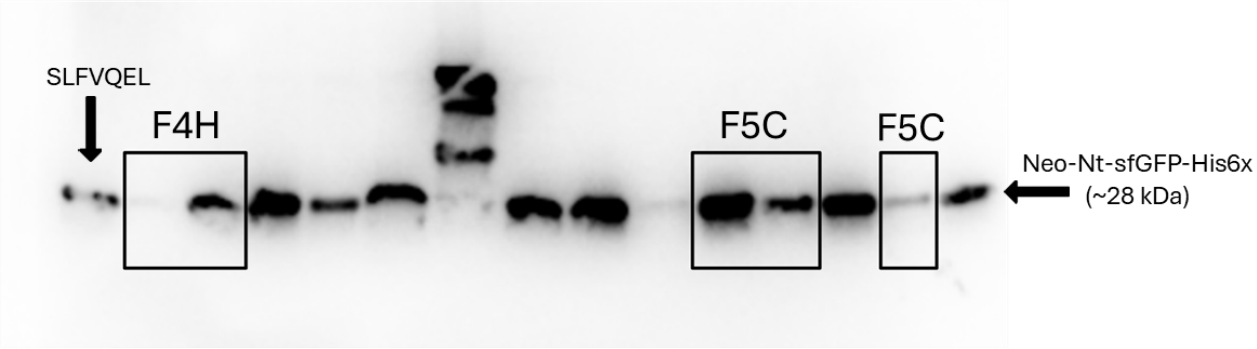
oSS1051 FWD	TTCTGCGTCTGCGTGGAGGAGGTTTGCGC TTACGCATCAGTGACTTCATCgcatccaag	GLRLR NT FWD
oSS1052 FWD	TTCTGCGTCTGCGTGGAGGATTGGTGCGC CTGGGGATCAGTGACTTCATCgcatccaag	LVRLG NT FWD
oSS1053 FWD	TTCTGCGTCTGCGTGGAGGAGGCAGTGTT CTTCGTATCAGTGACTTCATCgcatccaag	GSVLR NT FWD
oSS1054 FWD	TTCTGCGTCTGCGTGGAGGACACTGGATG CAAAAAATCAGTGACTTCATCgcatccaag	HWMQK NT FWD
oSS1055 FWD	TTCTGCGTCTGCGTGGAGGAAACATTCAG ATGAAAATCAGTGACTTCATCgcatccaag	NIQMK NT FWD
oSS1056 FWD	TTCTGCGTCTGCGTGGAGGATCATTGTGG AAATACATCAGTGACTTCATCgcatccaag	SLWKY NT FWD
oSS1057 FWD	TTCTGCGTCTGCGTGGAGGATCTTACATG ATGGATATCAGTGACTTCATCgcatccaag	SYMMD NT FWD
oSS1058 FWD	TTCTGCGTCTGCGTGGAGGAATCAAGAAA ACGCAGATCAGTGACTTCATCgcatccaag	IKKTQ NT FWD
oSS1059 FWD	TTCTGCGTCTGCGTGGAGGACTTGCGGC CAGCTACATCAGTGACTTCATCgcatccaag	LAASY NT FWD
oSS1060 FWD	TTCTGCGTCTGCGTGGAGGACTTGACCGT AATAGCATCAGTGACTTCATCgcatccaag	LDRNS NT FWD
oSS1061 FWD	TTCTGCGTCTGCGTGGAGGATTAGCCGTC CCGCGTATCAGTGACTTCATCgcatccaag	LAVPR NT FWD
oSS1062 FWD	TTCTGCGTCTGCGTGGAGGATTAGGCCC CCGTTCAATCAGTGACTTCATCgcatccaag	LRPRS NT FWD
oSS1063 FWD	TTCTGCGTCTGCGTGGAGGATTAGGGGAT TCAGTCATCAGTGACTTCATCgcatccaag	LGDSV NT FWD
oSS1064 FWD	TTCTGCGTCTGCGTGGAGGACAGCAGTGT AAATGGATCAGTGACTTCATCgcatccaag	QQCKW NT FWD
oSS1065 FWD	TTCTGCGTCTGCGTGGAGGACATCAGGC CCAGTGGATCAGTGACTTCATCgcatccaag	HQAQW NT FWD

oSS1066 FWD	TTCTGCGTCTGCGTGGAGGATTCCAAGCG CAATGGATCAGTGACTTCATCgcatccaag	FQAQW NT FWD
oSS1067 FWD	TTCTGCGTCTGCGTGGAGGATGGGGAGC AGCAGTGATCAGTGACTTCATCgcatccaag	WGGAV NT FWD
oSS1068 FWD	TTCTGCGTCTGCGTGGAGGACGTGGGGC CGCCGTTATCAGTGACTTCATCgcatccaag	RGAAV NT FWD
oSS1069 FWD	TTCTGCGTCTGCGTGGAGGACTGGGCGC AGCAGTTATCAGTGACTTCATCgcatccaag	LGAAV NT FWD
oSS1070 FWD	TTCTGCGTCTGCGTGGAGGATTTTTGTGTT TTGAAATCAGTGACTTCATCgcatccaag	FLCFE NT FWD
oSS1071 FWD	TTCTGCGTCTGCGTGGAGGACGCTGGTG CTTCTTGATCAGTGACTTCATCgcatccaag	RWCFL NT FWD
Other cloning oligos		
oSS055 REV	TCCTCCACGCAGACGCAG	NT cloning
oSS195 FWD	gagtgtatactggcttactatgttggca	pEVOL mid BB cloning
oSS196 REV	tgccaacatagtaagccagtatacactc	pEVOL mid BB cloning

1126 **Supplemental File 2.** Uncropped Western blots







1127

1128 A consecutive lane was skipped in Figure 5C due to compounding factors that may promote
1129 stability for a neutrally charged motif, such as the presence of repetitive small amino acids at P2-
1130 P5 (FAAAISDFI).

1131

1132