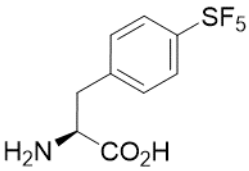
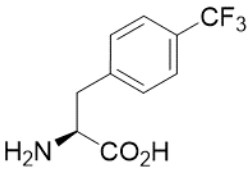
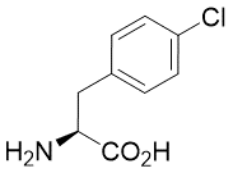
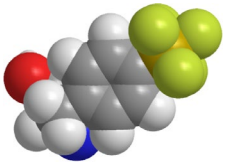
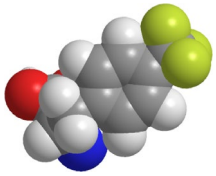
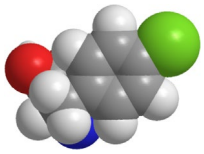
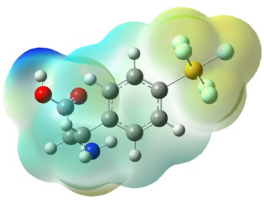
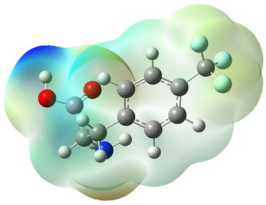
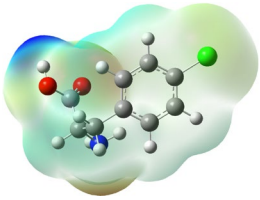
Supporting Information

Genetic Encoding of para-Pentafluorosulfanyl Phenylalanine: A Highly Hydrophobic and Strongly Electronegative Group for Stable Protein Interactions

Haocheng Qianzhu, Adarshi P. Welegedara, Holly Williamson, Amy E. McGrath, Mithun C. Mahawaththa, Nicholas E. Dixon, Gottfried Otting and Thomas Huber*

Table of Contents

Table S1. Comparison of the chemical properties of the unnatural amino acids used in this study. Molecular electrostatic potential surfaces were calculated by Gaussian09 software packages, with density functional theory at B3LYP level with 6-311++G (d,p) basis set in vacuum.[1]

	SF_5Phe	CF_3Phe	ClPhe
			
Space-filling model			
V_{vdW} of the functional group ($Å^3$)[2-4]	95.57	39.80	22.45
Electrostatic potential surface[1]			
Relative electronegativity of the functional group[5-6]	3.65	3.36	-
Hansch hydrophobicity of the functional group (π)[7]	1.51	1.09	0.76
Dipole moment of the side chain (D)[8]	3.44	2.60	-

Functional aminoacyl-tRNA synthetase screening

Construction of library plasmid pBK-MjYRS

A previously published *Mj*TyrRS mutant (CNPheRS) with polysubstrate specificity was used as library template **(Figure S1)**.[9] Five residues in the active site (Leu32, Val65, Met109, Ala159, Leu162) were fully randomized and Gly158 was mutated to a mixture of glycine and serine, known to reduce the affinity to tyrosine as false positive substrate.[10] A single amino acid substitution (D286R) was made to increase the overall aminoacylation rate.[11] A mutant *GlnS* promoter (*GlnS'*) was used for the *Mj*TyrRS gene[11], and the *lpp* promoter for an optimized version of ^{Tyr}tRNA$_{CUA}$ (opt-*Mj*YtRNA).[12]

A pBK-CNPheRS plasmid was first constructed to serve as the template for library plasmid pBK-MjYRS (**Figure S2a**), by introducing the CNPheRS (D286R)-encoding gene[11] into a pBK vector. This plasmid has a kanamycin resistance (*Kan^R*) marker and a pBR322 (*ColE1*) origin of replication. One extra restriction enzyme digestion site for *Sal*I was introduced. An expression cassette for an optimized version of ^{Tyr}tRNA$_{CUA}$ (opt-*Mj*YtRNA)[12] under the control of the *lpp* promoter was cloned into the vector in the opposite direction of the *Kan^R* and the synthase gene, while a mutant *GlnS* promoter (*GlnS'*)[11] was used for the synthase gene transcription.

To construct the plasmid DNA library encoding several aminoacyl tRNA synthestase (aaRS) variants, randomized triplet NNK codons (N = A + G + T + C and K = G + T) were introduced at the intended mutation sites using doped oligonucleotide primers containing these codons. Residues Tyr32, Val65, Met109, Ala159, Leu162 were targeted for full randomization and Gly158 was mutated to a mixture of Gly and Ser. This library theoretically encodes 6.4×10^6 variants of *Mj*TyrRS mutants. In-house Vent DNA polymerase was used to conduct overlap-extension PCR.

The final assembled fragment was digested with restriction enzymes *Nde*I and *Sal*I, gel-purified and ligated back into the pBK vector using T4 DNA ligase to afford the library of pBK-MjYRS variants. The ligation products were passed through the Wizard PCR clean-up system (Promega) and transferred into *E. coli* DH10B cells (10^9 competency) by electroporation. Electroporated cells were recovered in supra-optimal broth (SOB) medium for 1 h at 37ºC and then transferred into 1 L 2YT medium with kanamycin (50 μg/mL) and grown at 37ºC (180 rpm) to OD_{600nm} = 1.0. To estimate the library size, 50 μL of recovered SOB culture was removed and subjected to dilution in Milli-Q H_2O, plated on Luria-Bertani (LB) agar plates with kanamycin (50 μg/mL), and grown overnight at 37ºC. Based on the colony numbers on these plates, pBK-MjYRS contained approximately 10^8 individual transformants. DNA sequencing of 10 colonies revealed no significant bias. Library plasmid DNA was then purified from the amplified 1 L 2YT culture by a Zippy Maxiprep kit and used in the subsequent selection. All primers and gene fragments were synthesized by Integrated DNA Technologies (IDT).

Construction of selection plasmid pBAD-GRFP

A selection plasmid was constructed with a reporter for fluorescent screening, using a fusion of mNeoGreen green fluorescent protein (GFP) and mCherry red fluorescent protein (RFP)

with a linker interrupted by an amber codon (**Figure S2b**).[13-14] The gene of this GFP-RFP fusion protein (GRFP) was placed in the pBAD plasmid under the control of the *araBAD* promoter. This plasmid has an ampicillin-resistance marker and a *p15A* origin of replication. The optimal excitation and emission wavelengths of the green fluorescence are 506 and 560 nm, respectively, and it is 582 and 640 nm, respectively, for the red fluorescence. Fluorescence scanning was performed by a plate reader (SpectraMax M2, Molecular Devices). The level of red fluorescence was used to monitor amber codon read-through, because the background fluorescence from cellular components is very low within the RFP emission range. This double reporter allowed using the green fluorescence to identify healthy cell populations on the cell sorter and to normalize cell density without an additional OD_{600nm} absorbance measurement.

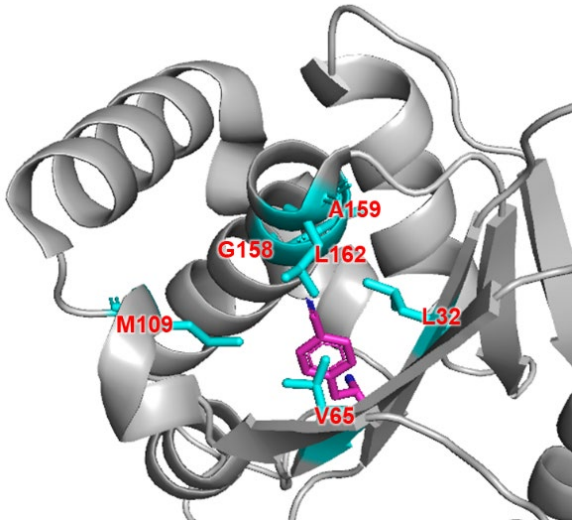
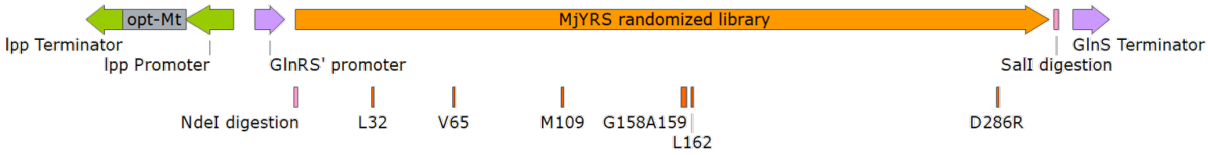


Figure S1. Crystal structure of the amino acid binding pocket of *p*-cyano-phenylalanine tRNA synthetase (CNPheRS; PDB ID: 3QE4[9]). Five residues (Leu32, Val65, Met109, Ala159, Leu162) are identified, which are within 6.5 Å of the *para*-position of the aromatic ring of *p*-cyano-phenylalanine (CNPhe). Gly158 is important for the affinity to tyrosine.[10]

(a)



(b)

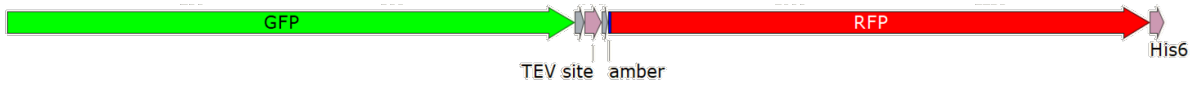


Figure S2. (a) Construct of library plasmid pBK-MjYRS. (b) Construct of the GRFP selection reporter in pBAD-GRFP under the *araBAD* promoter.

Screening of functional SF_5PheRS by FACS cell sorting

To carry out selection, the library plasmid pBK-MjYRS was transformed into *E. coli* DH10B cells harboring the selection plasmid pBAD-GRFP. The recover culture after transformation was directly inoculated into two flasks with 20 mL LB medium containing 100 μg/mL ampicillin and 50 μg/mL kanamycin, supplied with 0.2% arabinose. Each "+" sample contained 1 mM uAA, while the "–" samples were not provided with uAA. Overnight expression at 37°C led to a high level of GFP and a well-detectable level of RFP. Cells were resuspended in 15 mL PBS buffer (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4) after harvest. A 100-fold dilution yielded a concentration suitable for cell sorting by FACS on an Aria II high speed cell sorter (BD Biosciences). General gating was based on monitoring forward and side scattering to select the expected size of bacteria. Only cells displaying high GFP levels were considered healthy and sorted according to the level of RFP chosen as cut-off in each selection round. Positive selection rounds identified those *Mj*TyrRS mutants in the library, which were capable of reading through the amber codon. Negative selection rounds identified *Mj*TyrRS mutants unable to support amber codon read-through without the addition of uAA. Sorted cells from each round were directly inoculated into another 20 mL fresh LB medium containing 100 μg/mL ampicillin, 50 μg/mL kanamycin and 0.2% arabinose, with or without uAA, to iteratively repeat the overnight expression.

Three rounds of selection were carried out for SF_5PheRS selection (**Figure S3a**). SF_5Phe used in this study was purchased from Oakwood Products (CAS: 1266124-33-7). Each positive sample contained 1mM SF_5Phe.The top 7.5% of cells with high RFP levels were selected from the 1P+ round, collecting 2.0×10^6 cells from 9.2×10^7 events in 50 min. The cells collected were subjected to the following round of negative selection (2N–), from where cells with low RFP expression levels (48.7% of the total) were selected, collecting 2.0×10^6 of cells from 4.1×10^6 events in 5 min. These cells were aliquoted to inoculate media with positive (3P+) and negative (3P–) conditions. These two samples displayed a dramatic difference in fluorescence, with 68.1% of the cells displaying high RFP levels in the 3P+ sample compared to 2.2 % in the 3P– sample using the same gating conditions. Cells with high RFP levels were selected from the 3P+ sample, collecting 1.0×10^6 cells from 1.6×10^6 events in 5 min.

The cells selected from the third positive round were allowed to recover on LB agar plates containing 100 μg/mL ampicillin and 50 μg/mL kanamycin and individual clones analyzed using 96-well plates. 160 candidates were inoculated into both positive (+1 mM SF_5Phe) and negative (–uAA) growth conditions. The fluorescence level was then measured after expression overnight, using a plate reader (SpectraMax M2, Molecular Devices) and normalizing the RFP read-out by the level of GFP. In the negative condition, most candidates (156 out of 160 colonies) displayed low RFP concentrations near background level, while most of them showed a significantly higher RFP level in the positive condition. The 32 candidates with the highest RFP level in the positive condition were chosen for sequencing. Every one of them displayed different mutations (**Table S2**). From these candidates, four of them (SF20, SF61, SF87, SF100) with the highest RFP level under positive condition and low RFP level under negative condition were chosen, and another five of them were randomly picked (SF19, SF29, SF88, SF95, SF109) for further characterization.

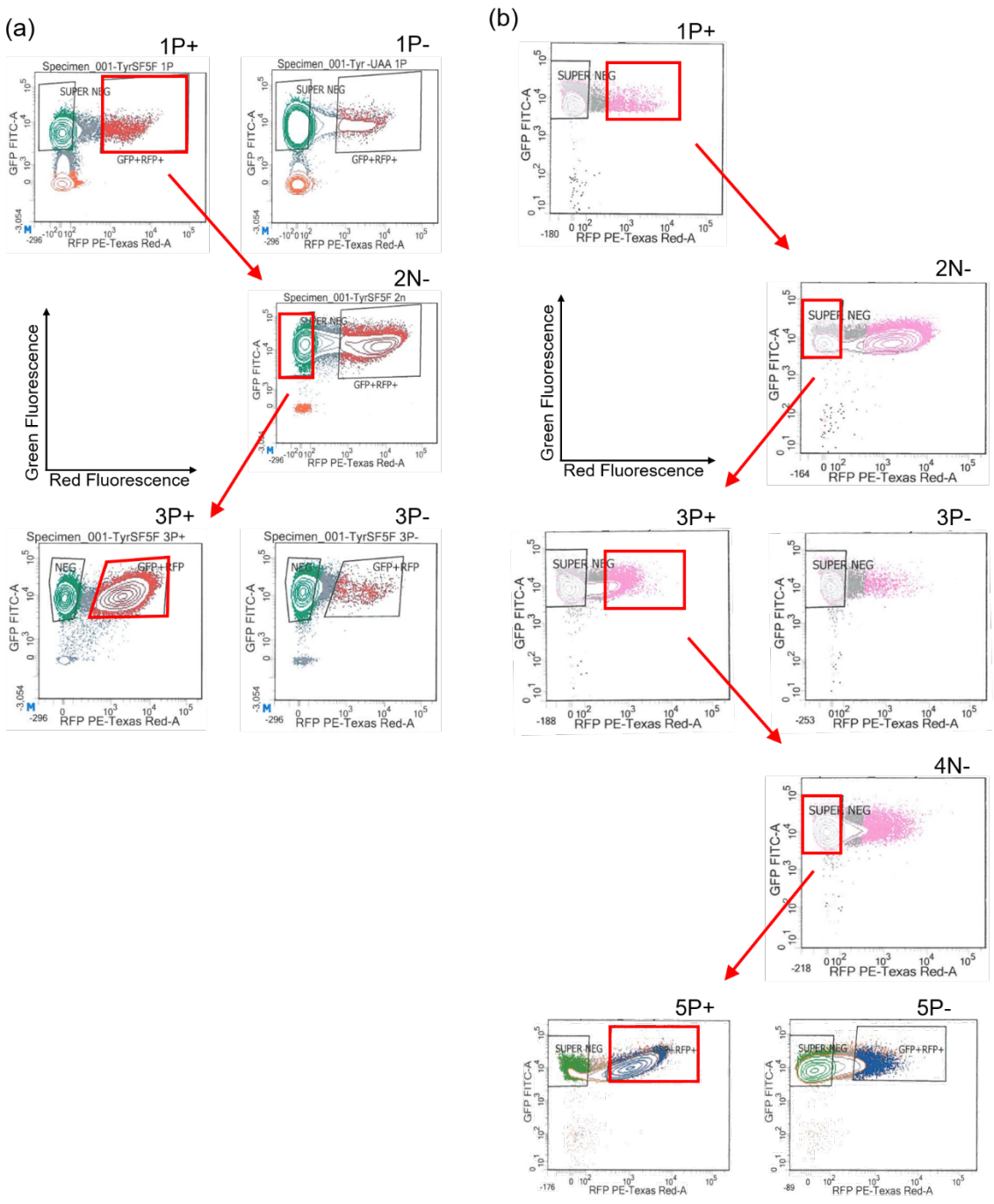


Figure S3. Flow of enrichment of bacterial cells with active and specific (a) SF_5PheRS (b) STyrRS enzymes via FACS cell sorting. Gating of the actual collected R+/R– population is shown by red frame. Looser gating was set in positive selection rounds, while tighter gating was applied in negative rounds.

Table S2. Mutations found in the 32 selected SF_5PheRS and corresponding red fluorescence level under positive growth conditions, normalized by the green fluorescence level. SF20, SF61, SF87, SF100 were potentially best candidates with the highest efficiency and high specificity. They were further characterized together with another five randomly picked candidates (SF19, SF29, SF88, SF95, SF109) by mass spectrometry analysis.

CNPheRS	L32	V65	M109	G158	A159	L162	RFP/GFP +uAA	RFP/GFP -uAA
SF11	V	G	G	G	Y	T	53.61	2.09
SF16	F	V	D	G	G	V	40.08	3.38
SF18	V	S	E	S	V	D	51.94	2.76
SF19	V	V	K	S	C	Y	41.33	2.07
SF20	**G**	**C**	**D**	**G**	**F**	**L**	**68.37**	**1.95**
SF21	H	V	G	G	V	T	71.31	4.91
SF28	V	A	Q	G	G	F	59.43	1.14
SF29	A	A	G	S	V	L	60.01	1.5
SF30	G	G	C	G	Y	V	45.19	2.66
SF38	V	V	C	G	P	S	51.88	1.50
SF54	G	S	E	S	L	R	52.77	0.76
SF60	A	G	S	G	Y	S	53.91	1.47
SF61	**V**	**S**	**A**	**G**	**V**	**K**	**74.44**	**2.17**
SF65	V	C	L	S	I	T	44.15	2.31
SF69	L	C	W	G	M	N	43.87	2.79
SF70	L	P	G	G	C	I	59.64	6.82
SF71	A	A	F	G	V	F	60.79	1.07
SF73	E	A	E	G	V	T	41.78	3.11
SF76	V	V	R	G	V	A	40.25	0.95
SF83	V	C	K	G	L	G	44.01	0.96
SF86	V	N	A	G	Y	V	67.47	7.81
SF87	**G**	**I**	**G**	**G**	**L**	**Y**	**69.5**	**3.85**
SF88	V	V	R	G	Y	S	70.87	14.77
SF95	A	A	E	G	G	L	48.56	3.50
SF100	**V**	**V**	**P**	**G**	**F**	**F**	**70.34**	**1.01**
SF101	V	Q	H	G	Y	S	50.48	4.58
SF106	V	S	V	G	C	T	41.23	1.16
SF109	A	S	A	G	V	S	42.13	2.11
SF110	L	G	C	G	S	F	47.53	1.85
SF141	H	V	M	G	R	F	40.14	1.01
SF157	L	P	G	G	S	W	57.31	1.98
SF160	A	A	G	S	V	M	54.44	1.14

Screening of functional STyrRS by FACS cell sorting

To compare our selection system with previously published methods, we carried out tRNA synthetase screening with *O*-sulfo-tyrosine (STyr), which previously has been shown to be incorporated by a *Mj*YRS mutant sTyrRS (Tyr32Leu, Leu65Pro, Asp158Gly, Ile159Cys, Leu162Lys).[15] During FACS screening (**Figure S3b**), 7.3% of cells with high RFP levels were selected from the 1P+ round and subjected to 2N– round with no STyr presented, where 20.0% of them with low RFP levels were collected for the next round. At the third round, the R+ population started to show difference between 3P+ and 3P– samples (18.2% *vs* 2.8%), telling that the target population did exist but has not been enriched enough. These 18.2% from 3P+ passed through another 4N– round, where 86.3% were collected. At the fifth round, the ratio of R+ between 5P+ and 5P– samples reached 91.5%:10.0%. The diversity of sequences in the 5P+ fraction is estimated to be approximately 1/10 of what we observed in the P3+ fraction of the SF_5PheRS selection. These 91.5% cells from 5P+ were allowed to recover on LB agar plates containing 100 μg/mL ampicillin and 50 μg/mL kanamycin and individual clones analyzed using 96-well plates. 128 candidates were inoculated into both positive (+1 mM STyr) and negative (–uAA) growth conditions, a control sample harboring the published sTyrRS was inoculated as well. The overnight expression of fluorescence reporters was measured by a plate reader and RFP read-out was normalized by the level of GFP. 25 candidates with the highest RFP level in the positive condition and low RFP level (near background) in the negative condition were chosen for sequencing (**Table S3**). Two selected synthetases SY19 (Tyr32Leu, Leu65Ser, Met109Cys, Asp158Gly, Ile159Cys, Leu162Trp) and SY43 (Tyr32Leu, Leu65Ser, Met109Ala, Asp158Gly, Ile159Cys, Leu162Trp) were confirmed by mass spectrometry analysis to specifically incorporate STyr.

Notably, all of the choose candidates had equal or higher RFP level compared to the published sTyrRS. Out of the chosen 25 STyrRS candidates, 23 sets of mutation combinations were individually different and one sequence was observed 3 times. The selection outcome with large sequence diversity highlights the potential of our high-throughput screening method over traditional plate-based selection. Some recent research has used FACS based approaches as well for synthetase engineering and identified multiple synthetase variants.[16-17] On the cell sorter, the selection pressure is readily controlled by changing the gating strategy, which allows adjustment to small sample batch variations. Most importantly, the required amount of unnatural amino acid is greatly reduced compared to plate-based screening, which reduces cost and makes selection feasible even when compounds are difficult to synthesize. In example, throughout the here described screening process less than 20 mg *O*-sulfo-tyrosine was used, and the sorting time was less than 3 hours in total.

Table S3. Mutations found in the 25 selected STyrRS and the previously published sTyrRS, and corresponding red fluorescence level under positive growth conditions, normalized by the green fluorescence level. SY19, SY26 and SY30 possessing the same sequence (Tyr32Leu, Leu65Ser, Met109Cys, Asp158Gly, Ile159Cys, Leu162Trp), together with a close analogue SY43 (Tyr32Leu, Leu65Ser, Met109Ala, Asp158Gly, Ile159Cys, Leu162Trp) were highlighted. All 25 selected STyrRS mutants show higher incorporation of *O*-sulfo-tyrosine into the RFP reporter than the previously published candidate sTyrRS. The incorporation of *O*-sulfo-tyrosine with SY19 and SY43 were confirmed by mass spectrometry analysis.

CNPheRS	L32	V65	M109	G158	A159	L162	RFP/GFP	W108
sTyrRS	L	P	Q(wildtype)	G	C	K	26.09	F(wildtype)
SY2	L	N	S	G	S	L	62.35	
SY6	V	T	Q	G	A	Q	66.51	
SY13	L	S	Q	G	S	M	31.96	
SY15	V	N	Q	G	S	M	32.36	
SY19	L	S	C	G	C	W	51.37	
SY20	L	G	G	G	A	S	39.12	
SY26	L	S	C	G	C	W	46.41	
SY27	Q	S	Q	G	S	R	30.50	
SY28	Q	G	H	G	Q	T	29.51	
SY30	L	S	C	G	C	W	47.49	
SY31	V	S	M	G	S	L	31.53	
SY33	D	K	Q	G	A	R	37.02	
SY35	L	N	Q	G	S	V	35.50	
SY36	L	N	S	G	C	R	40.57	
SY37	L	S	Q	G	G	L	41.56	
SY41	Q	S	N	G	A	W	33.36	
SY43	L	S	A	G	C	W	34.98	
SY45	V	S	G	G	Y	R	41.12	
SY46	L	S	N	G	A	L	28.59	
SY47	V	N	Q	G	A	L	29.03	
SY56	L	S	Q	G	S	Q	40.27	
SY57	D	K	Q	G	M	I	38.80	
SY61	L	C	Q	G	S	T	43.39	
SY62	L	C	Q	G	V	Q	42.88	
SY64	V	T	G	G	Y	R	49.22	

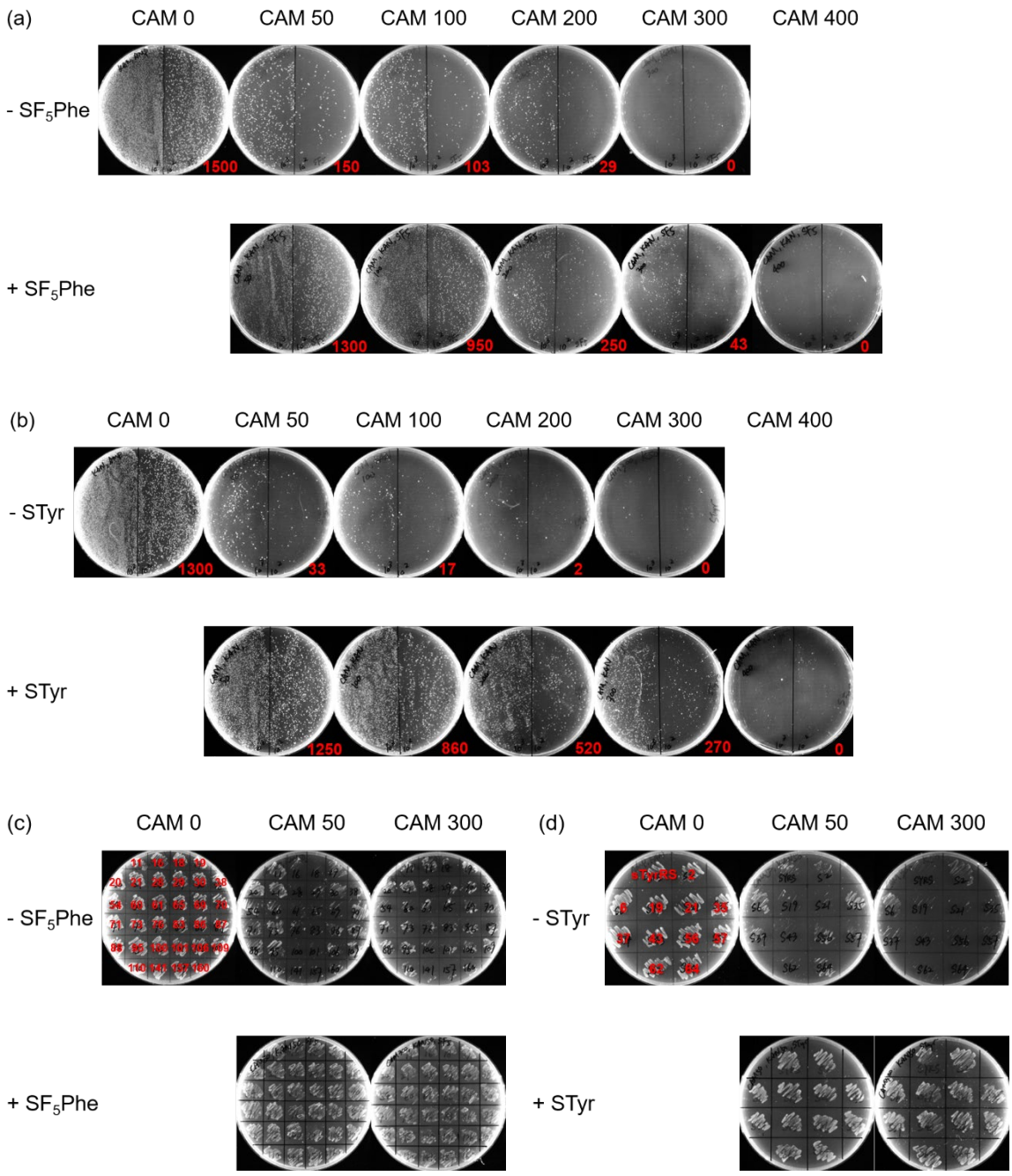


Figure S4. The selection plasmid pBAD-GRFP harbored a chloramphenicol acetyl transferase (CAT) gene with an amber mutation at a permissive site providing resistance of cells to chloramphenicol upon amber codon readthrough.[18] (a) A pool of potential SF_5PheRS candidates selected from 3P+ sample were recovered on LB agar plates with different level of antibiotics (same cell amount on each plate, right half with 10 times dilution of left half). The ratio of colony numbers between +/– SF_5Phe indicates a relatively pure true positive population. (b) A pool of potential STyrRS candidates selected from 5P+ sample were recovered on LB agar plates with different level of antibiotics. With two more selection rounds compared to the SF_5PheRS selection, the true positive population of STyrRS was larger. The sequence confirmed (c) SF_5PheRS (d) STyrRS candidates were individually streaked on plates with different level of antibiotics, showing extremely high incorporation efficiency and specificity.

Characterization of selected SF_5Phe proteins

Mass spectrometry

Mass spectrometry was conducted using an Orbitrap Elite Hybrid Ion Trap-Orbitrap mass spectrometer (Thermo Scientific, USA) coupled with anUltiMate S43000 UHPLC (Thermo Scientific, USA). Approximately 7.5 pmol of sample was injected to the mass analyser via an Agilent ZORBAX SB-C3 Rapid Resolution HT Threaded Column (Agilent, USA).

Expression and purification of H6RFP-SF5Phe

A His_6-amber(TAG)-RFP (H6RFP) reporter gene was placed in the pBAD plasmid under the control of the *araBAD* promoter. pBK plasmids harboring SF_5PheRS candidates were co-transformed with pBAD-H6RFP into *E. coli* DH10B cells. For each mutant, successfully transformed cells were grown in 20 mL LB medium containing 100 μg/mL ampicillin and 50 μg/mL kanamycin at 37ºC overnight, supplied with 0.2% arabinose and 1 mM SF_5Phe. The harvested cells were resuspended in 5 mL of buffer A (50 mM Tris-HCl, pH 7.5, 300 mM NaCl, 10 mM imidazole, 5% glycerol) and lysed by 2 passes through an Avestin Emulsiflex C5 high pressure homogenizer (ATA Scientific). After centrifuging the lysate at 15,000 *g* for 45 min to remove the cell debris, the supernatant was directly loaded onto a 1 mL Ni-NTA column (GE Healthcare, USA). The bound protein was eluted with buffer B (50 mM Tris-HCl, pH 7.5, 300 mM NaCl, 300 mM imidazole, 5% glycerol). The eluted protein was concentrated and washed by ultrafiltration using Amicon Centricon concentrators with a molecular weight cut-off (MWCO) of 10 kDa. Buffer A was used for washing. The incorporation specificity of 10 SF_5PheRS candidates (SF19, SF61, SF87, SF100; SF19, SF29, SF87, SF88, SF95, SF109) was analyzed by mass spectrometry (**Figure S5**).

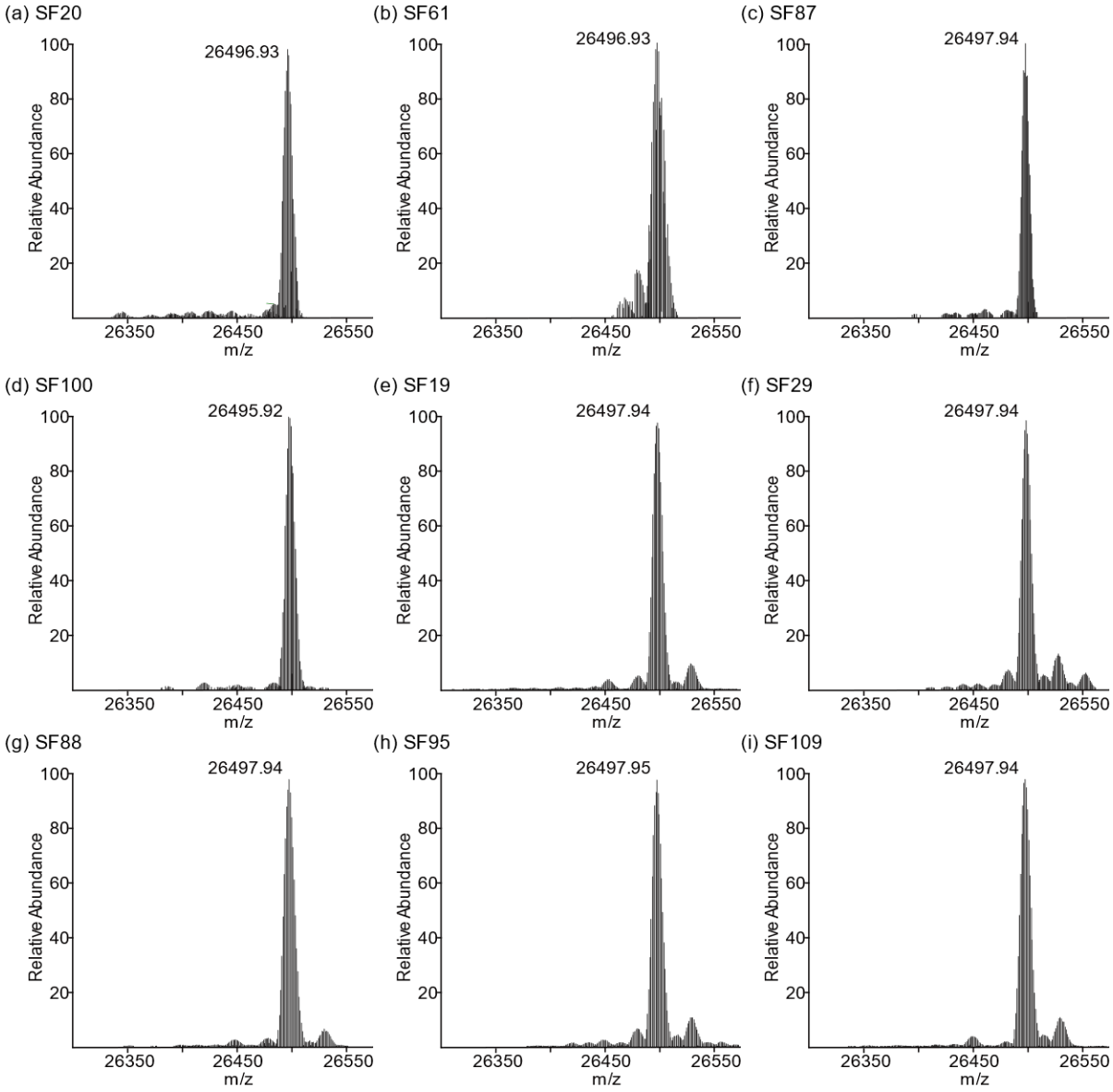


Figure S5. Mass spectrometric analysis for H6RFP-SF_5Phe expression using the selected SF_5PheRS mutants: (a) SF19, (b) SF61, (c) SF87, (d) SF100, (e) SF19, (f) SF29, (g) SF88, (h) SF95, (i) SF109. The calculated mass for H6RFP-SF_5Phe is 26,498.86 Da (mCherryRFP minus 22 Da to form the fluorophore).

Expression and purification of Ubq-cons1-uAA

The expression plasmid pEVOL-SFRS was constructed to contain the SF61 gene and opt-*Mj*YtRNA. It has a chloramphenicol resistance marker and a *p15A* origin of replication, and the extra synthetase expression cassette under *GlnS* promoter was removed.[11] SF61 expression is under control of the *araBAD* promoter and opt-*Mj*YtRNA is controlled by the *proK* promoter. Plasmid pEVOL-CNRS was constructed in the same way but harboring the CNPheRS gene. A fusion protein of ubiquitin was constructed, where the last glycine residue of ubiquitin was mutated to alanine, the cons1 motif was appended with the C-terminal phenylalanine residue available for mutation due to an amber codon (TAG), and the construct was followed by the Mxe-intein with C-terminal His_6-tag.[19] The resulting gene (Ubq-cons1-int; **Figure S6**) was cloned into the *Nde*I and *EcoR*I sites of the ampicillin resistance T7 expression vector pETMSCI[20] to yield the pETMCSI-Ubq-cons1-int plasmid.

pEVOL-SFRS was co-transformed with pETMCSI-Ubq-cons1-int into *E. coli* BL21(DE3) cells and starter culture grown for 16 h at 37ºC in LB medium supplemented with 100 µg/mL ampicillin and 33 µg/mL chloramphenicol. Overnight cultures were inoculated into fresh LB medium (1:100 dilution), supplemented with ampicillin , chloramphenicol and 0.2% arabinose, and grown at 37ºC until the OD_{600} reached 0.6-0.8. Next, isopropyl β-D-1-thiogalactopyranoside (IPTG) was provided together with SF_5Phe or CF_3Phe to reach final concentrations of 1 mM. The cells were harvested following overnight expression at room temperature. Similarly, ClPhe was incorporated using pEVOL-CNRS. After harvesting, the cells were treated as described in the previous section and the protein purified via 1 mL Co-NTA columns (GE Healthcare, USA). The eluted proteins were treated with 200 mM β-mercaptoethanol at room temperature overnight to cleave the intein-His_6 part from the rest of the protein. The proteins in 3 kDa dialysis membranes were dialyzed against 50 mM Tris-HCl, pH 7.5, 300 mM NaCl at 4ºC to ensure removal of all imidazole and β-mercaptoethanol. The samples were loaded onto 1 mL Co-NTA columns equilibrated with buffer A and the flow-through containing the Ubi-cons1 products were concentrated using Amicon ultracentrifugation centrifugal tubes with a MWCO of 3 kDa. The purification process was checked by SDS-PAGE (**Figure S7b**) and the final product was confirmed by mass spectrometry (**Figure S7a**). CF_3Phe was efficiently and specifically incorporated by SF61, indicating that our selection method, with less rounds and lower surviving pressure, allows poly-specificity of the screening outcome.

Figure S6. Construct for Ubq-cons1-uAA expression in pETMCSI-Ubq-cons1-int.

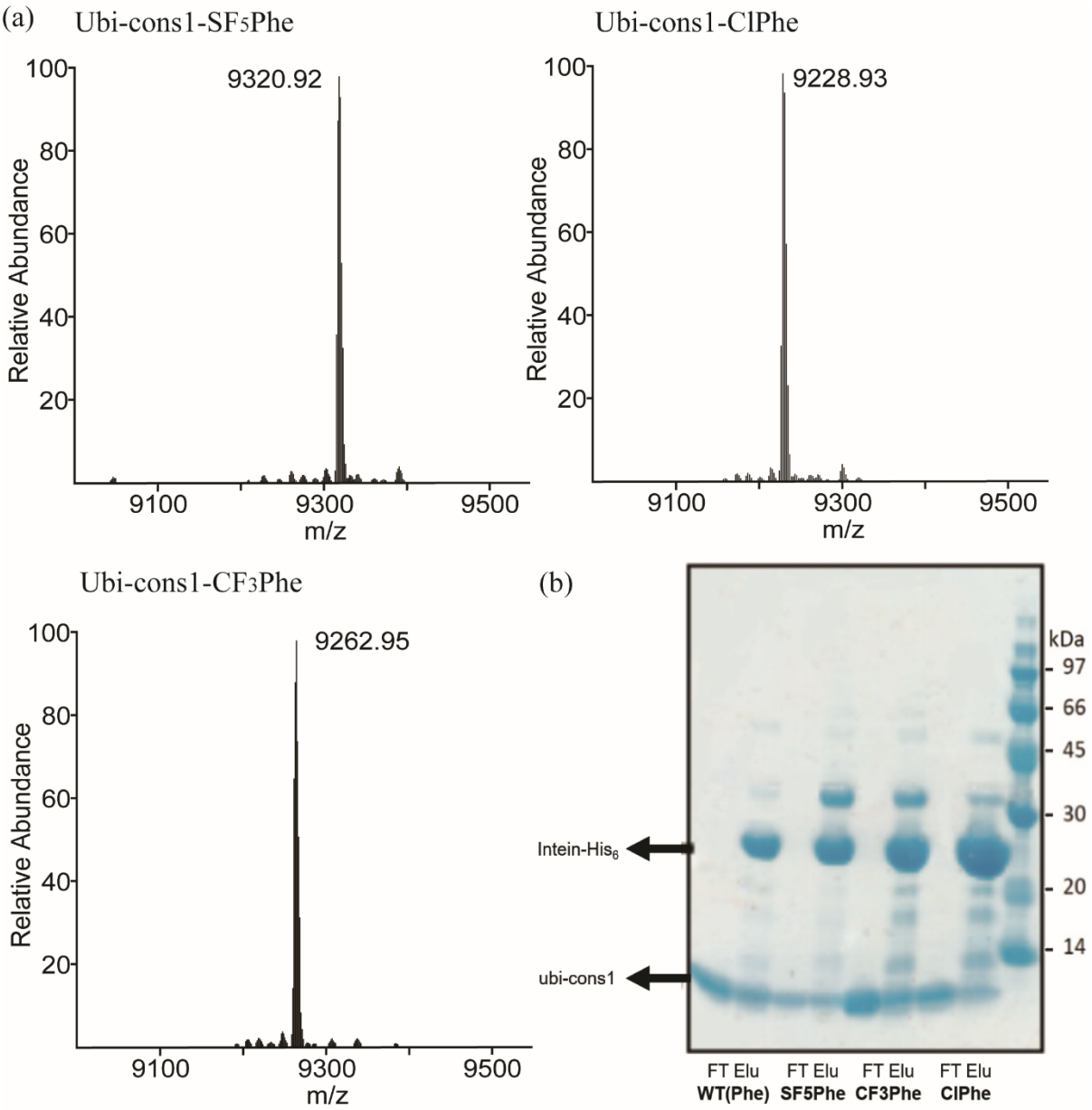
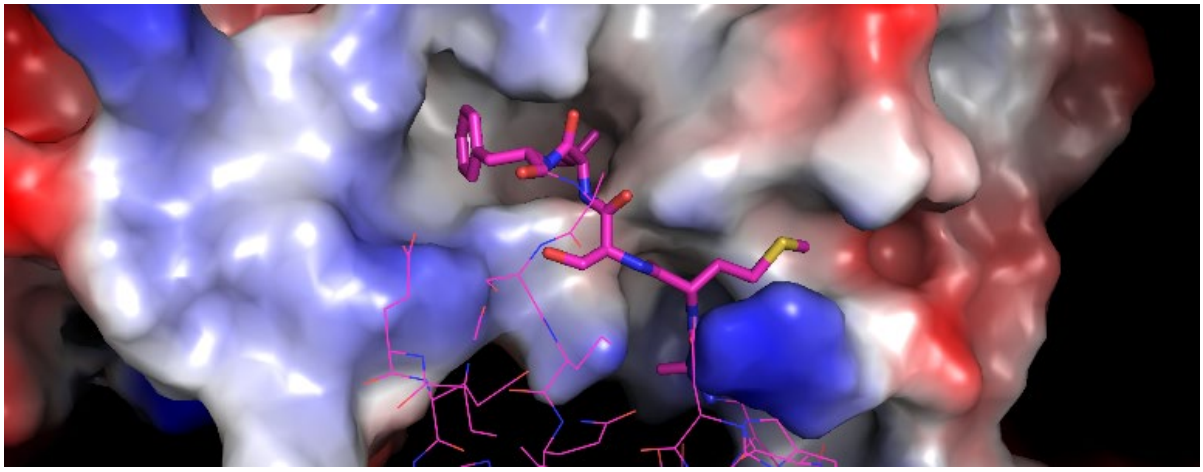


Figure S7. (a) Mass spectrometric analysis of Ubi-cons1-uAA mutants made using SF61 and CNPheRS. The calculated mass for Ubi-cons1-WT(Phe) is 9195.61 Da, 9321.66 Da for Ubi-cons1-SF_5Phe, 9263.61 Da for Ubi-cons1-CF_3Phe, and 9230.05 Da for Ubi-cons1-ClPhe. (b) SDS-PAGE gel illustrating the purity of Ubi-cons1 mutants in the flow-through (FT) fractions. Some of the product stayed in the elution fraction (Elu).

(a)



(b)

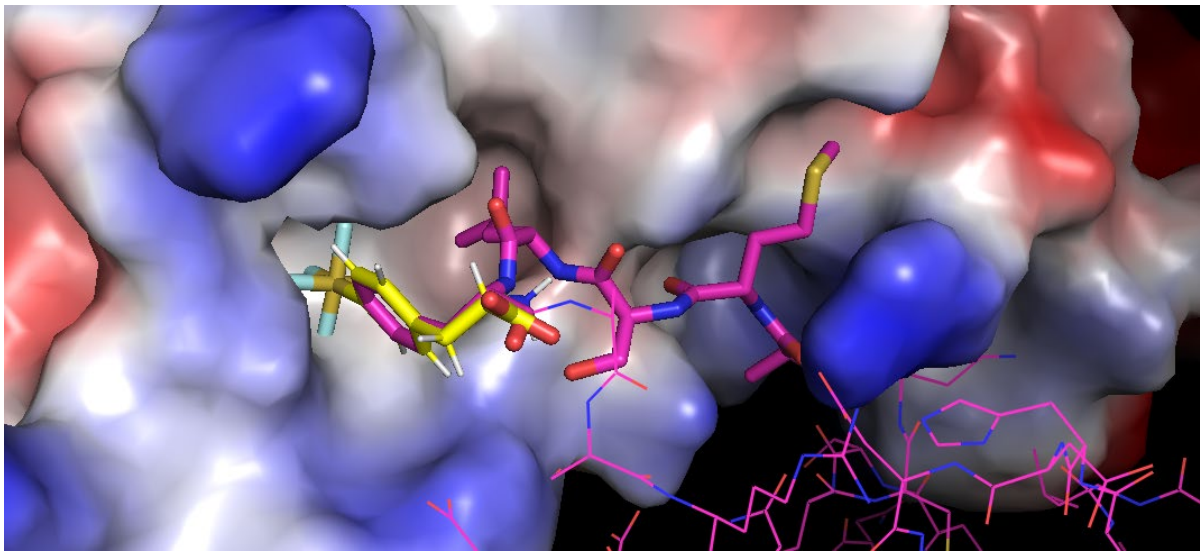


Figure S8. Computationally derived models of the β-clamp dimer with the clamp binding motif highlighted by bold lines. (a) The clamp binding motif of the δ subunit of *E. coli* DNA polymerase III as observed in the crystal structure (PDB ID: 1JQL[21]). (b) Model built with the C-terminal phenylalanine residue replaced by SF_5Phe. In both models, the last residue fits tightly in the deep pocket of the binding site.

Competition fluorescence polarization assays

Fluorescence polarization experiments were carried out essentially as described previously.[22] All experiments were conducted on a POLARstar Omega plate reader (BMG Labtech) using un-treated black sterile 96-well plates (Grenier, USA). The fluorescent tracer used was Ac-QLK-(5FAM)-LF (GL Biochem, China) at 10 nM and the *E. coli* β-clamp protein at 80 nM (monomer concentration). Inhibition of tracer binding to the β-clamp was measured using a two-fold dilution series of competing ligands in buffer over ranges of either 80–0.16 μM or 40–0.078 μM. The positive controls of known CBM peptides, AcQLDLF and AcQLSLPL, were diluted two-fold in DMSO with the final concentration of 4% DMSO used in the assays. The buffer used was 35 mM HEPES, pH 7.4, 1 mM EDTA, 1 mM dithiothreitol, 0.007% (*v*/*v*) nonidet-P40 in a final volume of 200 μL. Experiments were carried out in duplicates for high concentration samples or triplicates for low concentration samples.

Data analysis

Blank control (buffer), negative control (buffer and tracer), and positive control (buffer, tracer, and β-clamp) were used for data normalization. Measured mP values were converted to percent inhibition (%) using Eq. (1) and plotted against competing ligand concentrations to determine IC_{50} values using Eq. (2). Curves were fit using GraphPad Prism.

$$\text{Percent inhibition (\%)} = \frac{Y_{max}-X}{Y_{max}-Y_{min}} \times \frac{100}{1} \quad (1)$$

where Y_{max} is the mP value of the tracer-β complex, × is the mP value following addition of the competing peptide, and Y_{min} is the mP value for the tracer alone.

$$Y = Y_{min} + \frac{Y_{max}-Y_{min}}{1+\left([X]^h/\mathrm{IC}_{50}{}^h\right)} \quad (2)$$

where Y is the % inhibition, Y_{max} is the maximum inhibition, Y_{min} is the minimum inhibition, IC_{50} is the concentration of competing peptide that gives Y = 50%, [X] is the concentration of the competing peptide, and h is the Hill slope.

Expression and purification of Ubq-E18-uAA

The ubiquitin gene with an amber mutant at positions 18 was cloned into the pCDF vector to form pCDF-Ubq-E18. The expression of this amber-interrupted ubiquitin was under the control of a T7 promoter. The plasmid has a spectinomycin-resistance marker and a CloDF13 origin of replication. An expression plasmid pEVOL-CFRS harboring the published *Mj*YRS mutant tfm-PheRS (Tyr32Gln, Leu65Gln, Phe108Gln, His109Glu, Asp158Ser, Leu162Ala), which selectively incorporates CF_3Phe, was constructed as described.[23] *E. coli.* B95 cells were co-transformed with pCDF-Ubq-E18 and pEVOL-SFRS/ pEVOL-CFRS to inoculate 150 mL of fresh LB medium supplemented with spectinomycin, chloramphenicol and grown at 37ºC until the OD600 reached 0.6.[24] Then 0.2% arabinose and 1 mM SF_5Phe or CF_3Phe was provided, and the cell culture was allowed to shake at 37ºC for another 30min. After the culture cooled to room temperature, 1mM IPTG was provided to induce reporter expression. After harvesting, cells were treated as described in the previous section and the protein purified with 1 mL Ni-NTA columns (GE Healthcare, USA). The purification process was checked by SDS-PAGE (**Figure S9b**) and the final product was confirmed by mass spectrometry (**Figure S9a**). Three protein samples were expressed: Ubq-E18-SF_5Phe with SF61 (yield 20.93 mg/L); Ubq-E18-CF_3Phe with SF61 (yield 19.00 mg/L); Ubq-E18-CF_3Phe with tfm-PheRS (yield 17.20 mg/L). The result provide further evidence that SF61 is cross-specific to CF_3Phe and can incorporate both SF_5Phe and CF_3Phe into proteins with high yield.

β-cyclodextrin binding study

Purified Ubq-E18-uAA samples were dialyzed to NMR buffer (50 mM Tris-HCl, pH 7.5, 100 mM NaCl). 0.52 mM of Ubq-E18-SF_5Phe, 0.46 mM of Ubq-E18- CF_3Phe and 1mM free SF_5Phe amino acid were respectively titrated with β-cyclodextrin (**Figure S10**). Chemical shifts of the 4 equatorial fluorines of the SF_5 group and fluorines of CF_3 group were measured at 25ºC on a Bruker 400 MHz AVIII NMR spectrometer equipped with a broadband probe. 1D ^{19}F-NMR spectra were recorded without 1H-decoupling.

The dissociation constant K_d was determined by fitting Eq. (3) to the observed chemical shift changes (**Figure S10a-c**).[25] Δδ is the observed change in chemical shift, $\Delta\delta_{max}$ is the maximal chemical shift difference at saturation, [L] is the concentrations of ligand (β-cyclodextrin) and [P] is the concentrations of protein (Ubq-E18-uAA mutants or free SF_5Phe amino acid).

$$\Delta\delta = \Delta\delta_{max} \times \frac{([L]+[P]+K_d) - (([L]+[P]+K_d)^2 - 4[L][P])^{1/2}}{2[P]} \quad (3)$$

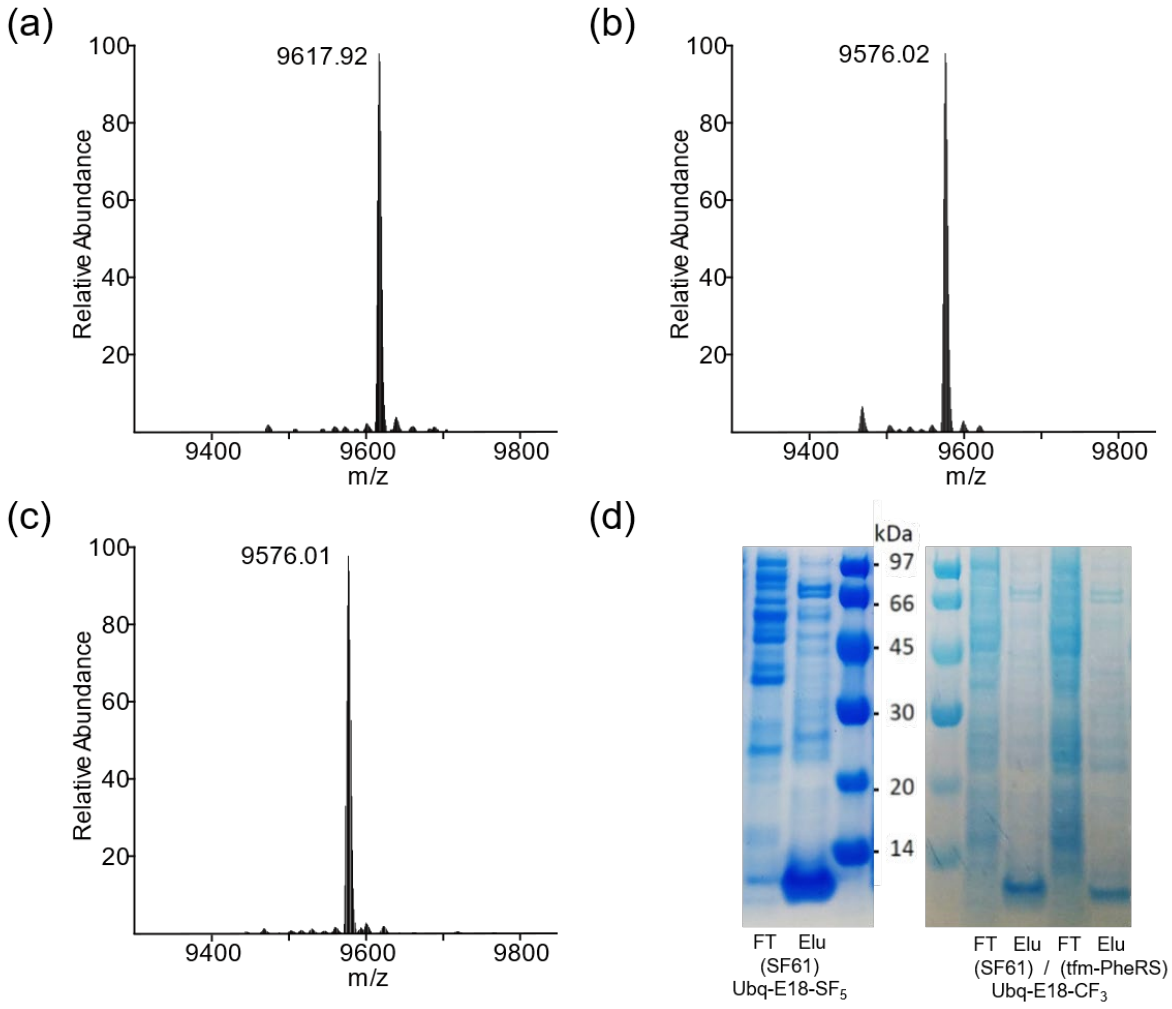


Figure S9. Mass spectrometric analysis of (a) Ubi-E18-SF_5Phe made by SF61 (b) Ubi-E18-CF_3Phe made by SF61 (c) Ubi-E18-CF_3Phe made by tfm-PheRS. The calculated mass for Ubi-E18-SF_5Phe is 9619 Da and for Ubi-E18-CF_3Phe is 9577 Da. (d) SDS-PAGE gel showing high yield and purity of Ubi-E18-uAA mutants.

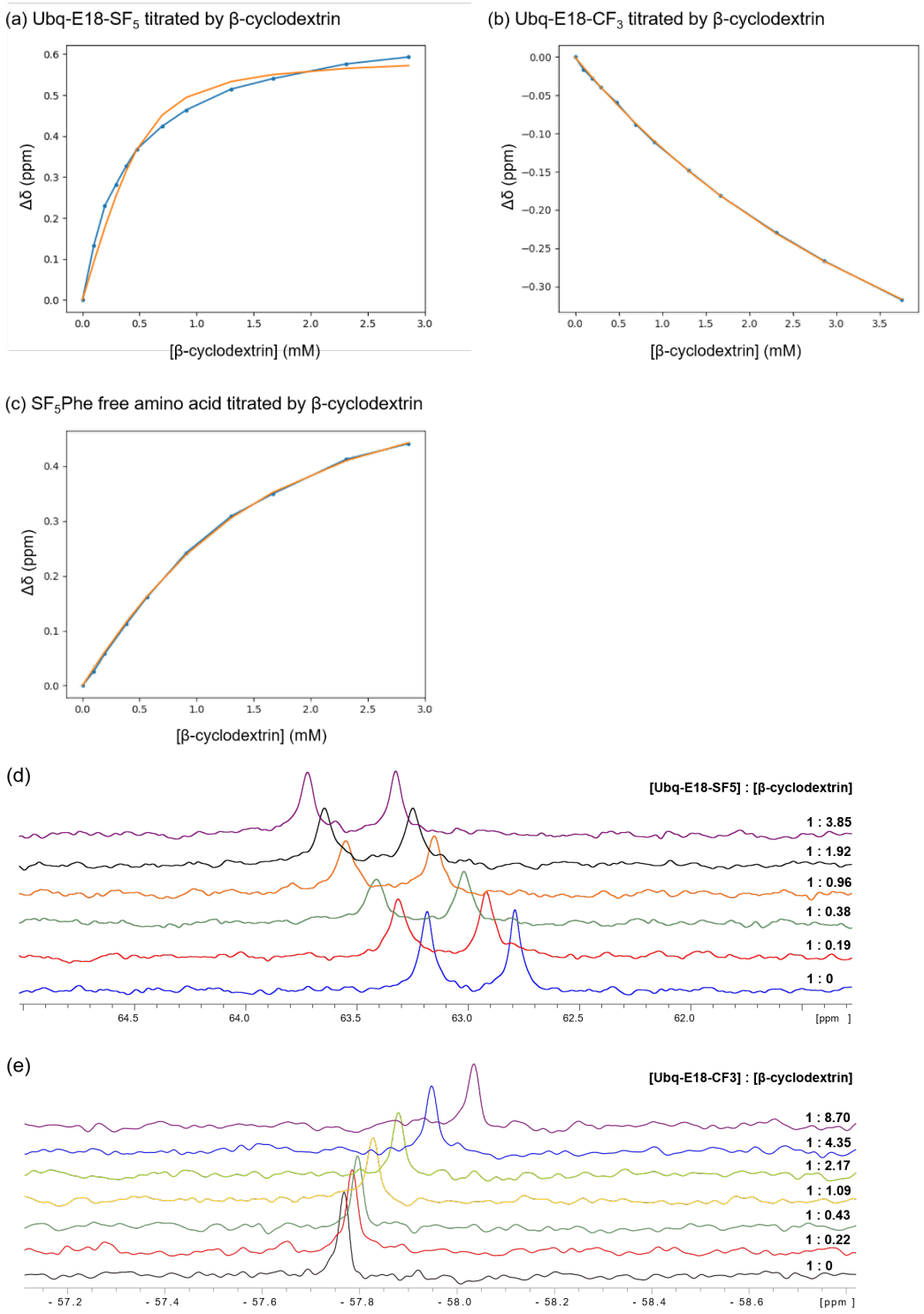


Figure S10. Changes in chemical shift of the observed fluorine spins with response to increasing β-cyclodextrin concentration. The binding curve corresponds to the best fit using Eq. (3). The calculated K_d value of β-cyclodextrin binding (a) towards Ubi-E18-SF_5Phe was 0.11 mM (b) towards Ubi-E18-CF_3Phe was 4.85 mM (c) towards free SF_5Phe amino acid was 0.83 mM. (d) 1D ^{19}F NMR spectra of 0.52 mM of Ubq-E18-SF_5Phe with β-cyclodextrin in NMR buffer. Protein-to-ligand titration ratios are indicated. (e) 1D ^{19}F NMR spectra of 0.46 mM of Ubq-E18- CF_3Phe with β-cyclodextrin.

Expression and purification of PpiB-F27,98-SF_5Phe

A double-amber mutant at positions 27 and 98 of the *ppiB* gene was cloned into the pET3a vector, which features an ampicillin resistance marker and a *ColE1* replication origin. This plasmid pET3a-PpiB was co-transformed with pEVOL-SFRS into *E. coli* BL21(DE3)Star cells (Invitrogen). The recovered culture was used to inoculate 250 mL LB medium containing 100 µg/mL ampicillin and 33 µg/mL chloramphenicol for growth at 37ºC until the OD_{600} reached 0.6. Expression was induced by 1 mM IPTG and 1 mM SF_5Phe was provided. The cell culture was left at room temperature overnight with shaking at 180 rpm. Following harvesting and cell lysis, the protein was purified using a 1 mL Ni-NTA column. The eluted product was confirmed by SDS-PAGE and mass spectrometry (**Figure S12**). The yield of this double-SF_5 labeled PpiB sample was 29.9 mg of purified protein per liter of growth medium. The protein sample was then buffer-exchanged (50 mM Tris-HCl, pH 7.5, 100 mM NaCl) for NMR analysis.

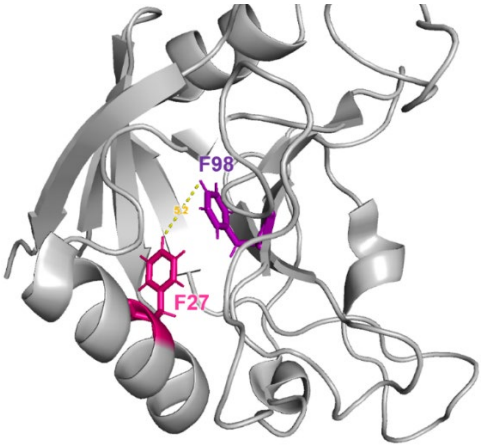


Figure S11. Crystal structure of PpiB (PDB ID: 2RS4[26]). The side chains of Phe27 and Phe98 are located in close proximity in the hydrophobic core of the protein, bringing the H^{ζ} atoms within 6 Å.

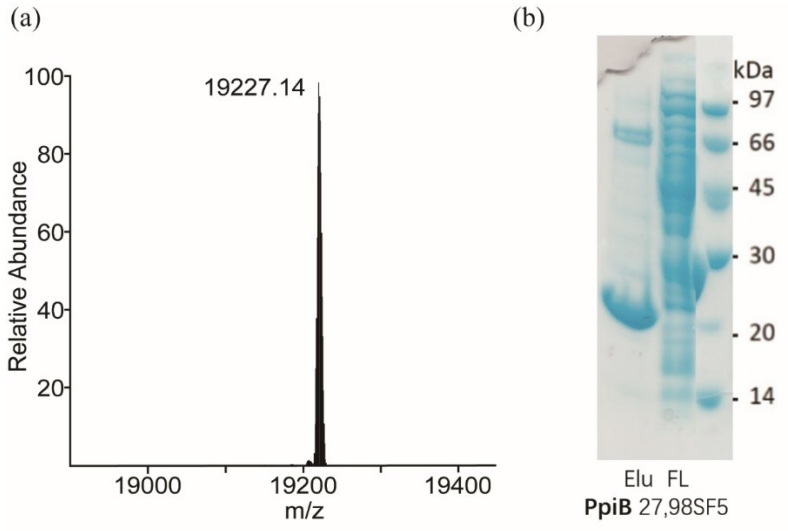


Figure S12. (a) Mass spectrometric analysis of the PpiB-F27,98-SF_5Phe mutant expressed with the help of the selected synthetase SF61. The calculated mass of the protein product is 19228.32 Da. (b) SDS-PAGE gel showing an intense band of the target protein.

NMR measurements

All NMR spectra were recorded at 25°C on a Bruker 400 MHz AVIII NMR spectrometer equipped with a broadband probe. ^{19}F-NMR spectra were recorded without ^{1}H-decoupling. The 1D ^{19}F-NMR spectrum shown underneath was recorded in 1.7 h. The 2D ^{19}F–^{19}F NOESY spectra of Figure 4 in the main text were recorded in 2 h each, using t_1 = 7 ms and t_2 = 67 ms. NMR spectra were calibrated relative to internal trifluoroacetate at -75.25 ppm.

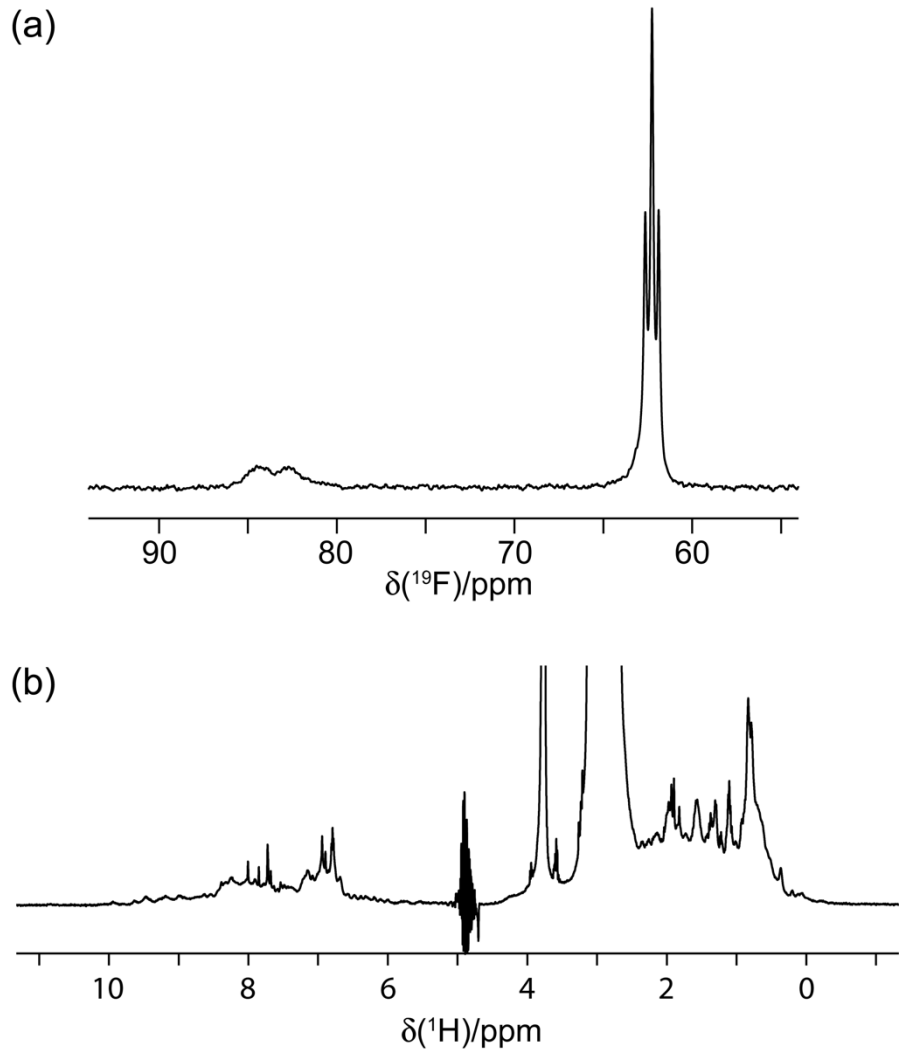


Figure S13. 1D NMR spectra of a 0.56 mM solution of PpiB-27,98-double-SF_5Phe in 90% H_2O/10% D_2O (50 mM Tris-HCl, pH 7.5, 100 mM NaCl). (a) The ^{19}F-NMR spectrum shows that the signals of the two SF_5 groups appear at slightly different chemical shifts, with the four equatorial fluorine atoms of each SF_5 group featuring two overlapping doublets at about 62 ppm, and the axial fluorine atoms displaying two unresolved quintets at about 83 ppm. The spectrum illustrates the much greater peak heights associated with the equatorial than the axial fluorines. (b) 1D ^{1}H-NMR spectrum. The chemical shift dispersion observed confirms that the protein is folded. The truncated signals in the range between 2 and 4 ppm are from components of the buffer.

Deep sequencing analysis of the selected SF_5PheRS family

5 pairs of forward and reverse primers that are complementary upstream and downstream of the mutation region of *Mj*TyrRS library were designed with overhang adapters:

Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-[locus-specific sequence]

Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[locus-specific sequence]

Plasmid extractions of cell pellet recovered from 1P+, 1P–, 2N–, 3P+, 3P– sorted samples were used as the library templates. In-house Vent DNA polymerase was used to amplify the library fragments with 20 cycles. PCR products were purified by AMPure XP beads (Beckman Coulter) and used as templates for Index PCR to attach Illumina sequencing adapters by using Nextera XT individual index primer sets (Illumina). Concentrations of the indexed library fragments were measured via Bioanalyzer DNA 1000 chip (Agilent Technologies). The normalized libraries were combined and submitted for Myseq next-generation sequencing analysis (Illumina). The output data were aligned, and the 1P– recovered sample was treated as the initial library.

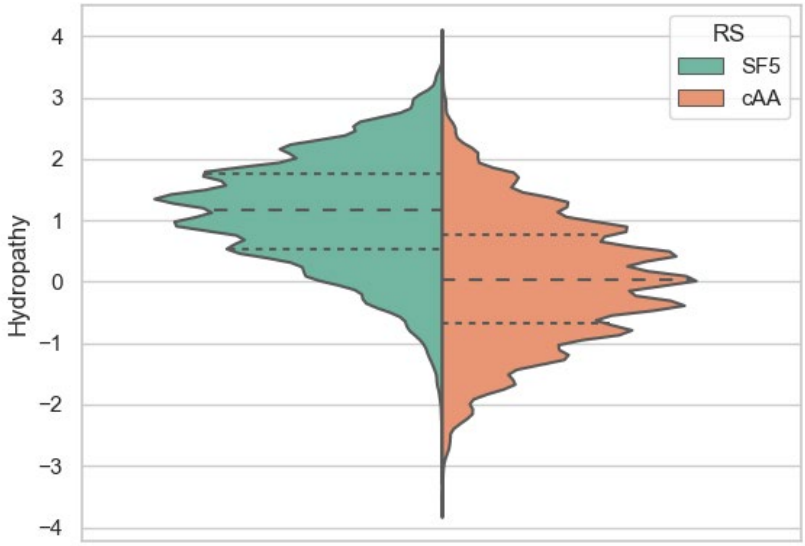


Figure S14. Distribution plots of the average Kyte and Doolittle hydropathy index[27] of the six randomized sites in the library. The dashed lines separate the lower, median, and upper quartiles. Left (green): Functional SF_5Phe tRNA synthetases after the third positive FACS selection round (3P+). A total of 146,816 unique sequences was identified from 429,314 deep sequencing reads. Right (orange): Functional tRNA synthetases with activities towards canonical amino acids obtained after the first positive FACS selection round in the absence of SF_5Phe (1P–). A total of 775,513 unique sequences was identified from 1,851,956 deep sequencing reads. The active sites of the SF_5Phe tRNA synthetases are significantly more hydrophobic.

Table S4. Amino acid sequences of the reporter proteins used in this study.

	Plasmid encoded in	Amino acid sequence
GRFP	pBAD-GRFP	MVSKGEEDNMASLPATHELHIFGSINGVDFDMVGQGTGNPNDGY EELNLKSTKGDLQFSPWILVPHIGYGFHQYLPYPDGMSPFQAAMV DGSGYQVHRTMQFEDGASLTVNYRYTYEGSHIKGEAQVKGTGFP ADGPVMTNSLTAADWCRSKKTYPNDKTIISTFKWSYTTGNGKRY RSTARTTYTFAKPMAANYLKNQPMYVFRKTELKHSKTELNFKEW QKAFTDVMGMDELYKGSAGENLYFQGGEF*ASSEDVIKEFMRFK VRMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDI LSPQFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVV TVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWEASTER MYPEDGALKGEIKMRLKLKDGGHYDAEVKTTYMAKKPVQLPGA YKTDIKLDITSHNEDYTIVEQYERAEGRHSTGAHHHHHH*
H6RFP	pBAD-H6RFP	MHHHHHH*ASSEDVIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPY EGTQTAKLKVTKGGPLPFAWDILSPQFQYGSKAYVKHPADIPDYL KLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTN FPSDGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGG HYDAEVKTTYMAKKPVQLPGAYKTDIKLDITSHNEDYTIVEQYER AEGRHSTGA*
Ubq-cons1-int	pETMSCI-Ubq-cons1-int	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAG KQLEDGRTLSDYNIQKESTLHLVLRLRGAQLDL*CITGDALVALPE GESVRIADIVPGARPNSDNAIDLKVLDRHGNPVLADRLFHSGEHPV YTVRTVEGLRVTGTANHPLLCLVDVAGVPTLLWKLIDEIKPGDYA VIQRSAFSVDCAGFARGKPEFAPTTYTVGVPGLVRFLEAHHRDPD AQAIADELTDGRFYYAKVASVTDAGVQPVYSLRVDTADHAFITNG FVSHATGGGGSHHHHHH*
Ubq-E18-amber	pCDF-UbqE18	MQIFVKTLTGKTITLEV*PSDTIENVKAKIQDKEGIPPDQQRLIFAGK QLEDGRTLSDYNIQKESTLHLVLRLRGGSHHHHHH*
PpiB-F27,98-amber	pET3a-PpiB	MVTFHTNHGDIVIKTFDDKAPETVKN*LDYCREGFYNNTIFHRVIN GFMIQGGGFEPGMKQKATKEPIKNEANNGLKNTRGTLAMARTQA PHSATAQ*FINVVDNDFLNFSGESLQGWGYCVFAEVVDGMDVVD KIKGVATGRSGMHQDVPKEDVIIESVTVSEHHHHHH*
CNPheRS	pBK-CNRS/ pEVOL-CNRS	MDEFEMIKRNTSEIISEEELREVLKKDEKSALIGFEPSGKIHLGHYL QIKKMIDLQNAGFDIIIVLADLHAYLNQKGELDEIRKIGDYNKKVF EAMGLKAKYVYGSEWMLDKDYTLNVYRLALKTTLKRARRSMEL IAREDENPKVAEVIYPIMQVNGAHYVGVDVAVGGMEQRKIHMLA RELLPKKVVCIHNPVLTGLDGEGKMSSSKGNFIAVDDSPEEIRAKI KKAYCPAGVVEGNPIMEIAKYFLEYPLTIKRPEKFGGDLTVNSYEE LESLFKNKELHPMRLKNAVAEELIKILEPIRKRL*
SF61	pBK-MjYRS/ pEVOL-SFRS	MDEFEMIKRNTSEIISEEELREVLKKDEKSAVIGFEPSGKIHLGHYL QIKKMIDLQNAGFDIIIVLADLHAYLNQKGELDEIRKIGDYNKKVF EAMGLKAKYVYGSEWALDKDYTLNVYRLALKTTLKRARRSMELI AREDENPKVAEVIYPIMQVNGVHYKGVDVAVGGMEQRKIHMLA RELLPKKVVCIHNPVLTGLDGEGKMSSSKGNFIAVDDSPEEIRAKI KKAYCPAGVVEGNPIMEIAKYFLEYPLTIKRPEKFGGDLTVNSYEE LESLFKNKELHPMRLKNAVAEELIKILEPIRKRL*

References

1. Aakeröy, C. B.; Wijethunga, T. K.; Desper, J.; Đaković, M., Crystal engineering with iodoethynylnitrobenzenes: a group of highly effective halogen-bond donors. *Crystal Growth & Design* **2015,** *15* (8), 3853-3861.
2. Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M., Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. *The Journal of organic chemistry* **2003,** *68* (19), 7368-7373.
3. Leroux, F., Atropisomerism, biphenyls, and fluorine: a comparison of rotational barriers and twist angles. *ChemBioChem* **2004,** *5* (5), 644-649.
4. Sitzmann, M., 1, 3, 4-Oxadiazoles with SF5-containing substituents. *Journal of Fluorine Chemistry* **1995,** *70* (1), 31-38.
5. Bergmann, D.; Hinze, J., Electronegativity and charge distribution. In *Electronegativity*, Springer: 1987; pp 145-190.
6. Sæthre, L. J.; Berrah, N.; Bozek, J. D.; Børve, K. J.; Carroll, T. X.; Kukk, E.; Gard, G. L.; Winter, R.; Thomas, T. D., Chemical insights from high-resolution X-ray photoelectron spectroscopy and ab initio theory: Propyne, trifluoropropyne, and ethynylsulfur pentafluoride. *Journal of the American Chemical Society* **2001,** *123* (43), 10729-10737.
7. Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M., The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *Journal of the American Chemical Society* **1963,** *85* (18), 2817-2824.
8. Sheppard, W. A., Alkyl-and arylsulfur trifluorides. *Journal of the American Chemical Society* **1962,** *84* (16), 3058-3063.
9. Young, D. D.; Young, T. S.; Jahnz, M.; Ahmad, I.; Spraggon, G.; Schultz, P. G., An evolved aminoacyl-tRNA synthetase with atypical polysubstrate specificity. *Biochemistry* **2011,** *50* (11), 1894-1900.
10. Xie, J.; Liu, W.; Schultz, P. G., A genetically encoded bidentate, metal‐binding amino acid. *Angewandte Chemie* **2007,** *119* (48), 9399-9402.
11. Young, T. S.; Ahmad, I.; Yin, J. A.; Schultz, P. G., An enhanced system for unnatural amino acid mutagenesis in E. coli. *Journal of Molecular Biology* **2010,** *395* (2), 361-374.
12. Chatterjee, A.; Sun, S. B.; Furman, J. L.; Xiao, H.; Schultz, P. G., A versatile platform for single- and multiple-unnatural amino acid mutagenesis in Escherichia coli. *Biochemistry* **2013,** *52* (10), 1828-1837.
13. Shaner, N. C.; Lambert, G. G.; Chammas, A.; Ni, Y.; Cranfill, P. J.; Baird, M. A.; Sell, B. R.; Allen, J. R.; Day, R. N.; Israelsson, M., A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nature methods* **2013,** *10* (5), 407-409.
14. Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N.; Palmer, A. E.; Tsien, R. Y., Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature biotechnology* **2004,** *22* (12), 1567-1572.
15. Liu, C. C.; Schultz, P. G., Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nature Biotechnology* **2006,** *24* (11), 1436-1440.
16. Lin, A. E.; Lin, Q., Rapid Identification of Functional Pyrrolysyl-tRNA Synthetases via Fluorescence-Activated Cell Sorting. *International Journal of Molecular Sciences* **2019,** *20* (1), 29.
17. Hohl, A.; Karan, R.; Akal, A.; Renn, D.; Liu, X.; Ghorpade, S.; Groll, M.; Rueping, M.; Eppinger, J., Engineering a Polyspecific pyrrolysyl-tRnA Synthetase by a High throughput fAcS Screen. *Scientific reports* **2019,** *9* (1), 1-9.
18. Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G., Expanding the genetic code of Escherichia coli. *Science* **2001,** *292* (5516), 498-500.
19. Batjargal, S.; Walters, C. R.; Petersson, E. J., Inteins as traceless purification tags for unnatural amino acid proteins. *Journal of the American chemical society* **2015,** *137* (5), 1734-1737.

20. Neylon, C.; Brown, S. E.; Kralicek, A. V.; Miles, C. S.; Love, C. A.; Dixon, N. E., Interaction of the Escherichia coli replication terminator protein (Tus) with DNA: a model derived from DNA-binding studies of mutant proteins by surface plasmon resonance. *Biochemistry* **2000,** *39* (39), 11989-11999.
21. Jeruzalmi, D.; Yurieva, O.; Zhao, Y.; Young, M.; Stewart, J.; Hingorani, M.; O'Donnell, M.; Kuriyan, J., Mechanism of processivity clamp opening by the delta subunit wrench of the clamp loader complex of E. coli DNA polymerase III. *Cell* **2001,** *106* (4), 417-428.
22. McGrath, A. E.; Martyn, A. P.; Whittell, L. R.; Dawes, F. E.; Beck, J. L.; Dixon, N. E.; Kelso, M. J.; Oakley, A. J., Crystal structures and biochemical characterization of DNA sliding clamps from three Gram-negative bacterial pathogens. *Journal of structural biology* **2018,** *204* (3), 396-405.
23. Jackson, J. C.; Hammill, J. T.; Mehl, R. A., Site-specific incorporation of a 19F-amino acid into proteins as an NMR probe for characterizing protein structure and reactivity. *Journal of the American Chemical Society* **2007,** *129* (5), 1160-1166.
24. Mukai, T.; Hoshi, H.; Ohtake, K.; Takahashi, M.; Yamaguchi, A.; Hayashi, A.; Yokoyama, S.; Sakamoto, K., Highly reproductive Escherichia coli cells with no specific assignment to the UAG codon. *Scientific reports* **2015,** *5* (1), 1-9.
25. Williamson, M. P., Using chemical shift perturbation to characterise ligand binding. *Progress in nuclear magnetic resonance spectroscopy* **2013,** *73*, 1-16.
26. Takeda, M.; Jee, J.; Ono, A. M.; Terauchi, T.; Kainosho, M., Hydrogen exchange study on the hydroxyl groups of serine and threonine residues in proteins and structure refinement using NOE restraints with polar side-chain groups. *J Am Chem Soc* **2011,** *133* (43), 17420-7.
27. Kyte, J.; Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **1982,** *157* (1), 105-132.