

Discovery of a 29-amino-acid reactive abiotic peptide for selective cysteine arylation

Authors: Ethan D. Evans, Bradley L. Pentelute*

Affiliations: Department of Chemistry, Massachusetts Institute of Technology, 77
Massachusetts Avenue, Cambridge, MA, 02139

*correspondence to:
Bradley L. Pentelute
Email: blp@mit.edu
Tel. (+1) 617 324 0180

Table of Contents

1. Materials
2. Methods
 - 2.1. Liquid chromatography-mass spectrometry
 - 2.2. Peptide synthesis, cleavage and purification
 - 2.3. Capture agent synthesis
 - 2.4. mRNA display
 - 2.5. qPCR analysis
 - 2.6. HTS analysis
 - 2.7. Kinetics and LCMS analysis
 - 2.8. Protein expression and purification
 - 2.9. Labeling site determination
3. References
4. Experiment data, tables and figures
 - 4.1. Sequence and mass of peptides used
 - 4.2. HTS Levenshtein clusters, library heatmaps, FCPF analysis and statistics
 - 4.3. Kinetics chromatograms and plots
 - 4.4. Determination of labeling location
 - 4.5. Analysis of truncated MP01
 - 4.6. Analysis of MP01 with urea
 - 4.7. MP-SrtA labeling, TEV cleavage and SrtA reactions

1. Materials

Chemicals and enzymes

Pentafluorophenyl sulfide was purchased from Santa Cruz Biotechnology (Dallas, TX). 1,4-Dithio-DL-threitol (DTT), 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU), Fmoc-L-Ala-OH, Fmoc-L-Cys(trt)-OH, Fmoc-L-Asp(tBu)-OH, Fmoc-L-Glu(tBu)-OH, Fmoc-L-Phe-OH, Fmoc-L-Gly-OH, Fmoc-L-His(Boc)-OH, Fmoc-L-Ile-OH, Fmoc-L-Lys(Boc)-OH, Fmoc-L-Leu-OH, Fmoc-L-Met-OH, Fmoc-L-Asn(Trt)-OH, Fmoc-L-Pro-OH, Fmoc-L-Gln(Trt)-OH, Fmoc-L-Arg(Pbf)-OH, Fmoc-L-Ser(tBu)-OH, Fmoc-L-Thr(tBu)-OH, Fmoc-L-Val-OH, Fmoc-L-Trp(Boc)-OH, Fmoc-L-Tyr(tBu)-OH, Fmoc-L-Lys(biotin)-OH, 2-chlorotriyl chloride resin were purchased from Chem-Impex International (Wood Dale, IL). H-rink-amide chemmatrix Hyr resin was obtained from PCAS BioMatrix, Inc (Quebec, Canada). (7-Azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (PyAOP) was purchased from P3 BioSystems (Louisville, KY). Tris(2-carboxyethyl)phosphine hydrochloride was purchased from Hampton Research (Aliso Viejo, CA). Carboxy-PEG₁₂-thiol was purchased from Thermo Fisher Scientific. SuperScript II reverse transcriptase and RNase OUT were purchased from Invitrogen (Carlsbad, CA), while Taq polymerase and T7 RNA polymerase (and their associated buffers) were obtained from New England Biolabs (Ipswich, MA). Flexi rabbit reticulocyte lysate along with rNTPs and dNTPs were purchased from Promega (Madison, WI). *N, N*-dimethylformamide (DMF), acetonitrile (ACN), diethyl ether were purchased from VWR (Radnor, PA). Trifluoroacetic acid (TFA) was obtained from Sigma-Aldrich. Other chemicals listed were purchased from either Sigma-Aldrich or VWR and used as received.

2. Methods

2.1. Liquid chromatography-mass spectrometry (LCMS)

For the remainder of this manuscript, solvent A will refer to water with 0.1% (v/v) TFA, while B will be acetonitrile with 0.1% (v/v) TFA. TIC refers to total ion current in the LCMS chromatogram. The majority of LCMS chromatograms and mass spectra were obtained using an Agilent 6520 ESI-Q-TOF mass spectrometer using method 1 unless otherwise noted (MS/MS analysis was conducted on an Agilent 6550 iFunnel Q-TOF mass spectrometer). Software used for LCMS analysis was the Agilent MassHunter package and deconvolution was performed using maximum entropy.

Method 1:

LC method: 0-2 minutes 5% B, 2-11 minutes 5-65% B linear ramp, 11-12 minutes 65% B, 0.8mL/min flow rate.

Column: Zorbax 300SB C₃ column (2.1 x 150mm, 5 μ m), 40°C

MS parameters: positive electrospray ionization (ESI).

Method 2:

LC method: 0-3 minutes 5% B, 3-17 minutes 5-95% B linear ramp, 17-18 minutes 95% B, 0.8mL/min flow rate.

Column: Zorbax 300SB C₁₈ column (2.1 x 150mm, 5 μ m), 40°C

MS parameters: positive ESI

Method 3:

LC method: 0-3 minutes 5% B, 3-15 minutes 5-80% B linear ramp, 15-16 minutes 80% B, 0.8mL/min flow rate.

Column: Zorbax 300SB C₁₈ column (2.1 x 150mm, 5μm), 40°C

MS parameters: positive ESI, MS off at 11 minutes

2.2. Peptide synthesis, cleavage and purification

Peptides were synthesized using an automated flow peptide synthesizer built in house¹ on a 0.09mmol scale using Fmoc-SPPS chemistry on H-rink amide chemmatrix Hyr resin. General synthesis was performed at 90°C using the following protocol with a 80mL/min flow rate: 15s amino acid coupling (0.14M HATU, 0.2M amino acid, 10% (v/v) *N,N*-diisopropylethylamine (DIEA), 4.8mL total), 38s 12mL DMF wash, 34s 11.2mL 20% piperidine (v/v) in DMF deprotection with a final 38s 12.8mL DMF wash. Deviations from this protocol included: Arg and Phe couplings that were completed using PyAOP instead of HATU and the HHHHHHRL sequence found on MP01-full that was synthesized using a 40mL/min coupling at 70°C with all other parameters the same. Following synthesis, peptides were cleaved from the resin and side-chain deprotected using a mixture of 94% TFA, 2.5% (v/v) 1,2-ethanedithiol (EDT), 2.5% (v/v) water and 1% (v/v) triisopropylsilane for 7 minutes at 60°C. Peptides were there triturated three times using cold diethyl ether. The resulting precipitate was then dissolved in 50% A: 50% B and lyophilized.

Crude peptides were then dissolved in the minimal amount of 95% A: 5% B and purified by reverse phase (RP) HPLC using an Agilent Zorbax C₃ column (21.2 x 250 mm, 7μm) using a linear gradient from 95% A: 5% B to 55% A: 45% B over 120 minutes at a flow rate of 7mL/min. Fractions were analyzed for purify by RP-LCMS using method 1.

2.3. Capture Agent Synthesis

Conjugation of carboxy-PEG-thiol to pentafluorophenyl sulfide

A solution consisting of 5mM carboxy-(PEG)₁₂-thiol, 500mM pentafluorophenyl sulfide, 20mM triphenylphosphine and 230mM DIEA in acetonitrile was vortexed and left at room temperature for 4 hours. The reaction was then diluted with 10.6x volume of 95% A: 5% B, solid phase extracted and lyophilized. The resulting material was analyzed by LCMS (Figure S1, method 2). This product will later be referred to as mCA (modified CA)

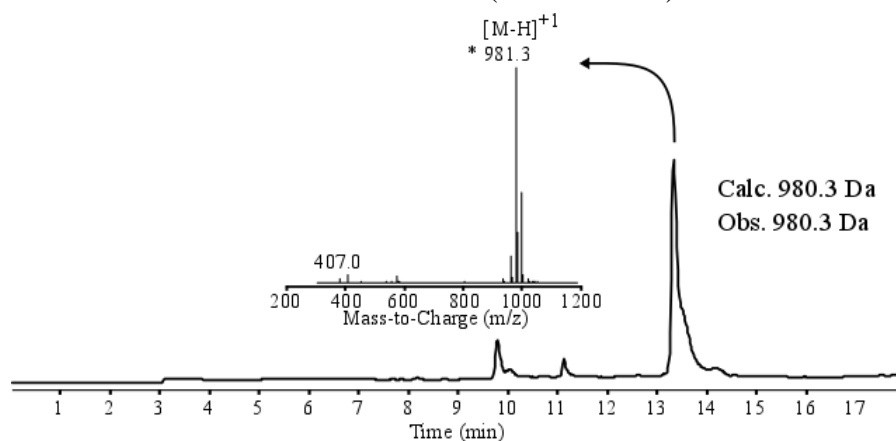


Figure S1. LCMS trace of the 4 hour reaction mixture, the peak at ~13.5m corresponds to the reaction product.

Conjugation of CT-Peg-pfp sulf to Lys(bio)-2-chloro trityl chloride resin

17.7mg of 2-chlorotrityl chloride resin (0.6-0.7mmol/g) was reacted with 51.4mg Fmoc-L-Lys(biotin)-COOH in 1mL DMF with 71 μ l DIEA. The solution was sparged with argon and left overnight. The resin was washed with DMF, DCM and dried. The Fmoc group was removed with 500 μ l of a 20% piperidine in DMF solution for 30 minutes at room temperature followed by DMF washes. 20.43mg of pentafluorophenyl sulfide-PEG-COOH was coupled to the 32.3mg of dried lysine attached resin with 490 μ l DMF, 0.4M HATU and 98 μ l DIEA. This was left for 2 hours at room temperature and then washed and dried *in vacuo*. The capture agent was cleaved from the resin with a two hour, room temperature treatment of 95% TFA, 2.5% water, 2.5% TIPS; the cleavage cocktail was evaporated and then 4mL of 50% A:50% B was added and the resulting solution lyophilized. Crude mass obtained = 17.5mg (Figure S2, method 2).

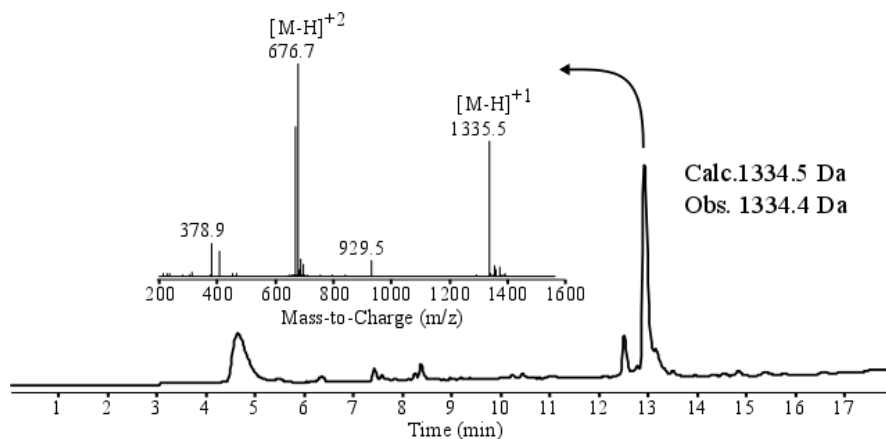


Figure S2. LCMS trace of the resin cleaved material, the peak at ~13m corresponds to the desired reaction product.

The capture agent was purified on a C3 with the following method: 10 minutes at 95% A:5% B, 30 minutes of a 1% B increase per minutes to 65% A:35% B, followed by a 150 min, 0.25% B per minute gradient to 35% A :65% B. Fractions were analyzed by LCMS, pure fractions

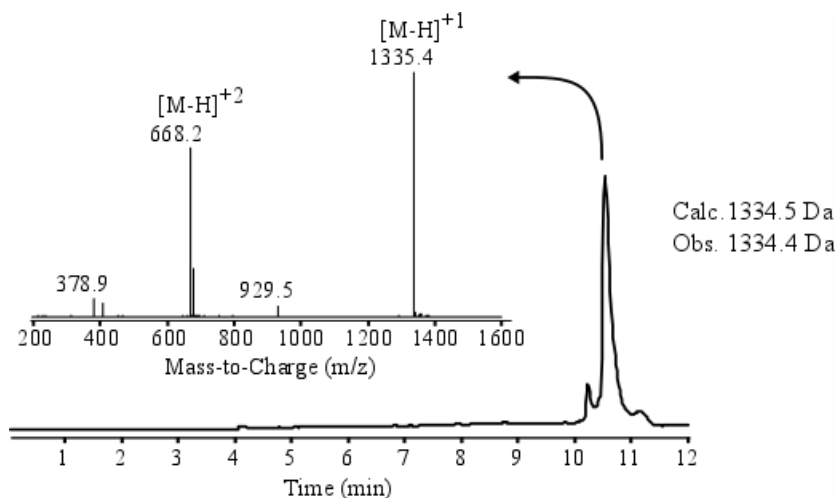


Figure S3. LCMS trace of the HPLC purified capture agent.

were combined, lyophilized and the final material was characterized by LCMS (Figure S3, method 1).

Major impurity:

Seen in figure S3 is a small peak to the left of the primary peak, this peak is believed to be an oxidized capture agent (oxidation in the biotin moiety) as its mass is 16 Da greater than that for the CA (see LCMS/MS analysis below). This impurity peak is seen in kinetics time courses and elutes immediately prior to the CA.

Capture Agent NMR

A resynthesized stock of the CA was diluted in DMSO-d₆ and analyzed by NMR using either 282 or 500MHz.

¹⁹F NMR (282 MHz, DMSO-d₆) δ -135.46 (d, *J* = 24.4 Hz), -135.97 (dd, *J* = 26.2, 10.7 Hz), -136.67 (dd, *J* = 27.3, 9.8 Hz), -153.35 (t, *J* = 22.5 Hz), -163.58 (t, *J* = 22.8 Hz).

¹³C NMR (150MHz, DMSO) δ 173.73, 171.87, 170.17, 162.76, 158.13 (d, *J* = 31.4 Hz), 147.90, 147.20 (d, *J* = 14.9 Hz), 146.28, 145.62, 145.52, 142.89, 141.18, 137.33 (d, *J* = 247.0 Hz), 117.23 (t, *J* = 20.6 Hz), 111.27 – 108.75 (m), 107.20 – 104.61 (m), 99.56, 69.94, 69.81, 69.75, 69.72, 69.67, 69.60, 69.53, 66.77, 61.08, 59.25, 55.45, 51.71, 39.87, 38.20, 35.82, 35.24, 33.80, 30.87, 28.81, 28.26, 28.07, 25.35, 22.87.

¹H NMR (500 MHz, DMSO-d₆) δ 12.54 (s, 1H), 8.09 (d, *J* = 7.8 Hz, 1H), 7.76 (t, *J* = 5.6 Hz, 1H), 6.43 (s, 1H), 6.37 (s, 1H), 4.30 (dd, *J* = 7.8, 4.9 Hz, 1H), 4.14 (ddd, *J* = 12.1, 7.9, 4.9 Hz, 2H), 2.53 – 2.47 (m, 1H), 3.65 – 3.32 (m, 52H), 3.16 (t, *J* = 5.9 Hz, 2H), 3.13 – 3.05 (m, 1H), 3.05 – 2.93 (m, 2H), 2.82 (dd, *J* = 12.4, 5.0 Hz, 1H), 2.57 (d, *J* = 12.4 Hz, 1H), 2.46 – 2.27 (m, 2H), 2.04 (t, *J* = 7.4 Hz, 2H), 1.78 – 1.13 (m, 11H).

* For the peaks at 3.65-3.32 and 1.78-1.13 we believe there to be water and grease contaminants respectively, thus throwing off the integrated value of protons.

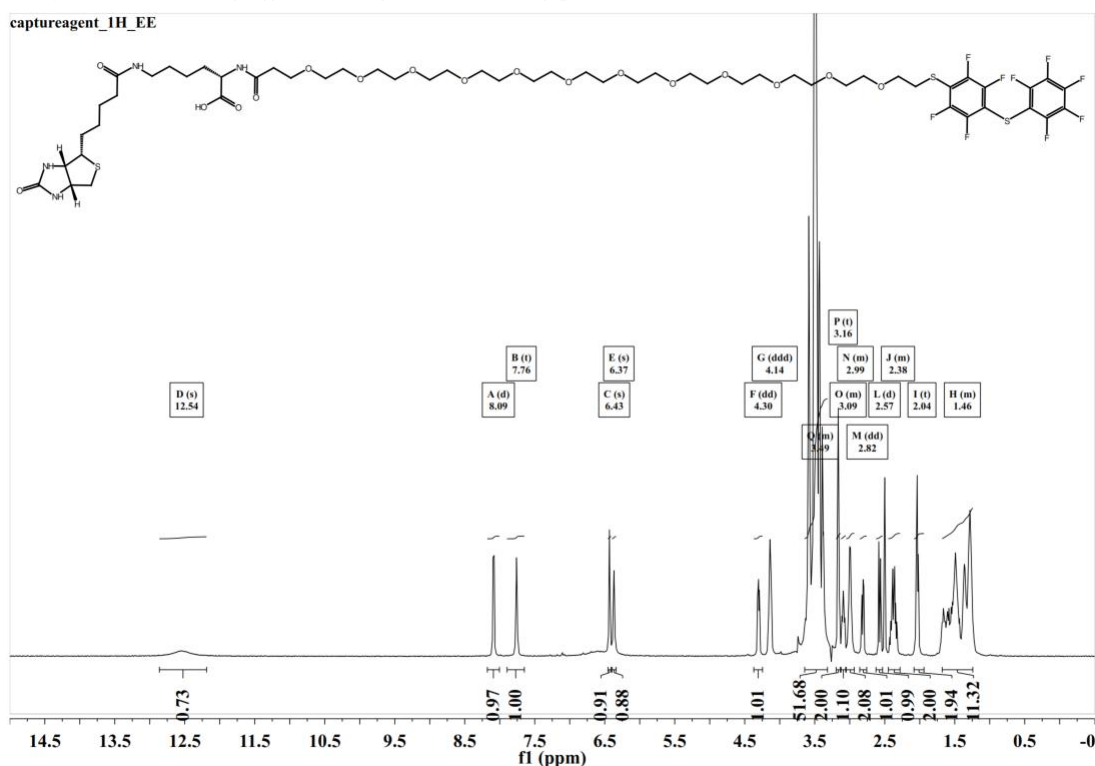


Figure S4. ¹H NMR spectrum for the selection CA in DMSO-d₆.

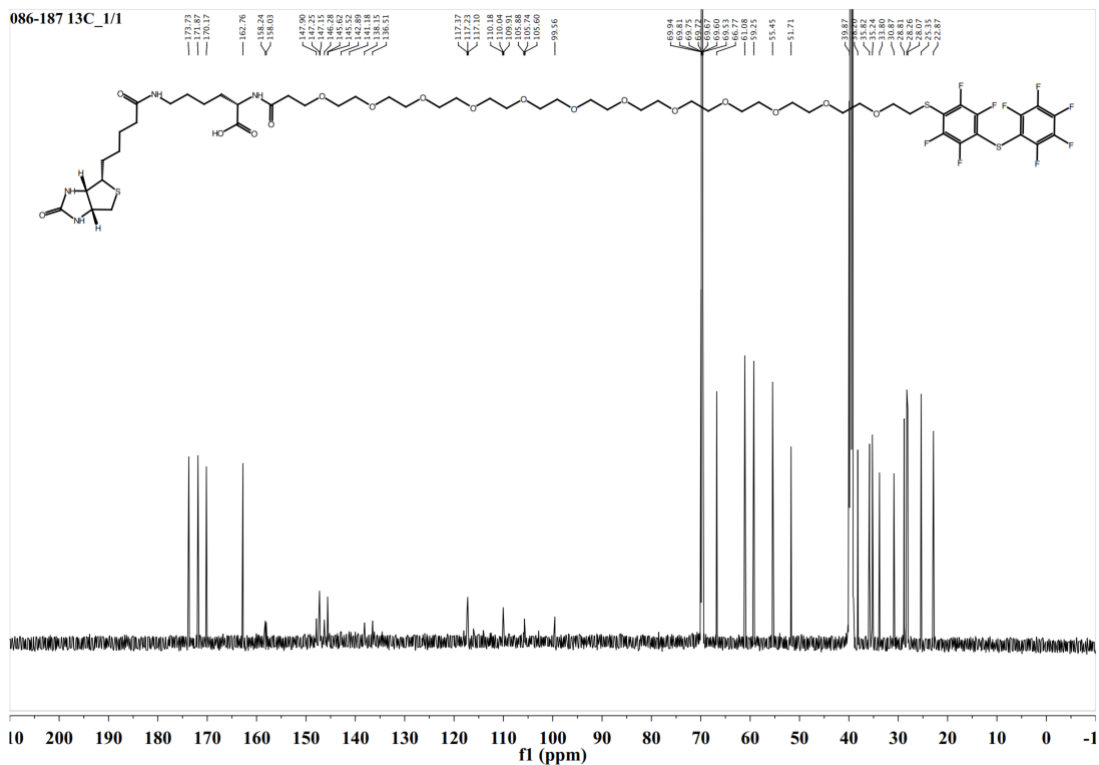


Figure S5. ¹³C NMR spectrum for the selection CA in DMSO-d6

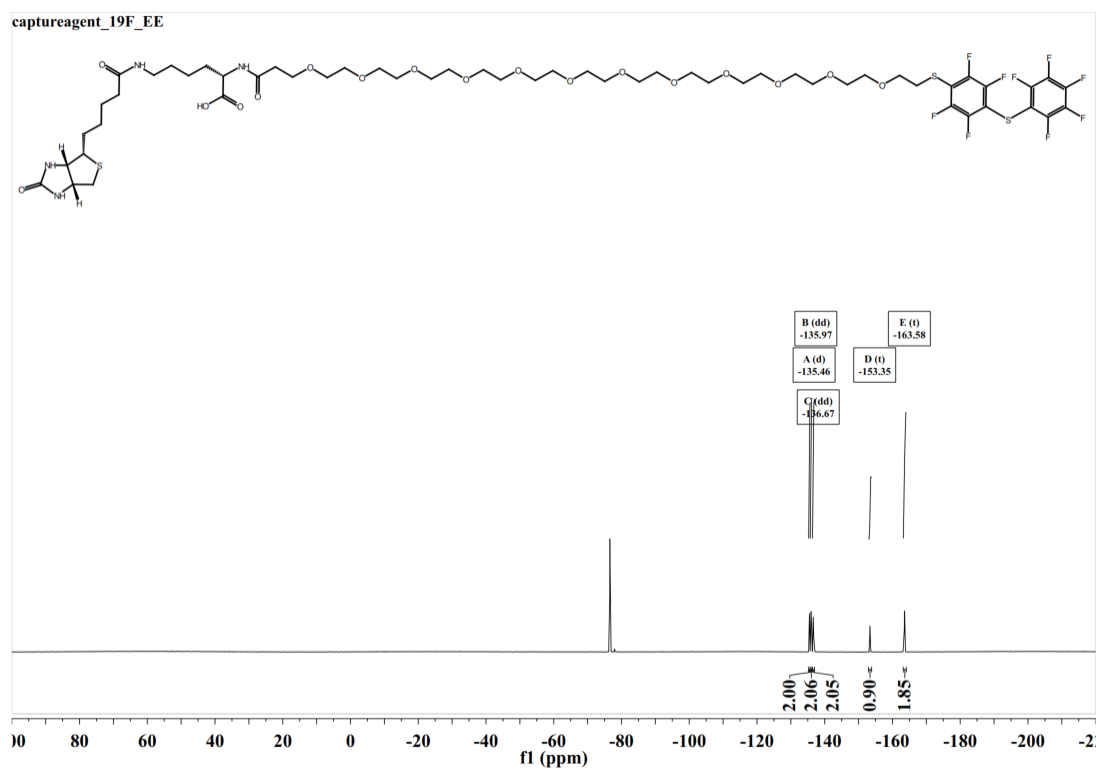


Figure S6. ¹⁹F NMR spectrum for the selection CA in DMSO-d6.

2.4. mRNA Display

Library Design

The library was designed to display a 30mer random peptide with the 14th-17th amino acids being doped as 40-50% FCPF with the following 188mer DNA sequence:

5' – TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA ATT ACA ATG NNS
NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS WWZ WYZ XXZ WWZ NNS
NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS GGC TCC GGT AGC TTA
GGC CAC CAT CAC CAT CAC CAC CGG CTA TAG GTA GCT AG – 3'

The doped FCPF was created with the following A:T:G:C ratios during DNA synthesis: W-(1:7:1:1), X-(1:1:1:7), Y-(1:1:7:1) while the G:C ratio for Z was (1:9). For this selection the following primers and oligonucleotides were purchased and used:

Library^a: 5' – TCA CTA TAG GGA CAA TTA CTA TTT ACA ATT ACA ATG NNS NNS NNS
NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS WWZ WYZ XXZ WWZ NNS NNS NNS
NNS NNS NNS NNS NNS NNS NNS NNS NNS NNS GGC TCC GGT AGC TTA GGC – 3'

F1^b: 5' – TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA ATT ACA – 3'

R1^b: 5' – CTA GCT ACC TAT AGC CGG TGG TGA TGG TGA TGG TGG CCT AAG CTA
CCG GAG CC – 3'

RT^b: 5' – TTT TTT TTT TTT TTT GTG ATG GTG ATG GTG GCC TAA – 3'

Psoralen Oligo^a: 5' – Psoralen C6-(uag ccg gug)²'-OMe-AAA AAA AAA AAA AAA-2x
Spacer9-ACC-Puromycin – 3'

Oligos were purchased from either the Keck Oligonucleotide Synthesis facility at Yale^a (New Haven, CT) or Integrated DNA Technologies^b (Coralville, IA)

Selection round protocols

The following is the complete selection protocol with parts from the manuscript methods repeated for clarity.

Initial library construction:

The single stranded DNA library was converted to the desired length dsDNA library in 10mL of PCR reaction (6 cycles of 30s 52°C annealing, 1 min 72°C extension, 30s 95°C denaturing): 25nM library, 1μM F1 and R1 primers, 0.2mM dNTPs, 1x thermo pol buffer and 2.5U/μl Taq in individual 100μl total. The library was then phenol/chloroform extracted, 1-butanol concentrated and ethanol precipitated.

1st round transcription:

A reaction containing 50nM dsDNA template, 1mM ATP, CTP, UTP, GTP, T7 buffer (1x), 0.4U/μl RNase OUT and 3U/μl T7 polymerase (carrying forward ~ 7.6x10¹³ sequences) was left at 37°C for ~15 hours. This was then purified by 6% denaturing PAGE and passively eluted. The eluted RNA was concentrated with 1-butanol and ethanol precipitated.

1st round crosslinking:

Crosslinking was performed with the following reaction: 3 μ M RNA, 7.5 μ M psoralen oligo, 1x XL buffer (100mM KCl, 1mM spermidine, 1mM EDTA, 20mM HEPES pH 7.5) as previously described.^{2,3} The mixture was heated to 70°C for 5 minutes, cooled to RT slowly, then placed on ice for >1 min. Then 100 μ l reactions were crosslinked in individual wells of a 96 well plate at 4°C with 365nm light for 20 minutes. All samples were then combined, concentrated with 1-butanol and run on 6% denaturing purification gel. For the first round, $\sim 1.28 \times 10^{-8}$ moles of RNA were input into the crosslinking reaction; thus, assuming a 2% recovery between crosslinking and input into the selection step along with a 3x oversampling, this amount would produce roughly 5.2×10^{13} unique sequences for the first round.

1st round translation:

A bulk translation was performed using the following salt optimized mixture: 28nM XL-RNA, 12.5 μ M amino acid mixture without met (AA-met), 12.5 μ M AA-leu, 3.5mM DTT, \sim 1mM Mg(OAc)₂, 140mM KCl, 0.2U/ μ l RNase OUT and 40% rabbit reticulocyte lysate. This was left at 30°C for 1.5 hours, subsequently salts were added to give \sim 50mM Mg²⁺ and 550mM K⁺, the reaction was then left for 42 minutes at room temperature, and finally placed in a -20°C freezer for 14 hours.

1st round oligo dT purification:

6x, 1mL suspensions of oligo d(T)₂₅ magnetic beads (New England Biolabs), were used for purification. The total round 1 translation was split evenly and the same protocol was performed for each of the six bead slurries. The translation mixture was combined with \sim 6x of dT binding buffer (20mM tris pH 7.5, 500mM NaCl, 1mM EDTA, 0.1% tween 20), added to a bead sample and rocked at room temperature for 1.5 hours. The beads were then washed 1x with 15 mL binding buffer, 3x 10mL wash buffer (20mM tris pH 7.5, 500mM NaCl, 1mM EDTA) and 1x 10mL low salt buffer (20mM tris pH 7.5, 200mM NaCl, 1mM EDTA, each 'wash' incorporated a 15 min incubation). Finally, 1mL of 20mM tris (pH 7.5) was added per bead suspension and all six were combined. This final mixture was placed at 65°C for 4 minutes upon which time the supernatant was removed. Then 5mL of 10mM tris buffer was added to the beads and the heating protocol was repeated. The two supernatants were combined and the concentration of RNA was determined by uv-vis spectrophotometry. The solution was filtered through a 0.22 μ m filter, concentrated on a 10K Amicon Ultra centrifugal filter (EMD Millipore) and ethanol precipitated.

1st round reverse transcription:

Reverse transcription was performed with the following conditions: 0.5mM dNTPs, 1.5 μ M RT primer, 10mM DTT, 1x 1st strand buffer, 2U/ μ l RNase OUT, 5U/ μ l SSII and the suspended mRNA-peptide. The RNA and primer were heated together at 65°C for 5 min first, then cooled to room temperature and finally placed on ice. Then the rest of the components were added and the mixture was incubated at 42°C for 55 min.

1st round Ni-NTA:

2mL of Ni-NTA agarose bead slurry were combined with the reverse transcription reaction along with 12mL of Ni-NTA binding buffer (100mM NaH₂PO₄, 6M Guan HCl, 0.2% triton X-100, 5mM β -mercaptoethanol, pH 8) and rocked at 4°C for 1 hour. The resin was washed with 3x 10mL wash buffer (100mM NaH₂PO₄, 0.2% triton X-100, 5mM β -mercaptoethanol, 300mM NaCl). Then 1mL aliquots of elution buffer (50mM NaH₂PO₄, 300mM NaCl, 5mM β -mercaptoethanol, 250mM imidazole) were added 8 times, each with a 5 min incubation. The elutions were combined, concentrated on a 10K filter and ethanol precipitated.

1st round selection:

The pellet was diluted in the round one selection mixture (1mL total) containing: 1x selection buffer (25mM HEPES-KOH pH 7.5, 100mM NaCl, 5mM CaCl₂, 5mM MgCl₂, 0.01% triton X-100), and 50μM capture agent at ~80nM RNA-peptide – this was termed the non-reduced library. This reacted 15 hours at room temperature during which time a sample for qPCR was removed (for round 1 input cDNA). The reaction was washed on a 10K filter until the concentration of free capture agent was ~0.12μM in 550μl. The concentrated selection mixture was added to ~1mg of pre-blocked (1x selection buffer and 2mg/mL yeast tRNA (Roche, Switzerland)) Pierce streptavidin magnetic beads and rotated at room temperature for 1 hour. The supernatant was removed and the beads were washed twice with 200μl, 1x selection buffer, these two washes were then combined with the first supernatant (giving a total volume of ~950μl) to which 50μl of 1mM capture agent and 2μl of 1M DTT (giving ~2mM) was added. This new, reduced reaction was left at room temperature for 18 hours – likewise an ‘input’ cDNA sample was removed for qPCR. Following concentration and capture agent dilution, the reduced library was similarly pulled down with ~0.8mg streptavidin beads. To elute both the reduced and non-reduced libraries from the beads following the initial pulldown, 1mL washes of 1x selection buffer were performed ten times, then the cDNA was eluted 4x with 100μl of 10mM tris pH 7.45 at 95°C for 3 min each and combined. This cDNA was used for the ‘output’ of round one.

1st round PCR:

Standard PCR conditions were used (30s at 95°C, 30s at 58°C, and 35s at 72°C) and both libraries were amplified for 16 rounds using the F1 and R1 primers. The mixture was then phenol and chloroform extracted, 1-butanol concentrated and ethanol precipitated. The pellet was then diluted in 10mM tris, 50mM NaCl and quantified by native PAGE densitometry.

Round 2:

Selection steps through Ni-NTA purification were performed in a similar manner to the first round for both the reduced and non-reduced selections. However, for all steps, scaled down reaction sizes were used as it was no longer necessary to carry the entire volume of each step through. This round diverged from the previous one in the selection step. The precipitated libraries were suspended in 1x selection buffer with or without 2mM DTT and a sample of round 2 ‘input’ cDNA was removed. These mixtures were then added to 0.15mg blocked streptavidin magnetic beads and incubated for 1 hour. The supernatant was removed and combined with the supernatants of four washes of the beads (all using 1x selection buffer), to this, capture agent was added, giving a 50μM final concentration and ~2.8μM RNA-peptide. The resulting mixture was left for 1 hour at room temperature. The capture agent concentration was then reduced using a 10K filter. For the pull down, 1mg streptavidin magnetic beads were washed, blocked and finally the selection mixture was added and incubated at RT for 1 hour. The beads were then washed 6x 1mL of 10mM tris at RT, then 4x 1mL 10mM tris with a 1 min incubation at 40°C. cDNA was eluted 4x, 50μl 10mM tris at 95°C and then PCR amplified.

Round 3:

Round three proceeded similarly to the previous rounds through the oligo dT purification step. Subsequently, excess salts from the dT purification were removed on a 10K concentrator, and the entire mixture for each selection was spun to 40μl. To this was added 20μl 5x selection buffer, 35μl water and 5μl 1mM capture agent, this mixture sat 30 min at room temperature (the ‘reduced’ library selection step did not have any DTT). This mixture was washed four times with water on a 10K filter; however, before the third spin, the entire solution (plus water wash) was removed from the filter and heated to 65°C for 2 min (to help remove excess capture agent) and then spun.

After the selection step a standard reverse transcription and Ni-NTA purification (both scaled appropriately) were performed and the final solution was then filtered until the imidazole was $\sim 0.131\mu\text{M}$. Here an 'input' cDNA sample was removed from the non-reduced library. The reduced library was reselected in the same final volume, concentration and time as before but with 2mM DTT, an 'input' cDNA sample was also removed. This 'redo' selection step was spun on a 10K concentrator to remove excess capture agent. Then both libraries were added to 0.2mg of prewashed and blocked streptavidin magnetic beads and incubated for 1 hour. These beads were then washed 5x 1mL at RT, 5x 1mL with 1 min at 40°C then eluted 4x, 50 μl 3 min elutions at 95°C. The two libraries were then PCR amplified.

Round 4:

The round four transcription was performed with 70nM template for 5 hours. This was then gel purified, crosslinked and translated. A standard oligo dT purification was done and then the samples were concentrated. Next a solution of 1x selection buffer, 100 μM capture agent and $\pm 3\text{mM}$ DTT (depending on the library) was created. After 30 minutes, excess capture agent was removed and a standard reverse transcription and Ni-NTA purified were performed. Following this, the solutions were spin filtered until there was only $\sim 0.12\mu\text{M}$ imidazole. 'Input' samples for qPCR were then taken. Next 0.25mg streptavidin magnetic beads were added and incubated for 1 hour. The beads were then washed 3x 1mL of 10mM tris and 50mM NaCl at room temperature, 7x 1mL with a 1 min 42°C incubation. Finally four elutions of 50 μl at 95°C were performed, and the cDNA was PCR amplified.

Round 5:

Round five followed a scaled down version of round one until the selection step. The libraries were suspended in 1x selection buffer, and the reduced selection received 2mM TCEP (instead of DTT), these mixtures were then added to prewashed beads for a negative selection and left for 15 min at room temperature. The supernatant was again incubated with a fresh batch of blocked beads for 15 minutes. The supernatant was removed, and both resins were washed with 1x buffer which was then combined with the original supernatant to which was added capture agent to a final concentration of 50 μM ; this reacted for 30 minutes. Following excess capture agent removal on a spin filter, round 'input' qPCR samples were removed. The remaining solution was added to 1mg of washed and blocked streptavidin beads for 1 hour. The beads were then washed 4x 1mL at room temperature, 6x 1mL at 42°C with 1 min incubation, the cDNA was eluted and PCR amplified like normal.

2.5. qPCR analysis

Quantitative PCR was performed at MIT's BioMicroCenter on a Light Cycler 480 II Real-Time PCR machine. To create a standard curve for each round, a sample of known concentration, reverse transcribed RNA was diluted to give a range of DNA concentrations (~ 4 orders of magnitude). PCR mixes were composed of 1 μM primers, 50% (2x) SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA), DNA template and water. Each reaction was split into three wells for triplicate measurements of C_p values, which were then averaged. Selection round yields were determined using the C_p values from samples of the selection step input and cDNA elution. A yield for each round was determined based off the known volumes of each step and the

standard curve correlating C_p and standard DNA concentration, these data were then plotted per round (Figure S7).

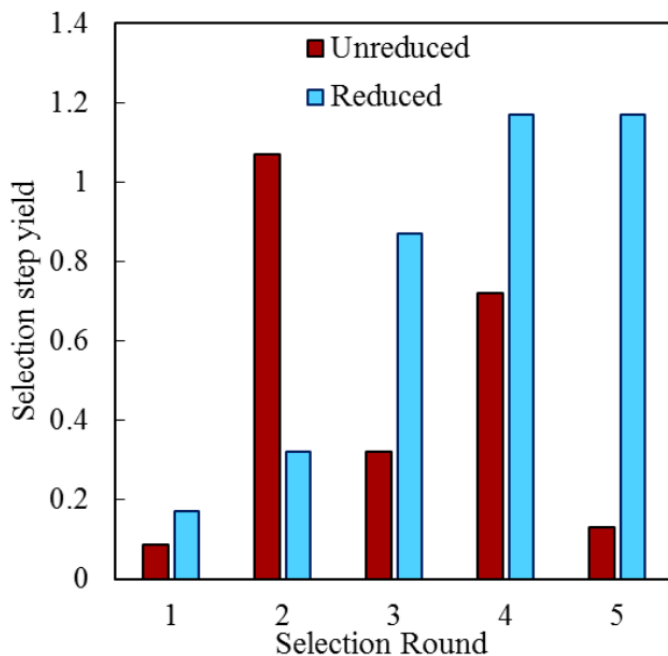


Figure S7. qPCR analysis of the two libraries used during the selection process.

2.6. HTS analysis

MiSeq (2x150bp) Illumina sequencing was performed. The FASTQ data was analyzed via custom python scripts that first combined pairs and filtered out DNA sequences possessing less than 85% Q30 Phred scores with ambiguous bases being determined by the higher Q-score base of the pair. Sequences were then translated into amino acid sequences, filtered again based off length and the presence of part of the C terminal fixed region. Sequences were then analyzed according to frequency and size of Levenshtein families with an edit distance less than five.

2.7. Kinetics and LCMS analysis

5 μ l kinetics time point samples were made to capture the initial reaction rate and quenched with the addition of 95 μ l of 49.75% H₂O, 49.75% Acetonitrile and 0.5% TFA. Time points within the linear range of the instrument were used for kinetics analysis. A second order kinetics rate constant (k_2) was extracted by fitting the data to the following equation:

$$k_2(A_0 - B_0)t = \ln\left(\frac{B_0A_t}{A_0B_t}\right)$$

For this, A_0 refers to the initial capture agent concentration, A_t is its concentration at the given time point (for kinetics data, this was obtained from the values of B_t as the TIC values of A_t were outside the linear range of the LCMS, B_0 is the initial peptide concentration and B_t signifies the peptide concentration when the sample was taken. The rate constant is an average of three measurements at different concentration (here error is estimated from the deviation in the three point estimates) while for the rest of the sequences it is estimated from a single MP concentration (for these samples, error bars represent the error determined from a linear regression fit).

2.8. Protein Expression and purification

Full length MP01 was appended to the sequence of Sortase A (with a TEV cleavage site between the two) and then placed into a pET-SUMO vector (Thermo Fisher) following factory protocols. Expression was performed in 1L cultures (30 μ g/mL kanamycin) of appropriately transformed *E. Coli* BL21(DE3), after an O.D. value of 0.5 was obtained at 37°C, the cultures were cooled to 16°C and induced with the addition of 0.2mM IPTG. Expression was conducted for 5.5 hours at 16°C followed by cell pelleting with 10 minute centrifugation at 7,000 RPM. The cell pellet was suspended in 25mL of Ni-NTA binding buffer (50mM Tris pH 8.1 150mM NaCl) with one protease inhibitor cocktail tablet (Roche Diagnostics, Switzerland), 20mg lysozyme (Calbiochem) and ~2mg DNase I (Sigma-Aldrich). Following sonication the cellular debris was removed by centrifugation at 17,000 RPM for 30 minutes. The supernatant was loaded directly onto a 5mL HisTrap FF crude Ni-NTA column (GE Healthcare, UK), following binding the column was washed with 25mL Ni-NTA binding buffer, 25mL Ni-NTA washing buffer (50mM Tris pH 8.1, 150mM NaCl, 500mM imidazole) and eluted with 10mL of Ni-NTA elution buffer (50mM Tris pH 8.1, 150mM NaCl, 500mM imidazole). The protein was then desalted on a HiPrep 26/10 Desalting Column (GE Healthcare, UK). Following concentrating, the SUMO group was removed by addition of 30 μ g of SUMO protease per mg of protein at 4°C overnight.

MP01-SrtA was further purified by anion exchange chromatography after being exchanged into buffer A (20mM HEPES pH 8.5, 1mM DTT). MP01-SrtA was loaded onto a HiTrap Q HP column (GE Healthcare, UK) in buffer A and eluted with buffer B (buffer A + 2M NaCl) during a 400mL linear gradient from 0-30% B. Fractions for the protein were combined and concentrated on a 10K spin filter and buffer exchanged into 0.5x selection buffer. The concentration were determined spectrophotometrically using 280nm light and extinction coefficients obtained from the ProtParam tool on web.expasy.org and a sample was taken for LCMS analysis (Figures S4 and S5). The protein was aliquoted and flash frozen in liquid nitrogen.

SUMO-MP01-SrtA (Calc. 36473.84). MP01-SrtA (Calc. 23075.6, green):

MGSSHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKI
KKTTPLRRLMEAFKRQKEMDSLRFYDGIQADQTPEDLDMEDNDIIEAHREQIGG
MHQKYKMTKDCFFSFLAHHKRRKLYPMSGSGSLGHHHHHHRLGENLYFQGGDPNSQA
KPQIPKDKSKVAGYIEIPDADIKEPVYPGPATSEQLNRGVSF AEENESLDDQNISIAGHTFI
DRPNYQFTNLKAAKKGSMVYFKVGNETRYKMTSIRNVKPTDVEVLDEQKGKDKQLT
LITCDDYNEKTGVWETRKIFVATEVK

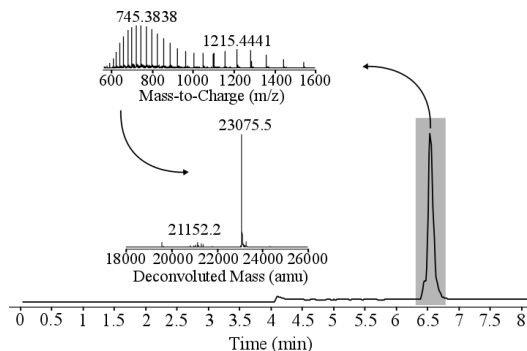


Figure S8. LCMS chromatogram and deconvoluted MS for MP01-SrtA.

2.9. Labeling site determination

To determine the location of labeling for MP01 a reaction of 0.3mM MP, 0.3mM CA, 1x selection buffer, 5mM TCEP pH 7.4 and reacted 24 hours. Next the peptide was digested with 0.2 mg/mL trypsin and chymotrypsin for 24 hours at 37°C. Fragments were then analyzed by LCMS/MS on an Agilent 6550 iFunnel Q-TOF mass spectrometer. Similarly, the CA by itself was analyzed with the same LCMS/MS protocol.

3. References

- (1) Mjalis, A.J.; Thomas III, D.A.; Simon, M.D.; Adamo, A.; Beaumont, R.; Jensen, K.F.; Pentelute, B.P. *Nat. Chem. Bio.* **2017**, *13*, 464-466.
- (2) Kurz, M.; Gu, K.; Lohse, P. A. *Nucleic Acids Res.* **2000**, *28* (18), e83.
- (3) Seelig, B. *Nat. Protoc.* **2011**, *6* (4), 540.

4. Experimental data, tables and figures

4.1. Sequence and mass of peptides used

All peptides were chemically synthesized as C-terminal amides using a rink linker as described.

Table S1. Name, sequence, calculated and observed mass of peptides.

Name	Sequence	Calculated mass	Observed mass
MP01-Full	MHQKYKMTKDCFFSFLAHHKRRKLYPMSGSGSLGHHHHHHRL	5078.5	5078.6
MP01-C	MHQKYKMTKDCFFSFLAHHKRRKLYPMSG	3585.8	3585.7
GCPG	GCPGGLLNK	984.6	984.6
S-pep-1	YALPSTGG	763.4	763.4
S-pep-2	GGGGGAGYLLGKINLKALAALAKKIL	2465.5	2465.5
MP01-T	KMTKDCFFSFL	1364.7	1364.7

4.2. HTS Levenshtein clusters, library heat maps, FCPF analysis and statistics

Levenshtein clusters (edit distance = 5), showing sequence, red point mutants relative to the cluster parent sequence and number of times each peptide appeared in the sequencing.

MP01 –

Reduced:

MHQKYKMTKDCFFSFLAHHK**Q**RKLYPMSG: 1, MHQKYKMTKDCFFSFLA**H**RKKRKLYPMSG: 2, MHQKYKMTKDCFF**P**FLAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFLAHHK**M**RKLYPMSG: 1, MHQKYKMTKDCFFSFLAHH**R**KRKLYPMSG: 1, MHQKYK**V**TKDCFFSFLAHHKRRKLYPMSG: 1, M**H**RKYKMTKDCFFSFLAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFLAHHKRRKLYP**M**GG: 1, MHQKYKMTKDCFFSFLAHHKRRK**S**YPMSG: 1, MHQKYKMTKDCFFS**S**LAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFL**S**HHKRRKLYPMSG: 1, MHQKYK**M**AKDCFFSFLAHHKRRKLYPMSG: 4, MHQKYKMTKDCFFSFLAHHKRRKLYPMSG: 185, M**H**RKYK**M**KDCFFSFLAHHKRRKLYPMSG: 1, MHQKY**E**MTKDCFFSFLAHHKRRKLYPMSG: 1, M**Y**QKYKMTKDCFFSFLAHHKRRKLYPMSG: 1, MHQ**H**KMTKDCFFSFLAHHKRRKLYPMSG: 2, MHQKYKMT**E**DCFFSFLAHHKRRKLYPMSG: 1, MHQKYKMT**R**DCFFSFLAHHKRRKLYPMSG: 3, MHQKYKMT**N**CFFSFLAHHKRRKLYPMSG: 1, MHQKYK**I**TKDCFFSFLAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFLAHHKRRKLYP**M**NG: 1, MHQKYKMTKDCFFSFLAHHKRRKLYP**T**SG: 2, MHQKYKMTKDCFFSFLA**Y**HKRRKLYPMSG: 2, M**Q**QKYKMTKDCFFSFLAHHKRRKLYPMSG: 1, MHQ**C**KMTKDCFFSFLAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFLAHHK**R**LYPMSG: 1, MHQKYKMTKDCFFSFL**T**HHKRRKLYPMSG: 1

MHQKYKMTKDCFLSFLAHHKRRKLYPMSG: 3, MHQKYKMTKDCFFSFLAHHKRRKLYPMSG: 3
 MHQKYKMTKDCFSFLAHHKRRKLYPMSG: 1, MHQKYKMTKDCFFSFLAHHKRRKLYPMSG: 1
 MRQKYKMTKDCFFSFLAHHKRRKLYPMSG: 2, MHQKYKMTKDCFFSFLAHHKRRKLYPMSG: 1

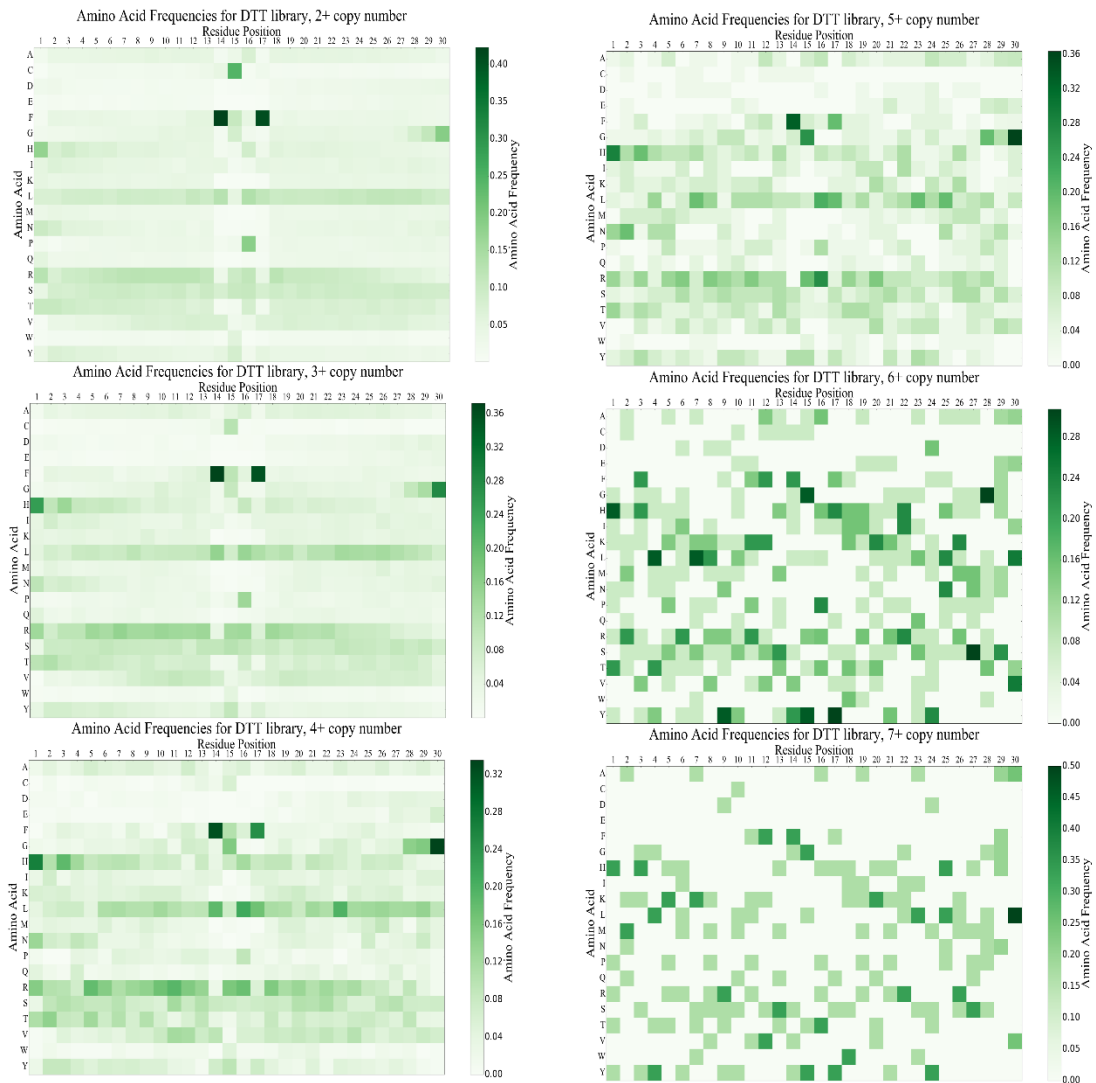


Figure S9. Amino acid frequency heat maps for the reduced library for sequences possessing a copy number greater than a minimum cutoff.

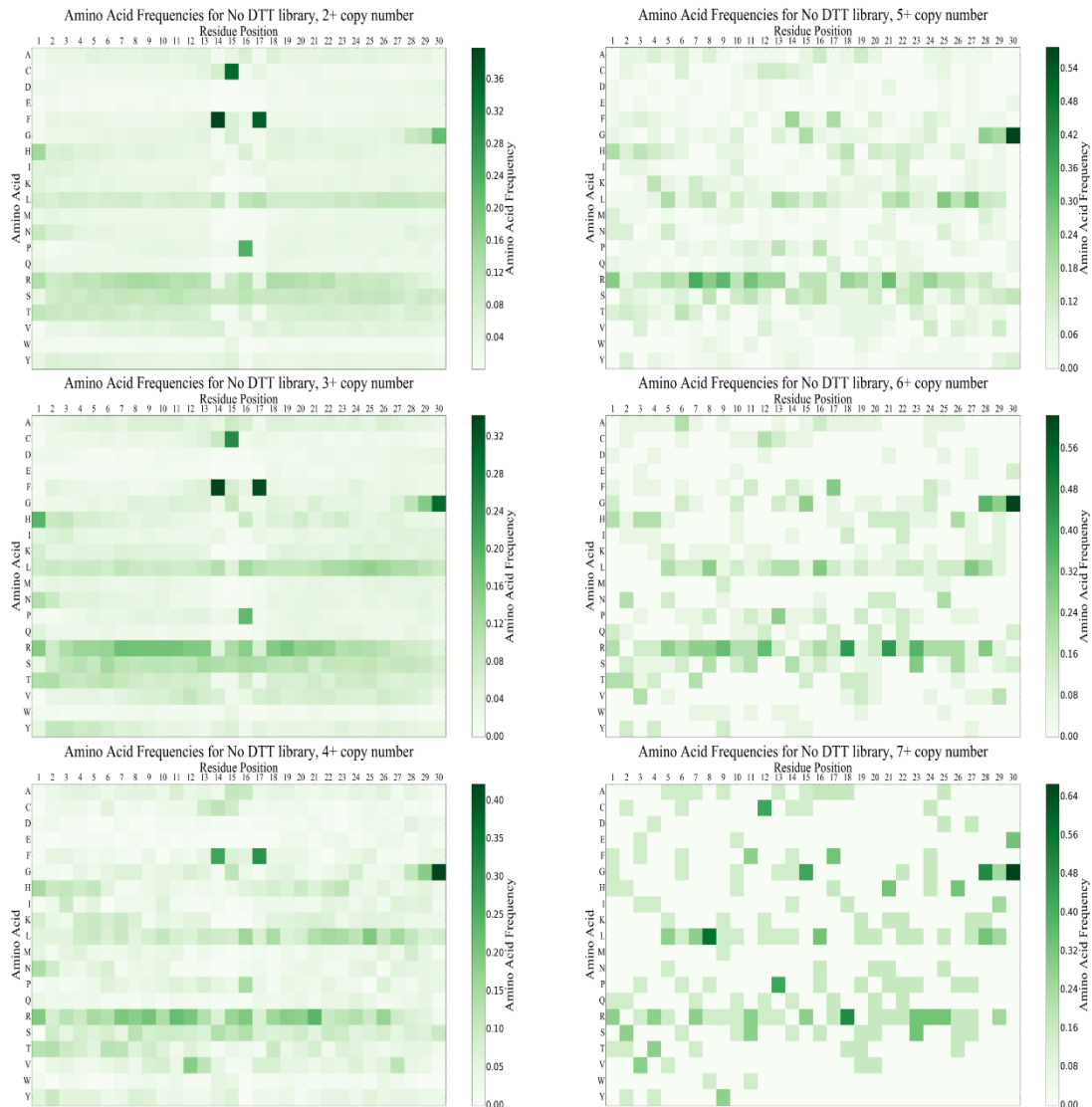


Figure S10. Amino acid frequency heat maps for the non-reduced library for sequences possessing a copy number greater than a minimum cutoff.

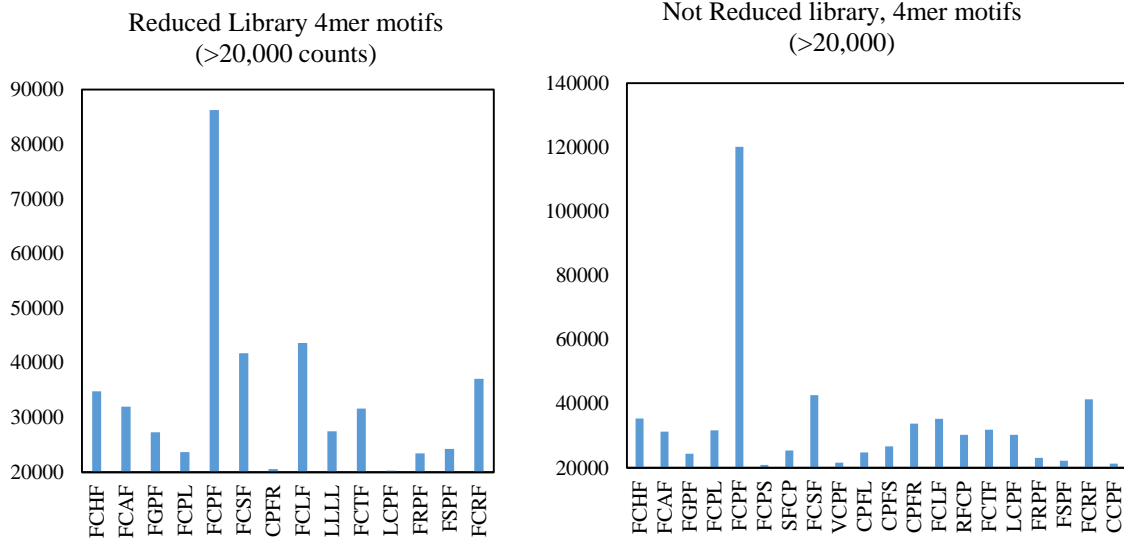


Figure S11. Motif analysis for both libraries depicting the most frequent 4mer motifs.

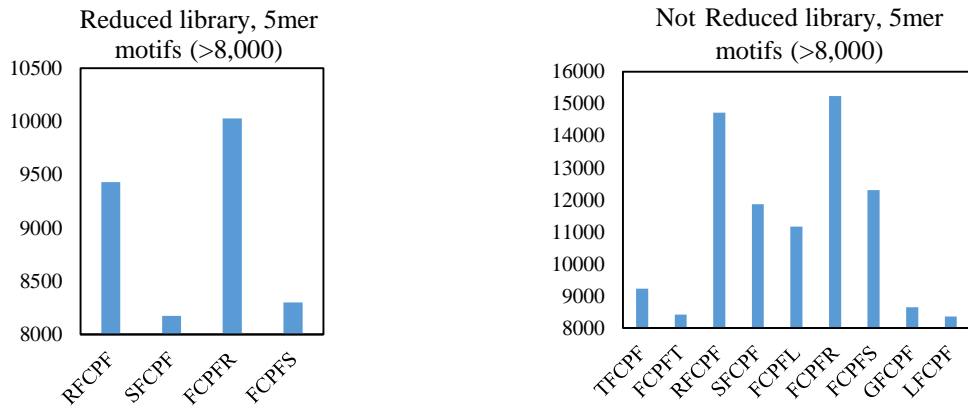


Figure S12. Motif analysis for both libraries depicting the most frequent 5mer motifs.

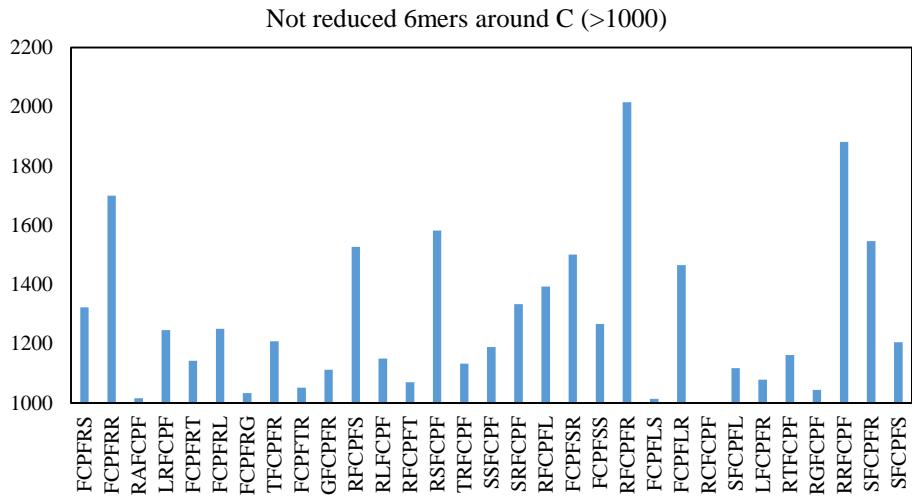
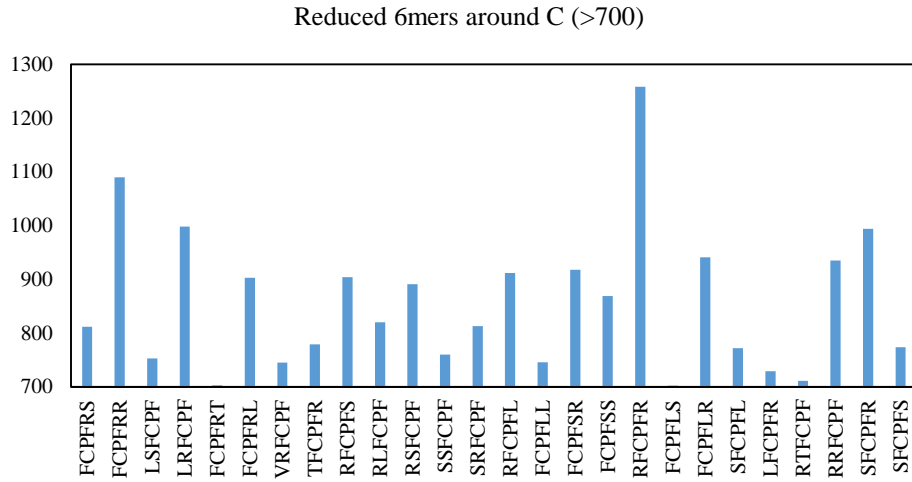


Figure S13. 6mer motif analysis for sequences with a cysteine for both libraries (reduced library top, not reduced bottom).

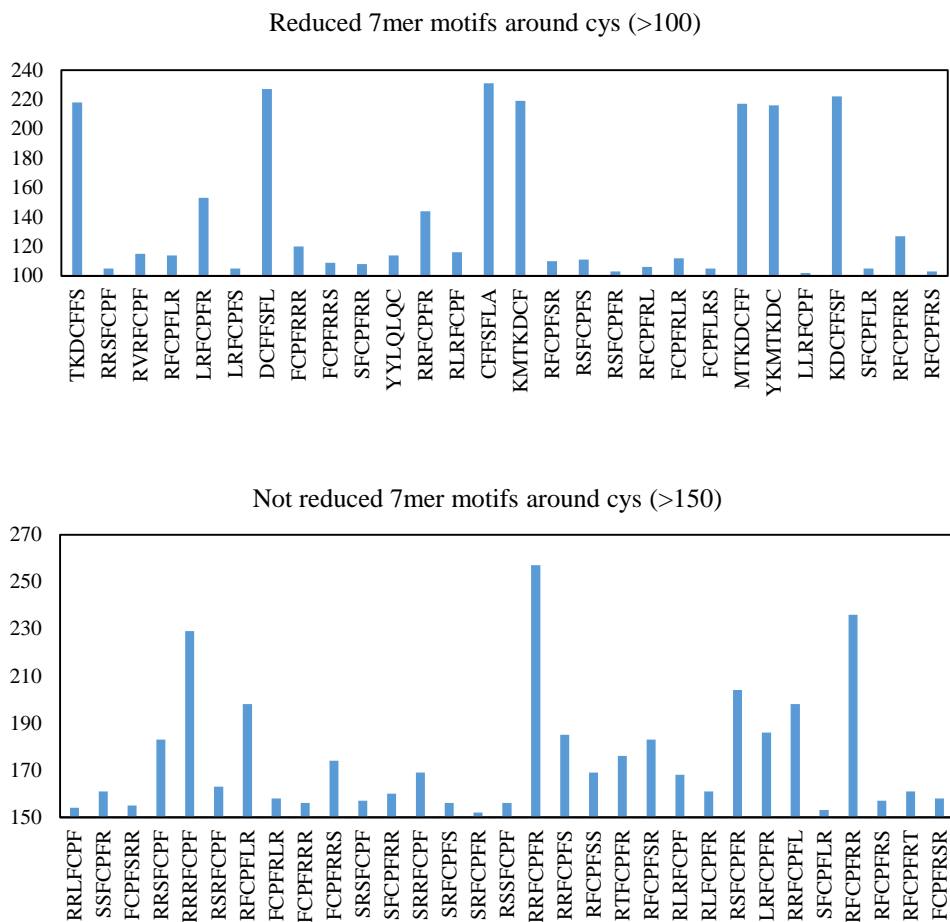


Figure S14. 7mer motif analysis for sequences containing cysteines for both libraries (top – reduced library, bottom – not reduced library). Most prominent motifs seen in the reduced library originate from the sequence of MP01

Table S2. General library statistics

Reduced library		Not reduced library	
Total count	3945597	Total count	3923881
Library size with cysteine	2457256	Library size with cysteine	3061270
Library size without cysteine	1488341	Library size without cysteine	862611
Counts of FCPF	86277	Counts of FCPF	120081

4.3. Kinetics chromatograms and plots

Aliquots were quenched at select time points with 19x (by volume) of a mixture with 49.75% water, 49.75% acetonitrile, 0.5% trifluoroacetic acid and analyzed by LCMS. For all except the GCPG control, the LC method was method 1.

Note:

For the following chromatograms in this section, the CA elutes at ~10m (on LC method 1) in all and is not labeled after the first two kinetics TIC traces. Likewise, the +16 Da impurity from the CA elutes immediately before it and is likewise not labeled for the majority of the traces.

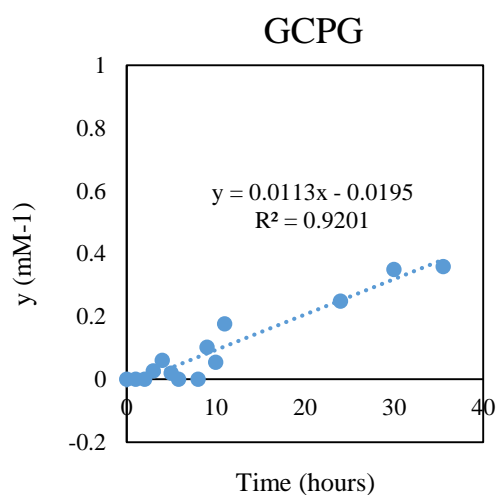
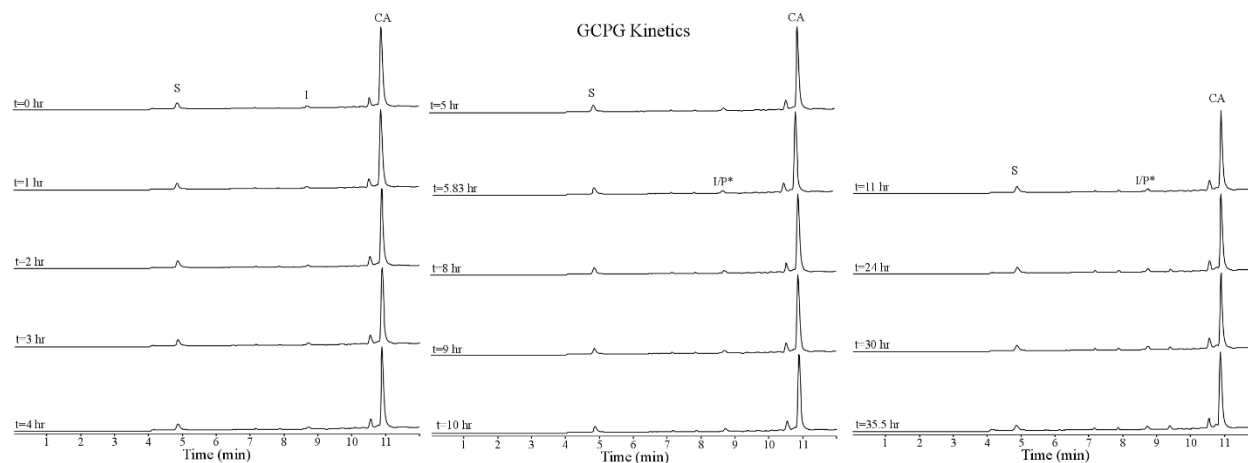


Figure S15. TIC trace (top) and kinetic analysis for the GCPG control peptide for measuring the background rate of reaction. All negative values were adjusted to 0 for the beginning times. I – impurity seen on the LCMS column and not in the reaction mixture, S – starting material, P – singly labeled product. LC method: 0-2 minutes 5% B, 2-11 minutes 5-65% B linear ramp, 11-12 minutes 65% B, 0.8mL/min flow rate.

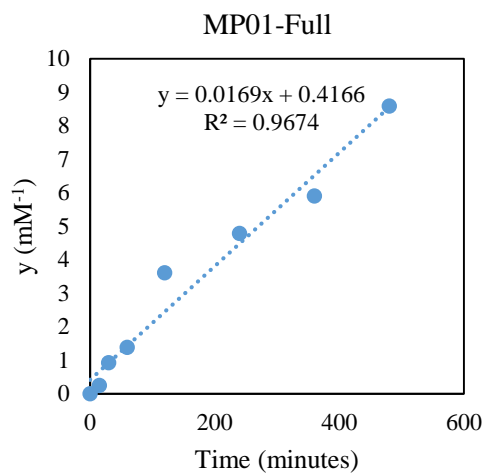
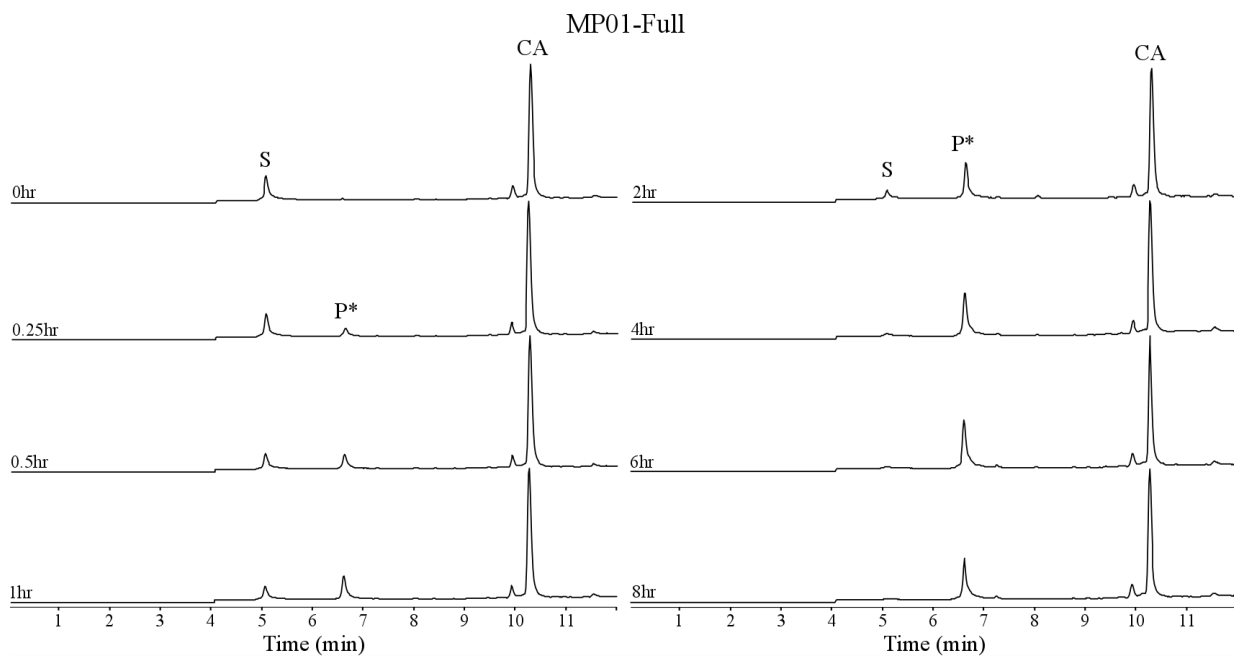


Figure S16. LCMS TIC traces for the kinetic analysis of full length (C-terminal constant region containing) MP01 (top). S – starting material, P* – singly labeled product. Below is the plot of integrated TIC peak area to determine a second order rate constant.

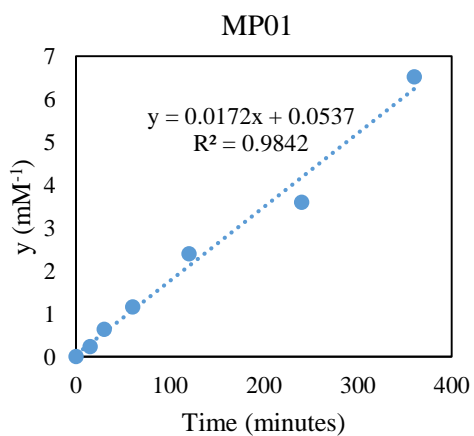
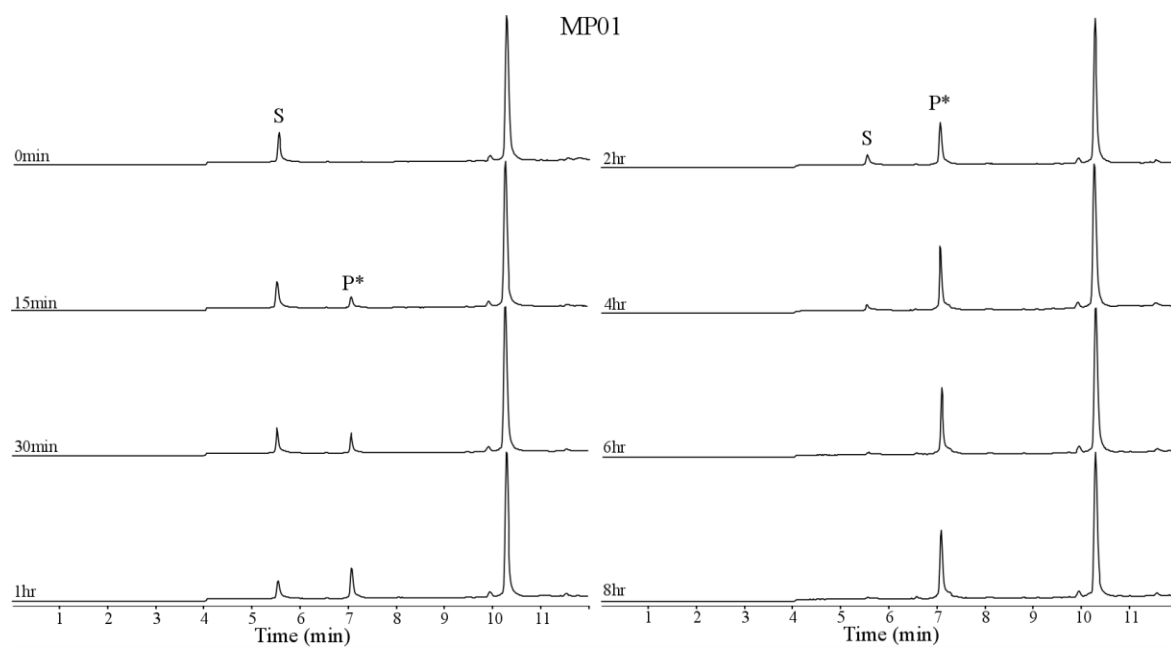


Figure S17. LCMS TIC traces for the kinetic analysis of MP01 lacking the C-terminal constant region (top). S – starting material, P* – singly labeled product. Below is the plot of integrated TIC peak area to determine a second order rate constant.

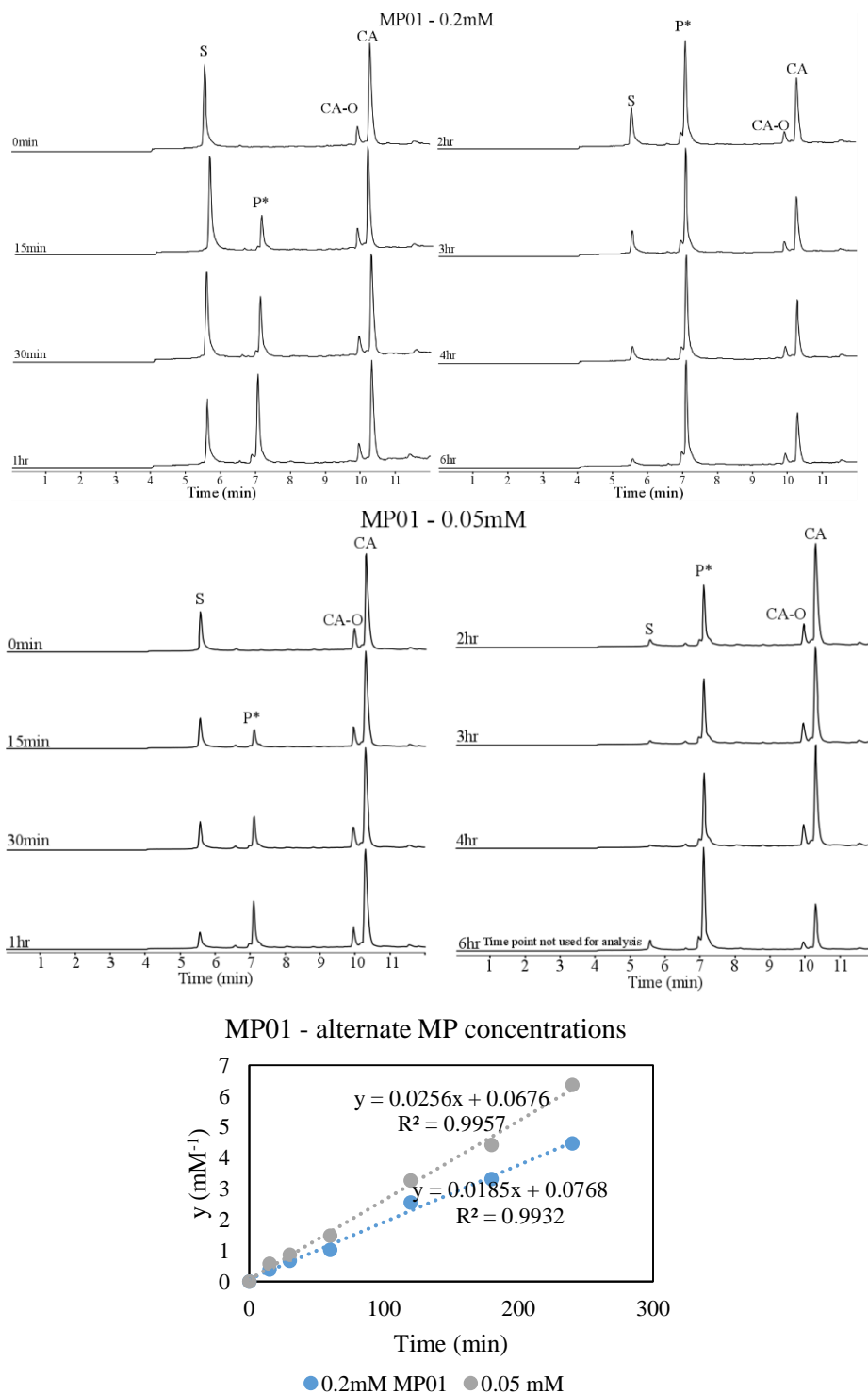


Figure S18. LCMS TIC traces for the kinetic analysis of MP01 lacking the C-terminal constant region using 0.2mM (top) or 0.05mM MP01 (middle). S – starting material, P* – singly labeled product, CA – capture agent, CA-O – oxidized CA. Below is the plot of integrated TIC peak area to determine a second order rate constant. The final time point for the 0.05mM was believed to be an error as all integrations were significantly off from reasonable values.

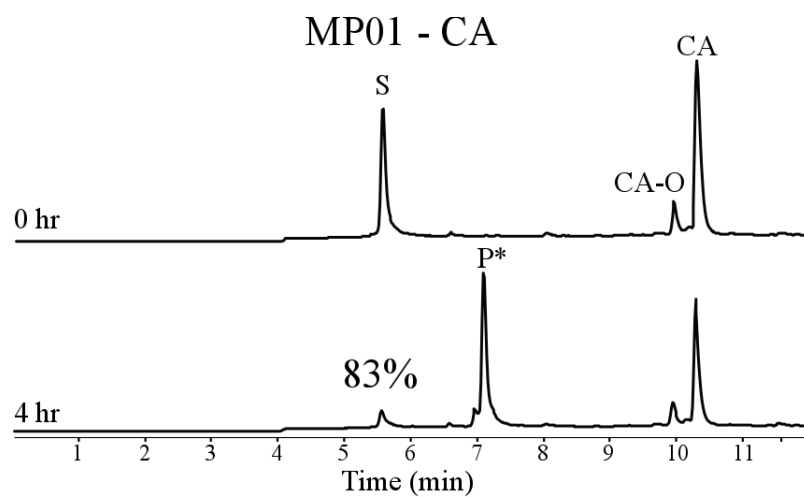


Figure S19. LCMS TIC traces for the CA conversion analysis of MP01 lacking the C-terminal constant region (top). S – starting material, P* – singly labeled product.

4.4. Determination of labeling location

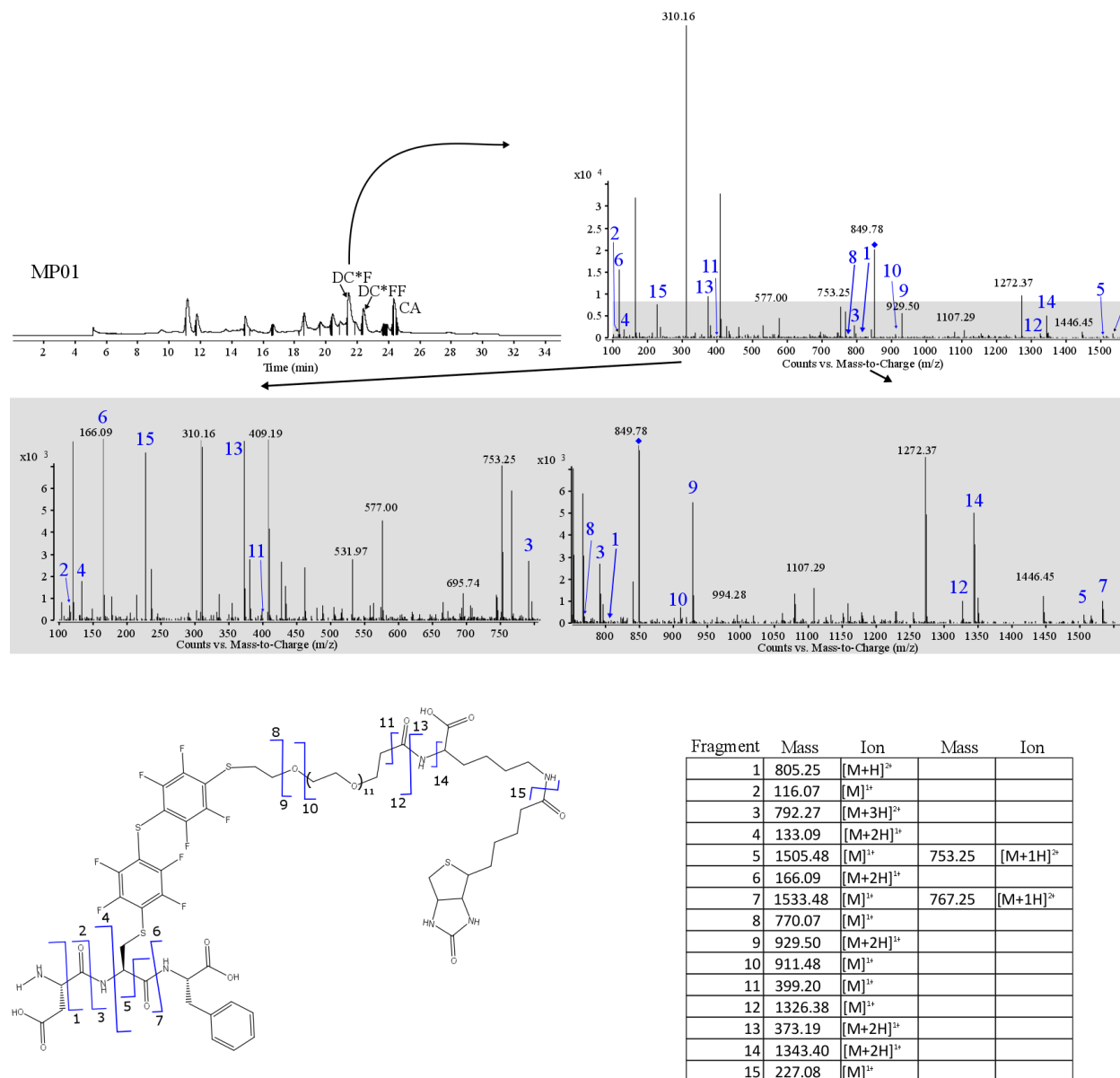
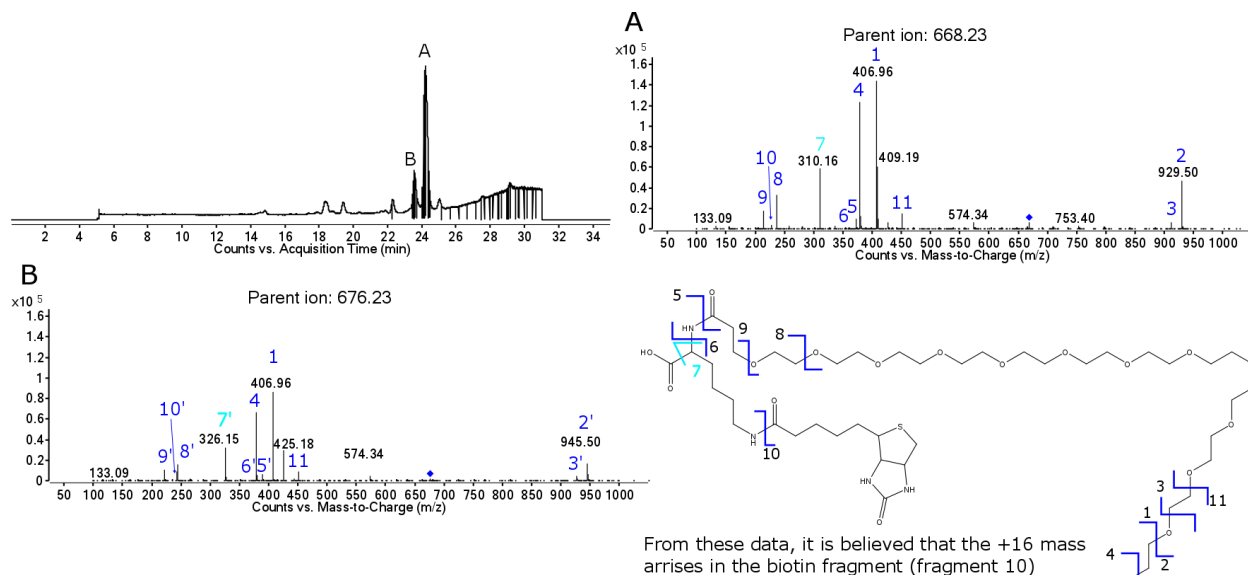


Figure S20. LCMS/MS analysis of the digested MP01 post labeling. Shown is the secondary MS for the parent ion corresponding to the $[M+2H]^{2+}$ ion of the fragment sequenced as DC*F where * is the covalently bound CA (calc. 849.78 Da, obs. 849.78 Da for the second charge state). The table in the lower right displays the observed masses corresponding to the fragments depicted in the structure.



A			B			
Main product			side product			
parent	668.09	$[M+2H]^{2+}$	parent	676.20	$[M+2H]^{2+}$	
ion			ion			
1	406.96	$[M]^{1+}$	1	406.96	$[M]^{1+}$	
2	929.50	$[M+2H]^{1+}$	2	missing	2'	945.49
3	911.49	$[M]^{1+}$	3	missing	3'	927.48
4	378.93	$[M]^{1+}$	4	378.93	$[M]^{1+}$	
5	373.19	$[M+2H]^{1+}$	5	missing	5'	389.18
6	355.18	$[M-1H]^{1+}$	6	missing	6'	371.17
7	310.16	$[M-1H]^{1+}$	7	missing	7'	326.15
8	236.12	$[M+1H]^{2+}$	8	missing	8'	244.11
9	214.10	$[M+1H]^{2+}$	9	missing	9'	222.10
10	227.08	$[M]^{1+}$	10	missing	10'	243.08
11	450.99	$[M]^{1+}$	11	450.99	$[M]^{1+}$	
	seen but not identified			seen but not identified		
	409.19			425.18		

Figure S21. LCMS/MS analysis of the CA. Shown is the secondary MS for the parent ions corresponding to peaks A and B, the charge state of the two peaks is the $[M+2H]^{2+}$ ion of the CA (A) and side product (B). Masses are: calc. 668.22 Da, obs. 668.23 Da for peak A and obs. 676.23 Da for peak B, a calculated mass is not given as the modification has not yet been determined (it is believed to be oxidation on biotin based on this data). The table at the bottom displays the observed masses corresponding to the fragments depicted in the structure. For peak B, fragments corresponding to the mass of an ion observed in A with an additional 16 Da are labelled as the corresponding ion prime (ex $2 \rightarrow 2'$) all of which are localized to the biotin containing fragments. Shown in teal is a potential fragment to explain the 310.16 mass observed in all spectra.

4.5. Analysis of truncated MP01

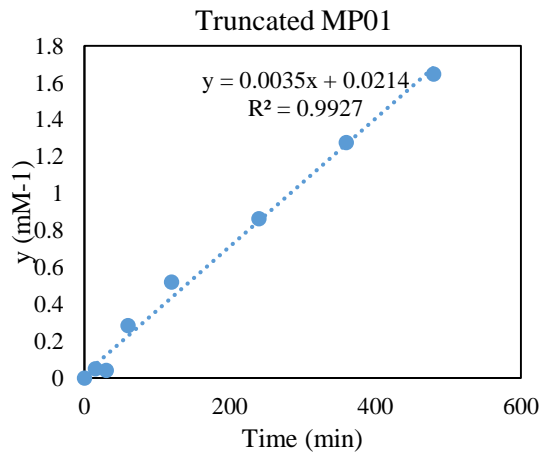
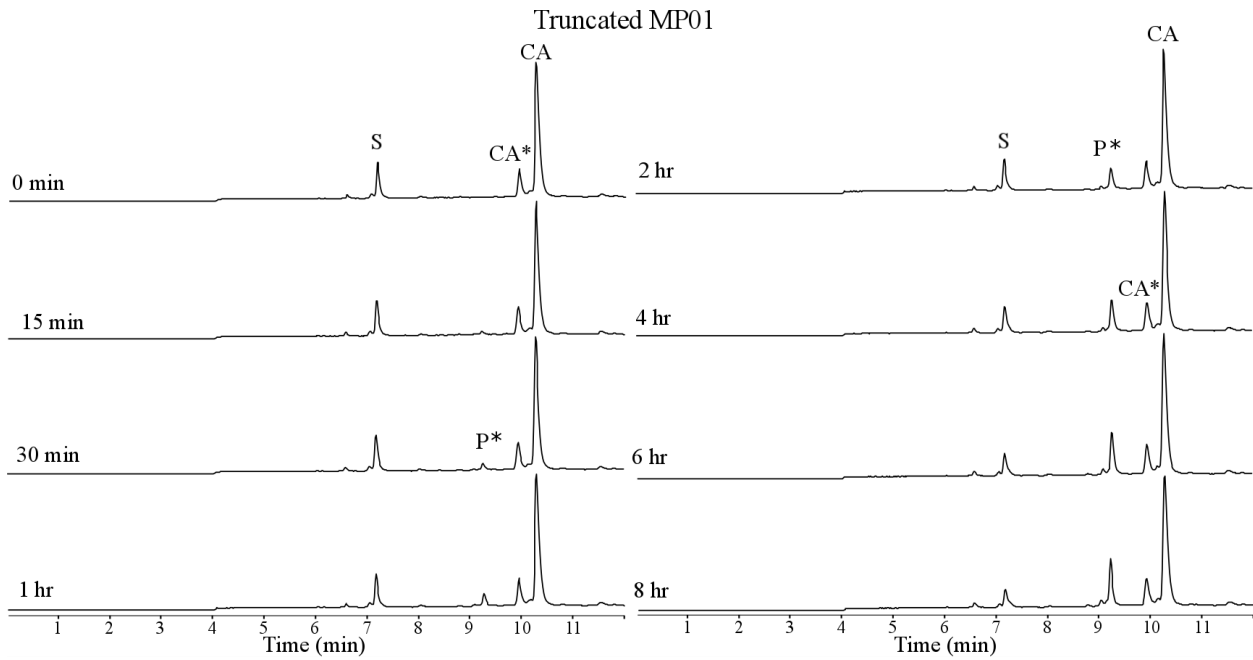


Figure S22. LCMS TIC traces for the kinetic analysis of the truncated MP01 (MP01-T) sequence (KMTKDCFFSFL) using 0.1mM MP01-T, 0.5mM CA, 5mM TCEP, 1x selection buffer, pH 7.4. S – Starting peptide, P* – single labeled product, CA – capture agent, CA* - CA impurity (oxidized). Below is the plot of integrated TIC peak area to determine a second order rate constant.

4.6. Analysis of MP01 with urea

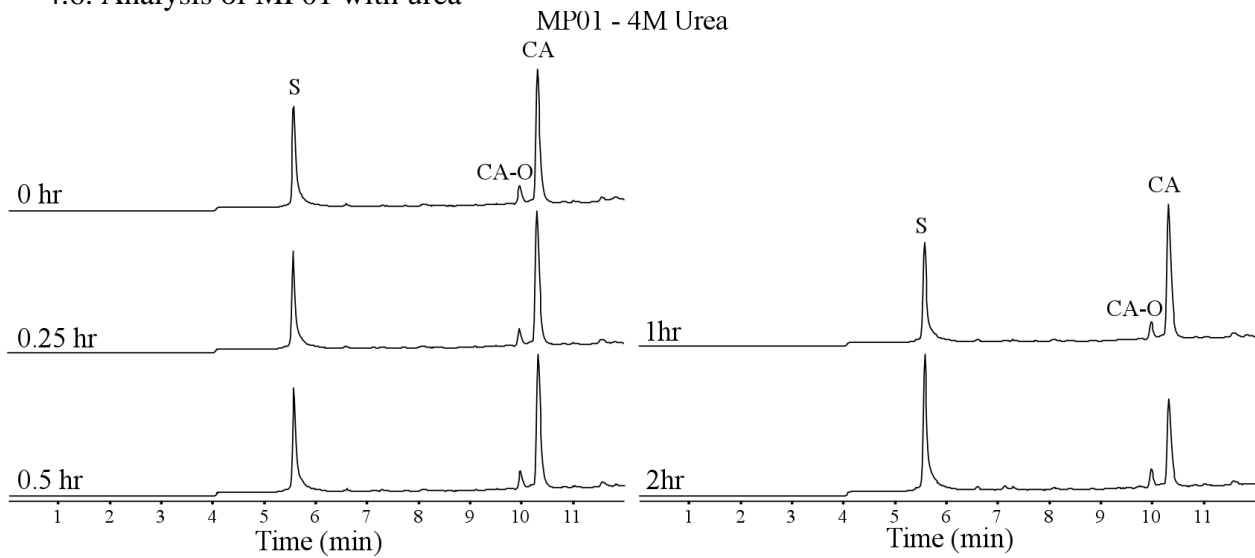


Figure S23. LCMS TIC traces for the kinetics analysis of 0.1mM MP01 with 0.5mM CA, 5mM TCEP, 4M urea and 1x selection buffer at pH 7.4. S – Starting peptide, CA – capture agent, CA-O – oxidized capture agent.

4.7. MP-SrtA labeling, TEV cleavage and SrtA reactions

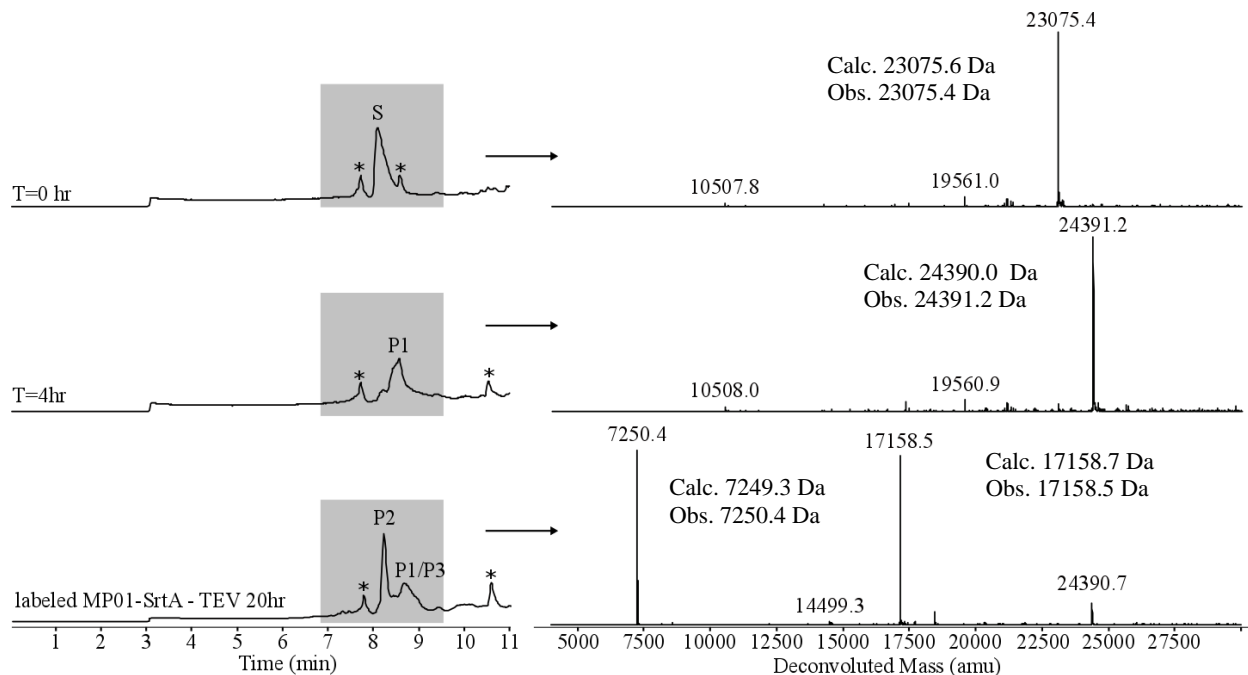


Figure S24. MP01-SrtA labeling showing the initial and 4 hour labeled chromatograms plus their deconvoluted MS (time window shown in gray) with the accompanying TEV cleavage (bottom). LCMS analysis was performed using method 3 with the MS turned off at 11 minutes to avoid over saturating the MS detector with CA. * refers to small molecules on the column and are not related to the protein labeling. S – starting material, P1 – capture agent labeled MP01-SrtA, P2 – TEV cleaved SrtA, P3 – TEV cleaved labeled MP01.

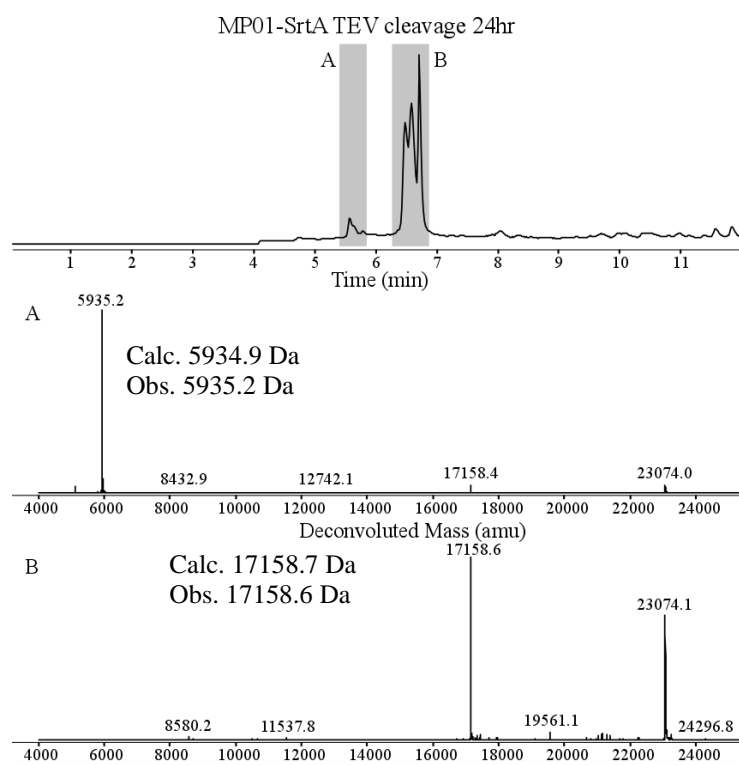


Figure S25. Top, LCMS TIC traces for 24 hour TEV cleaved, unreacted MP01-SrtA fusion. Highlighted areas correspond to the deconvoluted MS below.

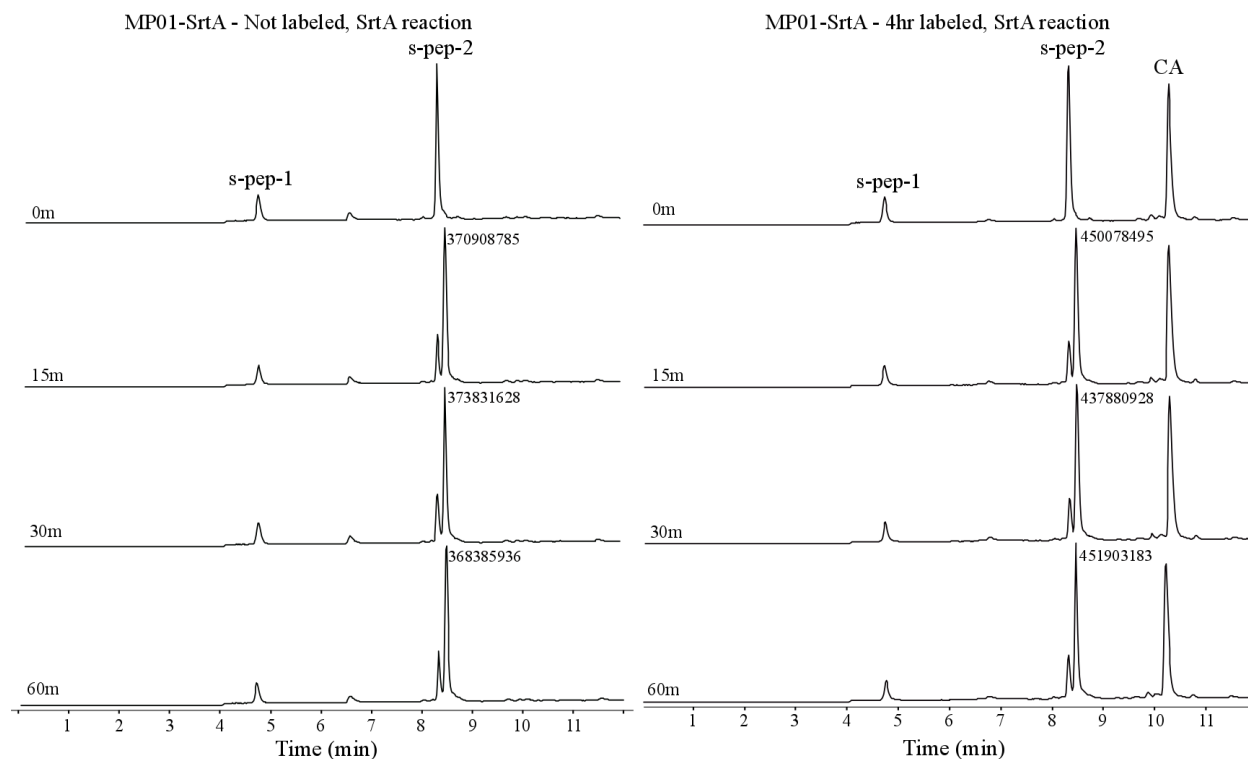


Figure S26. Labeled and unlabeled MP01-SrtA, Sortase reactivity assay, integration areas are shown for the product peak only. Reactions consisted of 5 μ M MP01-SrtA (either from a 4 hour CA reaction or unreacted), 250 μ M S-pep-2 and 50 μ M S-pep-1 and 10mM CaCl₂. Samples were taken at listed times, P – ligated product, with its integrated area displayed, CA – capture agent. S-pep-1 and s-pep-2 were chosen because they were known to be good substrates for the Sortase enzyme with good chromatographic behavior.