Electronic Supplementary Information (ESI).

# Design of antimicrobial peptides containing non-proteinogenic amino acids using multi-objective Bayesian optimisation

Yuki Murakami [a], Shoichi Ishida [a], Yosuke Demizu [a, b] and Kei Terayama [a, c, d] *

[a]Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Yokohama, Kanagawa 230-0045, Japan

[b]Division of Organic Chemistry, National Institute of Health Sciences, 3-25-26, Tonomachi, Kawasaki, Kanagawa 210-9501, Japan

[c]RIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[d]MDX Research Center for Element Strategy, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, 226-8501, Japan.

GitHub repository: https://github.com/ycu-iil/MODAN

* Correspondence: terayama@yokohama-cu.ac.jp

**Contents**

S1. Construction of surrogate models

S1-1. String representation of the natural and non-proteogenic amino acids and side-chain stapling

MODAN can handle various amino acids by registering codes, typically one or two letters, for the target amino acids and the SMILES strings on the side chains of the amino acids. Table S1 shows the names and codes of the prepared amino acids, including the NPAAs and side-chain staples used in this study. The standard code was used for the natural amino acids. (*S*)-2-(4-Pentenyl)alanine (S5) and (*R*)-2-(7-Octenyl)alanine (R8) were registered as the amino acids for the building blocks of side-chain stapling. Each peptide is represented by a sequence of the defined codes.

**Table S1** List of the amino acids and their codes used in this study, including NPAAs and the amino acids used as the building blocks for side-chain stapling.

| Natural amino acid | | Non-proteinogenic amino acid | |
|---|---|---|---|
| Name | Code | Name | Code |
| Alanine | A | Ornithine | O |
| Cysteine | C | L-homoserine-(O-allyl) | X0 |
| Aspartic acid | D | Diaminobutyric acid | J |
| Glutamic acid | E | Sarcosine | X2 |
| Phenylalanine | F | 1-Aminocyclopentanecarboxylic acid | B |
| Glycine | G | α-Aminoisobutyric acid | U |
| Histidine | H | Aminocyclohexyl carboxylic acid | Z |
| Isoleucine | I | | |
| Lysine | K | Amino acid for the building blocks of side-chain stapling | |
| Leucine | L | Name | Code |
| Methionine | M | (*S*)-2-(4-Pentenyl)alanine | S5 |
| Asparagine | N | (*R*)-2-(7-Octenyl)alanine | R8 |
| Proline | P | | |
| Glutamine | Q | | |
| Arginine | R | | |
| Serine | S | | |
| Threonine | T | | |
| Valine | V | | |
| Tryptophane | W | | |
| Tyrosine | Y | | |

S1-2. Preparation of the molecular fingerprints as the input features for the surrogate models

In this study, the peptide sequences were converted into two types of 2D molecular representations using SMILES: standard and Skip-7. For the standard molecular representation, peptide sequences consisting of amino acids codes in Table S1 were converted to SMILES strings, as shown in Fig. S1. For side-chain stapling, when S5 or R8 appears in a sequence, S5 is inserted four or seven residues apart, respectively. The two building blocks were then connected to form a side-chain staple. For the Skip-7 representation, a peptide sequence is converted to the standard SMILES representation, and then its α carbons are connected to 7 residues apart to take into account the effect of the alpha-helical structure. Here, the α carbons were converted to phosphorus or sulfur.

Three types of molecular fingerprints were calculated for the two-molecular representations. The generated SMILES strings were converted into MACCS keys and two types of count-based on the Morgan fingerprints. The MACCS key has 166 bits, which represent key structural descriptors. The maximal radii of two and four were used to generate two types of count-based on the Morgan fingerprints under 1024 dimensions.

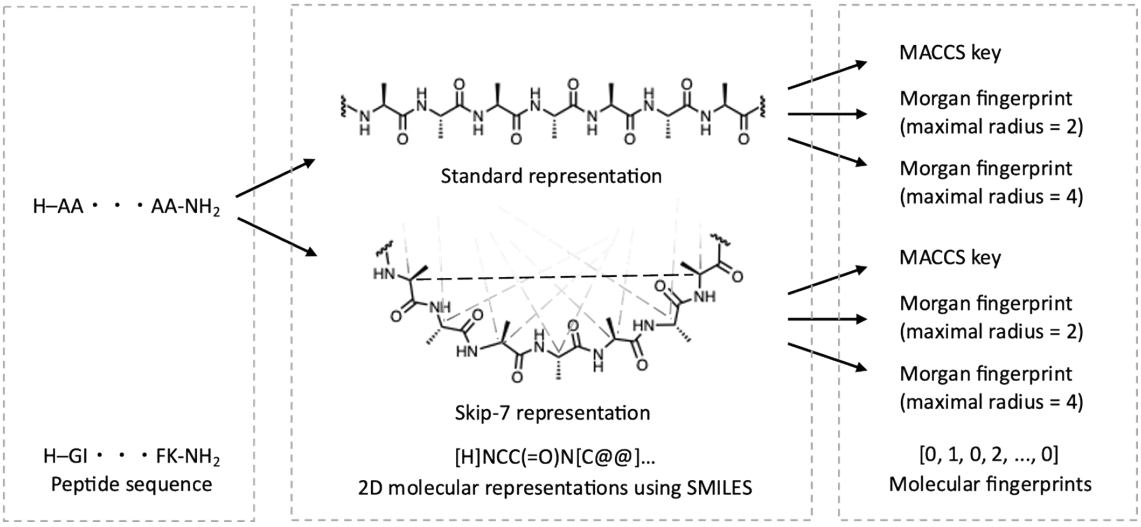



**Fig. S1** Flow showing the conversion from a peptide sequence to molecular fingerprints via 2D molecular representations using SMILES.

S1-3. The dataset used in this study

This study used an in-house dataset as the initial dataset, which is available at https://github.com/ycu-iil/MODAN/blob/main/data/Dataset_MODAN_initial.xlsx. The numbers of peptide data points for antimicrobial activities against each bacterium and haemolysis in the initial dataset were as follows: 40 for *Escherichia coli* NBRC 3972, 66 for *E. coli* DH5α, 80 for *Pseudomonas aeruginosa*, 47 for multiple-drug resistant *P. aeruginosa* (MDRP), 78 for *Staphylococcus aureus*, 22 for *Staphylococcus epidermidis*, and 57 for haemolysis.

S1-4. Evaluation of the prediction performances of the seven surrogate models

The prediction performances of the surrogate models were validated using the correlation coefficient between the predicted and experimental values via 10-fold cross-validation. Tables S2 and S3 present the prediction performances

of the trained surrogate models in the first and second rounds, respectively. In this study, the surrogate models with the best combination of SMILES representation and molecular fingerprints (Table S2) were used to recommend promising AMP candidates. For example, for haemolysis, a model with a Skip-7 representation and a Morgan fingerprint (radius = 4) was used.

**Table S2** Prediction performance of the surrogate models constructed in the first round.

| **SMILES** | **Fingerprint** | **Correlation coefficient** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | NBRC | DH5α | *P. aer.* | MDRP | *S. aur.* | *S. epi*. | Hemolysis |
| Standard | MACCS | 0.79 | 0.69 | 0.53 | 0.70 | 0.54 | **0.70** | 0.78 |
| Standard | Morgan (radius = 2) | 0.59 | 0.73 | **0.80** | 0.75 | 0.72 | 0.57 | 0.86 |
| Standard | Morgan (radius = 4) | 0.45 | 0.71 | 0.77 | 0.79 | **0.74** | 0.51 | 0.88 |
| Skip-7 | MACCS | **0.85** | 069 | 0.57 | 0.65 | 0.62 | 0.62 | 0.74 |
| Skip-7 | Morgan (radius = 2) | 0.51 | 0.71 | 0.71 | 0.75 | 0.64 | 0.28 | 0.86 |
| Skip-7 | Morgan (radius = 4) | 0.58 | **0.78** | 0.74 | **0.80** | 0.68 | 0.40 | **0.88** |

NBRC, *Escherichia coli* NBRC 3972; DH5α, *Escherichia coli* DH5α; *P. aer., Pseudomonas aeruginosa*; MDRP, Multiple-Drug Resistant *Pseudomonas aeruginosa*; *S. aur., Staphylococcus aureus; S. epi., Staphylococcus epidermidis*. "radius" in the fingerprint column represents the maximal radius.

**Table S3** Prediction performance of the surrogate models constructed in the second round.

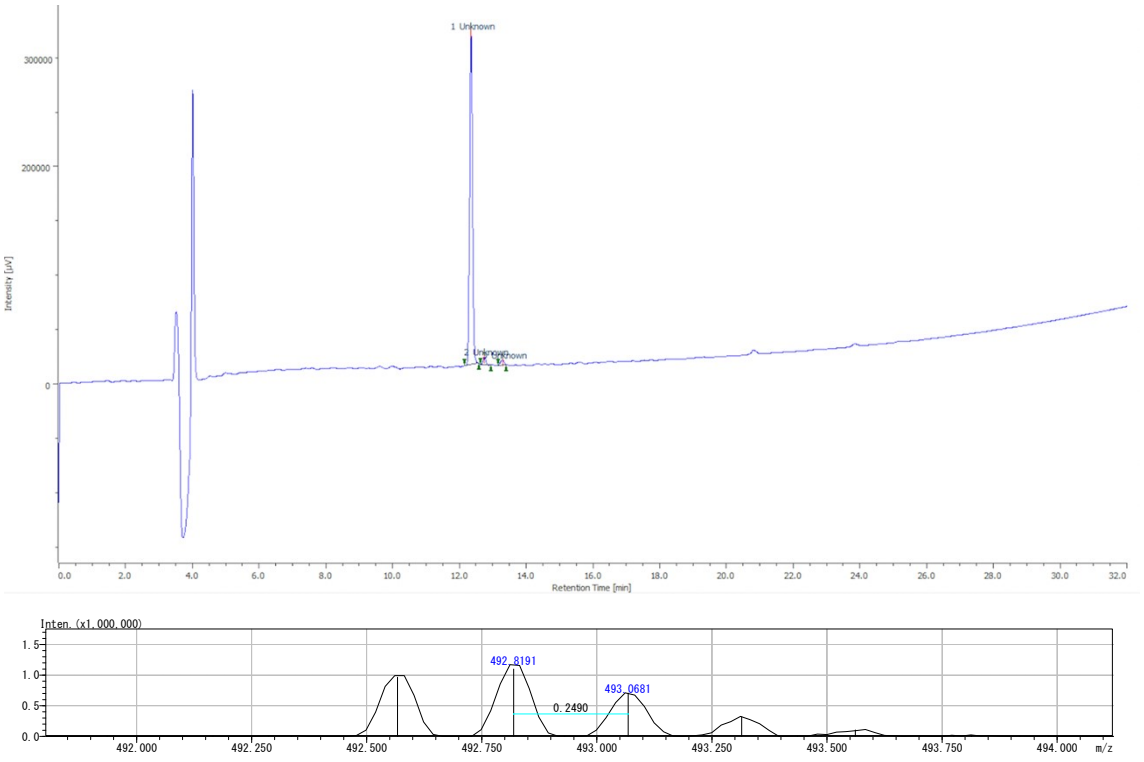
| **SMILES** | **Fingerprint** | **Correlation coefficient** | | | |
|---|---|---|---|---|---|
| | | DH5α | MDRP | *S. aur.* | Hemolysis |
| Standard | MACCS | 0.68 | 0.73 | 0.56 | 0.74 |
| Standard | Morgan (radius = 2) | 0.69 | 0.81 | 0.63 | 0.85 |
| Standard | Morgan (radius = 4) | 0.69 | 0.83 | 0.67 | 0.86 |
| Skip-7 | MACCS | 0.67 | 0.71 | 0.59 | 0.75 |
| Skip-7 | Morgan (radius = 2) | 0.70 | 0.80 | 0.61 | 0.87 |
| Skip-7 | Morgan (radius = 4) | 0.76 | 0.84 | 0.68 | 0.86 |

S2. LC-MS and HPLC data of the synthesised peptide data

Peptide initial, 17 amino acids

H-GIKKFLKSAKKFVKAFK-$NH_2$

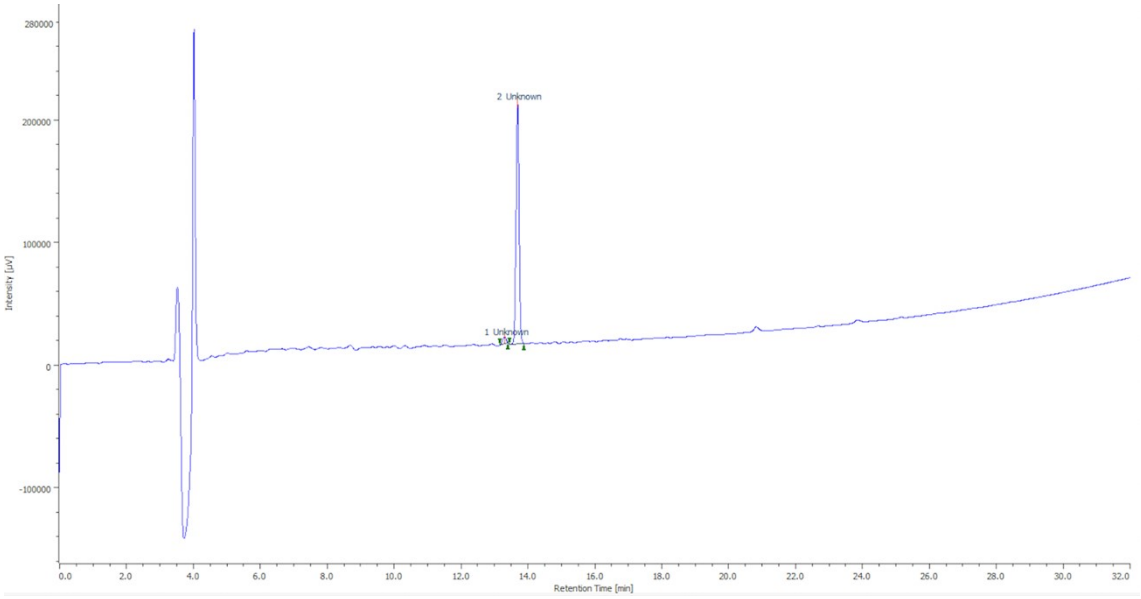
HRMS ($ESI^+$) calculated for $C_{97}H_{163}N_{25}O_{18}$ $[M+4H]^{4+}$:492.5652; observed:492.8191. Purity: 96.1%

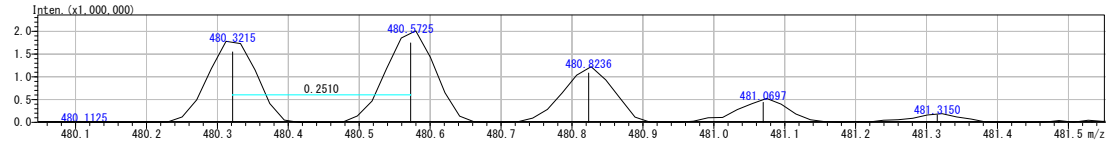


Peptide 1-1, 17 amino acids

H-GIKLLLKSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{94}H_{164}N_{24}O_{18}$ $[M+4H]^{4+}$:480.3164; observed:490.5725. Purity: 96.8%

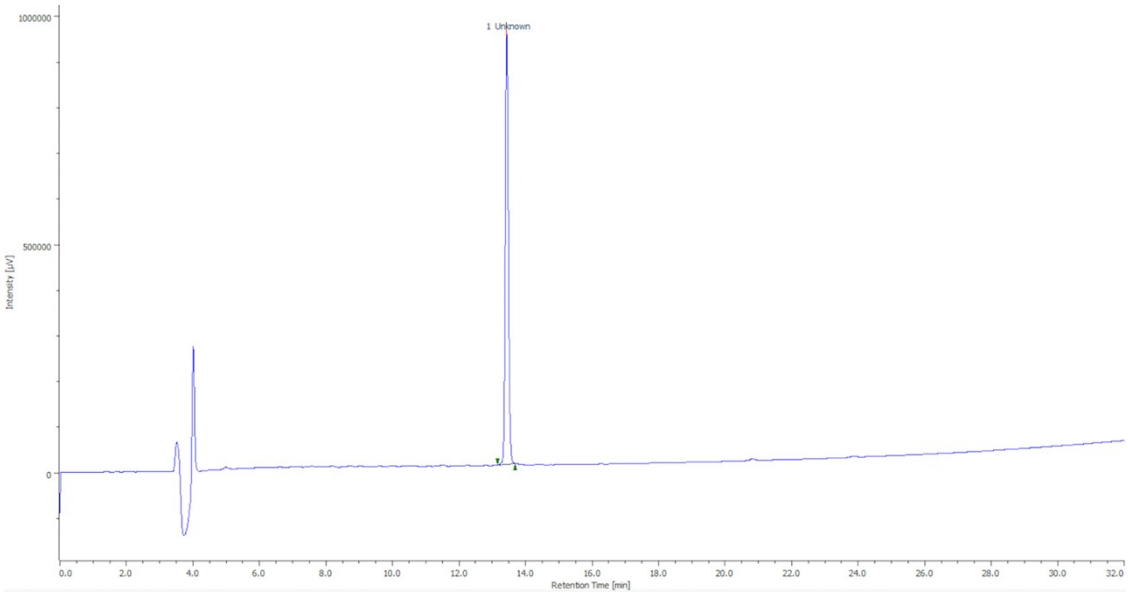


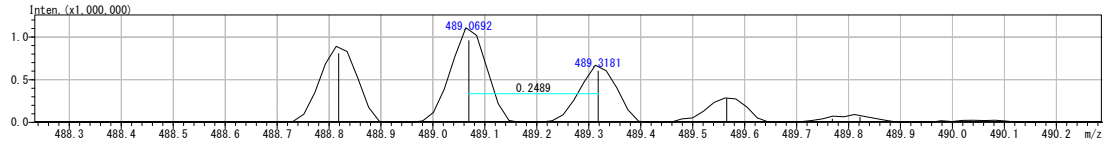



Peptide 1-2, 17 amino acids

H-GIKLFLKSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{97}H_{162}N_{24}O_{18}$ $[M+4H]^{4+}$:488.8125; observed:489.0692. Purity: >99.9%

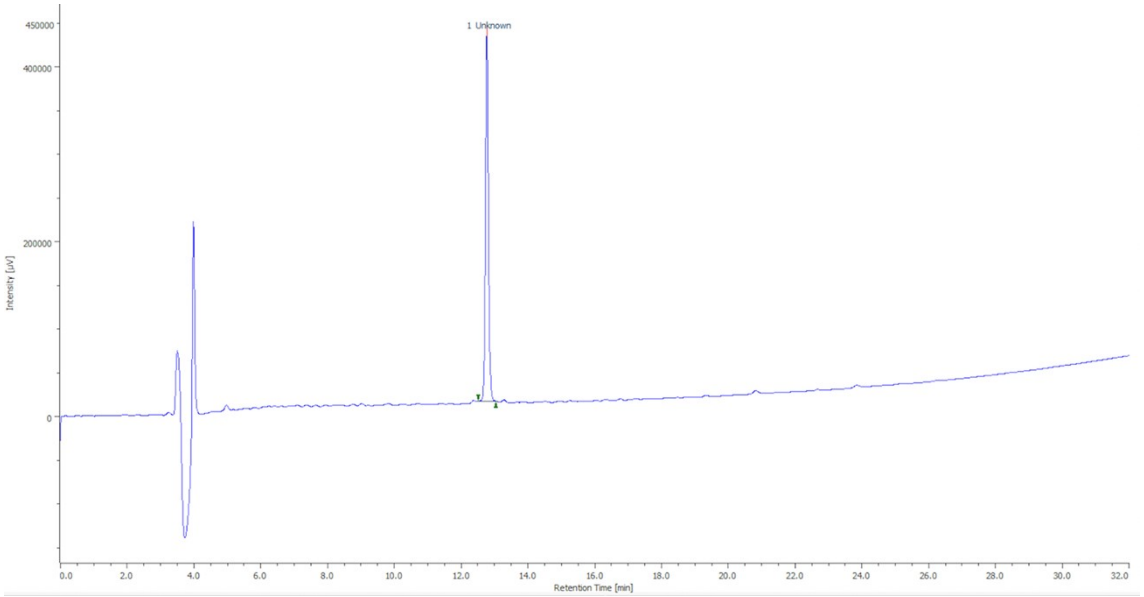


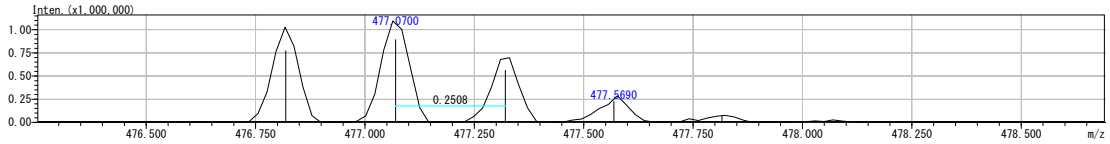



Peptide 1-3, 17 amino acids

H-GIKLVLKSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{93}H_{162}N_{24}O_{18}$ $[M+4H]^{4+}$:476.8125; observed:477.0700. Purity: >99.9%

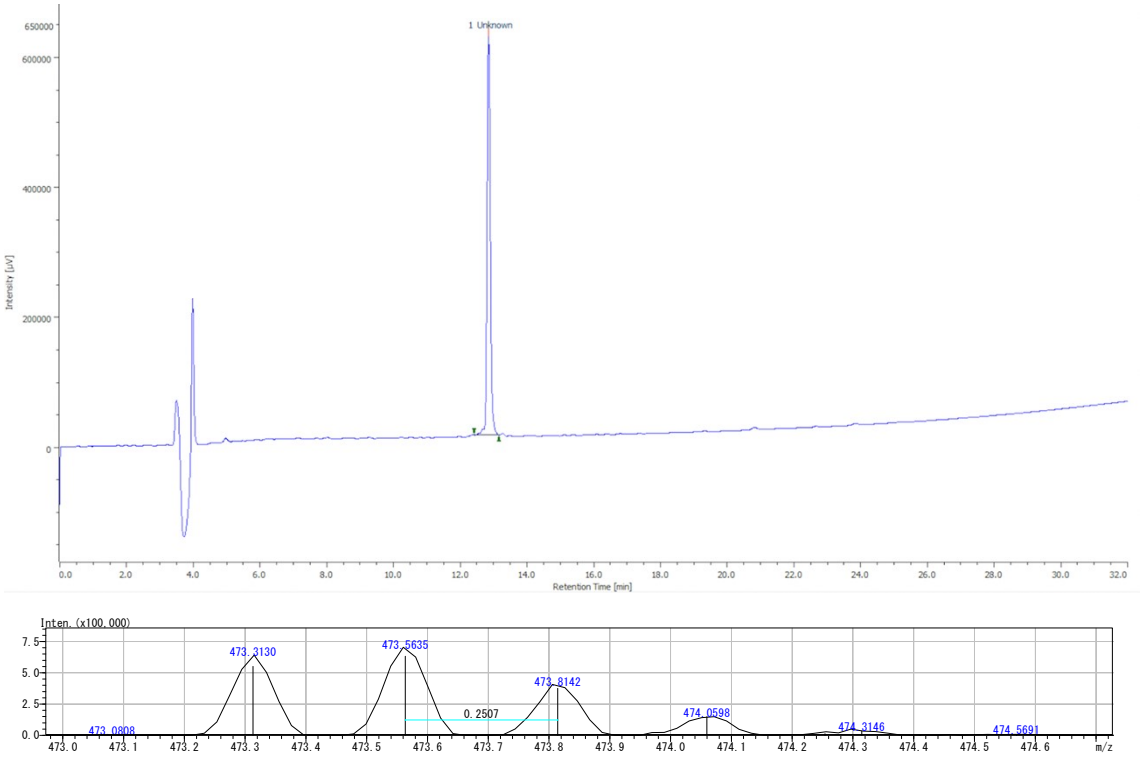


Peptide 1-4, 17 amino acids

H-GIKL (Aib) LKSAKKFVKAFK-$NH_2$

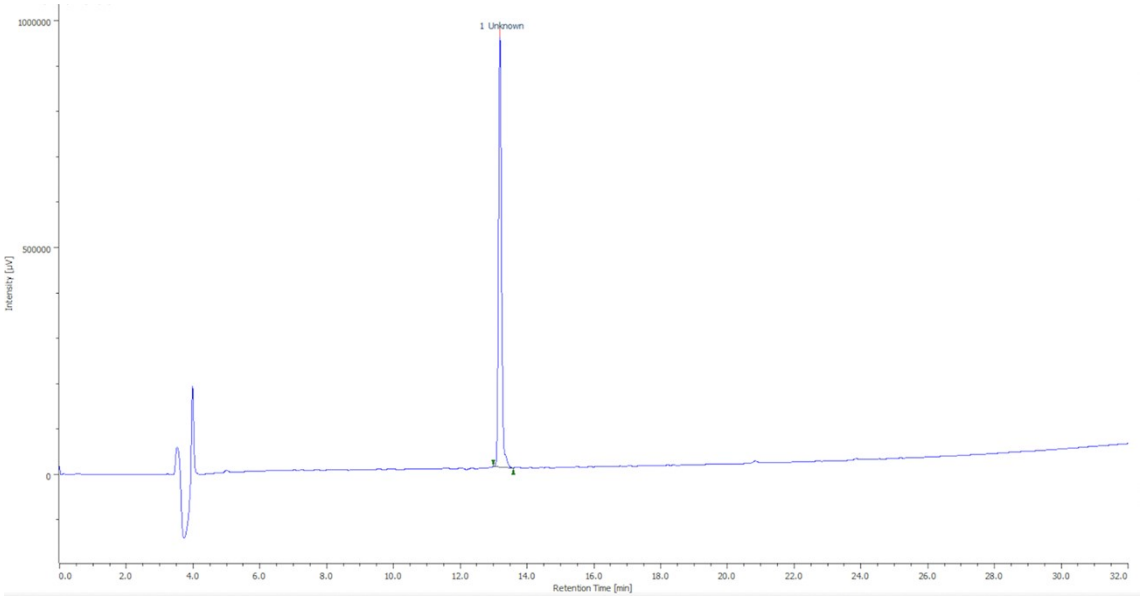
HRMS ($ESI^+$) calculated for $C_{92}H_{160}N_{24}O_{18}$ $[M+4H]^{4+}$:473.3086; observed:473.5635. Purity: >99.9%

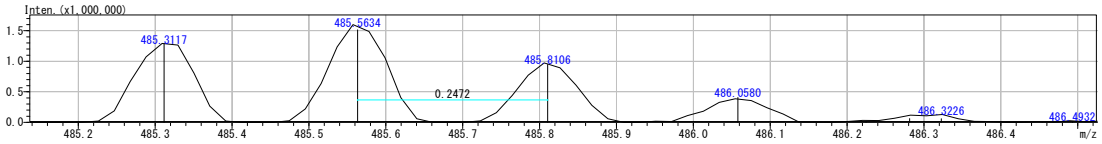


Peptide 1-5, 17 amino acids

H-GI (Orn) LFLKSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{96}H_{160}N_{24}O_{18}$ $[M+4H]^{4+}$:485.3086; observed:485.5634. Purity: >99.9%

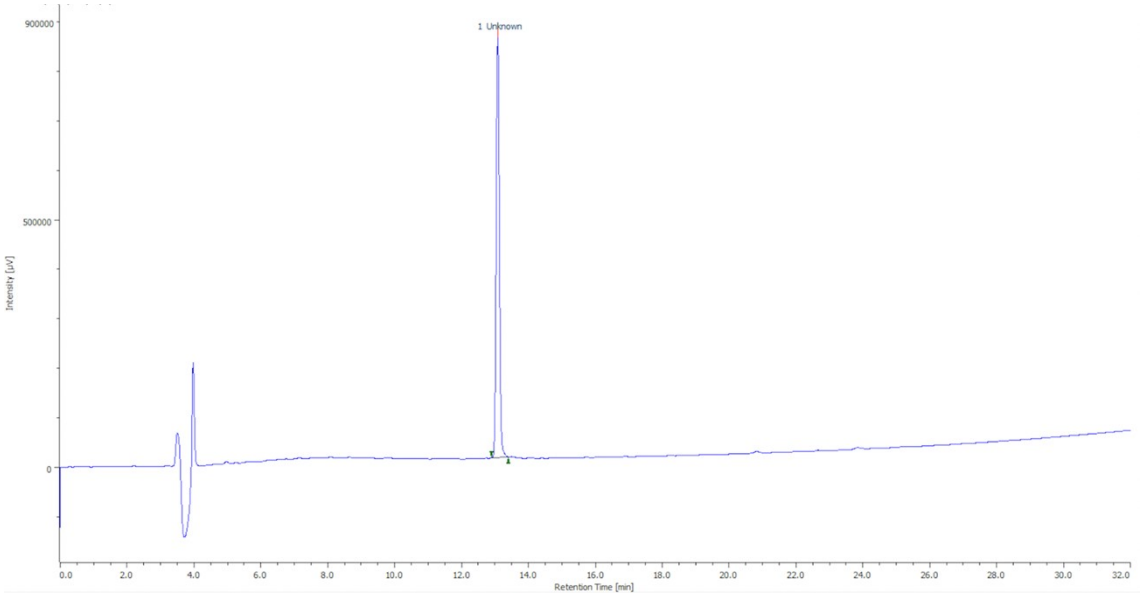


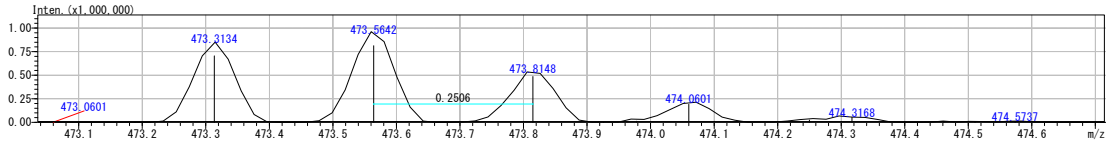



Peptide 1-6, 17 amino acids

H-GIKKFLKSAKL (Aib) VKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{92}H_{160}N_{24}O_{18}$ $[M+4H]^{4+}$:473.3086; observed:473.5462. Purity: >99.9%

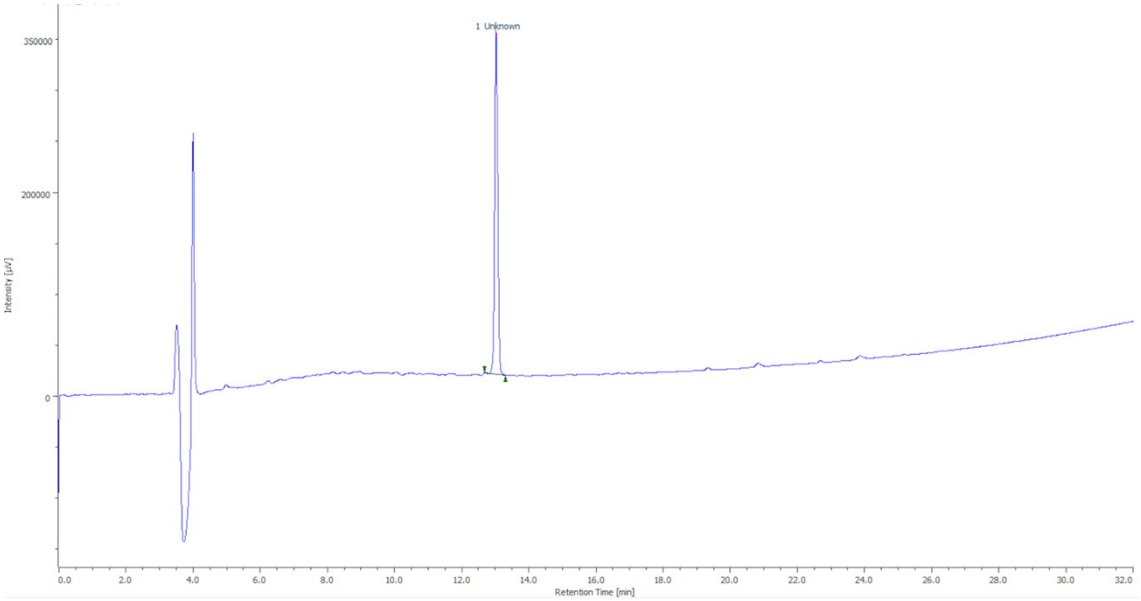


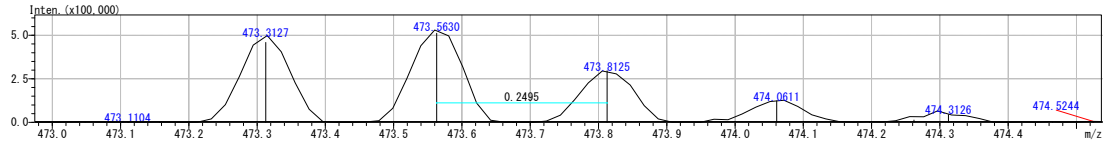



Peptide 1-7, 17 amino acids

H-GIKK (Aib) LKSAKLFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{92}H_{160}N_{24}O_{18}$ $[M+4H]^{4+}$:473.3086; observed:473.5630. Purity: >99.9%

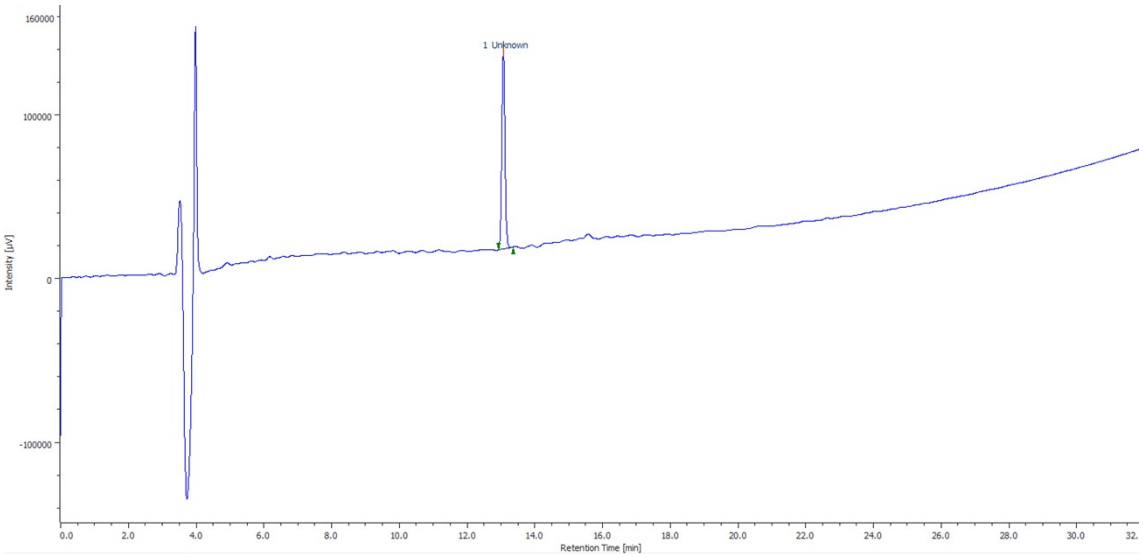


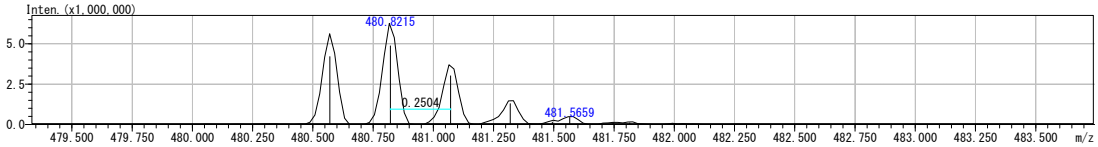



Peptide 2-1, 17 amino acids

H-GIKK (Aib) LKS (Aib) KKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{93}H_{163}N_{25}O_{18}$ $[M+4H]^{4+}$:480.5652; observed:480.8215. Purity: >99.9%

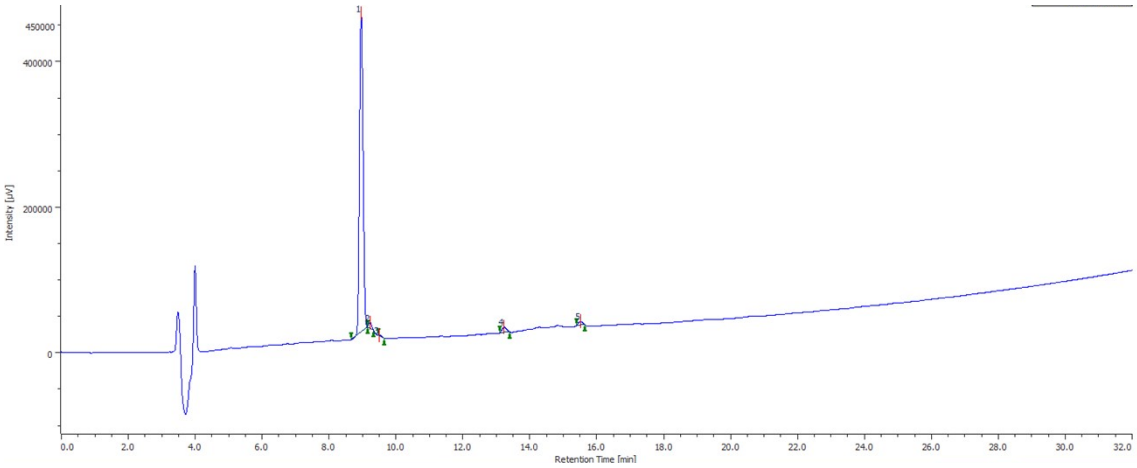


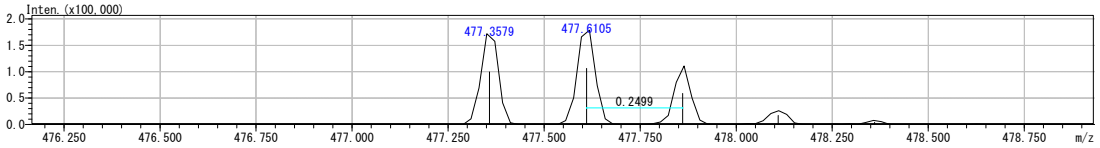



Peptide 2-2, 17 amino acids

H-GIKK (Orn) (Aib) KSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{91}H_{160}N_{26}O_{18}$ $[M+4H]^{4+}$:477.3101; observed:477.6105. Purity: 95.6%

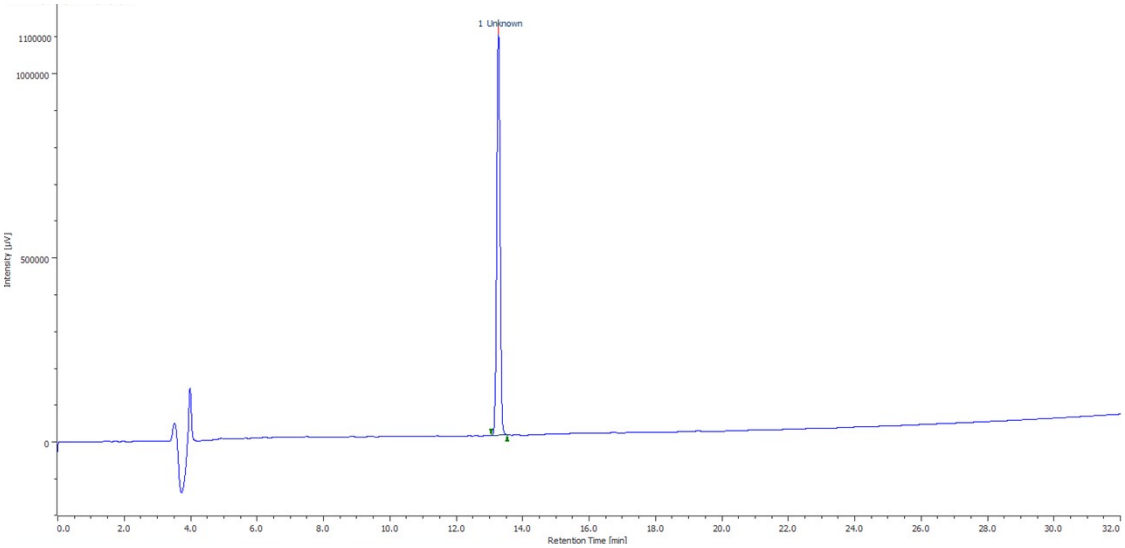


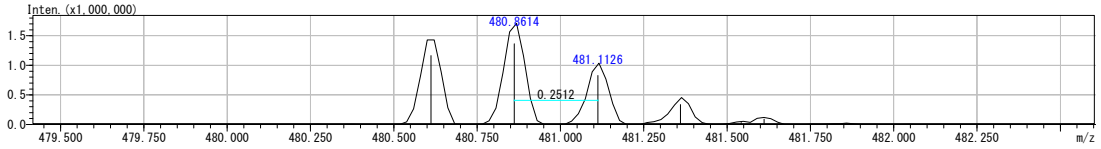



Peptide 2-3, 17 amino acids

H-GIKKFLKS (Aib) KK (Aib) VKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{93}H_{163}N_{25}O_{18}$ $[M+4H]^{4+}$:480.5652; observed:480.8614. Purity: >99.9%

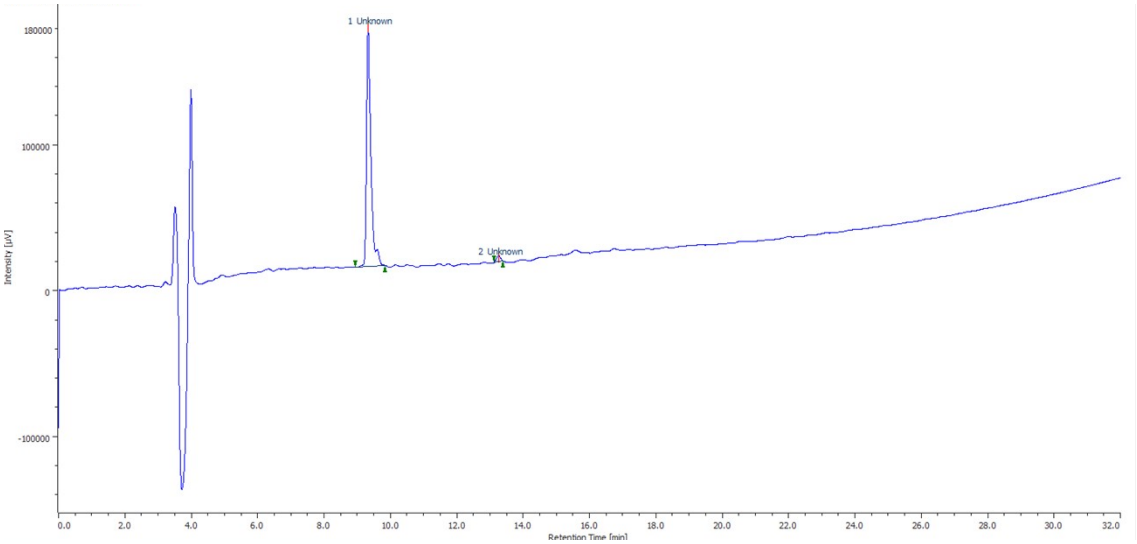


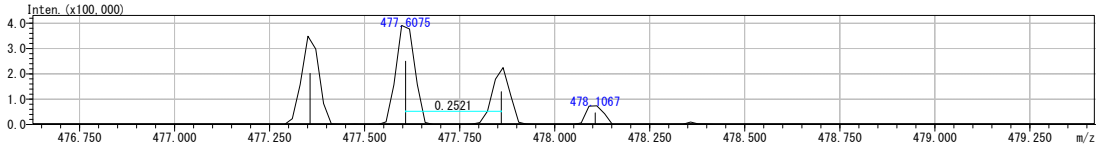



Peptide 2-4, 17 amino acids

H-GIKK (Aib) (Orn) KSAKKFVKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{91}H_{160}N_{26}O_{18}$ $[M+4H]^{4+}$:477.3101; observed:477.6075. Purity: 97.7%

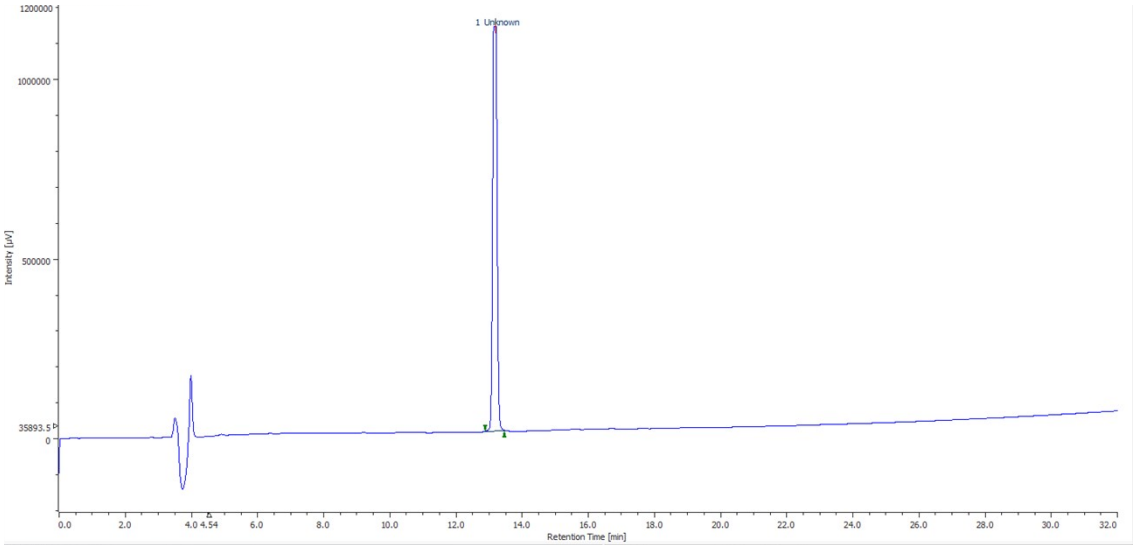


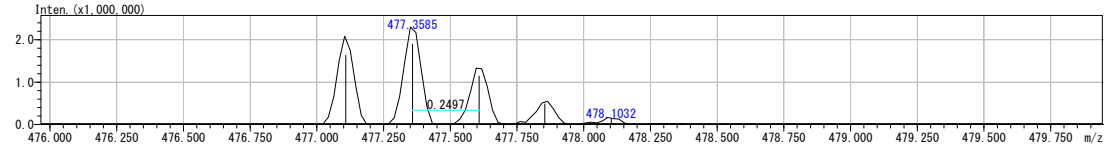



Peptide 2-5, 17 amino acids

H-GIKKFLKS (Aib) K (Orn) (Aib) VKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{92}H_{161}N_{25}O_{18}$ $[M+4H]^{4+}$:477.0613; observed:477.3585. Purity: >99.9%

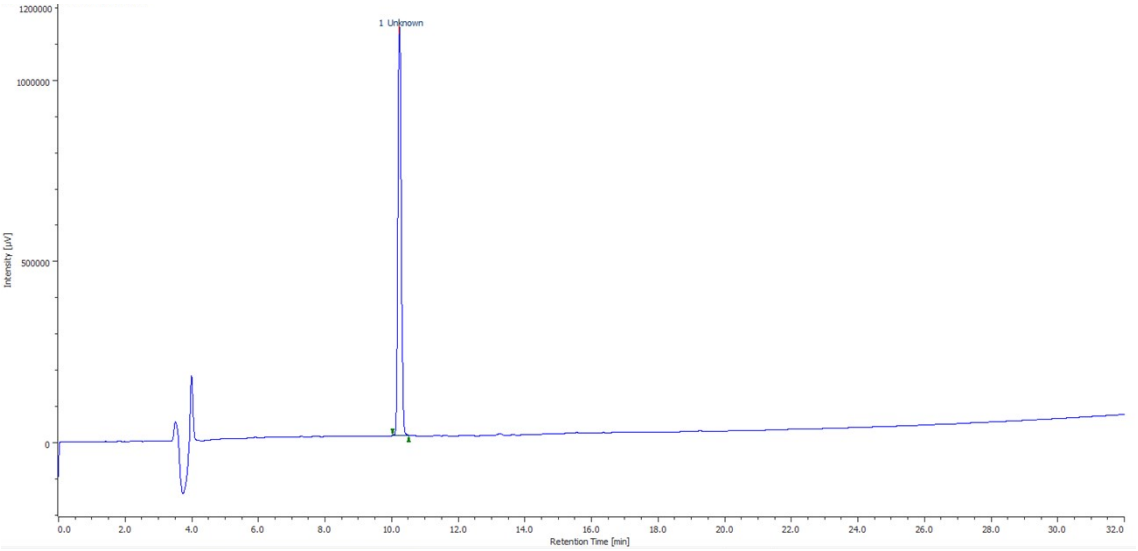


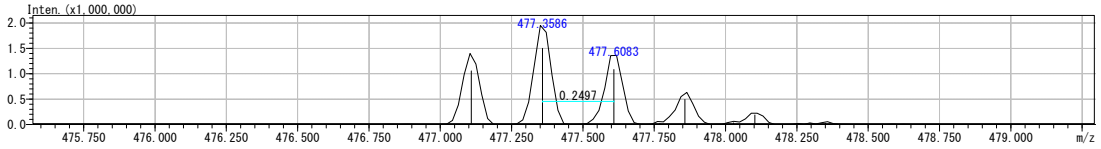



Peptide 2-6, 17 amino acids

H-GIKKFLKSAK (Orn) (Aib) (Orn) KAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{91}H_{160}N_{26}O_{18}$ $[M+4H]^{4+}$:477.3101; observed:477.3586. Purity: >99.9%

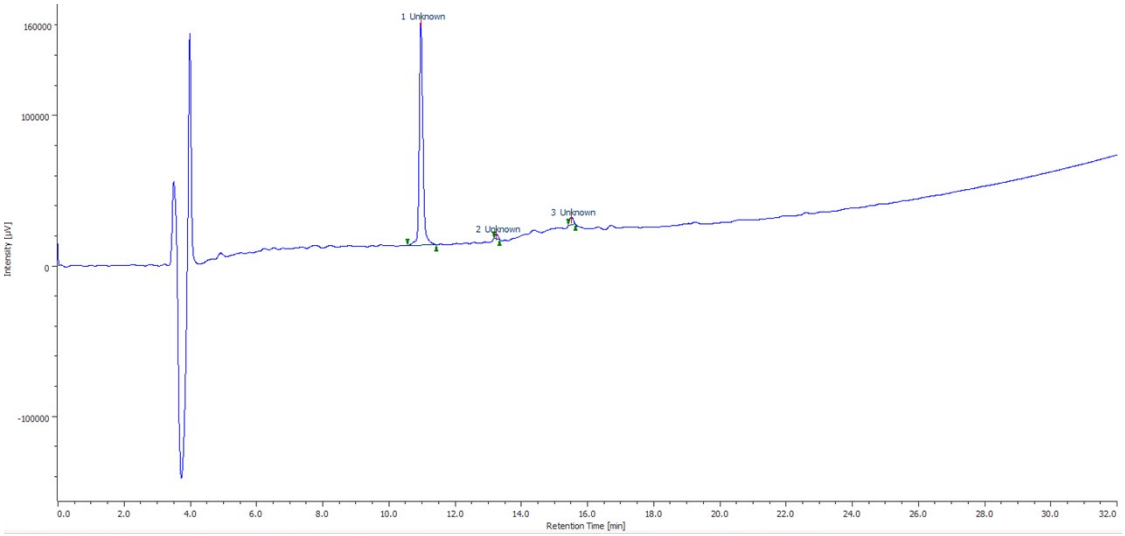


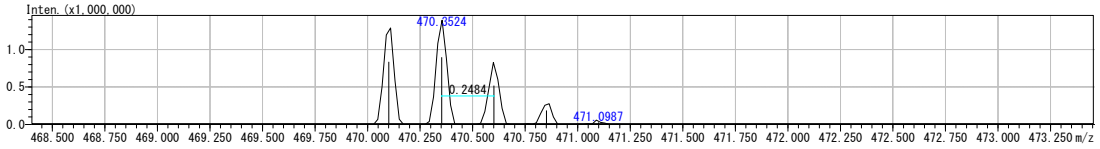



Peptide 2-7, 17 amino acids

H-GIKKFLKSAK (Aib) (Aib) (Orn) KAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{90}H_{157}N_{25}O_{18}$ $[M+4H]^{4+}$:470.0535; observed:470.3524. Purity: 95.0%

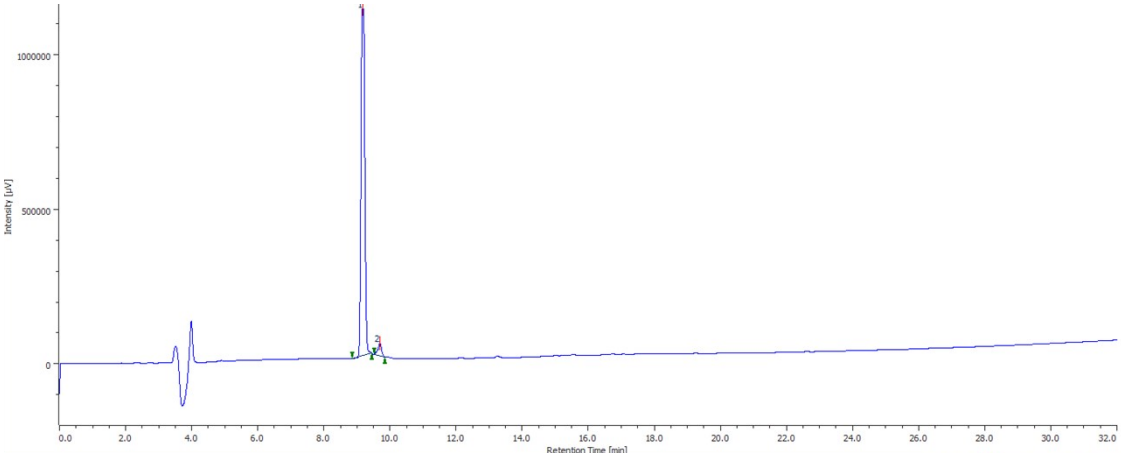


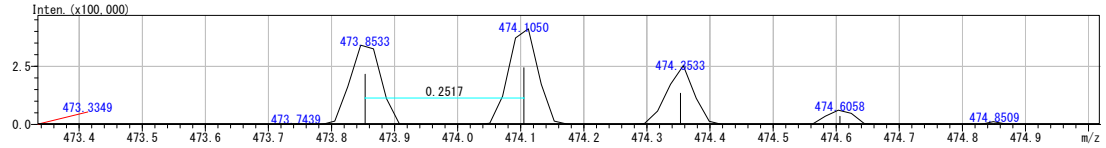



Peptide 2-8, 17 amino acids

H-GIKKF (Orn) KSAK (Orn) (Aib) VKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{90}H_{158}N_{26}O_{18}$ $[M+4H]^{4+}$:473.8062; observed:474.1050. Purity: 96.4%

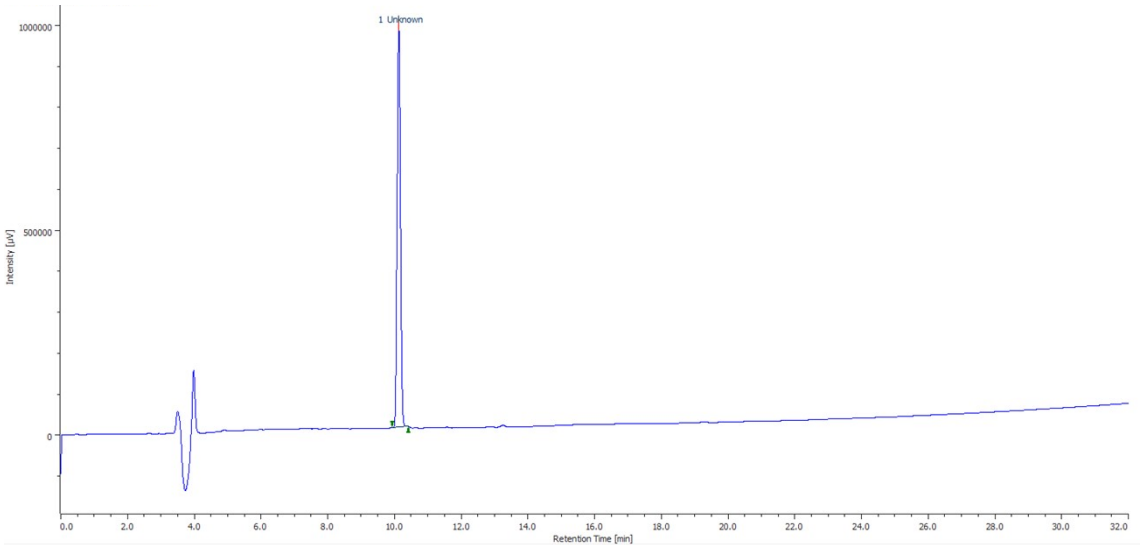


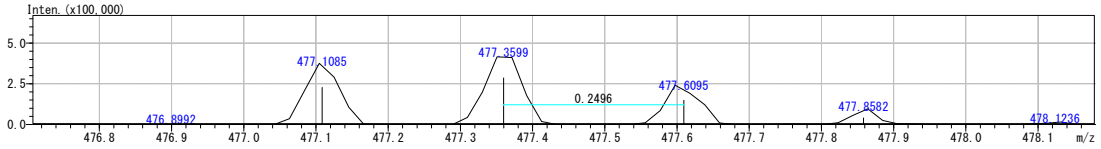



Peptide 2-9, 17 amino acids

H-GIKKFLKS (Orn) K (Aib) (Aib) VKAFK-$NH_2$

HRMS ($ESI^+$) calculated for $C_{92}H_{161}N_{25}O_{18}$ $[M+4H]^{4+}$:477.0613; observed:477.3599. Purity: >99.9%



S3. CD spectral analysis

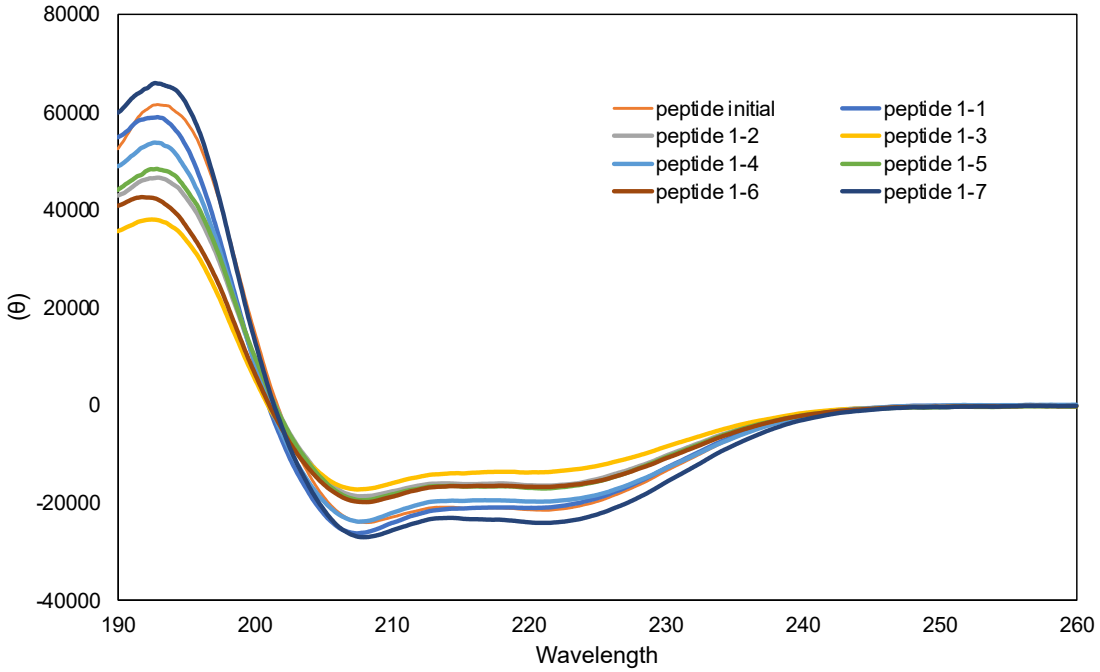



**Fig. S2** Secondary structure analysis using CD spectral data for the peptides in the first round.

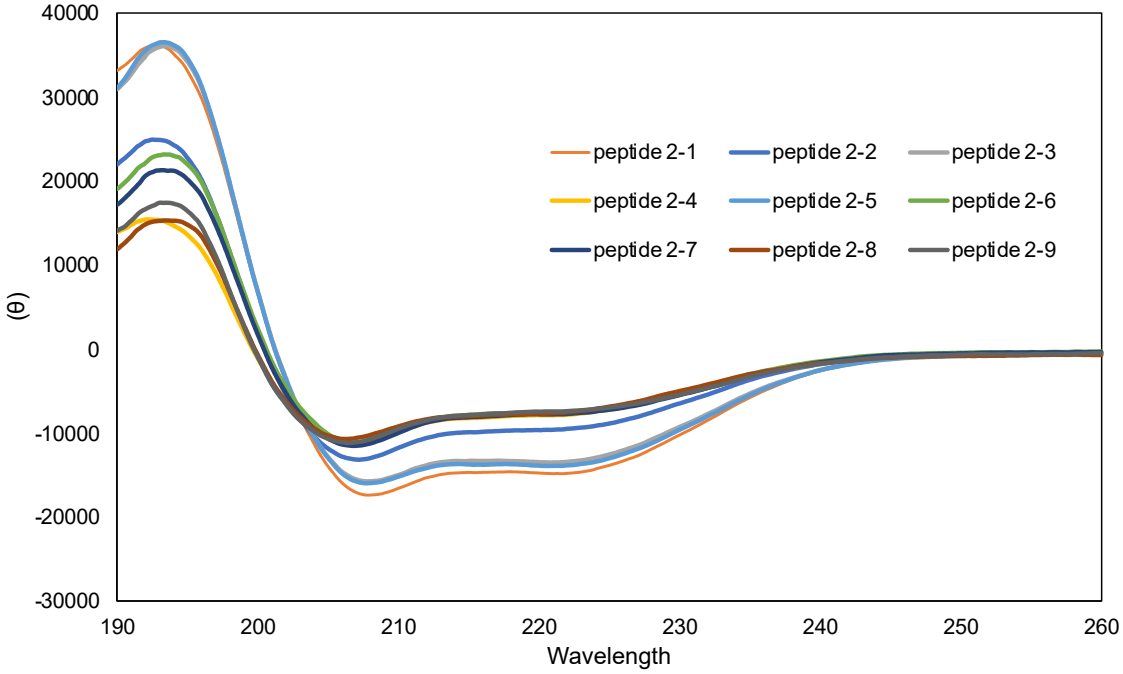



**Fig. S3** Secondary structure analysis using CD spectral data for the peptides in the second round.