ACS Medicinal Chemistry Letters

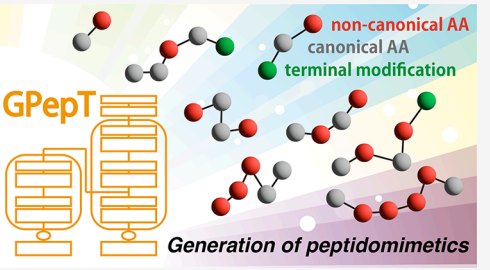
pubs.acs.org/acsmedchemlett
Letter

GPepT: A Foundation Language Model for Peptidomimetics Incorporating Noncanonical Amino Acids

Yuna Oikawa, Takanori Uzawa, Francois Berenger, Noriko Minagawa, Akiko Yumoto, Hideaki Takaku, Ryo Tamura, Yoshihiro Ito, and Koji Tsuda*

Cite This: ACS Med. Chem. Lett. 2025, 16, 1670−1675

ABSTRACT: Language models have been increasingly popular in therapeutic peptide generation, but molecular diversity remains limited due to reliance on the 20 canonical amino acids. We propose a language model that generates peptidomimetics incorporating noncanonical elements like noncanonical amino acids and terminal modifications. To accomplish this, we created a vocabulary of over 17,000 noncanonical elements by extracting them from chemical formulas stored in the ChEMBL database. Our pretrained language model, GPepT, showed improved diversity in molecular structures and chemical properties. To demonstrate its real-world application, we fine-tuned the model for antimicrobial peptides. Experimental validation revealed that one of the generated peptidomimetics exhibited effective antimicrobial activity, marking a successful case of AI-driven peptide development. GPepT is fully accessible on HuggingFace: https://huggingface.co/Playingyoyo/GPepT.



KEYWORDS: *Noncanonical Amino Acids, Amino Acids, Peptidomimetics, Protein, Peptide, Antimicrobial Peptides, RDKit, SMILES, GPT, AI, Language Model*

Peptides play a pivotal role in various biological functions, including antimicrobial, anticancer, and anti-inflammatory activities. Recent advancements in peptide engineering have sparked interest in developing novel peptides as therapeutic agents, with over 53 peptides—accounting for 10% of the 509 drugs approved by the FDA between 1999 and 2019—emerging in clinical applications.[1] To enable broader range of functions and binding properties, a diverse library of therapeutic peptides is crucially important.[2] There are at least three approaches to increasing diversity: 1) Collecting peptide sequences from a large pool of organisms.[3] 2) De novo sequence generation via machine learning.[4−8] 3) Incorporation of noncanonical elements.[9] In the first approach, Santos-Junior et al. successfully predicted nearly 1 million antimicrobial peptides from the global microbiome.[3] In the second approach, a variety of deep learning models have been developed. Earlier, separate generative models have been developed for distinct purposes. Examples include variational autoencoders,[4,6,10] recurrent neural networks,[5,8] generative adversarial networks[7] and transformers.[11] More recently, it has become more customary to build a foundation model pretrained with unlabeled data, which can then be fine-tuned to serve a specific purpose, e.g., PeptideBERT[11] and ProtGPT2.[12] The above-mentioned approaches create peptides only with 20 canonical amino acids; hence the chemical diversity of generated peptides is intrinsically limited.

In the third approach, the chemical space is expanded by introducing noncanonical elements. These elements, comprising noncanonical amino acids (ncAAs) and terminal modifications, give rise to ″peptidomimetics″—peptide-like molecules that transcend the limitations of conventional amino acid sequences.[2] Murakami et al.[9] extended their language model with few ncAAs, but the impact on diversity was insubstantial. Although subsets of common noncanonical elements exist in the literature (see, e.g., Goettig et al.[13]), there is no comprehensive collection of noncanonical elements, let alone tokens representing them. This deficiency of tokens presents a significant barrier to generating chemically diverse amino acid sequences using contemporary language models.[14]

Chemical compound databases, particularly ChEMBL,[15] contain peptidomimetics with previously underutilized noncanonical elements that may prove valuable in therapeutics (Figure 1a). CHEMBL2407177, synthesized by Murugan et al.,[16] demonstrates the utility of a Histidine-derived ncAA in creating antimicrobial therapeutics with enhanced proteolytic stability and negligible hemolytic activity. Marine-derived amino acid substituents, specifically brominated variant, are featured in a synthetic antifungal peptidomimetic

Received: June 13, 2025
Revised: July 8, 2025
Accepted: July 17, 2025
Published: July 22, 2025

https://doi.org/10.1021/acsmedchemlett.5c00375
ACS Med. Chem. Lett. 2025, 16, 1670−1675

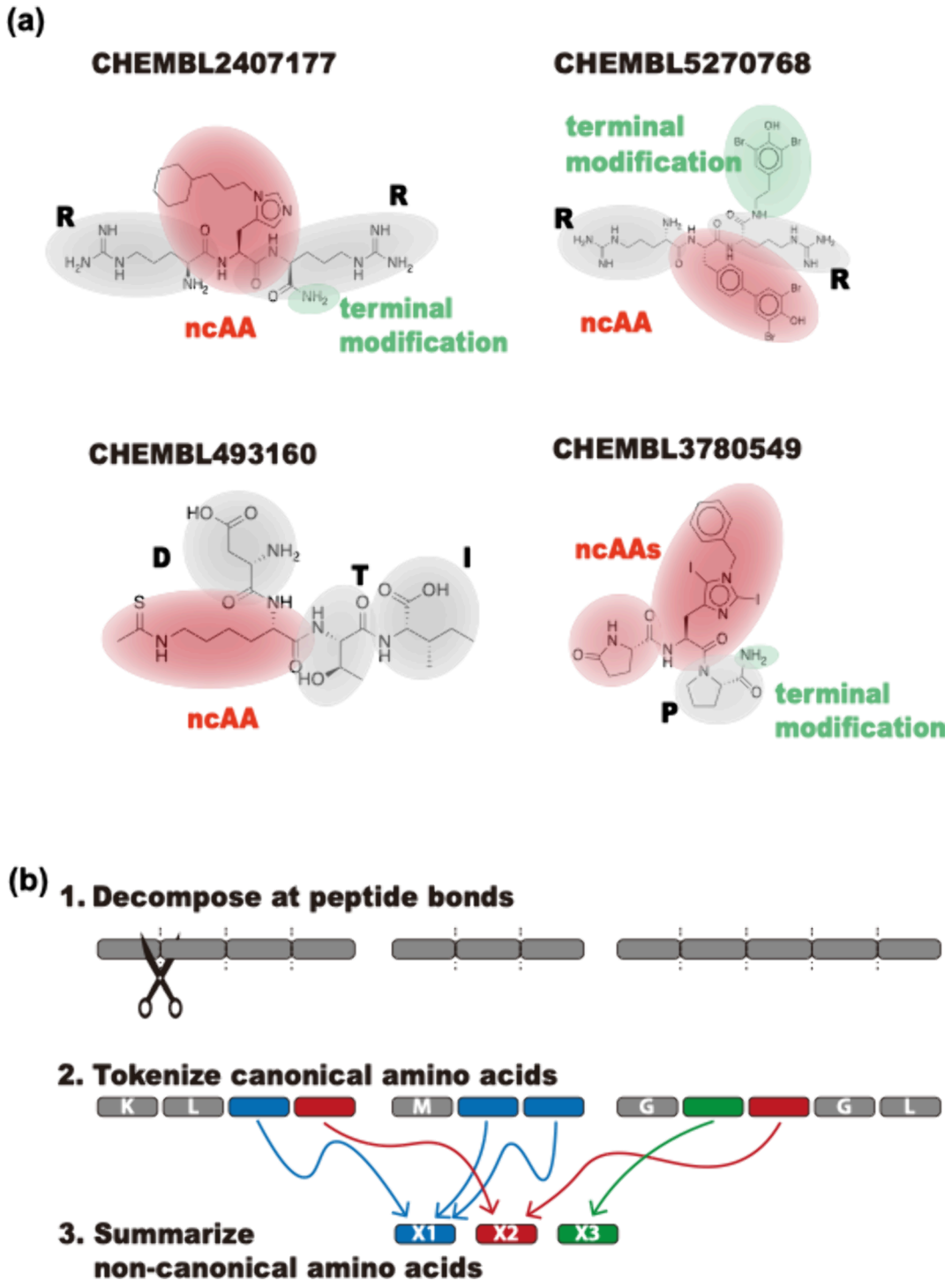


Figure 1. (a) Examples of uncommon amino acids found in ChEMBL database. (b) Function of Monomerizer: It decomposes peptides and peptidomimetics, represented as chemical formulas, into canonical and noncanonical amino acids, and tokenizes the newly identified noncanonical elements.

CHEMBL5270768.[17] is reported to be a potent inhibitor of SIRT1, a protein linked to type 2 diabetes and heart disease, as well as SIRT2, which may be involved in glioma tumorigenesis and Parkinson's disease. CHEMBL3780549, a thyrotropin-releasing hormone analogue, incorporates novel amino acids proposed by Meena et al.[18] Notably, these ncAAs remain absent in major chemical databases including PubChem,[19] limiting their accessibility to the broader scientific community.

In this paper, we construct a comprehensive vocabulary of noncanonical elements by detecting them in a large number of chemical formulas and use it to build a foundation model for peptidomimetics. The vocabulary was created with a Python-based software we named Monomerizer that decomposes peptides/peptidomimetics represented as chemical formulas into amino acids and tokenize noncanonical elements (Figure 1b). As a result of applying Monomerizer to ChEMBL-registered molecules, noncanonical elements were obtained,

https://doi.org/10.1021/acsmedchemlett.5c00375
ACS Med. Chem. Lett. 2025, 16, 1670−1675

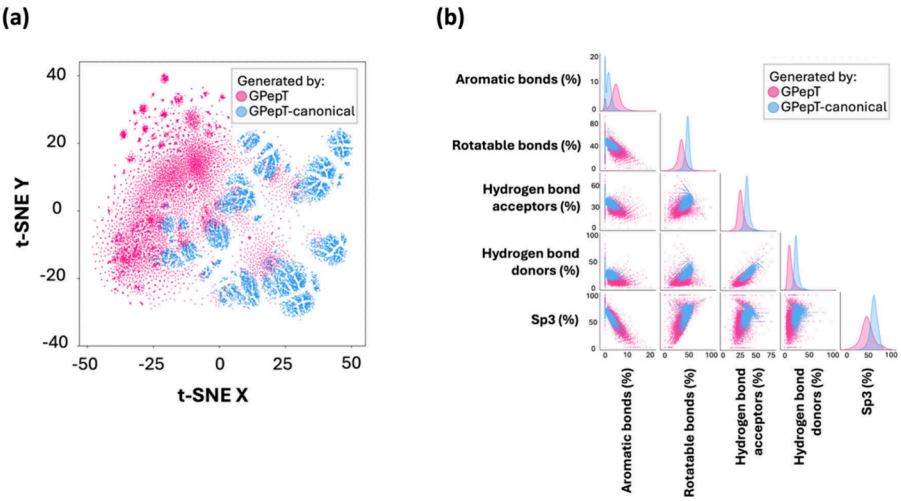


Figure 2. Comparison of amino acid sequences generated by GPepT and GPepT-canonical. (a) t-SNE visualization of Morgan fingerprints. (b) Distribution of physiochemical properties of the amino acid sequences.

and unique tokens were assigned to them. Next, the chemical formulas were converted to sequences via the vocabulary and used to pretrain a standard transformer language model.[12] The pretrained model is called Generative pretrained Peptidomimetics Transformer (GPepT). Using GPepT without any fine-tuning, tens of thousands of peptidomimetics were generated. The chemical diversity of generated sequences was found to be greatly enhanced in terms of both structural features and chemical properties (Figure S2).

To demonstrate the relevance of our study to real-world peptide development, we conducted a case study on designing antimicrobial peptides. Our model was fine-tuned with the sequences that showed antimicrobial activity against *Escherichia coli*. Among the generated sequences, five peptides proceeded to synthesis after synthesizability assessment by experts. One peptide, containing D-Tryptophan, exhibited antimicrobial activity against *E. coli*, making it one of the first successful cases of AI-generated peptidomimetics.

Monomerizer accepts a molecule in SMILES (Simplified Molecular Input Line Entry System) format,[20] a textual representation of chemical structures that encodes atoms and bonds as a string. Any molecule with fewer than two peptide bonds are rejected. Then, canonical amino acids are removed from the molecule by SMARTS (SMILES Arbitrary Target Specification)-based template matching,[21] which allows for specifying substructural patterns to identify and manipulate specific parts of a molecule. Each of the remaining parts are classified into two categories: 1) ncAAs, 2) terminal modifications, according to the presence or absence of a backbone. Once all molecules are processed, identical fragments are summarized and assigned unique tokens. Tokens $X_1, \ldots, X_n$ are assigned to all ncAAs, and tokens $Z_1, \ldots, Z_n$ to all terminal modifications. See Section S1 for algorithmic details.

By applying Monomerizer to all bioactivity-labeled 2,409,270 molecules on ChEMBL,[22] we identified 11,243 ncAAs and 6465 terminal modifications. 7157 (63.7%) of the ncAAs and 2811 (43.5%) of the terminal modifications were not registered in ChEMBL or PubChem as individual molecules. Using these tokens, 42,743 molecules were successfully converted into sequences, 38,138 (89.2%) of which were peptidomimetics containing at least one of the noncanonical elements. This collection of sequences, which we designate as *Data set P*, serves as the foundation for our subsequent analyses.

We developed GPepT by adapting the GPT-2 large transformer decoder from HuggingFace, consisting of 36 layers and a dimensionality of 1280, for sequence design of peptidomimetics. To handle our elements, we reinitialized the pretrained weights and built a custom tokenizer specifically designed to tokenize each element—canonical or non-canonical—rather than words or subwords as in traditional natural language processing. We then used HuggingFace's run_clm.py script,[23] which facilitates next-token prediction training with a single command, to train GPepT on our Data set P. Training adhered to standard configurations, including cross-entropy loss for autoregressive generation, a base learning rate of 1×10^{-5}, and Adam optimization, a widely used stochastic gradient descent method that adapts learning rates for each parameter using estimates of first and second moments of the gradients,[24] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. To assess the influence of noncanonical elements, we trained two versions of the model: GPepT, trained on the full data set, and GPepT-canonical, trained only on sequences composed of canonical amino acids.

After training, novel sequences were sampled with a repetition penalty 1.5. Occasionally, invalid sequences were generated where a terminal modification appeared in the middle. Sampling continued until 10,000 valid sequences were obtained. We converted the sequences back to a chemical formula and represented them as Morgan fingerprints, fixed-length binary vectors encoding molecular substructures used widely for chemical similarity and machine learning tasks. Using t-SNE (t-distributed Stochastic Neighbor Embedding),[25] we visualized the reduced high-dimensional data of Morgan fingerprints in two dimensions for the generated sequences by GPepT and GPepT-canonical (Figure 2a). The sequences generated by GPepT are more dispersed than those from GPepT-canonical, highlighting the superior chemical diversity provided by noncanonical elements. Figure 2b shows the individual and joint distributions of five physicochemical properties: fraction of aromatic bonds, fraction of rotatable bonds, fraction of hydrogen bond acceptors, fraction of hydrogen bond donors, and the fraction of carbon atoms that are SP3 hybridized. The increased diversity is evident in

https://doi.org/10.1021/acsmedchemlett.5c00375
ACS Med. Chem. Lett. 2025, 16, 1670−1675

these properties as well. Figures S1 and S2 present similar results for individual noncanonical elements and Data set P, respectively.

Antimicrobial resistance attributed to 1.27 million deaths in 2019.[26] Antimicrobial peptides (AMPs), key components of innate immunity, have emerged as candidates in the fight against resistant pathogens.[27] While antimicrobial peptides (AMPs) have emerged as promising candidates due to their role in innate immunity, bacterial resistance to AMPs, though rare, has been documented.[28] Nature has addressed this challenge through the incorporation of ncAAs in AMPs, expanding their functional diversity and potentially reducing resistance development.[29] To demonstrate how our language model could contribute to solving this real-world problem, we present a case study developing antimicrobial peptides.

In *Data set P*, 205 sequences are labeled as antimicrobial against *E. coli*. The weights pretrained on GPepT were fine-tuned with the 205 sequences to generate antimicrobial peptidomimetics. 500 sequences were generated using the fine-tuned model. First, we ranked all noncanonical elements according to the enrichment score, i.e., the fraction of generated sequences including the element divided by the fraction of sequences in *Data set P* including it. If the score is high, it implies that the element is preferred more after fine-tuning. Top five elements are shown in Figure 3. The top one's enrichment score is 33, suggesting that it is strongly related to antimicrobial activity.

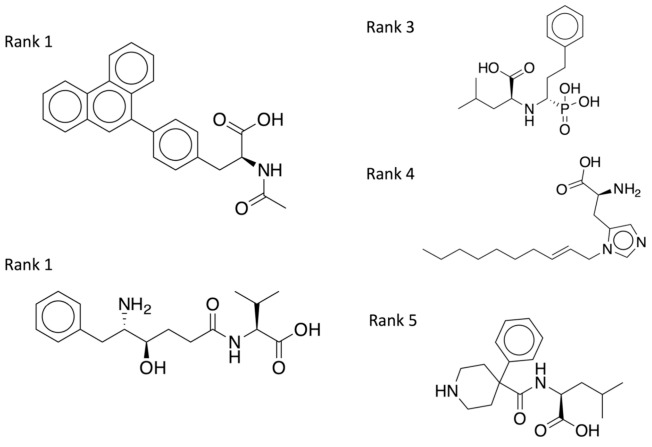


Figure 3. Top 5 enriched noncanonical elements. They correspond to the following tokens: X4857, X10616, X8507, X9886, X8517. Their enrichment scores are 33, 33, 30, 29, and 27. Only X4857 is found in PubChem.

For experimental validation, we selected five sequences of length 3−50 (Pep1−5; Table S1) based on the commercial availability of the included noncanonical elements and synthesizability. Pep1, 3, and 5 were successfully synthesized and purified to a degree suitable for antimicrobial testing. No unexpected or unusually high safety hazards were encountered. As shown in Figure 4a, Pep1 demonstrated potent antibacterial activity against *E. coli* with a minimum inhibitory concentration (MIC) of 50 μg/mL, effectively outperforming its canonical counterpart "WWWWWKZ0" (Z0 = Amide) (MIC > 100 μg/mL) (Figure 4a). The circular dichroism (CD) spectrum of Pep1, particularly the appearance of a distinct negative peak at 225 nm and a positive peak at approximately 235 nm, is atypical for canonical secondary structures and may suggest

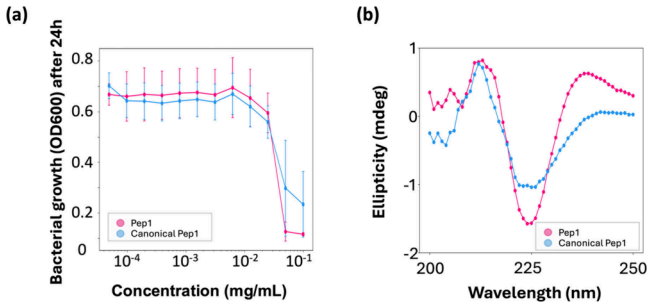


Figure 4. Experimental validation of GPepT-generated peptidomimetics Pep1. Canonical Pep1 refers to the peptide where noncanonical elements in Pep1 are replaced with canonical ones. a) Bacteria growth (OD600) after 24 h against peptide concentration. b) Circular dichroism spectra.

unique conformational features induced by the ncAA X556 (D-tryptophan). These peaks may arise from $\pi-\pi$ interactions involving the D-Trp residue or exciton coupling effects, which may contribute to increased local rigidity or to a functionally relevant structural orientation of the peptide. Such CD signatures in the 225−235 nm region have been associated with intramolecular exciton interactions involving aromatic side chains, as reported by Zsila.[30] While detailed structural elucidation would require further spectroscopic or computational studies, the observed features may indicate that ncAA incorporation stabilizes local conformations in a functionally relevant manner. Antimicrobial testing results for Pep3 and Pep5 are presented in Figure S3.

Our work has revealed the untapped potential of noncanonical elements in peptide drug discovery. The identification of 11,243 ncAAs represents a significant expansion of the peptide building block repertoire. This comprehensive tokenization of noncanonical elements addresses a critical gap in the field, enabling the systematic exploration of expanded peptide chemical space. Existing models such as PeptideBERT[11] can be expanded using our tokens of noncanonical elements.

We demonstrated a fine-tuning strategy to generate peptidomimetics for specific purposes. Pep1 outperformed its canonical counterpart in antimicrobial activity, but there is room for improvement. In fact, nearly 80% of our ncAAs that do not possess primary amine, making them incompatible with peptidomimetics synthesis using standard methods. This limitation underscores the gap between computational predictions and practical synthesis in peptide engineering. Nature enhances peptide diversity through post-translational modifications, suggesting that synthetic approaches could adopt similar strategies. For example, proline analogues have shown promise in modifying peptide backbones,[31] while in vitro ribosomal translation systems could allow the integration of D-, β-, or γ-amino acids, broadening access to our ncAA discoveries. Developing new synthesis techniques will be key to leveraging these computational findings in the real-world peptide engineering.

In conclusion, our work bridges computational peptide design with practical therapeutic development through the systematic exploration of noncanonical elements. While challenges in synthesis methods remain, our successful demonstration of a biologically active, language model-generated antimicrobial peptidomimetic validates this approach. As synthesis capabilities evolve, this expanded chemical

https://doi.org/10.1021/acsmedchemlett.5c00375
ACS Med. Chem. Lett. 2025, 16, 1670−1675

space promises to accelerate the development of peptide-based therapeutics with improved drug-like properties, offering new possibilities to address challenging therapeutic needs.

■ ASSOCIATED CONTENT

Data Availability Statement

The code of Monomerizer is available at https://github.com/tsudalab/Monomerizer. The noncanonical elements and the sequences of peptidomimetics are available at https://zenodo.org/records/14175750. GPepT is fully accessible on HuggingFace: https://huggingface.co/Playingyoyo/GPepT.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsmedchemlett.5c00375.

Algorithmic details of mMonomerizer, comparison of noncanonical amino acids (ncAAs), terminal modifications and canonical amino acids (cAAs) mined from ChEMBL, comparison of peptidomimetics and peptides mined from ChEMBL (Data set P),experimental details about peptide synthesis and measurement, valid peptidomimetics chosen for antimicrobial activity test, and bacteria growth (OD600) after 24 h against peptide concentration (Pep3 and Pep5) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Koji Tsuda − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba 305−0044, Japan; RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan;* orcid.org/0000-0002-4288-1606; Email: tsuda@k.u-tokyo.ac.jp

Authors

Yuna Oikawa − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*

Takanori Uzawa − *Emergent Bioengineering Materials Research Team, RIKEN Center for Emergent Matter Science, Wako, Saitama 351-0198, Japan; RIKEN Cluster for Pioneering Research, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0001-6042-513X

Francois Berenger − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*

Noriko Minagawa − *Emergent Bioengineering Materials Research Team, RIKEN Center for Emergent Matter Science, Wako, Saitama 351-0198, Japan*

Akiko Yumoto − *Emergent Bioengineering Materials Research Team, RIKEN Center for Emergent Matter Science, Wako, Saitama 351-0198, Japan*

Hideaki Takaku − *RIKEN Cluster for Pioneering Research, Wako, Saitama 351-0198, Japan*

Ryo Tamura − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba 305−0044, Japan; RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan;* orcid.org/0000-0002-0349-358X

Yoshihiro Ito − *RIKEN Cluster for Pioneering Research, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0002-1154-253X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsmedchemlett.5c00375

Author Contributions

Y.O., T.U., and K.T. conceived the idea and designed the research. Y.O., F.B., R.T., and K.T. developed the computational methods. Y.O., T.U., N.M., A.Y., and H.T. performed biological experiments. Y.I., R.W., and K.T. planned and supervised the study. All authors contributed to the preparation of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by JST ERATO JPMJER1903, JST CREST JPMJCR21O2 and MEXT JPMXP1122712807. We thank Yuya Takeda and Yuta Tomokiyo for their invaluable technical assistance.

■ ABBREVIATIONS USED

ncAA, noncanonical amino acid; GPepT, Generative pretrained Peptidomimetics Transformer; SMILES, Simplified Molecular Input Line Entry System; SMARTS, SMILES Arbitrary Target Specification; CD, Circular Dichroism; β_1, exponential decay rate for the first moment estimates in Adam optimizer; β_2, exponential decay rate for the second moment estimates in Adam optimizer; AMP, Antimicrobial peptide; t-SNE, t-distributed Stochastic Neighbor Embedding

■ REFERENCES

(1) Chen, C. H.; Lu, T. K. Development and Challenges of Antimicrobial Peptides for Therapeutic Applications. *Antibiotics* **2020**, *9* (1), 24.

(2) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic peptides: current applications and future directions. *Signal Transduct. Target. Ther.* **2022**, *7* (1), 48.

(3) Santos-Júnior, C. D.; Torres, M. D. T.; Duan, Y.; Rodríguez del Río, Á.; Schmidt, T. S. B.; Chong, H.; Fullam, A.; Kuhn, M.; Zhu, C.; Houseman, A.; et al. Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell* **2024**, *187* (14), 3761−3778.

(4) Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; dos Santos, C.; Chen, P.-Y.; et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **2021**, *5* (6), 613−623.

(5) Tran, D. P.; Tada, S.; Yumoto, A.; Kitao, A.; Ito, Y.; Uzawa, T.; Tsuda, K. Using molecular dynamics simulations to prioritize and understand AI-generated cell penetrating peptides. *Sci. Rep.* **2021**, *11* (1), 10630.

(6) Tučs, A.; Berenger, F.; Yumoto, A.; Tamura, R.; Uzawa, T.; Tsuda, K. Quantum Annealing Designs Nonhemolytic Antimicrobial Peptides in a Discrete Latent Space. *ACS Med. Chem. Lett.* **2023**, *14* (5), 577−582.

(7) Tucs, A.; Tran, D. P.; Yumoto, A.; Ito, Y.; Uzawa, T.; Tsuda, K. Generating Ampicillin-Level Antimicrobial Peptides with Activity-Aware Generative Adversarial Networks. *ACS Omega* **2020**, *5* (36), 22847−22851.

(8) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine learning designs non-hemolytic antimicrobial peptides. *Chem. Sci.* **2021**, *12* (26), 9221−9232.

(9) Murakami, Y.; Ishida, S.; Demizu, Y.; Terayama, K. Design of antimicrobial peptides containing non-proteinogenic amino acids using multi-objective Bayesian optimization. *Digit. Discovery* **2023**, *2* (5), 1347−1353.

(10) Szymczak, P.; Możejko, M.; Grzegorzek, T.; Jurczak, R.; Bauer, M.; Neubauer, D.; Sikora, K.; Michalski, M.; Sroka, J.; Setny, P.; et al.

Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.* **2023**, *14* (1), 1453.
(11) Guntuboina, C.; Das, A.; Mollaei, P.; Kim, S.; Barati Farimani, A. PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction. *J. Phys. Chem. Lett.* **2023**, *14* (46), 10427−10434.
(12) Ferruz, N.; Schmidt, S.; Höcker, B.; Ferruz, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Comm.* **2022**, *13* (1), 4348.
(13) Goettig, P.; Koch, N. G.; Budisa, N. Non-Canonical Amino Acids in Analyses of Protease Structure and Function. *Int. J. Mol. Sci.* **2023**, *24* (18), 14035.
(14) Simon, E.; Swanson, K.; Zou, J. Language models for biological research: a primer. *Nat. Methods* **2024**, *21* (8), 1422−1429.
(15) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100.
(16) Murugan, R. N.; Jacob, B.; Kim, E.-H.; Ahn, M.; Sohn, H.; Seo, J.-H.; Cheong, C.; Hyun, J.-K.; Lee, K. S.; Shin, S. Y.; et al. Non hemolytic short peptidomimetics as a new class of potent and broad-spectrum antimicrobial agents. *Bioorg. Med. Chem. Lett.* **2013**, *23* (16), 4633−4636.
(17) Craig, A. J.; Ermolovich, Y.; Cameron, A.; Rodler, A.; Wang, H.; Hawkes, J. A.; Hubert, M.; Björkling, F.; Molchanova, N.; Brimble, M. A.; et al. Antimicrobial Peptides Incorporating Halogenated Marine-Derived Amino Acid Substituents. *ACS Med. Chem. Lett.* **2023**, *14* (6), 802−809.
(18) Meena, C. L.; Thakur, A.; Nandekar, P. P.; Sharma, S. S.; Sangamwar, A. T.; Jain, R. Synthesis and biology of ring-modified l-Histidine containing thyrotropin-releasing hormone (TRH) analogues. *Eur. J. Med. Chem.* **2016**, *111*, 72−83.
(19) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373−D1380.
(20) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(21) Daylight Chemical Information Systems, I. *A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed June 1, 2025).
(22) *Chembl activities. ChEMBL Database*. European Bioinformatics Institute (EMBL-EBI). https://www.ebi.ac.uk/chembl/web_components/explore/activities/ (accessed Dec 15, 2023).
(23) *run_clm.py*; 2021. https://github.com/huggingface/transformers/blob/main/examples/tensorflow/language-modeling/run_clm.py (accessed May 1, 2025).
(24) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* 2014, arXiv.1412.6980.
(25) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(26) Murray, C. J. L.; Ikuta, K. S.; Sharara, F.; Swetschinski, L.; Aguilar, G. R.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **2022**, *399* (10325), 629−655.
(27) Fjell, C. D.; Hiss, J. A.; Hancock, R. E.; Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discovery* **2012**, *11* (1), 37−51.
(28) Nizet, V. Antimicrobial Peptide Resistance Mechanisms of Human Bacterial Pathogens. *Curr. Issues Mol. Bio.* **2006**, *8* (1), 11−26.
(29) Garg, N.; Oman, T. J.; Andrew Wang, T.-S.; De Gonzalo, C. V. G.; Walker, S.; van der Donk, W. A. Mode of action and structure-activity relationship studies of geobacillin I. *J. Antibiot.* **2014**, *67* (1), 133−136.
(30) Zsila, F. Far-UV circular dichroism signatures indicate fluorophore labeling induced conformational changes of penetratin. *Amino acids* **2022**, *54* (7), 1109−1113.
(31) Kubyshkin, V.; Davis, R.; Budisa, N. Biochemistry of fluoroprolines: the prospect of making fluorine a bioelement. *Beilstein J. Org. Chem.* **2021**, *17*, 439−460.

https://doi.org/10.1021/acsmedchemlett.5c00375
ACS Med. Chem. Lett. 2025, 16, 1670−1675