Supplementary Information for:

Learning Structure Activity Relationships (GE-SAR) of the Wittig Reaction from Genetically-Encoded substrates

Kejia Yan[a], Vivian Triana[a], Sunil Vasu Kalmady[b], Kwami Aku-Dominguez[a], Sharyar Memon[c], Alex Brown[a], Russ Greiner[b,d], Ratmir Derda[a]*

[a] Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2, Canada.
[b] Department of Computer Science, University of Alberta, Alberta, AB T6G 2E8, Canada.
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G 1H9, Canada.
[d] Alberta Machine Intelligence institute, Alberta, AB T5J 3B1, Canada.

Corresponding author email: ratmir@ualberta.ca

Table of Contents

Abbreviations 5
1. Synthetic methods.....6
1.1 *General chemistry information* 6
1.2 *Solid peptide synthesis*.....6
1.3 *Oxidation of N-terminal serine peptides* 7
1.4 *Kinetics of selected peptides*.....7
Supplementary Table S1. List of aldehydes and reaction rates.....8
1.4 *Hydrazone ligation experiments*.....9
1.5 *In situ E/Z selectivity determination*.....9
2. Biochemistry methods.....9
2.1 *General biochemistry information* 9
2.2 *Phage functionalization*.....9
2.2.1 *$SXCX_3C$ and SX_4 1% biotinylation and purification* 9
2.2.2 *$SXCX_3C$ and SX_4 0.1% biotinylation and purification* 10
2.2.3 *$SXCX_3C$ and SX_4 oxidations and purification* 10
2.3 *Phage pull-down experiments* 10
2.3.1 *1% and 0.1% SX_4 and $SXCX_3C$ selections*.....10
Figure S1: DNA sequences of PCR amplification protocol for Illumina deep sequencing.11
2.4 *PCR of phage*.....11
2.4.1 *PCR of NaOH elutions*.....11
2.4.2 *PCR of inputs*.....11
3. Data analysis.....12
3.1 *General data processing methods* 12
3.2 *Processing of illumina data*.....12
Supplementary Table S2 12
Supplementary Table S3 13
4. DFT computation 13
5. Machine Learning.....13
Figure S2. Analysis of the Wittig reaction in a population of peptide aldehydes. 15

Scheme S1. Analysis of the relationship between modification, panning and deep sequencing. 16

Figure S3. Comparison of 20×20 plots with different combinations of amino acid positions. 18

Figure S4. Experimental kinetic traces for HCO-X_4 peptides with high "Deep Conversion" values. 19

Figure S5. Experimental kinetic traces for HCO-X4 peptides with medium "Deep Conversion" values. 20

Figure S6. Experimental kinetic traces for HCO-X_4 peptides with low "Deep Conversion" values. 21

Figure S7. Experimental kinetics for HCO-WWXX with no Illumina counts. 22

Figure S8. Experimental kinetics for HCO-PXXX sequences.....23

Figure S9. Proline-Alanine scan to determine position and quantity of hydrogen-bond donors for stabilization of OPA transition state. 24

Figure S10. Example of SPXXX LCMS traces for determination of rate of intramolecular cyclization that produces a byproduct. 25

Figure S11. Gibbs free energy barrier of model peptides 26

Figure S12. Geometries of *trans (E)* TS1 model peptides. 27

Figure S13. Geometry of *cis (Z)* TS1 model peptides. 28

Figure S14. Kinetics of hydrazone ligation for HCO-WWRR at [H_2SO_4] = 65 mM. 29

Figure S15. Kinetics of hydrazone ligation for HCO-PPAA at [H_2SO_4] = 65 mM. 30

Figure S16. Kinetics of hydrazone ligation for HCO-WWRR at pH 5 31

Figure S17. Kinetics of hydrazone ligation for HCO-PPAA at pH 5 32

Figure S18. E/Z selectivity for HCO-WWRR (1:1 *E/Z*). 33

Figure S19. E/Z selectivity for HCO-HWFP (1:9 *E/Z*). 34

Figure S20. E/Z selectivity for HCO-PPAA. (*E/Z* 4.5:1). 35

Figure S21. E/Z selectivity for HCO-AA (*E/Z* 4.5:1).. 36

Figure S22. E/Z selectivity for HCO-SarcosineSarcosine (*E/Z* 3.6:1). 37

Figure S23. (a) Wittig reaction on SXCXXXC phage libraries. 38

Figure S24. Distribution of log DC value (DC = "deep conversion") of $SXCX_3C$ library. 39

Figure S25. Kinetic traces for sequences of $SXCX_3C$ selection 40

Figure S26. Distribution of log DC value. 41

Figure S27. A single decision tree for illustrative purposes. 42

Figure S28. Number of sequence patterns vs threshold number of sequences used as input features for the models. 43

Figure S29. Screenshot of the DC prediction app, http://44.226.164.95/. 44

Figure S30. Probability distributions for predictions and corresponding F1-scores for different probability thresholds. 45

Table S4. Performance metrics of the two classifiers for a 5-fold cross validation. 46

Table S5. Reaction rates of fast, average, and slow peptide sequences predicted by machine learning measured by HPLC. 47

HPLC purity and LCMS traces of synthesized peptides 48

Figure S31. Summary for HCO-WWRR synthesis 48

Figure S32. Summary for HCO-WWGP synthesis 49

Figure S33. Summary for HCO-QWLH synthesis 50

Figure S34. Summary for HCO-WWGL synthesis 51

Figure S35. Summary for HCO-WWPQ synthesis 52

Figure S36. Summary for HCO-WLPR synthesis 53

Figure S37. Summary for HCO-LWYR synthesis 54

Figure S38. Summary for HCO-WIVR synthesis 55

Figure S39. Summary for HCO-HWFP synthesis 56

Figure S40. Summary for HCO-ALRV synthesis 57

Figure S41. Summary for HCO-APAA synthesis 58

Figure S42. Summary for HCO-PPAA synthesis 59

Figure S43. Summary for HCO-PPLA synthesis 60

Figure S44. Summary for HCO-PPPA synthesis 61

Figure S45. Summary for HCO-PRLP synthesis 62

Figure S46. Summary for HCO-PQPL synthesis 63

Figure S47. Summary for HCO-PPPL synthesis 64

Figure S48. Summary for HCO-PPPP synthesis 65

Figure S49. Summary for HCO-PYPA synthesis 66

Figure S50. Summary for HCO-PAAA synthesis 67

Supporting information references 68

Abbreviations

dNTP	Deoxyucleoside Triphosphate
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EtOH	Ethanol
ESI	Electrospray Ionization
HILIC	Hydrophilic Interaction Chromatography
HRMS	High-resolution Mass Spectrometry
MOPS	3-(N-morpholino)propanesulfonic acid
MeCN	Acetonitrile
NMR	Nuclear Magnetic Resonance
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
RNAse	Ribonuclease
UPLC-MS	Ultra Performance Liquid Chromatography- Mass Spectrometry

1. Synthetic methods

1.1 *General chemistry information*

Chemical reagents and solvents were purchased from Sigma-Aldrich or Fisher Scientific unless noted otherwise. Reagents for peptide synthesis were purchased from ChemPep; model peptides were synthesized using standard Fmoc solid phase synthesis as described below. Ylide Ester Biotin (YEB) was synthesized as previousl reported.[1] HRMS (ESI) spectra were recorded on Agilent 6220 oaTOF mass spectrometer using either positive or negative ionization mode. Characterization of reaction crude was performed with UPLC-MS using a C18 column (Phenomenex Kinetex 1.7 μm EVO C18, 2.1×50 mm) running with a gradient of water/acetonitrile with 0.1% formic acid from 98/2 at 0 min to 40/60 at 5 min under a flow rate of 0.5 mL/min. HPLC kinetics were performed in an Agilent 1100 Series using a Thermo Scientific C18 column (Hypersil GOLD, 3 μm, 50×2.1mm). ^{1}H and ^{13}C NMR spectra were acquired on a 600 MHz four channel Agilent VNMRS spectrometer, equipped with a z gradient HCN probe and using VNMRJ 4.2A as the acquisition software. Suppression of the H_2O signal was performed using either presaturation or excitation sculpting.[2]

1.2 *Solid peptide synthesis*

Rink Amide AM resin (200 mg, 0.91 mmol/g, 0.18 mmol) was weighed into a Poly-Prep® chromatography column. The column was set up on a vacuum manifold. The manifold was equipped with a three-way stopcock that allows draining of the solvent by vacuum filtration and agitation of the resin by nitrogen bubbling.[3] CH_2Cl_2 (3 mL) was added to the dried resin for swelling. After 15 min, the solvent was drained by vacuum aspiration. The resin was washed with DMF (3 mL) and the protective Fmoc group was cleaved with 20% (v/v) piperidine in DMF (3 mL) for 1 min. The treatment was repeated for 10 min using fresh 20% (v/v) piperidine in DMF (3 mL). The resin was washed with DMF (4 × 3 mL). Fmoc-protected amino acid (0.73 mmol, 4 eq.) in DMF (1 mL) and HBTU (276 mg, 0.73 mmol, 4 eq.) in DMF (1 mL) was added to the resin followed by N,N-diisopropylaminoethylamine (DIPEA, 0.25 mL, 1.46 mmol, 8 eq.). After 30 min of agitation with nitrogen, the reagents were removed by vacuum aspiration and the resin was washed with DMF (4×3 mL). The Fmoc-deprotection, amide coupling, and washing steps were repeated consecutively as described above to elongate the peptide sequence. After final Fmoc-deprotection, the resin was washed with DMF (5 ×3 mL), followed by CH_2Cl_2 (5×3 mL). The resin was left on the manifold for 10 min to dry under the vacuum. A cleavage cocktail containing TFA/H_2O/phenol/TIPS [3 mL, 85/5/5/5 (v/v/w/v)] was added to the resin. The column was left on a rocker for 2 h to cleave the peptide then the solution was collected, and the resin was rinsed with TFA (1 mL). The combined cleavage mixture was added dropwise to ice cold diethyl ether (20 mL) in a 50 mL centrifuge tube. The mixture was incubated on ice for 30 min then centrifuged for 5 min at 3000 rpm. Supernatant was decanted and the precipitates were resuspended in cold diethyl ether (10 mL). The centrifugation and washing steps were repeated 2 times. The precipitates were air-dried and then left under vacuum overnight. Typical yield: 30–150 mg.

Crude peptide SAA (NH_2-Ser-Ala-Ala-$CONH_2$) was dissolved in water (5 mL). The solution was injected into a preparative HILIC-HPLC system. A gradient of solvent A (MQ deionized water, 0.1% (v/v) TFA) and solvent B (MeCN, 0.1% (v/v) TFA) was run at a flow rate of 13 mL/min as shown below.

Time (min)	Eluent B (%)
0	99.5

2	99.5
18	30
21	0
26	0
28	100
30	100

The fractions containing target peptides were identified using mass spectrometry ESI LCMS. MeCN was removed by evaporation under reduced pressure. The aqueous solution was lyophilized to yield the peptide as white powder (32 mg, 16 %, the yield was calculated amount of resin(200 mg) and its loading (0.91 mmol/g) and theoretical amount of SAA peptide (MW = 247) that could be made on theis amount of resin with this loading is 199 mg).

1.3 *Oxidation of N-terminal serine peptides*

This is a representative procedure for peptide SHAWD: The peptide solution in MeCN/water (16.2 mg in 1mL, 26.4 mM, 1 eq.) was incubated in ice for 5 minutes. $NaIO_4$ solution in water (14.1 mg in 282 μL, 66 mM, 2.5 eq) were added. The mixture was incubated on ice for 20 minutes and injected into preparative HILIC column for quenching and purification. filtered by a 0.2 micron filter and injected into preparative HPLC (HILIC, 0-2 min 99.5 % MeCN, 2-18 min 99.5-30 % MeCN, 18-21 min 30 % MeCN, 21-26 min 0 % MeCN, 26-28 min 0-100 % MeCN, 28-30 min, 100% MeCN). Collected fraction was lyophilized and peptide HCO-HAWD was obtained as a light yellow solid (13.0 mg, 84%). Apparent purity (95%) was estimated by LC-MS with UV-Vis detector at 214 nm. The MeCN was evaporated via speedvac and the oxidized peptide was dried using lyophilization.

When the peptide contains a methionine, a 2 mM solution of peptide (in 1× PBS) was kept in ice while 1.5 equivalents of NaOH were dissolved in water. The NaOH was added and after 10 seconds, the reaction was quenched with 12 equivalents of glutathione. The mixture was incubated in ice for ten minutes and then injected on C18 or HILIC column for purification. Acetonitrile was removed with speedvac and the oxidized peptides are lyophilized.

1.4 *Kinetics of selected peptides*

This is a representative procedure for aldehyde HCO-WWRR: The HPLC purified aldehyde from the previous step was dissolved water (10.3 μL, 7.7 mM, 1 eq) was diluted into 79.3 μL of MOPS buffer (200 mM, pH 6.5). The reaction was initiated by adding by the addition of YEB in acetonitrile (38.7 mM, 10.2 μL, 38.7 mM, 5 eq) was sampled at 5 time points, 1 M HCl (1 μL) was added to quench the reaction and analyzed by ESI-LC-MS to calculate the percentage conversion. Integration of the UV absorbance peaks corresponding to the YEB peak and the Wittig product peptide peak were compared to monitor the rate of product formation.

All reactions were performed under pseudo-first-order condition. Fitting of the kinetic curve to the equation $A_t = 1 - e^{-k*[YEB]*t}$ yielded the second-order rate constant k, where A_t is the absorbance at 214 nm for the product at time t and [YEB] is 4 mM. Curve fitting was done using MatLab scripts 95% upper and lower confidence interval limits.[1] Half time of the reaction ($t_{1/2}$) was calculated as $t_{1/2} = 1/k[A]_0$, where $[A]_0$ is the initial concentration of hydrazine used. The reactions were conducted on 100 μL scale to

measure the kinetics only. Rates for all reactions (± 95% CI) are summarized in Fig. 2 and supplementary table S1**.** At all measurements, the final percentage of acetonitrile was ≤ 2%, since rate and stereoselectivity of the Wittig reaction is highly dependent on the type of solvent.[4] Also, the volumes of aliquots were kept the same at all times to ensure all peptides were measured at the same pH. Self-catalysis and cross-catalysis tests were performed by same protocol.

Supplementary Table S1. List of aldehydes and reaction rates

Unique entries	Peptide	Reaction rate (k) $M^{-1}s^{-1}$	More info in Fig
1	HCO-WWRR	0.90±0.04	S7, S31
2	HCO-WWGP	0.57±0.09	S7, S32
3	HCO-WWPQ	0.56±0.01	S4, S35
4	HCO-WWGL	0.50±0.20	S4, S34
5	HCO-WLPR	0.21±0.01	S4, S36
6	HCO-LWYR	0.30±0.02	S4, S37
7	HCO-QWLH	0.29±0.06	S4, S33
8	HCO-WIVR	0.17±0.04	S4, S38
9	HCO-HWFP	0.17±0.04	S5, S39
10	HCO-ALRV	0.22±0.04	S5, S40
11	HCO-AAAP	0.22±0.04	S5
12	HCO-PRLP	0.17±0.07	S8, S45
13	HCO-PQPL	0.15±0.03	S8, S46
14	HCO-PYPA	0.13±0.03	S8, S49
15	HCO-APAA	0.10±0.03	S5, S41
16	HCO-PAAA	0.12±0.03	S8, S50
17	HCO-PPAA	0.020±0.002	S6, S42
18	HCO-PPPA	0.024±0.002	S6, S44
19	HCO-PPLA	0.017±0.005	S6, S43
20	HCO-PPPL	0.014±0.003	S6, S47
21	HCO-PPPP	0.020±0.003	S6, S48
22	HCO-WYFT	0.93±0.28	
23	HCO-LYAR	0.14±0.38	
24	HCO-WFFP	0.13±0.13	
25	HCO-WYAP	0.12±0.36	
26	HCO-VWTA	0.12±0.60	
27	HCO-FPWE	0.074±0.21	
28	HCO-GIIE	0.007±0.007	
29	HCO-RYIP	0.003±0.005	
30	HCO-YCKADC	0.44±0.16	S25
31	HCO-KCETFC	0.23±0.02	S25
32	HCO-QCYWRC	0.18±0.13	S25
33	HCO-QCYESC	0.14±0.033	S25
34	HCO-FCQGKC	0.15±0.003	S25
35	HCO-AA	0.093±0.02	
36	HCO-SarSar	0.021±0.01	

1.4 *Hydrazone ligation experiments*
This is a representative procedure for aldehyde HCO-WWRR: 2,4-dinitrophenylhydrazone solution (0.1 mg in 1.1 mL of concentrated H_2SO_4, 5.1 mmol) was slowly added to a mixture of 5.7 mL EtOH (95%)/1.6 mL H_2O to generate a 61 mM solution of hydrazine. HCO-WWRR solution (10.4 µL, 7.7 mM, 1.0 eq.) were mixed with 73.5 µL of H_2O, 10 µL of EtOH were added and then of hydrazine solution (6.5 µL, 61 mM, 5 eq.) were added. The reaction was sampled at 5-6 time points, an aliquot of the reaction (2 µL) was taken out and quenched by dilution with 198 µL of water and analyzed by ESI-LC-MS to calculate the percentage conversion. Quenching was performed at the times specified in Fig S14-S17. An aliquot of each quenched solution was injected into UHPLC-MS to obtain traces. Area percentages were extracted and the pseudo first order rate constant and kinetic traces were obtained by using the MATLAB script.

1.5 *In situ E/Z selectivity determination*
This is a representative procedure for aldehyde HCO-WWRR: HCO-WWRR aliquot (51.7 µL, 7.7 mM, 1.0 eq, final concentration 0.8 mM) were mixed with 391.6 µL of PB 200 mM at pH 6.5 in an NMR tube. 55 µL of D_2O were added. YEB solution (51.7 µL, 38.7 mM, 5.0 eq, final concentration 4.0 mM) was added. The final volume was 550 µL (D_2O 10%). Solvent suppression was performed and ^1H NMR spectra were collected at the specified time intervals.

2. Biochemistry methods

2.1 *General biochemistry information*
The SX_4 and $SXCX_3C$ library was bulk-amplified from libraries generated in previous report[5, 6] and the phage clone that displays SWYD peptide on its surface and reporter neon green clones for internal control (Fig S3) were produced as described in our previous report.[6] The sequencing files were uploaded to https://48hd.cloud/ and the links are attached below.
SX_4 library: https://48hd.cloud/file/20150201-57OOneOO-RD
$SXCX_3C$: SB1-SxCxxxC http://www.48hd.cloud/file/20170228-22OOooOS-HD

2.2 *Phage functionalization*

2.2.1 *$SXCX_3C$ and SX_4 1% biotinylation and purification*
10^{12} pfu of library were taken to a final volume of 100 µL with MOPS (200 mM). To this solution, 1 µL of freshly prepared $NaIO_4$ (6 mM in water) was added and the mixture was incubated on ice for 8 minutes. Then, 1 µL of methionine (50 mM in water) was added and the mixture was incubated for 20 minutes at room temperature. 100 µL of YEB (0.8 mM in MOPS 200 mM) were added and the reaction was incubated for 10 minutes at room temperature. After this, 800 µL of acetate buffer (200 mM) and 200 µL of PEG NaCl were added and the mixture was kept at 4 °C for at least 12 hours. After this, the sample was centrifuged at 21100 xg for 30 minutes and the supernatant was discarded. The sample was further centrifuged for 5 minutes and remaining supernatant was removed. The phage pellet was re-suspended in 200 µL of acetate buffer (10 mM) and incubated at room temperature for 15 minutes to allow phage to completely dissolve. After this, remaining solid debris was removed by centrifugation at 21100 xg for 5 minutes. The supernatant was transferred to a clean epi-tube and kept at 4 °C for use in pull-down screenings.

2.2.2 *$SXCX_3C$ and SX_4 0.1% biotinylation and purification*

10^{12} pfu of library were taken to a final volume of 100 µL with MOPS 200 mM. To the solution, 1 µL of freshly prepared $NaIO_4$ (6 mM in water) was added and the mixture was incubated in ice for 8 minutes. Then, 1 µL of methionine (50 mM in water) was added and the mixture was incubated for 20 minutes. 100 µL of YEB (0.8 mM in MOPS 200 mM) were added and the reaction was incubated for 10 minutes at room temperature. After this, PEG purification was performed as indicated in section 2.2.1 and the phage was stored at 4 °C, ready to be used in screenings.

2.2.3 *$SXCX_3C$ and SX_4 oxidations and purification*

10^{12} pfu of library were taken to a final volume of 100 µL with MOPS 200 mM. To the solution, 1 µL of freshly prepared $NaIO_4$ (6 mM in water) was added and the mixture was incubated in ice for 8 minutes. Then, 1 µL of methionine (50 mM in water) was added and the mixture was incubated for 20 minutes. After this, PEG purification was performed as indicated in section 2.2.1 and the phage was stored at 4 °C, ready to be used in screenings.

2.3 *Phage pull-down experiments*

2.3.1 *1% and 0.1% SX_4 and $SXCX_3C$ selections*

The sets X, Y, Z illustrated in Fig 1e or Fig S3 were prepared as follows.

Set X: ~10^9 pfu of biotinylated, purified library were mixed with ~10^5 pfu of internal control mixture neon green phage (in the case of 1% selections) and the volume was taken to 1 mL with BSA 2% in acetate buffer in a 1.5 mL microcentrifuge tube.

Set Y: ~10^9 pfu of biotinylated, purified library were mixed with ~10^5 pfu of internal control mixture (in the case of 1% selections) and the volume was taken to 1 mL with BSA 2% + 0.9 mM biotin in acetate buffer in a 1.5 mL microcentrifuge tube.

Set Z: ~10^9 pfu of oxidized, purified library were mixed with ~10^5 pfu of internal control mixture neon green phage (in the case of 1% selections) and the volume was taken to 1 mL with BSA 2% in acetate buffer in a 1.5 mL microcentrifuge tube.

All mixtures were prepared in triplicates. 100 µL of streptavidin magnetic beads per each sample (9 in total) were washed three times with 1 mL of acetate buffer (10 mM) and suspended in 1 mL of BSA 2% in acetate buffer.

All mixtures and bead suspensions were incubated for 30 minutes at room temperature in a Labquake Tube Rotator. After this, 10 µL aliquots of each phage mixture (9 tubes in total) were taken for tittering and PCR of inputs. The bead suspensions (9 tubes in total) were placed in a magnet and the BSA supernatants were discarded. Each phage mixture (990 µL each one) was added to one corresponding tube of beads and incubated at room temperature in a Labquake Tube Rotator for 1 hour. The unbound phage was washed from the beads with Tween 0.1% in acetate buffer (10 washes of 1 mL each one) by using a Kingfisher magnetic bead washer (ThermoFisher Scientific).

The washed beads were suspended in acetate buffer (10 mM), placed in a magnet and the supernatant was discarded (this step was necessary to remove any remaining Tween that might interfere with PCR). The rinsed beads were suspended in 20 µL of NaOH solution (10 mM) and incubated at room temperature for one hour in a Labquake Tube Rotator. After that, suspensions were placed in a magnet and the supernatant was transferred to a 0.6 mL microcentrifuge tube containing 20 µL of 1X Phusion® high fidelity (HF) buffer to give a final volume of 40 µL of phage elution. 2 µL are taken for tittering of this elution and the rest is stored 4 °C for use in PCR (Section 2.4).

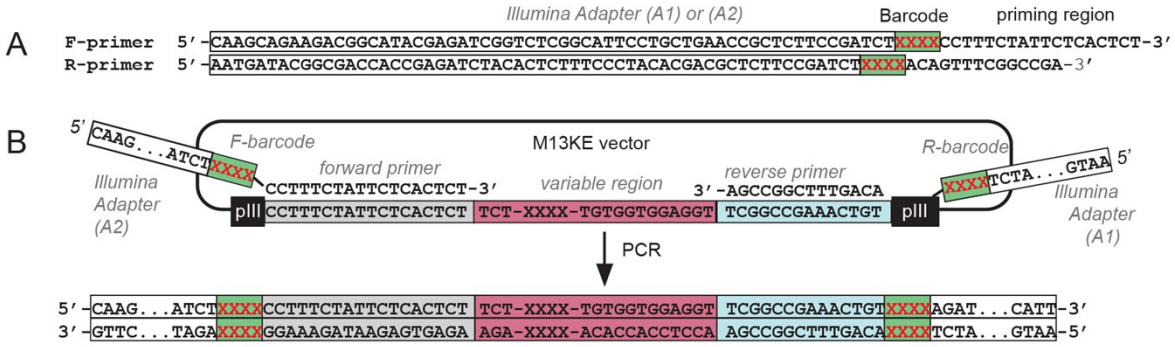


Figure S1: DNA sequences of PCR amplification protocol for Illumina deep sequencing (A) Primers used for amplifying ligated or naïve oligonucleotide DNA. XXXX denotes 4-nucleotidelong barcodes used to trace multiple samples in an Illumina sequencing experiment. (B) Generation of PCR product. Alignment of forward and reverse primers to 18-bp and 14-bp sequences flanking the variable region at the N-terminus of the pIII gene in M13KE vector, respectively.

2.4 *PCR of phage*

2.4.1 *PCR of NaOH elutions*

The 20 µL of DNA template solution in Nuclease free water was amplified in total volume of 60 µL. A typical 60 µL reaction mixture contained:

1. 5× Phusion buffer 10 µL
2. 10 mM dNTPs 1 µL
3. Phusion® Polymeras 1 µL
4. Forward primer (10 µM) 2.5 µL
4. Reverse primer (10 µM) 2.5 µL
6. Template solution 20 µL
7. Nuclease free water 23 µL

Cycling was performed using the following thermocycler settings:

a) 95°C 30 s
b) 95°C for 10 s
c) 60.5 °C 15 s
d) 72 °C 30 s
e) repeat b)-d) for 25 cycles
f) 72 °C 5 min
g) 4 °C hold

The PCR products for each elution were mixed before gel electrophoresis characterization.

2.4.2 *PCR of inputs*

The 5 µL of DNA template solution in Nuclease free water was amplified in total volume of 50 µL. A typical 50 µL reaction mixture contained:

1. 5x Phusion buffer 10 µL
2. 10 mM dNTPs 1 µL
3. Phusion® Polymeras 1 µL
4. Forward primer (10 µM) 2.5 µL
4. Reverse primer (10 µM) 2.5 µL

6. Template solution 5 µL
7. Nuclease free water 28 µL

Cycling was performed using the following thermocycler settings:
a) 95°C 30 s
b) 95°C for 10 s
c) 60.5 °C 15 s
d) 72 °C 30 s
e) repeat b)-d) for 25 cycles
f) 72 °C 5 min
g) 4 °C hold

3. Data analysis

3.1 *General data processing methods*

Data analysis was performed in MATLAB. Core scripts are available as part of the Supplementary Data.rar. Data storage cloud http://48hd.cloud/ was implemented in Linux-Apache-MySQL-Python (LAMP) architecture and details of this implementation are beyond the scope of this report. Prior to analysis, "test" and "control" datasets were retrieved from the http://48hd.cloud/ server as tables of peptides, DNA, and raw sequencing counts and combined into a master table (VT_unfiltered_Feb.txt) which is proved in Supplementary Data.rar.
The 20×20 plots were previously reported by our lab[5, 7] and were generated by the MATLAB scripts (plot20x20_generation.m and plot20x20_generation_SXCXXXC in Supplementary Data.rar)

3.2 *Processing of illumina data*

The Gzip compressed FASTQ files were downloaded from BaseSpace™ Sequence Hub. The files were converted into tables of DNA sequences and their counts per experiment. Briefly, FASTQ files were parsed based on unique multiplexing barcodes within the reads discarding any reads that contained a low-quality score. Mapping the forward (F) and reverse (R) barcoding regions allowing no more than one base substitution each and F-R read alignment allowing no mismatches between F and R reads yielded DNA sequences located between the priming regions. The files with DNA reads, raw counts, and mapped peptide modifications were uploaded to http://48hd.cloud/ server. Each experiment has a unique alphanumeric name (e.g., 20170829-09WIooPA-VT) and unique static URL: for example https://48hd.cloud/file/777)
URL links to sequencing data used by plot20x20_generation.m and plot20x20_generation_SXCXXXC.m to generate SX_4 and $SXCX_3C$ plots could be found in **Supplementary Table S2** and **S3**.

Supplementary Table S2

Set	type	File name	URL
Z unmodified	input	20170829-09OOooPA	https://48hd.cloud/file/775
	output	20170829-09OOsaSP	https://48hd.cloud/file/775
X modified	input	20170829-09WIooPA	https://48hd.cloud/file/777

capture	output	20170829-09WIsaSP	https://48hd.cloud/file/778
Y modified with blocked beads	input	20170829-09WIooPA	https://48hd.cloud/file/777
	output	20170829-09WIsaSP	https://48hd.cloud/file/778

Supplementary Table S3

Set	type	File name	URL
Z unmodified	input	20170829-22OOooPA	https://48hd.cloud/file/793
	output	20170829-22OOsaSP	https://48hd.cloud/file/794
X modified capture	input	20170829-22WIooPA	https://48hd.cloud/file/795
	output	20170829-22WIsaSP	https://48hd.cloud/file/849
Y modified with blocked beads	input	20170829-22WIooPA	https://48hd.cloud/file/795
	output	20170829-22WIsaSP	https://48hd.cloud/file/849

The corresponding file names under VT1-VT36 in VT_unfiltered_Feb.txt used in plot20x20_generation.m can be found in Supplementary Data.rar as VT_unfiltered_Feb_naming.xlsx. Barcodes used for sequencing (Fig S1) are listed for each file can be used to track individually.

4. DFT computation

All the optimized structures of reactants, products and transition state were obtained using density functional theory (DFT). The geometries were optimized in the gas phase using the B3LYP[8-11] function with 6-31G(d) basis set.[12, 13] All the optimizations were performed using default convergence criteria. The initial structure of TS1 and TS2 was built based on the structure from Robiette *et al*.[4] To confirm the TS conformation minima or first-order transition states, to determine Gibbs free energies (at 298 K), vibrational frequencies were computed for all the optimized structures. The Gaussian 09 suite of software[14] was used to perform all the electronic structure computations. All the output files are attached in Supplementary Data.rar.

5. Machine Learning

Feature engineering: The DC values for each tetramer peptides were used as input data to the MATLAB script in Supplementary Data.rar as MakeMLinput.m. Algorithm info is available at https://github.com/derdalab/GESAR. Quantitative chemical properties of each amino acid in the tetramer sequences were added as new columns, specifically z-scale descriptors (3×4 AA position = 12 features)[15] and VHSE (Vectors of Hydrophobic, Steric, and Electronic properties) descriptors (8×4 AA position = 32 features)[16] to the table.

Apart from the chemical properties, we added sequence patterns based on permutations of 20 amino acids among 4 positions within the tetramer sequences. We used the following strategy to generate features

based on sequence patterns. First, we computed the cartesian products with 1, 2 or 3 elements with repetitions as illustrated below:

- 3 elements: XXX, XXY, XYZ.. [8000]
- 2 elements: XX, XY, YZ.. [400]
- 1 element: X, Y.. [20]

Then, we generated sequence patterns (example as followed), X denotes any of the 20 amino acids

- 3 element: AAAX, AAXA, AXAA, XAAA [4 × 8000]
- 2 element: AAXX, AXXA, XXAA, AXAX, AXXA, XAXA [6 × 400]
- 1 element: XXXA, XXAX, XAXX, AXXX [4 × 20]

Then, we created features based on the lookup of these patterns. If a pattern is present in a tetramer instance, we give a value of 1 or 0. This procedure yielded 34,480 patterns, of which 2,396 were finally retained by setting a threshold of at least 20 tetramer instances that match a particular pattern.

We split the data in to HIGH or LOW deep conversion subgroups where HIGH and LOW are defined by the highest and lowest 5% of the Log_2DC values obtained by experimental observation (Figure S5). This means that there were two splits of the data where the first split was made up of HIGH (top 5%) and NOT-HIGH (95%), and the second was made up LOW (Bottom 5%) and NOT-LOW (95%). We used these class labels as our target labels to train our models. The scripts used to create these labels and pre-process the data can be found in the github repo under Final_evaluation_CLF_Models.ipynb file.

Training: Using a gradient boosted ensemble of decision trees (XGBoost[17]), we trained two binary classifiers: one to identify sequences belonging to the HIGH subgroup and the other to the LOW subgroup. The two splits of the data created earlier were used to train two independent classifiers. 5-fold stratified cross-validation was used to evaluate the performance of our models. When one of the five groups is used to evaluate the model while the other four groups are used to train the model.

Evaluation: The average evaluation scores collected from each reshuffle of the data are used to evaluate the performance of the model. Stratified 5-fold includes splitting the data so that class distribution is preserved across folds. We evaluated our models on several metrics including area under the receiver operating characteristic curve (ROC AUC) Scores, 81.2 ± 0.4 and 73.7 ± 0.8 for HIGH and LOW, Accuracy, F1-Scores, Precision and Recall (Supplementary table S4).

Prediction: To predict the sequences not observed in our experimental dataset, we created a new dataset that consists of about 100,000 sequences and their corresponding features. We computed these features using the process specified in feature engineering. Using our two independent classifier models, we predicted two independent probabilities for each of these sequences: probability of sequence belonging to LOW class and to HIGH class. We then set a threshold of probability = 0.5 to decide if a sequence belongs to the corresponding class or not, which means a sequence can be labelled as HIGH or NOT-HIGH (LOW or NOT-LOW for the other). Setting different threshold values can change the F1-score for each of the classifiers as seen Figure S13. The threshold is a trade off between precision and recall. A higher threshold allows us to minimize false positives while a lower threshold minimizes false negatives. Setting the threshold at 0.5 allows us to have a high confidence in predictions and maximize the F1-score for classifiers. I.e if a sequence had a probability of greater than or equal 0.5 for HIGH and less than 0.5 for low, we labelled it as HIGH. If it had a probability of greater than or equal to 0.5 and less than 0.5 for HIGH, we labelled it as low. If a sequence had any other combination of probabilities, we labelled it as the medium (middle 90%).

We used these labels to create a 20×20 plot that allows us to visualize the predictions of about 100,000 amino acid sequences (Fig. 4a). The 20×20 plots were created by a Python script can be found under helper_functions.py in our Github repository.

Deployment: We deployed the saved models into a public web app available at http://44.226.164.95/ which allows users to get the probability of sequences belonging to both the HIGH and LOW class as well as their position within the 20×20 plot of the observed data.

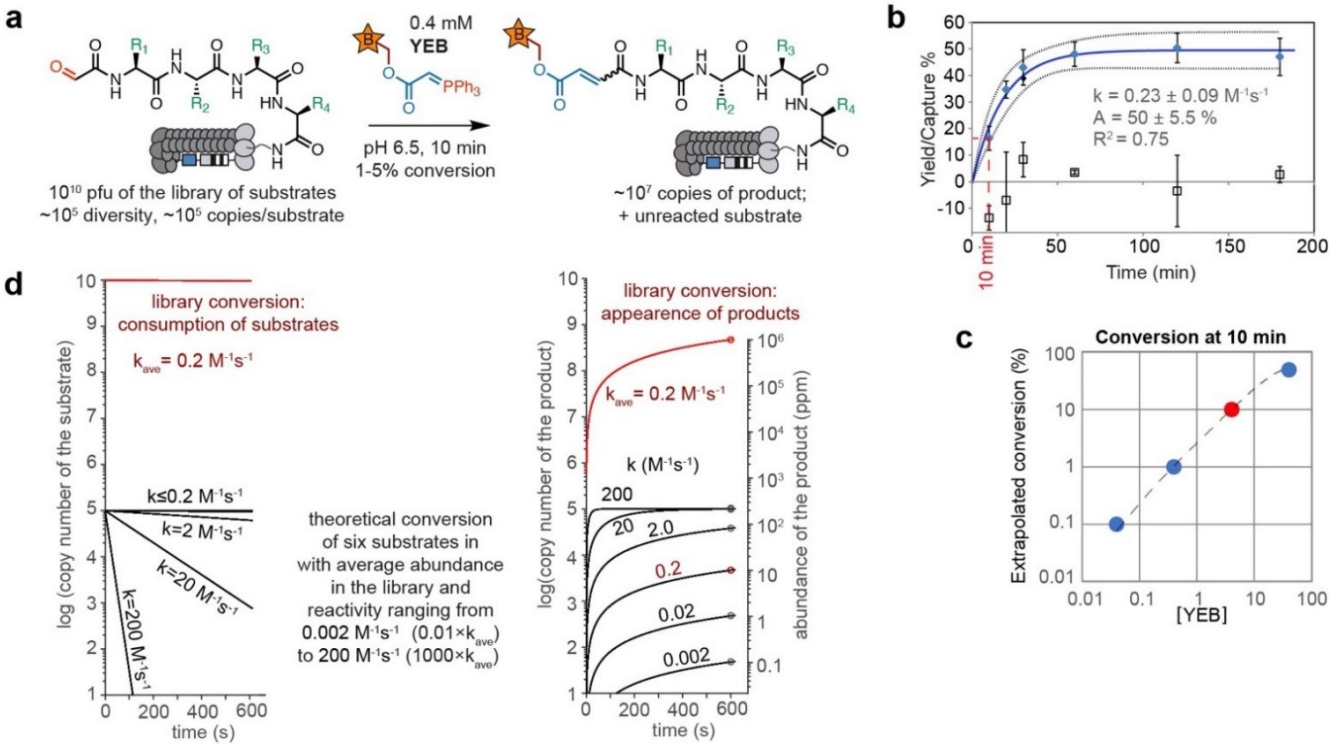
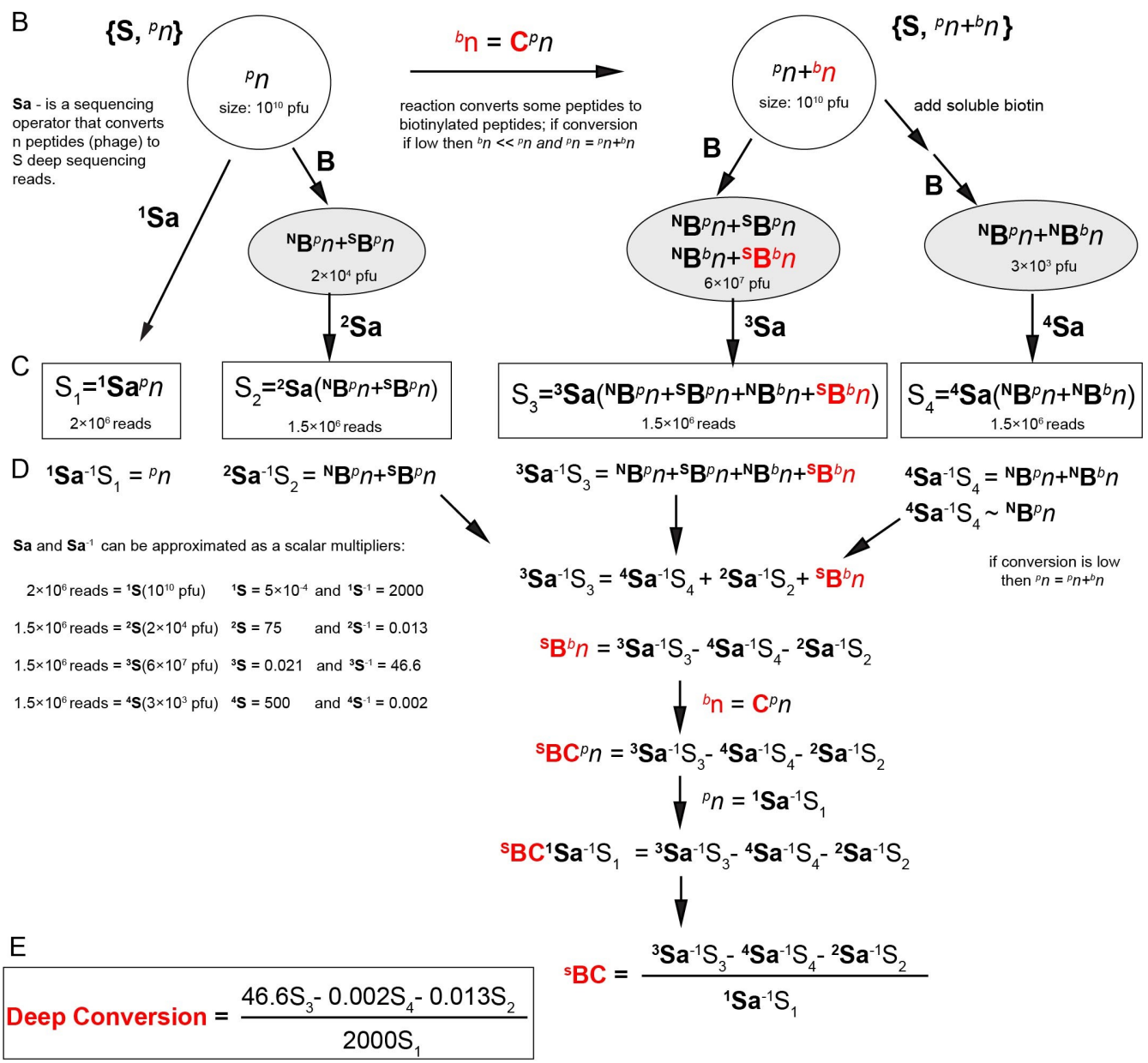


Figure S2. Analysis of the Wittig reaction in a population of peptide aldehydes.

(a) Scheme of the reaction oh phage library. (b) Population reaction rate measured by biotin capture and reported in our previous publication.[1] (c) The experimentally measured population rate constant can be used to extrapolate the conversion of the population at 10 minutes in the presence of different concentrations of YEB. (d) Individual members of the library react at rates that are faster or slower than the average population rate. We modeled both disappearance of the starting material and appearance of the products in such kinetically heterogeneous population. Specifically, we used a realistic input: a 10^{10} total number of the phage in the library. To simplify the visualization, we assumed that every member of the library is present in the same concentration. In such input, each phage clone is present at a copy number 10^5 and conversion of all clones to the product can be easily followed. After 10 minutes of reaction, it is possible to use the relative abundance of the products to distinguish the substrates with rate constants from 0.01 to 100 of the average population rate of 0.2 $M^{-1}s^{-1}$. Detection of substrates slower than 0.002 $M^{-1}s^{-1}$ is difficult because the copy numbers of the products approach single digits and their relative abundance is <0.1 parts/million. Detection of such product by deep sequencing requires significant depth (tens of millions of reads) and it would be hampered by sequencing noise. Differentiation of substrates faster than 20 $M^{-1}s^{-1}$ is not possible because they all reach completion under these conditions. Fast reactivities can be differentiated by shifting the detection time to the earlier time points; however, this would make it impossible to measure the slow reactivity. It is simple to show that the dynamic range of the detection—0.01 to 100 times of the average population rate—represents the upper theoretical limit at the 10^{10} pfu input and 10^6-10^7 read depth. To improve it, one must increase the number of phage clones in the input beyond 10^{10} pfu and increase the depth of sequencing beyond several million sequencing reads.



Scheme S1. Analysis of the relationship between modification, panning and deep sequencing. A) Definitions. B) Library of peptides can be represented as a multiset or tuple of sequence list (ordered set) and copy-vector that describes a copy number for each sequence as integers (0, 1, 2 .). Operations applied to the library such as **B**inding (to streptavidin beads), **C**onversion (to biotinylated peptide) and **Sa**mpling of the library for sequencing are described as operators **B**, **C**, **Sa** acting on the copy-vector.[18] Under the assumption that the operation (reaction, binding, sequencing) applied to one peptide does not interfere with the operation on another peptide, all operators are simple diagonal matrices. Circles

represent phage libraries in solution and grey ovals represent libraries that bound to streptavidin beads and then were eluted by NaOH treatment. C) Sequencing of four different libraries produce four sequencing multisets described as rectangles. Just like phage library, a sequencing multiset is an ordered list of sequences and a matched vector of integer copy number (S_1-S_4). These S_1-S_4 vectors are copies of peptides observed in the sequencing. D) A hypothetical inverse sequencing operator $\mathbf{Sa^{-1}}$ can be introduced to derive a deep conversion vector (E) which describes the conversion of all peptide sequences observed in the sequencing vectors S_1-S_4.

Note 1: To allow the last step in the derivation (D), the operator $\mathbf{Sa^{-1}}$ was approximated as a scalar (number). This number is a ratio of the number of phage particles used in sequencing over the total number of reads observed in sequencing. The division of vectors in (E) and last step of (D) is element-wise division that yields another vector.

Note 2: realistic copy number of peptides (i.e., titer of phage particles in the libraries) and total sequencing depth (total number of reads) were used to illustrate the values for **Sa** and $\mathbf{Sa^{-1}}$ scalars.

Note 3: The derivation shows that the "deep conversion" vector is a combination of two factors: (i) conversion of peptide to biotinylated peptide described by diagonal elements of **C** operator and (ii) capture of individual biotinylated peptides by streptavidin beads described by diagonal elements of $^{S}\mathbf{B}$ operator. Under the assumption that all biotinylated peptides bind to streptavidin bead with the same efficiency, all diagonal elements of the $^{S}\mathbf{B}$ operator are the same number and $^{S}\mathbf{B}$ can be approximated as one number that describes the fraction of biotinylated library that was captured by streptavidin beads and subsequently eluted by NaOH.

Note 4: The value of the multiplication scalars in the final equation (E) show that contribution of sequencing data from datasets 2 and 4 are negligible. Conclusions are dominated by sequencing dataset 3 (capture of biotinylated peptides). The simplest explanation can be seen in the sheer size of the phage particles in the sample used to derive dataset 3 and factors of 1000-10000 decrease in the number of phage particles in the control samples 2 (capture of non-biotinylated peptides) and 4 (capture of biotinylated peptides by biotin blocked beads)

Note 5: any approximation of operators by scalars ignores random sampling events, important notion that sequence copy numbers are integers (it allows for <1 sequence to be considered) and it ignores sequence-specific events like PCR bias. It is not ideal, but it is a workable approximation for the purpose of this manuscript.

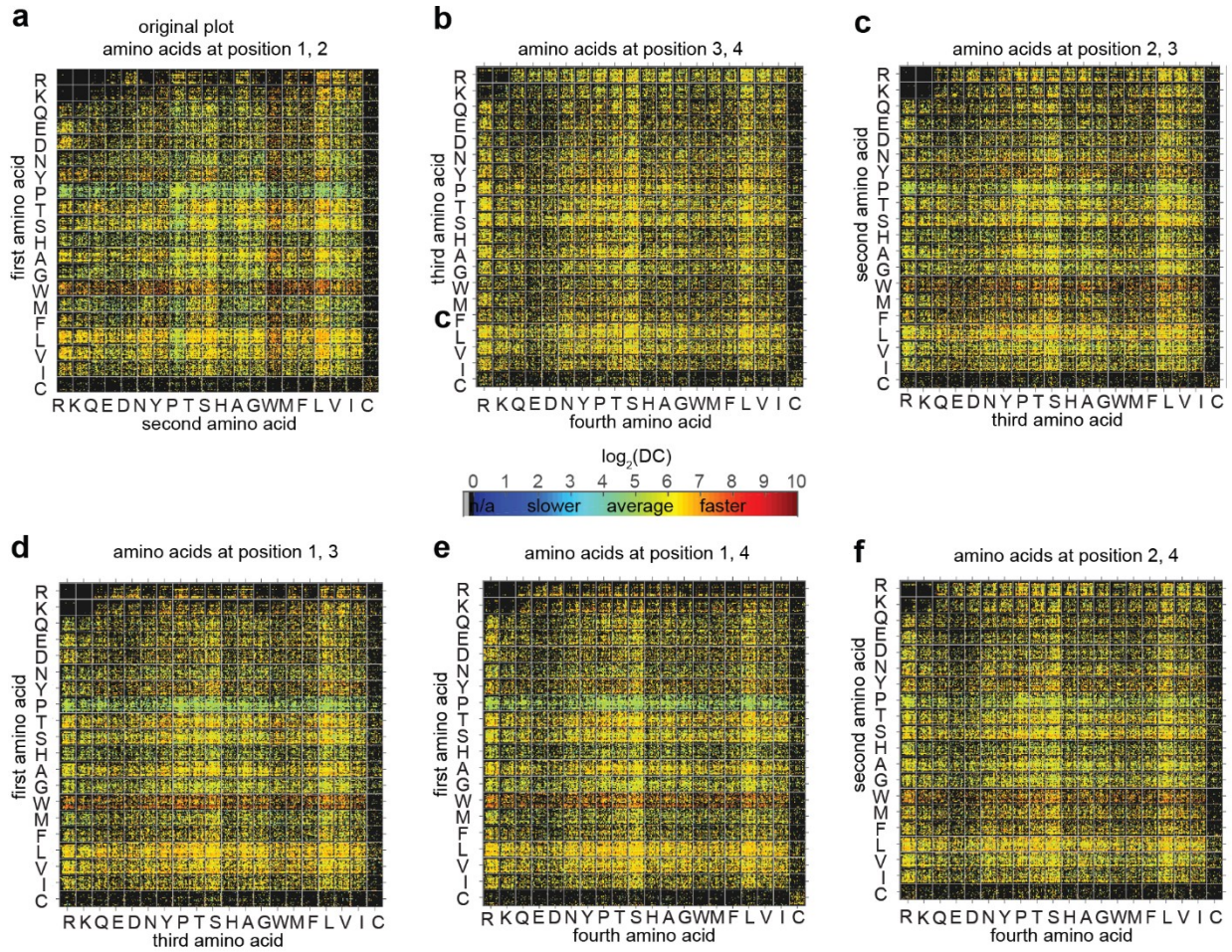


Figure S3. Comparison of 20×20 plots with different combinations of amino acid positions. (a) 20×20 plot duplicated from Fig. 1f. (b-f) 20×20 plots with different amino acid combinations generated from same data. It is clear that only with amino acids at first and second position, there is a clear pattern between amino acids and DC values.

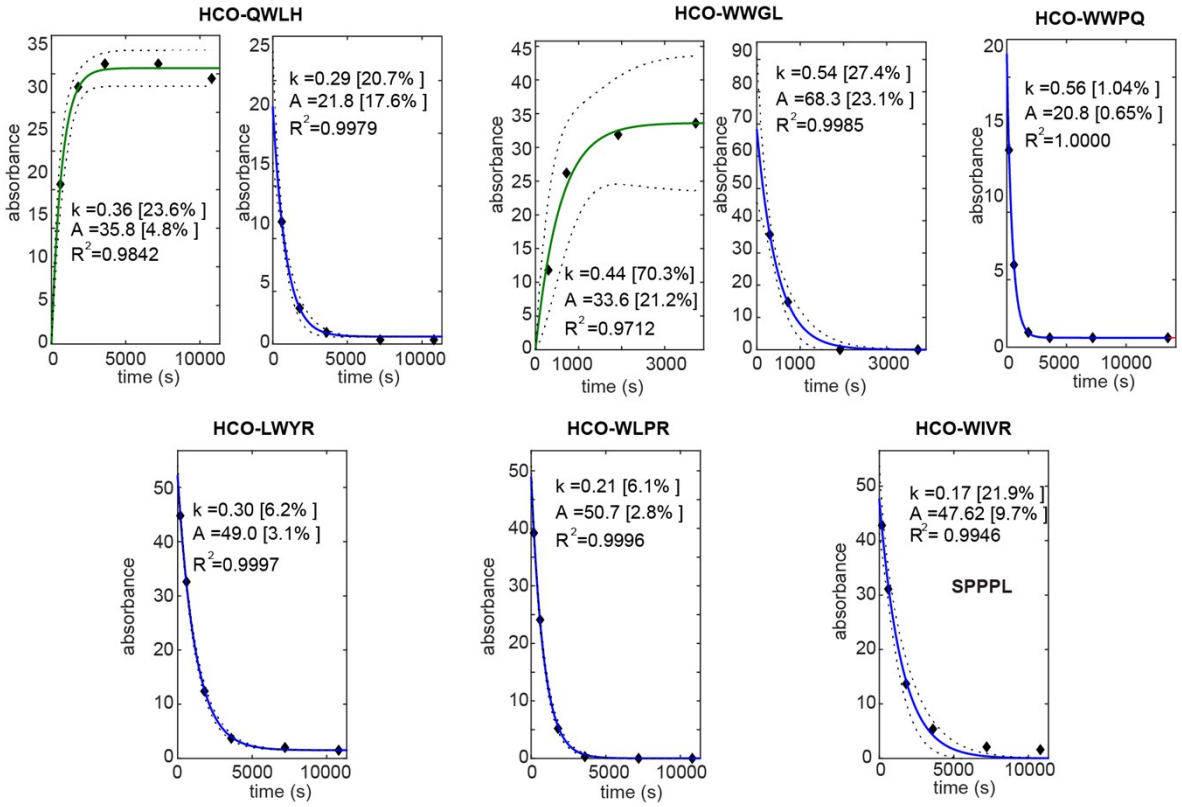


Figure S4. Experimental kinetic traces for HCO-X_4 peptides with high "Deep Conversion" values. Deep Conversion values of those peptides can be found in Supplementary Data.rar as MLinput.txt.

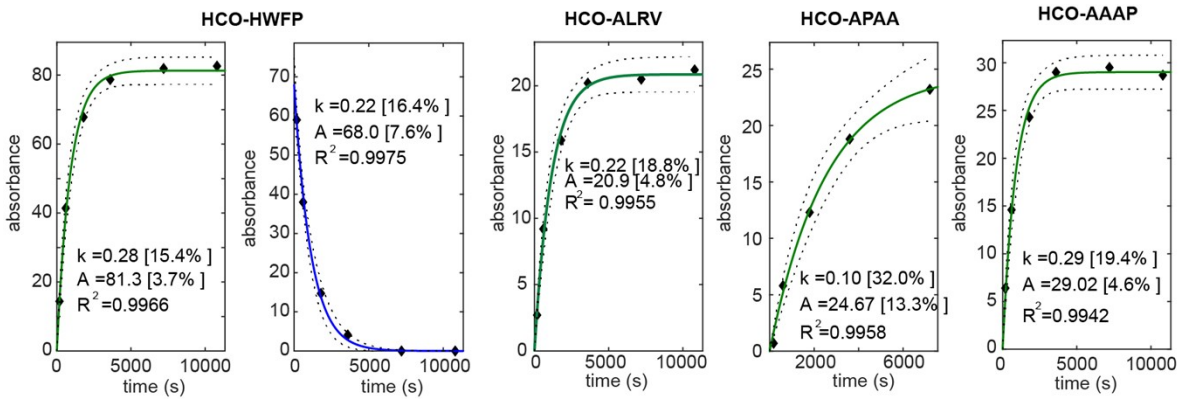


Figure S5. Experimental kinetic traces for HCO-X4 peptides with medium "Deep Conversion" values. Deep Conversion values of those peptides can be found in Supplementary Data.rar as MLinput.txt.

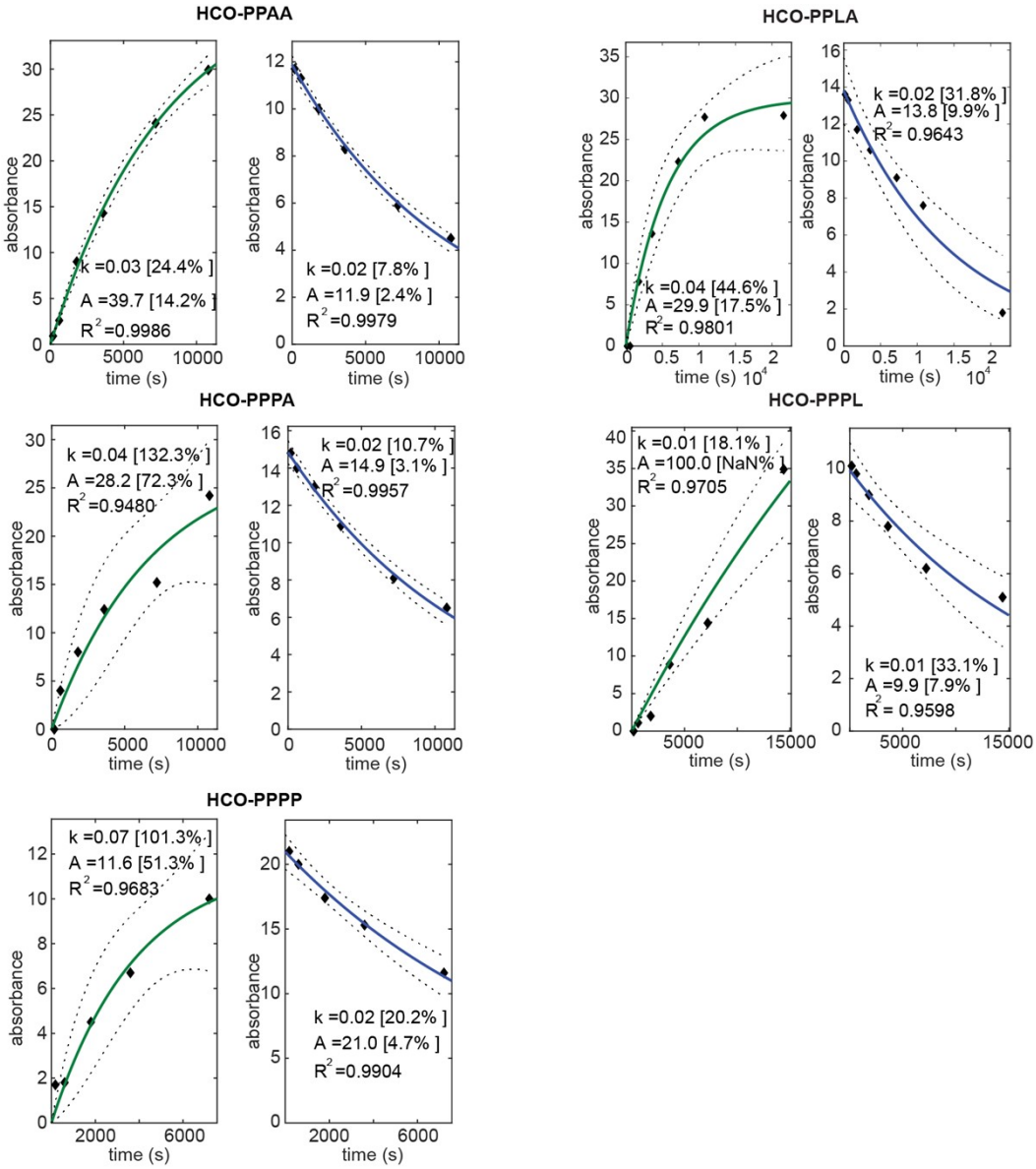


Figure S6. Experimental kinetic traces for HCO-X_4 peptides with low "Deep Conversion" values. Deep Conversion values of those peptides can be found in Supplementary Data.rar as MLinput.txt.

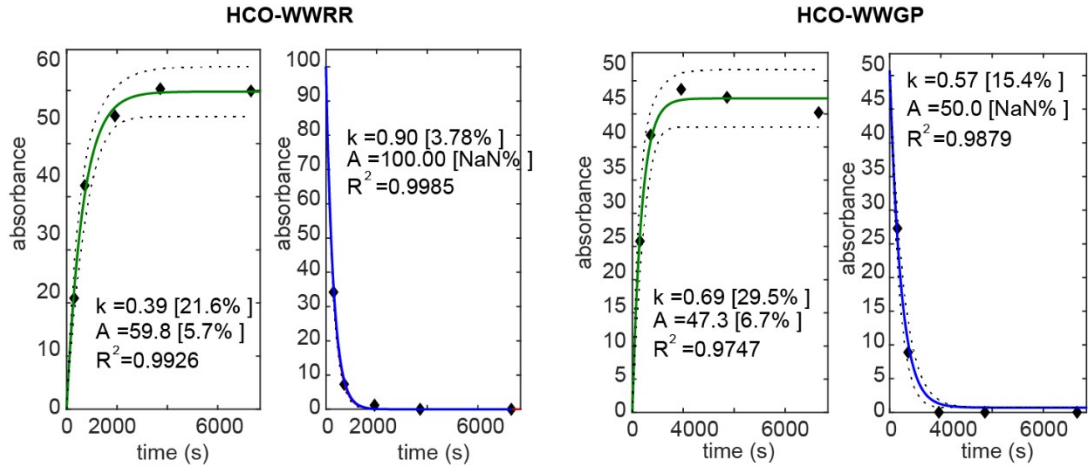


Figure S7. Experimental kinetics for HCO-WWXX with no Illumina counts.

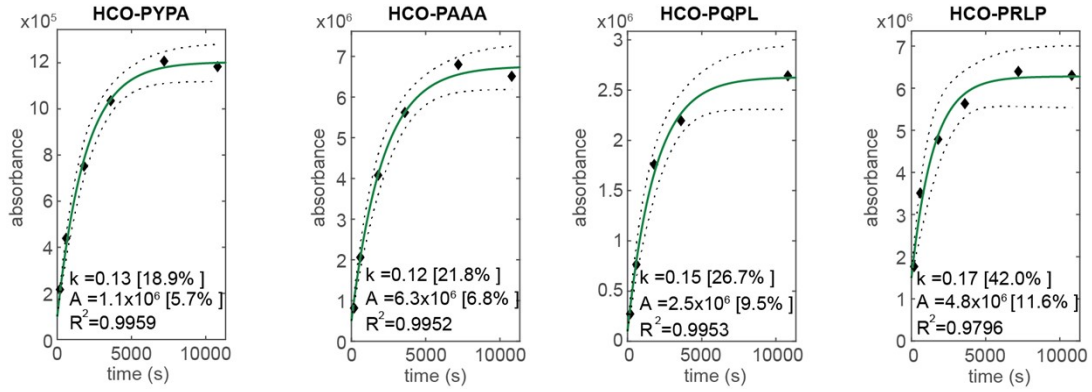


Figure S8. Experimental kinetics for HCO-PXXX sequences.

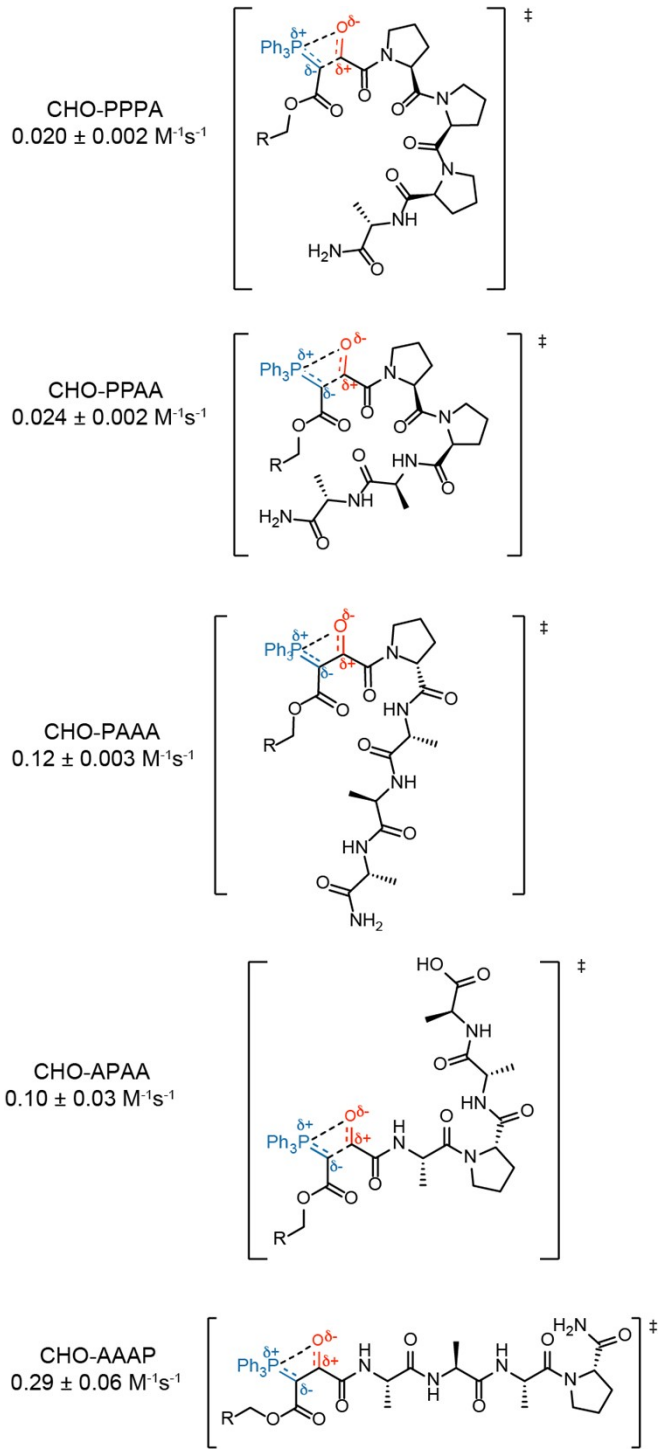


Figure S9. Proline-Alanine scan to determine position and quantity of hydrogen-bond donors for stabilization of OPA transition state.

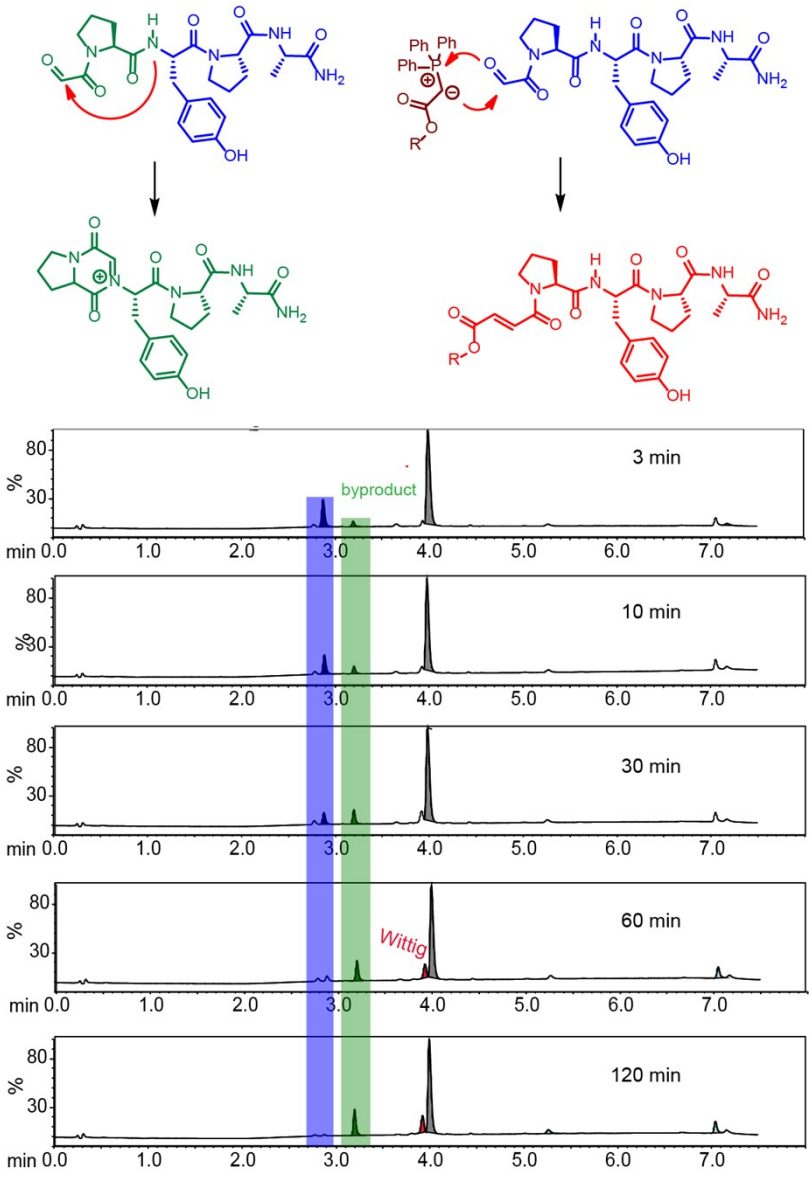


Figure S10. Example of SPXXX LCMS traces for determination of rate of intramolecular cyclization that produces a byproduct.

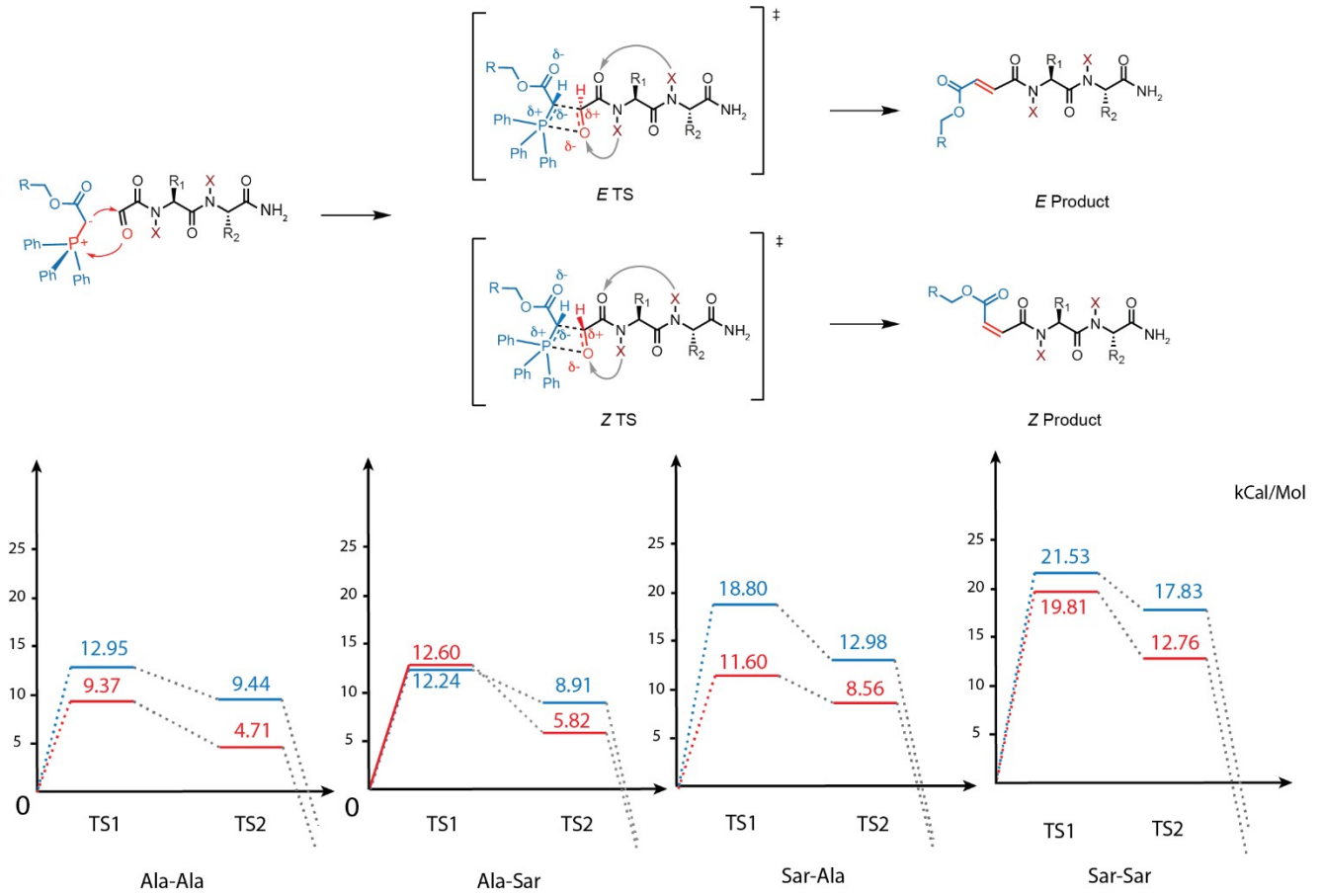


Figure S11. Gibbs free energy barrier of model peptides HCO-AlaAla, HCO-AlaSar, HCO-SarAla, HCO-SarSar (Sar = Sarcosine) as determined using B3LYP/6-31(g) in gas phase. E configurations are shown in red while Z configurations are shown in blue. Structures of TS1 are provided in Fig S12 and S13. Cartesian coordinates are provided in output files in Supplementary Data.rar as 'DFT calculation'.

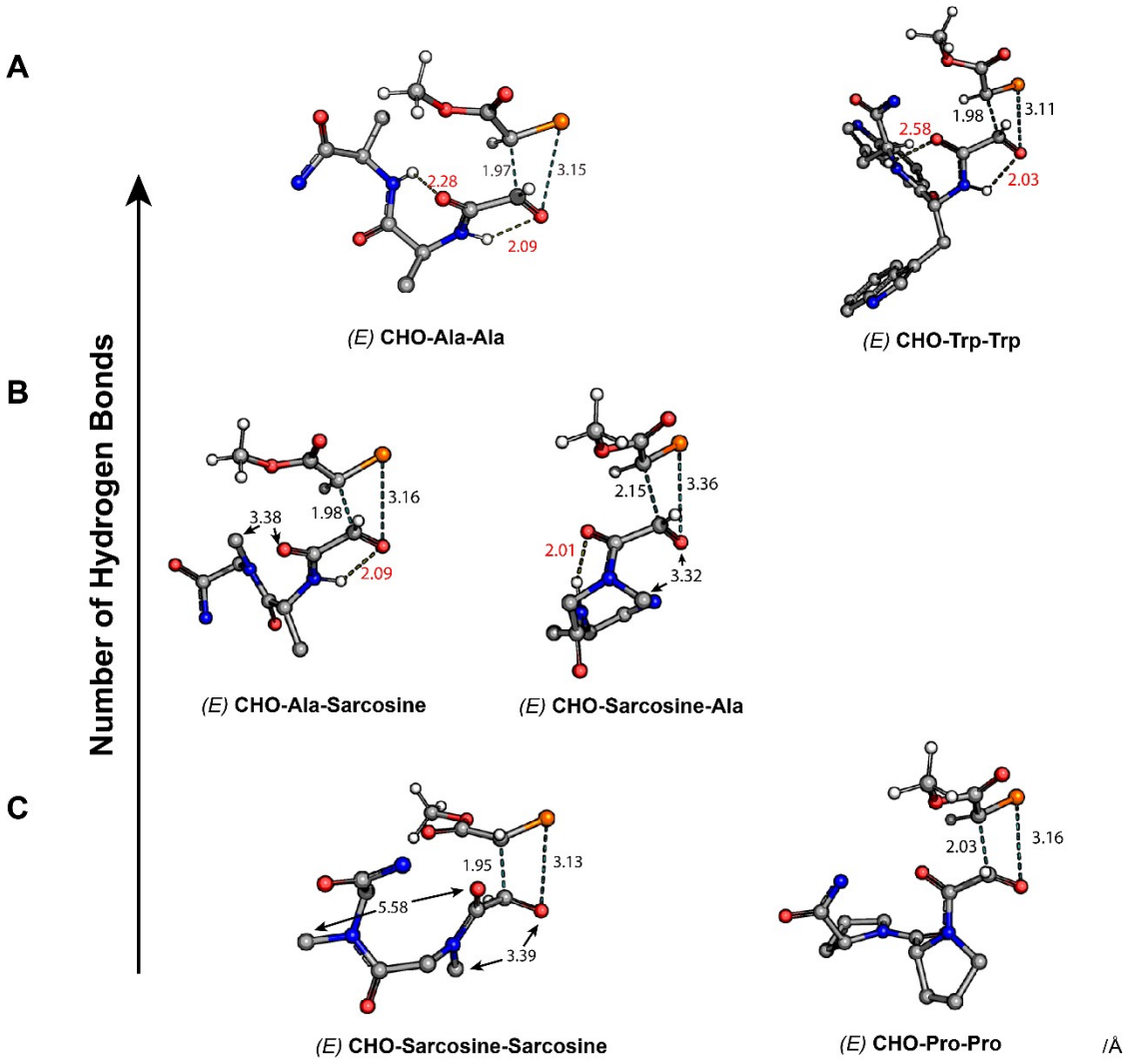


Figure S12. Geometries of *trans (E)* TS1 model peptides as determined using B3LYP/6-31(g) in gas phase. Important bond and distances are indicated, and Hydrogen bonds are labelled in red.

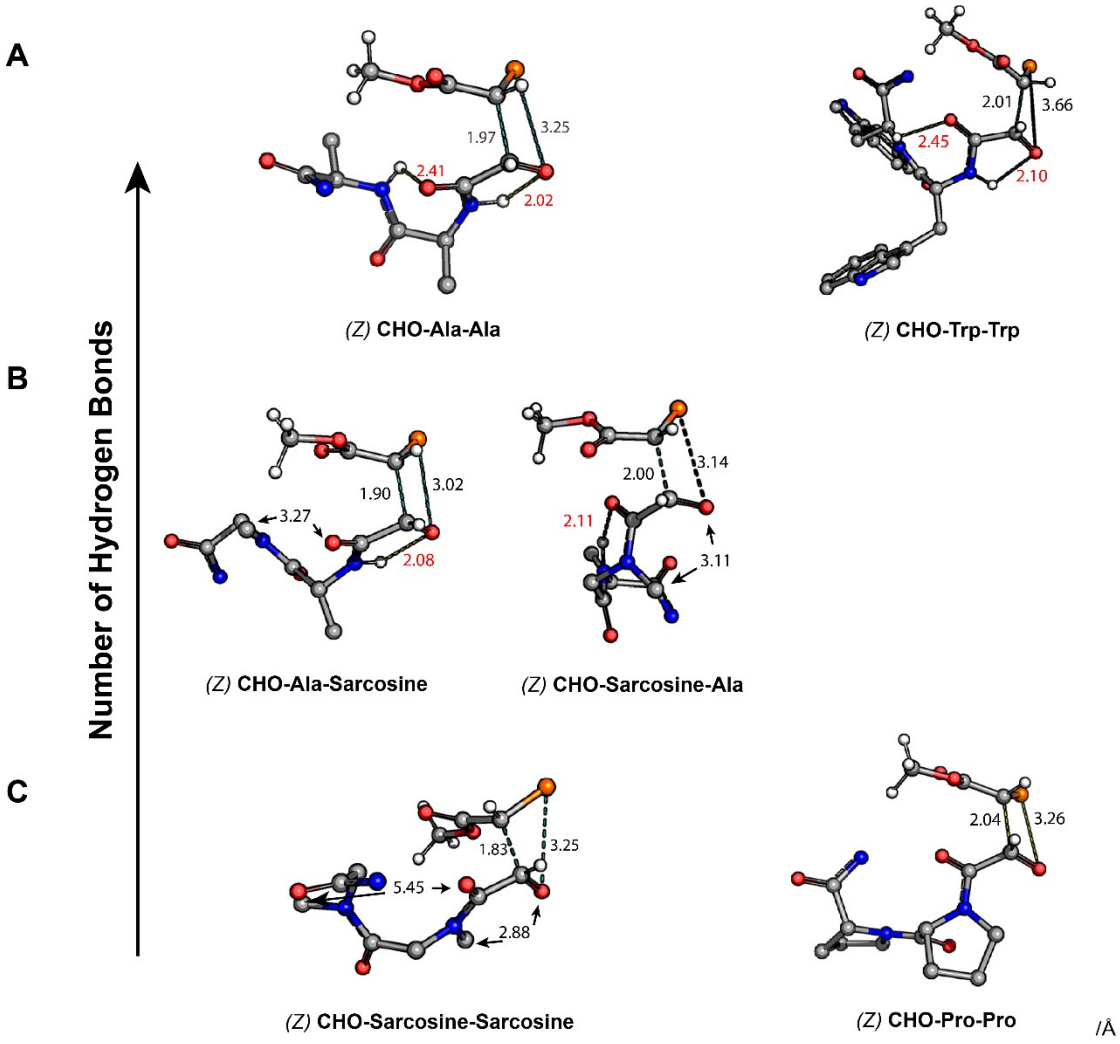


Figure S13. Geometry of *cis (Z)* TS1 model peptides as determined using B3LYP/6-31(g) in gas phase. Important bond and distances are indicated, and Hydrogen bonds are labelled in red.

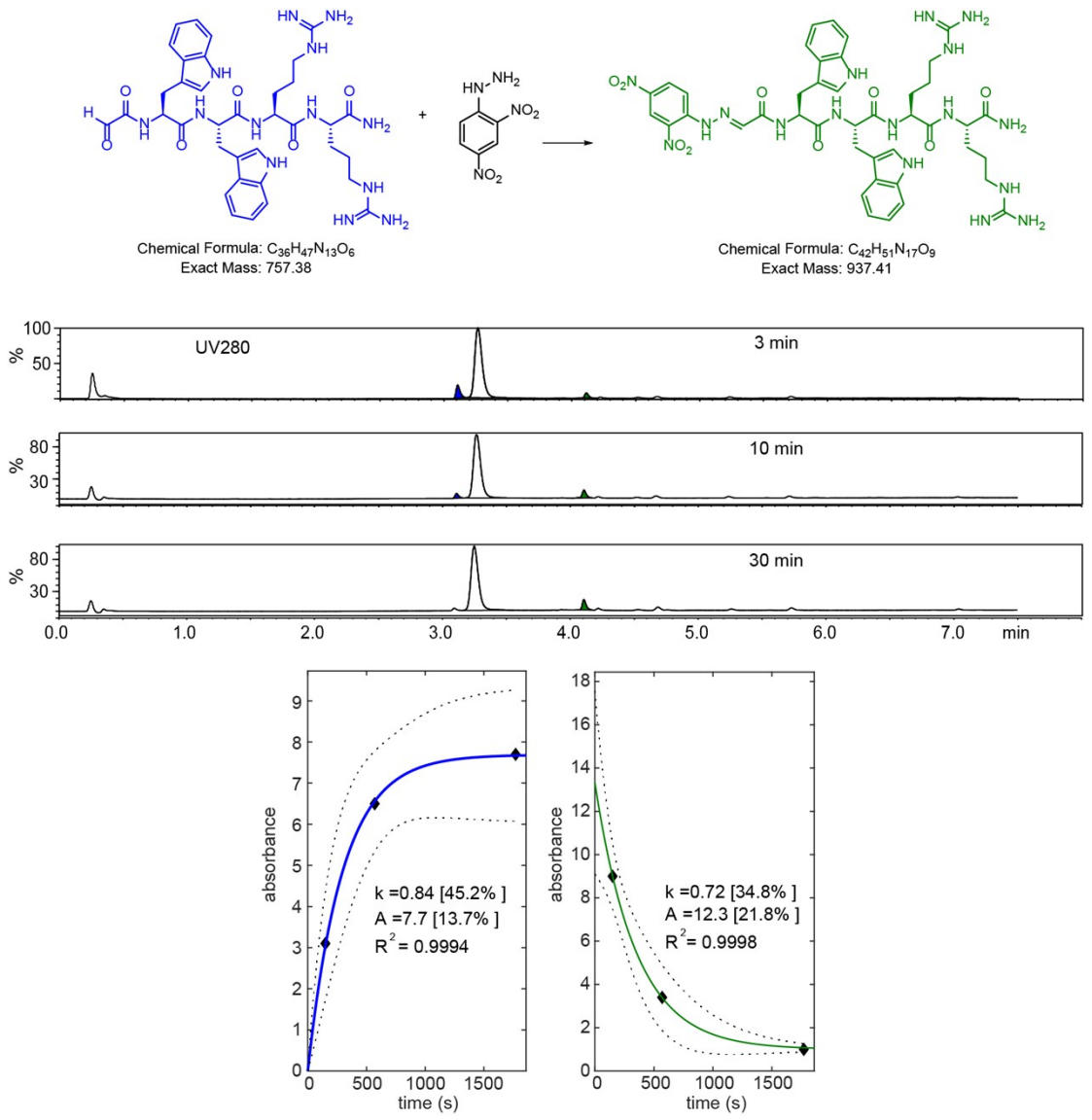


Figure S14. Kinetics of hydrazone ligation for HCO-WWRR at [H_2SO_4] = 65 mM.

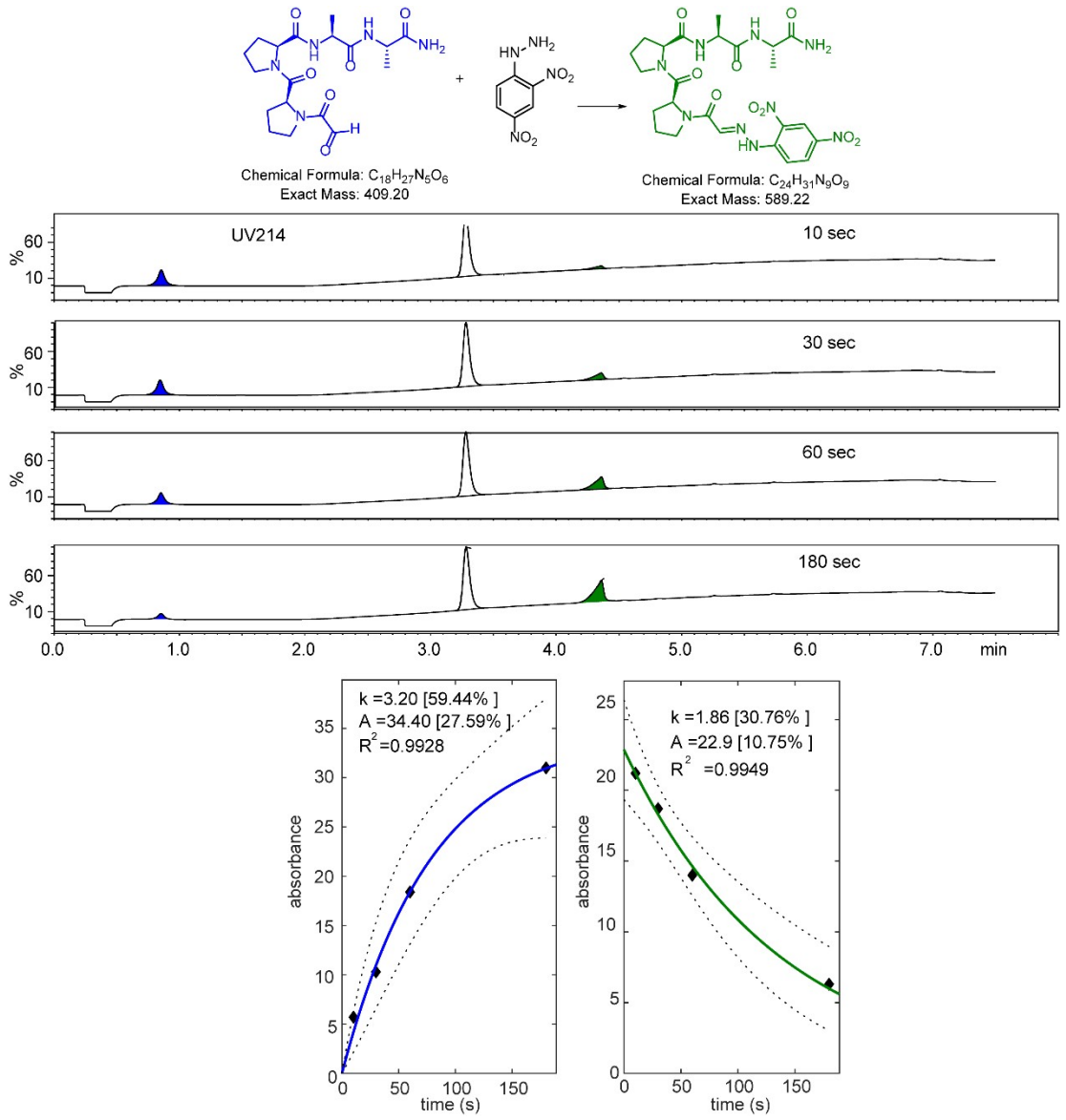


Figure S15. Kinetics of hydrazone ligation for HCO-PPAA at [H_2SO_4] = 65 mM.

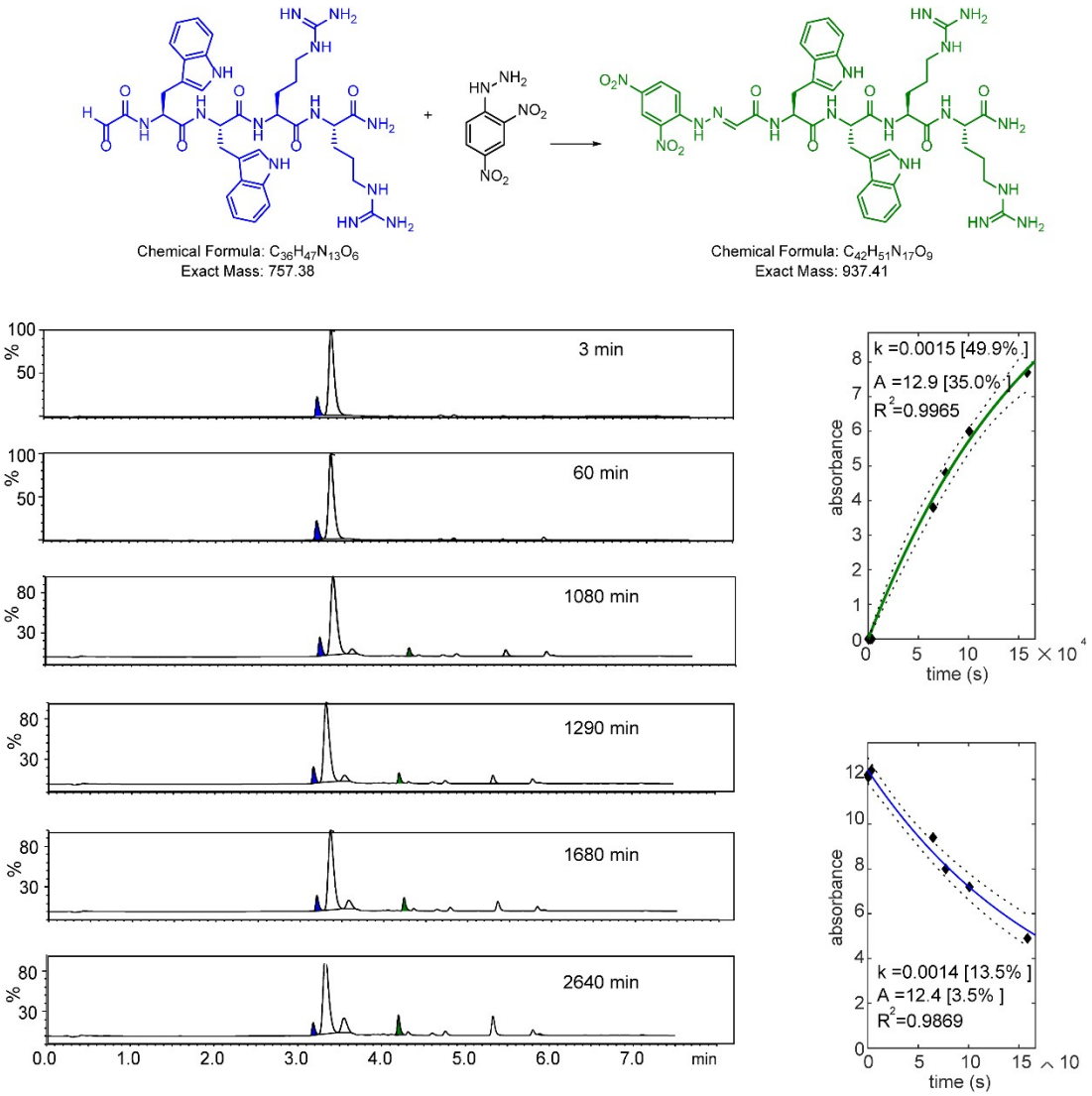


Figure S16. Kinetics of hydrazone ligation for HCO-WWRR at pH 5

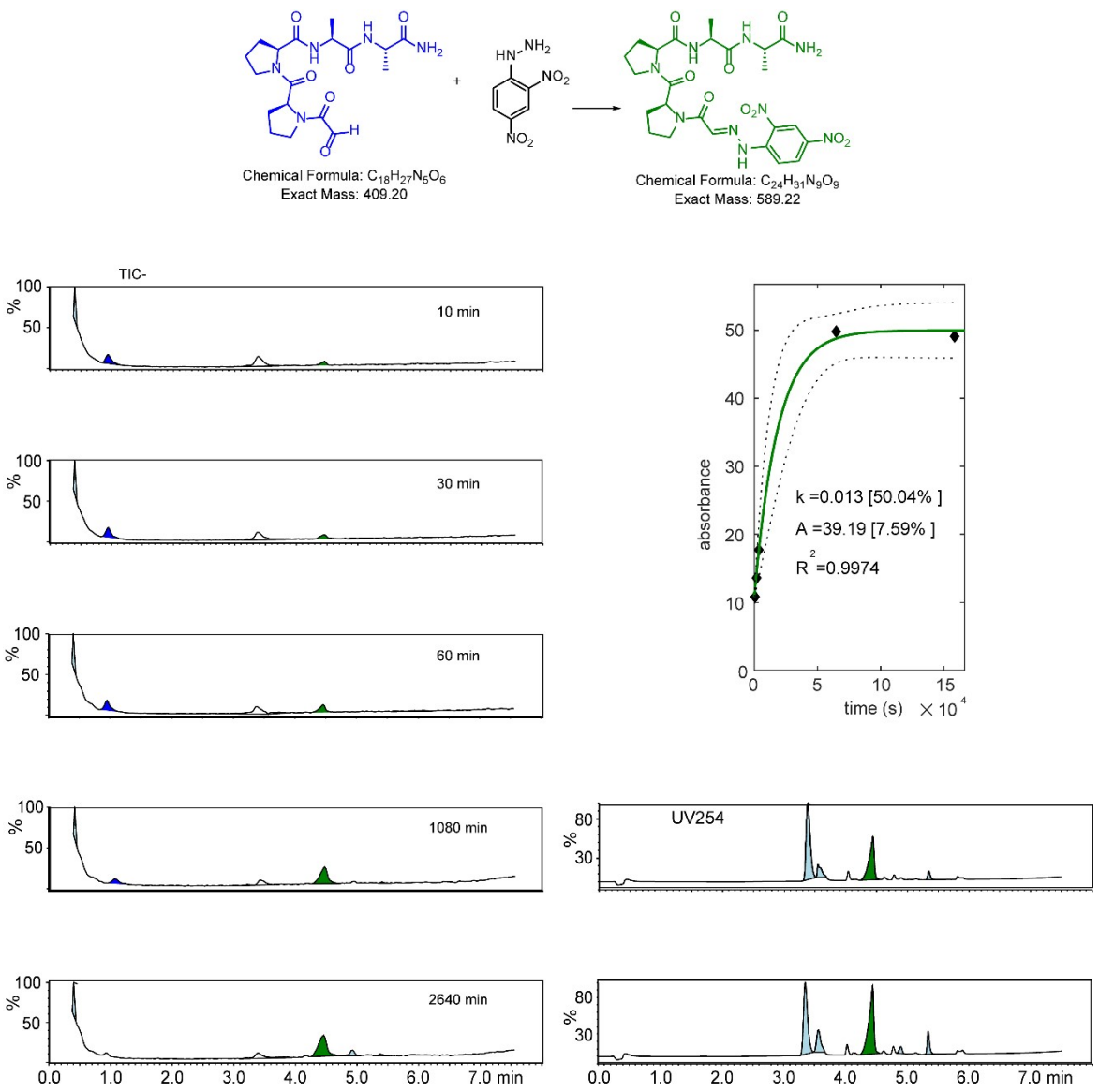


Figure S17. Kinetics of hydrazone ligation for HCO-PPAA at pH 5

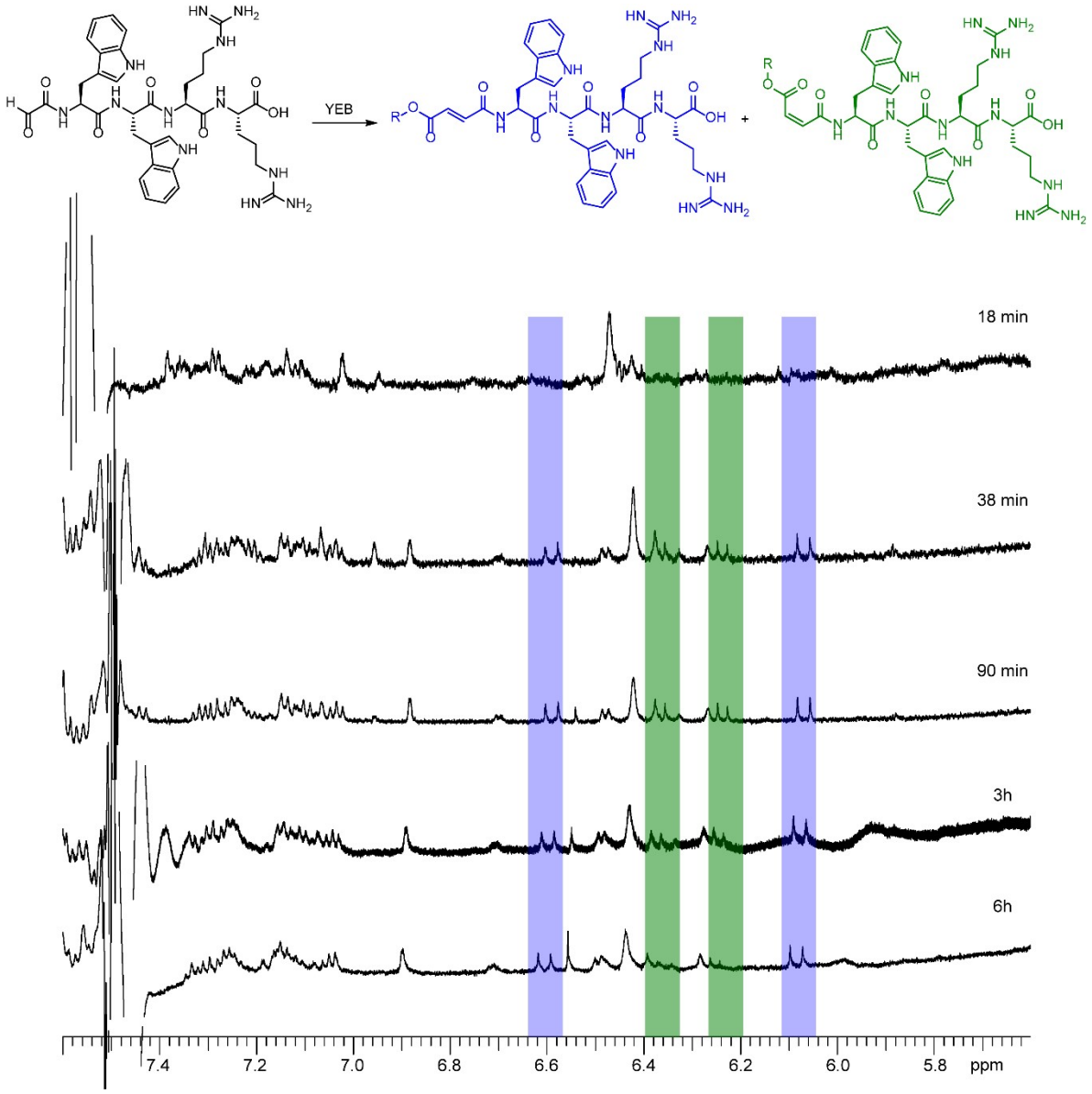


Figure S18. E/Z selectivity for HCO-WWRR (1:1 *E/Z*). NMR multiples for alkene protons in *E* product are highlighted in purple while *Z* product are highlighted in green.

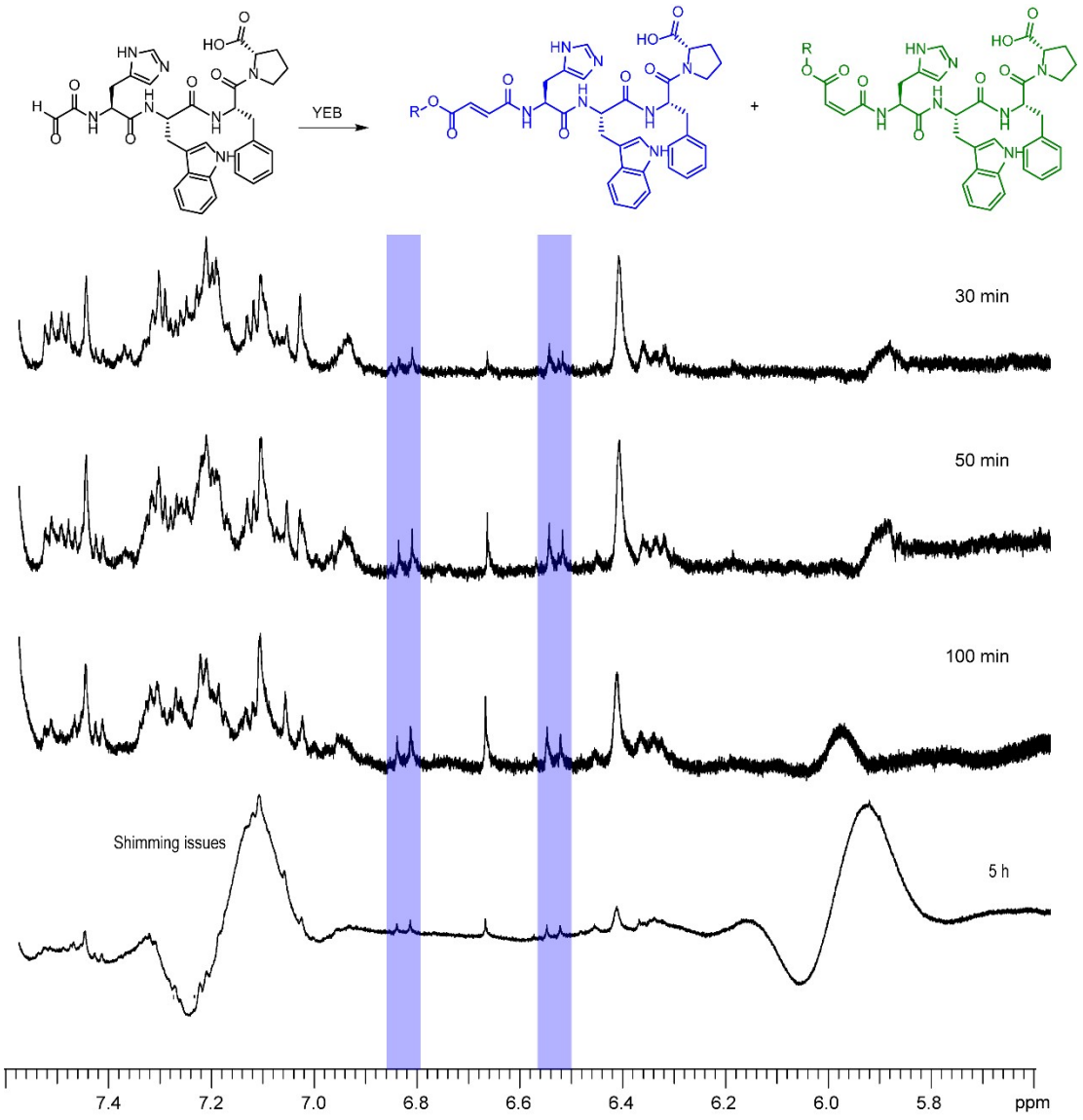


Figure S19. E/Z selectivity for HCO-HWFP (1:9 *E/Z*). NMR multiples for alkene protons in *E* product are highlighted in purple while *Z* product are highlighted in green.

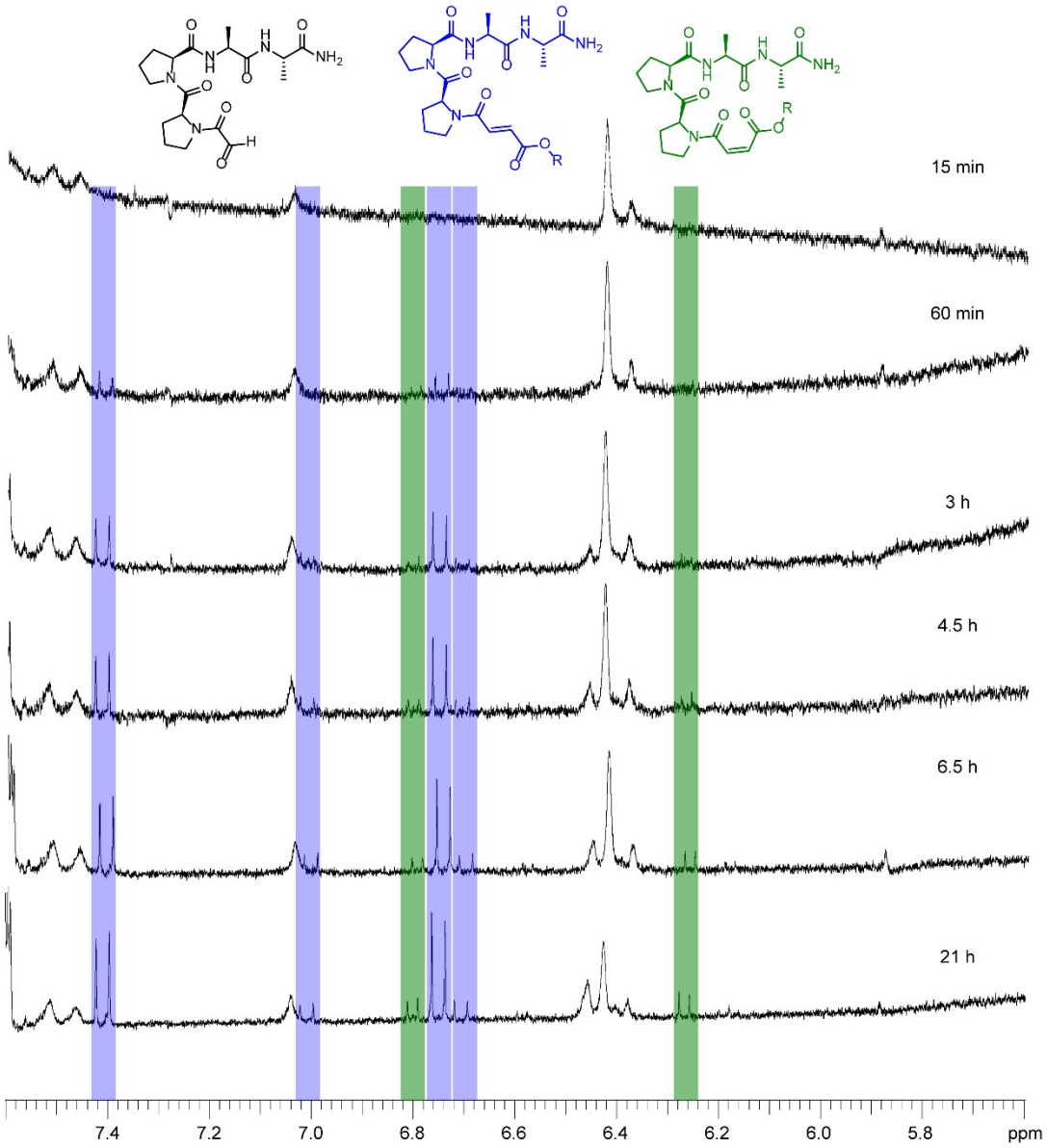


Figure S20. E/Z selectivity for HCO-PPAA. (*E/Z* 4.5:1). NMR multiples for alkene protons in *E* product are highlighted in purple while *Z* product are highlighted in green.

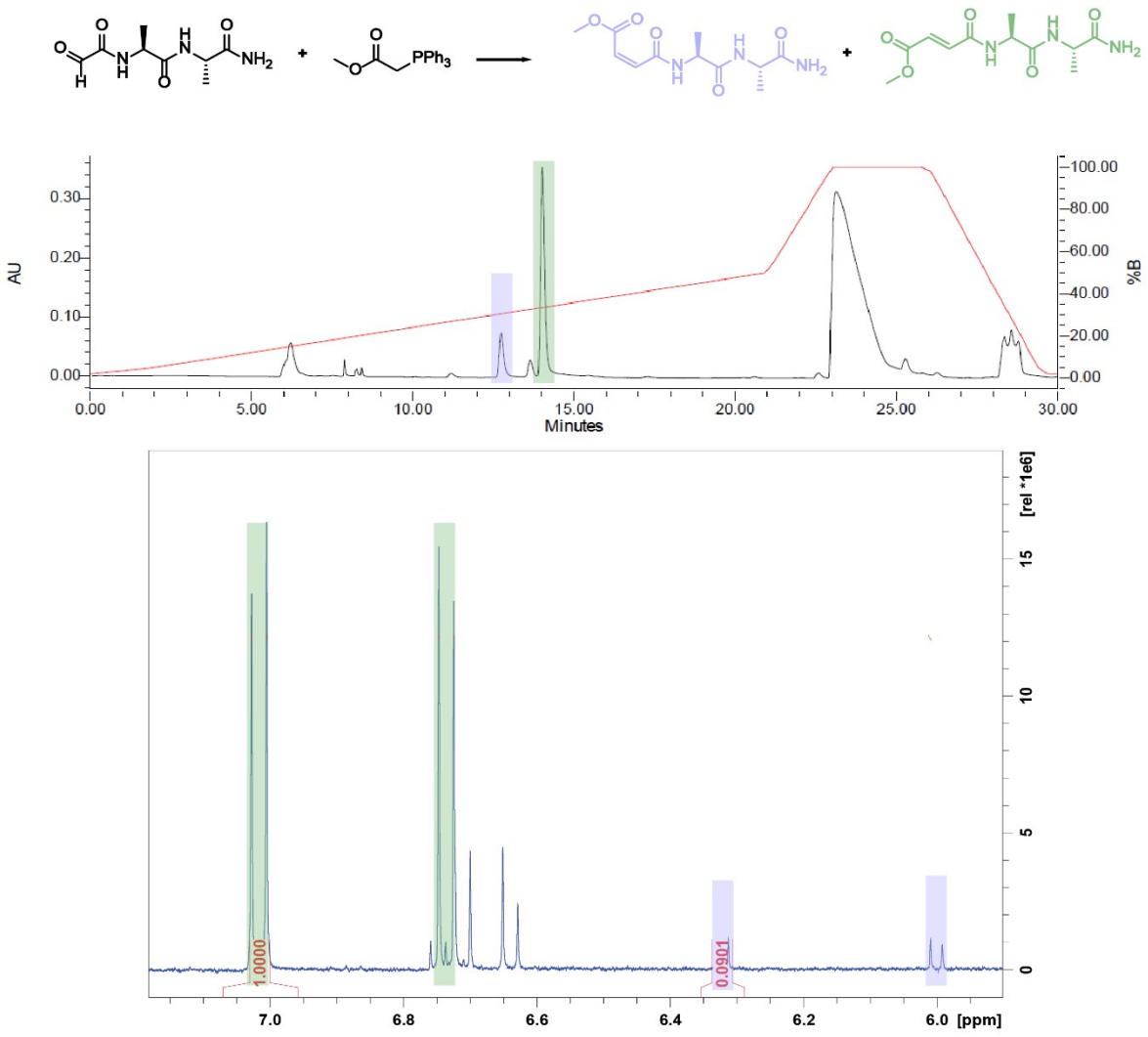


Figure S21. E/Z selectivity for HCO-AA (*E/Z* 4.5:1). NMR multiples for alkene protons in *E* product are highlighted in purple while *Z* product are highlighted in green.

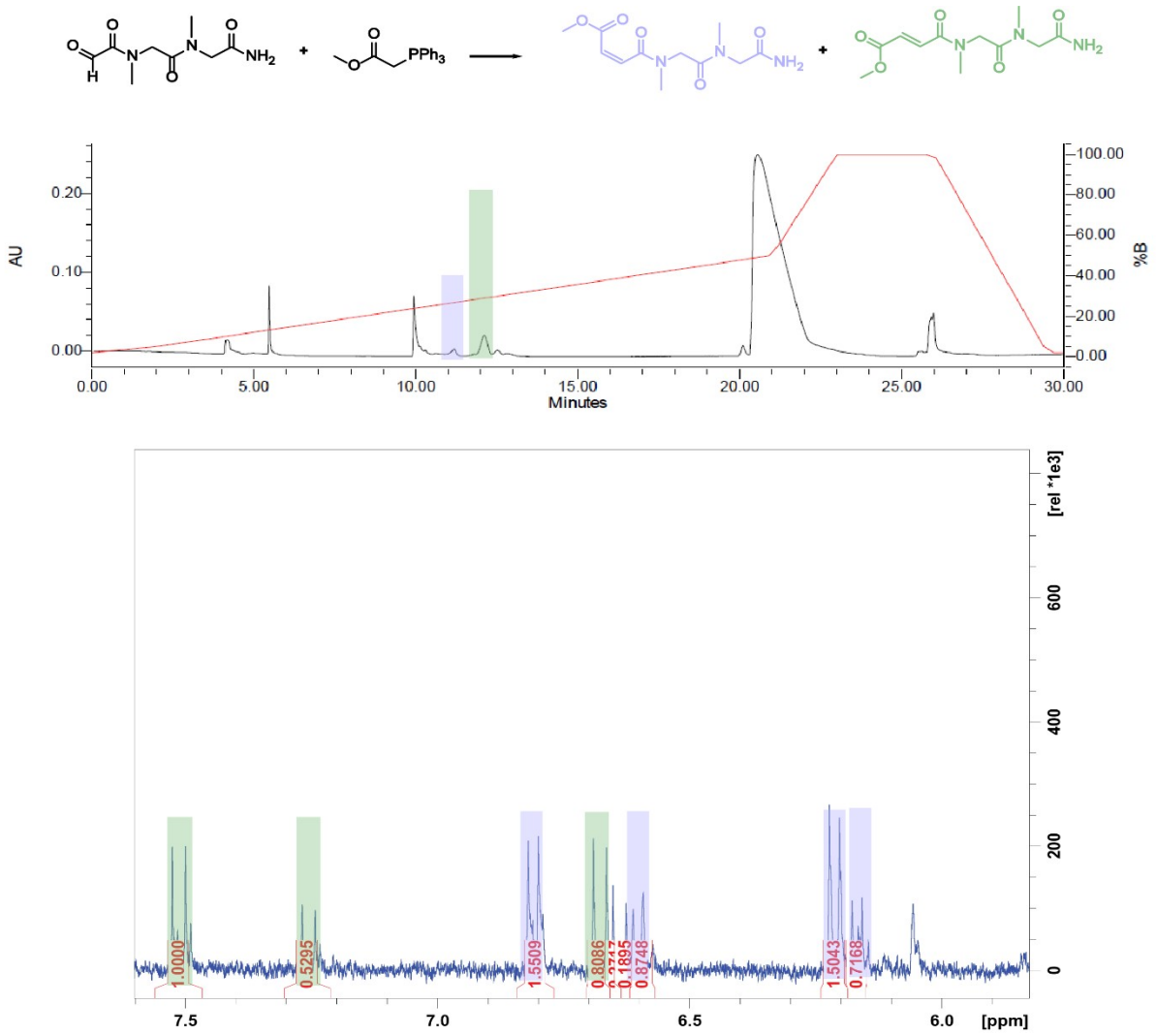
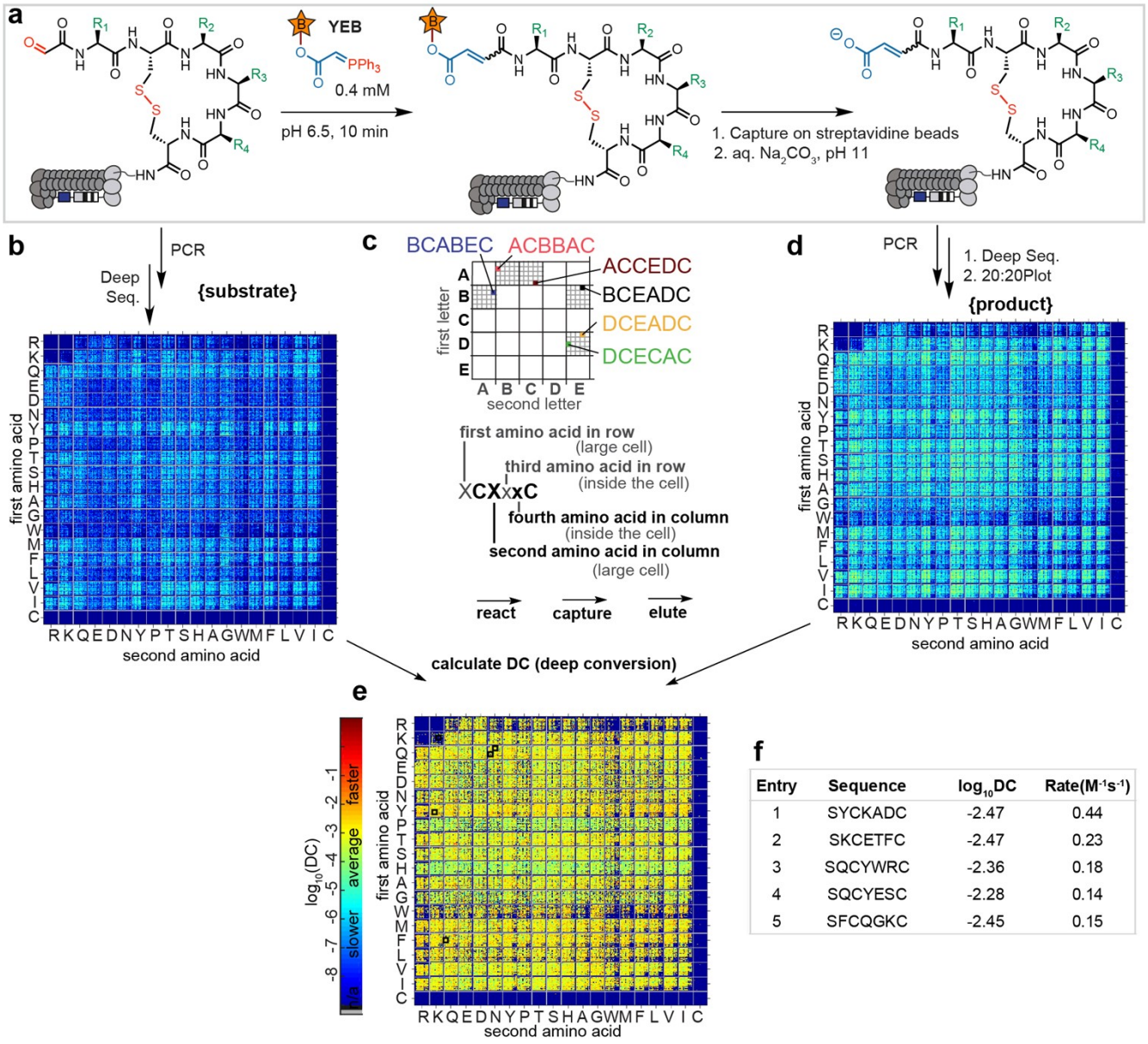


Figure S22. E/Z selectivity for HCO-SarcosineSarcosine (*E/Z* 3.6:1). NMR multiples for alkene protons in *E* product are highlighted in purple while *Z* product are highlighted in green.



Entry	Sequence	$log_{10}DC$	Rate($M^{-1}s^{-1}$)
1	SYCKADC	-2.47	0.44
2	SKCETFC	-2.47	0.23
3	SQCYWRC	-2.36	0.18
4	SQCYESC	-2.28	0.14
5	SFCQGKC	-2.45	0.15

Figure S23. (a) Wittig reaction on SXCXXXC phage libraries. (b) 20×20 plots displaying library-wide DC values from peptide substrates and the Wittig product. (c) How to read a 20×20 plot. (d) After reacting 10 min, the biotinylated Wittig products were captured by streptavidin beads and ready for sequencing. (e) 20×20 plots for $SXCX_3C$ selections. (f) Experimental reaction rates for $SXCX_3C$ peptides.

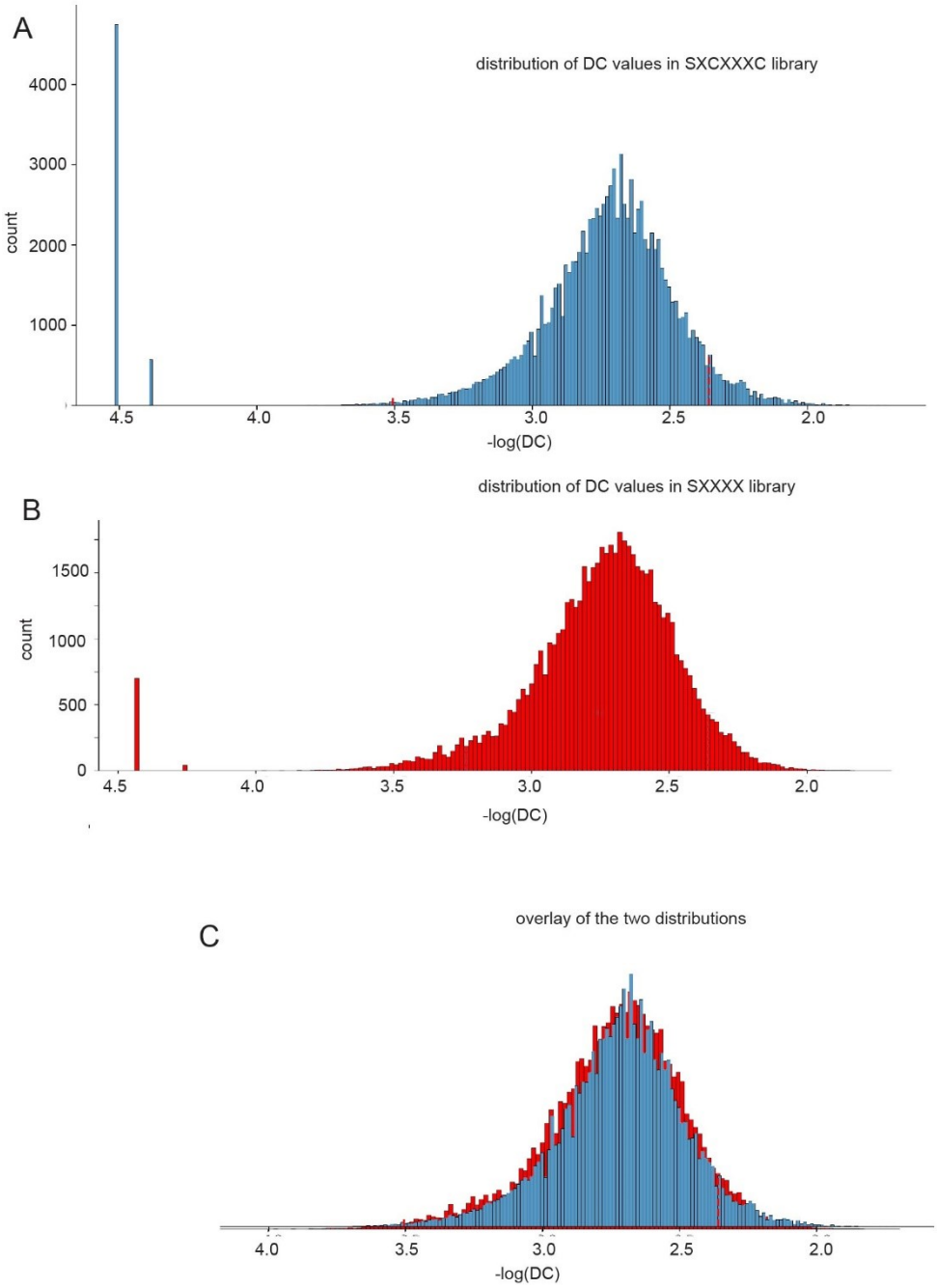


Figure S24. Distribution of log DC value (DC = "deep conversion") of $SXCX_3C$ library. When overlapping the distribution of $SXCX_3C$ (blue) library and SX_4 library (red), $SXCX_3C$ library shows a narrower distribution pattern.

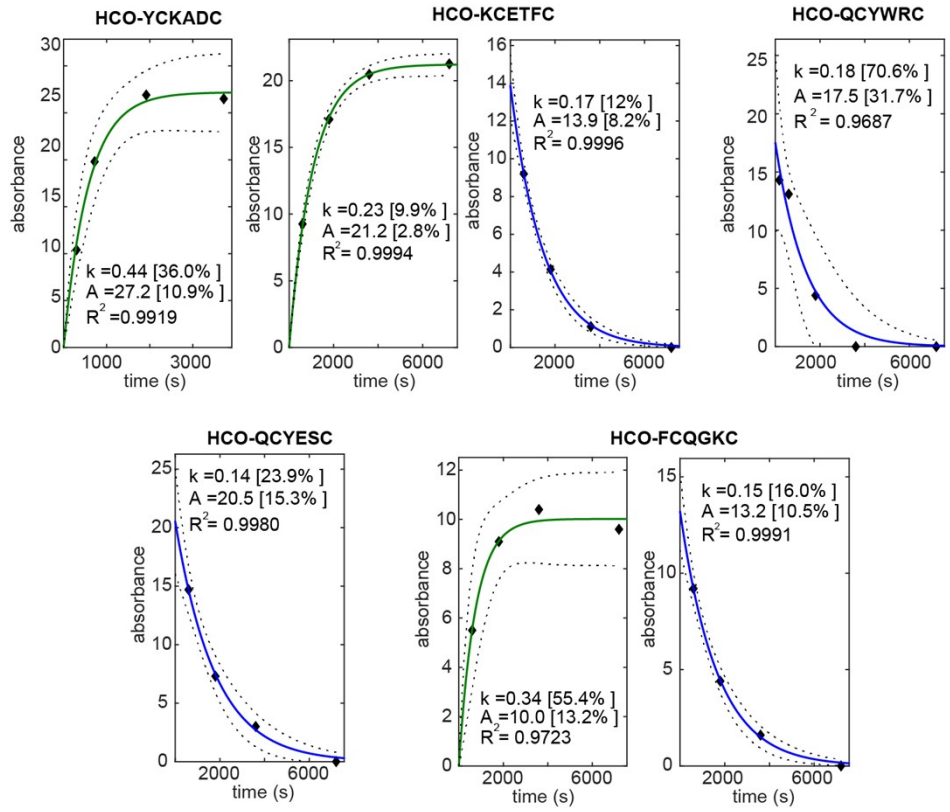


Figure S25. Kinetic traces for sequences of $SXCX_3C$ selection

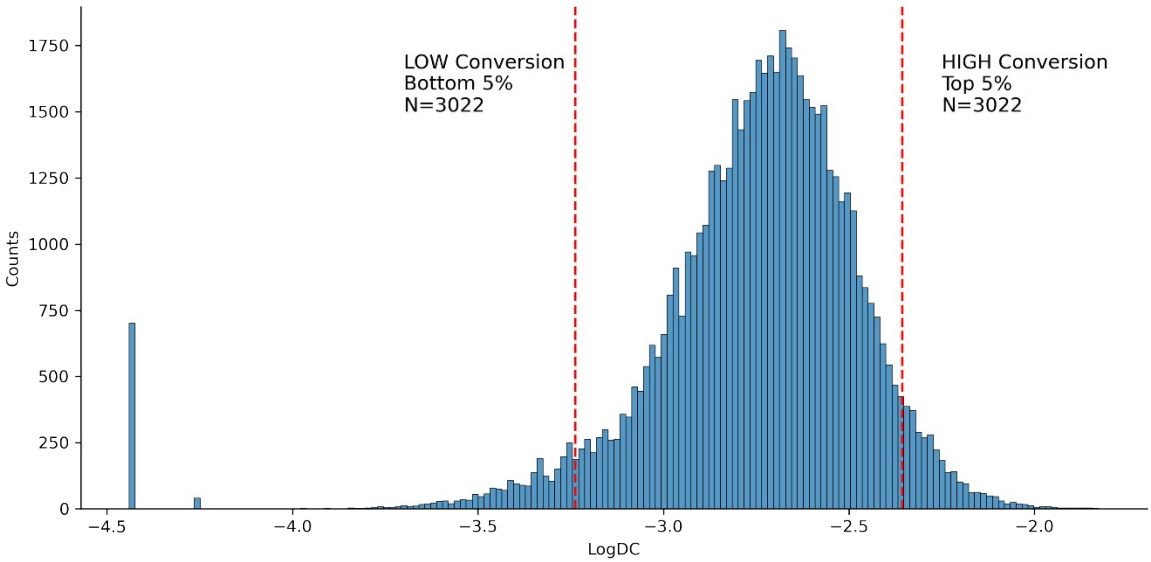


Figure S26. Distribution of log DC value.
Histogram of the top 5% and bottom 5% of the sequences based on the DC values. This threshold was used to define labels for peptides for the classification task.

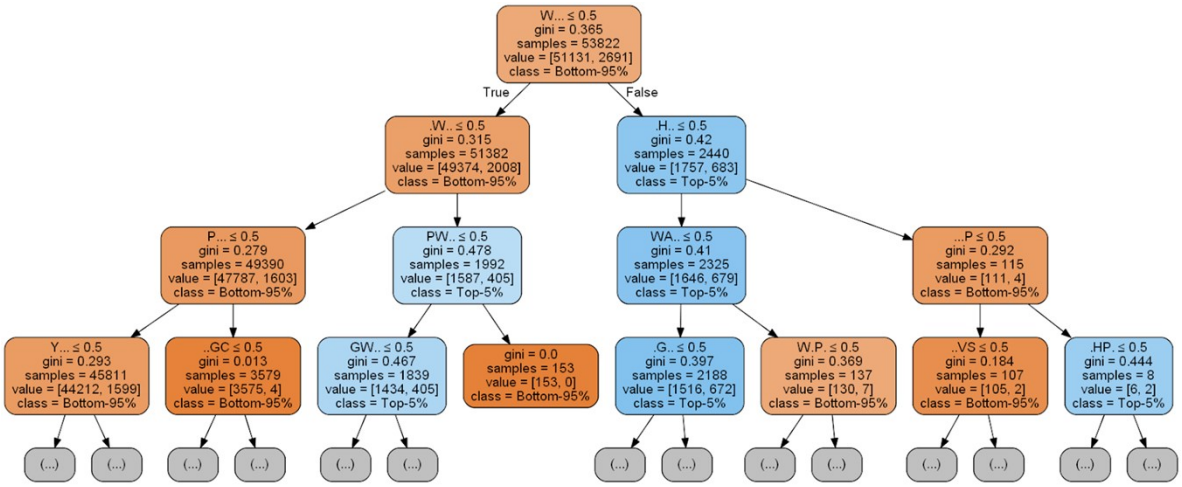


Figure S27. A single decision tree for illustrative purposes.

The XGBoost algorithm uses an ensemble of trees to generate a more accurate output for the classification task. Decision trees can be used to deduce a simple rule based model that navigates through a sequence pattern. Impurity at each node is shown below using the gini index. Wildcard base is shown as '.' e.g.: (W…) is WXXX where X can be any of the 20 amino acid bases. Value of 0 and 1 is assigned to the absence and presence of pattern respectively (Therefore ≤ 0.5 means absence of a particular pattern).

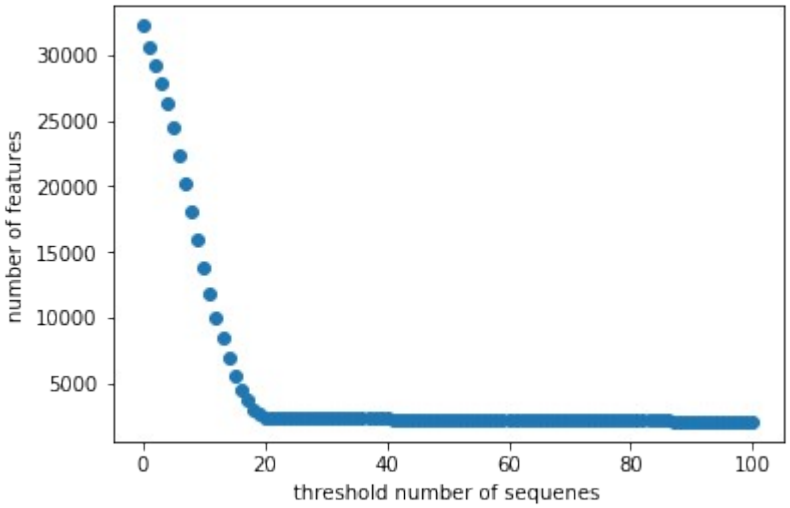


Figure S28. Number of sequence patterns vs threshold number of sequences used as input features for the models.
This threshold was based on the number of instances with each sequence pattern appearing among the observed sequences. When the threshold for the minimum number of instances was set to 20, it reduced the number of sequence patterns used as input features from 34,480 to 2,396.

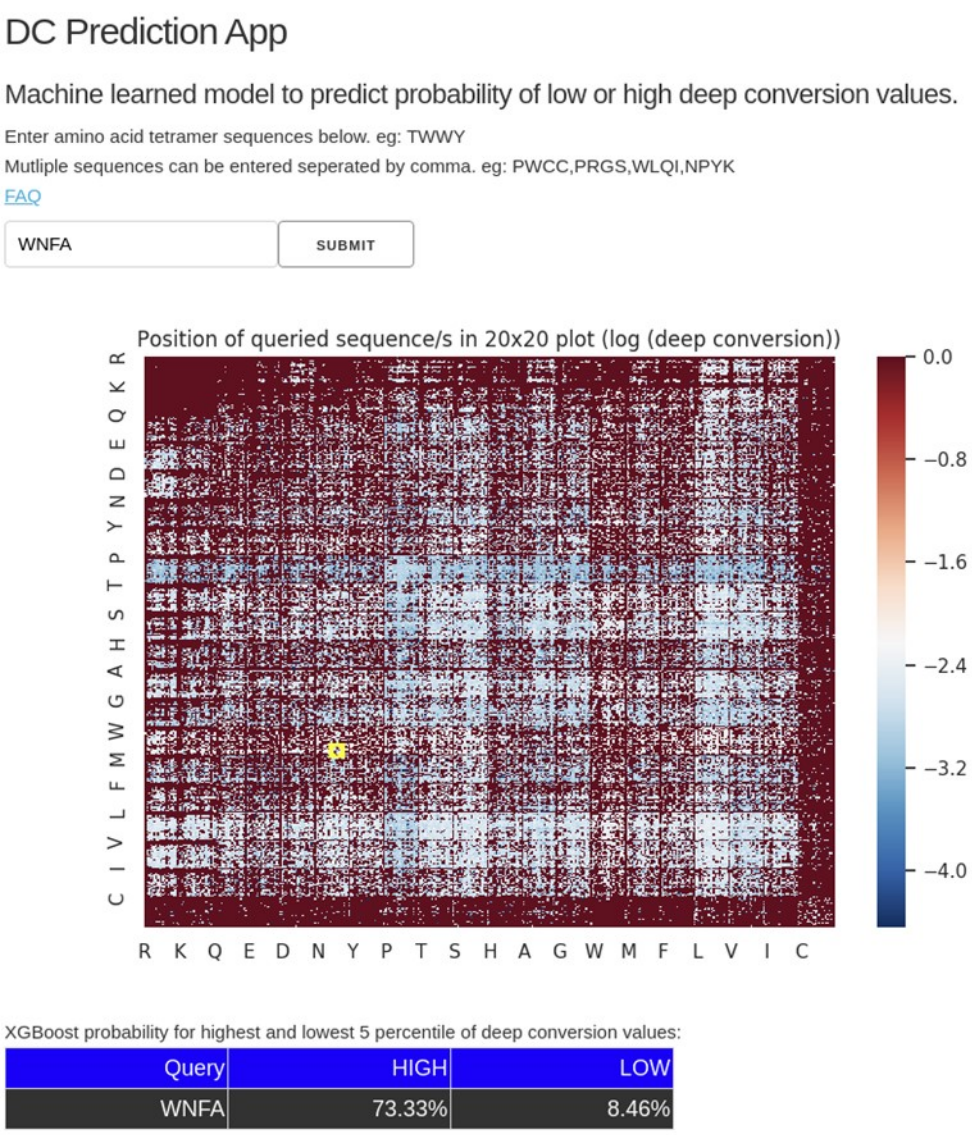


Figure S29. Screenshot of the DC prediction app, http://44.226.164.95/.

The app allows users to provide sequences and predicts the probability of a sequence belonging to the HIGH and LOW DC groups. HIGH and LOW are learnt using two separate models. Hence, the scores do not necessarily add up to 1. Users can use their judgement to compare the scores and decide which of the two is more likely. However, these scores are not calibrated probabilities.

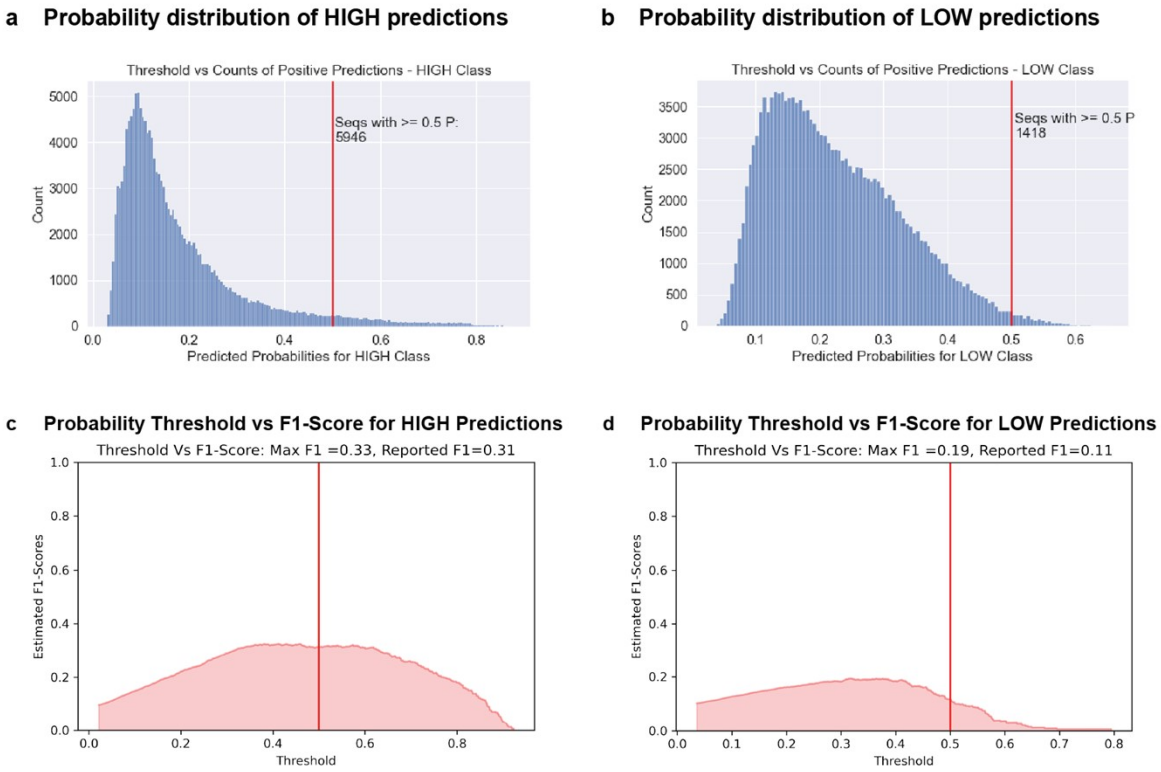


Figure S30. Probability distributions for predictions and corresponding F1-scores for different probability thresholds. a) Probability distribution for predicted probabilities from the HIGH classifier model for predicted (unobserved) data. b) Probability distribution for predicted probabilities from the LOW classifier model for predicted (unobserved) data. c) Probability threshold vs F1-scores for HIGH classifier. d) Probability threshold vs F1-scores for LOW classifier.

Metric	High DC Mean ± SD	Low DC Mean ± SD
Accuracy	92.5 ± 0.2	90.1 ± 0.2
AUC-ROC	**81.2 ± 0.4**	**73.7 ± 0.8**
F1-Score*	33.7 ± 0.9	19.0 ± 0.9
Precision	30.4 ± 1.2	16.1 ± 0.9
Recall	37.9 ± 1.1	23.2 ± 0.9

Table S4. Performance metrics of the two classifiers for a 5-fold cross validation.
* F1-Score is the harmonic mean of precision and recall. The highest value for F1 is 1.0 which indicates perfect precision and recall and the lowest is 0 which happens if either precision or recall is 0.

Query	Prediction	HIGH	LOW	k ($M^{-1}s^{-1}$)	assessment
WYFT	fast	90.80%	12.28%	0.93	true
WWGP	fast	92.61%	2.43%	0.92	true
WWRR	fast	88.00%	2.19%	0.90	true
LYAR	average	27.82%	3.43%	0.14	true
WFFP	fast	92.69%	12.65%	0.13	false
WYAP	fast	91.41%	4.53%	0.12	false
VWTA	average	48.01%	6.56%	0.12	true
FPWE	slow	2.43%	66.23%	0.074	true
HAWD	slow	5.55%	55.49%	0.020	true
GIIE	slow	4.07%	59.88%	0.007	true
RYIP	fast	73.48%	20.26%	0.003	false (class switch)

Table S5. Reaction rates of fast, average, and slow peptide sequences predicted by machine learning measured by HPLC.

We added true/false assessment for each predicted sequence as following:

"True" label was used if experimentally determined rate coincided with the predicted class:
Example 1: Sequence predicted to react fast indeed reacted faster than average.
Example 2: Sequence predicted to react slow indeed reacted slower than average.
Example 3: Sequence predicted to be not fast and not slow (i.e., average) had an average reaction rate.

We added label "false" if experimentally determined rate did not belong to the predicted class. We note that two sequences predicted to be "fast" were experimentally determined to be "not fast" (i.e., average) whereas one sequence predicted to be "fast" reacted significantly slower than average (i.e., it belongs to a "slow" class).

HPLC purity and LCMS traces of synthesized peptides

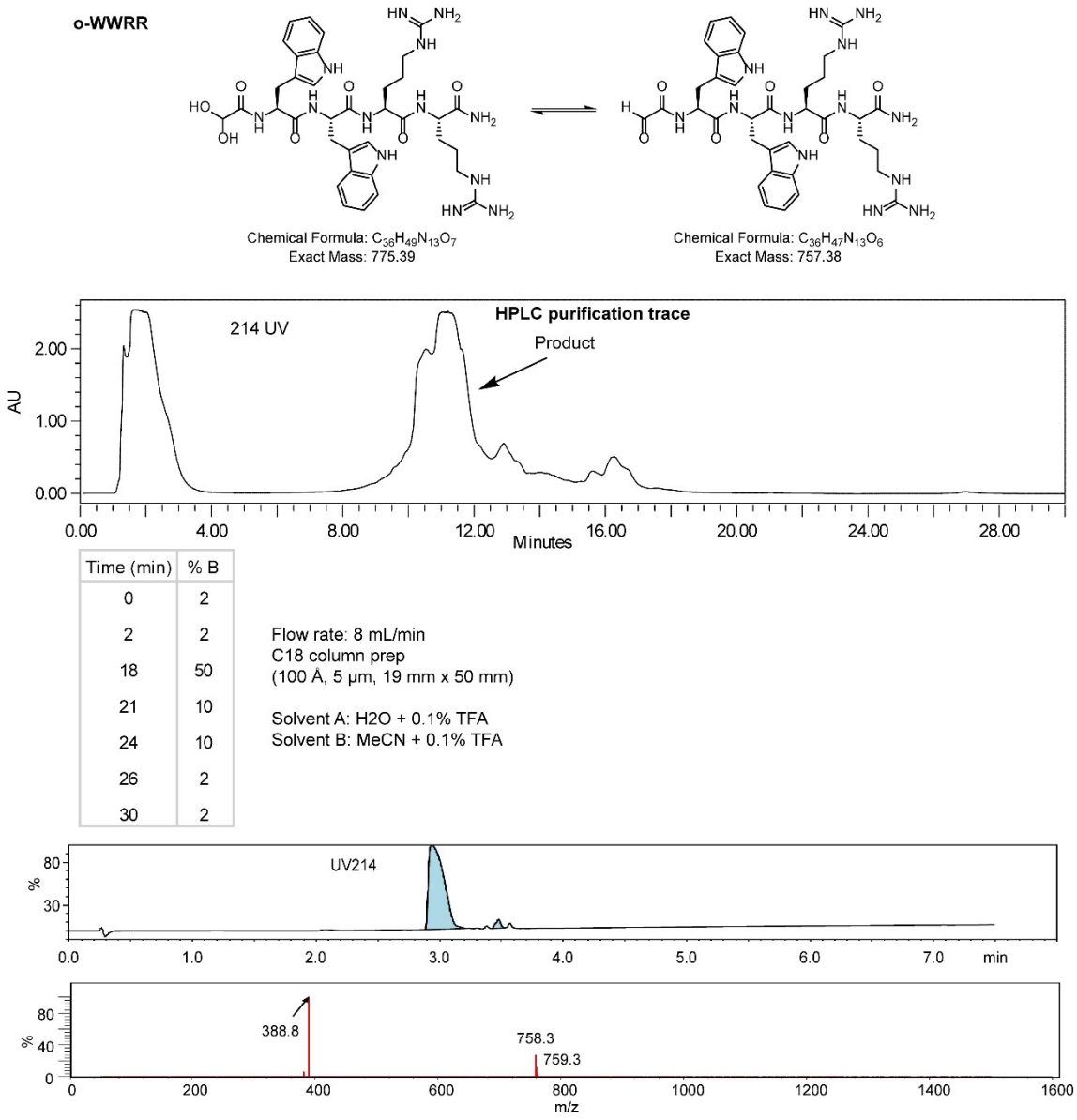


Figure S31. Summary for HCO-WWRR synthesis

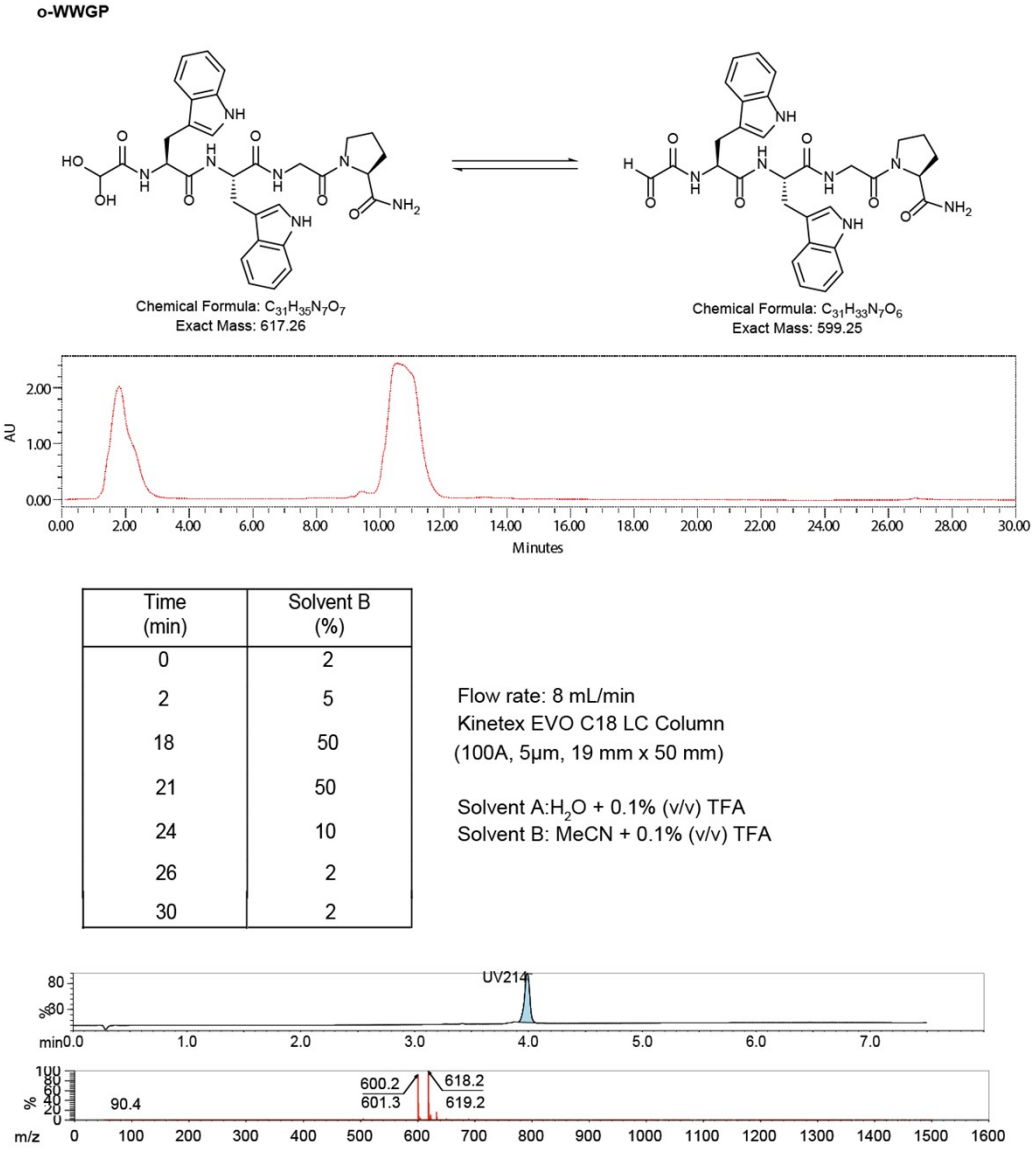


Figure S32. Summary for HCO-WWGP synthesis

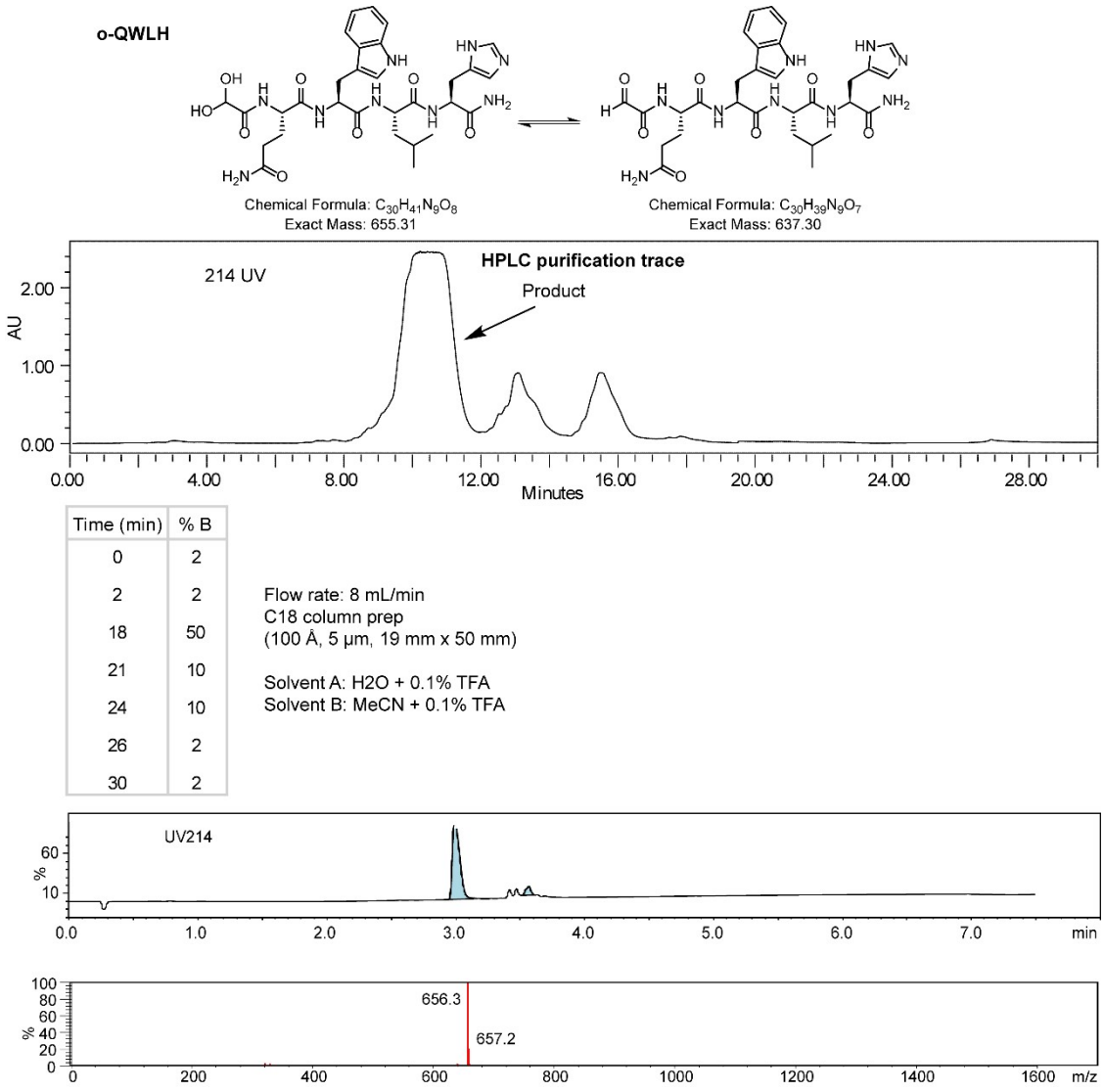


Figure S33. Summary for HCO-QWLH synthesis

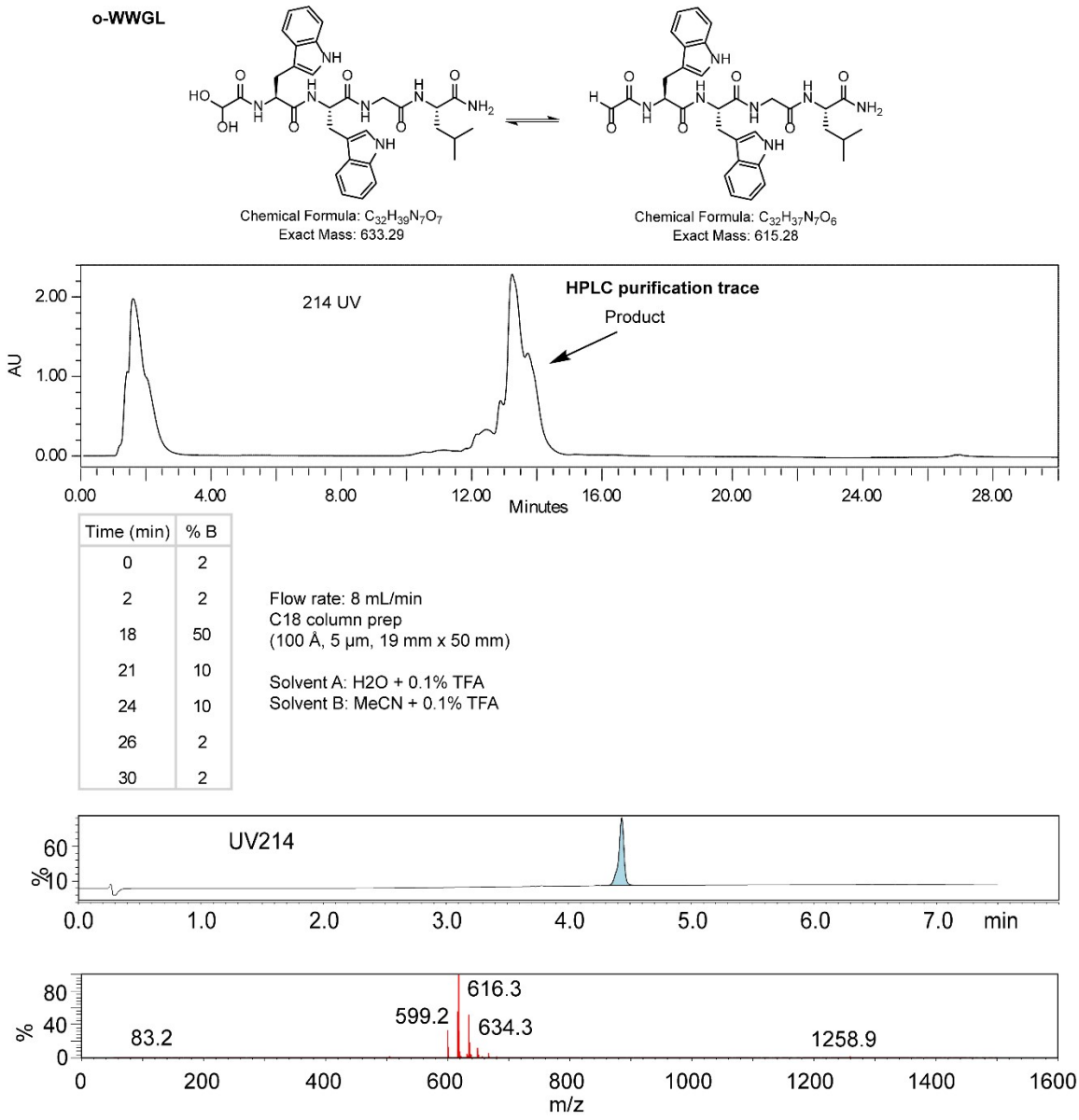


Figure S34. Summary for HCO-WWGL synthesis

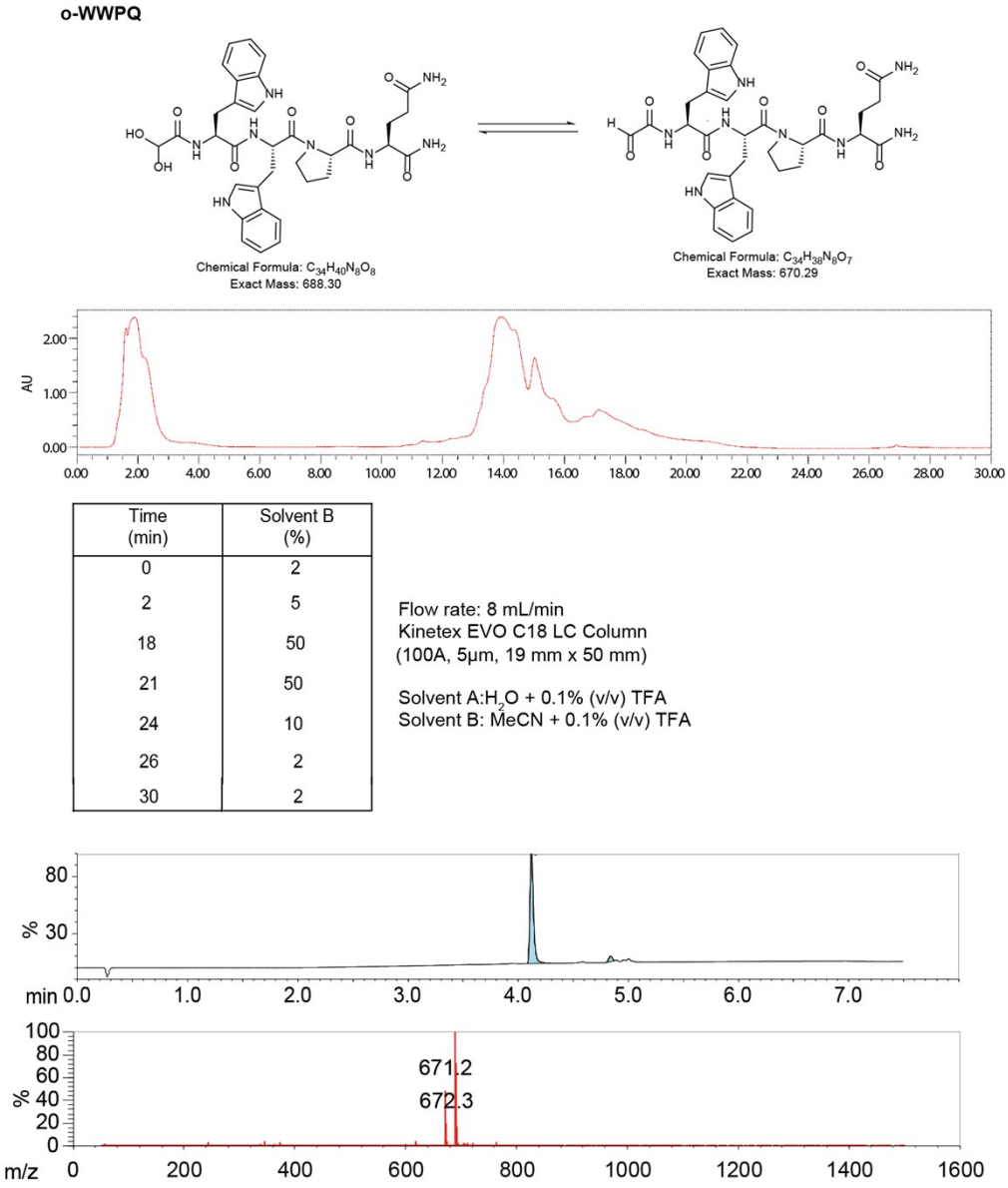


Figure S35. Summary for HCO-WWPQ synthesis

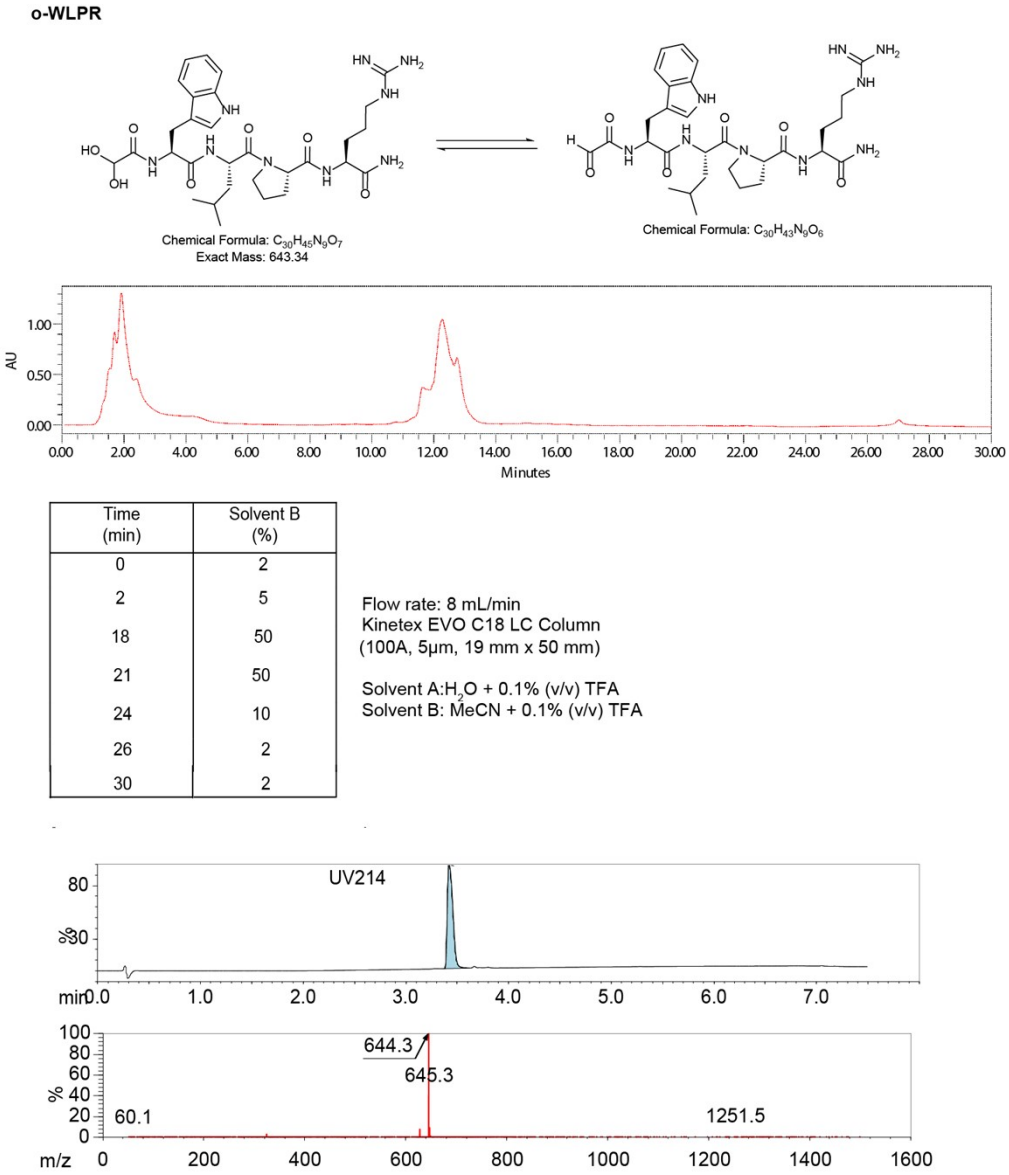


Figure S36. Summary for HCO-WLPR synthesis

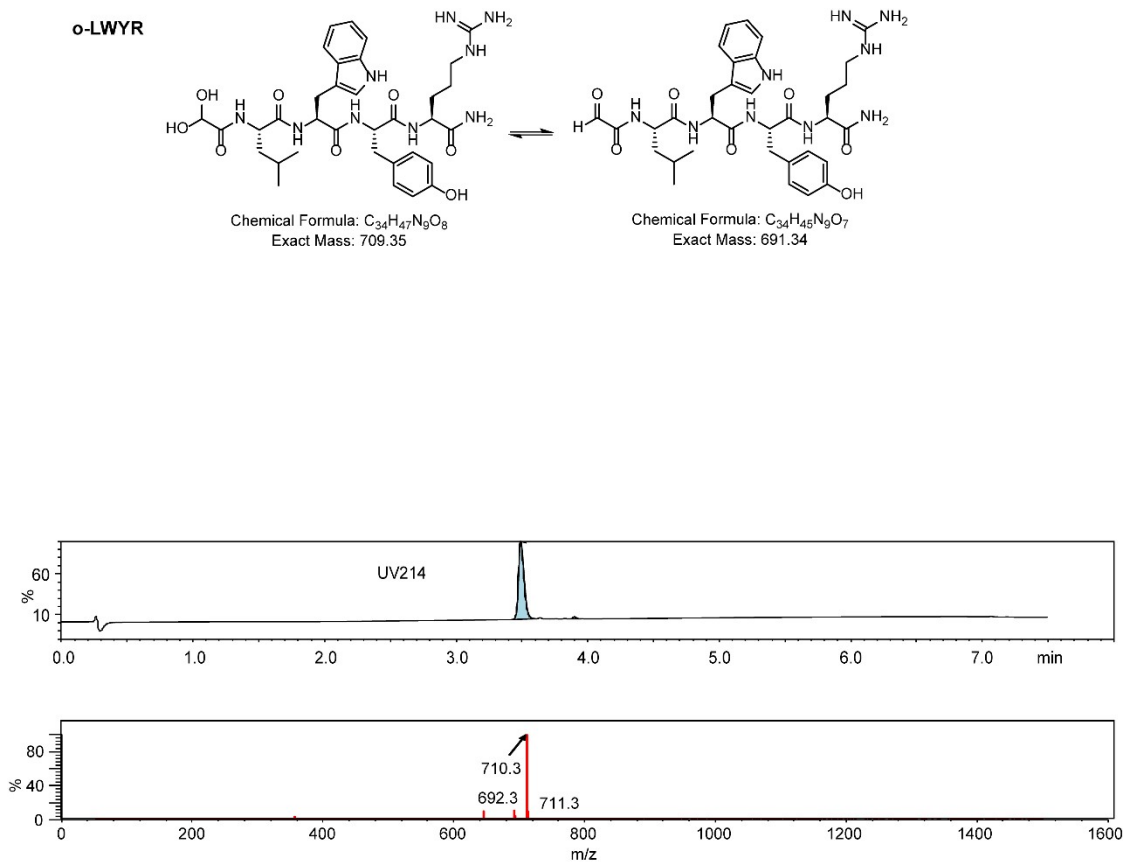


Figure S37. Summary for HCO-LWYR synthesis

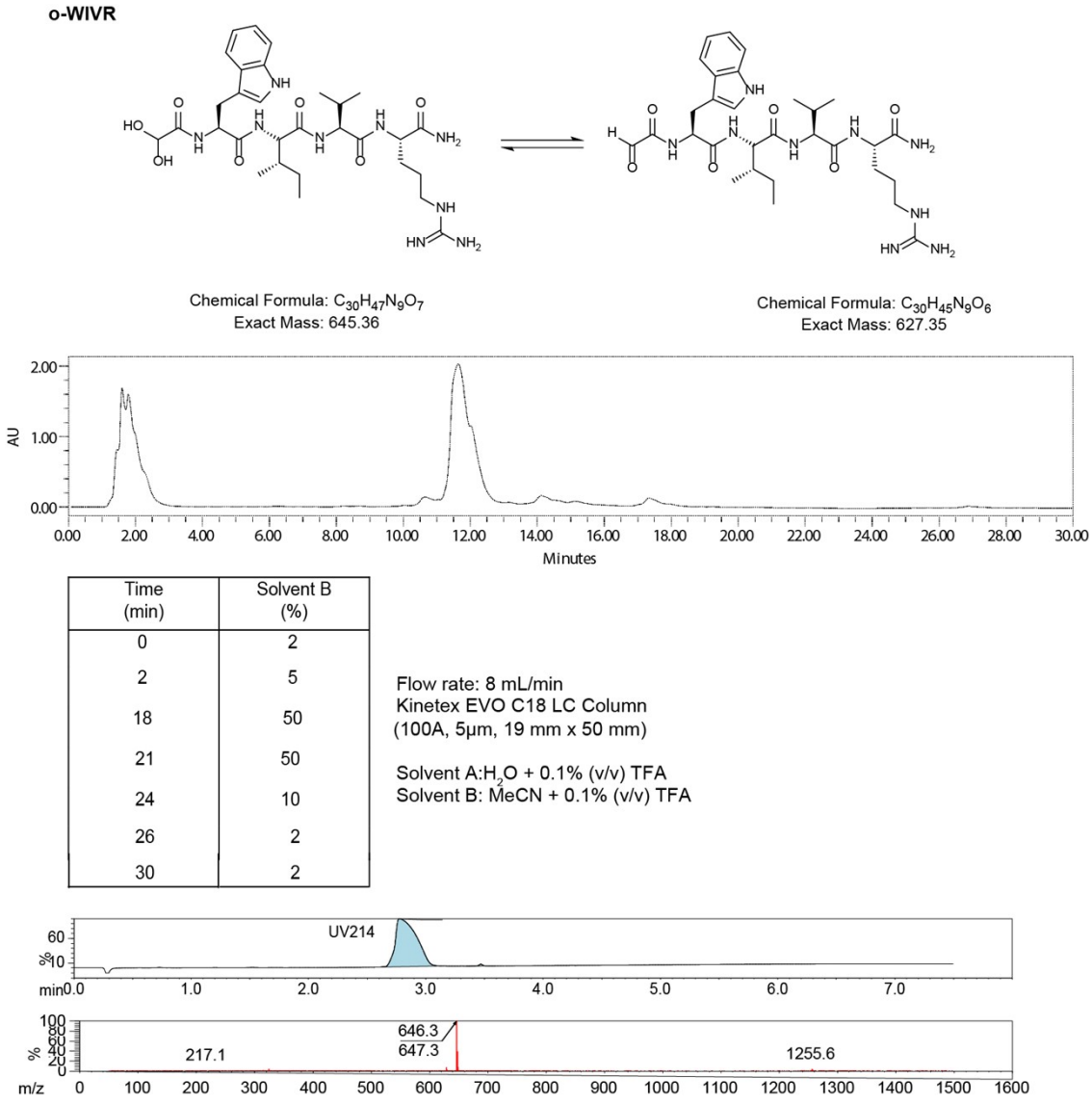


Figure S38. Summary for HCO-WIVR synthesis

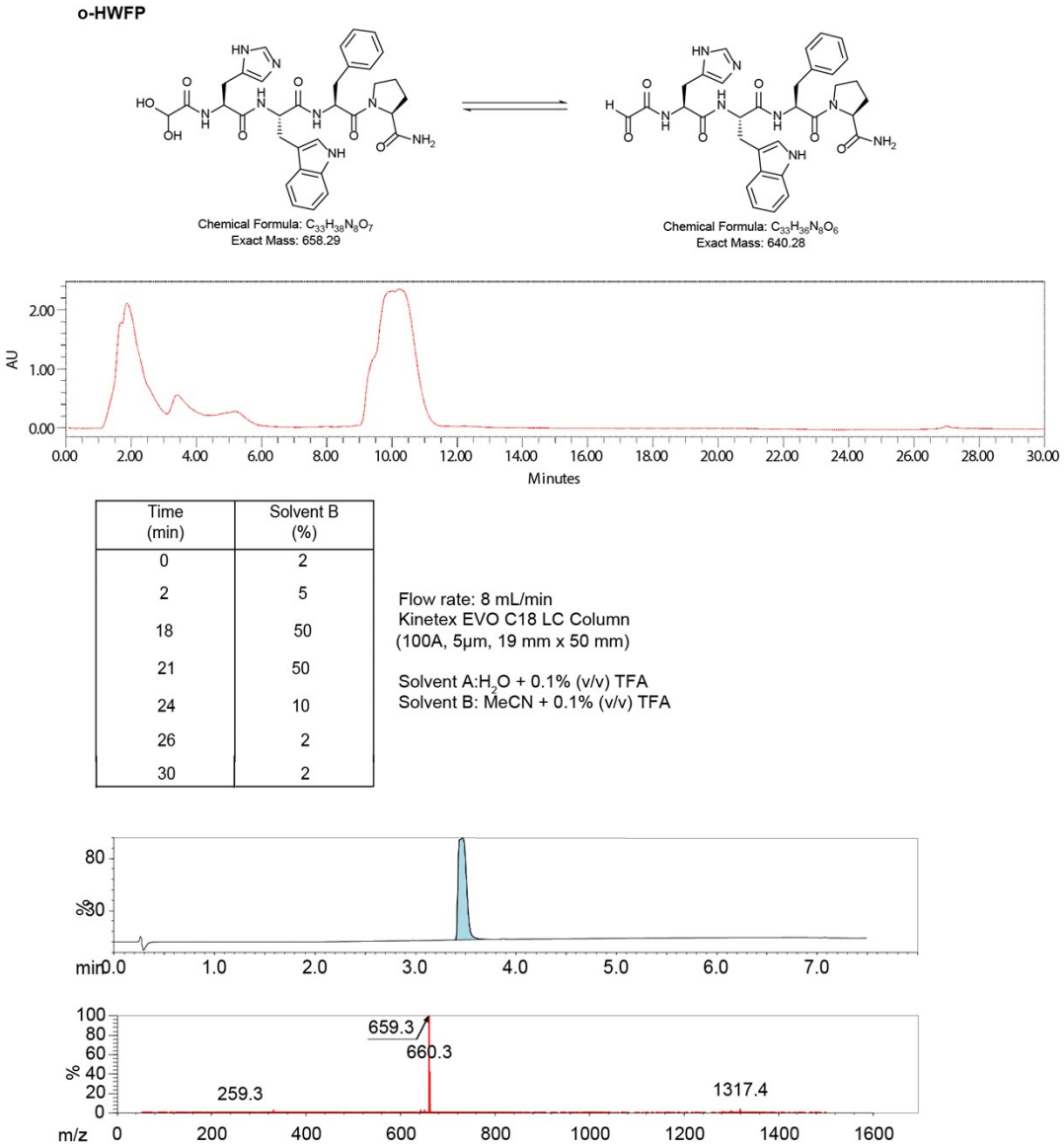


Figure S39. Summary for HCO-HWFP synthesis

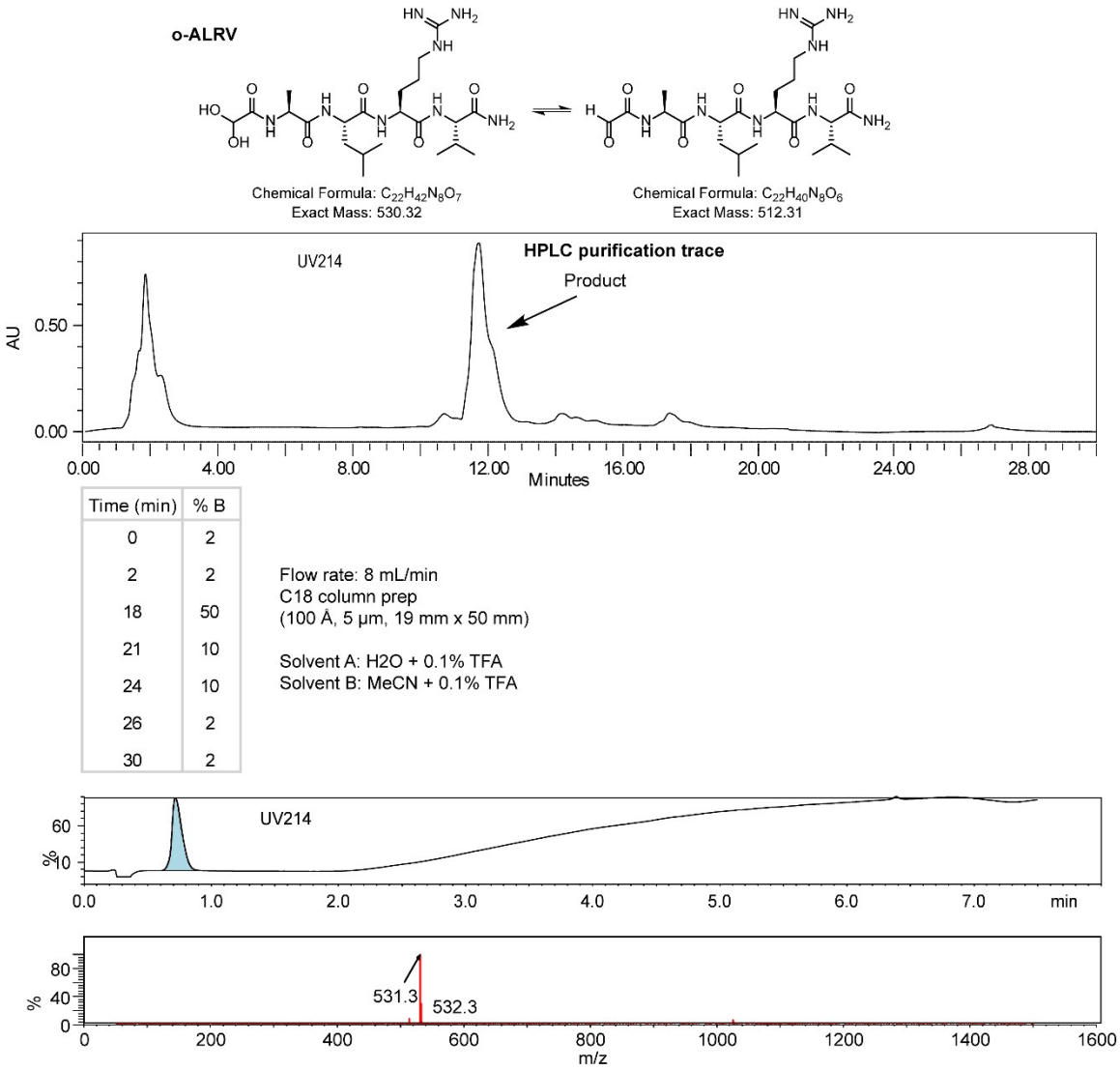


Figure S40. Summary for HCO-ALRV synthesis

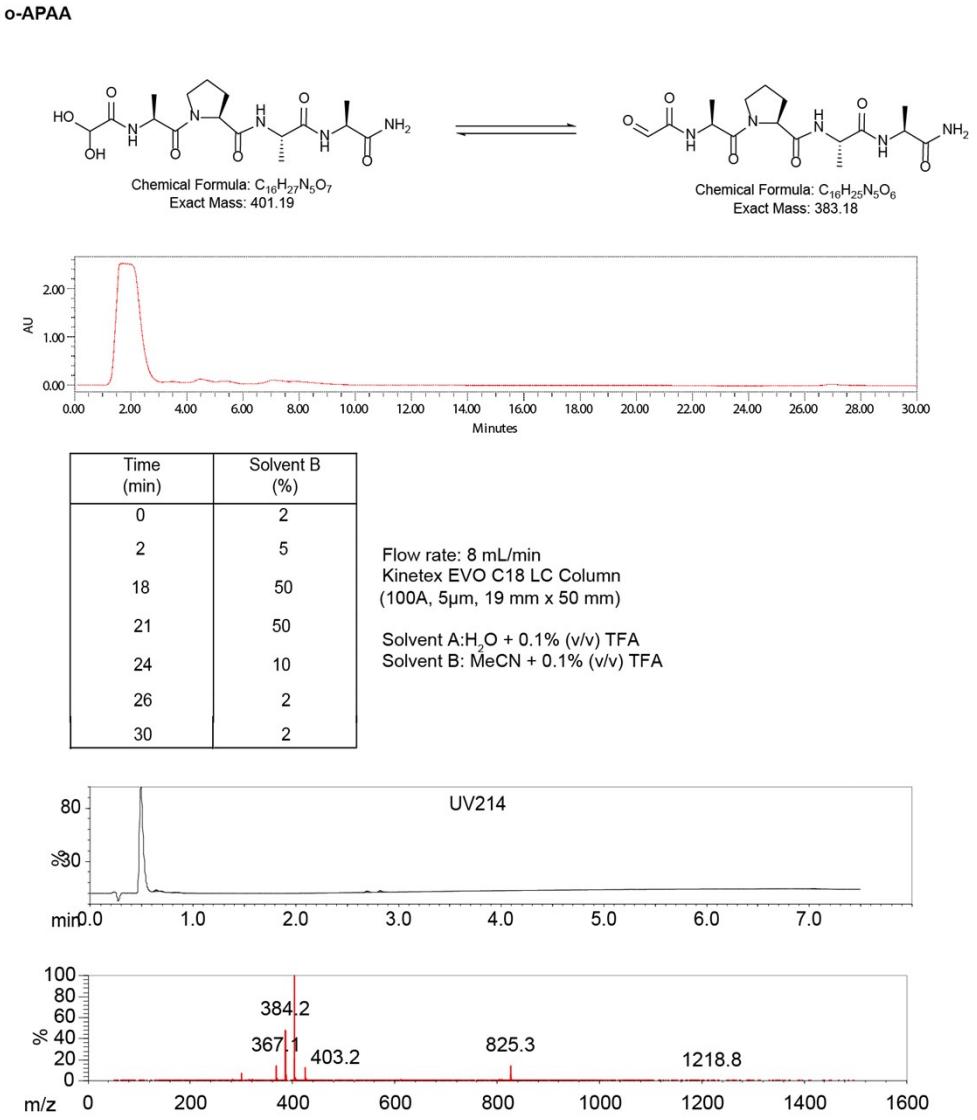


Figure S41. Summary for HCO-APAA synthesis

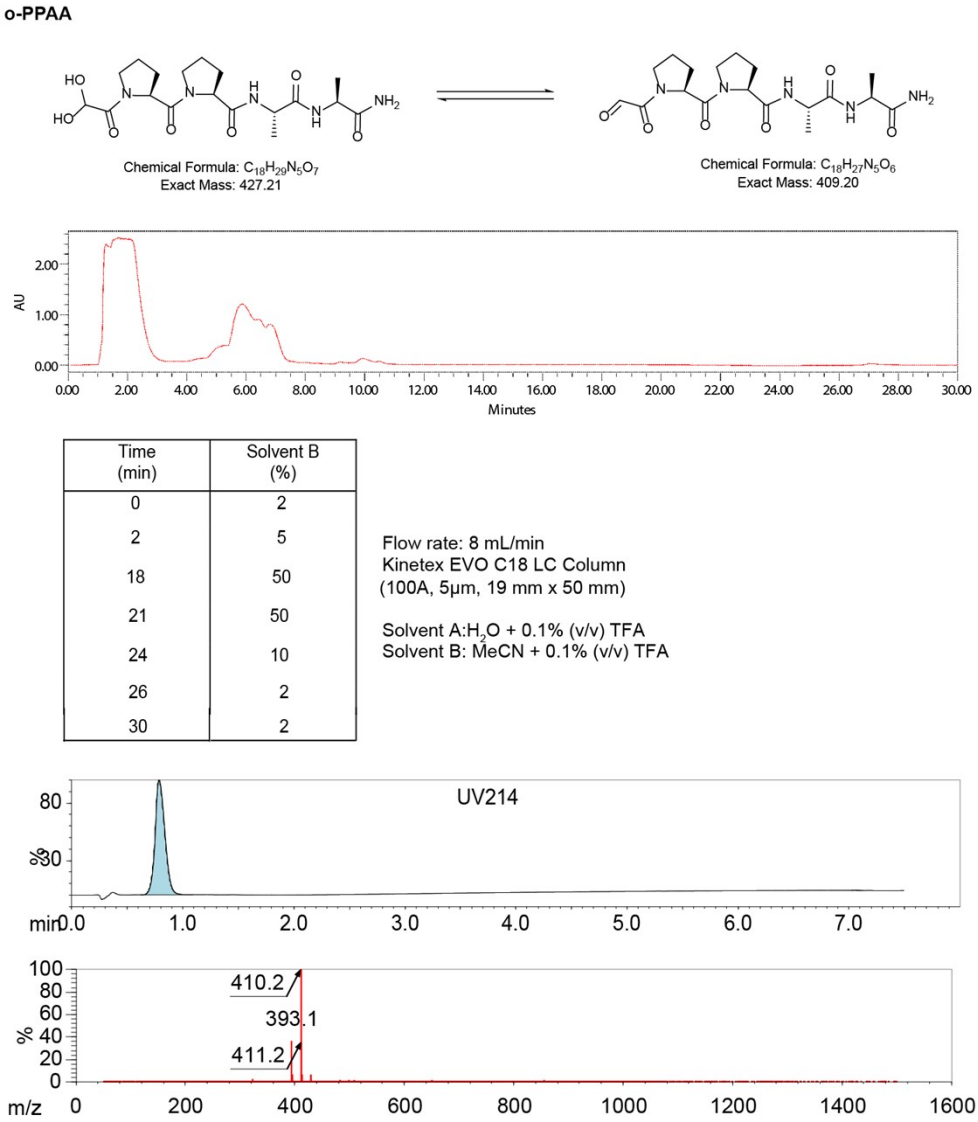


Figure S42. Summary for HCO-PPAA synthesis

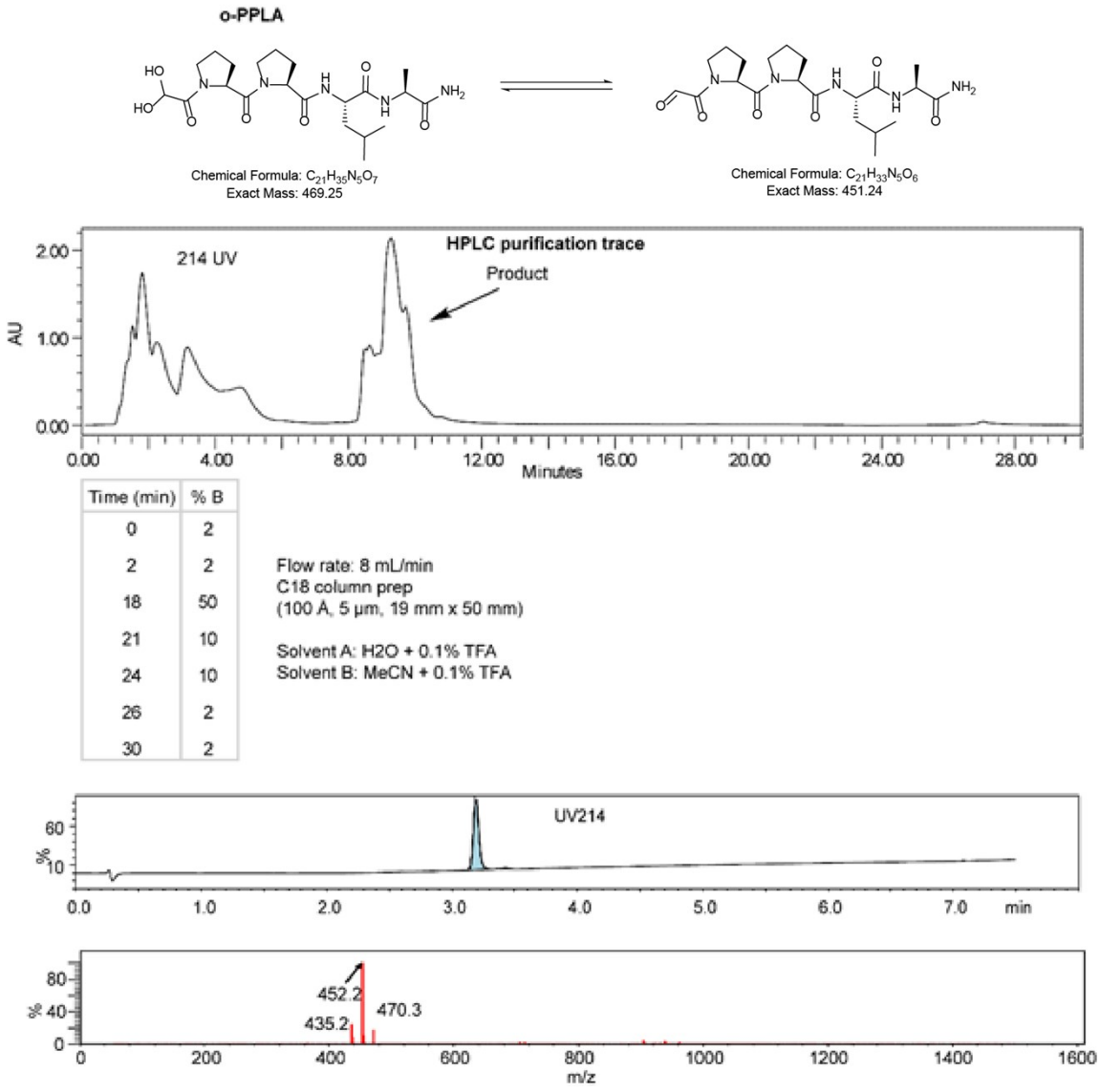


Figure S43. Summary for HCO-PPLA synthesis

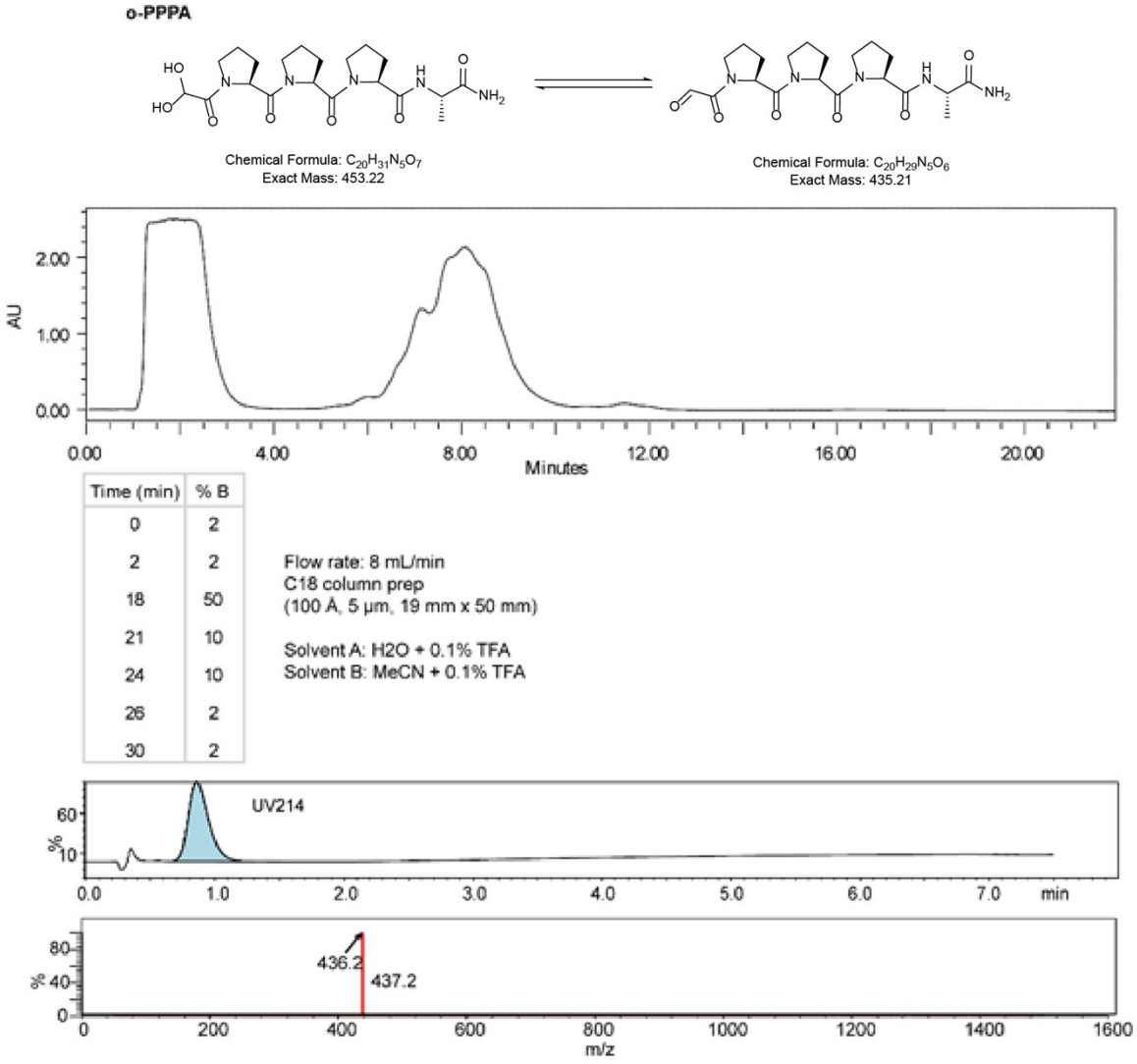


Figure S44. Summary for HCO-PPPA synthesis

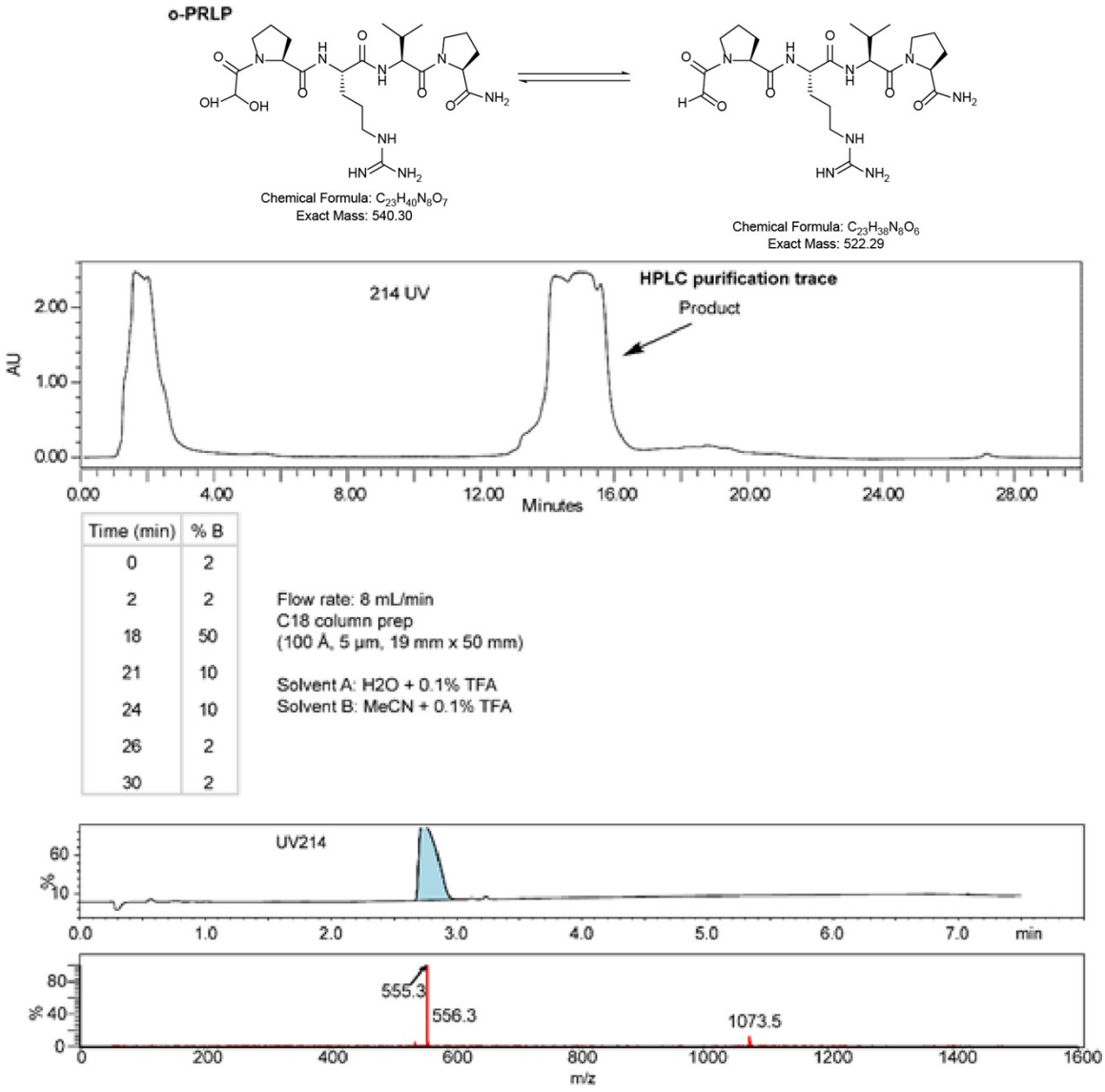


Figure S45. Summary for HCO-PRLP synthesis

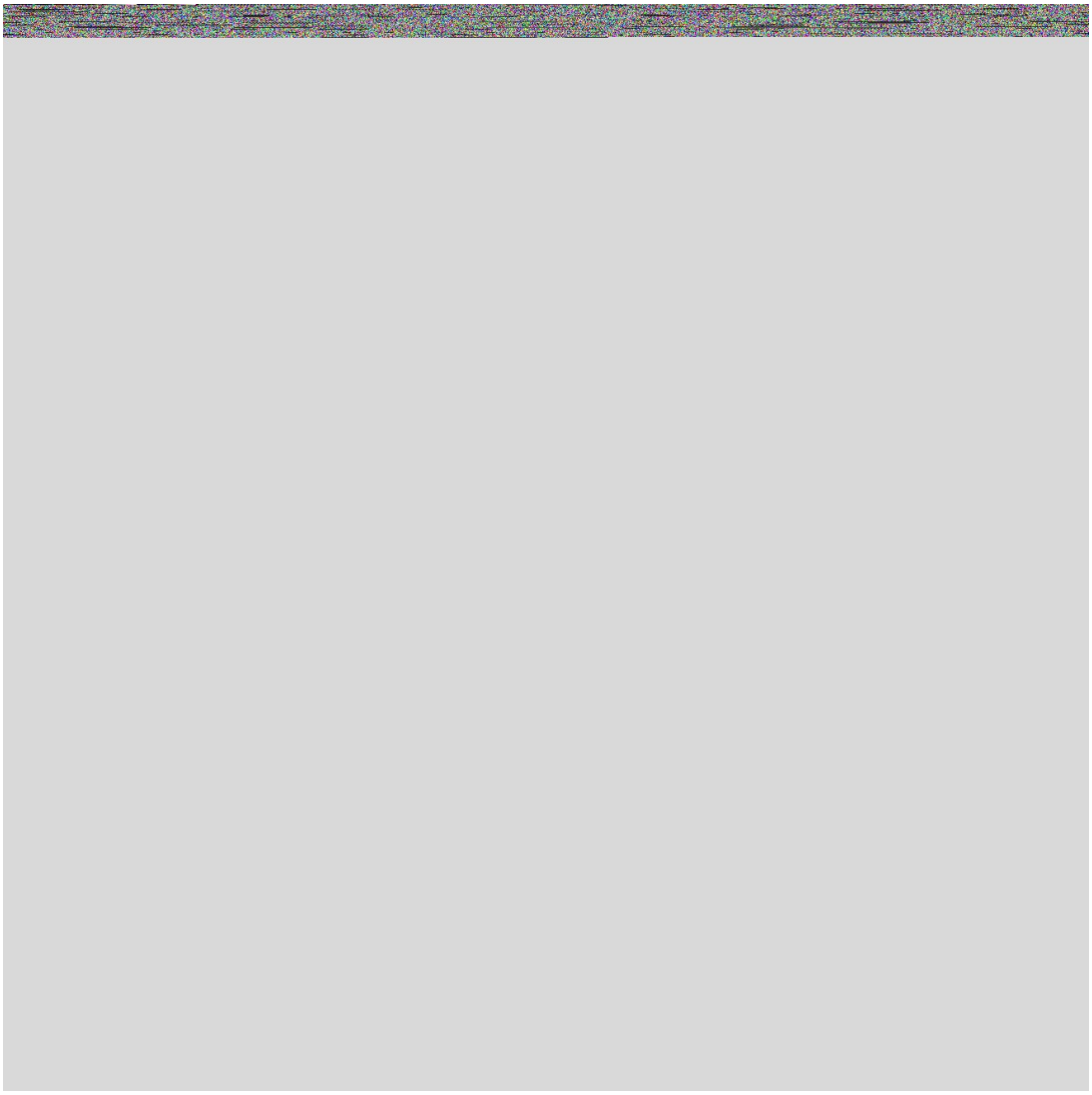


Figure S46. Summary for HCO-PQPL synthesis

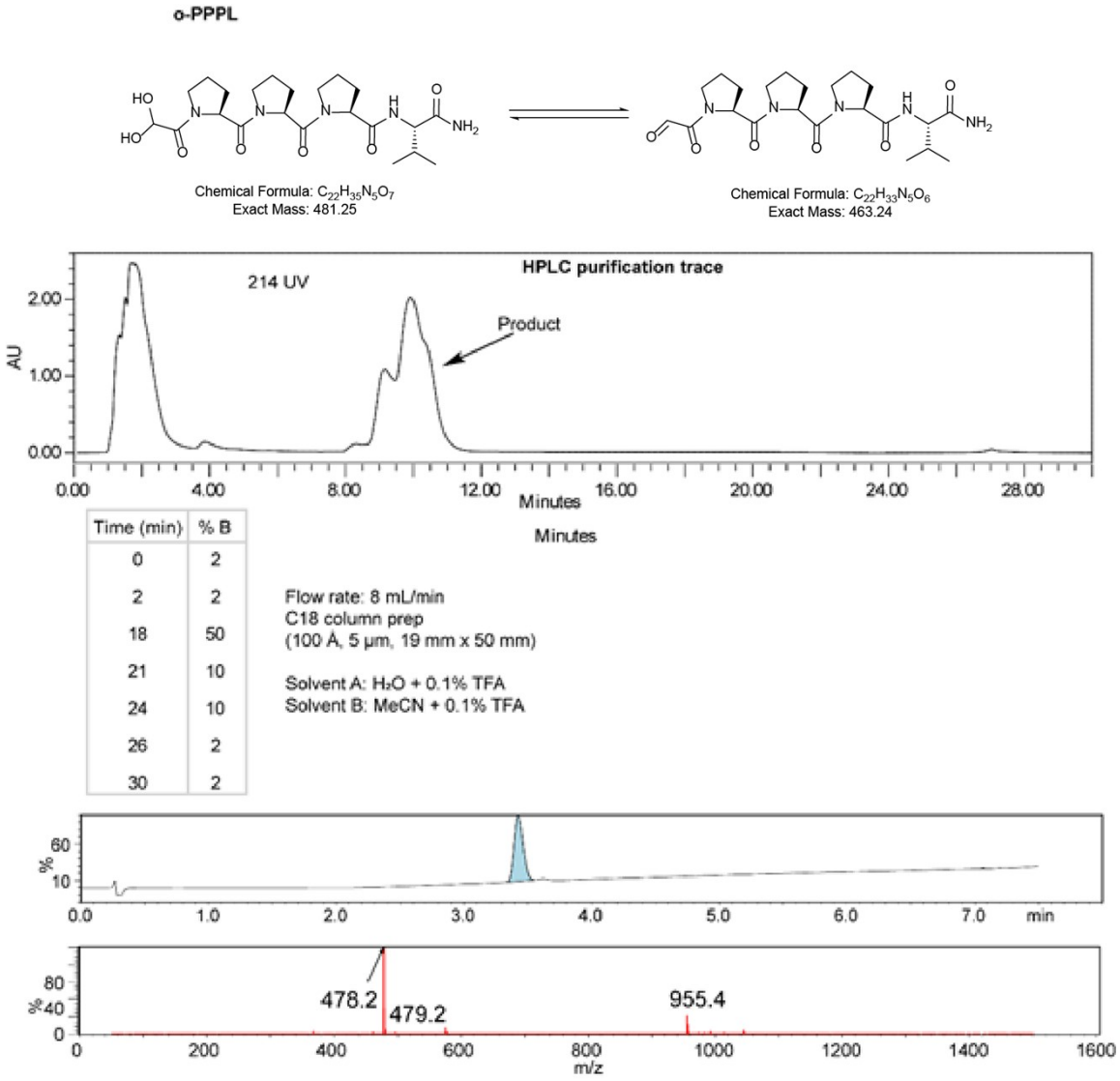


Figure S47. Summary for HCO-PPPL synthesis

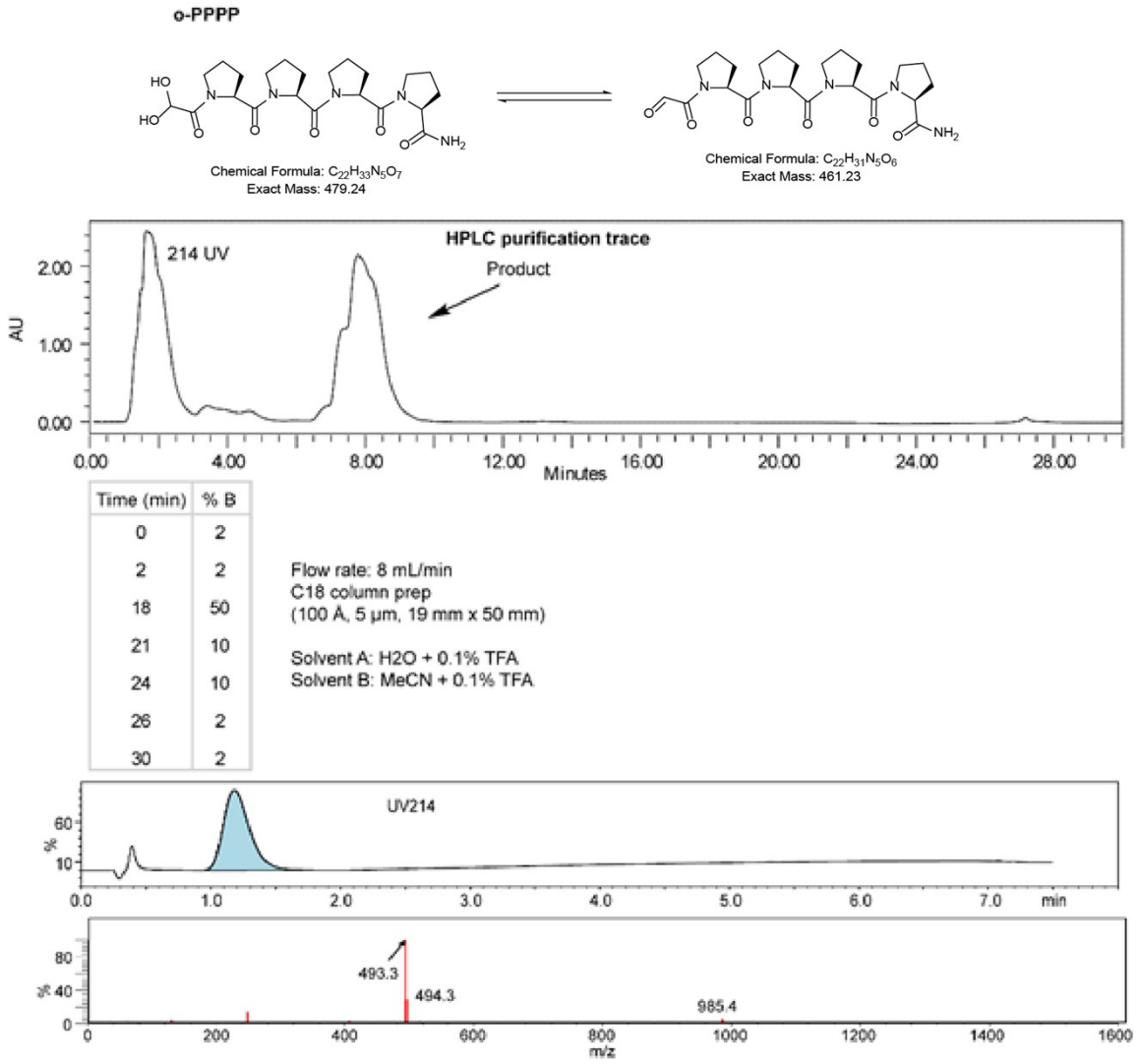


Figure S48. Summary for HCO-PPPP synthesis

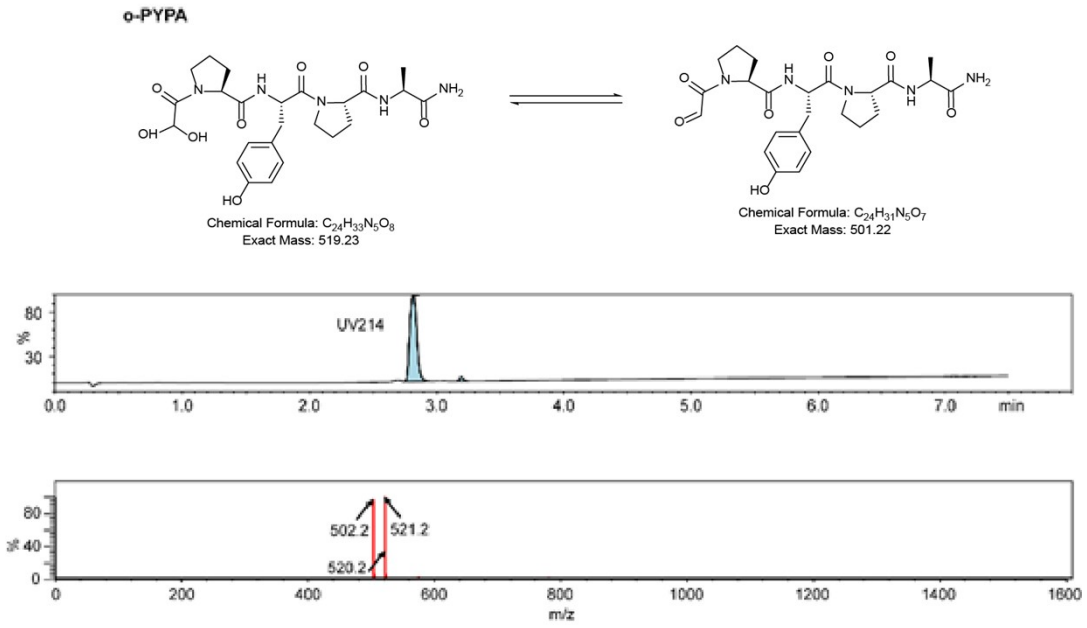


Figure S49. Summary for HCO-PYPA synthesis

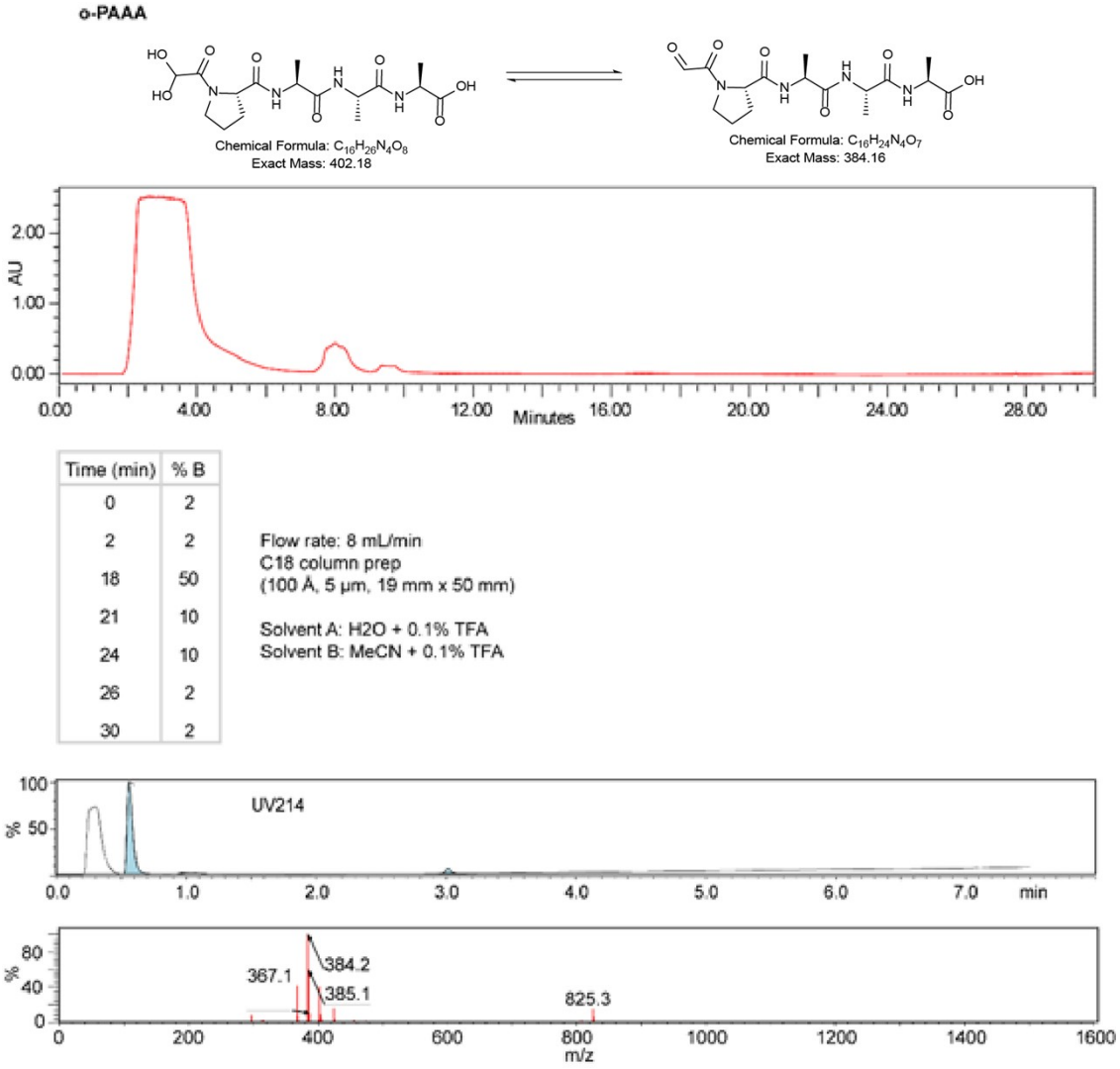


Figure S50. Summary for HCO-PAAA synthesis

Supporting information references

1. V. Triana and R. Derda, *Org. Biomol. Chem.*, 2017, **15**, 7869-7877.
2. T. L. Hwang and A. J. Shaka, *J. Magn. Reson., Series A*, 1995, **112**, 275-279.
3. Y.-W. Kim, T. N. Grossmann and G. L. Verdine, *Nature Protocols*, 2011, **6**, 761-771.
4. R. Robiette, J. Richardson, V. K. Aggarwal and J. N. Harvey, *J. Am. Chem. Soc.*, 2006, **128**, 2394-2409.
5. B. He, K. F. Tjhung, N. J. Bennett, Y. Chou, A. Rau, J. Huang and R. Derda, *Sci. Rep.*, 2018, **8**, 1214.
6. M. Sojitra, S. Sarkar, J. Maghera, E. Rodrigues, E. J. Carpenter, S. Seth, D. Ferrer Vinals, N. J. Bennett, R. Reddy, A. Khalil, X. Xue, M. R. Bell, R. B. Zheng, P. Zhang, C. Nycholat, J. J. Bailey, C.-C. Ling, T. L. Lowary, J. C. Paulson, M. S. Macauley and R. Derda, *Nat. Chem. Biol.*, 2021, **17**, 806-816.
7. K. F. Tjhung, P. I. Kitov, S. Ng, E. N. Kitova, L. Deng, J. S. Klassen and R. Derda, *J. Am. Chem. Soc.*, 2016, **138**, 32-35.
8. A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372-1377.
9. C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785-789.
10. S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200-1211.
11. P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623-11627.
12. W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257-2261.
13. R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724-728.
14. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. M. Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian 09 (Revision A.02) Gaussian, Inc., Wallingford CT, 2016.
15. C. Ge, E. Spånning, E. Glaser and Å. Wieslander, *Mol. Plant*, 2014, **7**, 121-136.
16. J. Xie, Z. Xu, S. Zhou, X. Pan, S. Cai, L. Yang and H. Mei, *PLOS ONE*, 2013, **8**, e74506.
17. T. Chen and C. Guestrin, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.
18. W. L. Matochko and R. Derda, *Comput. Math. Methods Med.*, 2013, **2013**, 491612.