生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.220878

Apr. 25, 2023, 39(4): 1815-1824

·生物技术与方法·

基于 BERT 与 Text-CNN 的抗菌肽识别方法

徐小放[1,2]，杨春德[1]，舒坤贤[3]，袁新普[4]，李默程[5]，朱云平[2*]，陈涛[2*]

1 重庆邮电大学计算机科学与技术学院，重庆 400065
2 军事科学院军事医学研究院生命组学研究所 国家蛋白质科学中心(北京)北京蛋白质组研究中心 蛋白质组学国家重点实验室，北京 102206
3 重庆邮电大学 大数据生物智能重庆市重点实验室，重庆 400065
4 解放军总医院 第一医学中心普通外科医学部，北京 102206
5 国防科技大学计算机学院 量子信息研究所兼高性能计算国家重点实验室，湖南 长沙 410073

徐小放, 杨春德, 舒坤贤, 袁新普, 李默程, 朱云平, 陈涛. 基于 BERT 与 Text-CNN 的抗菌肽识别方法[J]. 生物工程学报, 2023, 39(4): 1815-1824.
XU Xiaofang, YANG Chunde, SHU Kunxian, YUAN Xinpu, LI Mocheng, ZHU Yunping, CHEN Tao. An antibacterial peptides recognition method based on BERT and Text-CNN[J]. Chinese Journal of Biotechnology, 2023, 39(4): 1815-1824.

摘　要：抗菌肽(antimicrobial peptides, AMPs)广泛存在于生命体中，是一种具有广谱抗菌活性、免疫调节功能的小分子多肽。抗菌肽不易产生耐药性，适用范围广，具有极大的临床价值，是传统抗生素的有力竞争者。识别抗菌肽是抗菌肽研究领域中的重要研究方向，湿实验法在进行大规模抗菌肽识别时存在成本高、效率低、周期长等难点，计算机辅助识别法是抗菌肽识别手段的重要补充，如何提升准确率是其中的关键问题。蛋白质序列可以被近似地看作是由氨基酸组成的语言，运用自然语言处理(natural language processing, NLP)技术可能提取到丰富的特征。本文将自然语言处理领域中的预训练模型 BERT 和微调结构 Text-CNN 结合，对蛋白质语言进行建模，提供了开源可用的抗菌肽识别工具，并与已发表的 5 种抗菌肽识别工具进行了比较。结果表明，优化“预训练-微调”策略带来了准确率、敏感度、特异性和马修相关系数的整体提升，为进一步研究抗菌肽识别算法提供了新思路。

关键词：蛋白质；抗菌肽；语言模型；预训练

资助项目：国家重点研发计划(2021YFA1301603)
This work was supported by the National Key Research and Development Program of China (2021YFA1301603).
*Corresponding authors. E-mail: ZHU Yunping, zhuyunping@ncpsb.org.cn; CHEN Tao, taochen1019@163.com
Received: 2022-11-04; Accepted: 2023-02-17; Published online: 2023-03-02

An antibacterial peptides recognition method based on BERT and Text-CNN

XU Xiaofang[1,2], YANG Chunde[1], SHU Kunxian[3], YUAN Xinpu[4], LI Mocheng[5], ZHU Yunping[2*], CHEN Tao[2*]

1 The School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2 State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Institute of Lifeomics, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 102206, China
3 Chongqing Key Laboratory on Big Data for Bio-Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
4 Department of General Surgery, First Medical Center, Chinese PLA General Hospital, Beijing 102206, China
5 State Key Laboratory of High Performance Computing, Institute for Quantum Information, College of Computer, National University of Defense Technology, Changsha 410073, Hunan, China

Abstract: Antimicrobial peptides (AMPs) are small molecule peptides that are widely found in living organisms with broad-spectrum antibacterial activity and immunomodulatory effect. Due to slower emergence of resistance, excellent clinical potential and wide range of application, AMP is a strong alternative to conventional antibiotics. AMP recognition is a significant direction in the field of AMP research. The high cost, low efficiency and long period shortcomings of the wet experiment methods prevent it from meeting the need for the large-scale AMP recognition. Therefore, computer-aided identification methods are important supplements to AMP recognition approaches, and one of the key issues is how to improve the accuracy. Protein sequences could be approximated as a language composed of amino acids. Consequently, rich features may be extracted using natural language processing (NLP) techniques. In this paper, we combine the pre-trained model BERT and the fine-tuned structure Text-CNN in the field of NLP to model protein languages, develop an open-source available antimicrobial peptide recognition tool and conduct a comparison with other five published tools. The experimental results show that the optimization of the two-phase training approach brings an overall improvement in accuracy, sensitivity, specificity, and Matthew correlation coefficient, offering a novel approach for further research on AMP recognition.

Keywords: protein; antibacterial peptides; language model; pre-training

近年来，抗菌素耐药性(antimicrobial resistance, AMR)已经成为全世界共同面对的危机[1]。抗菌肽(antimicrobial peptides, AMPs)广泛分布于自然界中，是一种具有广谱抗菌活性、免疫调节功能的小分子多肽，且不易产生耐药性，是应对多重耐药菌的重要手段[2-6]。当前，抗菌肽的识别主要分为湿实验法与计算机辅助识别法。然而涉及蛋白质的湿实验法设计复杂、耗时较长且成本高昂[7-9]，使用湿实验法进行大规模抗菌肽识别具有一定挑战性。因此，计算机辅助识别方法对于抗菌肽识别任务具有重大的现实意义与实用价值。

通过计算机辅助识别方法，可以低成本筛选与识别抗菌肽。目前，已有数款基于信息学的

抗菌肽识别工具发表。iAMP-2L[10]基于伪氨基酸组成(pseudo amino acid composition, PseAAC)和模糊 K 近邻(fuzzy K-nearest neighbor, FKNN)算法构建了一个抗菌肽的多标签分类器；MAMPs-Pred[11]利用了随机森林(random forests, RF)等机器学习方法构建了一个抗菌肽识别工具并在样本欠平衡情况下进行了讨论。Youmans 等[12]和 AMPScan Vr.2 工具[13]均使用了深度学习相关的方法，其中 Youmans 等[12]引入了一个双向的长短期记忆(bidirectional long short-term memory, Bi-LSTM)神经网络用于抗菌肽识别，AMPScan Vr.2[13]则采取了将卷积神经网络(convolutional neural networks, CNN)和长短期记忆(long short-term memory, LSTM)神经网络相结合的方式。这些方法成功地减轻了湿实验识别的压力，但仍存在一定的改进空间。例如，有监督学习方法在面对无标签数据时稍显乏力。Zhang 等[14]首次将预训练模型 BERT[15]引入了抗菌肽识别领域，可以利用大量无标签数据进行预训练。

蛋白质序列可以由氨基酸字母串表示，因此可以类似地看作是一种蛋白质的语言[16-18]，运用自然语言处理(natural language processing, NLP)技术可能提取到更丰富的特征。Transformer[19]是在 NLP 领域中的重要创新，在多个 NLP 任务上达到了顶尖水准[20-23]。BERT 是基于 Transformer 的编码层(encoder)构建的一个双向表示编码器，进一步刷新了 SQuAD v1.1[24]等 11 个 NLP 任务上的最佳成绩[25-26]。本文将 BERT 与经过修改的 Text-CNN[27]相结合，进一步加强 NLP 技术与抗菌肽识别之间的联系，通过对微调(fine-tuning)层的改进，提升了抗菌肽计算机辅助识别方法的准确率，提供了一种准确可用的抗菌肽识别的方法。

1 材料与方法

1.1 数据集

预训练模型 BERT 通过学习大量无标签数据，降低了对有标签数据的依赖。本研究使用了由 Zhang 等[14]预训练的 BERT 模型以保证实验的公平性，该 BERT 模型由 UniProt[28]下载的约 55 万条蛋白质序列数据训练而成。由于抗菌肽识别领域中尚未形成金标准数据集，本文使用了 4 款已发表的抗菌肽识别工具(iAMP-2L[10]、MAMPs-Pred[11]、Bi-LSTM[12]和 AMPScan Vr.2[13])的数据集作为基准数据集，并在各个数据集上分别进行微调与测试。此外，本文提供了一个新的微调数据集，该数据集收集了上述 4 个数据集的负样本并从 DRAMP[29]中收集了 5 500 条抗菌肽作为正样本，经过数据整合与清洗，删除了重复和可能存在错误注释的样本。各数据集样本量如表 1 所示。TensorFlow[30]和 PyTorch[31]通用的可移植数据文件以及更详细的数据集信息可在 GitHub 项目(github.com/shallFun4Learning/BERT-CNN-AMP)中获取。

1.2 实验配置

本实验依托国家蛋白质科学中心(北京)的计算节点完成，具体配置如表 2 所示。

1.3 方法

本方法遵循“预训练-微调”架构，由预训练模型 BERT 与改进的 Text-CNN 组成。

Tenney 等[32]研究表明，BERT 可以对不同层次的信息进行建模，并具有消除歧义表示的能力；类似地，Clark 等[33]经过实验证明，BERT 具有抽取语法信息和关注重点词汇的能力。Zhang 等[14]利用 BERT 对蛋白质序列进行语言建模，驱使 BERT 学习氨基酸之间的相互信息，提供蕴含丰富特征的高维编码，以有利于抗菌肽识别等下游任务的训练。

表 1 样本量统计

Table 1 Statistics of sample

Datasets	Training set			Testing set		
	Positive	Negative	Total	Positive	Negative	Total
From iAMP-2L	879	879	1 758	920	920	1 840
From MAMPs-Pred	2 617	2 617	5 234	284	1 382	1 666
From AMPScan Vr.2	1 066	1 066	2 132	712	712	1 424
From Bi-LSTM	2 087	2 087	4 174	522	634	1 156
From DRAMP	4 310	1 271	5 581	1 067	328	1 395

表 2 实验环境

Table 2 Experimental environment

Name	Type
OS	CentOS Linux release 7.6.1810 (Core)
CPU	Intel(R) Xeon(R) Gold 6132 CPU @ 2.60 GHz
Memory	Samsung M393A4K40CB2-CVF 32GB
GPU	Tesla V100-PCIE-16GB-LS

BERT 对蛋白质语言进行建模的流程如图 1 所示：首先，对蛋白质序列数据进行预处理，用“[CLS]”和“[SEP]”作为开始与分割的特殊标记，随机将 15%的氨基酸词汇用“[MASK]”标记遮蔽；然后将由氨基酸组成的蛋白质序列转为张量(tensor)表示，并嵌入位置信息与分块信息；随后，传入由多个 Transformer 的编码层组成的 BERT，输出该蛋白质序列的高维表示；最后，通过自监督学习的方式促使 BERT 模型学习氨基酸词汇之间的关系。预训练阶段主要存在两个任务：(1) 根据双向的氨基酸信息对被遮蔽的词汇进行预测。(2) 判断一个氨基酸片段是否为另一个氨基酸片段的后续，即是否在同一条蛋白质序列中按顺序紧密相连。

“预训练-微调”结构在多项任务中展现了出色的性能[34-36]。Text-CNN 是一种结构简单而有效的微调架构。以蛋白质序列为例，传统 Text-CNN 的结构如图 2 所示：对单个氨基酸词汇的编码进行卷积得到特征图，再对特征图进行最大池化操作后传入全连接层计算，最后针对分类等具体的任务进行处理。该架构并非本方法选择的最终结构，出于保留编码中更多信息的考虑，本方法舍弃了池化层。

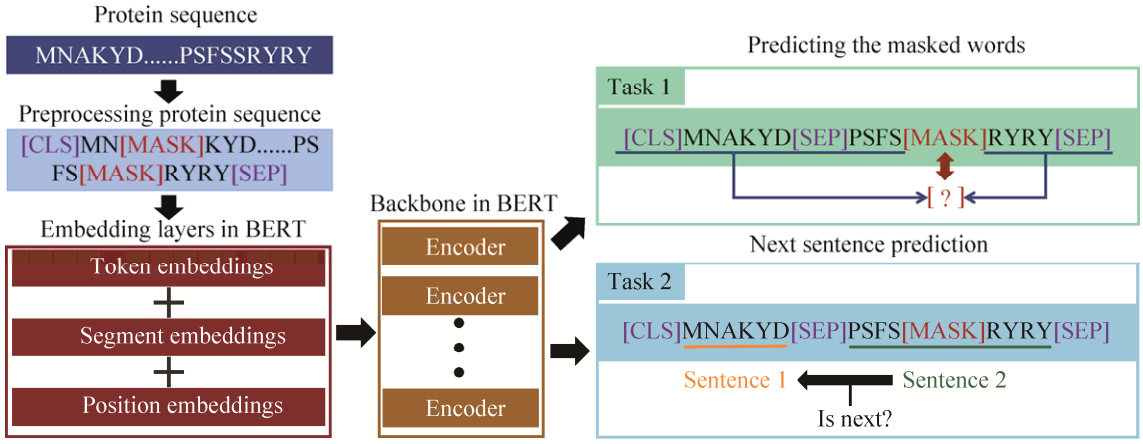

图 1 BERT 预训练流程

Figure 1 Overall pre-training for BERT.

本方法的整体流程如图 3 所示。BERT 发展初期，研究者们通常认为“[CLS]”标记会捕捉整个序列的信息[37-39]。因此，早期的分类任务中往往只使用“[CLS]”标记的信息，但这并非最佳实践[40-42]。本文改进了 Text-CNN 结构并对 BERT 的整体输出进行处理，以期获得更优秀的抗菌肽识别能力，具体流程如下：(1) 将目标肽段添加标记符后传入 BERT；(2) 采用 3 种不同尺寸的卷积核对 BERT 的输出进行卷积并拼接，卷积过程中采用复制(replicate)的方式进行填充(padding)；(3) 利用全连接层对拼接结果进行降维和分类处理。

本文涉及的代码和可用模型文件在 GitHub (github.com/shallFun4Learning/BERT-CNN-AMP)中提供下载。

2 结果与分析

Holdout 检验是一种常见的检验方法[43-45]，该方法将数据集切分为训练集和测试集进行检验。训练集用于模型的参数学习，测试集则用于检验模型的学习效果，模型在训练时应对测试集完全不可见。为了检验本方法的有效性，

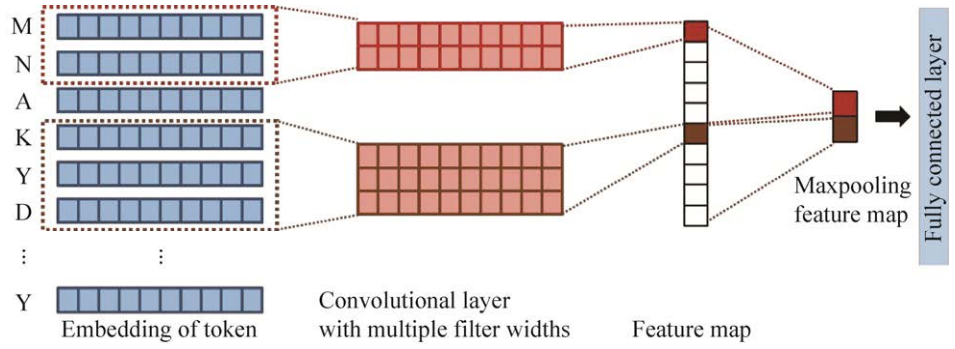

图 2 Text-CNN 结构图

Figure 2 The structure of Text-CNN.

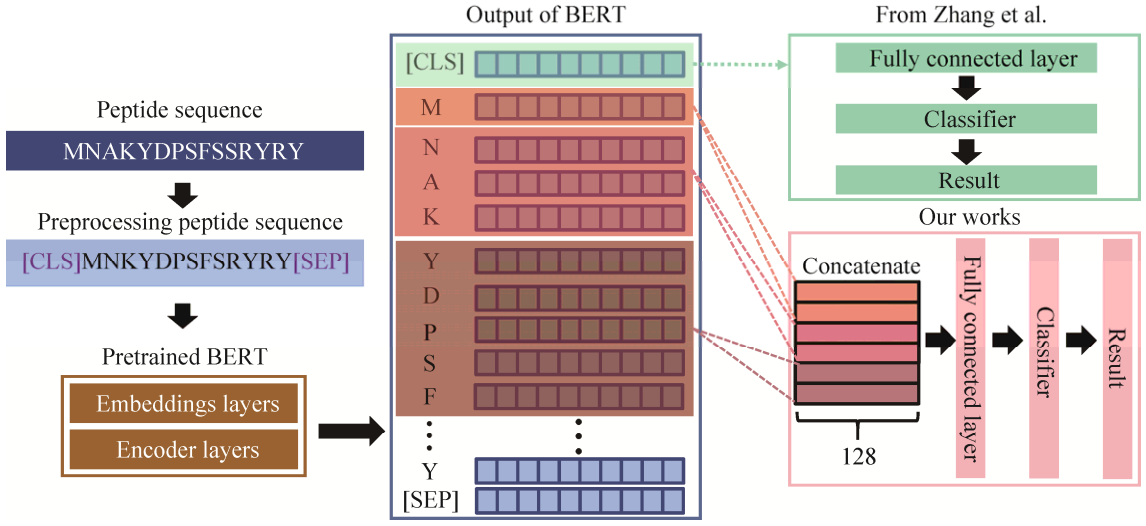

图 3 抗菌肽识别流程

Figure 3 The pipeline for AMP recognition.

本文与 Zhang 等[14]提出的方法在相同的条件下进行对比，分别在 5 个数据集(iAMP-2L[10]、MAMPs-Pred[11]、Bi-LSTM[12]、AMPScan Vr.2[13]和 DRAMP[29])上进行了独立测试。本实验使用与 Zhang 等[14]相同的数据集划分、学习率(learning rate)、训练周期(epochs)、批次大小(batch size)以及评估指标(表 3)，包括：敏感度(sensitivity, Sn)、特异性(specificity, Sp)、准确率(accuracy, Acc)和马修相关系数(Matthew correlation coefficient, Mcc)。敏感度和特异性分别指示对正负样本的识别能力，准确率代表整体上的识别能力，马修相关系数描述了算法的随机性，当其绝对值趋于 1 时，代表着预测结果与数据标签具有强相关性。

此外，本文引入了 Focal Loss[46]作为可选的损失函数。Focal Loss 可以平衡学习权重，赋予模型聚焦困难样本的能力。当数据集样本来源不同或分布不平均时，使用 Focal Loss 作为损失函数可以带来模型鲁棒性的提升，改善模型在特异性指标上的表现，满足组学时代对特异性指标的要求。

如表 4 所示，本实验在上述 5 个数据集上各进行 8 次重复实验并取平均值，标注为“/”的部分为原算法未提供相关指标。表中第 1 列指示数据集来源，第 2 列为所用的模型，第 3 列至第 6 列分别是准确率、敏感性、特异性和马修相关系数 4 个评价指标，各个数据集上的最佳指标用粗体展示。

结果表明，本文提出的模型囊括了所有的最佳指标，启用 Focal Loss 后特异性指标进一步得到优化。在普遍关心的准确率指标上，与原算法相比最高提升 5.86%，与 Zhang 等[14]相比最高提升 0.81%；在组学时代较为关注的特异性指标上，与原算法相比最高提升 10.49%，与 Zhang 等[14]相比最高提升 1.96%。本工作强调了蛋白质语言预训练模型应用到抗菌肽领域的实施性，说明了优化微调层的有效性，验证了利用 Focal Loss 提升特异性指标的可行性，提供了抗菌肽识别的新方法。

3 讨论

抗菌肽是传统抗生素的有力竞争者，具有诸多天然优势[47-48]。湿实验法成本较高、周期较长，难以支撑大规模的抗菌肽鉴定。因此，计算机辅助识别法对抗菌肽研究具有重要的现实意义和实用价值。

作为前置的辅助手段，如何提高准确率是计算机辅助识别法中的关键问题。本文就先进的 NLP 技术与抗菌肽识别算法的融合应用进行了讨论：(1) 改进了微调层，引入 Text-CNN 技

表 3 评估指标

Table 3 Evaluation indicators

Indicators	Formulae
Sensitivity (Sn)	$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
Specificity (Sp)	$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$
Accuracy (Acc)	$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$
Mathew correlation coefficient (Mcc)	$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}}$

表 4　各数据集微调后的模型性能

Table 4　Performance of the models trained on each dataset

Dataset	Model	Acc (%)	Sn (%)	Sp (%)	Mcc
From iAMP-2L[10]	iAMP-2L	92.23	97.72	86.84	0.844 6
	Zhang et al.[14]	97.28	98.88	95.46	0.945 3
	Our model	97.77	**98.91**	96.69	0.955 7
	Our model with Focal Loss	**98.09**	98.85	**97.33**	**0.961 8**
From MAMPs-Pred[11]	MAMPs-Pred	84.16	83.10	84.40	/
	Zhang et al.[14]	85.35	85.12	85.41	0.601 1
	Our model	85.78	**85.27**	85.71	**0.606 0**
	Our model with Focal Loss	**85.85**	84.23	**85.85**	0.605 0
From Bi-LSTM[12]	Bi-LSTM	94.98	/	/	0.899 0
	Zhang et al.[14]	95.22	94.91	95.50	0.904 2
	Our model	**95.66**	95.73	95.51	0.912 0
	Our model with Focal Loss	95.65	**95.79**	**95.54**	**0.912 4**
From AMPScan Vr.2[13]	AMPScan Vr.2	91.01	89.89	92.13	0.820 4
	Zhang et al.[14]	92.60	90.75	94.48	0.851 3
	Our model	**92.99**	**91.44**	94.53	**0.856 7**
	Our model with Focal Loss	92.72	90.90	**94.55**	0.855 1
From DRAMP[29]	Zhang et al.[14]	92.47	95.02	84.20	0.791 3
	Our model	**92.73**	**95.03**	85.24	**0.799 0**
	Our model with Focal Loss	92.52	94.47	**86.16**	0.795 2

The best performance achieved by each dataset on each indicator is given in boldface, and "/" means that this indicator is not provided.

术，取得了更高的准确性、敏感性、特异性和马修相关系数，可以有效减轻湿实验法的负担，降低识别抗菌肽的成本。(2) 引入了 Focal Loss 作为新的损失函数，进一步提升了组学时代较为关心特异性指标。(3) Zhang 等[14]提供的代码基于较早的 TensorFlow[30]版本开发，本文提供了 PyTorch[31]版本的开源代码和模型，丰富了研究途径，相关研究人员可以更专注于算法本身而不必过度关注深度学习框架。

着眼未来，改进预训练流程是一个值得探讨的方向。在当前版本中，BERT 的嵌入层(embedding layers)只添加了位置信息和分块信息，但氨基酸词汇与自然语言词汇之间仍然存在差异。例如，氨基酸词汇拥有相对分子质量信息，可以考虑通过嵌入层将相对分子质量信息进行嵌入，引入更多的先验知识。通过类似的方式或许可以填补蛋白质语言与自然语言的沟壑，引入更多维度的信息，从而带来性能上的提升。

REFERENCES

[1] ASLAM B, WANG W, ARSHAD MI, KHURSHID M, MUZAMMIL S, RASOOL MH, NISAR MA, ALVI RF, ASLAM MA, QAMAR MU, SALAMAT MKF, BALOCH Z. Antibiotic resistance: a rundown of a global crisis[J]. Infection and Drug Resistance, 2018, 11: 1645-1658.

[2] MAGANA M, PUSHPANATHAN M, SANTOS AL, LEANSE L, FERNANDEZ M, IOANNIDIS A, GIULIANOTTI MA, APIDIANAKIS Y, BRADFUTE S, FERGUSON AL, CHERKASOV A, SELEEM MN, PINILLA C, deLa FUENTE-NUNEZ C, LAZARIDIS T, DAI TH, HOUGHTEN RA, HANCOCK REW, TEGOS GP. The value of antimicrobial peptides in the age of resistance[J]. The Lancet Infectious Diseases, 2020, 20(9): e216-e230.

[3] BROWNE K, CHAKRABORTY S, CHEN RX, WILLCOX MD, STCLAIR BLACK D, WALSH WR, KUMAR N. A new era of antibiotics: the clinical potential of antimicrobial peptides[J]. International Journal of Molecular Sciences, 2020, 21(19): 7047.

[4] LEI J, SUN LC, HUANG SY, ZHU CH, LI P, HE J, MACKEY V, COY DH, HE QY. The antimicrobial peptides and their potential clinical applications[J]. American Journal of Translational Research, 2019, 11(7): 3919-3931.

[5] 于伟康, 张珊珊, 杨占一, 王家俊, 单安山. 超分子多肽自组装在生物医学中的应用[J]. 生物工程学报, 2021, 37(7): 2240-2255.
YU WK, ZHANG SS, YANG ZY, WANG JJ, SHAN AS. Application of supramolecular peptide self-assembly in biomedicine[J]. Chinese Journal of Biotechnology, 2021, 37(7): 2240-2255 (in Chinese).

[6] HUAN YC, KONG Q, MOU HJ, YI HX. Antimicrobial peptides: classification, design, application and research progress in multiple fields[J]. Frontiers in Microbiology, 2020, 11: 582779.

[7] MEDEMA MH, FISCHBACH MA. Computational approaches to natural product discovery[J]. Nature Chemical Biology, 2015, 11(9): 639-648.

[8] LI X, WU M, KWOH C-K, NG S-K.Computational approaches for detecting protein complexes from protein interaction networks: a survey[J]. BMC Genomics,2010, 11 (1): 10.1186/1471-2164-11-S1-S3.

[9] KÜKEN A, NIKOLOSKI Z. Computational approaches to design and test plant synthetic metabolic pathways[J]. Plant Physiology, 2019, 179(3): 894-906.

[10] XIAO X, WANG P, LIN WZ, JIA JH, CHOU KC. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types[J]. Analytical Biochemistry, 2013, 436(2): 168-177.

[11] LIN Y, CAI Y, LIU J, LIN C, LIU X. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies[J]. BMC Bioinformatics,2019, 20 (8): 10.1186/s12859-019-2766-9.

[12] YOUMANS M, SPAINHOUR C, QIU P. Long short-term memory recurrent neural networks for antibacterial peptide identification[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). November 13-16, 2017, Kansas City, MO, USA. IEEE, 2017: 498-502.

[13] VELTRI D, KAMATH U, SHEHU A. Deep learning improves antimicrobial peptide recognition[J]. Bioinformatics, 2018, 34(16): 2740-2747.

[14] ZHANG Y, LIN JY, ZHAO LM, ZENG XX, LIU XR. A novel antibacterial peptide recognition algorithm based on BERT[J]. Briefings in Bioinformatics, 2021, 22(6): bbab200.

[15] DEVLIN J, CHANG M, LEE K, TOUTANOVA K. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: ar Xiv: 1810. 04805. https://arxiv.org/abs/1810.04805.

[16] OFER D, BRANDES N, LINIAL M. The language of proteins: NLP, machine learning & protein sequences[J]. Computational and Structural Biotechnology Journal, 2021, 19: 1750-1758.

[17] TRAN NH, ZHANG X, XIN L, SHAN B, LI M. *De novo* peptide sequencing by deep learning[J]. Proceedings of the National Academy of Sciences, 2017, 114 (31): 8247-8252.

[18] QIAO R, TRAN NH, XIN L, CHEN X, LI M, SHAN BZ, GHODSI A. Computationally instrument-resolution-independent *de novo* peptide sequencing for high-resolution devices[J]. Nature Machine Intelligence, 2021, 3(5): 420-425.

[19] VASWANI A, SHAZEER N, PARMAR N, USZKOREIT J, JONES L, GOMEZ AN, KAISER Ł, POLOSUKHIN I. Attention is all You need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[20] GILLIOZ A, CASAS J, MUGELLINI E, ABOU KHALED O. Overview of the transformer-based models for NLP tasks[C]//Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems. September 6-9, 2020. IEEE, 2020: 179-183.

[21] SINGH S, MAHMOOD A. The NLP cookbook: modern recipes for transformer based deep learning architectures[J]. IEEE Access, 2021, 9: 68675-68702.

[22] HUANG XN, BI N, TAN J. Visual transformer-based models: a survey[M]//Pattern Recognition and Artificial Intelligence. Cham: Springer International Publishing, 2022: 295-305.

[23] BIAN ZD, LI SG, WANG W, YOU Y. Online evolutionary batch size orchestration for scheduling deep learning workloads in GPU clusters[C]// Proceedings of the International Conference for High Performance Computing, Networking, Storage and

Analysis. November 14-19, 2021, St. Louis, Missouri. New York: ACM, 2021: 1-15.

[24] RAJPURKAR P, ZHANG J, LOPYREV K, LIANG P. Squad: 100 000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.

[25] ACHEAMPONG FA, NUNOO-MENSAH H, CHEN WY. Transformer models for text-based emotion detection: a review of BERT-based approaches[J]. Artificial Intelligence Review, 2021, 54(8): 5789-5829.

[26] KOROTEEV M.BERT: a review of applications in natural language processing and understanding[J]. arXiv preprint arXiv:2103.11943,2021.

[27] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746-1751.

[28] CONSORTIUM TU. UniProt: a worldwide hub of protein knowledge[J]. Nucleic Acids Research, 2019, 47(D1): D506-D515.

[29] SHI GB, KANG XY, DONG FY, LIU YC, ZHU N, HU YX, XU HM, LAO XZ, ZHENG H. DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides[J]. Nucleic Acids Research, 2022, 50(D1): D488-D496.

[30] ABADI M. TensorFlow: learning functions at scale[C]//Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming. September 18-24, 2016, Nara, Japan. New York: ACM, 2016: 1.

[31] PASZKE A, GROSS S, MASSA F, LERER A, BRADBURY J, CHANAN G, KILLEEN T, LIN Z, GIMELSHEIN N, ANTIGA L.Pytorch: an imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 2019, 32.

[32] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4593-4601.

[33] CLARK K, KHANDELWAL U, LEVY O, MANNING CD. What does bert look at? An analysis of bert's attention[C]//Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2019: 276-286.

[34] TIAN S, QI P, XUE HJ, SUN Y. C-BERTT: a BERT-based model for extractive summarization of Chinese online judge questions[C]//2021 Ninth International Conference on Advanced Cloud and Big Data (CBD). March 26-27, 2022, Xi'an, China. IEEE, 2022: 127-132.

[35] JIANG XC, SONG C, XU YC, LI Y, PENG YL. Research on sentiment classification for netizens based on the BERT-BiLSTM-TextCNN model[J]. PeerJ Computer Science, 2022, 8: e1005.

[36] LU DM. daminglu123 at SemEval-2022 task 2: using BERT and LSTM to do text classification[C]// Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Seattle, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 186-189.

[37] WANG B, KUO CC J. SBERT-WK: a sentence embedding method by dissecting BERT-based word models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2146-2157.

[38] SUN SQ, CHENG Y, GAN Z, LIU JJ. Patient knowledge distillation for BERT model compression[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4323-4332.

[39] CHOI H, KIM J, JOE S, GWON Y. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks[C]//2020 25th International Conference on Pattern Recognition (ICPR). January 10-15, 2021, Milan, Italy. IEEE, 2021: 5482-5487.

[40] KIM T, YOO KM, LEE SG. Self-guided contrastive learning for BERT sentence representations[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2528-2540.

[41] REIMERS N, GUREVYCH I, REIMERS N, GUREVYCH I, THAKUR N, REIMERS N, DAXENBERGER J, GUREVYCH I, REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-Networks[C]//Proceedings of the

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3982-3992.

[42] JIANG ZY, TANG R, XIN J, LIN J. How does BERT rerank passages? An attribution analysis with information bottlenecks[C]//Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Punta Cana, Dominican Republic. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 496-509.

[43] PAL K, PATEL BV. Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques[C]//2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). March 11-13, 2020, Erode, India. IEEE, 2020: 83-87.

[44] van der VELDEN BHM, JANSE MHA, RAGUSI MAA, LOO CE, GILHUIJS KGA. Volumetric breast density estimation on MRI using explainable deep learning regression[J]. Scientific Reports, 2020, 10: 18095.

[45] OLIVEIRA M, TORGO L, SANTOS COSTA V. Evaluation procedures for forecasting with spatiotemporal data[J]. Mathematics, 2021, 9(6): 691.

[46] LIN TY, GOYAL P, GIRSHICK R, HE KM, DOLLÁR P. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. IEEE, 2017: 2999-3007.

[47] LI WY, SEPAROVIC F, O'BRIEN-SIMPSON NM, WADE JD. Chemically modified and conjugated antimicrobial peptides against superbugs[J]. Chemical Society Reviews, 2021, 50(8): 4932-4973.

[48] MORETTA A, SCIEUZO C, PETRONE AM, SALVIA R, DARIO MANNIELLO M, FRANCO A, LUCCHETTI D, VASSALLO A, VOGEL H, SGAMBATO A, FALABELLA P. Antimicrobial peptides: a new hope in biomedical and pharmaceutical fields[J]. Frontiers in Cellular and Infection Microbiology, 2021, 11: 668632.

(本文责编 郝丽芳)