Supplementary Information for

Unsupervised machine learning leads to an abiotic picomolar peptide ligand

Joseph S. Brown[1], Somesh Mohapatra[2,‡], Michael A. Lee[1], Roman Misteli[1,†], Yitong Tseo[2], Nathalie M. Grob[1], Anthony J. Quartararo[1,#], Andrei Loas[1], Rafael Gomez-Bombarelli[2]*, and Bradley L Pentelute[1,3-5]*

[1] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[2] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[3] The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States
[4] Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[5] Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States

[#]Current address: Fog Pharmaceuticals, Inc, 30 Acorn Park Dr, Cambridge, MA 02140, USA.
[†]Current address: Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom.
[‡]Current address: Caterpillar Inc, 5205 N O'Connor Blvd Ste. 100, Irving, TX 75039, USA.

*Email: rafagb@mit.edu, blp@mit.edu

Table of Contents

Supplementary Data: All canonical peptides and noncanonical peptidomimetics discovered from AS-MS experiments as well as peptides sampled from the canonical libraries (presumed nonbinders) are provided within the Github repository: https://github.com/josephsbrown1/Peptide-Map/.

Table of abbreviations

Abbreviation	**Full name**
AGC	Automatic gain control
AggCl	Agglomerative clustering
ALC	Average local confidence
AS-MS	Affinity selection-mass spectrometry
BLI	Biolayer interferometry
Boc	tert-Butyloxycarbonyl
BSA	Bovine serum albumin
CID	Collision induced dissociation
CV	Column volume
Da	Dalton mass unit
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCM	Dichloromethane
DIPEA or DIEA	N,N-diisopropylethylamine
DMF	N,N-dimethylformamide
ECFP_6	Extended connectivity Fingerprint
ESI	Electrospray ionization
ESM-2	Evolutionary scale model-2
EThcD	Electron-transfer dissociation with higher-energy collision
FBS	Fetal bovine serum
Fmoc	9-fluorenylmethyloxycarbonyl
HATU	1-[Bis(dimethylamino) methyl-ene]- 1H-1,2,3-triazolo[4,5-b]-pyridinium 3-oxide hexafluoro-phosphate
HCD	Higher-energy CID
HPLC	high pressure or high performance liquid chromatography
K Buffer	Kinetics buffer
LCMS	Liquid chromatography-mass spectrometry
MDS	Multidimensional scaling
MeCN	Acetonitrile
MEME	Multiple Em for Motif Elicitation
MeOH	Methanol
NHS	N-Hydroxysuccinimide
nLC	Nano liquid chromatography
PBS	Phosphate buffer saline
PCA	Principal component analysis
PEG	Polyethylene glycol
PTM	Post-translational modification

SA	Streptavidin
SAR	Structure activity relationship
STREME	Sensitive, Thorough, Rapid, Enriched Motif Elicitation
TFA	Trifluoroacetic acid
Trt	Trityl
UMAP	Uniform manifold approximation
XSTREME	Extreme Sensitive, Thorough, Rapid, Enriched Motif Elicitation

1 Materials

Canonical Fmoc-protected amino acids Fmoc-L-Ala-OH, Fmoc-L-Arg(Pbf)-OH; Fmoc-L-Asn(Trt)-OH; Fmoc-L-Gln(Trt)-OH; Fmoc-L-Leu-OH; Fmoc-L-Lys(Boc)-OH; Fmoc-L-Pro-OH; Fmoc-L-Ser(t-Bu)-OH; Fmoc-L-Tyr(t-Bu)-OH, Fmoc-L-Asp-(Ot-Bu)-OH; Fmoc-L-Glu(Ot-Bu)-OH; Fmoc-Gly-OH; Fmoc-L-Phe-OH; Fmoc-L-Thr(t-Bu)-OH; and Fmoc-L-Val-OH were purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-L-His(Boc)-OH was purchased from Advanced ChemTech and used as received. Fmoc-Rink amide linker (4-[(R,S)-(2,4-dimethoxyphenyl)(Fmoc-amino)methyl]phenoxyacetic acid) was purchased from Chem Impex Inc (Wood Dale, IL) and used as received.

Table S1. Noncanonical amino acids used in this work with their associated protecting groups. Unless specified as synthetically produced, all were purchased and used as received.

Noncanonical amino acid	Abbreviation	1-Letter Abbreviation	Source
Fmoc-L-Phe(2-trifluoromethyl)-OH	2F3F	v	Chem Impex, Inc
Fmoc-3-fluoro-L-phenylalanine	3fF	m	Chem Impex, Inc
Fmoc-4-(Boc-amino)-L-phenylalanine	4AF	k	Chem Impex, Inc
Fmoc-Asn(GlcNAc(Ac)$_3$-β-D)-OH	Agn	X	Millipore Sigma
Fmoc-α-aminoisobutyric acid	Aib	b	Chem Impex, Inc
Fmoc-(4-aminomethyl) benzoic acid	Amb	h	Chem Impex, Inc
Fmoc-azetidine-3-carboxylic acid	Aza	a	Chem Impex, Inc
Fmoc-β-cyclopropyl-L-alanine	Cpa	d	Chem Impex, Inc
Fmoc-(4-tert-butyloxycarbonyl)-L-phenylalanine	Cxf	t	Chem Impex, Inc
Fmoc-3,4-difluoro-L-phenylalanine	DfF	r	Chem Impex, Inc
Fmoc-4-diethylphosphomethyl-L-phenylalanine	Dpf	z	Chem Impex, Inc
Fmoc-3,3-diphenyl-L-alanine	DPh	w	Chem Impex, Inc
Fmoc-L-HomoArg(Pbf)-OH	hArg	o	Chem Impex, Inc
Fmoc-L-homocitrulline	hCit	p	Chem Impex, Inc
Fmoc-O-tert-butyl-L-trans-4-hydroxyproline	Hyp	e	Chem Impex, Inc
Fmoc-L-methionine sulfone	Msn	l	Chem Impex, Inc
Fmoc-3-(1-naphthyl)-L-alanine	Nal	u	Chem Impex, Inc
Fmoc-pentafluoro-L-phenylalanine	PfF	y	Chem Impex, Inc
Fmoc-4-phenylpiperidine-4-carboxylic acid	Php	s	Chem Impex, Inc
1-Boc-piperidine-4-Fmoc-amino-4-carboxylic acid	Pip	f	Chem Impex, Inc
Fmoc-(S)3-amino-2-(phenylsulfonylamino)propionic acid	Psa	x	Chem Impex, Inc
Fmoc-O-benzylphospho-L-serine	pSer	n	Chem Impex, Inc
Fmoc-3-(4-thiazolyl)-L-alanine	Tha	i	Chem Impex, Inc
Fmoc-4-amino-tetrahydropyran-4-carboxylic acid	Thp	g	Chem Impex, Inc
Fmoc-(3S-)-1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid	Tic	j	Chem Impex, Inc
Fmoc-Bispyridinolysine-OH	Bpl	B	Synthesized, see Noncanonical Monomer Synthesis
Fmoc-D-Galactosyl-L-citrulline	Git	Z	Synthesized, see Noncanonical Monomer Synthesis

For the synthesis of noncanonical monomers (Bpl and Git, see *Noncanonical Monomer Synthesis*), Fmoc-Lys-OH was purchased from Ambeed Inc. Sodium triacetoxyborohydride, 2-pyridinecarboxaldehyde, 1,2-dichloroethane, methanol, (D)-(+)-galactose, acetic anhydride and pyridine were purchased from MilliporeSigma. Fmoc-Cit-OH was purchased from Chem-Impex International Inc (Wood Dale, IL). Coupling agent O-(7-azabenzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate (HATU, ≥97.0%) was purchased from P3 Biosystems (Lyndon, Kentucky).

Biosynthesis OmniSolv® grade N,N-dimethylformamide (DMF) was purchased from EMD Millipore (DX1732-1) and incubated with 1 pack of AldraAmine trapping agents (for 1000 – 4000 mL DMF, Sigma-Aldrich, catalog number Z511706) for 48 hours prior to use. Diisopropylethylamine (DIEA; 99.5%, biotech grade, catalog number 387649) and piperidine (ACS reagent, ≥99.0%) were purchased from Sigma-Aldrich. Formic acid (FA, 97%) was purchased from Beantown Chemical, Corp. Trifluoroacetic acid (HPLC grade, ≥99.0%), Diethyl ether (anhydrous, ACS reagent, ≥99.0%), acetonitrile (HPLC grade, ≥99.9%), Omnisolv® acetonitrile (LC-MS grade, AX0156-1), Omnisolv® water (LC-MS grade, WX0001-1) and were purchased from Sigma-Aldrich. Formic acid Optima LC/MS (A117) was purchased from Fisher Chemical. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 μm syringe filters were TISCH brand SPEC17984.

H-Rink Amide-ChemMatrix® (0.49 mmol/g) resin was purchased from PCAS Biomatrix (St-Jean-sur-Richelieu, Quebec, Canada) and 20 μm TentaGel® M NH_2 Monosized Amino Microsphere resin was purchased from Rapp Polymere Inc. (Tübingen, Germany). HyClone™ Fetal Bovine Serum (SH30071.03HI, heat inactivated) was purchased from GE Healthcare Life Sciences (Logan, UT) Dynabeads MyOne Streptavidin T1 magnetic microparticles were purchased from Invitrogen (Carlsbad, CA). Phosphate buffered saline (10x, Molecular biology grade) was purchased from Corning. Sodium chloride (ACS grade) was purchased from Avantor. Guanidine hydrochloride (Cat BP178) and sodium phosphate monobasic monohydrate were purchased from Fisher Scientific.

Mouse anti-hemagglutinin antibody (clone 12ca5) was purchased from Columbia Biosciences Corporation (Cat: 00-1722, Frederick, Maryland) biotin-$(PEG)_4$-NHS ester and biotin-$(PEG)_4$-propionic acid were purchased from ChemPep Inc. (Wellington, FL). Biotinylation of 12ca5 was performed as previously described.[1]

2 Peptide and peptidomimetic library synthesis

2.1 Canonical peptide library synthesis

A total of three libraries were prepared, each portioned into 5 aliquots each (15 aliquots total), with 12 sampled in affinity selection-mass spectrometry experiments. The procedure below describes the synthesis of a single library.

Total number of beads:	1 x 10^{9}
Size:	20 micron Tentagel M NH2 (Cat: M30202)
Library design:	X_{12}K-NH_2
Variable Positions	12
# of monomers	18 (Canonical 20 minus Ile,Cys)
Ala, Asp, Glu, Phe, Gly, His, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr	
Theoretical diversity	1.16 x 10^{15}
Redundancy	4.32 x 10^{-7}
Note: Redundancy is Total number of beads in each library / Theoretical diversity or 1.16 x 10^{15} / 1 x 10^{8} and speaks to the sampling rate of the theoretical sequence space available	

2.2 Noncanonical peptidomimetic library synthesis

A single library was prepared, each portioned into 5 aliquots (5 aliquots total), with 3 sampled in affinity selection-mass spectrometry experiments.

Total number of beads:	1 x 10^{9}
Size:	20 micron Tentagel M NH2 (Cat: M30202)
Library design:	X_{12}K-NH_2
Variable Positions	12
# of monomers	36
Noncanonical monomers: Aze, Aib, Bpl, Cpa, Hyp, Pip, Thp, Amb, Tha, Tic, 4AF, Msn, 3fF, pSer, hArg, hCit, DfF, Php, CxF, Nal, 2F3F, DPh, Psa, Agn, PfF, Dpf, Git Also, the following canonicals were included: Ala, Asp, Gly, His, Pro, Gln, Thr, Val, and Tyr, as well as Lys that was only included at the C-terminus position.	
Theoretical diversity	4.74 x 10^{18}
Redundancy	1.06 x 10^{-10}

Note: For both canonical and noncanonical library synthesis, these libraries are highly 'nonredundant,' meaning the theoretical sequence is under-sampled. The successful discovery of high-affinity peptide binders is dependent on the presence of the minimal required motif / sequence required for binding. Low-complexity binding motifs defined by 3-5 amino acids are readily discovered because they are statistically common even within a highly nonredundant library. Since the library is highly nonredundant, sequence isomers can be confidently identified and removed (see *Curation of AS-MS Data*) as they are highly unlikely to exist.

2.3 Solid-phase peptide library synthesis by split-pool synthesis

4.2 g of 20 µm TentaGel M NH2 resin (0.26 mmol/g, 1.1 mmol, 1.0 x 10^9 beads) was swollen in and washed with DMF (3x) within a 250mL peptide synthesis vessel (medium frit, 10-15 µm pore size, ChemGlass CG-1866-05). Fmoc-Rink amide linker (2.9 g, 5.4 mmol, 5 eq) was dissolved in HATU solution (0.38 M in DMF, 12.9 mL, 4.5 mmol), activated with DIEA (2.7 mL, 16 mmol) immediately prior to coupling, and added to resin bed. Coupling was performed for 30 min and then washed with DMF (2 x 100 mL). Fmoc removal was completed with 20% piperidine in DMF (1 x 50 mL flow wash followed by 2 x 50 mL, 5 min batch treatments). Resin was then washed with DMF (3 x 150 mL). This process of coupling and Fmoc deprotection was repeated with the Fmoc-Lys(Boc)-OH (2.54 g, 5.4 mmol, 5 eq).

The resin was then **split** for the coupling of randomized (“X”) positions with the library amino acids. The resin was suspended in DMF (50 mL) and carefully divided evenly among HSW Norm-Ject syringes (Torviq) mounted on Restek Resprep SPE vacuum manifolds equipped (Cat 26077) with valves for coupling of each amino acid monomer in the library (i.e., for canonical synthesis: 18 syringes; for noncanonical synthesis 36 syringes).

With the Resprep valves closed, Fmoc-protected amino acids (0.6 mmol, 10 eq relative to resin) in HATU solution (0.38 M in DMF, 1.4 mL, 0.54 mmol, 0.9 eq relative to amino acid) were activated with DIEA (1.2 mmol, 2 eq relative to amino acid) and each added to their respective split resin (theory: ~260 mg resin, 60 µmol). Couplings proceed for one hour minimum. For Fmoc-Bpl-OH, 5.0 equiv. of DIEA relative to amino acid was used. For precious amino acids, lower equivalents were used: Fmoc-Blp-OH (6.6 equiv.), Fmoc-Git(OAc)$_4$-OH (4.7 equiv.), Fmoc-Dpf-OH (3.8 equiv.) and Fmoc-Agn(OAc)$_3$-OH (2.3 equiv.) with extended coupling times up to three hours. After coupling was completed, the Resprep valves were opened to remove the excess coupling solution from the resin.

All resin was then **pooled** into the 250 mL peptide synthesis vessel and the syringes were washed (3 x 5 mL) to recombine all resin. Additional wash (2 x 100 mL) and Fmoc deprotection (1 x 50 mL flow wash followed by 2 x 50 mL, 5 min batch treatments) with 20% piperidine in DMF. Resin was washed with DMF (3 x 100 mL) and was then ready again for the next split cycle. The cycle was iterated 12 times total to accomplish the $X_{12}K$-NH_2 design.

2.3.1 Portioning

With the final N-terminal Fmoc group was removed, the resin was washed with DMF (150 mL), then suspended in DMF (~ 50 mL) and divided evenly among 5 aliquots in 20 mL syringes (2 x 10^8 peptides per aliquot). Then each were washed with DCM (3x) and dried under reduced pressure overnight. Resin was taken to perform experiment to validate the quality of the library, see *Library Validation Analysis*.

2.3.2 Cleavage from resin and solid phase extraction

Deacetylation of peracetylated noncanonical side-chains (Agn, Git) was carried out by treatment of resin with a solution of 5% anhydrous hydrazine in DMF for 16 h at ambient temperature. After deacetylation, the resin was washed with DMF (3x), DCM (3x), DMF (3x), MeOH (3x) and DCM (3x) and dried under reduced pressure.

Canonical libraries were globally deprotected and cleaved from resin with 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane, for 3 h at ambient temperature (~2 mL/mg of resin). Noncanonical libraries were globally deprotected and cleaved from resin with 85% (v/v) TFA, 5% (v/v) water, 5% (v/v) phenol and 5% (v/v) thioanisole for 2 h at ambient temperature (TIPS was found to reduce the GlcNAc of the Agn side chain).

The crude peptides were triturated with cold diethyl ether. Precipitated peptide was triturated (3x) with cold diethyl ether, dissolved in 50% acetonitrile in water (0.1% TFA), passed through a 0.2 µm nylon syringe filter, and lyophilized.

2.3.3 Solid-phase extraction

Crude lyophilized powders were resuspended in 5% acetonitrile in water (0.1% TFA) purified using Supelco Discovery® DSC-18 SPE Tubes (Millipore Sigma Cat: 52607-U). The SPE tube was first conditioned with 3 CV of acetonitrile (0.1% TFA) and then equilibrated with 5 CV of 5% acetonitrile in water (0.1% TFA). Then, the suspended crude was loaded (Maximum 150 mg crude peptide loaded onto 2 g bed mass) and washed with 10-12 CV of 5% acetonitrile in water (0.1% TFA). Peptides were eluted with 70% acetonitrile (0.1% TFA) and lyophilized.

2.3.4 Preparation of library stock solutions

Lyophilized, SPE-purified powders of libraries were each dissolved first in DMF and then diluted with 1x PBS to a final library concentration of 8 mM (~40 pM/member), and a final DMF concentration of 5% (v/v). Stock solutions were aliquoted out into low-bind tubes and stored at -80 °C. Aliquots were thawed on ice prior to use.

3 Library Validation Analysis

Canonical libraries were validated as previously described.[1] For the noncanonical library, 20 mg of resin was weighed out in a microcentrifuge tube and agitated for 16 h in 5% anhydrous hydrazine in DMF (100 mg/mL). The resin was then transferred to a 3 mL fritted Torviq syringe and washed with DMF (3x), DCM (3x), DMF (3x), MeOH (3x) and DCM (3x). The resin was suspended in DCM and transferred to a 15 mL conical tube and the solvent was evaporated under a stream of nitrogen.

For both the canonical and noncanonical libraries, 1.5 mg of dried resin was weighed out and suspended in DMF (5 mg/mL). From this stock suspension, 1.5 µL (estimated 877 beads) were transferred to a microcentrifuge tube, suspended in 200 µL cleavage solution. Canonical libraries were treated with 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane and heated to 60 ºC for 10 minutes. Noncanonical libraries were treated with 85% (v/v) TFA, 5% (v/v) water, 5% (v/v) phenol and 5% (v/v) thioanisole) and left at room temperature for 2 hours. The TFA was then evaporated under a stream of nitrogen and the remaining waxy oil was dissolved in 200 µL of 5% acetonitrile in water (0.1% TFA) and sonicated / vortex vigorously. The suspension was centrifuged at 21,300 rcf at room temperature. The supernatant was added onto a conditioned C18 STAGE tip (CDS Empore™ SDB-XC, Fisher Scientific Cat: 13-110-020) and purified according to the protocol of Rappsilber et al.[2] The eluting solvent was evaporated by vacuum centrifugation and the peptides were resuspended in 29 uL of 0.1% formic acid in water to enable the injection of 100 pg/peptide with 1 µL. The solution was centrifuged at 21,300 rcf at 4°C for 10 min and the supernatant was transferred to a MS vial for Orbitrap analysis. Upon analysis of the canonical and noncanonical libraries, the canonical library demonstrated near even monomer incorporation as previously reported.[1] However, within the noncanonical library, higher monomer variation was observed, with Bpl (Fmoc-Bispyridinolysine-OH) and PfF (Fmoc-pentafluoro-L-phenylalanine) showing poor incorporation at all positions. FfF (Fmoc-pentafluoro-L-phenylalanine) has previously been successfully incorporated into other noncanonical libraries. Additionally, the hydrazinolysis of for deacetylation of the glycan-mimetic functional groups (Agn, Git) was suspected to affect the slightly lower incorporation of Psa. Despite these shortcomings in the noncanonical library, it was used in AS-MS experiments as follows.

4 AS-MS experiments

Affinity selection-mass spectrometry (AS-MS) was performed manually as previously described with modifications[1] or with a KingFisher Duo Prime (Thermo Fisher Scientific).

For manual AS-MS, 100 µL of magnetic beads (1 mg; 0.13 nmol IgG binding capacity, MyOne Streptavidin T1 Dynabeads, Thermo Fisher Scientific Cat: 65602) were transferred to 1.7 mL plastic centrifuge tubes and washed 3 times with blocking buffer (10% fetal bovine serum (FBS) in 1x PBS pH 7.4 and 0.01% Tween20, 0.2 µm filtered) using a magnetic separation rack (NEB Cat: S1506S). Then, 1.2 to 2 eq of biotinylated anti-hemagglutinin antibody (clone 12ca5, Columbia Biosciences Cat: 00-1722) was incubated with the magnetic beads at approximately 0.5 µM. The resulting suspensions were incubated on a nutating mixer for 30 min at 4 ºC and then washed 3 times with blocking buffer.

Next, the affinity selection samples were prepared. The peptide library was depleted of 'bead binders.' In a new tube, the following were combined for a 1mL sample and scaled if needed for multiple replicates using the library: 100 uL of neat FBS, 550 uL of 1x PBS, 250 uL of library stock solution to provide 10 fmol/peptide, and 50 uL of pre-washed magnetic beads. This sample was incubated for 1 hour at 4 ºC. Then, this sample was then centrifuged at 21,300 rcf and the supernatant aliquoted to a new tube to provide the library depleted of peptides that bind to the magnetic beads with high affinity. Then, 1 mg (100 uL volume in blocking buffer) of the washed magnetic beads with 12ca5 immobilized was mixed with the pre-depleted library solution to provide a solution concentration of 100-130 nM of 12ca5 final. These affinity selection samples were then incubated at 4 ºC for 1 hour on a nutating mixer. Then, the samples were washed 3-6 times with cold 1x PBS pH 7.4 using a magnetic rack (~10 minutes contact time with buffer). The isolated beads were eluted using 2 x 100 uL of 6 M guanidine, 50 mM sodium phosphate pH 7.

For automated selections, a KingFisher Duo Prime was utilized with two (2) x 96 Deepwell Plates (Thermo Fisher, #95040450) in the following format, marked by rows. Three replicates were run by using three columns per library aliquot for 12 separate $X_{12}K$ libraries. The isolated peptides bound to the beads were eluted using 2 x 100 uL of 6 M guanidine, 50 mM sodium phosphate pH 7 in elution strips.

	Plate 1		Plate 2	
Row	Description	Vol, mL	Description	Vol, mL
A	Selection samples, see text	1	1x PBS, cold	1
B	Blocking buffer	1	1x PBS, cold	1
C	Blocking buffer	1	1x PBS, cold	1
D	Blocking buffer	1	1x PBS, cold	1
E	Biotinylated 12ca5	0.5	1x PBS, cold	1
F	Blocking buffer	1	1x PBS, cold	1
G	Blocking buffer	1	Comb for Kingfisher magnet	
H	Blocking buffer + beads	1		

	Elution strip 1		Elution strip 2	
Row	Description	Vol, mL	Elution strip 2	Vol, mL
N/A	6 M guanidine, 50 mM sodium phosphate, pH 7	0.1	6 M guanidine, 50 mM sodium phosphate, pH 7	0.1

For the “Selection samples” (Plate 1 Row A), the sample was prepared similarly to the manual selection. First, the peptide library was depleted of ‘bead binders.’ In a new tube, the following were combined for a each sample and scaled if needed for multiple columns / replicates: 100 uL of neat FBS, 550 uL of 1x PBS, 250 uL of library stock solution to provide 10 fmol/peptide, and 50 uL of pre-washed magnetic beads. This sample was incubated for 1 hour at 4 ºC. Then, this sample was then centrifuged at 21,300 rcf and the supernatant aliquoted to the 96 Deepwell plate to provide the library depleted of peptides that bind to the magnetic beads with high affinity.

For “Blocking buffer + beads” (Plate 1 Row H), 100 µL of magnetic beads were added to 900 uL of blocking buffer (10% fetal bovine serum (FBS) in 1x PBS pH 7.4 and 0.01% Tween20, 0.2 µm filtered).

For “Biotinylated 12ca5” (Plate 1 Row E), 500 uL of blocking buffer was added with the amount needed to provide 1.2-2 eq of 12ca5 from its stock solution (typically 10.4 uL of 12ca5 stock solution at 25 µM for 2 eq).

The following steps were programmed for affinity selection:

1. Collect comb from Plate 2, Row G
2. Wash beads by release beads (30 s, medium) in Plate 1, Row H, collect beads (3 x 1 second)
3. Wash beads as in Step 2 in Plate 1, Row G
4. Wash beads as in Step 2 in Plate 1, Row F
5. Release beads (20 s, medium) into Plate 1 Row E (30 minutes, mix slowly)
6. Wash beads as in Step 2 in Plate 1, Row D
7. Wash beads as in Step 2 in Plate 1, Row C
8. Wash beads as in Step 2 in Plate 1, Row B
9. Release beads into Plate 1, Row A, (1 hour, mix slowly)
10. Add plate 2, containing cold 1x PBS to the Kingfisher instrument
11. Collect beads from Plate 1, Row A (5 x 1 second)
12. Wash beads as in Step 2 in Plate 2, Row A
13. Wash beads as in Step 2 in Plate 2, Row B
14. Wash beads as in Step 2 in Plate 2, Row C
15. Wash beads as in Step 2 in Plate 2, Row D
16. Wash beads as in Step 2 in Plate 2, Row E
17. Wash beads as in Step 2 in Plate 2, Row F
18. Release beads into elution strip 1, 1 minute mix fast, collect beads (5 x 1 s)
19. Release beads into elution strip 2, 1 minute mix fast, collect beads (5 x 1 s)
20. Release beads and comb into Plate 2 Row G to end the program

Eluted peptide samples were then prepared for Orbitrap analysis by C18 STAGE tip (CDS Empore™ SDB-XC, Fisher Scientific Cat: 13-110-020) and purified according to the protocol of

Rappsilber et al.[2] The eluting solvent was evaporated by vacuum centrifugation and the peptides were resuspended in 12-13 uL of 0.1% formic acid in water. The solution was centrifuged at 21,300 rcf at 4°C for 10 min and the supernatant was transferred (leave behind 1.5 uL) to a MS vial for Orbitrap analysis. Usually, 4-5 uL were injected onto the Orbitrap Fusion Lumos whereas 2-3 uL were injected onto the Orbitrap Eclipse.

5 nLC-MS/MS

Nanoscale liquid chromatography tandem mass spectrometry (nLC-MS/MS) was performed using an EASY-nLC 1200 (Thermo Fisher Scientific) nano-liquid chromatography handling system connected to an Orbitrap Fusion Lumos or an Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo Fisher Scientific). Solvent A is water (0.1% formic acid) and solvent B is 80% acetonitrile in water (0.1% formic acid). Precolumn and analytical column equilibration with 8 µL of solvent A was performed at maximum of 1 µL/min or 600 bar. Samples were injected and loaded onto a nanoViper Trap Column (C18, 3 µm particle size, 100 A pore size, 20 mm x 75 µm ID; Thermo Fisher Scientific, Cat: 164946) for desalting with 12 µL of solvent A (maximum of 1 µL/min or 600 bar). The autosampler wash was 100 uL of solvent A. After trapping, samples were injected onto a PepMap RSLC C18 column (2 µm particle size, 15 cm x 50 µm ID; Thermo Fisher Scientific, Cat: ES901). The standard nano-LC method was run at 40 °C and a flow rate of 300 nL/min with the following gradient, expressed in % solvent B in solvent A: 1% to 41% over 120 minutes (*AS-MS Experiments*) or 90 minutes (*Library Validation Analysis* or other simple mixtures), move to 90% in 3 minutes, hold for 7 minutes, and then perform 2 “seesaw” washes (each comprising of moving to 20% over 3 minutes, holding at 20% for 3 minutes, moving to 90% for 3 minutes, and holding at 90% for 3 minutes).

Mass spectrometry acquisition was performed using an Orbitrap Fusion Lumos or an Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo Fisher Scientific) with positive mode, where the ion source settings was set by the tune parameters (Spray voltage usually ~ 2200 V with no Arb gas). The method to perform data-dependent acquisition has been iteratively optimized.

The standard AS-MS MS analysis method analyzes from 3-120 minutes, with an expected LC peak width of 20 seconds, default charge state of 3, and no internal mass calibration. Primary spectra acquisition in positive mode was observed by the Orbitrap with resolution = 120,000, using quadrupole isolation, 200-1400 m/z, RF Lens 30%, 250% AGC Target (auto injection time, usually < 10 ms), and 1 microscan. Secondary MS was performed with the following filters: Precursor selection range: 300-1200 m/z, MIPS: Peptide, Intensity threshold: 4e4, Charge state: 2-5 excluding undetermined charge states, Dynamic exclusion: exclude after 1 time for 30 seconds (10 ppm tolerance), Targeted mass exclusion of all peptides in the Pierce™ Peptide Retention Time Calibration Mixture (z = 2 and 3, Thermo Fisher Scientific, Cat: 88321). HCD and EThcD were completed. HCD used quadrupole isolation (1.3 m/z, no offset) at a fixed 28% collision energy and was observed on the Orbitrap with resolution = 30,000, Scan Range Mode: Define First Mass: 120 m/z, 600% AGC Target, maximum injection time 100 ms, and 2 microscans. EThcD used a charge filter of z ≥ 3, quadrupole isolation (1.3 m/z, no offset), using calibrated charge-dependent ETD activation, and supplemental HCD activation a fixed 25% collision energy and was observed on the Orbitrap with resolution = 30,000, Scan Range Mode: Define First Mass: 120 m/z, 600% AGC Target, maximum injection time 100 ms, and 2 microscans.

6 Curation of AS-MS Data

De novo analysis of sequencing data was performed as described previously for canonical libraries using PEAKS Studio 8.5 (Bioinformatics Solutions, Inc, ON, Canada).[1] Mass precursor correction was used. Auto *de novo* sequencing was performed using a 15 ppm precursor mass error and 0.02 Da fragment mass error. For canonical libraries, the following PTM modifications were used: fixed C-terminal amidation (-.98 Da) on lysine, and variable oxidation on methionine (+15.99 Da). For noncanonical libraries, the PTMs used are shown in Table S2. 20 candidate sequences were obtained for each preprocessed scan. Post-*de novo* data analysis was performed as previously described[3] to convert the PTMs to 1-letter encoding also in Table S2.

Table S2. Post-translational modification (PTM) utilized in PEAKS *de novo* sequencing analysis of noncanonical library. Where a single amino acid is modified (e.g., F modified to be F(+17.99) to represent 3fF), a fixed PTM is used. When the same amino acid can be modified to represent multiple noncanonical amino acids (e.g., alanine), a variable PTM was used.

Monomer	PTM	1-letter code
Aze	A(+12.00)	a
Aib	A(+14.02)	b
Cpa	A(+40.03)	d
Hyp	A(+42.01)	e
Pip	A(+55.04)	f
Thp	A(+56.03)	g
Amb	A(+62.02)	h
Tha	A(+82.98)	i
Tic	A(+88.03)	j
4AF	A(+91.04)	k
Msn	M(+31.99)	l
3fF	F(+17.99)	m
pSer	S(+79.97)	n
hArg	R(+14.02)	o
hCit	N(+57.06)	p
hCit	A(+100.06)	c
DfF	C(+80.04)	r
Php	E(+58.06)	s
CxF	A(+120.02)	t
Nal	L(+84.00)	u
2F3F	A(+144.02)	v
DPh	W(+37.02)	w
Psa	A(+155.00)	x
PfF	A(+165.98)	y
Dpf	A(+226.08)	z
Bpl	A(+239.14)	B
Agn	A(+246.09)	X
Git	A(+248.10)	Z

6.1 Removal of sequence isomers

After concatenating all data from *de novo* sequencing, the data was rigorously cleaned to remove poorly sequenced peptides and sequence isomers from the data, beyond what has previously been published.[3]

First, simple filters on the average local confidence of sequencing (ALC) and calculated ppm error of sequencing from PEAKS Studio 8.5 were applied: ALC > 85 (canonical) or > 80 (noncanonical) and absolute ppm error < 10 ppm were retained. Also, all duplicate peptides were removed. Also,

Second, all sequences were compared pairwise and marked for removal if they had the same precursor mass within 0.01 Da or had specific differences in precursor mass corresponding to 1) incorrect monoisotopic precursor selection (absolute delta of 1, 2, or 3 Da), oxidation (absolute delta of 16, 32), or sodium adduct (absolute delta of 22). Additionally, the peptides must have some amount of sequence similarity (empirically seen to work well on trial datasets with a similarity of 0.69 by difflib.SequenceMatcher in Python). Retention time differences were not considered in case the data was acquired using different gradients. The highest ALC peptide was retained, with the lowest ppm sequencing error as tie-breaker.

Third, all remaining sequences were compared pairwise and marked for removal based only on a very high degree of sequence similarity. Again using difflib.SequenceMatcher in Python, a peptide similarity of > 0.92 was only seen for sequence isomers with either a single amino acid replacement or a dipeptide swap with the X12K type of peptides. While rigorous and potentially overly conservative, this step often removes < 5% of the remaining data after the second step is completed.

With the canonical library, 4104 peptides were uniquely identified from AS-MS with high sequencing fidelity for unsupervised learning analysis.

With the noncanonical library, 17 peptides were uniquely identified from AS-MS with high sequencing fidelity for unsupervised learning analysis.

6.2 Characteristics of the peptides sampled from the original peptide libraries (presumed to be nonbinders)

From the library validation analysis of the canonical library, 5,047 peptides were identified by sampling the original library before AS-MS. In all cases except the sensitivity analysis in Figure 3, these peptides were added to PCA- and UMAP-constructed maps without re-learning. MDS is unable to add additional data to its sequence map without re-learning.

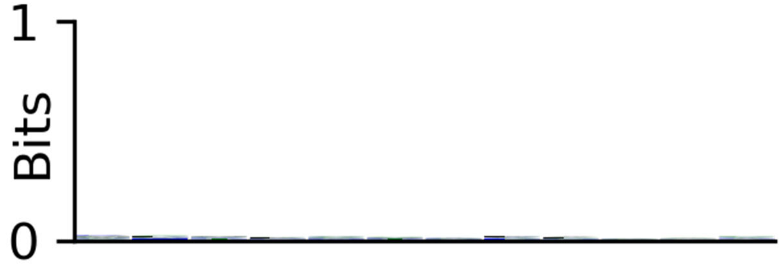


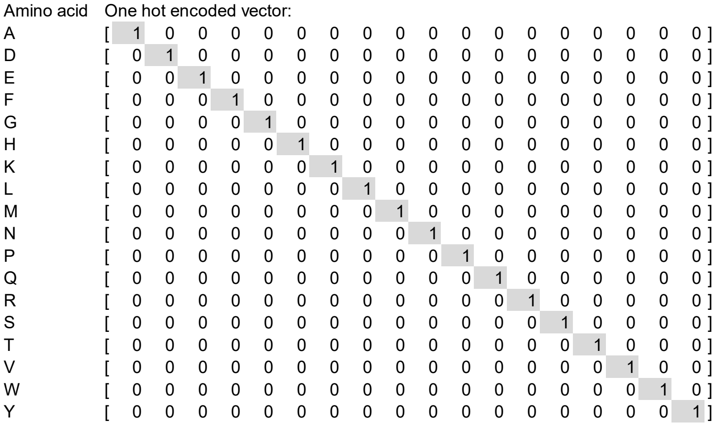
Figure S1. Logo plot of the peptides sampled from the X12K library, presumed to be nonbinders. Essentially no residues are shown, even at this zoomed y-scale, meaning that the peptides are largely random. This is corroborated by the unsupervised clustering seen during the motif detection testing in Section 13, where the library peptides largely show a diffuse sequence space when the AS-MS ligand dataset is not added.

7 Encoding of peptides for unsupervised analysis

7.1 One-hot encoding

Each amino acid was represented by the vectors seen below. A peptide was represented by concatenating these vectors together. Thus, each peptide was represented by a **vector 12 * 20 = 240 in length vector descriptor for each peptide**.

Table S3. One-hot encoding vectors for canonical amino acids



Amino acid	One hot encoded vector:																	
A	[1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0]
D	[0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0]
E	[0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0]
F	[0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0]
G	[0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0]
H	[0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0]
K	[0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0]
L	[0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0]
M	[0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0]
N	[0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0]
P	[0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0]
Q	[0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0]
R	[0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0]
S	[0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0]
T	[0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0]
V	[0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0]
W	[0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0]
Y	[0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1]

7.2 Physicochemical encoding

Each amino acid was represented by 12 physicochemical properties as reported from literature.[4] The reported properties were standardized before use. These properties included H11 and H12: hydrophobicity; H2: hydrophilicity; NCI: net charge index of side chains; P11 and P12: polarity; P2: polarizability; SASA: solvent-accessible surface area; V: volume of side chains; F: flexibility; A1: accessibility; E: exposed; T: turns; A2: antigenic. Hydrophobicity (H11 and H12) and polarity (P11 and P12) were calculated using two methods. The peptide was represented by concatenating the vectors of each amino acid together (**12 residues * 12 properties = 144 length vector descriptor for each peptide**)

7.3 ESM-2 encoding

ESM-2 is a protein language model that can be used for multiple applications where properties, structure, and function are derived from the input sequence, where the model was trained on the proteome (UniRef 50). Encoding was completed by extracting the amino acid embeddings of the peptides from 33rd layer of the pretrained "esm2_t33_650M_UR50D" model. From this layer, each embedding per amino acid is size 1280, and a peptide is represented by concatenating this output residue by residue, resulting in a **12 residues * 1280 sized embedding = 15,360 length vector descriptor for each peptide**. While this can seem large, N-grams encoding was also on this order of magnitude.

7.4 Fingerprint encoding

Extended connectivity Fingerprint encoding was used with bit-vectors of 256 length and radius = 3. Canonical and noncanonical amino acids were drawn in ChemDraw 21.0.0 with N-acetylation and N-methyl carboxamidation to replicate the featured of the amino acid integrated within a peptide. Histidine was drawn in its most common τ-tautomer form. Amino acids were exported as SMILES and canonicalized (standardized) in using molvs (standardize_smiles). The Fingerprint was the isolated using Chem.GetMorganFingerprintAsBitVect and Chem.MolFromSmiles. With an n-bit vector of 256, each peptide was represented as **12 residues * 256 bit-vector length = 3,072 length vector descriptor for each peptide**

Monomer	A	D	E	F	G	H	K	L	M	N	P	Q	R	S	T	V	W	Y
Unique Features	2	2	1	3	5	8	1	3	6	2	9	2	7	4	5	2	12	5
Shared Features	20	24	29	30	16	32	31	25	24	26	23	29	31	21	20	21	36	29
Sum Features	22	26	30	33	21	40	32	28	30	28	32	31	38	25	25	23	48	34

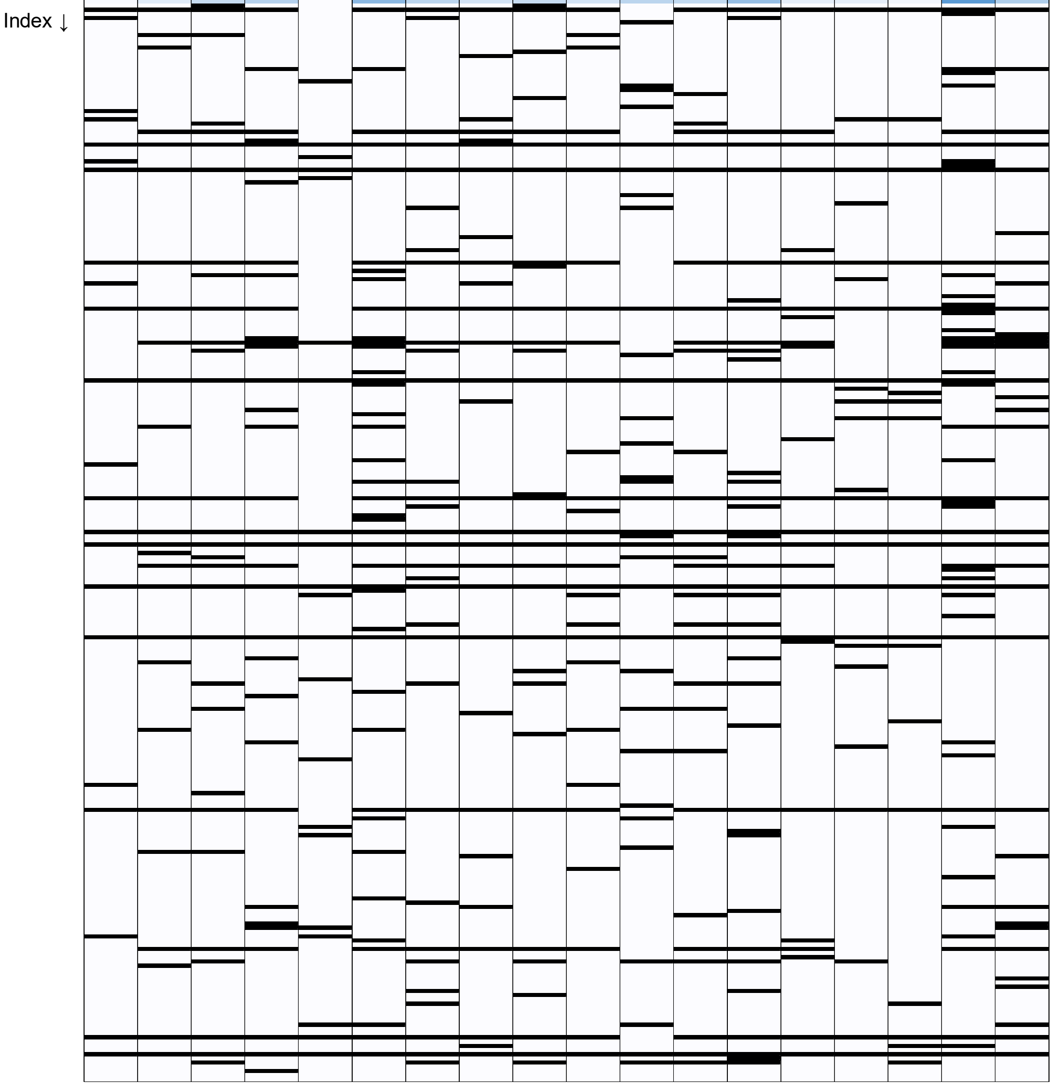


Figure S2. The Fingerprint encoding illustrates the similarities and number of unique features in canonical amino acids. Specifically, one can see the similarity in specific substructure features between amino acids, as well as the number of unique features.

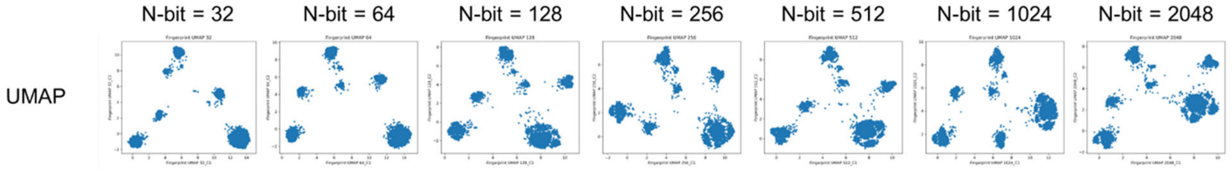


Figure S3. The Fingerprint radius of 3 is generally set for extended connectivity Fingerprint encoding for ECFP_6. However, the bit-vector length can and was varied to see if it affected the data ranging from 2^5 (32 bit vector length) - 2^{11} (2048 bit vector length). Low bit-vector length minimized the appearance of distinct clusters in some analyses of the AS-MS data (e.g., UMAP Fingerprint shown above). The bit-vector length of 256 length was seen to provide more distinct clusters within some of the sequence maps, and above this value, no additional resolution was seen.

7.5 N-grams encoding

N-grams encoding was completed by pre-calculating the observed n-mers in the dataset up to a maximum n-mer length of the full peptide length (12 residues), as described below. As pre-calculated (Figure S3), the entire **peptide was represented at once as a 138,622 length vector**, where each index of the vector describes an n-mer motif that is either present (1) or absent (0) in the peptide.

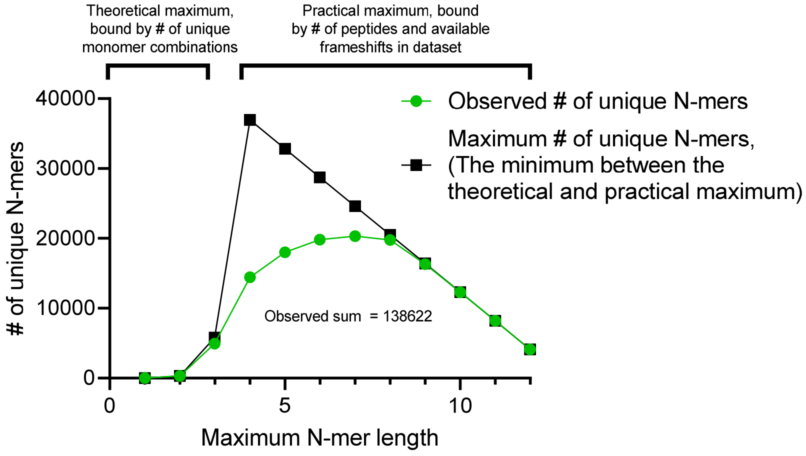


Figure S4. The number of unique N-grams for encoding versus the maximum N-gram length used. N-Grams encoding proceeds first by predetermining all n-mers (sometimes called k-mers) within the dataset. The theoretical number of n-mers is bound by the number of unique combinations of monomers and the maximum N-gram length (i.e., [# of monomers]$^{\text{Maximum N-gram length}}$), which up to a 12-mer length peptide would be 10^{15} n-mers. However, since the n-mer space is pre-calculated from the dataset, significantly fewer are actually observed than theoretically possible even with the maximum N-gram length set to the length of the peptides in the library. The practical maximum is the observed n-mers, bound by (# of peptides) x [1 + (Full Peptide Length – Maximum N-gram length)]. The true maximum is the minimum of the theoretical and practical maximum shown in the figure above in green.

8 UMAP dimensionality reduction hyperparameter optimization

UMAP is a user-friendly, non-linear dimensionality reduction technique that requires minimal optimization to use. However, UMAP embedding results are generally stochastic. Thus the random seed state was always fixed. Some variation in the embeddings was noticed due to the UMAP version, which was 0.5.3 for this work. Lastly, UMAP embeddings are affected by the order of the data within the datafile used (see UMAP shuffle samples leads to quit different result · Issue #268 · lmcinnes/umap) likely because data seen first is weighted more in the initialization of the manifold. Thus, the sequences from AS-MS were randomly shuffled, and then used throughout this work. Additionally, we have observed that exact embedding results can vary from computer to computer, but should remain generally similar.

The two main hyperparameters are n_neighbors and min_dist, and the distance metric setting.

First, n_neighbors balances the importance of the local vs global structure within the data. Low n_neighbors values (~1% of the dataset size) will provide results that focus on local structures, while large values seek to emphasize the global structures, losing fine local detail. This is observed by producing the UMAP embeddings versus n_neighbors (Figure S4).

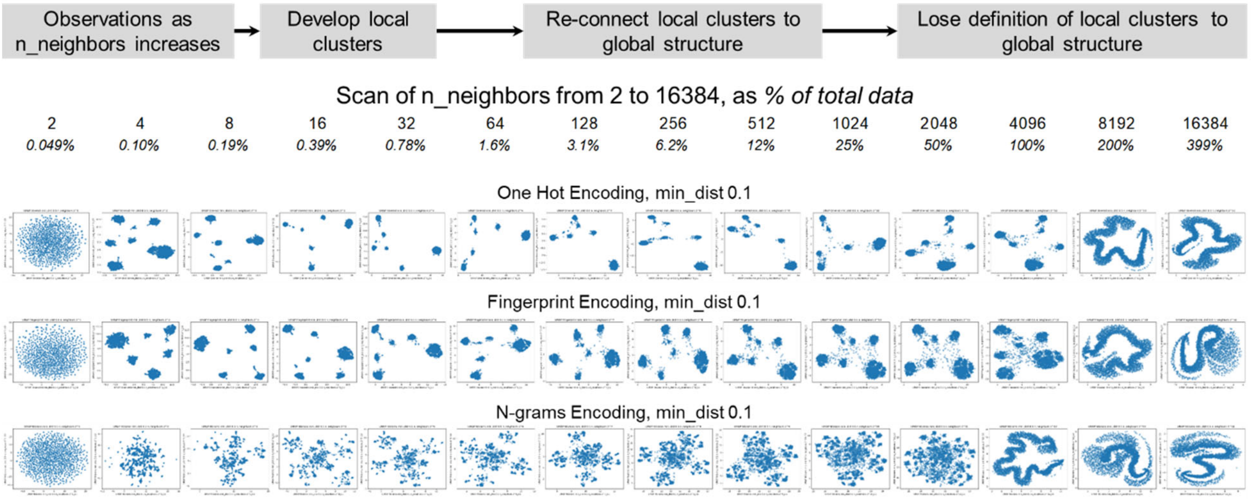


Figure S5. Scan of n_neighbors with UMAP using one-hot, Fingerprint, and N-grams encoding. Local clusters are rapidly and initially developed. As n_neighbors increases, local clusters are reconnected to the global structure of the data at an optimal n_neighbors. As n_neighbors grows to a significant percentage of the dataset (≥ 50%), the clusters begin to be obscured in the global structure unifying the peptides. Stable embeddings results were seen at n_neighbors throughout the dataset from (1.5 – 25%), so 6.2% (n_neighbors = 256) was taken as an optimal value.

Second, min_dist sets the minimum distance between points, meaning that tight local clusters are forced to be spread apart. The default of 0.1 was used for all analysis except for one-hot encoding, which showed exceptionally tight clusters, and so it was set to 0.4.

The distance metric was appropriately set based on the encoding type:[5] binary encoding method (one-hot, Fingerprint, and N-grams) used the Tanimoto distance metric, while continuous descriptors (evolutionarily-learned and physicochemical encoding) used the Euclidean distance metric.

9 Multidimensional scaling (MDS) results

Multidimensional scaling (MDS)[6] was used as the similarity mapping method. However, it is currently unable to incorporate additional results without re-learning. Thus, the dataset of randomly sampled peptides could not be added as it would cause MDS to learn over random sequence space combined with the AS-MS discovered space. Specifically, MDS does not have a .transform function in the current version used (scikit-learn, version 1.0.2), see https://github.com/scikit-learn/scikit-learn/issues/2887, and https://github.com/scikit-learn/scikit-learn/issues/15808 .

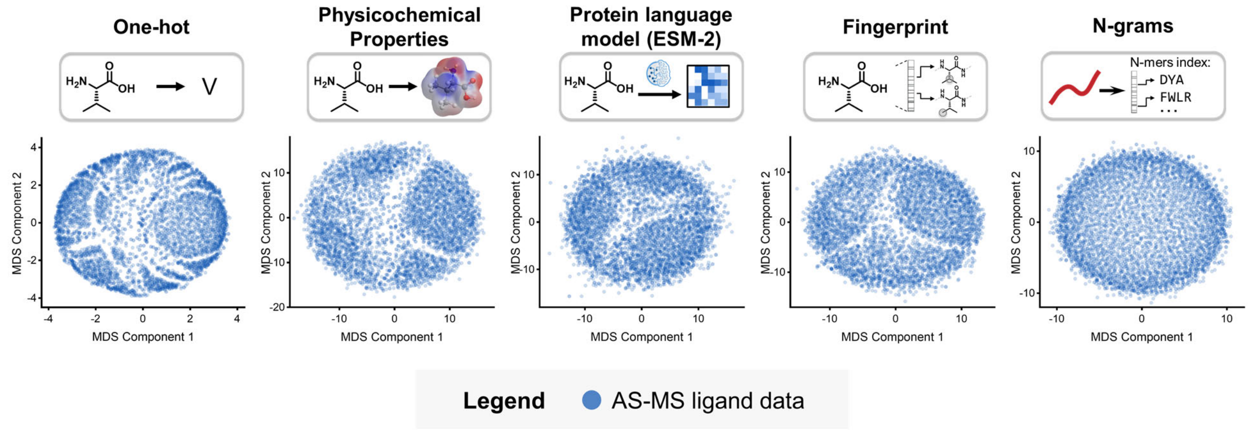


Figure S6. MDS dimensionality reduction versus encoding method of the AS-MS data.

10 Label definitions for 12ca5-specific and nonspecific binders

From the curated AS-MS data,12ca5-specific peptides are defined as *D..DYA* or *D..DYS* from the motif known in literature.[7,8] Note that "*" is a variable length wildcard, while "." is a single amino acid length wildcard.

Care was taken in defining nonspecific binders. From the full dataset, all *D..DYA* or *D..DYS* sequences were removed. Also, all possible mis-sequenced isobaric dipeptides based on of the D**DYA or D**DYS motif were removed. Isobaric was defined as within 10 ppm to match the *de novo* sequencing error tolerance. Sequences containing *DYA*, *DYS* and the commonly observed *PDY*, and *EDY* motifs, gapped isomers (e.g., *D.YA* and *D..YA*), and their dipeptide sequence isomers were removed for consideration as nonspecific binders. Lastly, sequence containing *D.D*, *D..D*, and *D…D* were also removed for consideration as nonspecific binders

All other sequences that were not considered 12ca5-specific or nonspecific were labeled as unknown.

Table S4. Number of peptides manually assigned in each class as defined in *Label definitions for 12ca5-specific and nonspecific binders*

12ca5-specific	**Nonspecific**	**Unknown**	**Total**
3512	139	453	4014

11 All dimensionality reduction results with manually added common motif labels

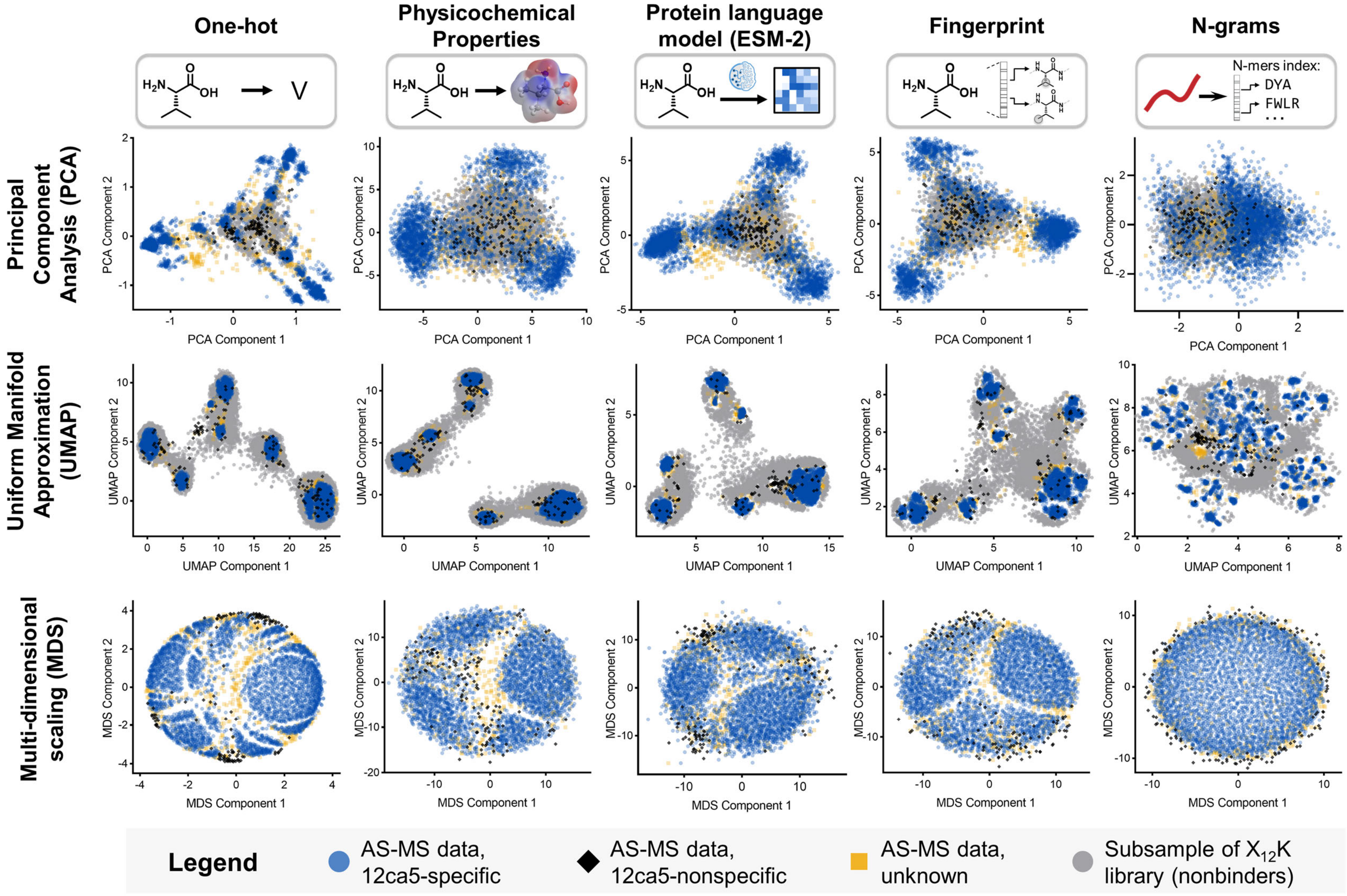


Figure S7. All dimensionality reduction results using all representation encodings with manually added common motif labels as described in SI Section 10: Label definitions for 12ca5-specific and nonspecific binders.

12 Information about all clusters from dimensionality reduction

Every report here on each combination of encoding and dimensionality reduction technique has the following:

1. The sequence map shown in the Main Text, with the manually categorized color-coded labels:
   a. **Common Motif in blue**, defined as D**DYA or D**DYS, where * is a single-character wildcard at any frameshift within a peptide,
   b. **Expanded Motif in orange**, defined as any reported motif that expands, deviates, or adds additional definition to the Common Motif, or
   c. **Weak in gray**, displays a weak signal, no clear motif.
2. The same sequence map with its respective automatous labels.
3. If any expanded motifs are observed in the analysis, a large plot reporting the centroid peptide from each cluster. While a single centroid peptide is reported here, the option is available to report more centroid peptides spread throughout the cluster.
4. A table of all information about each cluster including:
   a. Main text cluster number, if applicable
   b. Autonomously assigned cluster number
   c. The number of peptides in each cluster
   d. One centroid sequence. More centroids can be reported interspersed within each cluster.
   e. Consensus sequence, determined from each cluster with the requirement that the amino acid position shown must be present 33% or more in all of the peptides in the cluster, otherwise X.
   f. Logo of the cluster to infer Consensus sequence and Motif class, prepared using Logomaker.[9]
   g. Motif Class, assigned manually by inspecting the Logo.

Table S5: Report of automated cluster detection algorithm and parameters used from scikit-learn with either Agglomerative Clustering (AggCl) or Density-Based Spatial Clustering of Applications with Noise (DBSCAN).[10] The parameters used and reported here were found by scanning the parameters and inspecting the results.

Dimensionality Reduction Method	Encoding Method	Algorithm	eps	min_samples	# clusters observed
PCA	One-hot	AggCl		31	31
	Physicochemical	AggCl		5	5
	ESM-2	AggCl		6	6
	Fingerprint	AggCl		6	6
	N-grams	AggCl		2	2
UMAP	One-hot	DBSCAN	0.21	10	8
	Physicochemical	DBSCAN	0.21	10	7
	ESM-2	DBSCAN	0.1446	23	16
	Fingerprint	DBSCAN	0.1125	15	19
	N-grams	DBSCAN	0.1022	16	67

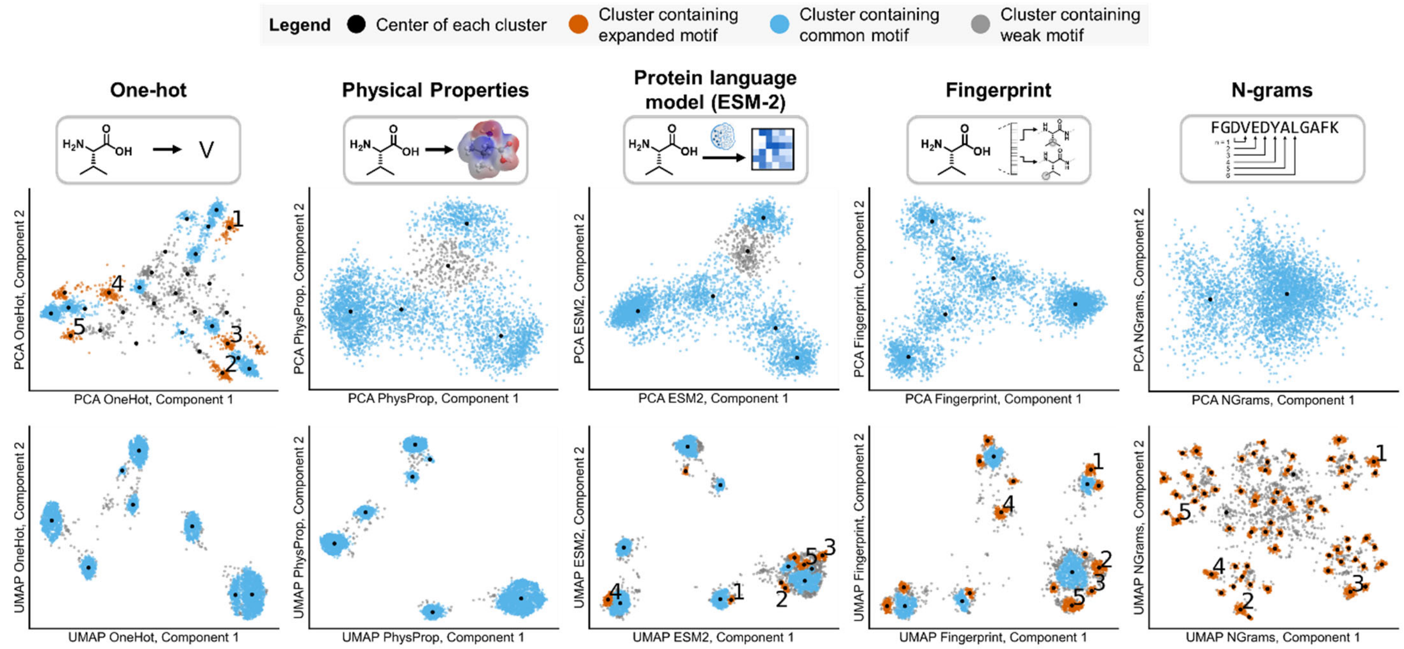


Figure S8. Summary of analyzing the motif of each cluster across all encoding and dimensionality reduction techniques. All sequence maps are shown, with the color-coded labels based on motif class. Motif class was manually categorized as Common Motif in blue, defined as D**DYA or D**DYS, where * is a single-character wildcard at any frameshift within a peptide Expanded Motif in orange, defined as any reported motif that expands, deviates, or adds additional definition to the Common Motif, or Weak in gray, displays a weak signal, no clear motif. Note that no cluster information is available to multi-dimensional scaling as the clusters had little-to-no definition, and could not be detected well with DBSCAN or Agglomerative clustering.

12.1 PCA, One-hot encoding cluster information

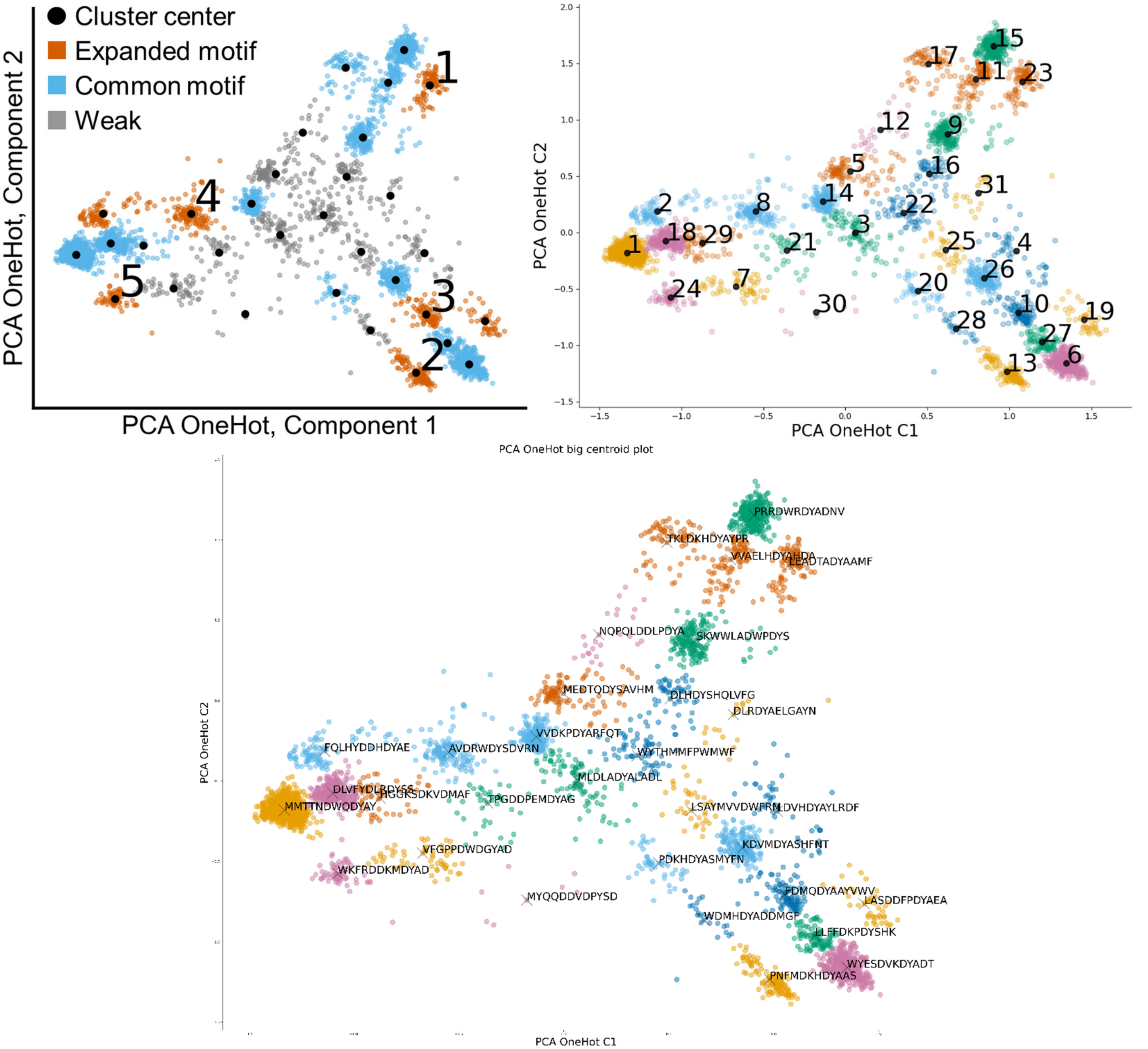
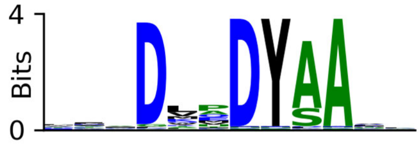
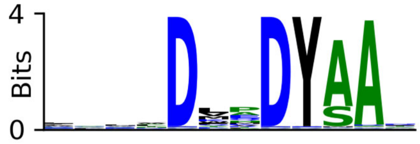


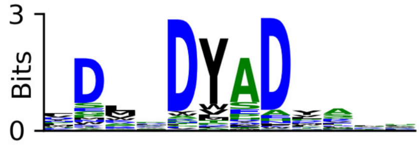
Figure S9. PCA decomposition of all AS-MS data encoded by one-hot encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {23,13,10,8,2,19,24}, respectively. Each cluster is colorblind color coded and labeled with a central point. **Bottom:** A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.

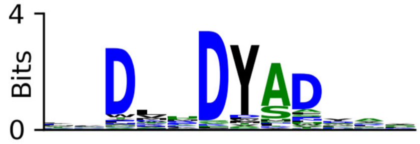
Table S6: Sequence logo report of all clusters detected from PCA dimensionality reduction using one-hot encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

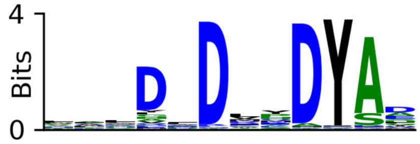
Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class

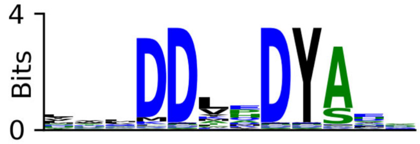
1	23	100	`LEADTADYAAMF,`
			`XXXDXXDYAAX`
					Expanded motif

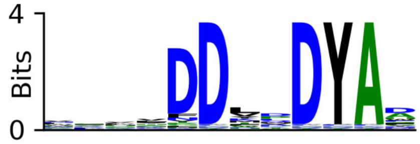
2	13	121	`PNFMDKHDYAAS,`
			`XXXXDXXDYAA`
					Expanded motif

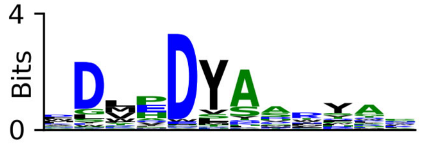
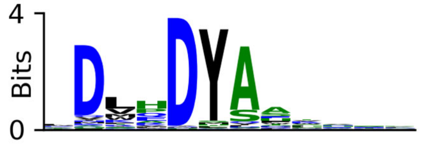
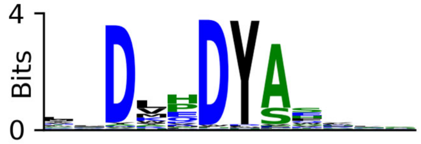
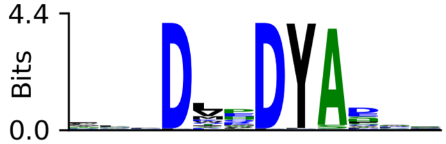
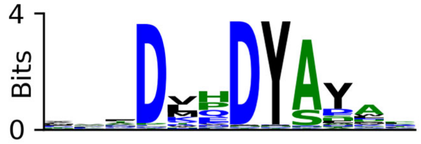
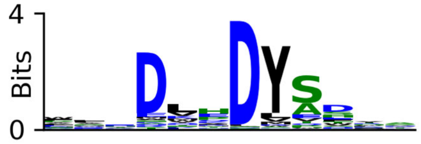
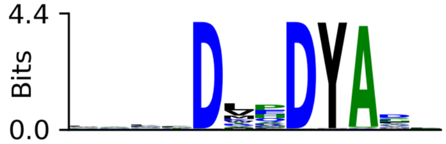
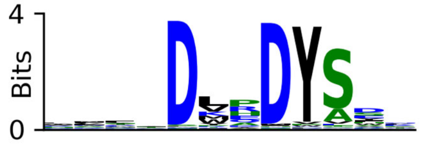
3	10	103	`FDMQDYAAYVWV,`
			`XDXXDYADXXX`
					Expanded motif

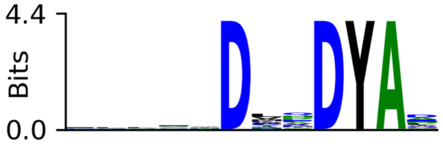
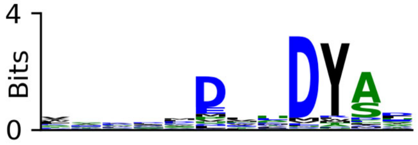
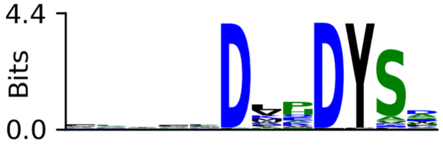
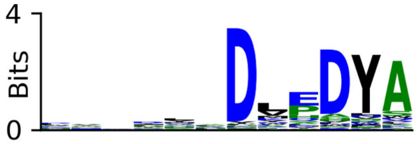
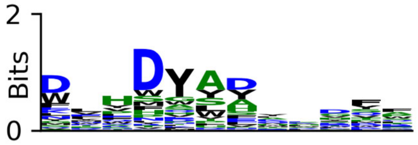
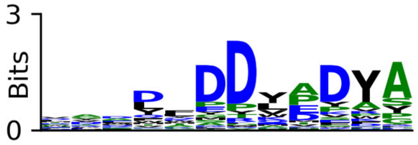
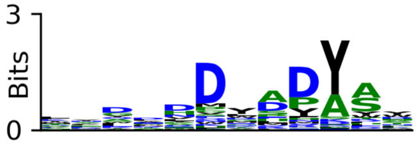
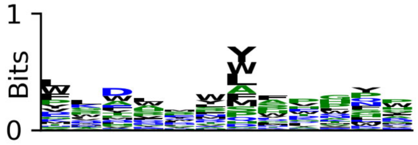
4	8	139	`AVDRWDYSDVRN,`
			`XXDXXDYADXX`
					Expanded motif

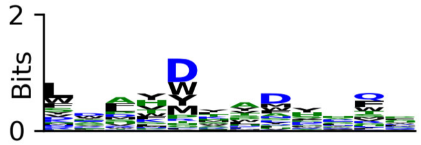
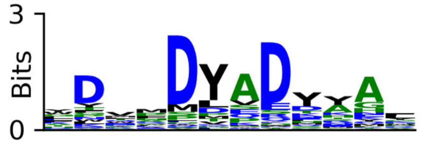
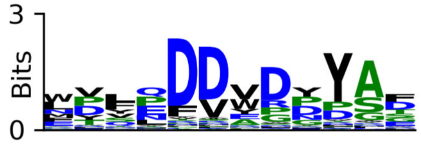
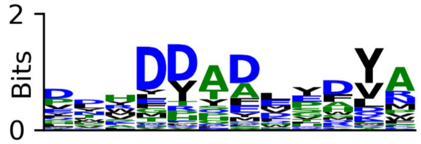
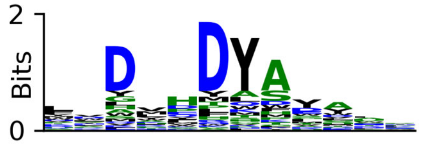
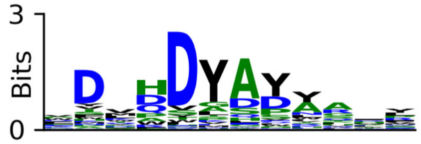
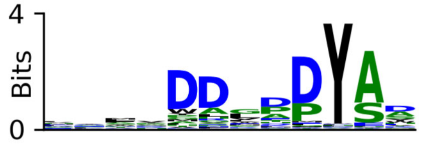
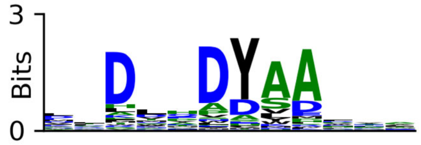
5	2	89	`FQLHYDDHDYAE,`
			`XXXDXDXXDYA`
					Expanded motif

	19	44	`LASDDFPDYAEA,`
			`XXXDDXXDYAX`
					Expanded motif

	24	65	`WKFRDDKMDYAD,`
			`XXXXDDXXDYA`
					Expanded motif

	20	47	PDKHDYASMYFN, XDXXDYAXXXX		Common motif
	26	183	KDVMDYASHFNT, XDXXDYAXXXX		Common motif
	14	194	VVDKPDYARFQT, XXDXXDYAXXX		Common motif
	15	303	PRRDWRDYADNV, XXXDXXDYAXX		Common motif
	17	72	TKLDKHDYAYPR, XXXDXXDYAYX		Common motif
	11	94	VVAELHDYAHDA, XXXDXXDYSXX		Common motif
	6	429	WYESDVKDYADT, XXXXDXXDYAX		Common motif
	27	96	LLFFDKPDYSHK, XXXXDXXDYSX		Common motif

	1	921	MMTTNDWQDYAY, XXXXXDXXDYA		Common motif
	29	55	HGGKSDKVDMAF, XXXXXDXXDYA		Common motif
	18	302	DLVFYDLRDYSS, XXXXXDXXDYS		Common motif
	9	187	SKWWLADWPDYS, XXXXXXDXXDY		Common motif
	16	56	DLHDYSHQLVFG, XXXDXXXXXXX		Weak
	12	25	NQPQLDDLPDYA, XXXXXDDXXDY		Weak
	21	38	TPGDDPEMDYAG, XXXXXDXXDYX		Weak
	22	62	WYTHMMFPWMWF, XXXXXXXXXXX		Weak

	25	37	LSAYMVVDWFRM, XXXXXXXXXXX		Weak
	28	24	WDMHDYADDMGF, XDXXDYADXXA		Weak
	30	8	MYQQDDVDPYSD, XXXXDDXDXYA		Weak
	31	18	DLRDYAELGAYN, XXXDXXXXXXX		Weak
	3	91	MLDLADYALADL, XXDXXDYXXXX		Weak
	4	43	LDVHDYAYLRDF, XDXXDYAXXXX		Weak
	7	59	VFGPPDWDGYAD, XXXXDDXXDYA		Weak
	5	99	MEDTQDYSAVHM, XXDXXDYAAXX		Weak

12.2 PCA, Physicochemical encoding cluster information

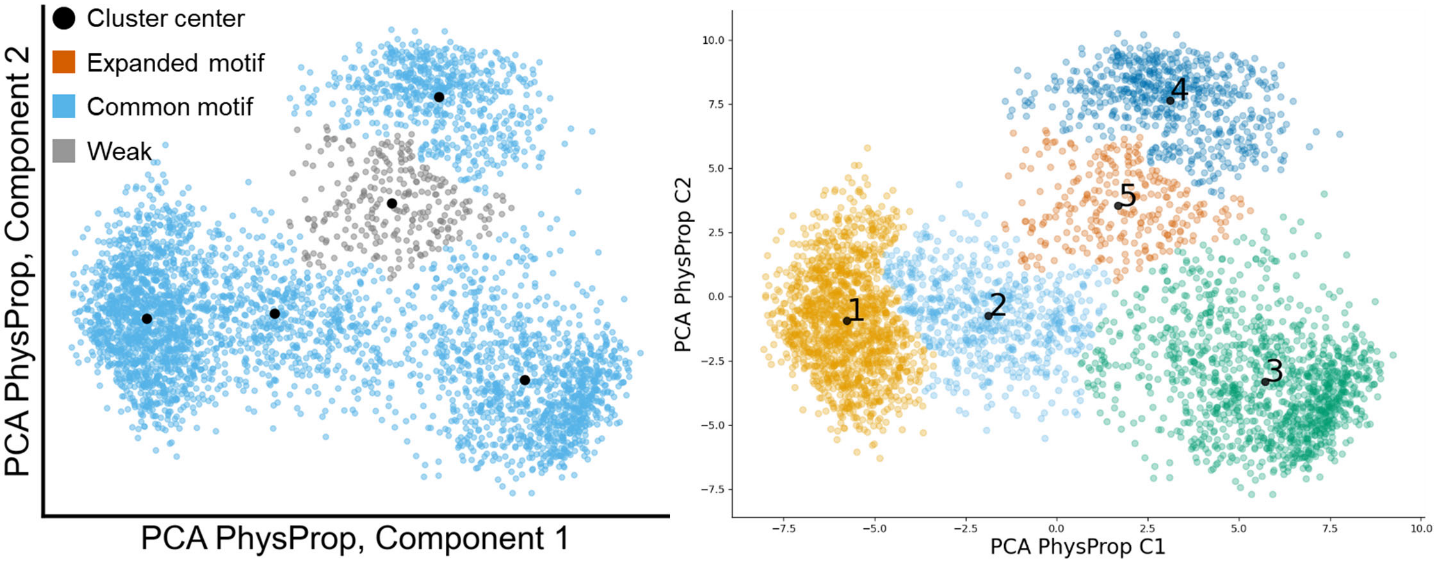
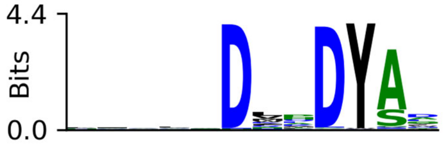
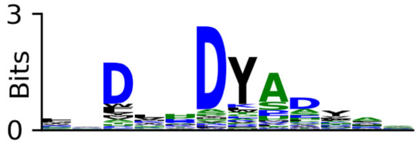
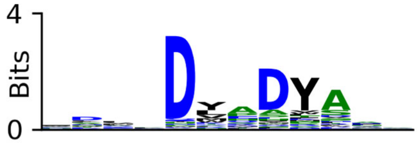
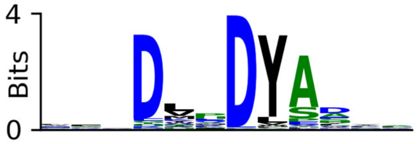
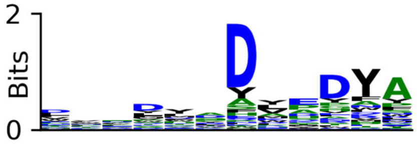


Figure S10. PCA decomposition of all AS-MS data encoded by Physicochemical encoding with automated cluster as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

Table S7: Sequence logo report of all clusters detected from PCA dimensionality reduction using Physicochemical encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
1	1403	LLQTQDYPDYSQ, XXXXXDXXDYA		Common motif
2	609	VFDLEDYAGRAP, XXDXXDYAXXX		Common motif
3	1197	YFNEDAPDYASP, XXXXDXXDYAX		Common motif

4	625	MPLDVGDYAAQN, XXXDXXDYAXX		Common motif
5	270	SPAVHHDVEDYA, XXXXXXDXXXX		Weak

12.3 PCA, ESM-2 encoding cluster information

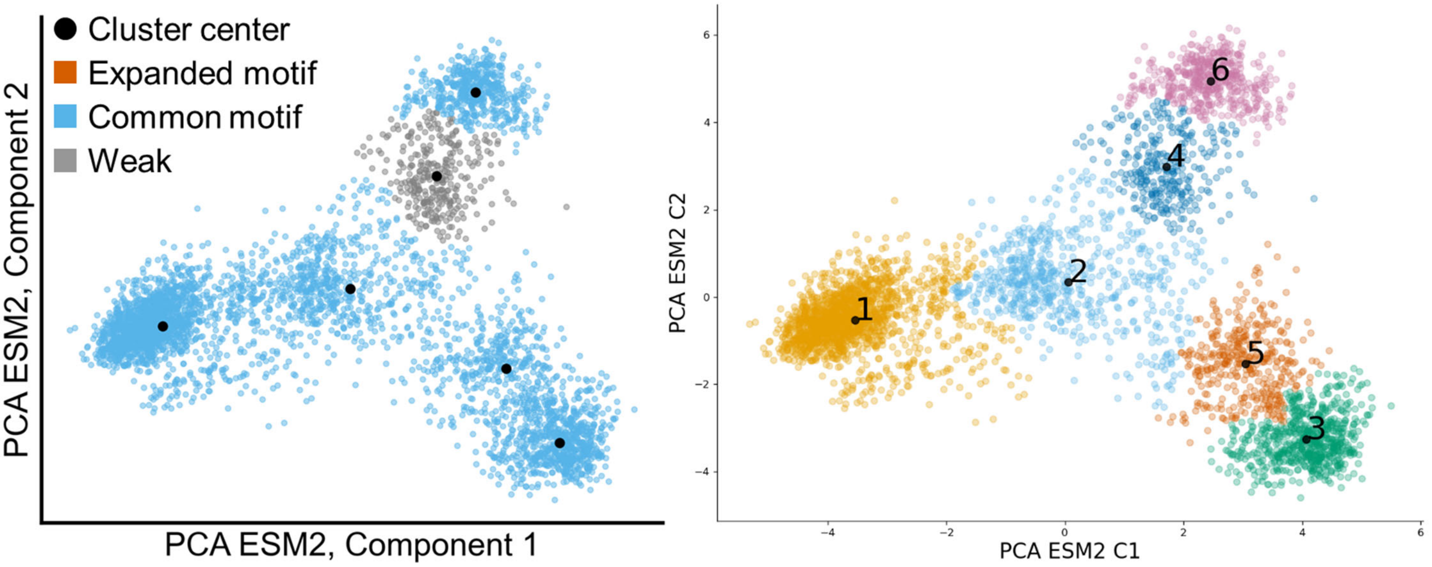
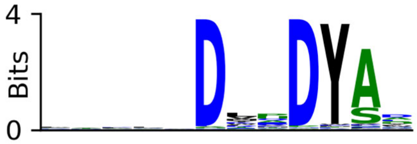
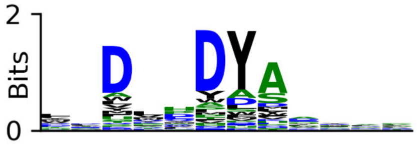


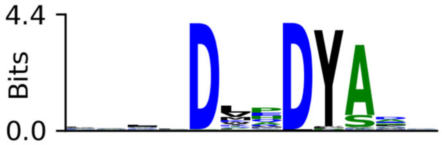
Figure S11. PCA decomposition of all AS-MS data encoded by ESM2 encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

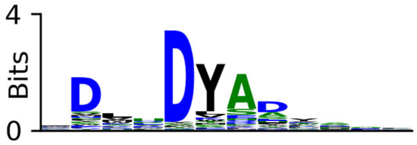
Table S8: Sequence logo report of all clusters detected from PCA dimensionality reduction using ESM2 encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

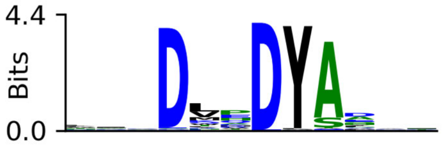
Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**

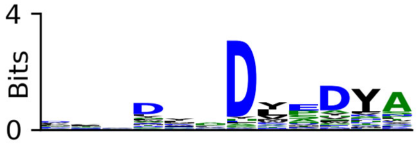
1	1599	`WFRAFDMEDYSD,`
		`XXXXXDXXDYA`
				Common motif

2	648	`LDDPADYAVGTK,`
		`XXDXXDYXXXX`
				Common motif

3	663	`HHTYDLPDYSFY,`
		`XXXXDXXDYAX`
				Common motif

5	389	`LDVQDYANVSES,`
		`XDXXDYAXXXX`
				Common motif

6	495	`YLMDLFDYAHKT,`
		`XXXDXXDYAXX`
				Common motif

4	310	`WDVFFPDYSHRP,`
		`XXXXXXDXXDY`
				Weak

12.4 PCA, Fingerprint encoding cluster information

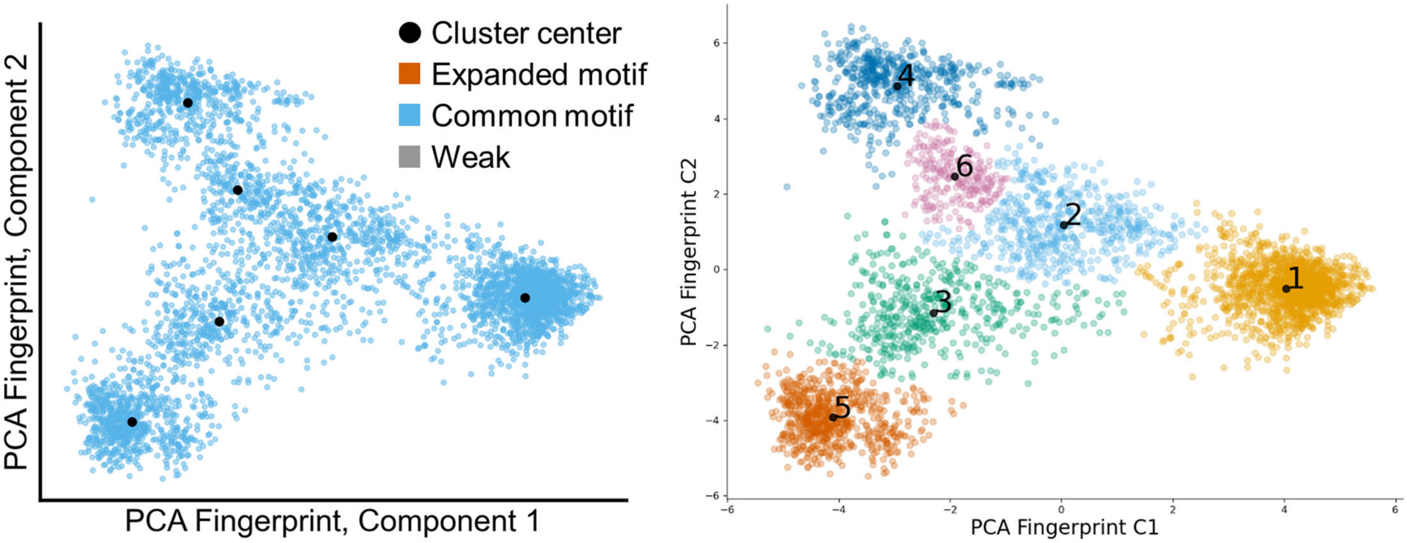
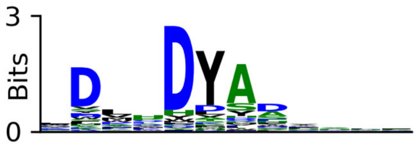
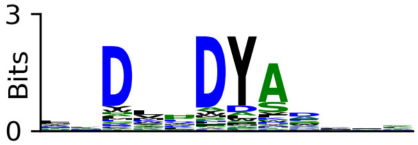


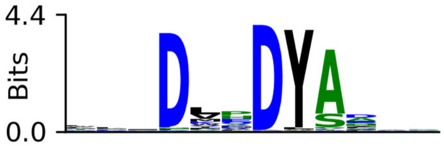
Figure S12. PCA decomposition of all AS-MS data encoded by Fingerprint encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

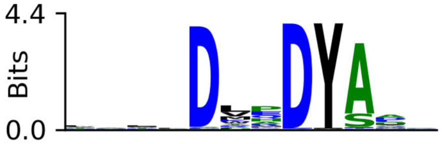
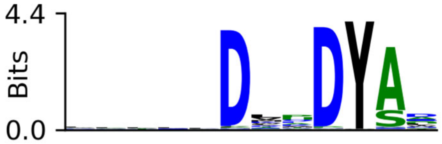
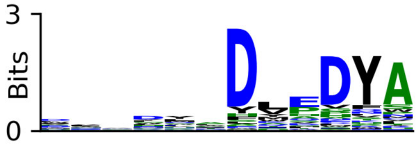
Table S9: Sequence logo report of all clusters detected from PCA dimensionality reduction using Fingerprint encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class

3	484	`FDRLDYSDQFFK,`
		`XDXXDYAXXXX`
				Common motif

2	572	`HADVQDYAFHYT,`
		`XXDXXDYAXXX`
				Common motif

4	575	`LDGDLWDYADTY,`
		`XXXDXXDYAXX`
				Common motif

5	703	FFLMDLWDYARS, XXXXDXXDYAX		Common motif
1	1521	LLKWVDKHDYAY, XXXXXDXXDYA		Common motif
6	249	KDHDYAYFMETR, XXXXXXDXXDY		Common motif

12.5 PCA, N-grams encoding cluster information

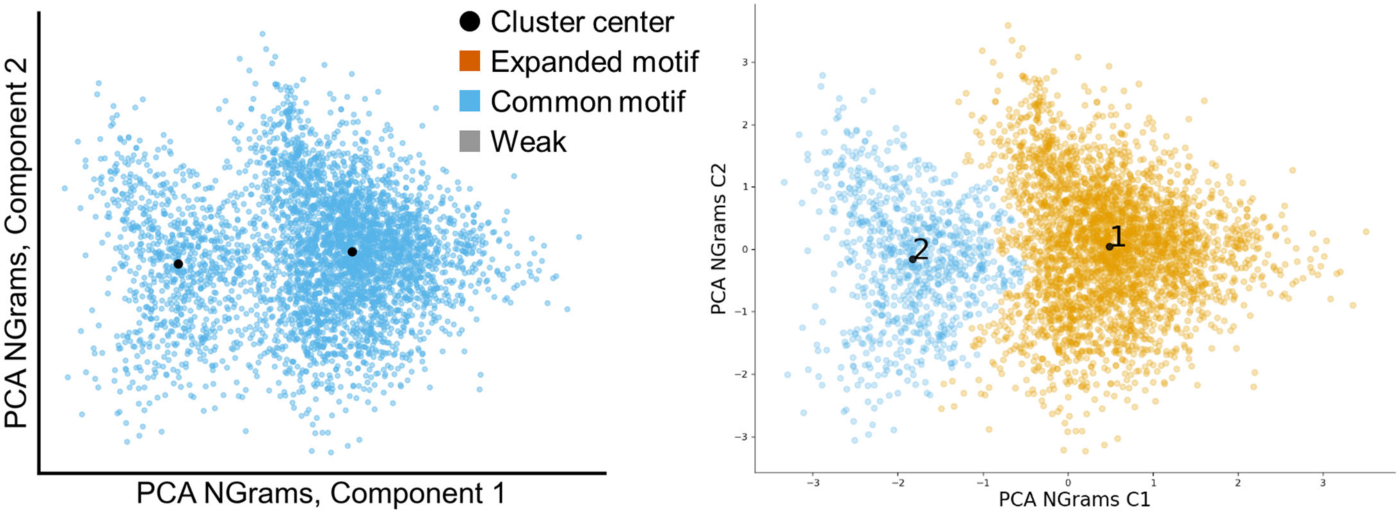
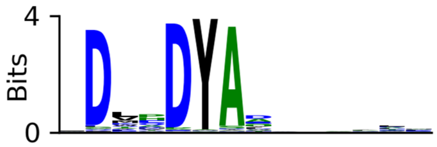
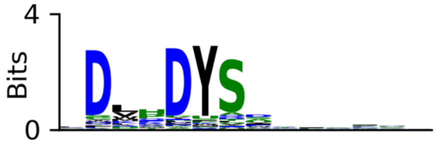


Figure S13. PCA decomposition of all AS-MS data encoded by N-grams encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

Table S10: Sequence logo report of all clusters detected from PCA dimensionality reduction using N-grams encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

NOTE: Because N-grams encodes peptides by the presence of their motifs, irrespective of frameshift, the logo plot displays the sequences aligned by ClustalW to the second position to show the motif.

Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	***ALIGNED*** Sequence Logo	Motif Class
1	3242	PSDLRDYAAGFF, XDXXDYAX-----		Common motif
2	862	QVDTRDYSDLYF, XDXXDYSX-----		Common motif

12.6 UMAP, One-hot encoding cluster information

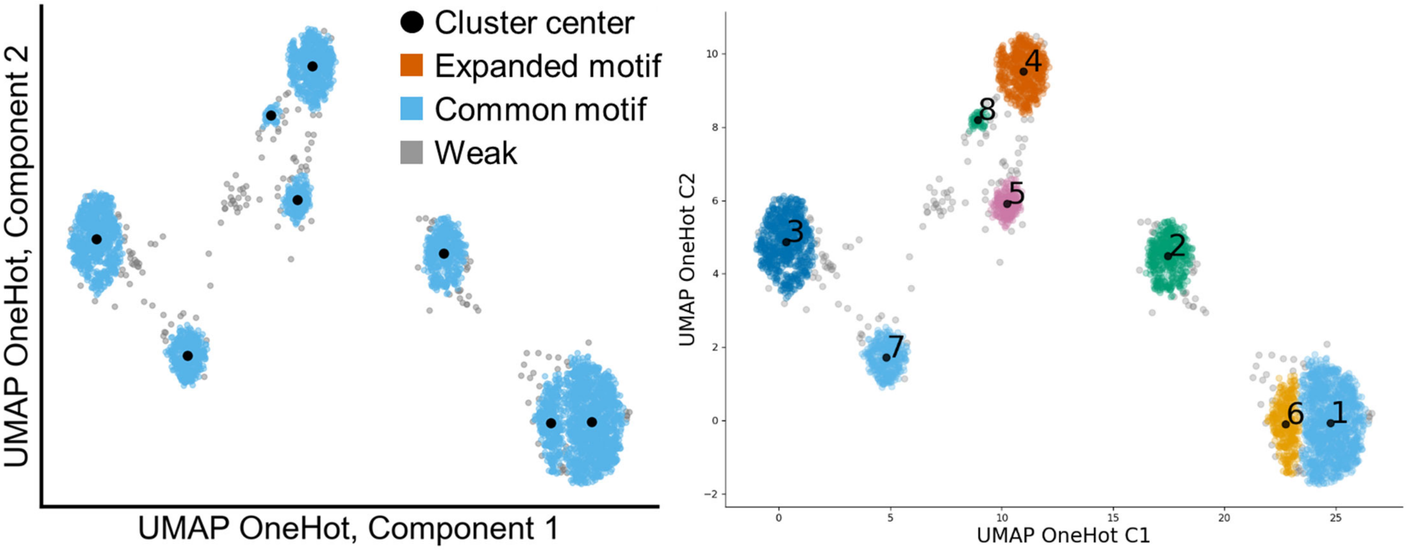
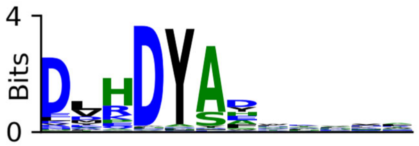
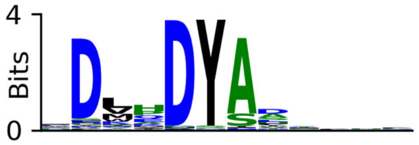
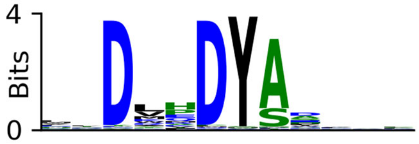
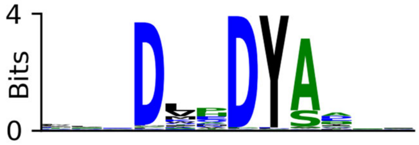
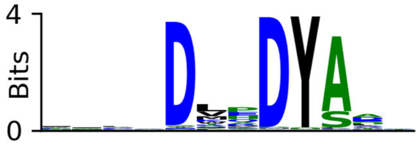
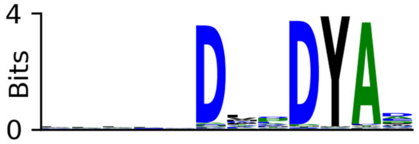
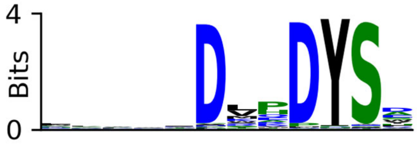
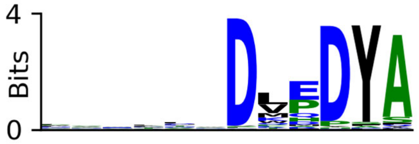
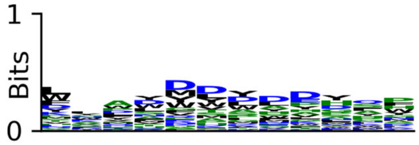


Figure S14. UMAP decomposition of all AS-MS data encoded by one-hot encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

Table S11: Sequence logo report of all clusters detected from UMAP dimensionality reduction using one-hot encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class
8	59	DVRDYAENDFLV, DXHDYAXXXXX		Common motif
7	354	LDMQDYAAGDWM, XDXXDYAXXXXX		Common motif

2	454	EGDAEDYAAFRG, XXDXXDYAXXX		Common motif
4	573	FNLDEQDYADTP, XXXDXXDYAXX		Common motif
3	739	FPVVDWEDYATW, XXXXDXXDYAX		Common motif
1	1230	SNEFSDMLDYAE, XXXXXDXXDYA		Common motif
6	323	FDLFLDVPDYSS, XXXXXDXXDYS		Common motif
5	209	LPGGFLDWEDYA, XXXXXXDXXDY		Common motif
0	163			Weak

12.7 UMAP, Physicochemical encoding cluster information

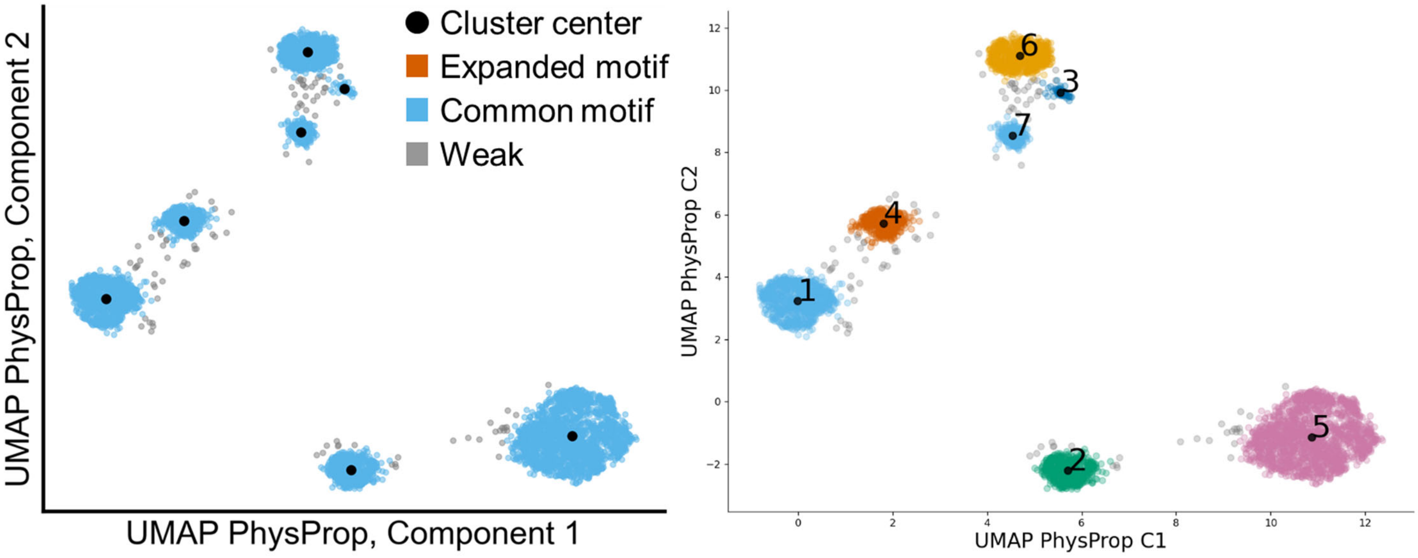
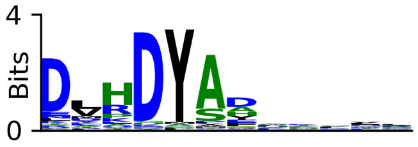
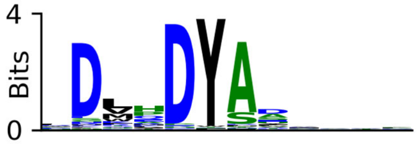


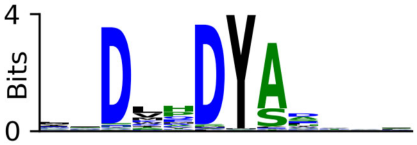
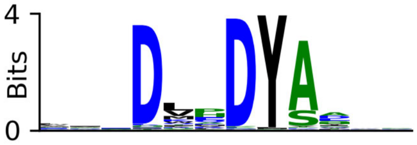
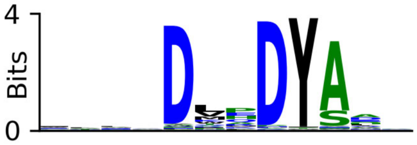
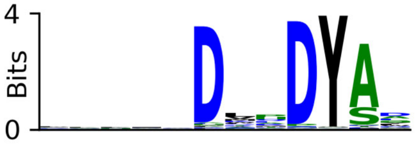
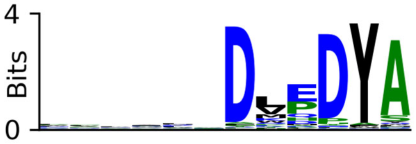
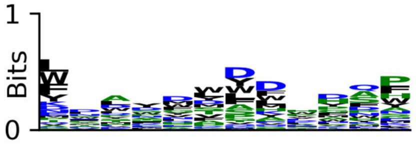
Figure S15. UMAP decomposition of all AS-MS data encoded by Physicochemical encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. ***No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif.*** Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.

Table S12: Sequence logo report of all clusters detected from UMAP dimensionality reduction using Physicochemical encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**

3	64	`DLKDYADNHWEA,`
		`DXXDYAXXXXX`
				Common motif

4	358	`ADMEDYAQNYPL,`
		`XDXXDYAXXXX`
				Common motif

2	465	FFDLPDYSVPKL, XXDXXDYAXXX		Common motif
6	578	PYLDMEDYAQLF, XXXDXXDYAXX		Common motif
1	756	LYWDDVEDYAEH, XXXXDXXDYAX		Common motif
5	1572	LDFGGDWPDYAH, XXXXXDXXDYA		Common motif
7	214	TPQMEADVDPYA, XXXXXXDXXDY		Common motif
0	97			Weak

12.8 UMAP, ESM-2 encoding cluster information

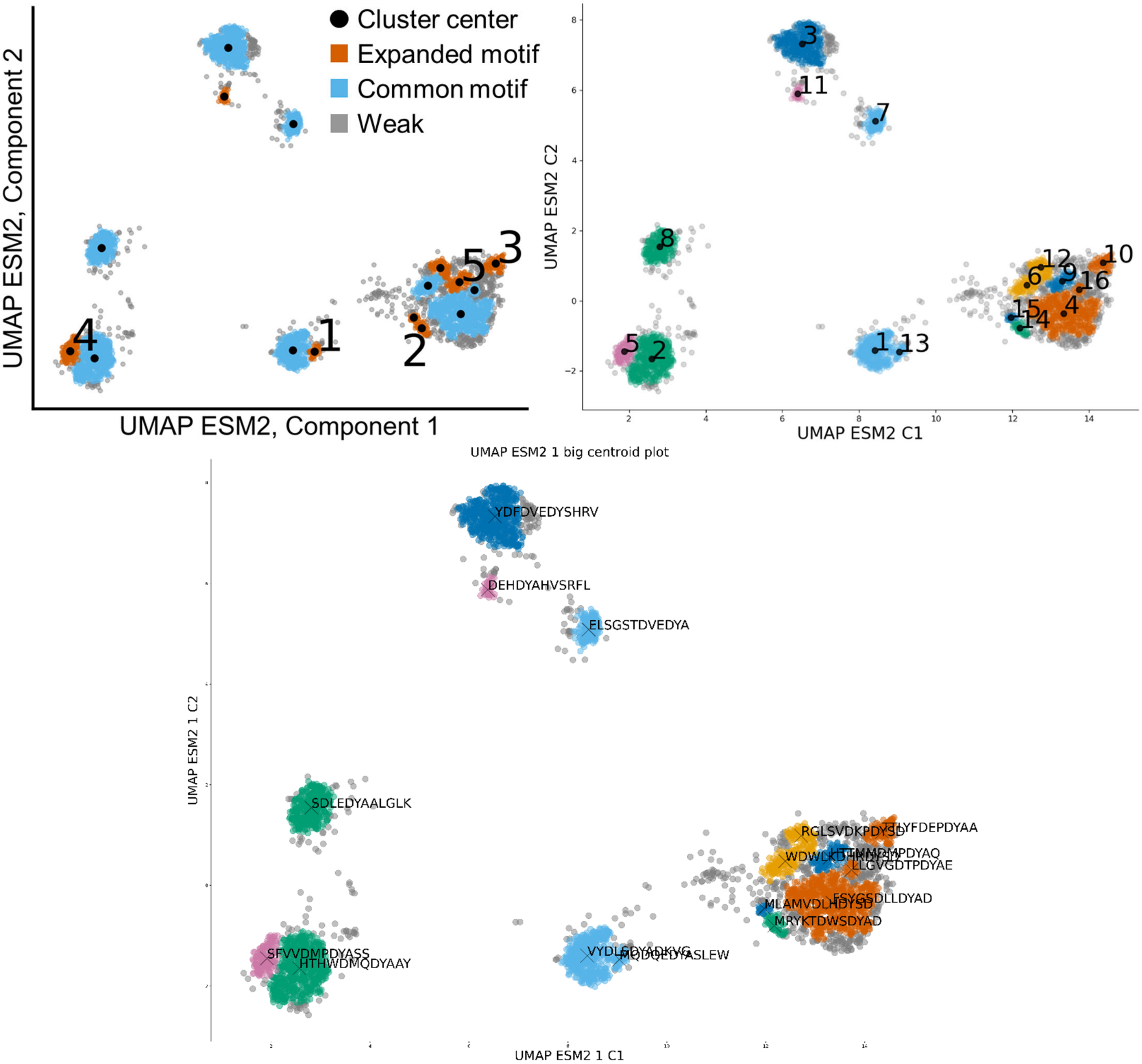
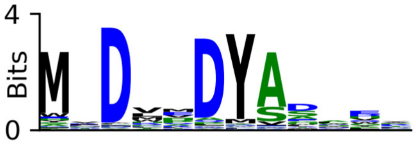
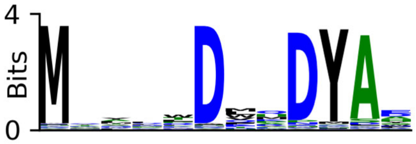


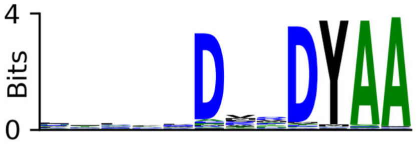
Figure S16. UMAP decomposition of all AS-MS data encoded by ESM-2 encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {13,14,15,10,5,9,12}, respectively. Each cluster is colorblind color coded and labeled with a central point. **Bottom:** A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.

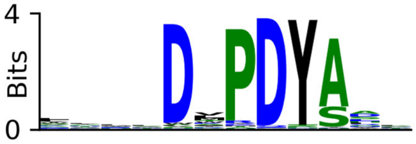
Table S13: Sequence logo report of all clusters detected from UMAP dimensionality reduction using ESM-2 encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

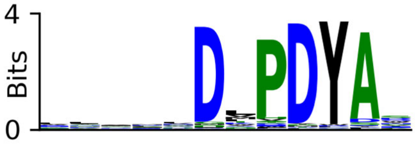
Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class

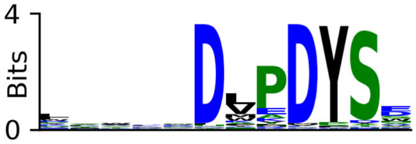
1	13	45	`MQDQEDYASLEW,`
			`MXDXXDYAXXX`
					Expanded motif

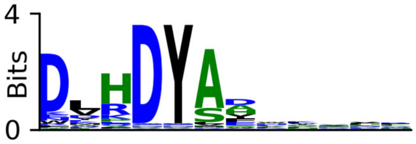
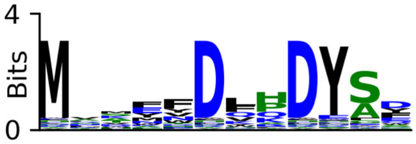
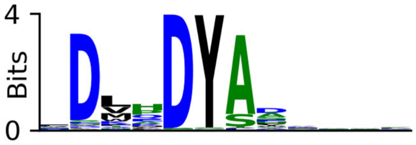
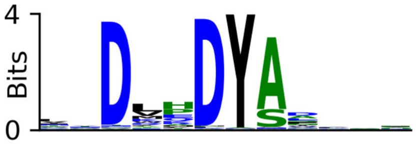
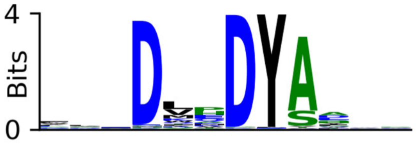
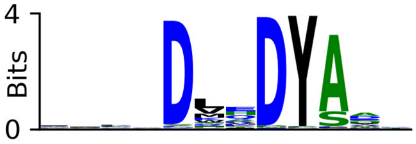
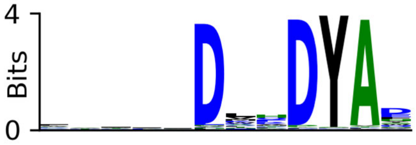
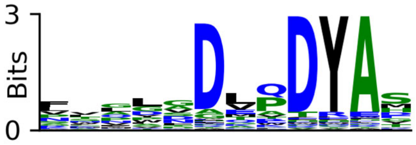
2	14	51	`MRYKTDWSDYAD,`
			`MXXXXDXXDYA`
					Expanded motif

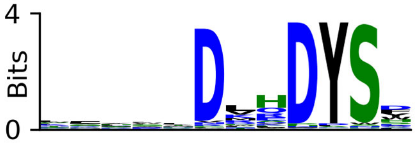
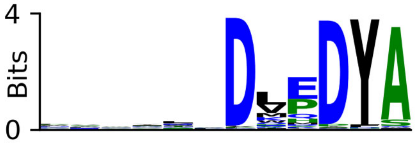
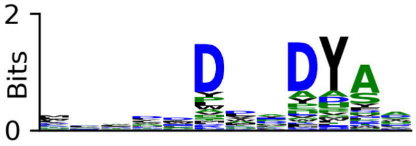
3	10	115	`TTLYFDEPDYAA,`
			`XXXXXDXXDYA`
					Expanded motif

4	5	149	`SFVVDMPDYASS,`
			`XXXXDXPDYAX`
					Expanded motif

5	9	109	`HTTMMDMPDYAQ,`
			`XXXXXDXPDYA`
					Expanded motif

	12	87	`RGLSVDKPDYSD,`
			`XXXXXDXPDYS`
					Expanded motif

	11	54	DEHDYAHVSRFL, DXHDYAXXXXX		Expanded motif
	15	22	MLAMVDLHDYSD, MXXXXDXXDYS		Expanded motif
	8	330	SDLEDYAALGLK, XDXXDYAXXXX		Common motif
	1	397	VYDLSDYADKVG, XXDXXDYAXXX		Common motif
	3	518	YDFDVEDYSHRV, XXXDXXDYAXX		Common motif
	2	564	HTHWDMQDYAAY, XXXXDXXDYAX		Common motif
	4	645	FSYGSDLLDYAD, XXXXXDXXDYA		Common motif
	16	24	LLGVGDTPDYAE, XXXXXDXXDYA		Common motif

	6	132	WDWLKDHRDYSD, XXXXXDXXDYS		Common motif
	7	191	ELSGSTDVEDYA, XXXXXXDXXDY		Common motif
	0	671			Weak

12.9 UMAP, Fingerprint encoding cluster information

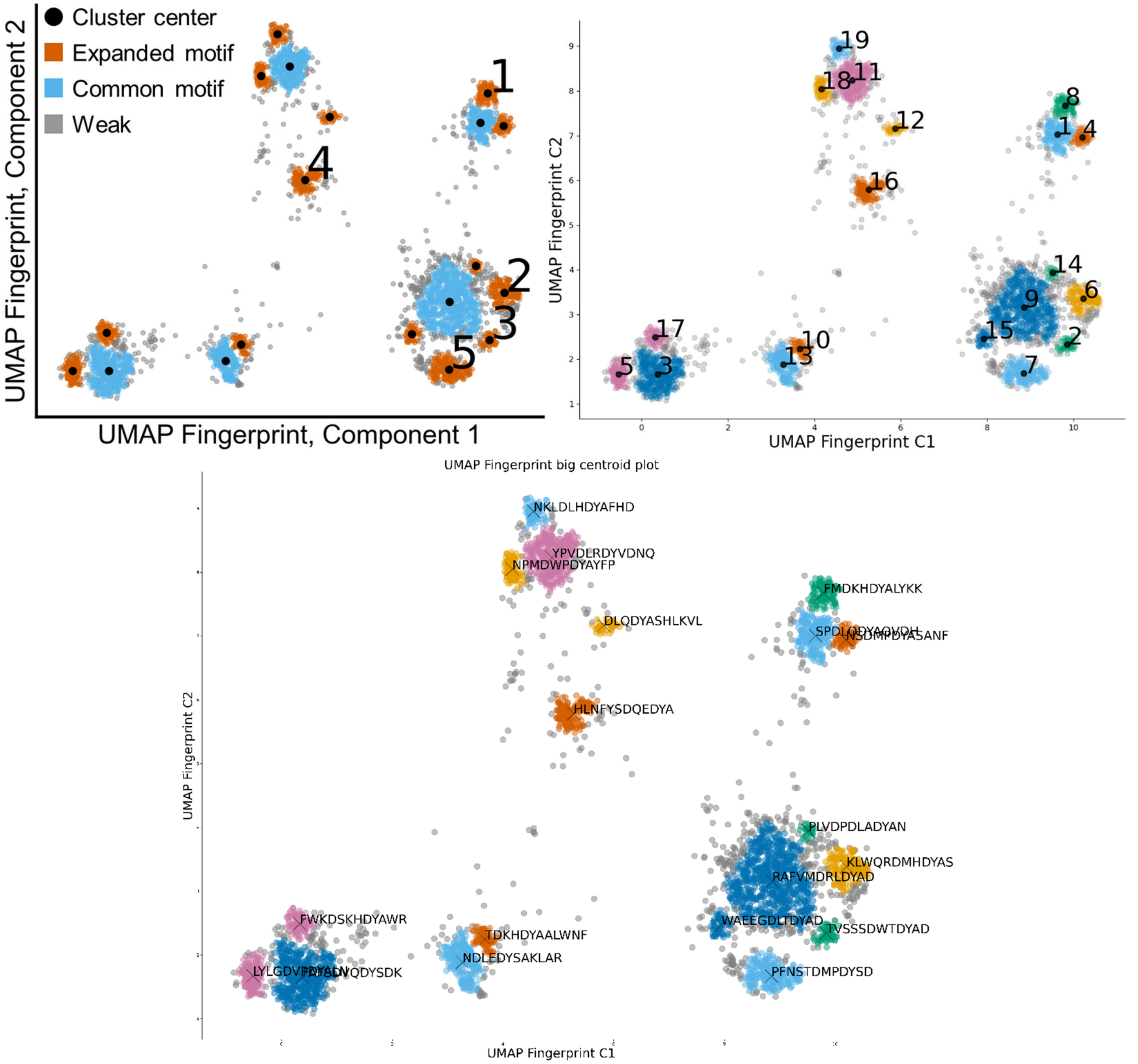
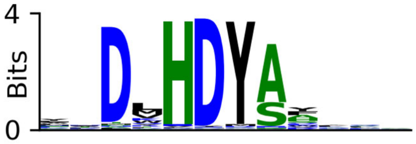
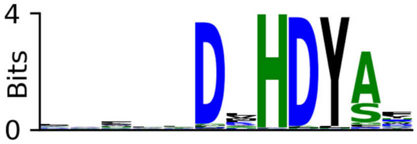
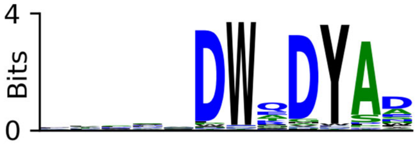
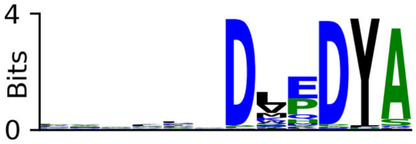
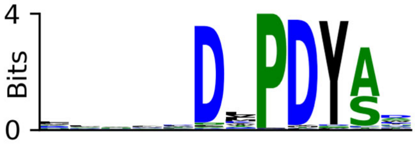
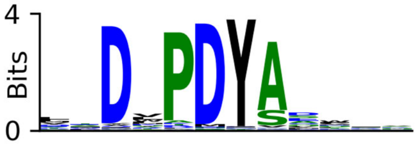
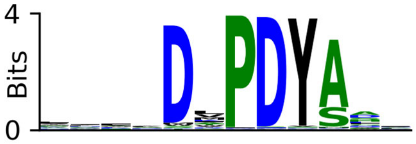
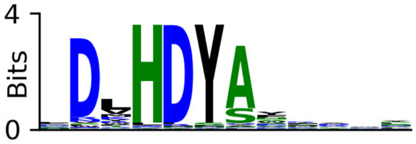
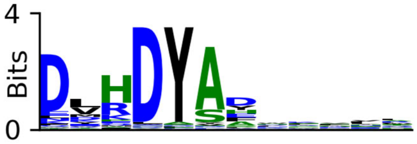
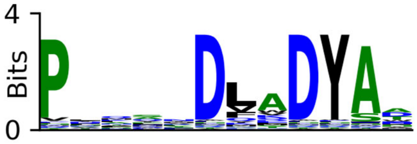
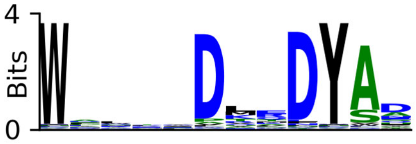
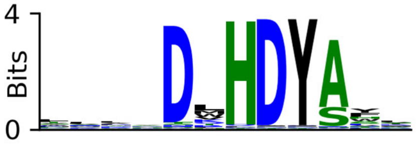
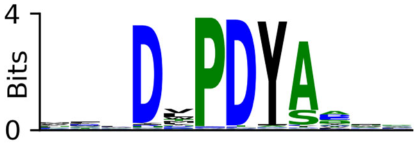
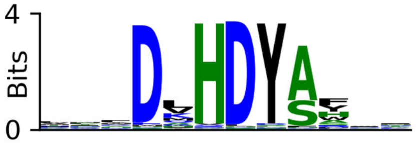
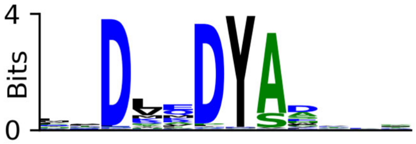
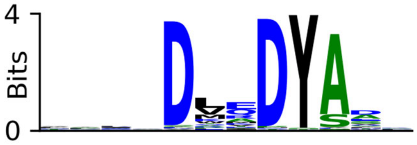
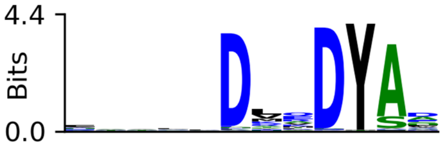
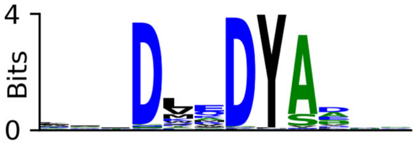
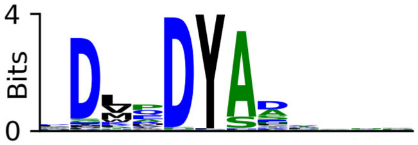
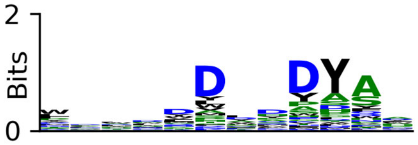


Figure S17. UMAP decomposition of all AS-MS data encoded by Fingerprint encoding with automated cluster detection as described in Table S6. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {8,6,15,2,16,5,7}, respectively. Each cluster is colorblind color coded and labeled with a central point. **Bottom:** A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.

Table S14: Sequence logo report of all clusters detected from UMAP dimensionality reduction using Fingerprint encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class
1	8	122	FMDKHDYALYKK, XXDXHDYAXXX		Expanded motif
2	6	172	KLWQRDMHDYAS, XXXXXDXHDYA		Expanded motif
3	2	68	TVSSSDWTDYAD, XXXXXDWXDYA		Expanded motif
4	16	193	HLNFYSDQEDYA, XXXXXXDXXDY		Expanded motif
5	7	223	PFNSTDMPDYSD, XXXXXDXPDYA		Expanded motif
	4	90	NSDMPDYASANF, XXDXPDYAXXX		Expanded motif

	5	149	LYLGDVPDYALN, XXXXDXPDYAX		Expanded motif
	10	81	TDKHDYAALWNF, XDXHDYAXXXX		Expanded motif
	12	56	DLQDYASHLKVL, DXHDYAXXXXX		Expanded motif
	14	29	PLVDPDLADYAN, PXXXXDLADYA		Expanded motif
	15	62	WAEEGDLTDYAD, WXXXXDXXDYA		Expanded motif
	17	90	FWKDSKHDYAWR, XXXXDXHDYAX		Expanded motif
	18	110	NPMDWPDYAYFP, XXXDXPDYAXX		Expanded motif
	19	92	NKLDLHDYAFHD, XXXDXHDYAXX		Expanded motif

	1	225	SPDLQDYAQVDH, XXDXXDYAXXX		Common motif
	3	435	FAFSDVQDYSDK, XXXXDXXDYAX		Common motif
	9	723	RAFVMDRLDYAD, XXXXXDXXDYA		Common motif
	11	336	YPVDLRDYVDNQ, XXXDXXDYAXX		Common motif
	13	243	NDLEDYSAKLAR, XDXXDYAXXXX		Common motif
	0	605			Weak

12.10 UMAP, N-grams encoding cluster information

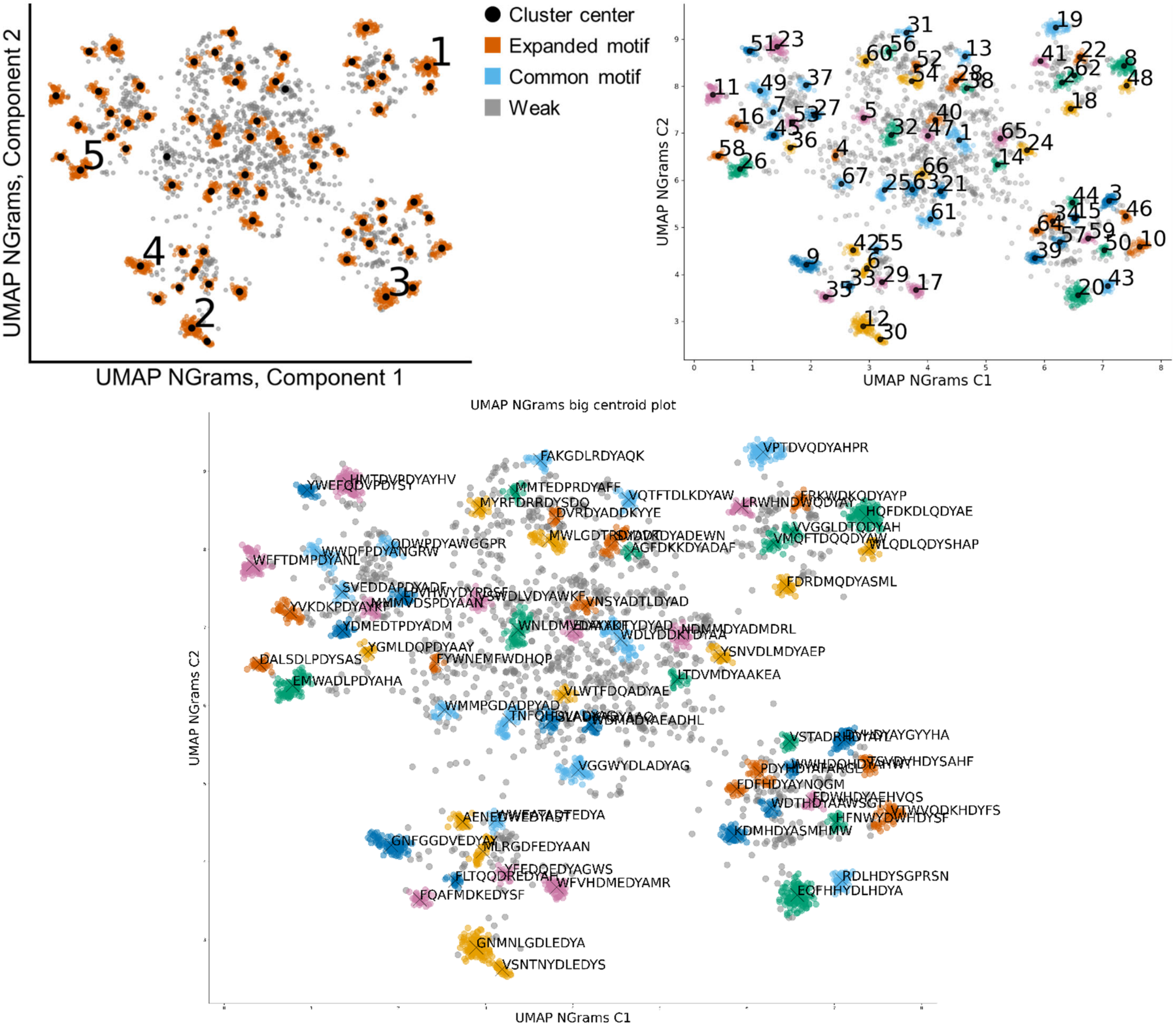
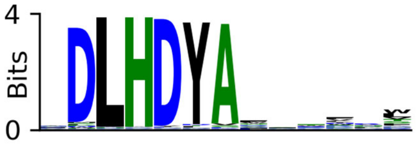
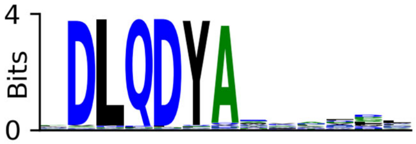
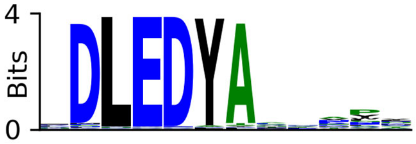
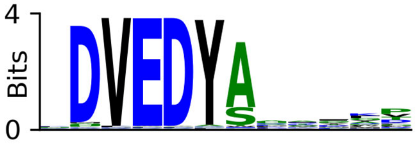
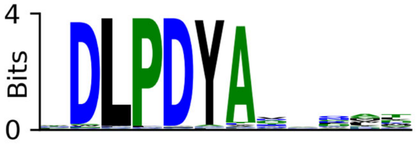
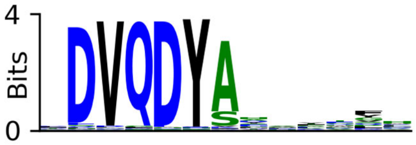


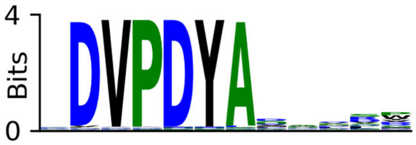
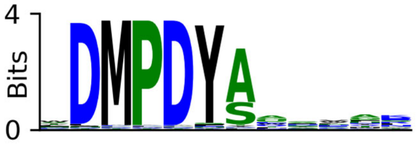
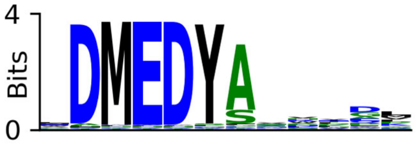
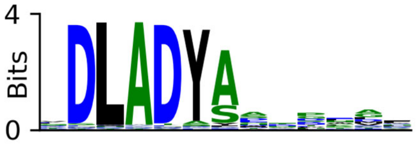
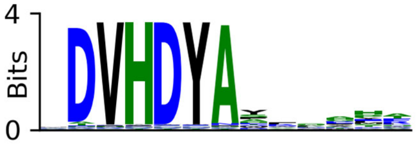
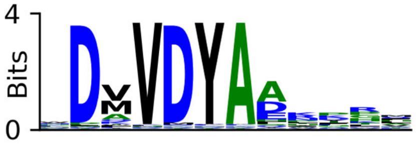
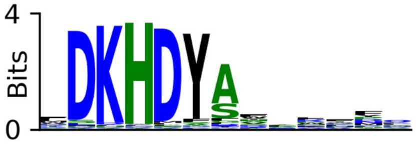
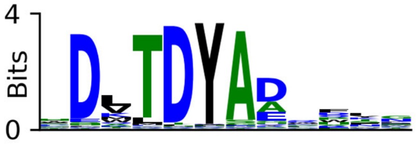
Figure S18. UMAP decomposition of all AS-MS data encoded by N-grams encoding with automated cluster detection as described in Table S5. **Top Left:** Figure as labeled in the main text. **Top Right:** The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {8,12,20,9,26,19,23}, respectively. Each cluster is colorblind color coded and labeled with a central point. **Bottom:** A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.

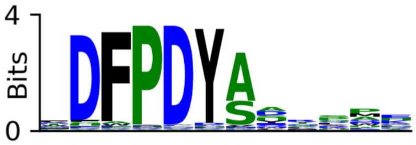
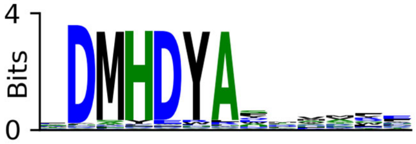
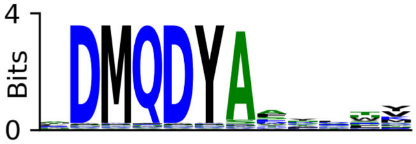
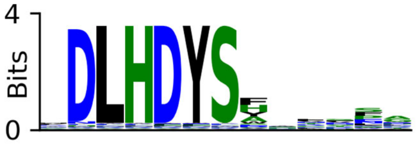
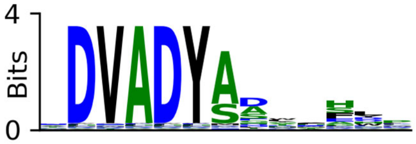
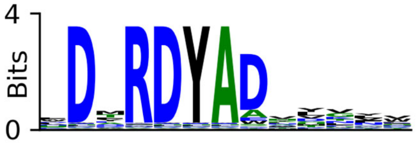
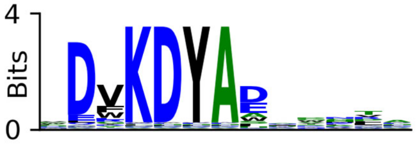
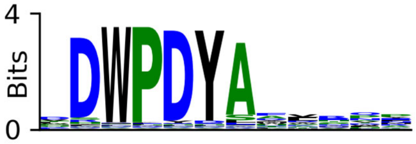
Table S15: Sequence logo report of all clusters detected from UMAP dimensionality reduction using N-grams encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus

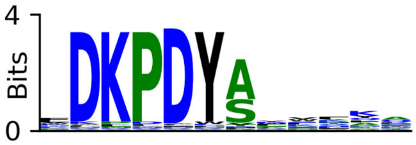
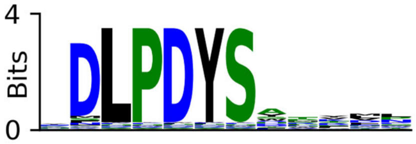
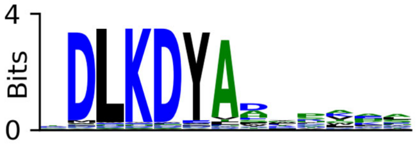
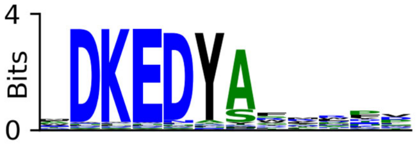
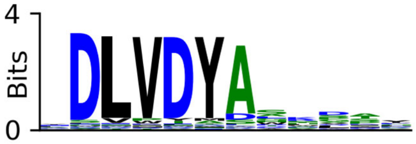
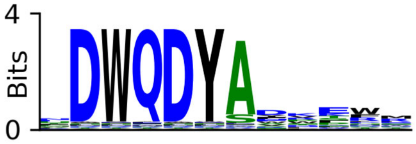
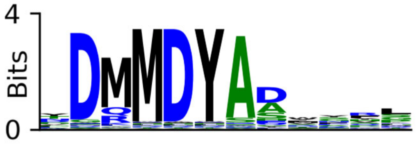
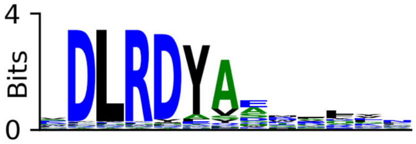
sequence, logos, and motif class. Details are described in *Information about all clusters from dimensionality reduction*.

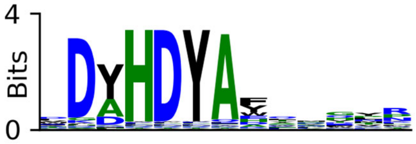
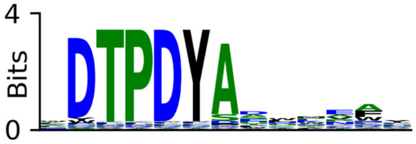
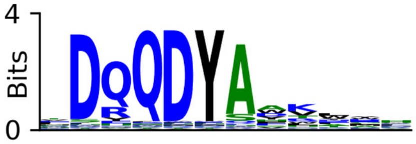
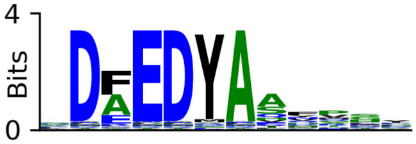
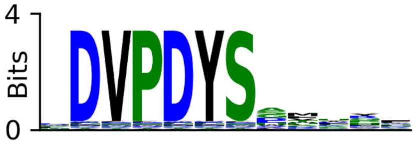
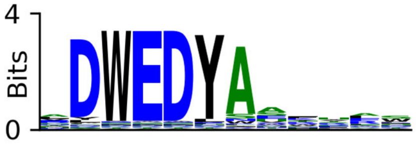
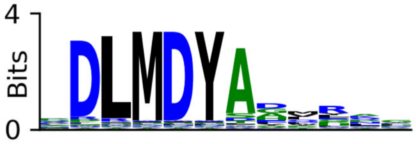
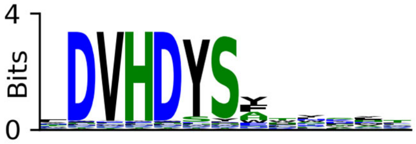
NOTE: Because N-grams encodes peptides by the presence of their motifs, irrespective of frameshift, the logo plot displays the sequences aligned by ClustalW to the second position to show the motif.

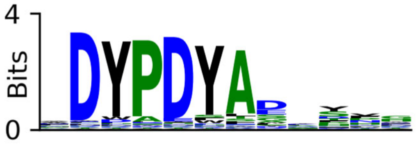
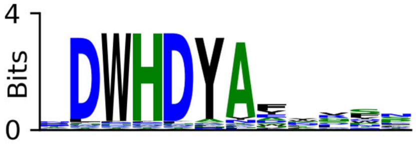
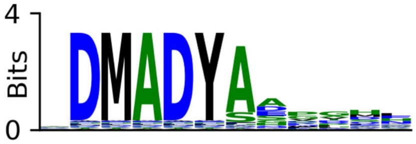
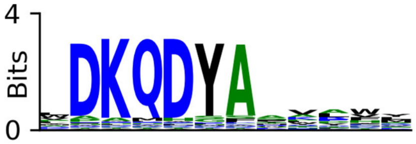
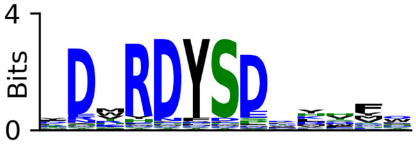
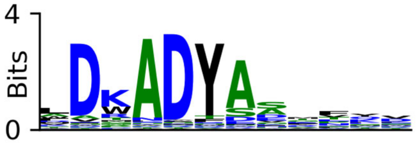
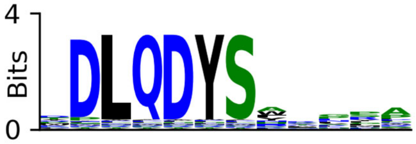
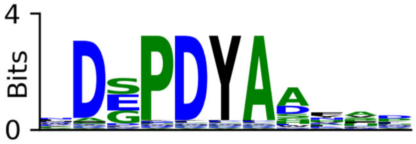
Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	ALIGNED Sequence Logo	Motif Class
1	20	121	EQFHHYDLHDYA, -XXXXXDLHDYAXXXX-		Expanded motif
2	8	121	HQFDKDLQDYAE, -XXXXXDLQDYAXXX--		Expanded motif
3	12	121	GNMNLGDLEDYA, -XXXXXDLEDYAXXX-		Expanded motif
4	9	104	GNFGGDVEDYAY, -XXXXXDVEDYAXXX-		Expanded motif
5	26	102	EMWADLPDYAHA, -XXXXXDLPDYAXXX-		Expanded motif
	19	97	VPTDVQDYAHPR, -XXXXXDVQDYAXXX--		Expanded motif

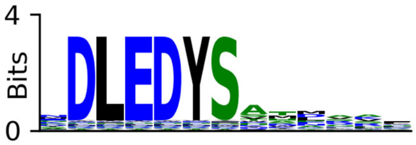
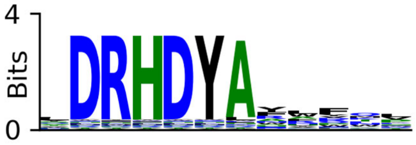
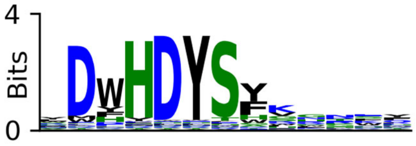
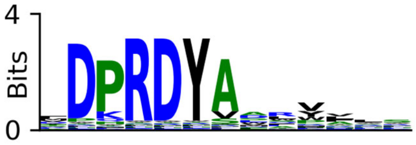
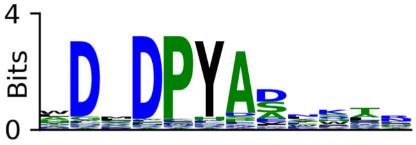
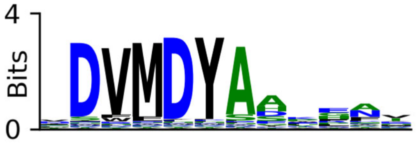
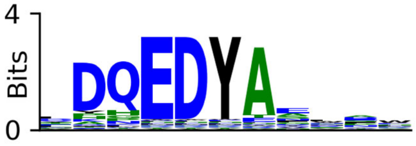
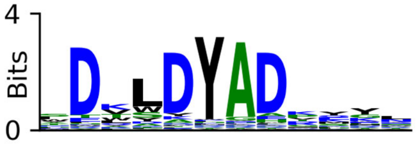
	23	94	HMTDVPDYAYHV, -XXXXXDVPDYAXXX-		HA tag
	11	80	WFFTDMPDYANL, -XXXXXDMPDYXXX--		Expanded motif
	17	71	WFVHDMEDYAMR, -XXXXXDMEDYAXX--		Expanded motif
	61	69	VGGWYDLADYAG, -XXXXXDLADYAXXX--		Expanded motif
	3	66	DVHDYAYGYYHA, --XXXXDVHDYAXXXXX-		Expanded motif
	32	64	WNLDMVDYAAKF, -XXXXXDXVDYAXXXX-		Expanded motif
	10	63	VTWVQDKHDYFS, -XXXXXDKHDYXXXXX--		Expanded motif
	1	62	WDLYDDKTDYAA, XXXXXDXTDYAXX--		Expanded motif

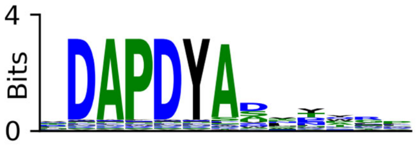
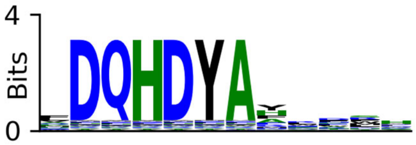
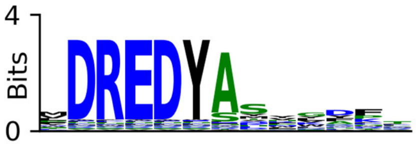
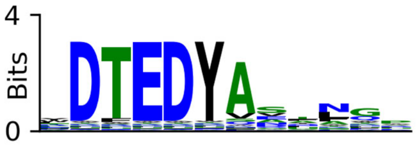
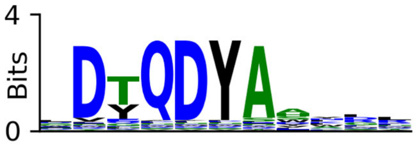
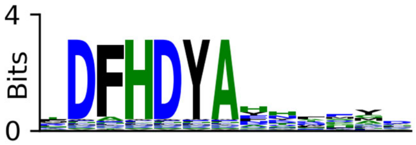
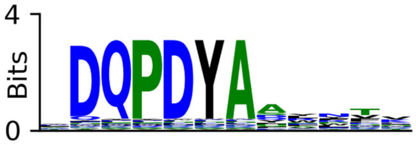
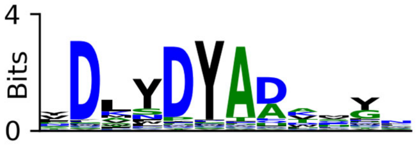
	49	55	WWDFPDYANGRW, XXXXXDFPDYXXXX-		Expanded motif
	39	52	KDMHDYASMHMW, -XXXXDMHDYAXXXX-		Expanded motif
	18	51	FDRDMQDYASML, -XXXXXDMQDYAXXX-		Expanded motif
	43	50	RDLHDYSGPRSN, -XXXXDLHDYSXXXX-		Expanded motif
	25	49	TNFQHDVADYAG, XXXXXDVADYAXXX--		Expanded motif
	54	48	MWLGDTRDYADT, XXXXXDXRDYADX---		Expanded motif
	28	47	SVDVKDYADEWN, XXXXXDXKDYAXXX--		Expanded motif
	37	46	QDWPDYAWGGPR, -XXXXXDWPDYAXXXX		Expanded motif

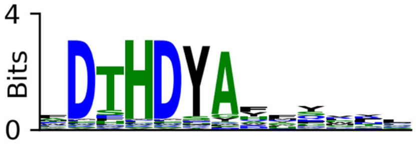
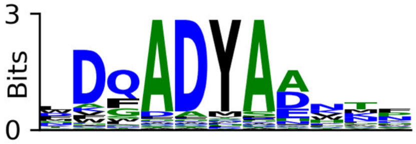
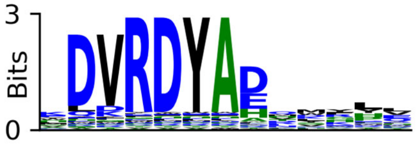
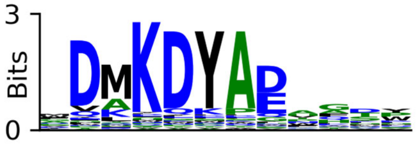
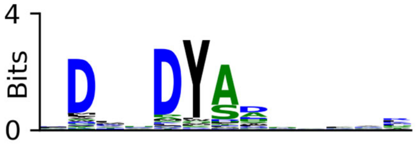
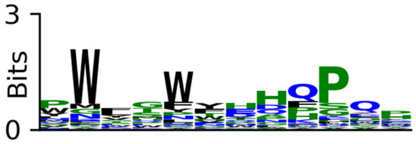
	16	45	YVKDKPDYAYKF, -XXXXXDKPDYXXX--		Expanded motif
	58	43	DALSDLPDYSAS, -XXXXXDLPDYSXXX-		Expanded motif
	13	42	VQTFTDLKDYAW, XXXXXDLKDYAXXX--		Expanded motif
	35	41	FQAFMDKEDYSF, -XXXXXDKEDYAXXX-		Expanded motif
	5	40	VSWDLVDYAWKF, -XXXXXDLVDYAXXX-		Expanded motif
	41	40	LRWHNDWQDYAY, XXXXXDWQDYAXX--		Expanded motif
	65	40	NDMMDYADMDRL, XXXXXDMMDYAXXX-		Expanded motif
	31	38	FAKGDLRDYAQK, -XXXXXDLRDYAXXXX-		Expanded motif

	34	38	PDYHDYAFARGL, XXXXXDXHDYAXXXX-		Expanded motif
	45	38	YDMEDTPDYADM, XXXXXDTPDYAXXX--		Expanded motif
	2	37	VMQFTDQQDYAW, -XXXXXDQQDYAXX--		Expanded motif
	6	37	MLRGDFEDYAAN, -XXXXXDXEDYAXX--		Expanded motif
	51	37	YWEFQDVPDYSY, XXXXXDVPDYSXXX-		Expanded motif
	42	36	AENEDWEDYAST, -XXXXXDWEDYAXXX-		Expanded motif
	24	34	YSNVDLMDYAEP, -XXXXXDLMDYAXX--		Expanded motif
	46	34	TSVDVHDYSAHF, XXXXXDVHDYSXXXX-		Expanded motif

	27	33	LPVHWYDYPDSF, -XXXXXDYPDYAXXX-		Expanded motif
	59	33	FDWHDYAEHVQS, -XXXXXDWHDYAXXXX		Expanded motif
	21	32	WDMADYAEADHL, XXXXXDMADYAXXX-		Expanded motif
	22	30	FRKWDKQDYAYP, --XXXXXDKQDYAXX--		Expanded motif
	60	30	MYRFDRRDYSDQ, -XXXXDXRDYSDXXX-		Expanded motif
	63	30	FSLADKADYAAQ, XXXXXDXADYAXX--		Expanded motif
	48	29	WLQDLQDYSHAP, -XXXXDLQDYSXXXX		Expanded motif
	53	29	MMMVDSPDYAAN, XXXXXDXPDYAXX-		Expanded motif

	30	28	VSNTNYDLEDYS, -XXXXXDLEDYSXXX--		Expanded motif
	44	28	VSTADRHDYAYL, XXXXXDRHDYAXXX-		Expanded motif
	50	28	HFNWYDWHDYSF, XXXXXDXHDYSXXXX-		Expanded motif
	56	27	MMTEDPRDYAFF, -XXXXXDPRDYAXX---		Expanded motif
	67	27	WMMPGDADPYAD, -XXXXXDXDPYAXX--		Expanded motif
	14	26	LTDVMDYAAKEA, -XXXXXDVMDYAXXX-		Expanded motif
	29	26	YFEDQEDYAGWS, -XXXXXDQEDYAXX-		Expanded motif
	40	26	VNSYADTLDYAD, XXXXXDXXDYADX--		Expanded motif

	7	25	SVEDDAPDYADF, -XXXXXDAPDYAXX---		Expanded motif
	15	25	WWHDQHDYAHWT, -XXXXDQHDYAXXX-		Expanded motif
	33	24	FLTQQDREDYAH, -XXXXXDREDYAXX---		Expanded motif
	55	24	WWEATADTEDYA, -XXXXXDTEDYAXX--		Expanded motif
	62	24	VVGGLDTQDYAH, XXXXXDXQDYAX--		Expanded motif
	64	24	FDFHDYAYNQGM, XXXXXDFHDYAXXXX-		Expanded motif
	36	23	YGMLDQPDYAAY, -XXXXXDQPDYAXXX-		Expanded motif
	47	23	ELAYYDTYDYAD, XXXXXDXXDYAXX--		Expanded motif

	57	23	WDTHDYAAWSGT, XXXXXDTHDYAXXXXX-		Expanded motif
	66	22	VLWTFDQADYAE, XXXXXDXADYAX--		Expanded motif
	52	18	DVRDYADDKYYE, XXXXXDVRDYAXXXXX-		Expanded motif
	38	16	AGFDKKDYADAF, XXXXXDXKDYAXXXX-		Expanded motif
	0	1102	, ----XXXXDXXDYXXXXX---		Weak
	4	16	FYWNEMFWDHQP, ---XXXXWXXXXXXXXX-		Weak

13 Motif-based clustering sensitivity of UMAP dimensionality reduction

For this analysis, specific data were isolated from the AS-MS data. Specifically, a variable number of unaligned peptides containing the *DLHDYA* motif were added to random library peptides (which do not contain the motif) for 5000 total. The motif *DLHDYA* was used since it was discovered by clustering of the 12ca5 AS-MS data, most clearly seen in the UMAP + N-grams encoding analysis.

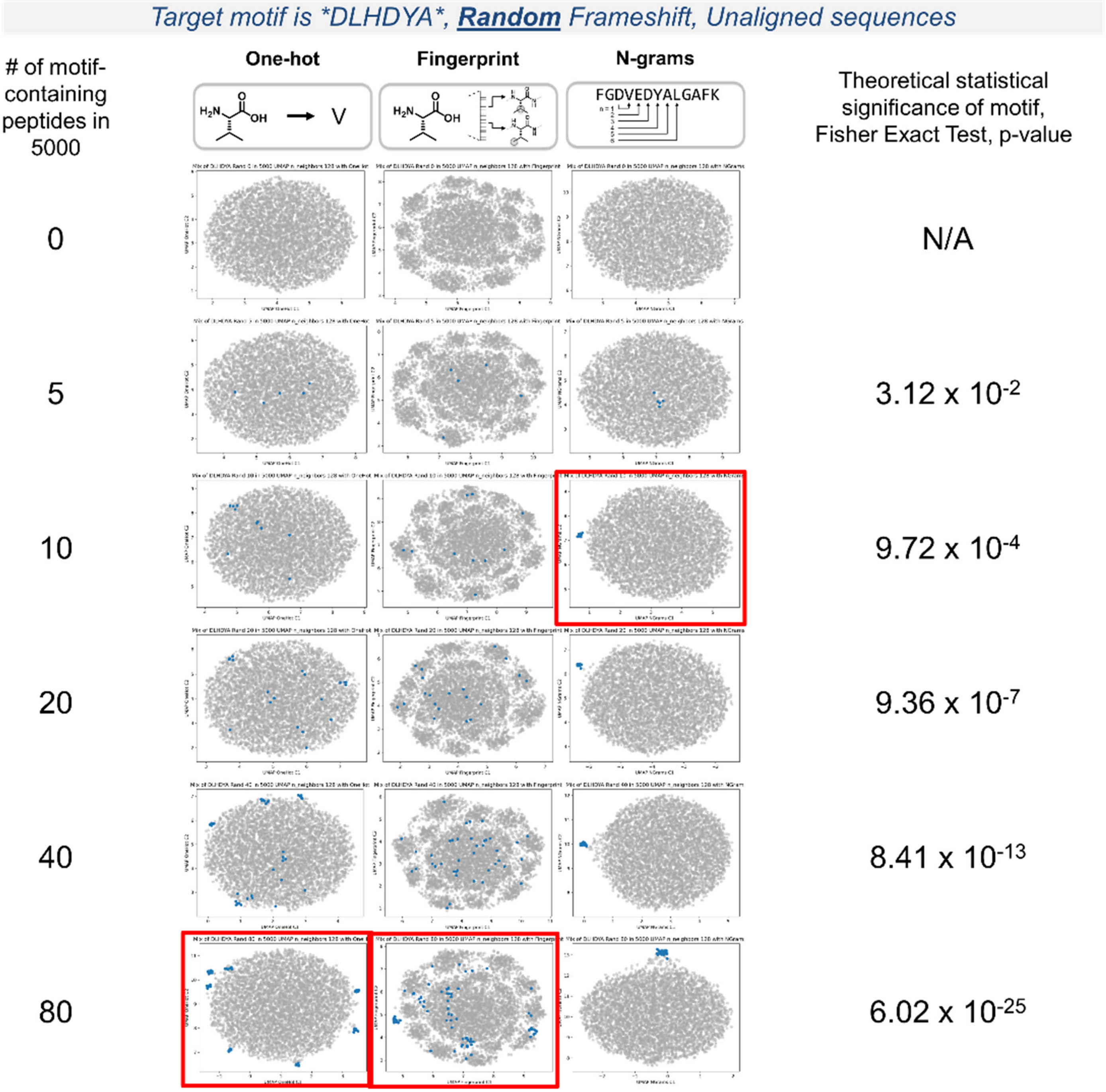


Figure S19. UMAP sensitivity to cluster and enable the detection and isolation of target peptides in a 5000-peptide dataset. Unaligned target peptides contain the high-affinity binding motif of *DLHDYA* at random frameshifts. N-grams demonstrates the lowest sensitivity, with only 10 peptides required for a distinct cluster to

appear. One-hot and Fingerprint encoding requires 80 and 160 peptides, respectively. This result is because N-grams encoding is performed irrespective of frameshift, whereas one-hot and Fingerprint encoding are frameshift sensitive. Thus, as the number of target peptides increases, one-hot and Fingerprint encoded UMAP sequence maps form seven clusters as the seven frameshifts of *DLHDYA* in a 12-mer variable region are populated to have at least 10 peptides in each cluster. A red box is placed to guide the readers eye to location in which clusters appear to form distinctly from the random library peptides. AS-MS peptides are shown in blue with random library peptides in gray. The theoretical statistical significance via Fishers Exact Test of each condition is shown,[11–13] indicating that at only 5 sequences, the peptides with the *DLHDYA* motif could be theoretically distinguished from the background (randomized input dataset), though 10 are required for a clear cluster to form.

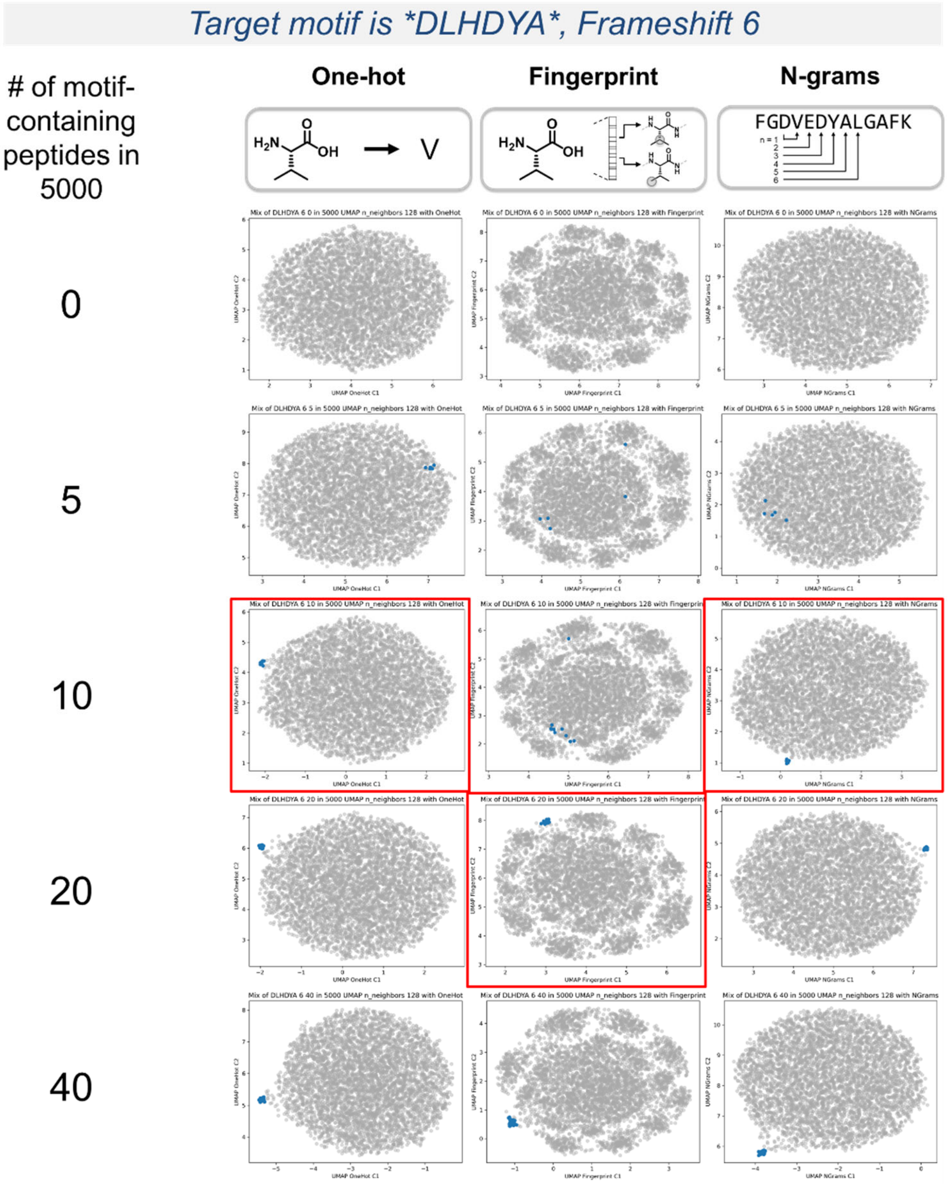


Figure S20. N-grams, one-hot, and Fingerprint encoding provide similar clustering sensitivity with target peptides containing a motif at the same frameshift. See Figure S17 for further details. A red box is placed to guide the readers eye to location in which clusters appear to form distinctly from the random library peptides. AS-MS peptides are shown in blue with random library peptides in gray.

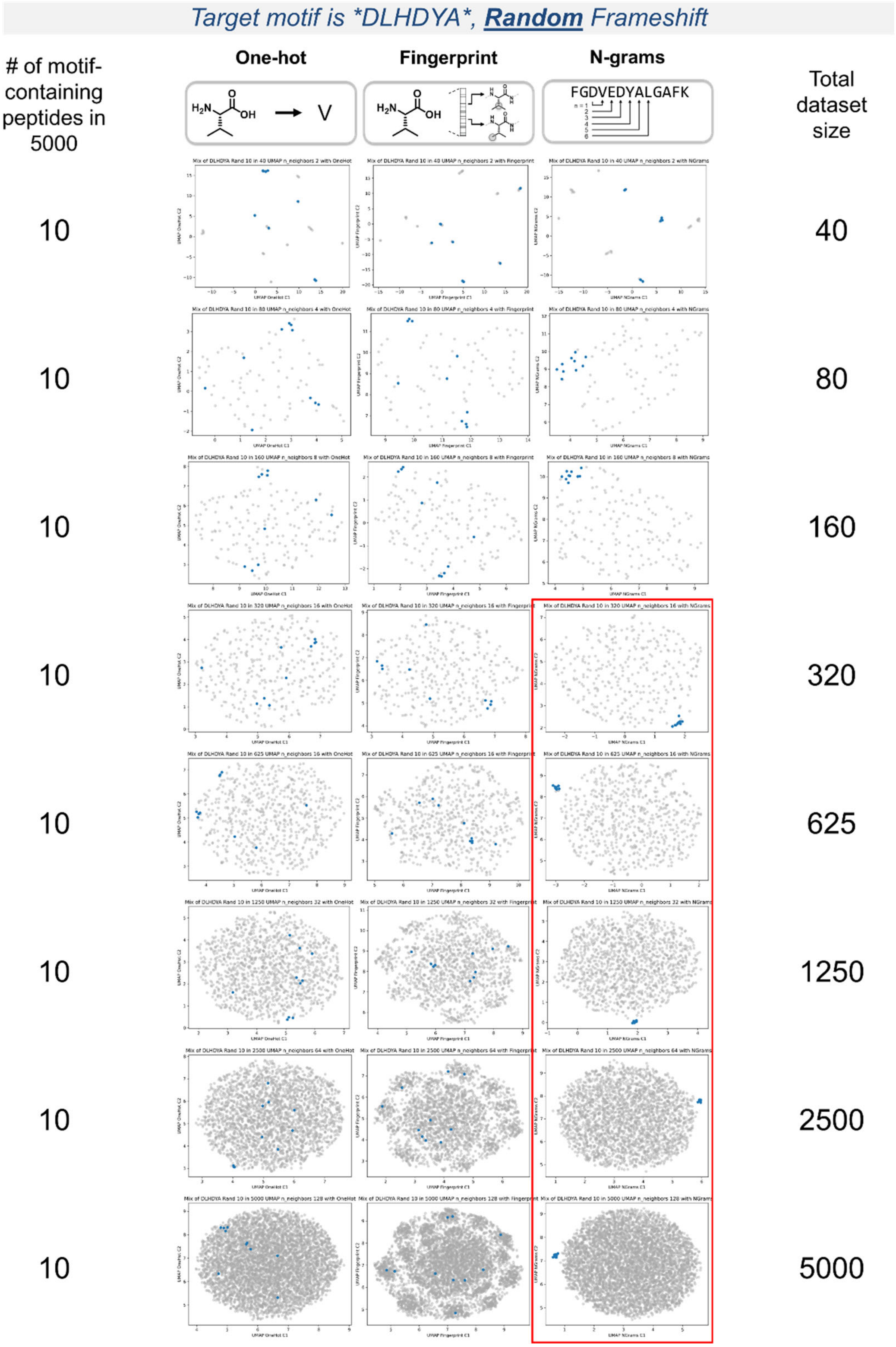


Figure S21. The construction of UMAP sequence space is affected by the total dataset size. At low dataset sizes, highly similar peptides can be dispersed on the sequence space map. Thus, augmenting the total dataset size with random library peptides can sometimes improve clarity of the clusters of similar peptides.

14 Comparison of motif-detection sensitivity with XSTREME

Motif discovery was performed using the XSTREME, part of the MEME Suite webserver.[13,14]

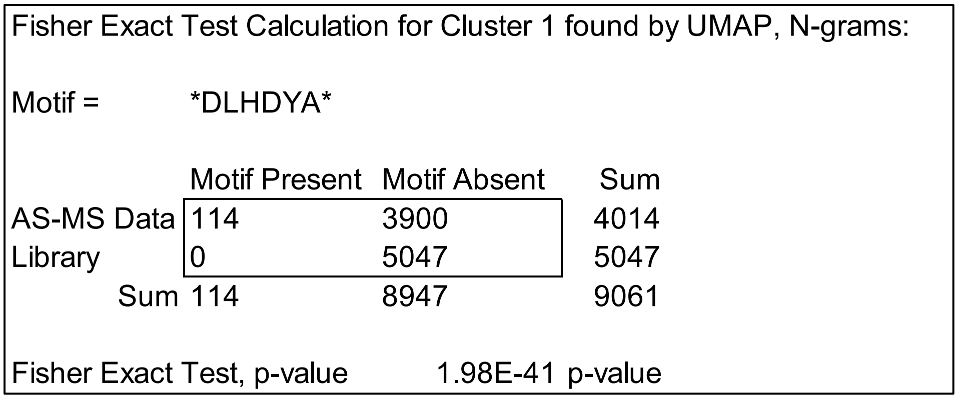
XSTREME combines:

- MEME, which discovers novel, ungapped motifs (recurring, fixed-length patterns) in sequences. MEME will split variable-length patterns into two or more separate motifs.
- STREME, which discovers ungapped motifs (recurring, fixed-length patterns) that are enriched in sequences or relatively enriched in comparison to a control dataset.

Two experiments were performed

1. The AS-MS data was input to XSTREME as the positive dataset with the randomly sampled library peptides as the negative dataset

The Fisher Exact Test can quantify the statistical significance of finding a specific motif, and is used by STREME when a background dataset is input. The motif *DLHDYA*, found in the clustering analysis using UMAP and N-grams encoding. The p-value is 1.98 x 10^{-41}, meaning it should be detected (see below)



Fisher Exact Test Calculation for Cluster 1 found by UMAP, N-grams:

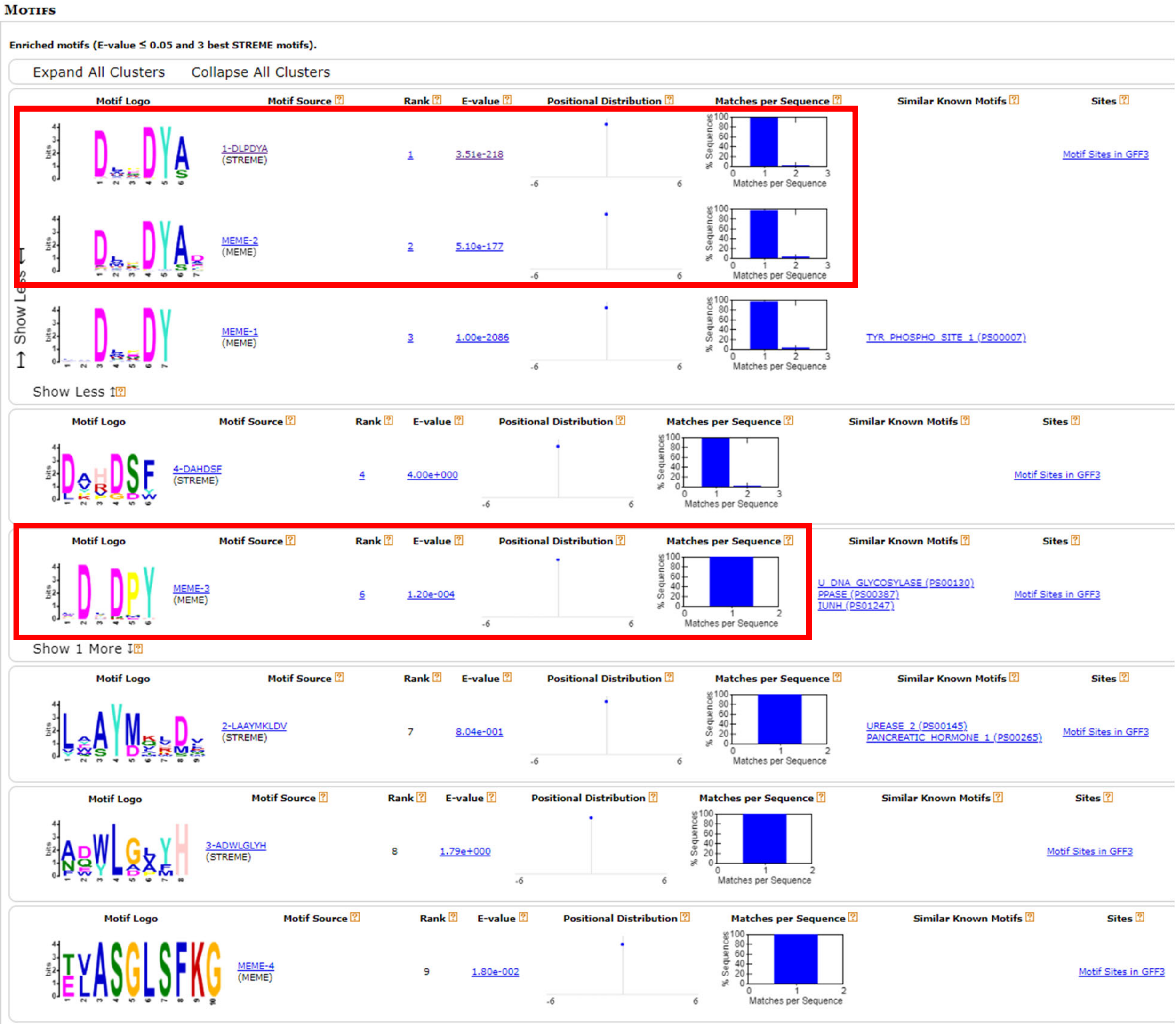
Motif = *DLHDYA*

	Motif Present	Motif Absent	Sum
AS-MS Data	114	3900	4014
Library	0	5047	5047
Sum	114	8947	9061

Fisher Exact Test, p-value 1.98E-41 p-value

2. The sensitivity of motif detection was determined using the same datasets in Figure S17, using either 5, 10, or 20 target peptides that contain a *DLHDYA* motif at random frameshifts.

14.1 XSTREME Experiment 1 (12ca5 AS-MS data vs library):



Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Expand All Clusters Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-DLPDYA (STREME)	1	3.51e-218				Motif Sites in GFF3
	MEME-2 (MEME)	2	5.10e-177				
	MEME-1 (MEME)	3	1.00e-2086			TYR_PHOSPHO_SITE_1 (PS00007)	

Show Less

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	4-DAHDSF (STREME)	4	4.00e+000				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-3 (MEME)	6	1.20e-004			U_DNA_GLYCOSYLASE (PS00130) PPASE (PS00387) IUNH (PS01247)	Motif Sites in GFF3

Show 1 More

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	2-LAAYMKLDV (STREME)	7	8.04e-001			UREASE_2 (PS00145) PANCREATIC_HORMONE_1 (PS00265)	Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	3-ADWLGLYH (STREME)	8	1.79e+000				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-4 (MEME)	9	1.80e-002				Motif Sites in GFF3

Figure S22. XSTREME motif detection result of motifs enriched in the AS-MS dataset (positive) relative to the randomly sampled library peptides (negative). **Boxed in red** are the common motif D**DYA, as well as D**DYAD* and D*DPY* which were the only expanded motif discovered with statistical significance.

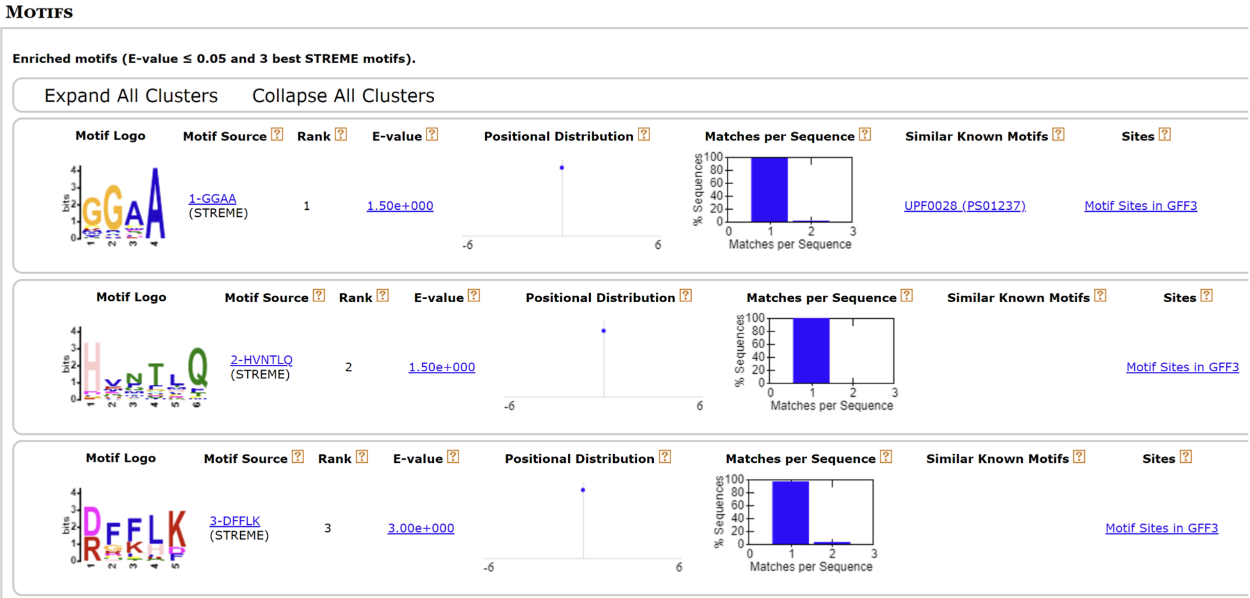
14.2 XSTREME Experiment 2 (Analysis of detection sensitivity of unaligned, motif-containing peptides):

Next, the detection sensitivity was assessed using the same datasets as in Figure S17 with 5, 10, and 20 target peptides, containing an unaligned *DLHDYA* motif in dataset of 5000 random library peptides.

For our clustering approach, all 5,000 sequences were input, whereas for XSTREME analysis, the same 5,000 input sequences were compared against a background dataset constructed from the randomization of the input sequences.

14.2.1 5 target peptides in 5000

XSTREME Summary (STREME + MEME):



MOTIFS

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

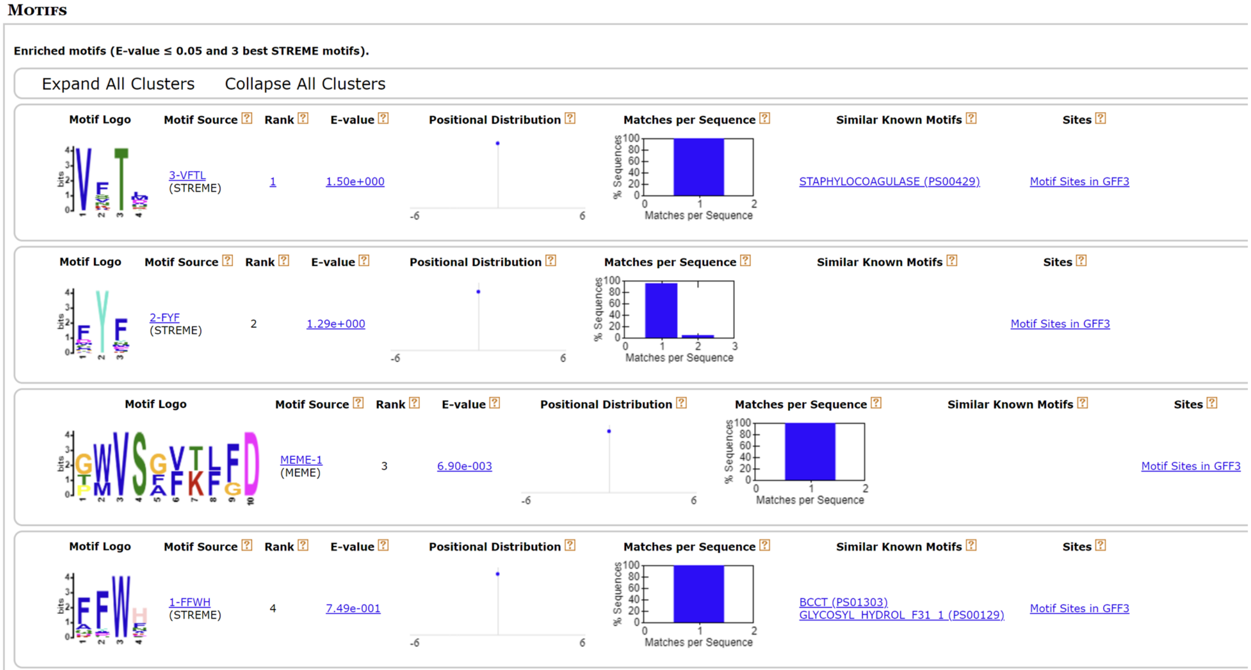
Expand All Clusters Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-GGAA (STREME)	1	1.50e+000			UPF0028 (PS01237)	Motif Sites in GFF3
	2-HVNTLQ (STREME)	2	1.50e+000				Motif Sites in GFF3
	3-DFFLK (STREME)	3	3.00e+000				Motif Sites in GFF3

Figure S23. The XSTREME results for motif discovery and detection using the dataset of 5 target peptides in 5000 random library peptides. None of the motifs are statistically significant and the 5 *DLHDYA* peptides were not identified. STREME reported all these motifs, and motifs evaluated by the Binomial Test, providing the p-value reported.

14.2.2 10 target peptides in 5000

XSTREME Summary (STREME + MEME):



Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

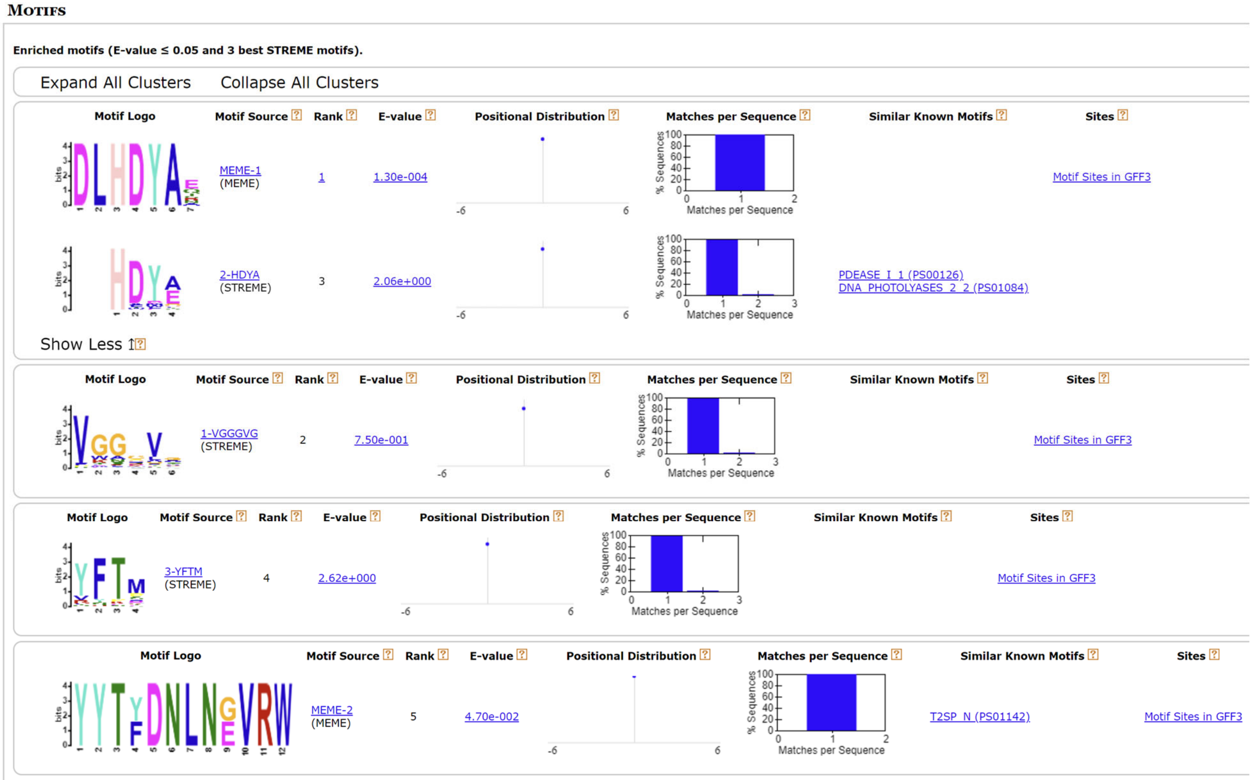
Expand All Clusters   Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	3-VFTL (STREME)	1	1.50e+000			STAPHYLOCOAGULASE (PS00429)	Motif Sites in GFF3
	2-FYF (STREME)	2	1.29e+000				Motif Sites in GFF3
	MEME-1 (MEME)	3	6.90e-003				Motif Sites in GFF3
	1-FFWH (STREME)	4	7.49e-001			BCCT (PS01303) GLYCOSYL_HYDROL_F31_1 (PS00129)	Motif Sites in GFF3

Figure S24. The XSTREME and STREME results for motif discovery and detection using the dataset of 10 target peptides in 5000 random library peptides. The 10 *DLHDYA* peptides were not identified. STREME reported motifs were evaluated by the Fisher Exact Test, providing the p-value (E-value * # of reported sequences) reported. MEME reported motifs were evaluated by E-value.

14.2.3 20 target peptides in 5000

XSTREME Summary (STREME + MEME):



Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Expand All Clusters   Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-1 (MEME)	1	1.30e-004				Motif Sites in GFF3
	2-HDYA (STREME)	3	2.06e+000			PDEASE_I_1 (PS00126) DNA_PHOTOLYASES_2_2 (PS01084)	

Show Less

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-VGGGVG (STREME)	2	7.50e-001				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	3-YFTM (STREME)	4	2.62e+000				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-2 (MEME)	5	4.70e-002			T2SP_N (PS01142)	Motif Sites in GFF3

Figure S25. The XSTREME and STREME results for motif discovery and detection using the dataset of 20 target peptides in 5000 random library peptides. The 20 *DLHDYA* peptides were identified, and were calculated to be statistically significant with an E-value of 1.3 x 10^{-4} from discovery by MEME analysis. This is twice as many as can be clearly seen by our clustering approach.

15 Augmentation of sequence maps with noncanonical peptides discovered by AS-MS

Table S16. Peptidomimetics discovered using AS-MS for purity and LCMS characterization see *Analytical characterization of all synthesized noncanonical peptidomimetics discovered by AS-MS*. For BLI characterization see *SI Section 19 Biolayer interferometry (BLI) measurements*, Table S17

Peptide #	Sequence, 1-letter code	ALC	Sequence, 3-letter code for noncanonicals													Binder / Nonbinder	KD, nM (Ave ± SD)
1	HoiDueDYAoxPK	90	H	hArg	Tha	D	Nal	Hyp	D	Y	A	hArg	Psa	P	Lys	Binder	44 ± 29
2	duiDueDYAoxPK	98	Cpa	Nal	Tha	D	Nal	Hyp	D	Y	A	hArg	Psa	P	Lys	Binder	75 ± 56
3	giibmDpoDYAiK	99	Thp	Tha	Tha	Aib	3fF	D	hCit	hArg	D	Y	A	Tha	Lys	Binder	3.1 ± 0.67
5	tzwksnYVkuliK	93	Cxf	Dpf	Dph	4Af	Php	pSer	Y	V	4Af	Nal	Msn	Tha	Lys	Binder	77 ± 57
15	pgYDwDVADYADK	91	hCit	Thp	Y	D	Dph	D	V	A	D	Y	A	D	Lys	Binder	3.9 ± 0.68
16	jVVdDQPDYAtlK	99	Tic	V	V	Cpa	D	Q	P	D	Y	A	Cxf	Msn	Lys	Binder	0.21 ± 0.15
17	xPAGDTPDYADmK	93	Psa	P	A	G	D	T	P	D	Y	A	D	3fF	Lys	Binder	4.4 ± 2.7
4	ovuxjvVrbevGK	94	hArg	2F3F	Nal	Psa	Tic	2F3F	V	DfF	Aib	Hyp	2F3F	G	Lys	Nonbinder	
6	ktGwzTQwpptZK	91	4Af	Cxf	G	Dph	Dpf	T	Q	Dph	hCit	hCit	Cxf	Git	Lys	Nonbinder	
7	jmHVGwhYAQAHK	90	Tic	3fF	H	V	G	Dph	Amb	Y	A	Q	A	H	Lys	Nonbinder	
8	irhTAsjViDYAK	88	Tha	DfF	Amb	T	A	Php	Tic	V	Tha	D	Y	A	Lys	Nonbinder	
9	uTxpzdpmmjTzK	87	Nal	T	Psa	hCit	Dpf	Cpa	hCit	3fF	3fF	Tic	T	Dpf	Lys	Nonbinder	
10	TNXfQYvoTYifK	84	T	N	Agn	Pip	Q	Y	2F3F	hArg	T	Y	Tha	Pip	Lys	Nonbinder	
11	iiAldjwTtswzK	84	Tha	Tha	A	Msn	Cpa	Tic	Dph	T	Cxf	Php	Dph	Dpf	Lys	Nonbinder	
12	NfXlKDbutvzdK	83	N	Pip	Agn	Msn	K	D	Aib	Nal	Cxf	2F3F	Dpf	Cpa	Lys	Nonbinder	
13	swrYPzTmjGexK	81	Php	Dph	DfF	Y	P	Dpf	T	3fF	Tic	G	Hyp	Psa	Lys	Nonbinder	
14	NrTzzdkYmjzTK	81	N	DfF	T	Dpf	Dpf	Cpa	4Af	Y	3fF	Tic	Dpf	T	Lys	Nonbinder	

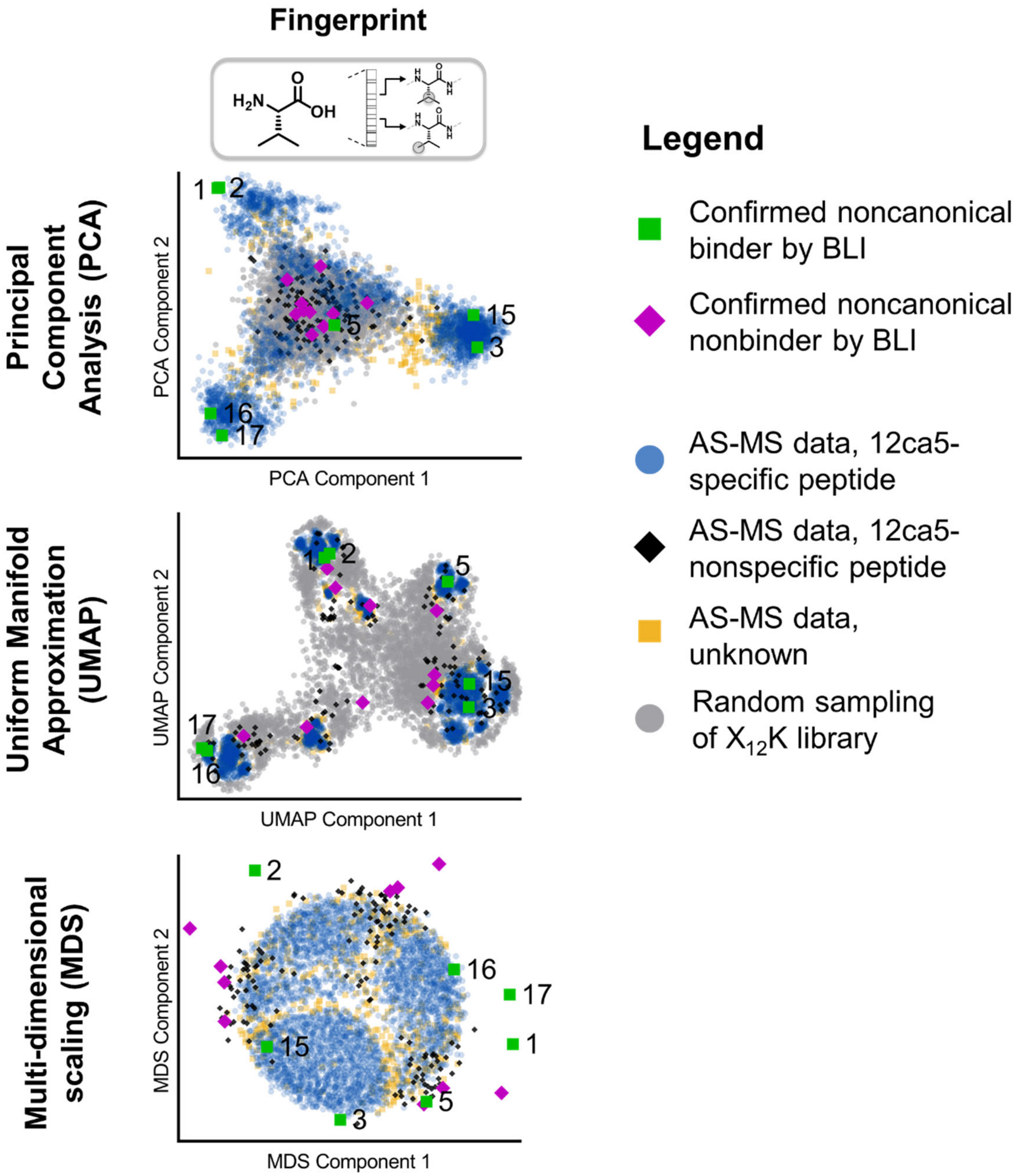


Figure S26. Augmentation of canonical sequence maps with noncanonical peptides discovered from AS-MS and experimentally evaluated using BLI to distinguish binders from nonbinders (see *Biolayer interferometry (BLI) measurements*). Peptides are labeled with their respective numbers. Also included are the 12ca5-based labels as defined in Label definitions for 12ca5-specific and nonspecific binders. Seventeen noncanonical peptides were added to the dataset and the sequence space was relearned and then the randomly sampled peptides from the canonical $X_{12}K$ library were added to the PCA and UMAP maps. The randomly sampled peptides cannot be added to MDS without re-learning.

16 Peptide synthesis and cleavage

Peptides and peptidomimetic α-carboxamides were manually synthesized in batch using 100 mg of H-Rink Amide ChemMatrix resin (0.49 mmol/g). Resin was swollen in amine-free DMF for a minimum of 10 minutes in HSW Norm-Ject syringe (Torviq) syringes mounted on a Restek Resprep SPE vacuum manifolds equipped (Cat 26077) with valves. For each coupling cycle, Fmoc-protected amino acids (5 eq, 0.245 mmol) were dissolved at 0.4 M in 0.38 M HATU (4.75 eq relative to resin, 0.95 eq relative to Fmoc-protected amino acid) in amine free DMF and sonicated or vortexed as needed. Diisopropylethyl amine (DIEA; 10 eq, 0.49 mmol, 85.4 µL) was added and the solution, hand mixed to form the active ester, and confirmed to return being visually transparent as a clear light yellow solution. Using the Restek manifold, the excess DMF was drained from the DMF-swelled resin. Then the solution containing the activated Fmoc-amino acid ester was added to the resin and incubated at room temperature for 45 minutes. After which, the resin was drained and washed 3 x with amine free DMF. Fmoc deprotection was completed using 20% piperidine in DMF (2 x 5 minutes), and then washed 3 x with amine free DMF. Then the next amino acid coupling cycle could proceed. After synthesis was complete, resins were washed 5 x with amine free DMF, 3 x DCM, vacuum was pulled on the dry resin to remove the DCM (5 minutes), and then the resin was dried under vacuum before cleavage.

Cleavage was performed in HSW Norm-Ject syringe (Torviq) syringes by using the syringe plunger to pull the cleavage solution onto the resin with a blunt tip needle and then capping the syringe. Global side chain deprotection and cleavage from solid support were carried out using solution of 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane, for 1 hour minimum at ambient temperature (~2 mL of deprotection solution / 100 mg of resin). Upon which, the crude peptide and cleavage solution was isolated from the syringe into a 15 mL Falcon tube and triturated with cold diethyl ether (~12 mL, chilled on dry ice). The peptide was then suspended in 50% acetonitrile in water (0.1% TFA) and lyophilized.

Peptide purification was completed using reverse-phase flash purification or with preparative high performance liquid chromatography purification (HPLC). For flash purification, a Biotage Selekt was used with a Biotage® Sfär C18 D - Duo 100 Å 30 µm 12 g column. One-third of the cleaved, lyophilized peptide mass (< 10 mg) was suspended in 0.9 to 1.8 mL of 20% MeCN in Water (0.1% TFA), centrifuged at 3.4k rcf for 10 minutes, and the supernatant was loaded onto the column and separated using using a gradient of 10% to 55% MeCN in Water (0.1% TFA) over 12-15 column volumes (CVs) and observed by UV absorption at 210 and 280 nm and fraction collected with 3 mL maximum fraction sizes. Peptides that exhibited close elution to deletion products or poor elution profiles were purified by preparative HPLC. Preparative HPLC was performed on an Agilent 1260 Infinity LC equipped with a 6130 single quadrupole mass spectrometer. Samples were prepared as described above, filtered using a 0.2 µm filter, and loaded onto a Zorbax 300SB C18 column (9.4 x 150 mm, 5 µm, 8 mL/min) with a C8 guard column using a automated injector and separated using 5% to 55% MeCN in Water (0.1% TFA) over 30 minutes with fractionation over the entire run using 62 fractions. Fractions were analyzed by LCMS and UPLC to assess purity.

17 Liquid-Chromatography Mass Spectrometry (LC-MS) analysis

LC-MS analysis was acquired using an Agilent 6550 MS Q-TOF mass spectrometer with Dual Agilent Jet Stream (AJS) ESI ion source in extended dynamic mode in mass range 100 - 3000 m/z with scan rate of 1.00 spectra/sec. An isopump delivered a reference ion mass (922.0098 m/z). The following instrument parameters were used: gas temperature 200 ºC, gas flow 14 L/min, nebulizer pressure 55 psig, sheath gas temperature 350 ºC, sheath gas flow 11 L/min. The following scan source parameters were used: VCap: 3500, nozzle voltage 1000 V, fragmentor 175, and Octopole RF Vpp 750. Column was a Zorbax 300SB C3, 2.1 × 150 mm, 5 µm kept at 40 ºC. The gradient utilized 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B), flow rate 0.5 mL/min, starting at 1% B in A running to 91% B in A over 7 minutes with 1 minute at 91% B in A and 1 minute post-time re-equilibration at 1% B in A. Data were analyzed in Agilent MassHunter Qualitative Analysis B.06.00.

18 Purity analysis by Ultra Performance Liquid Chromatography (UPLC)

LC analysis was performed with an Agilent 1260 LC system controlled by ChemStation software, using an Agilent Zorbax RRHD 300SB-C18, 2.1 x 50 mm, 1.8 μm (Cat: 857750-902) column at 40 ºC. The gradient utilized 0.1% trifluoroacetic acid (TFA) in water (solvent A) and 0.1% TFA in acetonitrile (solvent B). The flow rate was 0.5 mL/min, starting at 5% B in A running to 65% B in A over 11 minutes, moving to 90% B in A in 0.25 minute, holding for 1 minute, moving to 5% B in A in 0.05 minute, and re-equilibrating for 1.5 minutes. Approximately 1-10 ug of each peptide was injected for analysis for a target response of <1000 mAU. The absorbance at 214 nm was recorded and integrated using ChemStation software to report the purity relative to an equal volume injection of 50% acetonitrile in water.

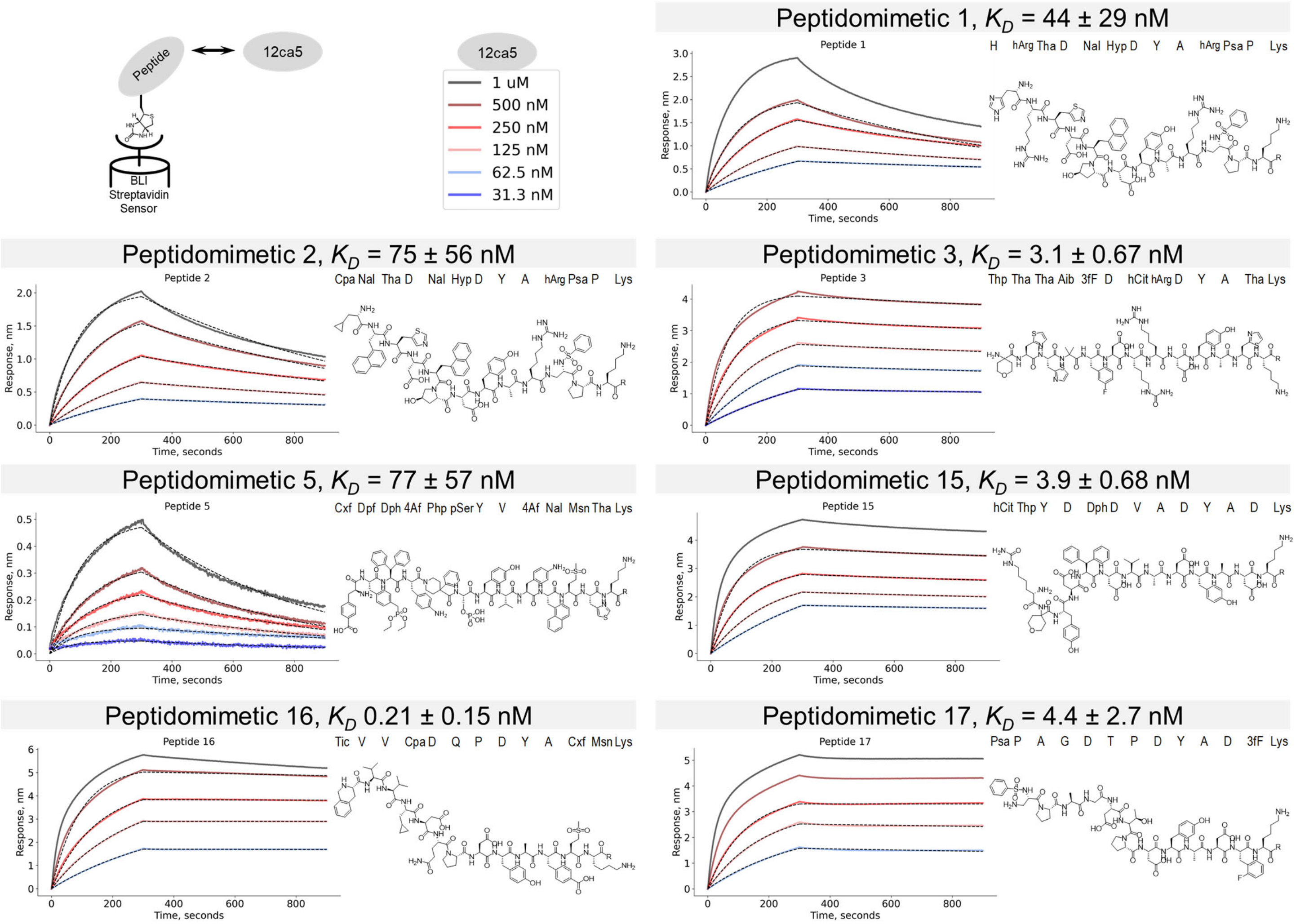
19 Biolayer interferometry (BLI) measurements

Ideally, proteins including 12ca5 would be immobilized and dipped into solutions of the peptides to test their binding activity. This immobilization orientation is preferred because it would use the same biotinylated 12ca5 used in AS-MS in the same orientation and avoid potential avidity affects. However, when immobilizing 12ca5 onto the BLI tip, insufficient signal was observed when dipping into solutions of known peptide binders. This lack of signal was attributed to the relatively small size of these peptides (e.g., ~2 kDa HA tag) to the size of the immobilized 12ca5 (~150 kDa). Thus, biotinylated peptides were prepared using a resin preloaded with GGSK(Biotin). To avoid avidity effects and use a 1:1 model, the ligand density of the immobilized biotinylated peptide or peptidomimetic on the BLI tip was immobilized slowly (over 300 s) up to ≤ 60% of saturation level.

BLI was carried out using the GatorBio GatorPlus Label-Free Analysis system using Greiner Bio-One 96-well Non-treated Black Polypropylene Microplates (FisherSci Cat 07-000-110) using Streptavidin (SA) Probes (GatorBio Cat 160002). All well solution conditions were prepared using kinetics buffer (K Buffer, 0.02% BSA and 0.02% Tween20 in 1x PBS pH 7.4, 0.2 µm filtered). SA tips were equilibrated in K Buffer for 15 minutes prior to analysis. Plate temperature was set to 30 °C with agitation speed at 1000 rpm during measurement and 200 µL well volumes were used.

During each run, sensor tips were equilibrated K buffer (120 seconds), then dipped into of 50–500 nM biotinylated peptide solution for peptides immobilization (300 seconds), with an additional well with no peptide as a control. Concentrations of the peptide immobilization solutions were surveyed beforehand and adjusted such that the peptide response signal (nm) arrived at 60% or less of its saturation level during 300 seconds of immobilization. This extra step was done to appropriately load the tip to minimize avidity effects during downstream association per manufacturer recommendation. Once loaded with peptides, the tips were then moved into wells containing various concentrations of 12ca5 (nonbiotinylated) for association measurement, with an additional well corresponding to a sensor tip with immobilized peptide with no protein as a control. After association (300 seconds), the tips were moved to a well with K buffer to obtain the dissociation (600 seconds). Peptide-only and protein-only conditions (concentration at 1000 nM) were used as references for background subtraction. The association and dissociation curves were fitted with the GatorOne Software (v 2.7.3.1013) using a 1:1 binding model (n ≥ 3 fit curves accepted with Full $R^2 > 0.8$ and $X^2 < 32$, see Table S17) to calculate the apparent dissociation constant (K_D, reported as the average of the fits ± standard deviation of the fits).

20 BLI Curves of all AS-MS discovered noncanonical peptidomimetics



Peptidomimetic 1, K_D = 44 ± 29 nM

Peptidomimetic 2, K_D = 75 ± 56 nM

Peptidomimetic 3, K_D = 3.1 ± 0.67 nM

Peptidomimetic 5, K_D = 77 ± 57 nM

Peptidomimetic 15, K_D = 3.9 ± 0.68 nM

Peptidomimetic 16, K_D 0.21 ± 0.15 nM

Peptidomimetic 17, K_D = 4.4 ± 2.7 nM

Figure S27. BLI sensorgrams of all binding peptides and peptidomimietics with their monomers and structures shown. Peptides were labeled with a SGGLys(Biotin)-NH2 (labeled as R) at the C-terminus. In the top left, the BLI assay format is shown, with biotinylated peptides immobilized and 12ca5 in solution at the concentrations shown. Note that Peptide 4, 6, 7, 8, 9, 10, 11, 12, 13, and 14 are nonbinders seen in Figure S26. The association and dissociation curves were fitted using a 1:1 binding model (n ≥ 3 fit curves accepted shown as black dashed lines with Full R^2 > 0.8 and X^2 < 32, see Table S17) to calculate the apparent dissociation constant (K_D).

Table S17. BLI Data Summary of all binding peptides and peptidomimetics in this work. Note that Peptide 4, 6, 7, 8, 9, 10, 11, 12, 13, and 14 are nonbinders.

Peptide 1

Dissociation Constant, K_D
44 ± 29 nM
Ave ± SD nM

12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
2000	1.52E-03	1.05E+04	1.45E-07	3.13	2.92	3.06	0.624	0.978	38.30
1000	1.27E-03	1.38E+04	9.21E-08	3.07	2.81	2.90	0.689	0.986	19.00
500	1.08E-03	1.80E+04	6.03E-08	2.28	2.03	1.99	0.584	0.993	4.22
250	7.84E-04	2.16E+04	3.62E-08	2.10	1.84	1.58	0.674	0.998	0.98
125	5.71E-04	2.63E+04	2.17E-08	1.68	1.43	0.99	0.594	0.999	0.12
62.5	3.44E-04	2.87E+04	1.20E-08	1.67	1.40	0.66	0.649	0.999	0.06

Peptide 2

Dissociation Constant, K_D
75 ± 56 nM
Ave ± SD nM

12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
2000	1.35E-03	7.60E+03	1.77E-07	2.54	2.33	2.43	0.790	0.983	16.20
1000	1.16E-03	1.01E+04	1.15E-07	2.24	2.01	2.02	0.780	0.991	5.86
500	9.71E-04	1.31E+04	7.41E-08	1.97	1.71	1.57	0.794	0.996	1.57
250	7.37E-04	1.63E+04	4.52E-08	1.60	1.36	1.05	0.842	0.999	0.29
125	5.81E-04	2.13E+04	2.73E-08	1.26	1.03	0.65	0.770	0.999	0.08
62.5	4.36E-04	3.23E+04	1.35E-08	0.92	0.75	0.40	0.743	0.998	0.06

Peptide 3

Dissociation Constant, K_D
3.1 ± 0.67 nM
Ave ± SD nM

12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
1000	9.01E-05	1.90E+04	4.73E-09	4.54	4.52	4.67	0.507	0.963	73.80
500	1.16E-04	2.81E+04	4.14E-09	4.19	4.16	4.21	0.513	0.986	30.10
250	1.37E-04	3.91E+04	3.51E-09	3.55	3.50	3.38	0.496	0.996	6.57
125	1.54E-04	4.97E+04	3.10E-09	3.08	3.01	2.58	0.499	0.999	1.17
62.5	1.54E-04	6.15E+04	2.51E-09	2.80	2.69	1.89	0.496	1.000	0.43
31.3	1.37E-04	6.01E+04	2.28E-09	2.69	2.51	1.13	0.505	1.000	0.15

Peptide 5

Dissociation Constant, K_D
77 ± 57 nM
Ave ± SD nM

M Conc.(nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
1000	1.85E-03	8.55E+03	2.16E-07	0.60	0.49	0.48	1.389	0.986	0.66
500	1.82E-03	1.17E+04	1.56E-07	0.44	0.34	0.31	1.314	0.991	0.18
250	1.50E-03	2.53E+04	5.92E-08	0.30	0.24	0.23	1.334	0.979	0.18
125	1.33E-03	5.63E+04	2.37E-08	0.19	0.16	0.15	1.345	0.955	0.16
62.5	8.12E-04	1.72E+05	4.72E-09	0.11	0.10	0.10	1.370	0.854	0.15
31.3	1.21E-03	4.34E+05	2.79E-09	0.05	0.05	0.05	1.332	0.768	0.08

Peptide 15

Dissociation Constant, K_D
3.9 ± 0.68 nM
Ave ± SD nM

12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
1000	9.35E-05	1.86E+04	5.01E-09	4.61	4.59	4.73	0.293	0.977	48.70
500	1.07E-04	2.59E+04	4.15E-09	3.78	3.75	3.76	0.266	0.994	11.40
250	1.28E-04	3.42E+04	3.73E-09	3.05	3.01	2.82	0.264	0.999	0.78
125	1.31E-04	4.20E+04	3.13E-09	2.76	2.70	2.17	0.256	1.000	0.17
62.5	1.08E-04	3.32E+04	3.26E-09	3.71	3.53	1.69	0.289	1.000	0.10

Peptide 16

Dissociation Constant, K_D
≤ 1 nM*
*Measured 0.21 ± 0.15 nM (Ave ± SD), out of range for instrument

12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
1000	1.08E-04	1.61E+04	6.71E-09	5.67	5.63	5.76	0.297	0.972	97.70
500	5.15E-05	2.12E+04	2.43E-09	5.26	5.24	5.11	0.299	0.993	30.40
250	8.47E-06	2.71E+04	3.12E-10	4.41	4.41	3.87	0.262	0.999	4.77
125	1.22E-07	3.33E+04	3.67E-12	4.06	4.06	2.93	0.303	1.000	0.72
62.5	1.32E-05	4.15E+04	3.18E-10	3.15	3.14	1.73	0.286	1.000	0.27

Peptide 17

Dissociation Constant, K_D
4.4 ± 2.7 nM
Ave ± SD nM

M Conc.(nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Assoc.X2
1000	NA	1.86E+04	NA	4.99	4.99	5.20	0.254	0.966	67.90
500	NA	2.54E+04	NA	4.37	4.37	4.41	0.257	0.991	19.50
250	2.83E-05	3.31E+04	8.54E-10	3.63	3.61	3.38	0.221	0.999	1.78
125	2.02E-04	3.99E+04	5.05E-09	3.36	3.23	2.59	0.269	0.999	0.19
62.5	3.10E-04	4.17E+04	7.44E-09	3.07	2.74	1.62	0.267	0.999	0.11

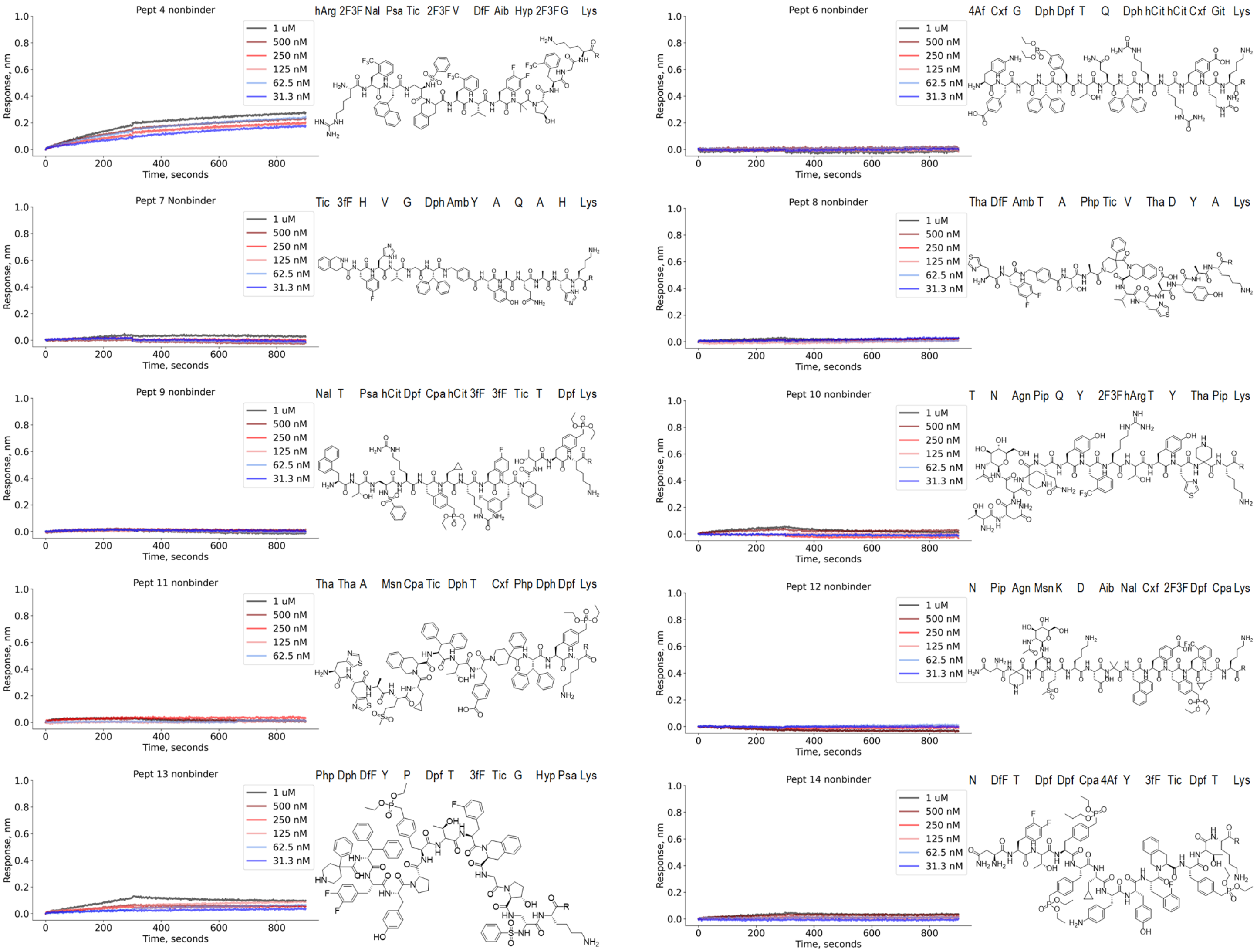


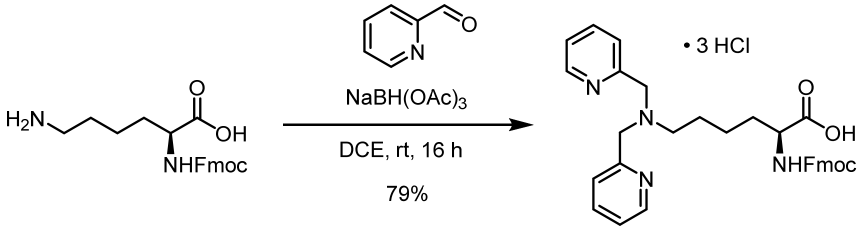
Figure S28. BLI sensorgrams of all nonbinding peptides and peptidomimietics with their monomers and structures shown. Peptides were labeled with a SGGLys(Biotin)-NH2 (labeled as R) at the C-terminus.

21 Noncanonical monomer synthesis

Reactions were monitored on glass-backed analytical thin-layer chromatography (TLC) plates (250 µm, 60 Å, SiliaPlate) containing a fluorescent indicator (254 nm). NMR spectra were recorded on a Bruker AVIII HD 400 MHz or Bruker Neo 500 MHz. ^{1}H NMR chemical shifts are reported in parts per million (ppm, δ scale) and are referenced to the residual protonated NMR solvent (DMSO-*d*6: δ 2.50). All ^{13}C spectra recorded are proton decoupled with chemical shifts reported in parts per million (ppm, δ scale) and are referenced to the carbon resonance of the NMR solvent (DMSO-*d*6: δ 39.5). ^{1}H NMR spectroscopic data are reported as follows: chemical shift in ppm (multiplicity, coupling constants J (Hz), assigned number of protons in molecule). The multiplicities are abbreviated with s (singlet), br. s (broad singlet), d (doublet), t (triplet), and m (multiplet). The chemical shift of all signals is reported as the center of the resonance range, except in the case of multiplets, which are reported as ranges in chemical shift. All raw fid files were processed, and the spectra analyzed using the program MestReNOVA 14.2 from Mestrelab Research S. L. High-resolution mass spectra were obtained on an Agilent Technologies 6550 Q-TOF LC/MS systems (see *Analysis methods with Liquid-Chromatography Mass Spectrometry (LC-MS)*).

21.1 Synthesis of Fmoc-Bpl-OH

N^2-(((9H-fluoren-9-yl)methoxy)carbonyl)-N^6,N^6-bis(pyridin-2-ylmethyl)-L-lysine (Fmoc-Bpl-OH) (2)



1 2

To a 0°C suspension of Fmoc-Lys-OH (1.50 g, 4.07 mmol, 1.0 eq.) and NaBH(OAc)$_3$ (2.59 g, 12.2 mmol, 3.0 eq.) in dichloroethane (22.6 mL) under nitrogen atmosphere, 2-pyridinecarboxaldehyde (0.965 mL, 1.09 g, 10.2 mmol, 2.5 eq.) was added and the resulting suspension was stirred at rt for 16 h. After checking the completion of the reaction by LC-MS, the suspension was cooled to 0°C and quenched by addition of MeOH (25 mL). The resulting solution was concentrated under reduced pressure, the residue redissolved in 4:1 MeCN/H$_2$O and purified by reverse phase column chromatography (Biotage® Sfär C18 D Duo 100 Å 30 µm 30 g, MeCN + 0.1% HCl : H$_2$O + 0.1% HCl = 1:9 → 4:1) to afford the title compound as dark yellow solid (2.12 g, 79%). 0.1% HCl was used rather than 0.1% trifluoroacetic acid to prevent against any possible trifluoracetylation during coupling.

ESI-HRMS: calc. $C_{33}H_{34}N_4O_4$ $[M+H]^+$ 551.2658 found 551.2714, 10.2 ppm error.

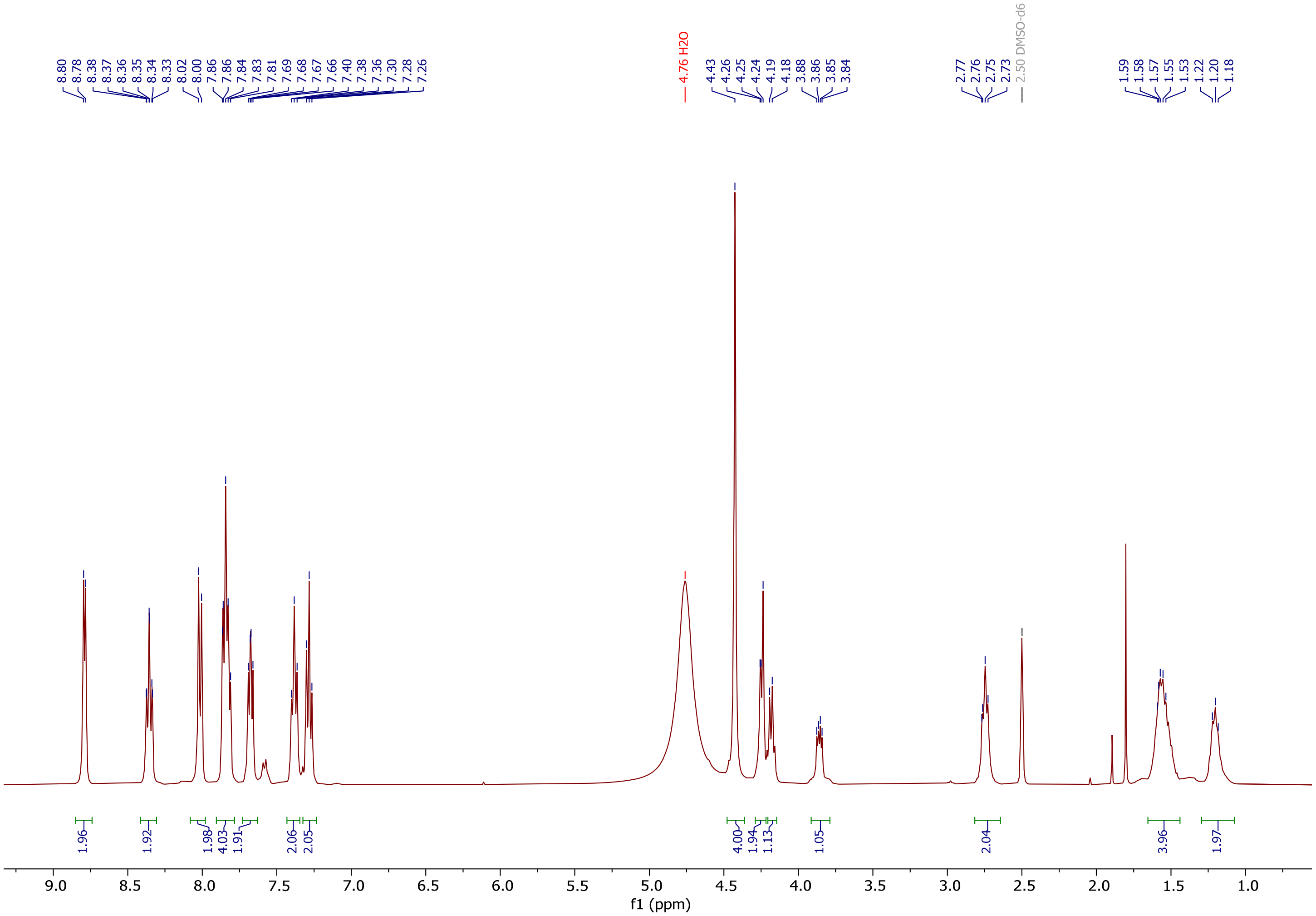


Figure S29. ^{1}H NMR (400 MHz, DMSO-*d6* + 1% D_2O) of Fmoc-Bpl-OH: δ 8.79 (d, *J* = 5.4 Hz, 2H), 8.40-8.31 (m, 2H), 8.01 (2H, *J* = 7.8 Hz, 2H), 7.89-7.79 (m, 4H), 7.72-7.64 (m, 2H), 7.38 (t, *J* = 7.4 Hz, 2H), 7.28 (t, *J* = 7.5 Hz, 2H), 4.43 (s, 2H), 4.29-4.14 (m, 3H), 3.90-3.82 (m, 1H), 2.75 (t, *J* = 8.3 Hz, 2H), 1.65-1.41 (m, 4H), 1.27-1.12 (m, 2H).

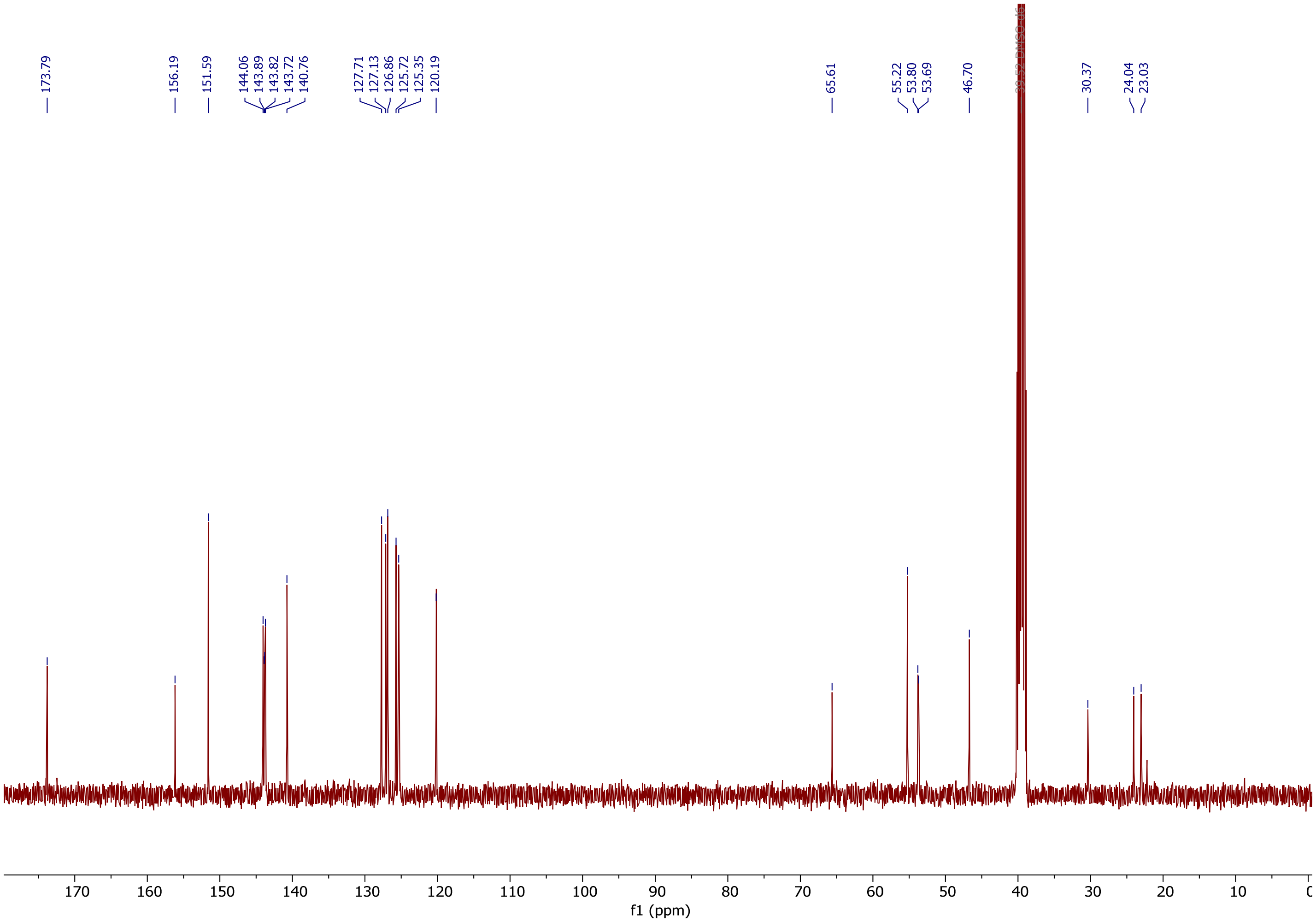
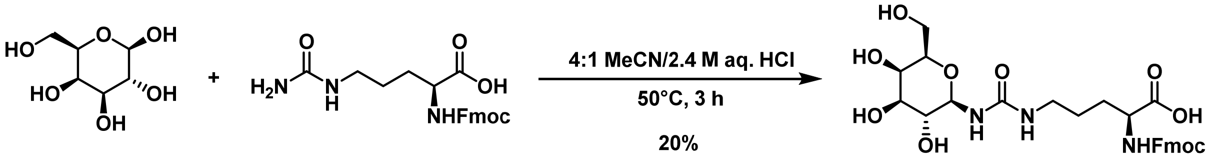


Figure S30. ^{13}C NMR (101 MHz, DMSO-*d*6) of Fmoc-Bpl-OH: δ 173.8, 156.2, 151.6, 144.1, 143.9, 143.7, 140.8, 127.7, 127.1, 126.9, 125.7, 125.4, 120.2, 65.61, 55.2, 53.8, 53.7, 46.7, 30.4, 24.0, 23.0.

21.2 Synthesis of Fmoc-Git-OH

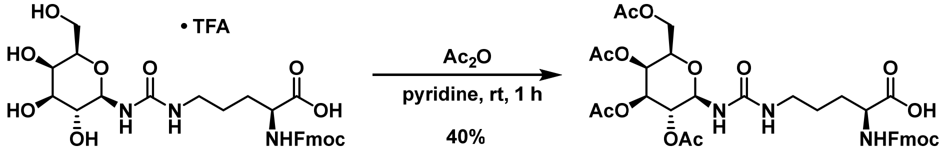
(S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5R,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid (5)



3 4 5

A suspension of D-(+)-galactose (1.50 g, 8.33 mmol, 1.0 eq.) and Fmoc-Cit-OH (4.30 g, 10.8 mmol, 1.30 eq.) in 4:1 MeCN/2.4 M aq. HCl was heated to 50°C for 3 h. The mixture was concentrated and purified by reverse phase column chromatography (Biotage® Sfär C18 Duo 100 Å 30 µm 30 g, MeCN + 0.1% TFA/H_2O + 0.1% TFA = 1:9 → 1:1) to afford the title compound as white solid (1.11 g, 20%) that was used for the next step without further purification.

(S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5S,6R)-3,4,5-triacetoxy-6-(acetoxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid (Fmoc-Git-OH) (6)



5 6

To a solution of (S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5R,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid TFA salt (1.11 g, 1.65 mmol, 1.0 eq.) in pyridine (8.24 mL), acetic anhydride (7.79 mL, 8.41 g, 82.4 mmol, 50 eq.) was added and the resulting solution was stirred at rt for 1 h. After completion of the reaction, the mixture was cooled to 0°C and quenched with 2.4 M aq. HCl. The suspension was diluted with Et_2O (50 mL) and the aqueous phase was extracted with Et_2O (5x). The combined organic layers were dried over anhydrous $MgSO_4$, concentrated under reduced pressure and purified by reverse phase column chromatography (Biotage® Sfär C18 D Duo 100 Å 30 µm 30 g, MeCN + 0.1% HCl : H_2O + 0.1% HCl = 1:19 → 4:1) to yield the title compound as white solid (481 mg, 40%).

ESI-HRMS: calc. $C_{35}H_{42}N_3O_{14}$ $[M+H]^+$ 728.2667 found 728.2666. -0.1 ppm error.

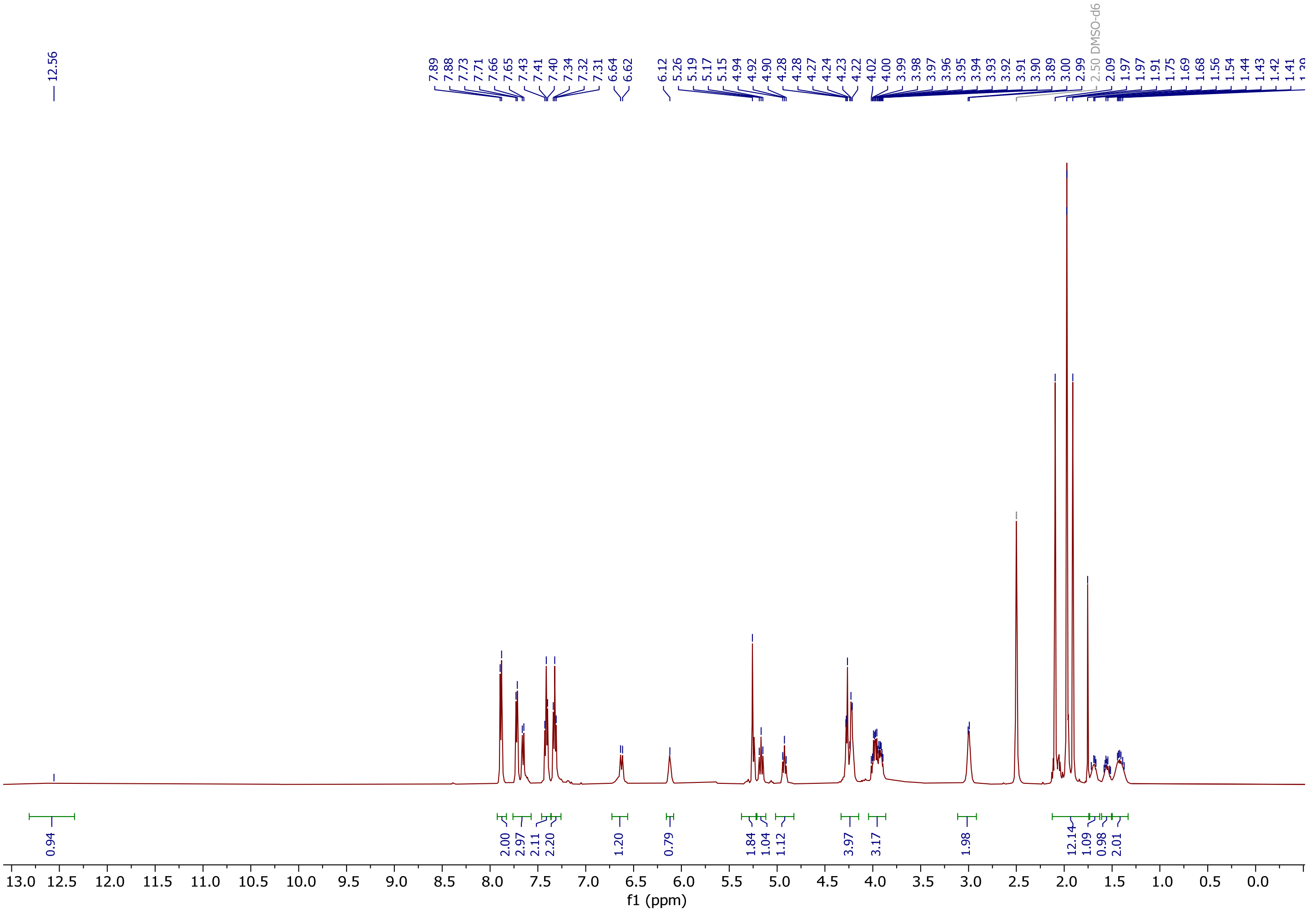


Figure S31. ^{1}H NMR (500 MHz, DMSO-*d*6) of Fmoc-Git-OH: δ 12.56 (br s, 1H), 7.89 (d, *J* = 7.5 Hz, 2H), 7.72 (d, *J* = 7.4 Hz, 2H), 7.65 (d, *J* = 8.0 Hz, 1H), 7.41 (t, *J* = 7.4 Hz, 2H), 7.32 (t, *J* = 7.4 Hz, 2H), 6.63 (d, *J* = 10.2 Hz, 1H), 6.12 (s, 1H), 5.32 – 5.21 (m, 2H), 5.17 (t, *J* = 9.7 Hz, 1H), 4.92 (t, *J* = 9.4 Hz, 1H), 4.32 – 4.15 (m, 4H), 4.08 – 3.82 (m, 3H), 4.08 – 3.82 (m, 2H), 2.09 (s, 3H), 2.02 – 1.92 (m, 6H), 1.91 (s, 3H), 1.75 – 1.63 (m, 1H), 1.62 – 1.49 (m, 1H), 1.47 – 1.39 (m, 2H).

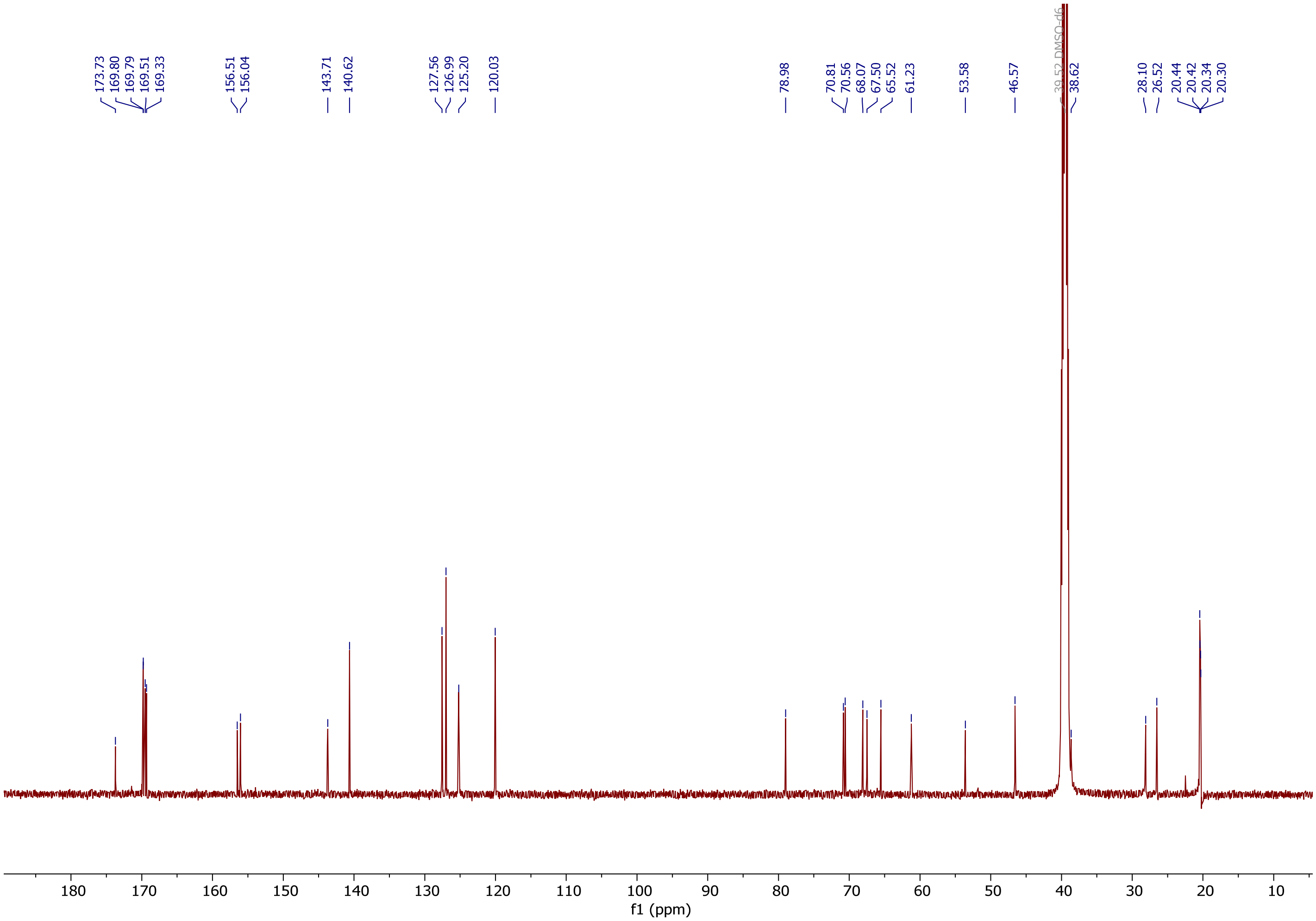
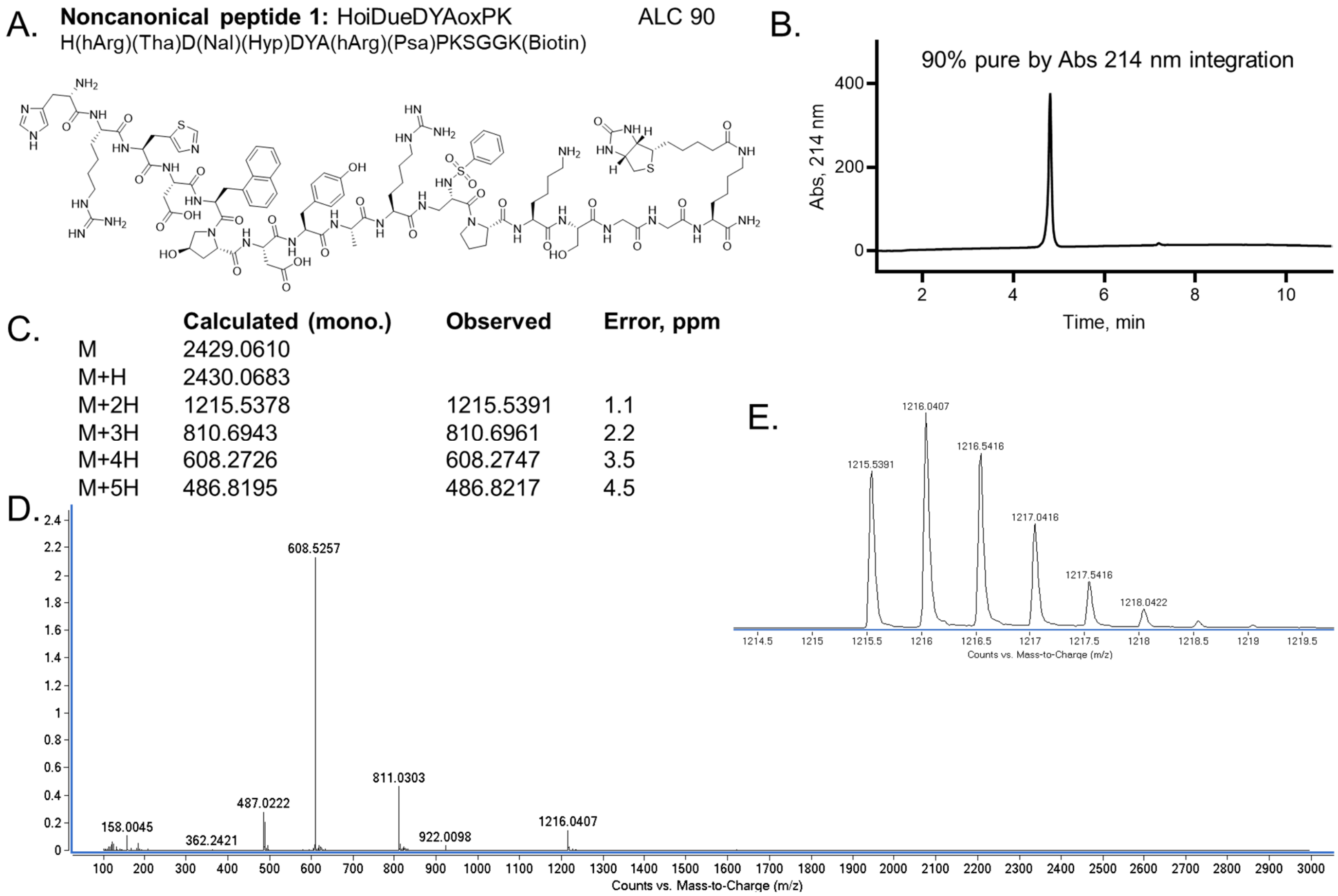


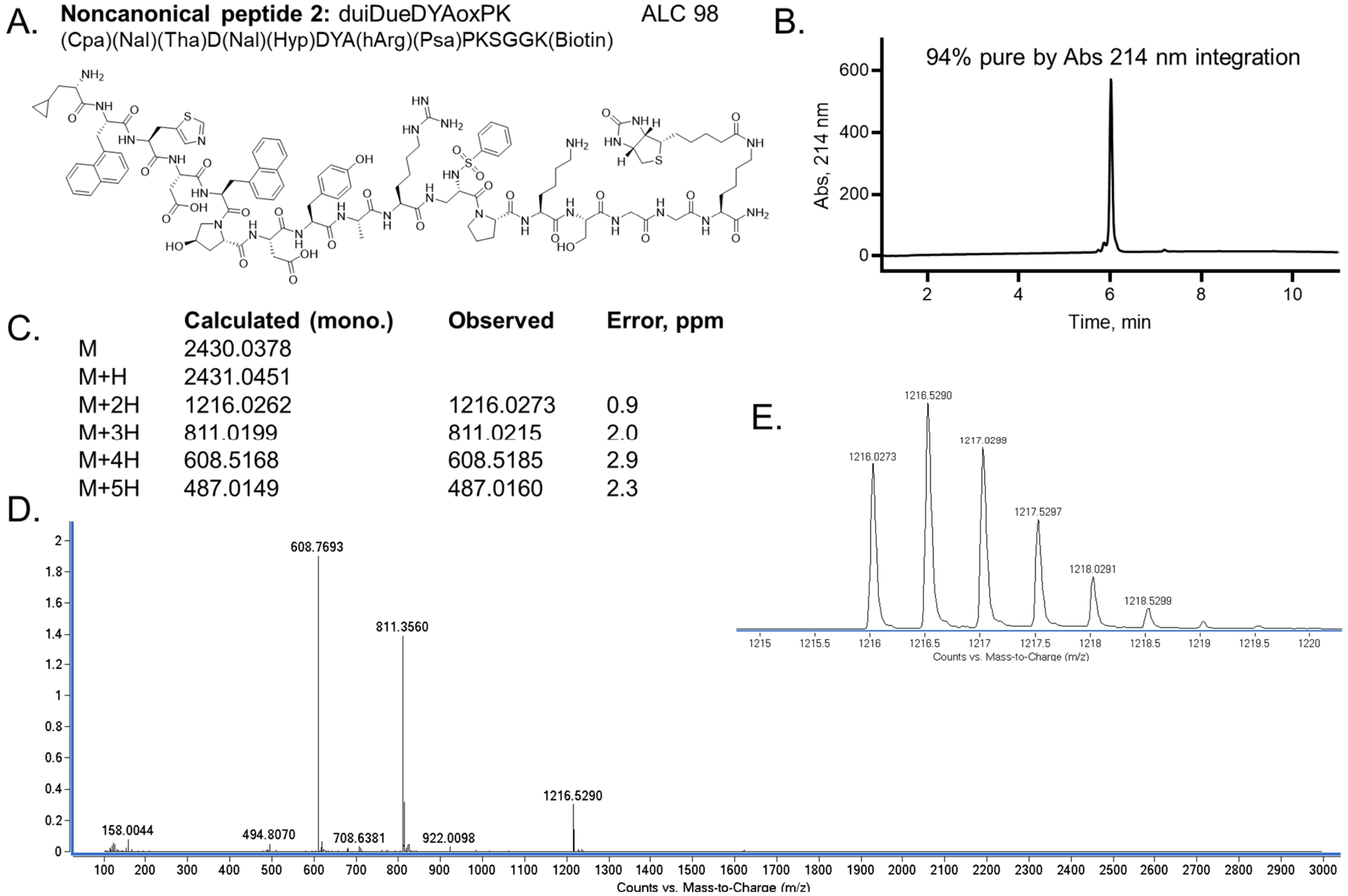
Figure S32. ^{13}C NMR (126 MHz, DMSO-*d*6) of Fmoc-Git-OH: δ 173.7, 169.8, 169.8, 169.5, 169.3, 156.5, 156.0, 143.7, 140.6, 127.6, 127.0, 125.2, 120.0, 79.9, 70.8, 70.6, 68.0, 67.5, 65.5, 61.2, 53.6, 46.6 38.6, 28.1, 26.5, 20.4, 20.4, 20.3, 20.3.

22 Analytical characterization of all synthesized noncanonical peptidomimetics discovered by AS-MS



	Calculated (mono.)	Observed	Error, ppm
M	2429.0610		
M+H	2430.0683		
M+2H	1215.5378	1215.5391	1.1
M+3H	810.6943	810.6961	2.2
M+4H	608.2726	608.2747	3.5
M+5H	486.8195	486.8217	4.5

Figure S33. Analytical characterization of purified Noncanonical **Peptide 1**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

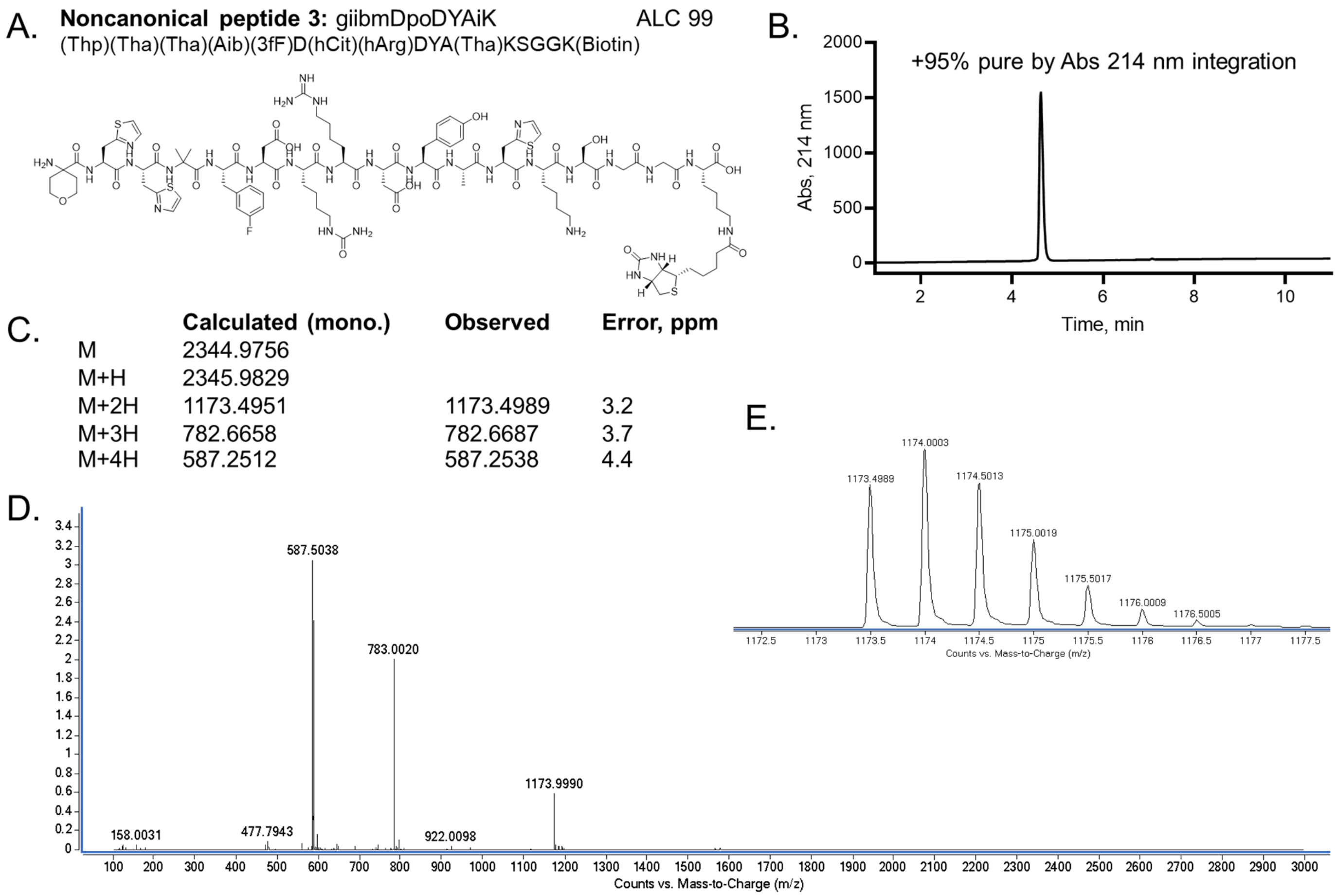


Noncanonical peptide 2: duiDueDYAoxPK ALC 98

(Cpa)(Nal)(Tha)D(Nal)(Hyp)DYA(hArg)(Psa)PKSGGK(Biotin)

	Calculated (mono.)	Observed	Error, ppm
M	2430.0378		
M+H	2431.0451		
M+2H	1216.0262	1216.0273	0.9
M+3H	811.0199	811.0215	2.0
M+4H	608.5168	608.5185	2.9
M+5H	487.0149	487.0160	2.3

Figure S34. Analytical characterization of purified Noncanonical **Peptide 2**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

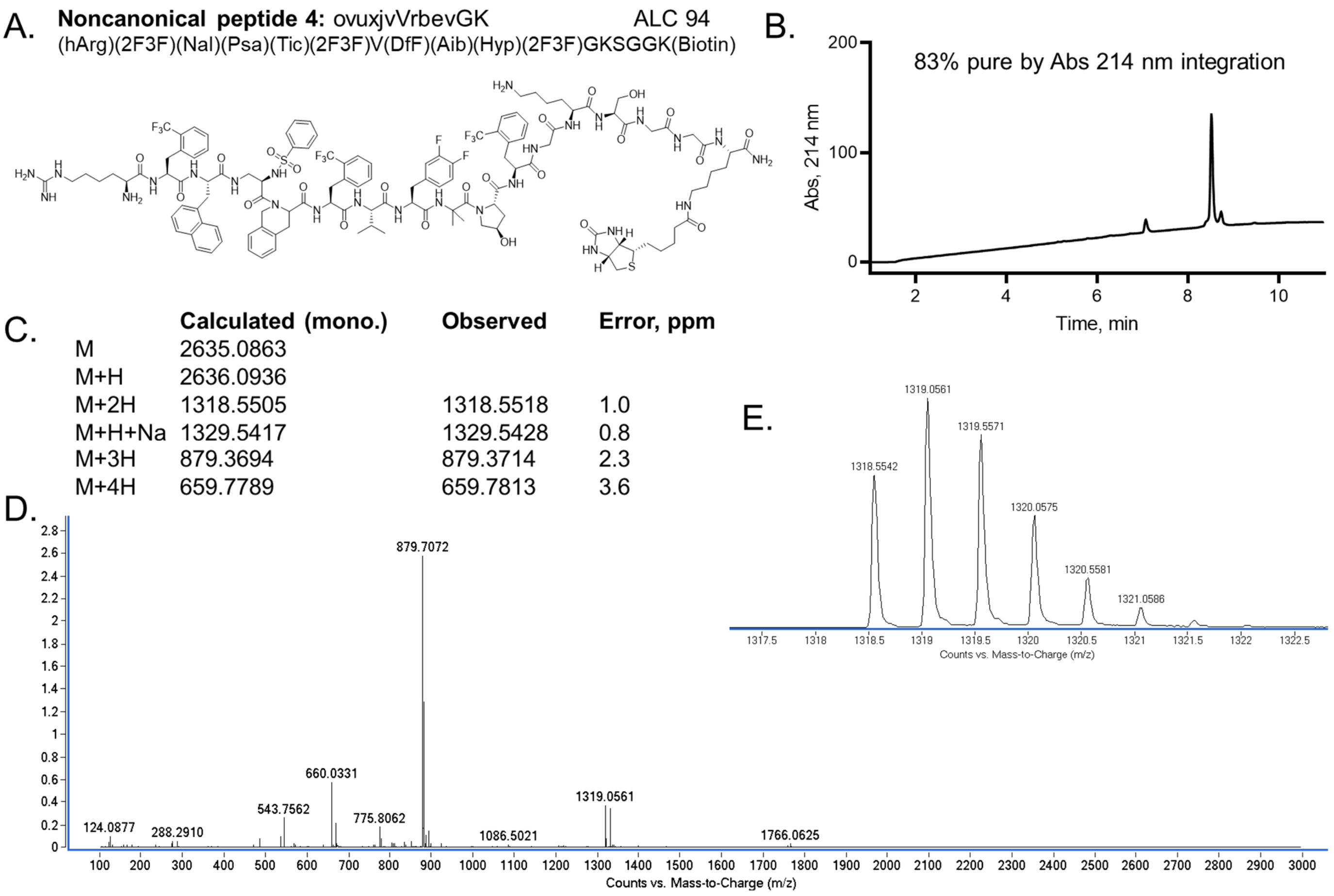


A. **Noncanonical peptide 3:** giibmDpoDYAiK ALC 99
(Thp)(Tha)(Tha)(Aib)(3fF)D(hCit)(hArg)DYA(Tha)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2344.9756		
M+H	2345.9829		
M+2H	1173.4951	1173.4989	3.2
M+3H	782.6658	782.6687	3.7
M+4H	587.2512	587.2538	4.4

Figure S35. Analytical characterization of purified Noncanonical **Peptide 3**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



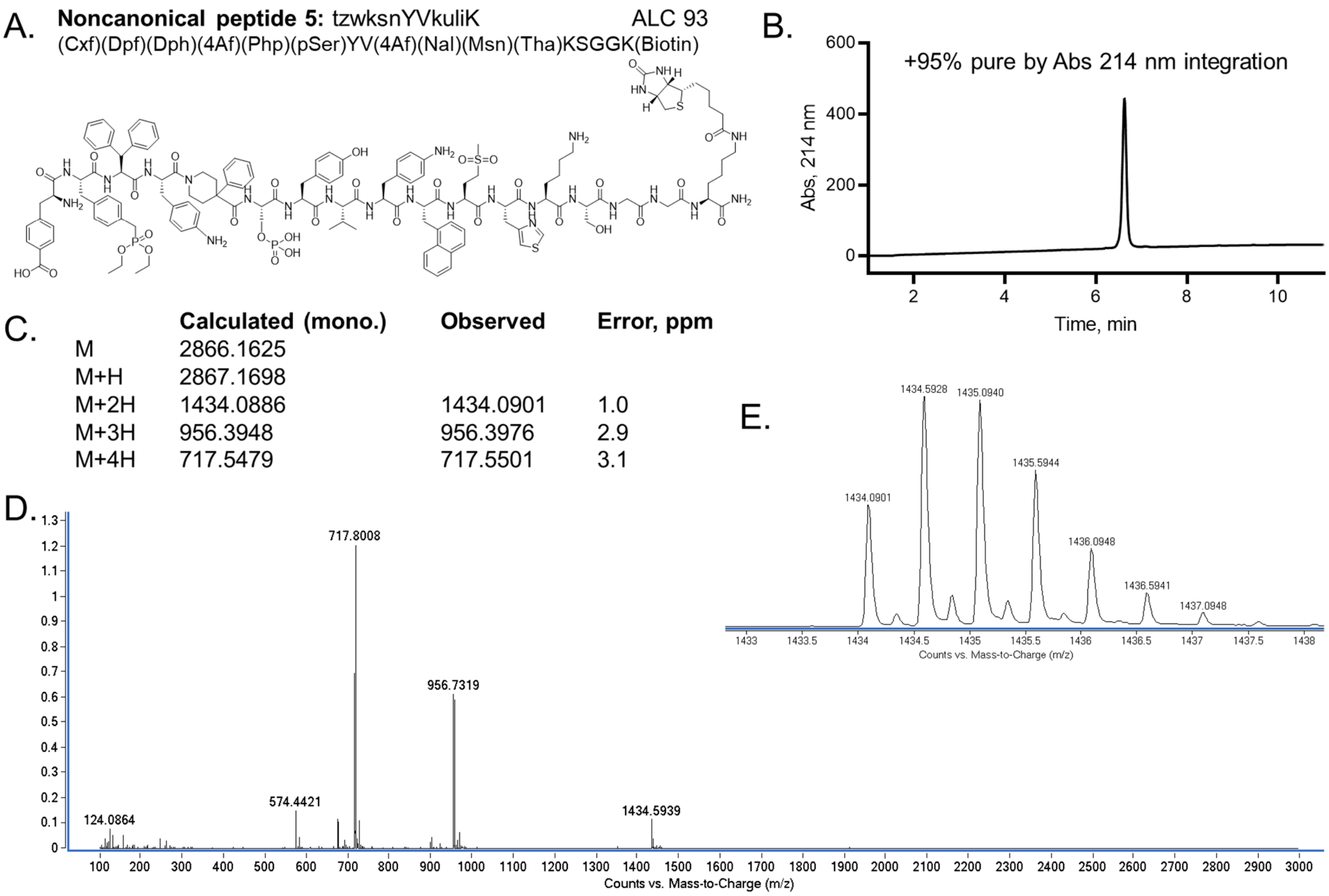
A. **Noncanonical peptide 4:** ovuxjvVrbevGK ALC 94

(hArg)(2F3F)(Nal)(Psa)(Tic)(2F3F)V(DfF)(Aib)(Hyp)(2F3F)GKSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2635.0863		
M+H	2636.0936		
M+2H	1318.5505	1318.5518	1.0
M+H+Na	1329.5417	1329.5428	0.8
M+3H	879.3694	879.3714	2.3
M+4H	659.7789	659.7813	3.6

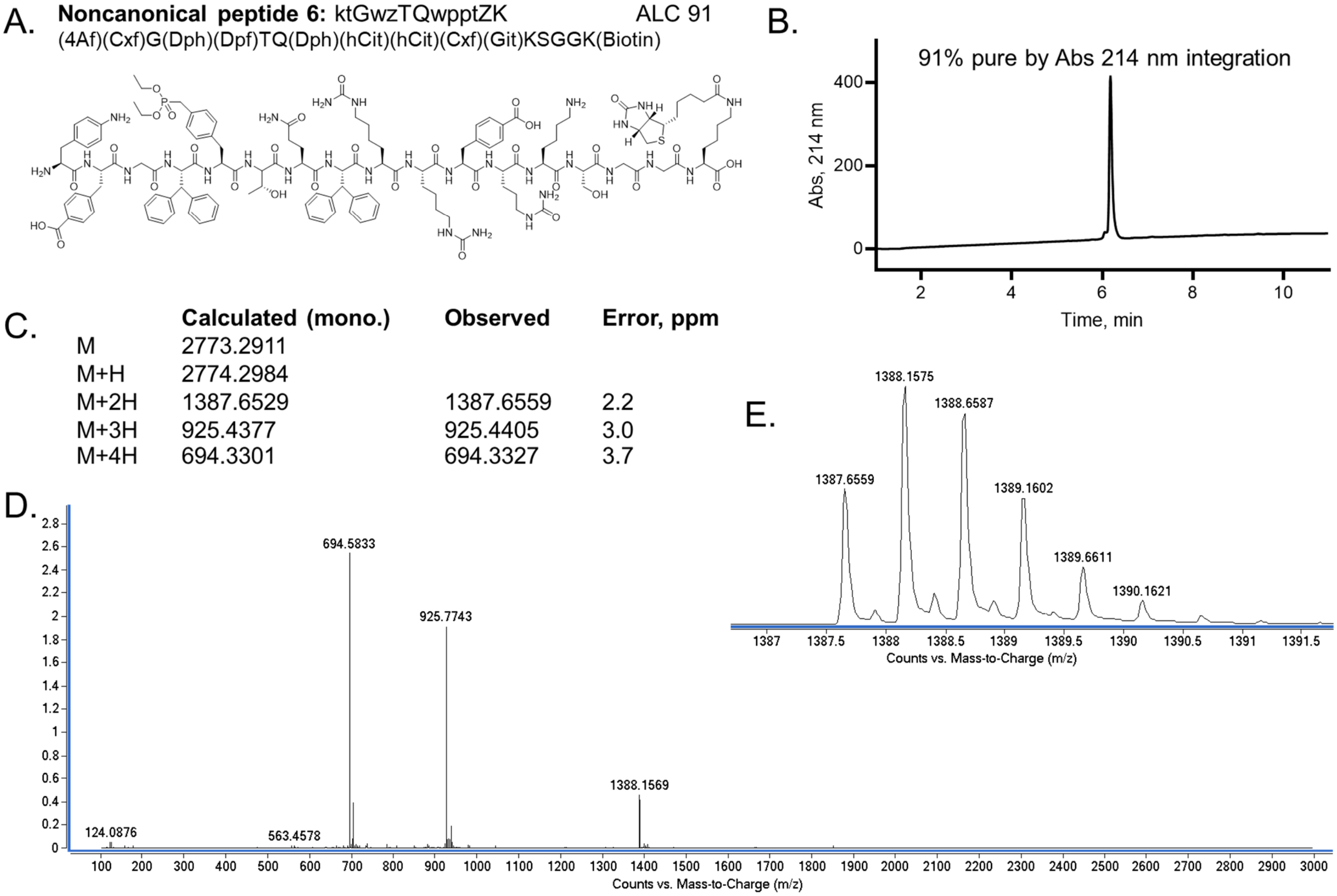
Figure S36. Analytical characterization of purified Noncanonical **Peptide 4**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



	Calculated (mono.)	Observed	Error, ppm
M	2866.1625		
M+H	2867.1698		
M+2H	1434.0886	1434.0901	1.0
M+3H	956.3948	956.3976	2.9
M+4H	717.5479	717.5501	3.1

5

Figure S37. Analytical characterization of purified Noncanonical **Peptide 5**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



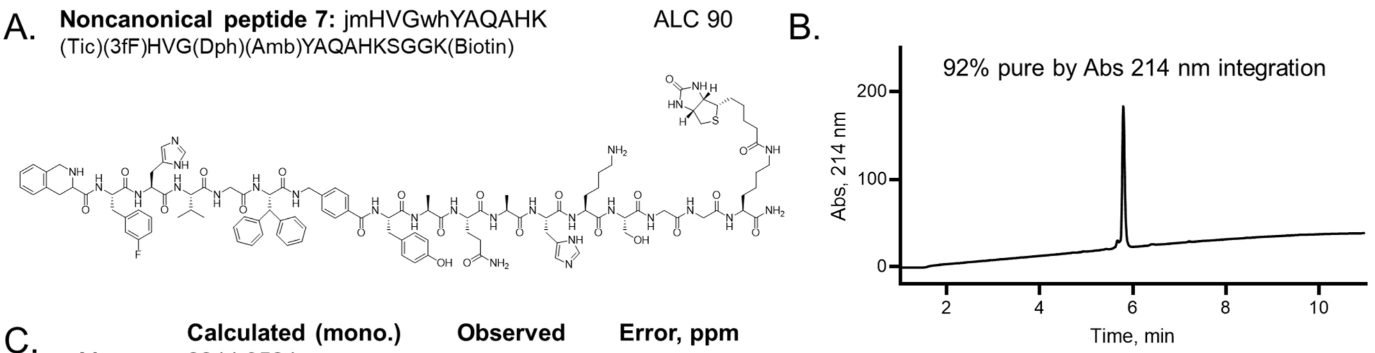
A. **Noncanonical peptide 6:** ktGwzTQwpptZK ALC 91

(4Af)(Cxf)G(Dph)(Dpf)TQ(Dph)(hCit)(hCit)(Cxf)(Git)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2773.2911		
M+H	2774.2984		
M+2H	1387.6529	1387.6559	2.2
M+3H	925.4377	925.4405	3.0
M+4H	694.3301	694.3327	3.7

Figure S38. Analytical characterization of purified Noncanonical **Peptide 6**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).


A.
Noncanonical peptide 7: jmHVGwhYAQAHK
ALC 90
(Tic)(3fF)HVG(Dph)(Amb)YAQAHKSGGK(Biotin)
B.
92% pure by Abs 214 nm integration
Abs, 214 nm
200
100
0
2
4
6
8
10
Time, min

C.

	Calculated (mono.)	Observed	Error, ppm
M	2244.0524		
M+H	2245.0600		
M+2H	1123.0337	1123.0343	0.5
M+3H	749.0249	749.0289	5.3
M+4H	562.0205	562.0252	8.4

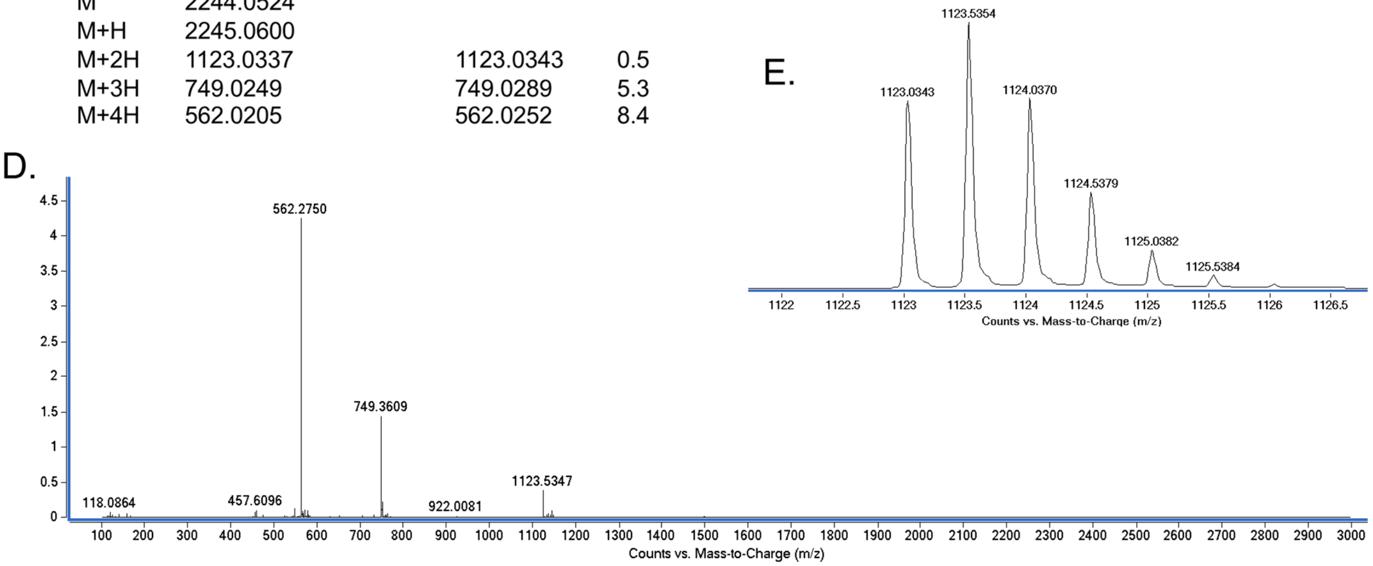
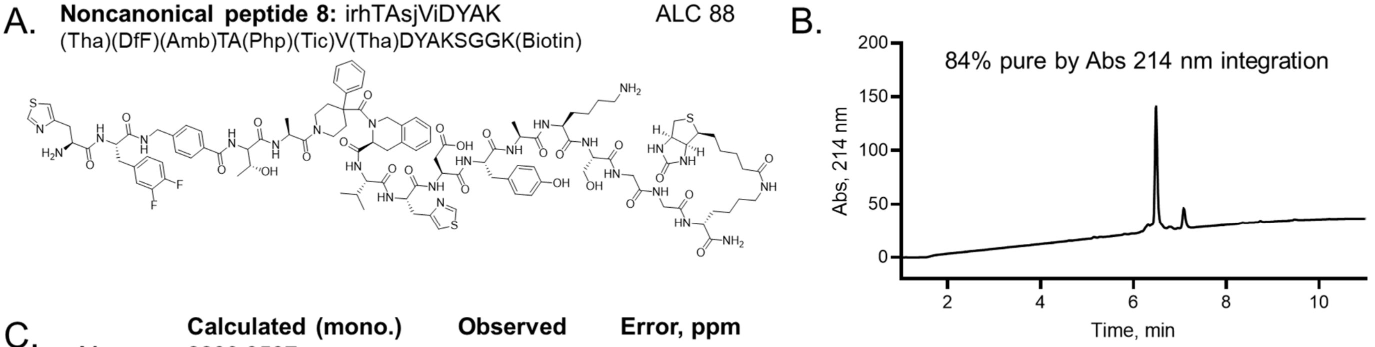


Figure S39. Analytical characterization of purified Noncanonical **Peptide 7**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

A. **Noncanonical peptide 8:** irhTAsjViDYAK ALC 88

(Tha)(DfF)(Amb)TA(Php)(Tic)V(Tha)DYAKSGGK(Biotin)



C.

	Calculated (mono.)	Observed	Error, ppm
M	2290.9597		
M+H	2291.9670		
M+2H	1146.4872	1146.4891	1.7
M+3H	764.6605	764.6623	2.4
M+4H	573.7472	573.7491	3.3

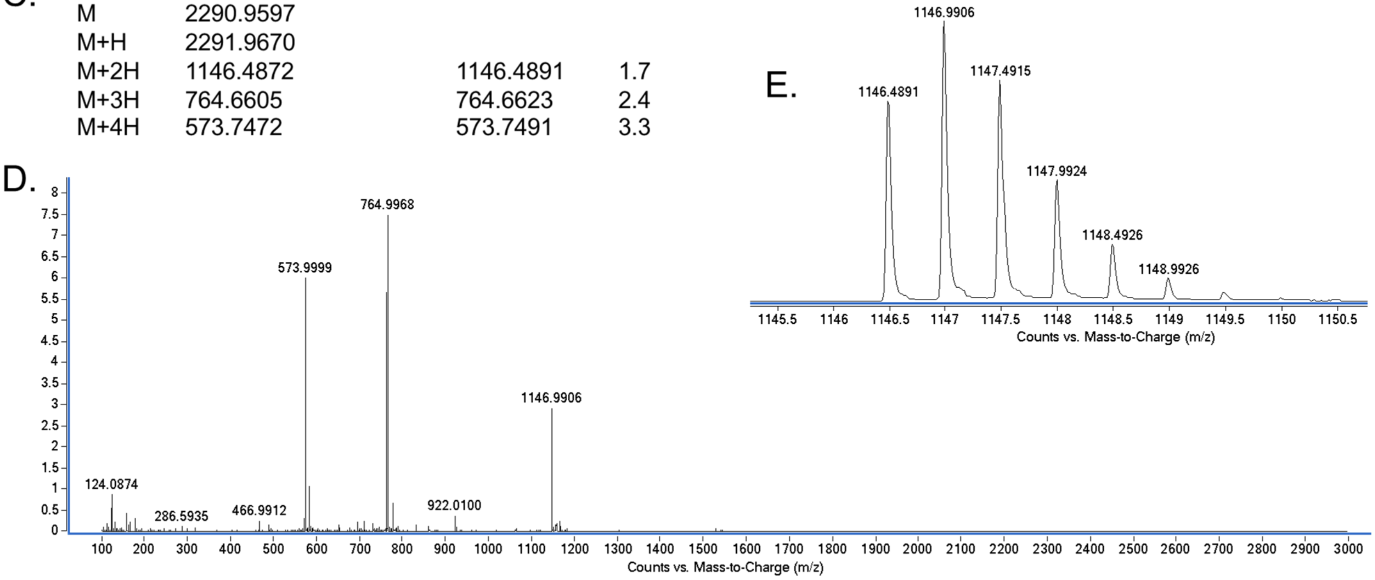
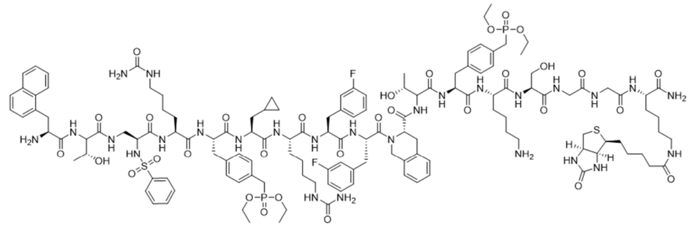
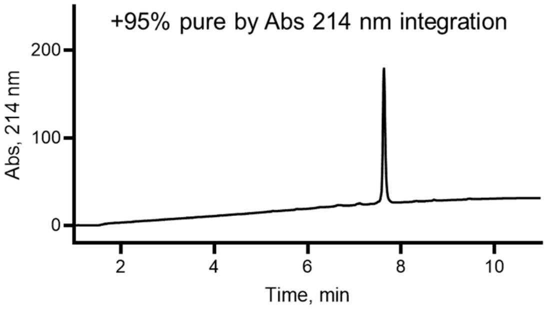


Figure S40. Analytical characterization of purified Noncanonical **Peptide 8**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

A. **Noncanonical peptide 9:** uTxpzdpmmjTzK ALC 87
(Nal)T(Psa)(hCit)(Dpf)(Cpa)(hCit)(3fF)(3fF)(Tic)T(Dpf)KSGGK(Biotin)



B.



C.

	Calculated (mono.)	Observed	Error, ppm
M	2862.2716		
M+H	2863.2789		
M+2H	1432.1431	1432.1469	2.7
M+3H	955.0978	955.1005	2.8
M+4H	716.5752	716.5771	2.7

D. E.

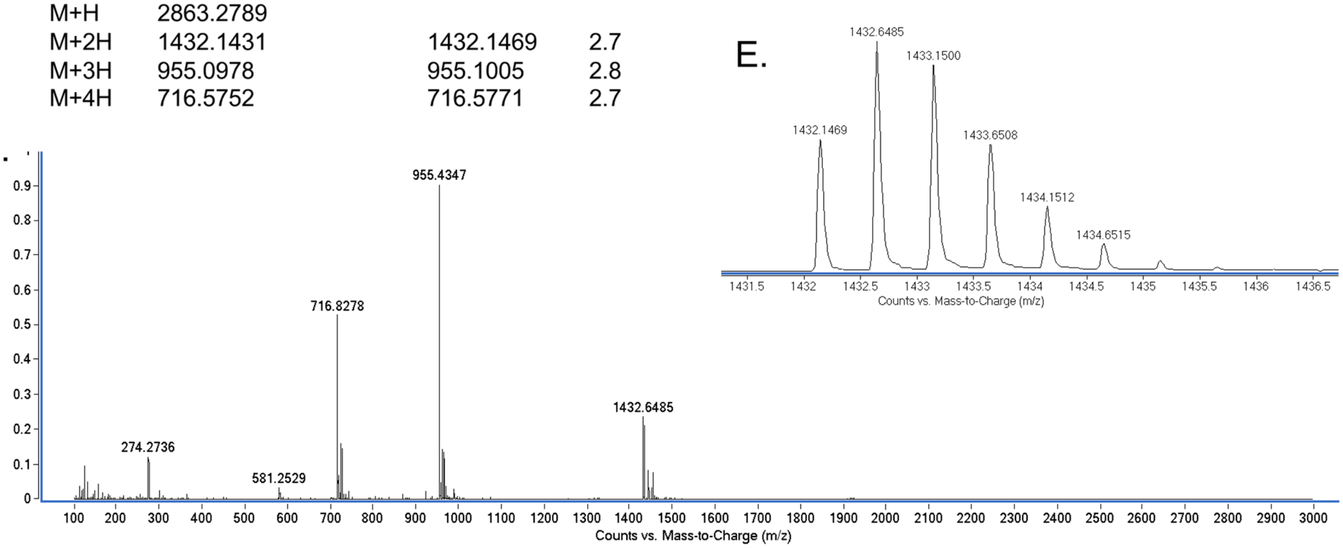
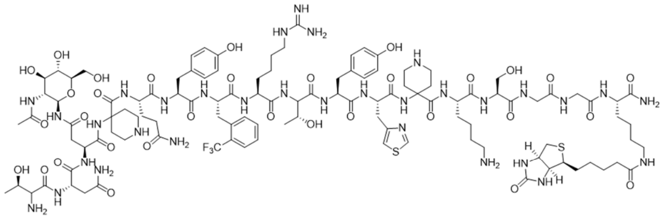


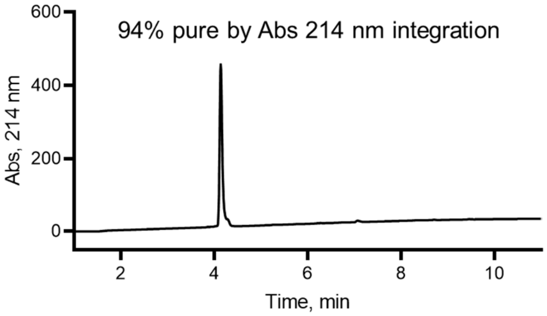
Figure S41. Analytical characterization of purified Noncanonical **Peptide 9**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

A. **Noncanonical peptide 10:** TNXfQYvoTYifK ALC 84

TN(Agn)(Pip)QY(2F3F)(hArg)TY(Tha)(Pip)KSGGK(Biotin)



B.



C.

	Calculated (mono.)	Observed	Error, ppm
M	2376.0862		
M+H	2377.0935		
M+2H	1189.0504	1189.0526	1.9
M+H+Na	1200.0417	1200.0445	2.3
M+3H	793.0360	793.0378	2.3
M+4H	595.0289	595.0299	1.7

D. E.

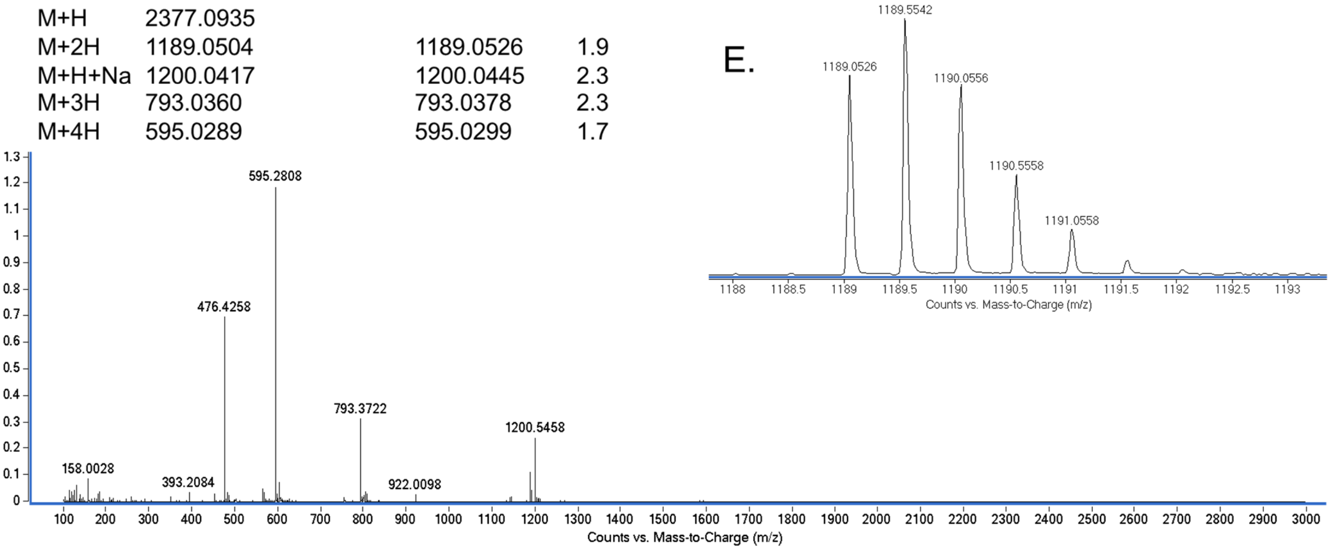
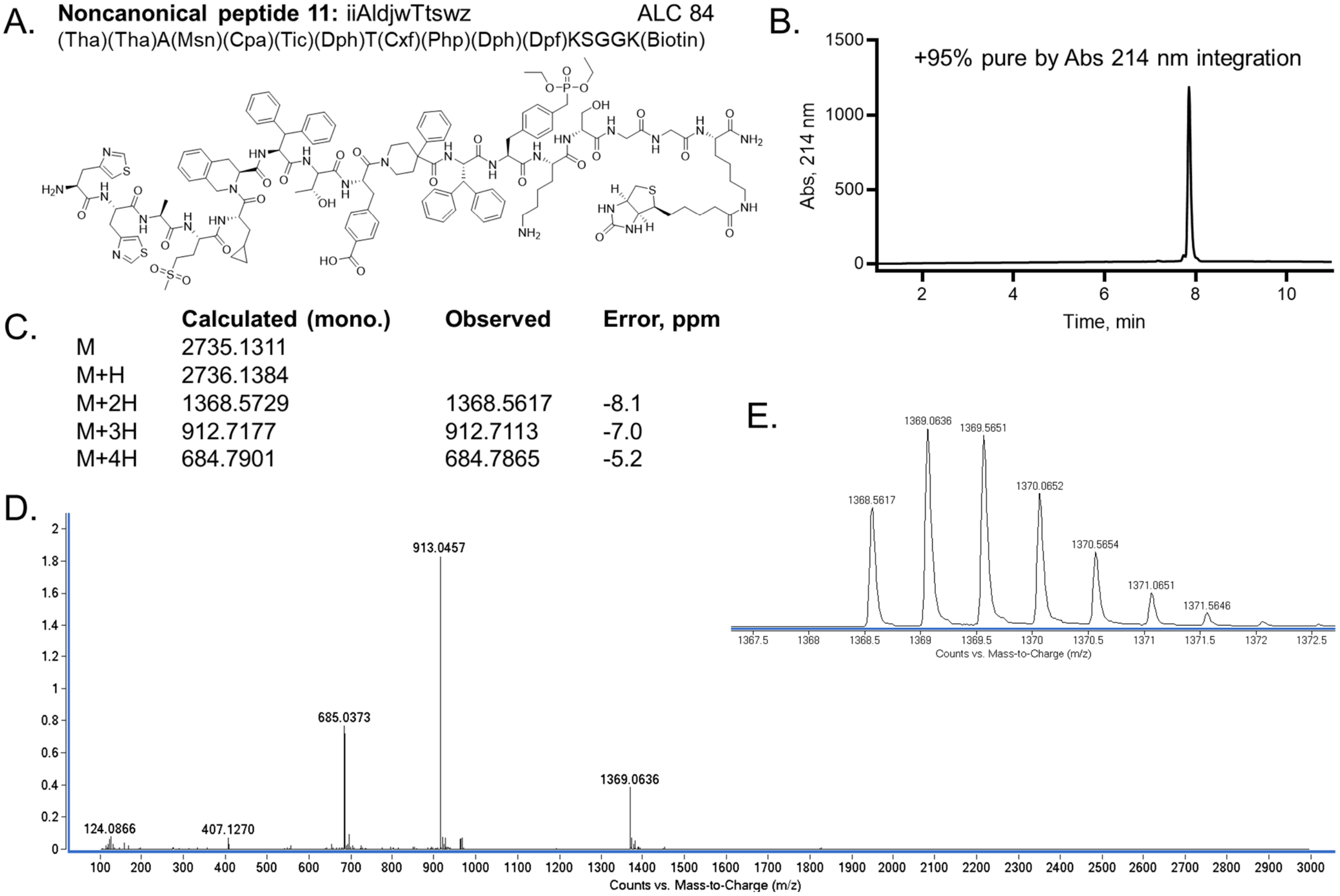


Figure S42. Analytical characterization of purified Noncanonical **Peptide 10**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



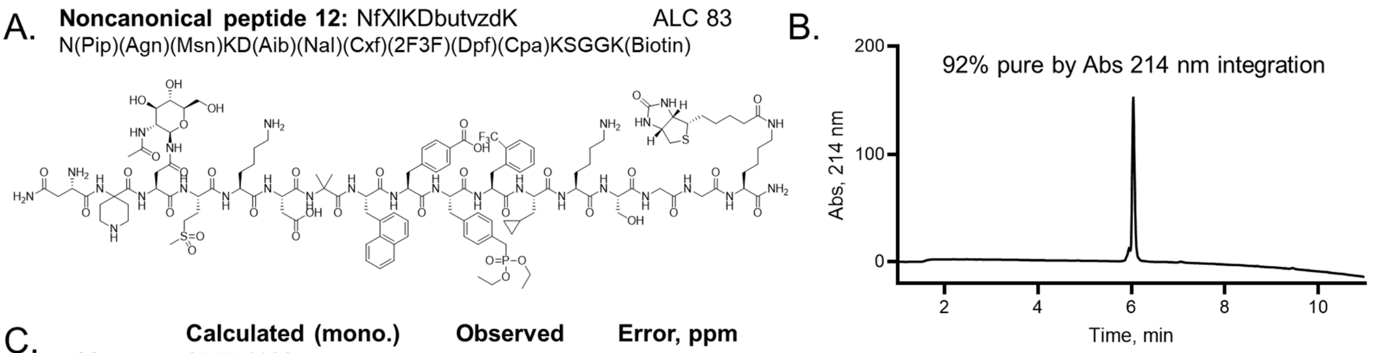
A. **Noncanonical peptide 11:** iiAldjwTtswz ALC 84

(Tha)(Tha)A(Msn)(Cpa)(Tic)(Dph)T(Cxf)(Php)(Dph)(Dpf)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2735.1311		
M+H	2736.1384		
M+2H	1368.5729	1368.5617	-8.1
M+3H	912.7177	912.7113	-7.0
M+4H	684.7901	684.7865	-5.2

Figure S43. Analytical characterization of purified Noncanonical **Peptide 11**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).


A.
Noncanonical peptide 12: NfXlKDbutvzdK
ALC 83
N(Pip)(Agn)(Msn)KD(Aib)(Nal)(Cxf)(2F3F)(Dpf)(Cpa)KSGGK(Biotin)
B.
92% pure by Abs 214 nm integration
Abs, 214 nm
200
100
0
2
4
6
8
10
Time, min

C.

	Calculated (mono.)	Observed	Error, ppm
M	2557.1183		
M+H	2558.1256		
M+2H	1279.5665	1279.5668	0.1
M+3H	853.3801	853.3816	1.8
M+4H	640.2869	640.2887	2.8

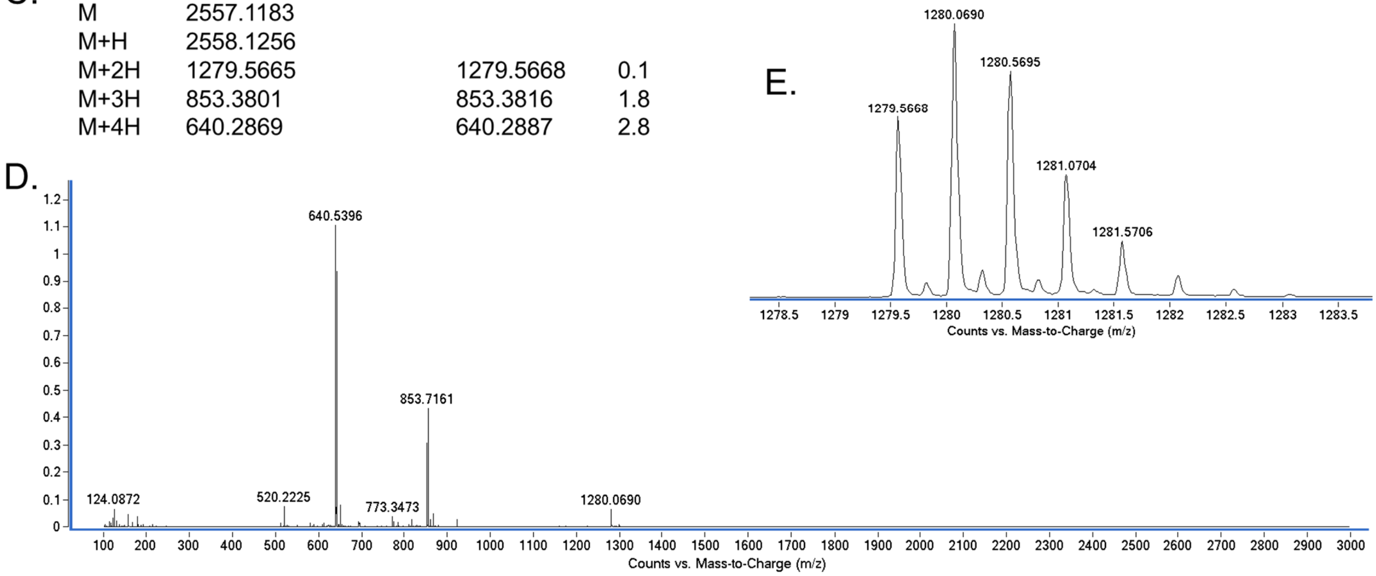
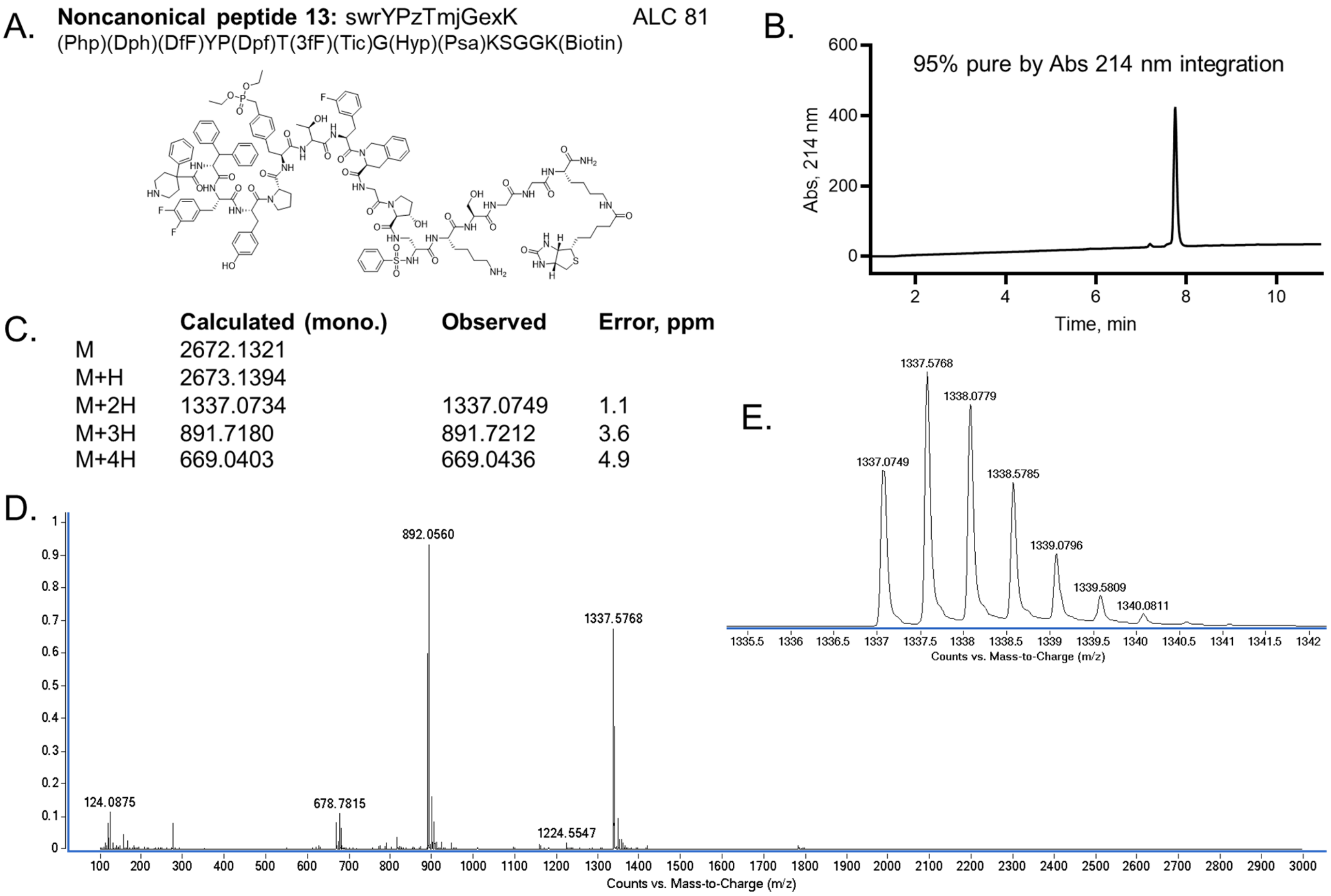


Figure S44. Analytical characterization of purified Noncanonical **Peptide 12**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



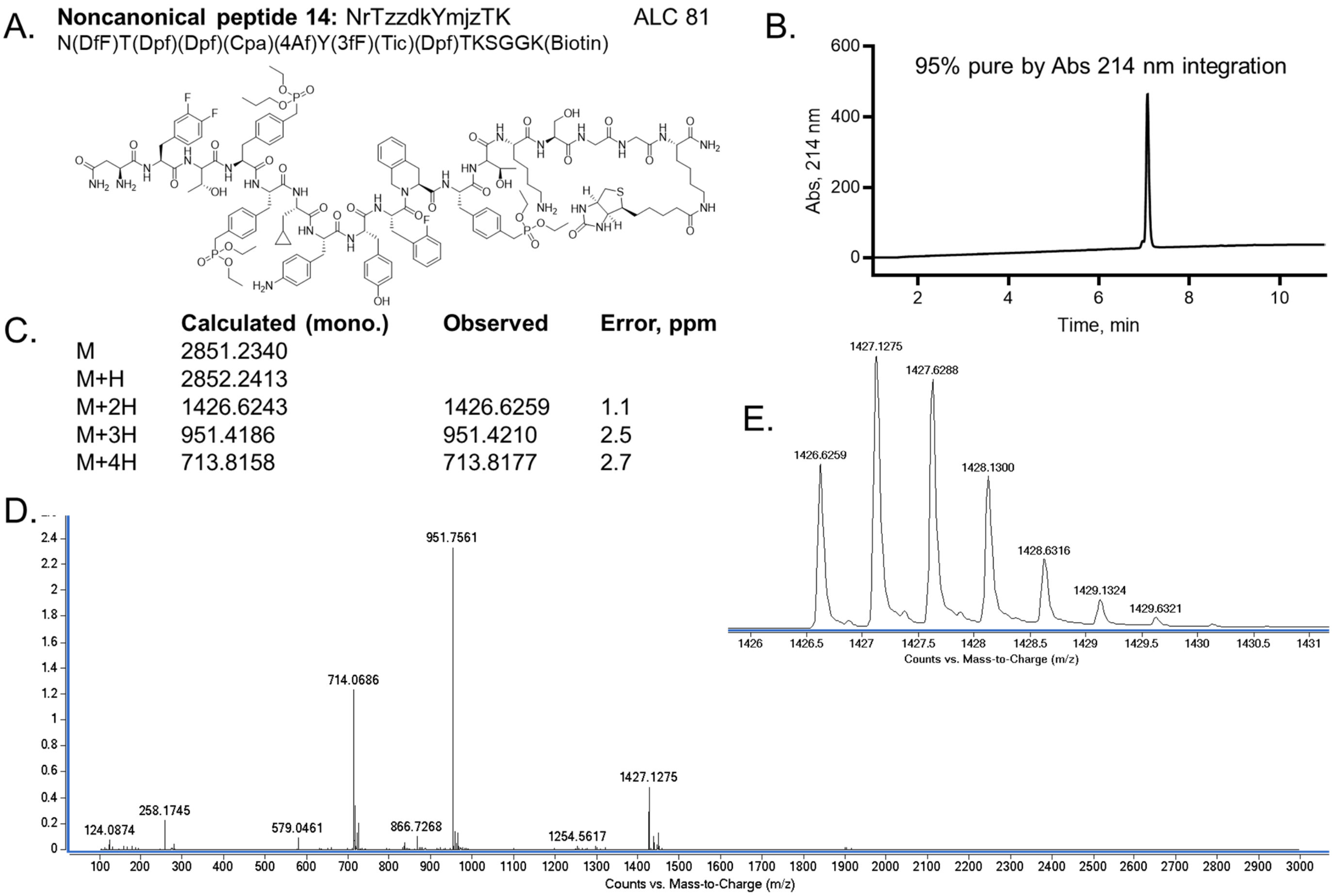
A. **Noncanonical peptide 13:** swrYPzTmjGexK ALC 81

(Php)(Dph)(DfF)YP(Dpf)T(3fF)(Tic)G(Hyp)(Psa)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2672.1321		
M+H	2673.1394		
M+2H	1337.0734	1337.0749	1.1
M+3H	891.7180	891.7212	3.6
M+4H	669.0403	669.0436	4.9

Figure S45. Analytical characterization of purified Noncanonical **Peptide 13**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



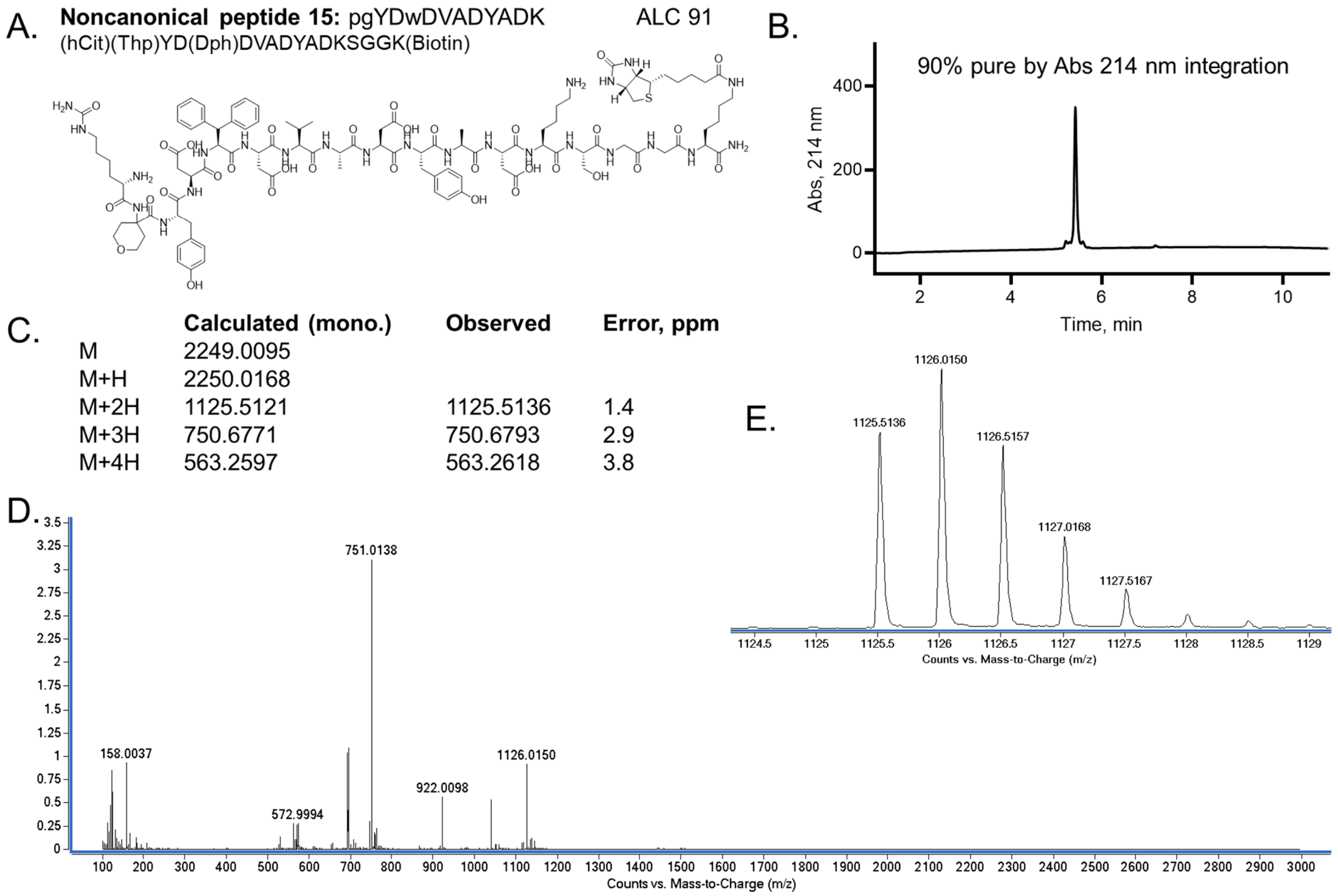
A. **Noncanonical peptide 14:** NrTzzdkYmjzTK ALC 81

N(DfF)T(Dpf)(Dpf)(Cpa)(4Af)Y(3fF)(Tic)(Dpf)TKSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2851.2340		
M+H	2852.2413		
M+2H	1426.6243	1426.6259	1.1
M+3H	951.4186	951.4210	2.5
M+4H	713.8158	713.8177	2.7

Figure S46. Analytical characterization of purified Noncanonical **Peptide 14**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



Noncanonical peptide 15: pgYDwDVADYADK ALC 91
(hCit)(Thp)YD(Dph)DVADYADKSGGK(Biotin)

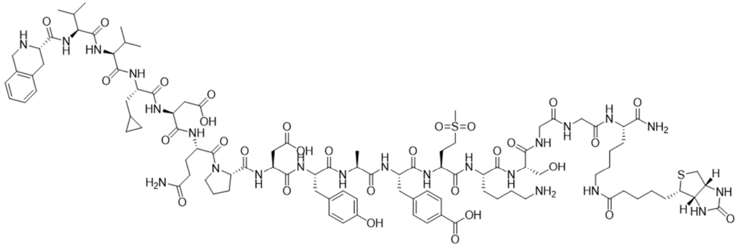
C.

	Calculated (mono.)	Observed	Error, ppm
M	2249.0095		
M+H	2250.0168		
M+2H	1125.5121	1125.5136	1.4
M+3H	750.6771	750.6793	2.9
M+4H	563.2597	563.2618	3.8

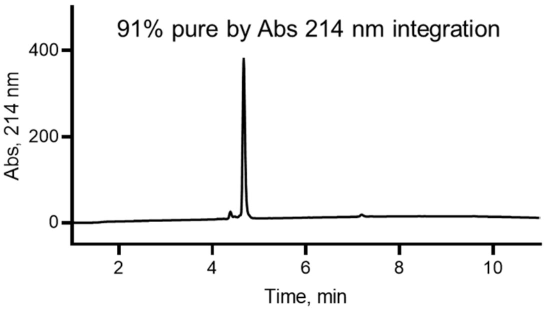
Figure S47. Analytical characterization of purified Noncanonical **Peptide 15**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

A. **Noncanonical peptide 16:** jVVdDQPDYAtlK ALC 99

(Tic)VV(Cpa)DQPDYA(Cxf)(Msn)KSGGK(Biotin)



B.



C.

	Calculated (mono.)	Observed	Error, ppm
M	2211.9965		
M+H	2213.0038		
M+2H	1107.0056	1107.0079	2.1
M+3H	738.3395	738.3416	2.9
M+4H	554.0064	554.0073	1.6

D.

E.

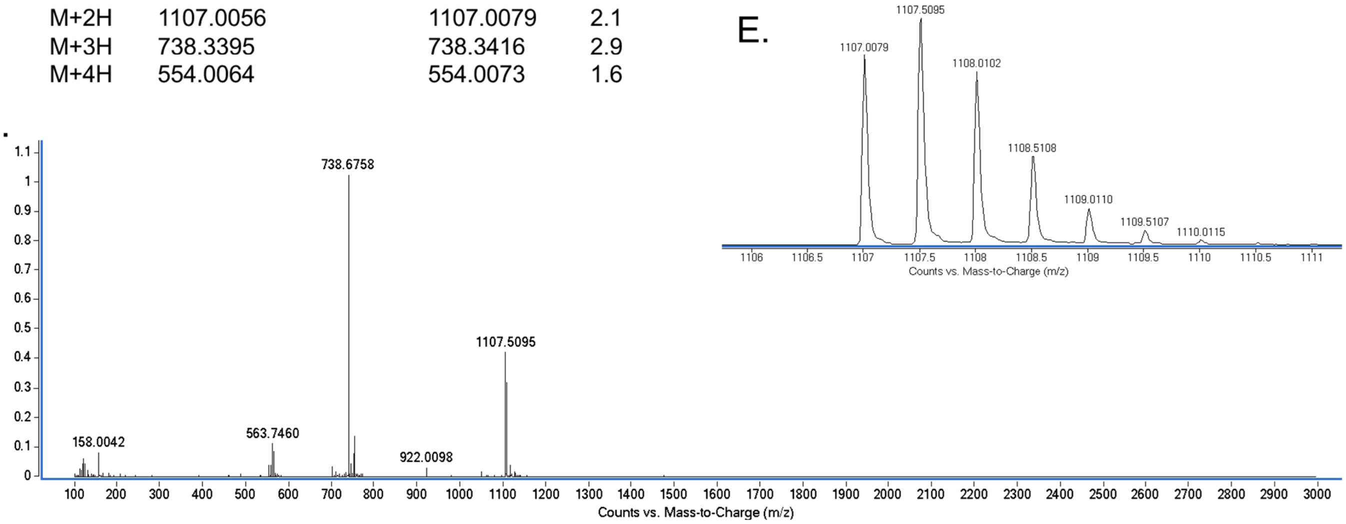
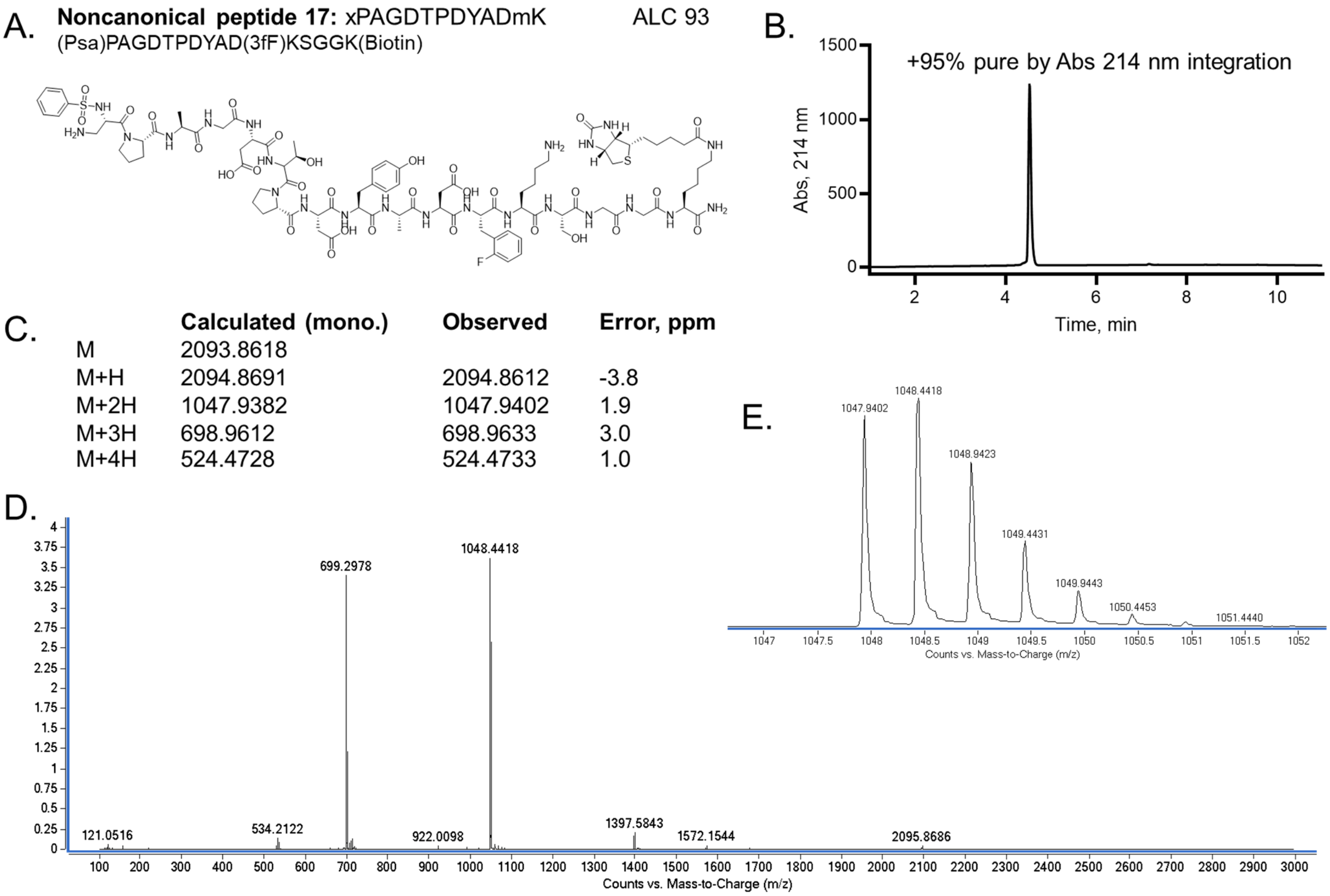


Figure S48. Analytical characterization of purified Noncanonical **Peptide 16**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).



A. **Noncanonical peptide 17:** xPAGDTPDYADmK ALC 93
(Psa)PAGDTPDYAD(3fF)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2093.8618		
M+H	2094.8691	2094.8612	-3.8
M+2H	1047.9382	1047.9402	1.9
M+3H	698.9612	698.9633	3.0
M+4H	524.4728	524.4733	1.0

Figure S49. Analytical characterization of purified Noncanonical **Peptide 17**. **A.** Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. **B.** Purity and UPLC chromatogram **C.** Calculated and observed monoisotopic masses with ppm error reported. **D.** Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and **E.** ~5 m/z zoom in on the lowest charge species observed (often z = 2).

23 References

1. Quartararo, A. J. *et al.* Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat Commun* **11**, 3183 (2020).
2. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols 2007 2:8* **2**, 1896–1906 (2007).
3. Vinogradov, A. A. *et al.* Library Design-Facilitated High-Throughput Sequencing of Synthetic Peptide Libraries. *ACS Comb Sci* **19**, 694–701 (2017).
4. Chen, K. H. & Hu, Y. J. Residue–Residue Interaction Prediction via Stacked Meta-Learning. *International Journal of Molecular Sciences 2021, Vol. 22, Page 6393* **22**, 6393 (2021).
5. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for Fingerprint-based similarity calculations? *J Cheminform* **7**, 1–13 (2015).
6. Cox, M. A. A. & Cox, T. F. Multidimensional Scaling. in *Handbook of Data Visualization* vol. 4 315–347 (Springer Berlin Heidelberg, 2008).
7. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science (1979)* **255**, 959–965 (1992).
8. Pinilla, C., Appel, J. R. & Houghten, R. A. Investigation of antigen-antibody interactions using a soluble, non-support-bound synthetic decapeptide library composed of four trillion (4 × 1012) sequences. *Biochemical Journal* **301**, 847–853 (1994).
9. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
10. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
11. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13 (2009).
12. Bailey, T. L. Discovering Sequence Motifs. in *Comparative Genomics. Methods in Molecular Biology* (ed. Bergman, N. H.) vol. 395 231–251 (Humana Press, 2007).
13. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).
14. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834–2840 (2021).