Supporting information:

GPepT: A foundation language model for peptidomimetics incorporating non-canonical amino acids

Yuna Oikawa[1], Takanori Uzawa[2,5], Francois Berenger[1], Noriko Minagawa[2], Akiko Yumoto[2], Hideaki Takaku[5], Ryo Tamura[1,3,4], Yoshihiro Ito[5], and Koji Tsuda[1,3,4,*]

[1] *Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan.*

[2] *Emergent Bioengineering Materials Research Team, RIKEN Center for Emergent Matter Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.*

[3] *Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba 305–0044, Japan.*

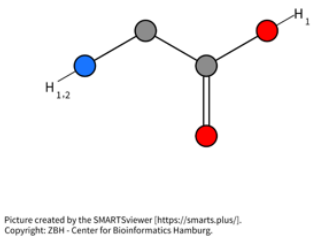
[4] *RIKEN Center for Advanced Intelligence Project, RIKEN, 1-4-1 Nihombashi, Chuo-ku, Tokyo, 103-0027 Japan.*

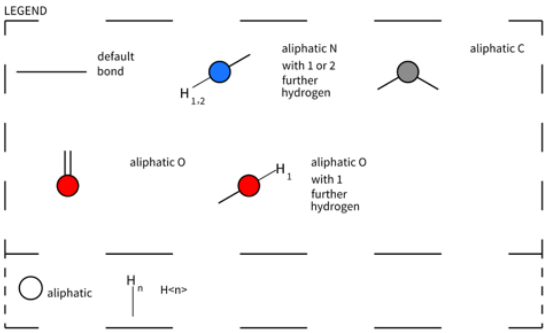
[5] *RIKEN Cluster for Pioneering Research, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.*

E-mail: tsuda@k.u-tokyo.ac.jp

Section S1: Algorithmic details of Monomerizer

1. Translation to a Molecular Representation: The structural formula is converted into a molecular representation. If the molecular representation contains multiple substructures, all smaller parts (e.g., ions) are removed.
2. Template Matching: The molecular representation is compared to template molecules of the 20 canonical amino acids. If a match is found, the corresponding atoms in the input molecule are labeled as such. To avoid incorrect partial matches, the template dictionary is organized from the largest to the smallest amino acids.
3. Peptide Bond Identification: Atoms that match the peptide bond template structures (shown below) are labeled as such. If the number of bonds is smaller than the minimum set by the user's preference, the procedure for that molecular representation is terminated. In this study, the minimum was set to 3.
4. Non-Canonical Fragments Identification: Atoms that remain unlabeled are identified as non-canonical fragments. A breadth-first search is performed to group these atoms by peptide bonds. If a labeled peptide bond is found within a ring structure, the process is terminated, as cyclic sequences are not the focus.
5. Fragment Isolation: All atoms labeled with canonical amino acids and peptide bonds are temporarily removed. The bonding sites of these fragments to neighboring amino acids are recorded for future use in the program.
6. Fragment Classification: Each fragment is classified as either an ncAA or a terminal modification based on whether they match any valid backbone template. Canonical SMILES of these monomers and the labeled peptide molecule are saved for later processing. After processing each input structural formula, the list of obtained non-canonical fragments undergoes deduplication depending on their tautomer hash.
7. Template Re-matching: A final round of labeling is performed on each output molecule, now including the obtained non-canonical fragments as templates. Groups of labeled atoms are checked for connections to peptide bonds, identifying groups with only peptide bonds connected, as the terminal of the sequence. Of them, the monomer with an N atom at the end is identified as the N-terminal (the start of the sequence). A breadth-first search is conducted starting from the N-terminus to determine the sequence of monomers.
8. Outputs results: The algorithm produces detailed output for both monomeric units and complete peptides. For each peptide and each monomer, we provide structural illustration and sequence representation. To maintain data integrity, Monomerizer removes invalid sequences containing misplaced terminal modifications. This time we also filtered out any monomers that cannot be found on PubChem as well as the sequences containing them.





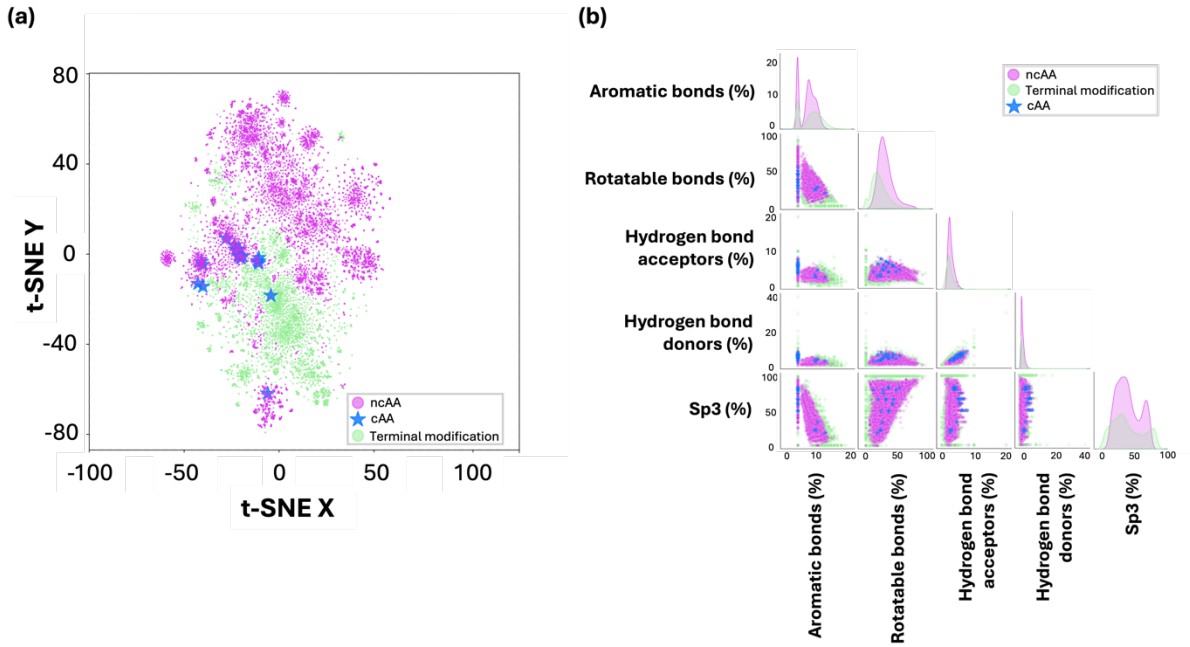


Figure S1: Comparison of non-canonical amino acids (ncAAs), terminal modifications and canonical amino acids (cAAs) mined from ChEMBL. (a) t-SNE visualization of Morgan fingerprints. (b) Distribution of physiochemical properties.

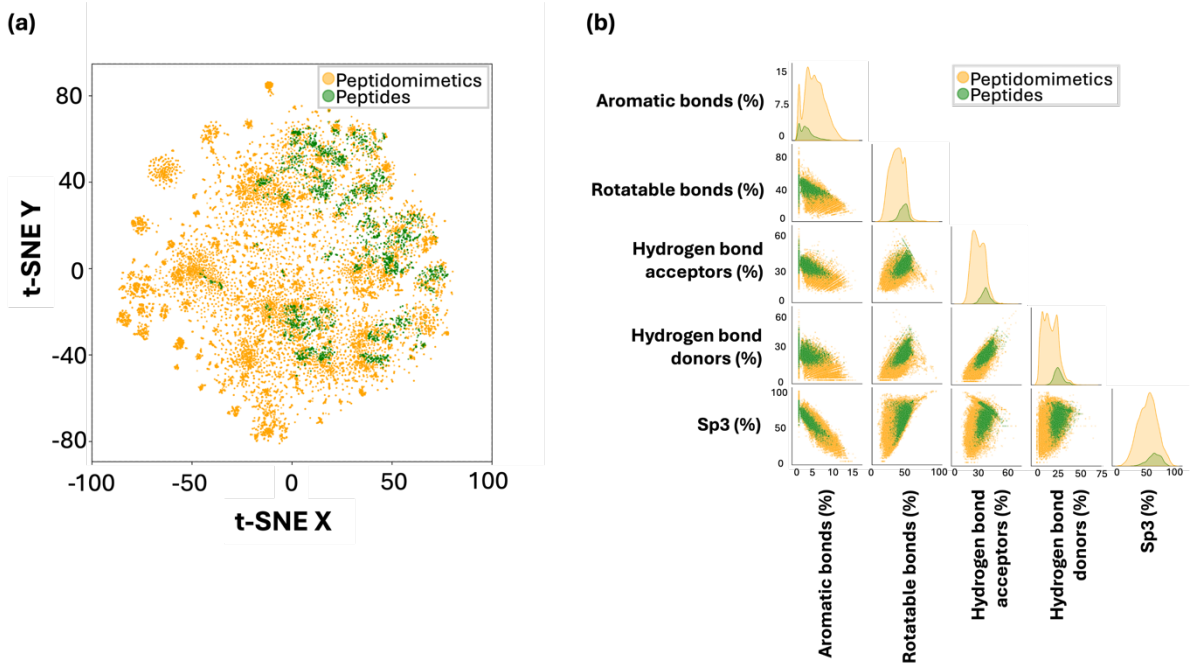


Figure S2: Comparison of peptidomimetics and peptides mined from ChEMBL (Dataset P). (a) t-SNE visualization of Morgan fingerprints. (b) Distribution of physiochemical properties.

Table S1: Valid peptidomimetics chosen for antimicrobial activity test.

Name	Sequence	Elements
Pep1	X556WX556WWKZ0	X556=D-Tryptophan, Z0=Amide
Pep2	X556WWWWWWWWWWWW	X556=D-Tryptophan
Pep3	LQKYYRRVRGGRX518F	X518=N6-(Trifluoroacetyl)-L-lysine
Pep4	QGRKQGRRX518	X518=N6-(Trifluoroacetyl)-L-lysine
Pep5	X3449WWWWWR	X3449=Citrulline

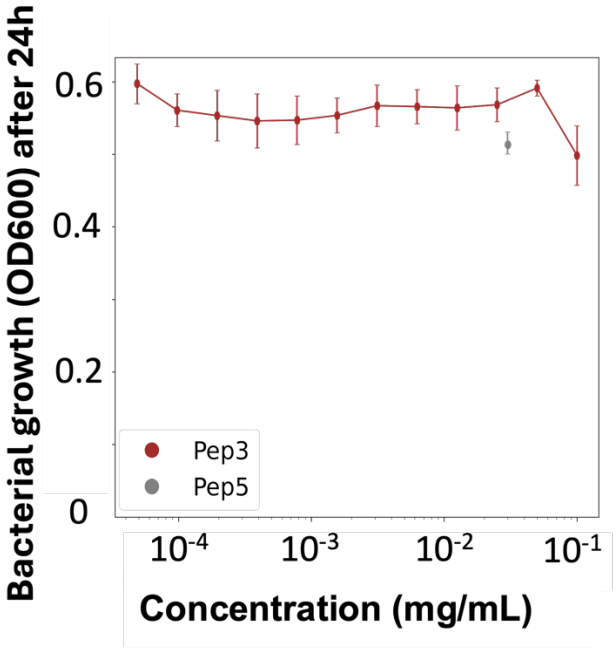


Figure S3: Bacteria growth (OD600) after 24 hours against peptide concentration (Pep3 and Pep5).