

MORE RELATED WORK

Anonymous authors

Paper under double-blind review

1 CANONICAL PEPTIDE PREDICTION

Machine learning models have been widely applied to predict therapeutically relevant properties of peptides, particularly canonical peptides. Existing approaches span classical machine-learning algorithms, graph neural networks (GNNs), recurrent neural networks (RNNs), Transformer architectures, and large protein language models (PLMs). Representative works include:

- **amPEPpy** (Lawrence et al., 2021) trains a Random Forest (RF) classifier on a diverse set of physicochemical descriptors to identify antimicrobial peptides, and employs a drop-column strategy to quantify feature importance.
- **HemoPI2** (Rathore et al., 2025) develops both classification and regression models to predict hemolytic peptides and their hemolytic concentration (HC_{50}). The classifier uses a hybrid RF-based approach incorporating motif features, whereas the regression model predicts HC_{50} based on protein language model embeddings.
- **PepExplainer** (Zhai et al., 2024) proposes an explainable graph neural network (GNN) framework for predicting and optimizing the bioactivity of macrocyclic peptides derived from selection-based libraries. It introduces a substructure-masking explanation (SME) mechanism to estimate atomic-level contributions and leverages transfer learning to improve IC_{50} prediction on enriched libraries.
- **AMPainter** (Dong et al., 2025) incorporates a high-accuracy antimicrobial activity predictor—HyperAMP, a hypergraph neural network (HGNN) model in which: nodes represent individual amino-acid residues, hyperedges represent 2-gram, 3-gram, or 4-gram contiguous fragments that correlate with AMP activity, and node features are 768-dimensional embeddings from the pretrained PLM Ankh, which outperforms alternatives such as ESM-1b and ProtT5.
- **APEX** (Wan et al., 2024) adopts a sequence-encoder + dual-decoder multi-task architecture. The encoder uses an RNN with attention to extract hidden sequence features—such as hydrophobic patterns or charge distribution—yielding a 566-dimensional representation. Its two decoders perform: regression (predicting MIC values against 34 bacterial species, with specific optimization for ESKAPE pathogens), and classification (distinguishing AMPs from non-AMPs using 5,093 AMPs and 5,500 non-AMPs curated from the DBAASP database).
- **AMPDesigner** (Wang et al., 2025b) trains a GPT-based large language model on peptide sequences, and shows that the resulting embeddings enable highly accurate MIC prediction and guide LLM-based peptide design.
- **PepBERT** (Du and Li, 2025) presents lightweight Transformer-based peptide language models trained from scratch on peptide-specific corpora. Available in small (1.86M) and large (4.9M) versions, PepBERT yields compact yet powerful peptide representations and outperforms or matches ESM-2 on eight of nine downstream peptide-property prediction tasks.
- **PeptideBERT** (Guntuboina et al., 2023) fine-tunes ProtBERT (Elnaggar et al., 2020) on peptide-only datasets and achieves state-of-the-art performance in predicting hemolysis and non-fouling properties.

2 NON-CANONICAL PEPTIDE PREDICTION

While the advances in ML approaches for canonical peptides highlight their effectiveness, modeling non-canonical peptides remains less explored, primarily due to the limited availability of data. Sev-

eral methods based on GNNs, pre-trained small-molecule language models, and transfer learning have been applied to address this challenge:

- **TransSAFP** (Liu et al., 2025) constructs a balanced antimicrobial activity prediction dataset across all 11 N-terminal types. The model undergoes pre-training on a public natural peptide dataset, followed by fine-tuning with the SAFP dataset for improved predictive performance.
- **PepLand** (Zhang et al., 2025) introduces a novel pre-training architecture using a multi-view heterogeneous graph neural network to analyze peptide stability, degradation resistance, and biomolecular interactions. It effectively predicts properties such as protein-protein interactions, permeability, solubility, and synthesizability.
- **PepDoRA** (Wang et al., 2024) is a unified peptide representation model that applies Weight-Decomposed Low-Rank Adaptation (DoRA) to fine-tune ChemBERTa-77M-MLM for peptide property prediction. It captures functional properties of both modified and unmodified peptides, enabling accurate predictions of membrane permeability, non-fouling behavior, hemolysis propensity, and target-specific binding.
- **PeptideCLM** (Feller and Wilke, 2025) introduces a peptide-specific chemical language model capable of encoding peptides with various chemical modifications, including non-canonical amino acids and cyclic structures. It outperforms models such as ChemBERTa and ChemBERTa-2, particularly in predicting the transmembrane diffusion of cyclic peptides.

3 CANONICAL SEQUENCE GENERATION

Peptide generation tasks can be broadly categorized into sequence generation and structure generation. The frameworks are typically unconditional, conditional, or reinforcement learning (RL)-based. Key high-quality works include:

- **Peptide-GPT** (Shah et al., 2024) is based on the ProtGPT2 architecture and is trained to generate peptides with specific bioactivities such as hemolytic activity, solubility, and anti-fouling behavior.
- **AMPDesigner** (Wang et al., 2025b) proposes a generative framework that integrates GPT, prompt tuning, knowledge distillation, and RL to ensure diversity and reduce computational cost, focusing on antimicrobial peptide generation.
- **PepDiffusion** (Wang et al., 2025d) uses a latent diffusion model combined with molecular dynamics to generate antimicrobial peptides optimized for activity and structural properties.
- **Evolgradient** (Wang et al., 2025a) treats activity as an optimizable objective and iteratively refines amino-acid sequences using gradient-guided optimization, similar to in-silico directed evolution.
- **AMPainter** (Dong et al., 2025) unifies peptide sequence optimization and generation within a deep reinforcement learning framework.
- **EvoPlay** (Wang et al., 2023) draws inspiration from self-play in Go, exploring the vast mutation space and using multiple structural/functional simulators to train an agent that enhances protein function through mutation.
- **PepMLM** (Chen et al., 2025) fine-tunes ESM-2 with a masked language modeling strategy by masking C-terminal peptide regions bound to target proteins, enabling de novo binder generation.
- **EvoBind2** (Li et al., 2025) designs linear and cyclic peptide binders using only target-protein sequences. A modified AlphaFold2 predicts complex structures, and the peptide is mutated iteratively to minimize interface distance while maximizing structural confidence (pLDDT).

4 NON-CANONICAL PEPTIDE GENERATION

Recent models extend beyond canonical amino acids to incorporate non-canonical residues and diverse chemical modifications:

- **PepTune** (Tang et al., 2025) is a masked diffusion language model trained to generate peptide SMILES sequences from scratch. It is pretrained on large datasets of non-canonical amino acids and cyclic peptides, and uses classifier-guided Monte Carlo tree search for multi-objective optimization (e.g., dual-target binding, solubility, membrane permeability).
- **GPepT** (Oikawa et al., 2025) creates a vocabulary of over 17,000 non-canonical building blocks mined from ChEMBL. A language model trained on this vocabulary generates peptidomimetic sequences with rich structural diversity.
- **HELM-GPT** (Zhang et al., 2012) is a GPT-based cyclic peptide generator using HELM notation. It incorporates reinforcement learning for property-directed optimization (e.g., transmembrane diffusion, KRAS binding), comparing ten predictive models and selecting RF/XGB for key properties.
- **PepINVENT** (Geylan et al., 2025) extends the REINVENT platform to non-natural amino-acid design, enabling de novo exploration of modified peptides via RL-based property optimization.
- **PepThink R1** (Wang et al., 2025c) combines LLMs, chain-of-thought supervision, and RL to directly reason over monomer-level substitutions with pharmacology-aware reward shaping.

REFERENCES

- Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pages 1–9, 2025.
- Ruihan Dong, Qiushi Cao, and Chen Song. Painting peptides with antimicrobial potency through deep reinforcement learning. *Advanced Science*, page e06332, 2025.
- Zhenjiao Du and Yonghui Li. Pepbert: Lightweight language models for peptide representation. *bioRxiv*, 2025. doi: 10.1101/2025.04.08.647838.
- Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020. doi: 10.1101/2020.07.12.199554.
- Aaron L Feller and Claus O Wilke. Peptide-aware chemical language model successfully predicts membrane diffusion of cyclic peptides. *Journal of Chemical Information and Modeling*, 65(2): 571–579, 2025.
- Gökçe Geylan, Jon Paul Janet, Alessandro Tibo, Jiazhen He, Atanas Patronov, Mikhail Kabeshov, Werngard Czechtizky, Florian David, Ola Engkvist, and Leonardo De Maria. Pepinvent: generative peptide design beyond natural amino acids. *Chemical Science*, 16(20):8682–8696, 2025.
- Chakradhar Guntuboina, Adrita Das, Parisa Mollaei, Seongwon Kim, and Amir Barati Farimani. Peptidebert: A language model based on transformers for peptide property prediction. *The Journal of Physical Chemistry Letters*, 14(46):10427–10434, 2023.
- Travis J Lawrence, Dana L Carper, Margaret K Spangler, Alyssa A Carrell, Tomás A Rush, Stephen J Minter, David J Weston, and Jessy L Labbé. ampeppy 1.0: a portable and accurate antimicrobial peptide prediction tool. *Bioinformatics*, 37(14):2058–2060, 2021.
- Qiuzhen Li, Efsthios Nikolaos Vlachos, and Patrick Bryant. Design of linear and cyclic peptide binders from protein sequence information. *Communications Chemistry*, 8(1):211, 2025.

- 162 Huayang Liu, Zilin Song, Yu Zhang, Bihan Wu, Dinghao Chen, Ziao Zhou, Hongyue Zhang, Sang-
163 shuang Li, Xinpeng Feng, Jing Huang, et al. De novo design of self-assembling peptides with
164 antimicrobial activity guided by deep learning. *Nature Materials*, pages 1–12, 2025.
- 165 Yuna Oikawa, Takanori Uzawa, Francois Berenger, Noriko Minagawa, Akiko Yumoto, Hideaki
166 Takaku, Ryo Tamura, Yoshihiro Ito, and Koji Tsuda. Gpept: A foundation language model for
167 peptidomimetics incorporating noncanonical amino acids. *ACS Medicinal Chemistry Letters*, 16
168 (8):1670–1675, 2025.
- 169 Anand Singh Rathore, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta, and Gajen-
170 dra PS Raghava. Prediction of hemolytic peptides and their hemolytic concentration. *Communi-
171 cations Biology*, 8(1):176, 2025.
- 172 Aayush Shah, Chakradhar Guntuboina, and Amir Barati Farimani. Peptide-gpt: Generative design of
173 peptides using generative pre-trained transformers and bio-informatic supervision. *arXiv preprint
174 arXiv:2410.19222*, 2024.
- 175 Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic
176 peptides with multi-objective-guided discrete diffusion. *ArXiv*, pages arXiv–2412, 2025.
- 177 Fangping Wan, Marcelo DT Torres, Jacqueline Peng, and Cesar de la Fuente-Nunez. Deep-learning-
178 enabled antibiotic discovery through molecular de-extinction. *Nature Biomedical Engineering*, 8
179 (7):854–871, 2024.
- 180 Beilun Wang, Peijun Lin, Yuwei Zhong, Xiao Tan, Yangyang Shen, Yi Huang, Kai Jin, Yan Zhang,
181 Ying Zhan, Dian Shen, et al. Explainable deep learning and virtual evolution identifies antimicro-
182 bial peptides with activity against multidrug-resistant human pathogens. *Nature Microbiology*, 10
183 (2):332–347, 2025a.
- 184 Jike Wang, Jianwen Feng, Yu Kang, Peichen Pan, Jingxuan Ge, Yan Wang, Mingyang Wang, Zhenx-
185 ing Wu, Xingcai Zhang, Jiameng Yu, Xujun Zhang, Tianyue Wang, Lirong Wen, Guangning Yan,
186 Yafeng Deng, Hui Shi, Chang-Yu Hsieh, Zhihui Jiang, and Tingjun Hou. Discovery of antimi-
187 crobial peptides with notable antibacterial potency by an llm-based foundation model. *Science
188 Advances*, 11(10):eads8932, 2025b. doi: 10.1126/sciadv.ads8932. Epub 2025 Mar 5.
- 189 Leyao Wang, Rishab Pulugurta, Pranay Vure, Yinuo Zhang, Aastha Pal, and Pranam Chatterjee.
190 Pepdora: A unified peptide language model via weight-decomposed low-rank adaptation. *arXiv
191 preprint arXiv:2410.20667*, 2024.
- 192 Ruheng Wang, Hang Zhang, Trieu Nguyen, Shasha Feng, Hao-Wei Pang, Xiang Yu, Li Xiao, and
193 Peter Zhiping Zhang. Pepthink-r1: Llm for interpretable cyclic peptide optimization with cot sft
194 and reinforcement learning. *arXiv preprint arXiv:2508.14765*, 2025c.
- 195 Yeji Wang, Minghui Song, Fujing Liu, Zhen Liang, Rui Hong, Yuemei Dong, Huaizu Luan, Xiaojie
196 Fu, Wenchang Yuan, Wenjie Fang, et al. Artificial intelligence using a latent diffusion model
197 enables the generation of diverse and potent antimicrobial peptides. *Science Advances*, 11(6):
198 eadp7171, 2025d.
- 199 Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng
200 Yang. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*,
201 5(8):845–860, 2023.
- 202 Silong Zhai, Yahong Tan, Cheng Zhu, Chengyun Zhang, Yan Gao, Qingyi Mao, Youming Zhang,
203 Hongliang Duan, and Yizhen Yin. Pepexplainer: An explainable deep learning model for
204 selection-based macrocyclic peptide bioactivity prediction and optimization. *European Journal
205 of Medicinal Chemistry*, 275:116628, 2024.
- 206 Ruochi Zhang, Haoran Wu, Chang Liu, Qian Yang, Yuting Xiu, Kewei Li, Ningning Chen, Yu Wang,
207 Yan Wang, Xin Gao, et al. Pepland: a large-scale pre-trained peptide representation model for a
208 comprehensive landscape of both canonical and non-canonical amino acids. *Briefings in Bioin-
209 formatics*, 26(4):bbaf367, 2025.
- 210 Tianhong Zhang, Hongli Li, Hualin Xi, Robert V Stanton, and Sergio H Rotstein. Helm: a hierar-
211 chical notation language for complex biomolecule structure representation. *Journal of Chemical
212 Information and Modeling*, 52:2796–2806, 2012.