基于文本卷积神经网络模型的抗菌药物发现

姚明丽[1]，高丁佳[2]，张洁[3]，李珊[2]，吴松[3]，司鑫鑫[1*]，夏杰[3*]（1. 江苏海洋大学药学院，江苏 连云港 222005；2. 南京航空航天大学经济与管理学院，南京 210016；3. 中国医学科学院北京协和医学院药物研究所，北京 100050）

摘要：目的 基于文本卷积神经网络（Text-Convolutional Neural Network，Text-CNN）算法，构建抗金黄色葡萄球菌（*Staphylococcus aureus*）活性的预测模型，通过虚拟筛选发现具有抑制 *S. aureus* 活性的苗头化合物。**方法** 从 ChEMBL 数据库中收集并整理了 26327 个标注有 *S. aureus* 活性数据的化合物，通过随机采样建立 10 组训练集和测试集，采用 Text-CNN 算法建立 10 个模型，通过模型评估选择性能最佳的模型，对该模型进行 Y-随机化检验和应用域分析。使用该模型虚拟筛选内部化合物库，确定潜在的抗菌化合物，并采用微量肉汤稀释法测定化合物的抗 *S. aureus* 活性。**结果** 名为 Text-CNN3 的机器学习模型具有良好的分类性能，该模型对于测试集的马修斯相关系数为 0.573，ROC 曲线下面积为 0.881。基于该模型的虚拟筛选和抗菌活性测试，发现了两个抗菌活性化合物 Y5 和 Y7，其对 *S. aureus* 的最低抑菌浓度（minimal inhibitory concentration，MIC）分别为 8 和 4 μg·mL^{-1}。**结论** 本研究建立的 Text-CNN3 模型可有效发现抗 *S. aureus* 化合物，所发现的苗头化合物 Y5 和 Y7 有进一步研究的意义和价值。

关键词：金黄色葡萄球菌；文本卷积神经网络；活性预测；最低抑菌浓度

doi：10.11669/cpj.2024.03.008　　**中图分类号：**R965.1　　**文献标志码：**A　　**文章编号：**1001－2494（2024）03－0249－07

Text-Convolutional Neural Network-based Discovery of Antibacterial Agents

YAO Mingli[1], GAO Dingjia[2], ZHANG Jie[3], LI Shan[2], WU Song[3], SI Xinxin[1*], XIA Jie[3*] (1. *School of Pharmacy, Jiangsu Ocean University, Lianyungang* 222005, *China*; 2. *School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*; 3. *Institute of Materia Medica, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing* 100050, *China*)

ABSTRACT: OBJECTIVE To build a text-convolutional neural network (Text-CNN)-based prediction model for anti-*Staphylococcus aureus* (*S. aureus*) activity and identify anti-*S. aureus* hits by virtual screening. **METHODS** A dataset containing 26327 compounds annotated with *S. aureus* activity data was collected and curated from the ChEMBL database. Ten pairs of training and test sets were generated by random partition for 10 times and then 10 models were built using the Text-CNN algorithm. The best-performing model was determined by model evaluation and further studied by Y-randomization test and applicability domain analysis. Following that, the best-performing model was used to virtually screen the in-house chemical library, by which the potential antibacterial agents were determined. The micro-broth dilution method was used to test anti-*S. aureus* activity of the potential hits. **RESULTS** The machine-learning model (named Text-CNN3) performed well in classification. Evaluated on the test set, its Mathews correlation coefficient was 0.573 and the area under the ROC curve was 0.881. With this model for virtual screening as well as antibacterial screening, compounds Y5 and Y7 were identified as antibacterial compounds, with minimum inhibitory concentrations (MIC) of 8 and 4 μg·mL^{-1}, respectively. **CONCLUSION** The Text-CNN3 model in this study is effective to identify anti-*S. aureus* compounds, while the antibacterial hits Y5 and Y7 are worthy of further study.

KEY WORDS: *Staphylococcus aureus*; Text-CNN; activity prediction; minimum inhibitory concentration

抗生素耐药性（antibiotic resistance）给人类健康带来巨大威胁，WHO 已经将其列为全球公共卫生的严重威胁之一。当前，抗生素耐药性逐步演变成了抗生素危机，全球每年有 70 万人死于该危机[1-2]。据推测，如果在抗生素研发方面没有显著进展，预计到 2050 年每年将有 1 000 万人死亡[3]。为应对抗生素耐药性，新药研发至关重要。众所周知，抗菌药物的研发是一个周期长、耗资高且成功率低的过程[4]。近 20 年来，全球范围内获批上市的抗菌药物品种仅有三种新结构类型，即恶唑

基金项目：中国医学科学院医学与健康科技创新工程重大协同创新项目资助（2021-I2M-1-069）

作者简介：姚明丽，女，硕士研究生　　研究方向：分子信息学与药物设计　　* **通讯作者：**司鑫鑫，女，博士，副教授　　研究方向：分子药理学；夏杰，男，博士，副研究员　　研究方向：分子信息学与药物设计

烷酮类(如利奈唑胺)、脂肽类(如达托霉素)、截短侧耳素类(如瑞他莫林)。其余上市的抗菌药物(如非达霉素、康替唑胺、苹果酸奈诺沙星)和正处于临床开发阶段的抗菌药物中,大部分药物属于以上类型或者临床使用的抗菌药物的衍生物[5]。相同类型的抗生素易产生交叉耐药,无法从根本上克服抗生素耐药性,因此亟需发现新结构类型的抗菌药物。

虚拟筛选技术是快速发现新结构类型的抗菌药物先导结构的有效手段。近年来,随着人工智能技术的快速发展,越来越多的研究人员开始采用该技术方法建立抗菌活性预测模型并开展虚拟筛选研究[6-9],发现了多个抗菌化合物(图 1)。比如,Wang 等[6]使用朴素贝叶斯、支持向量机、递归拆分和 k-最近邻法,基于理化性质描述符和分子指纹构建了近千个机器学习模型,并将性能最佳的模型应用于抗金黄色葡萄球菌(*Staphylococcus aureus*)化合物的虚拟筛选,并发现化合物 **C1** 和 **C2** 对 *S. aureus* 表现出较好的抗菌活性,其最低抑菌浓度(minimal inhibitory concentration,MIC)范围 4 ~ 16 μg · mL^{-1}。Stokes 等[7]采用有向消息传递深度神经网络(directed message passing nerural network, D-MPNN)建立了抗菌活性预测模型,并对 ZINC15 化合物库进行预测和筛选,发现了具有抗 *S. aureus* 活性的化合物 **C3** 和 **C4**,其 MIC 分别为 2 和 0.25 μg · mL^{-1}。

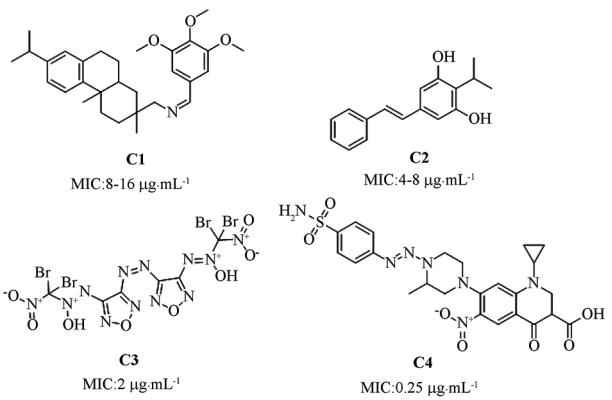


图 1 基于机器学习发现的具有抗金黄色葡萄球菌活性的代表性化合物

Fig. 1 Representative anti-*Staphylococcus aureus* compounds discovered by machine learning

2014 年,Kim 等[10]首次提出文本卷积神经网络(text-convolutional neural network,Text-CNN)算法并将其运用到文本分类任务中,取得了较好的应用效果。Text-CNN 算法的出现,使得将药物分子的简化分子线性输入规范(simplified molecular input line entry system,SMILES)表示作为输入建立活性预测模型成为可能。文献检索结果表明,目前还没有采用 Text-CNN 模型进行虚拟筛选并发现抗菌化合物的报道。基于此,本研究通过收集标注有 *S. aureus* 活性数据的化合物,对数据进行预处理后,采用 Text-CNN 算法构建了 10 个 Text-CNN 模型,随后通过模型评价确定性能最佳模型用于虚拟筛选,并进行体外抗菌活性测试,发现了具有抗 *S. aureus* 活性的苗头化合物。

1 材料和方法

1.1 Text-CNN

1.1.1 数据收集和整理 从 ChEMBL 29 数据库(https://www.ebi.ac.uk/chembl/)下载标注有 *S. aureus* 最低抑菌浓度的化学生物学数据 102,891 条。使用 RDKit(2019.09.3 版)和 MolVS(0.1.1 版)去盐、中和、SMILES 标准化[11]。仅保留具有确切 MIC 值的化合物,同一个化合物若存在多条 MIC 数据则取其平均值。根据类药性(drug-likeness)原则,仅保留相对分子质量不大于 600 且可旋转键不多于 20 的化合物。根据 MIC 值是否小于 32 μg · mL^{-1},将化合物标记为活性化合物和非活性化合物。随机抽取 80% 数据作为训练集,另外 20% 数据作为测试集。该随机拆分过程重复 10 次,获得 10 对训练集和测试集。

1.1.2 化合物 SMILES 表示及分词 化合物结构采用 SMILES 表示,随后对 SMILES 进行原子级分词(atom-level tokenization)提取词元(token)。具体操作包括:①将多字符元素符号(例如“Cl”及“Br”)视为单独的词元;②将用方括号表示的特殊符号(例如[O-]、[nH]等)也视为词元;③将剩余单个字符的原子视为词元。

首先,对 ChEMBL 数据集内的小分子化合物进行原子级分词,共得到 256 个词元并构成分词词典。同样地,将 *S. aureus* 数据集内的小分子化合物的 SMILES 进行分词得到词元,根据分词词典用词元的 ID 进行编码。经统计,本研究中的 *S. aureus* 数据集内小分子化合物 SMILES 的词元的数量最多为 91。为保持所有小分子化合物 SMILES 的编码长度一致,使用 keras(https://keras.io/)的 pad_sequences 函数将 SMILES 编码后填充至 91 个词元长度,用作 Text-CNN 模型的输

入。即对于长度不足的在末尾补0，而对于长度超过91的则进行截断。

1.1.3 网络基本架构 本研究采用的卷积神经网络架构如图2所示[12]。input为输入层，用于接受经编码后的SMILES张量；embedding为词嵌入层，将input层传递的输入转换到预定义的向量空间；conv_0、conv_1、conv_2都是一维卷积层，每一层设置若干个卷积核，通过卷积操作进行特征提取；随后，经过batch_normalization、batch_normalization_1、batch_normalization_2进行批规范化，避免梯度消失；然后，分别经过maxpool_0、maxpool_1、maxpool_2池化层进行最大池化，即对特征求最大值；concatenate层则对前面经卷积、批规范化、池化后得到的特征进行合并连接；dropout层通过设置dropout率，冻结部分权重，防止过拟合；最后，dense层通过接收特征向量作为输入，并采用合适的激活函数进行分类。本研究使用TensorFlow中集成的Keras API建立Text-CNN模型。

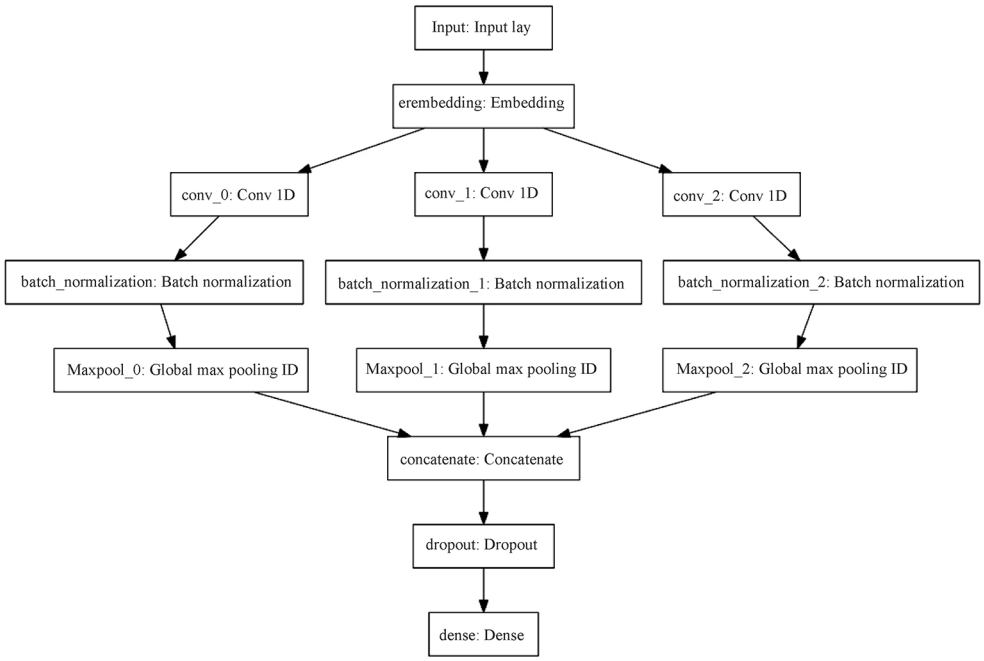


图2 文本卷积神经网络(Text-CNN)的基本架构

Fig. 2 The infrastructure of Text-CNN

1.1.4 模型超参数设置 如上所述，所有化合物的SMILES的词元长度被固定为91，因此此处设置输入大小为91。词嵌入层输出维度设置为50，即SMILES输入模型后，词嵌入层实现将维数为词典大小(256维)的输入映射到维数为50的实值向量。

如图2所示，本模型共设置3种一维卷积层，卷积核大小分别设置为2、3、4，步长均设置为1，激活函数为ReLU，且每种卷积核的个数设置为256个。在经过一维卷积层进行特征提取、批规范化和最大池化操作后，对特征进行合并连接，然后使用SoftMax函数进行二分类，输出两种可能状态，即有抗菌活性或无活性。模型损失函数采用sparse_categorical_crossentropy，优化器采用Adam算法，模型训练的迭代次数设置为40。

除此以外的超参数是采用10折交叉验证以及网格搜索来确定的。具体而言，无效神经元的比例(dropout rate)的搜索范围：0.1、0.2、0.3，批大小(batch size)的搜索范围为32、64、128，学习率(learning rate)的搜索范围为0.000 1、0.001、0.01。对于每一组超参数的组合，计算10次训练所得到的准确度的平均值。通过比较不同超参数组合各自在10折交叉验证中的准确度确定最佳超参数。

1.1.5 模型性能评估 本研究通过灵敏度(sensitivity，SE)、特异性(specificity，SP)[13]、马修斯相关系数(Matthews correlation coefficient，MCC)[14]以及受试者工作特征曲线(receiver operating characteristic，ROC)曲线下面积(area under the curve，AUC)[15]这四个指标对模型的性能进行评估。其

中，SE 又称真阳性率，是指实际为阳性的样本判断为阳性的百分比。SP 又称为真阴性率，是实际为阴性的样本判断为阴性的比例。真阳性（true positive，TP）代表活性化合物数量，假阳性（false positive，FP）表示将非活性化合物预测为活性化合物的数量，真阴性（true negative，TN）代表非活性化合物的数量，假阴性（false negative，FN）表示将活性化合物预测为非活性化合物的数量，MCC 则是以上 4 个方面的综合指标。AUC 代表阳性样本排在阴性样本前的概率，其值越高，说明模型对于阳性样本的排序效果越好，具体定义见公式 1 ~3：

$$SE = \frac{TP}{TP + FN} \quad \text{公式（1）}$$

$$SE = \frac{TN}{TN + FP} \quad \text{公式（2）}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{公式（3）}$$

1.1.6 Y-随机化检验　Y-随机化检验（Y-randomization test）用于验证模型的鲁棒性[16]。在本研究中，Y 随机化检验共进行 3 次。在每次检验中，保持训练集中的特征数据不改变，将训练集中活性和非活性标签进行随机打乱，从而生成一个新的训练集。然后，通过 10 折交叉验证以及网格搜索确定最优超参数后重新训练模型。在模型性能评估中，使用和原模型相同的测试集。如果重新训练得到的模型性能远不如原模型的性能，那么原模型的预测性能就不具有偶然性。

1.1.7 适用域分析　基于 MACCS 指纹计算测试集中每个化合物与训练集中任意化合物的谷本系数（tanimoto coefficient，*Tc*）。通过设置 *Tc* 阈值（从 1 到 0.5，逐步降低 0.05），逐步剔除测试集中与训练集化合物相似度超过阈值的化合物，使用模型预测排除相似结构后的测试集内化合物的活性，计算 AUC 值，由此研究模型性能随着多样性增加的变化趋势。

1.2 虚拟筛选

选取性能最好的抗菌活性预测模型虚拟筛选课题组内部的化合物库（含 435 个化合物），保留预测为有抗菌活性的化合物。通过计算每个预测有抗菌活性的化合物与上文所述的 *S. aureus* 数据集内活性化合物的相似度（即基于 MACCS 的 *Tc* 值）[17]，若与已知活性化合物两两之间高度相似（*Tc* 不小于 0.75 的分子被认为是已知活性化合物的类似物[18-20]），则从列表中删除该化合物。使用 Discovery Studio 软件（v16.1.0）中的“Cluster Ligands”模块根据 FCFP_6 指纹[21]将分子聚成 10 簇。最终，考虑化合物的可合成性等因素，挑选一定数量的化合物进行体外抗菌活性测试。

1.3 体外抗菌活性评价

采用微量肉汤稀释法[22]对化合物的体外抗菌活性进行评价，采用的代表性 *S. aureus* 为标准菌株 ATCC29213，活性指标为 MIC。

将 96 孔培养板灭菌后，第 1 列中加入 200 μL 的供试菌液，剩余 11 列都加 100 μL 菌液。然后将待测样品用 DMSO 溶解配制浓度至 1.6 mg·mL^{-1}，取 4 μL 溶液加入至含有 200 μL 菌液的孔里，采用 2 倍梯度稀释，将第 1 孔中含有化合物的菌液逐个加入到其余各孔（含 100 μL 菌液）中，各孔中的化合物终浓度分别为 32、16、8、4、2、1、0.50、0.25、0.125、0.062 5、0.031 25、0.015 625 μg·mL^{-1}。将 96 孔板转移至 37 ℃培养箱内培养 16 ~20 h。培养结束后，将 96 孔板置于超净台内，观察每个孔内细菌的生长情况，从第 1 个孔至最后一个孔，未见细菌生长的首个孔所对应的化合物浓度即为该化合物的 MIC。实验采用 3 复孔，设置空白对照组和阳性对照组（左氧氟沙星）。

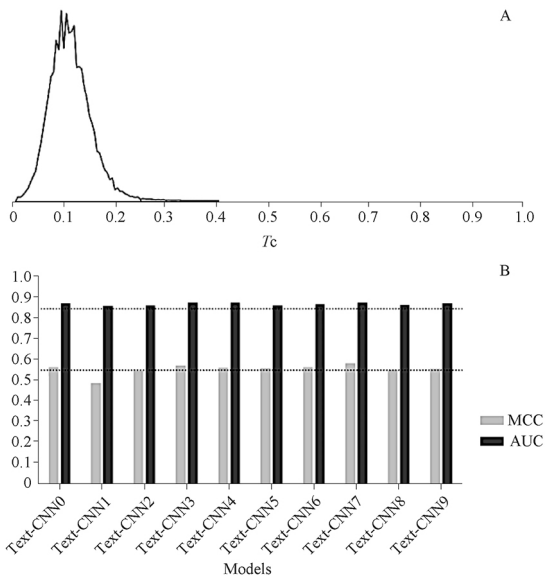
2 结　果

2.1 Text-CNN 模型

2.1.1 *S. aureus* 建模数据集　为了保持数据集内活性与非活性化合物的平衡，本研究设置 32 μg·mL^{-1}作为阈值。为此，抗菌活性化合物和非抗菌活性化合物的数量分别为 18 489 和 7 838。图 3A 显示了基于 Morgan2 指纹所计算的化合物两两间的结构相似度值 *Tc* 的分布。由该图可见，由于大多数化合物之间的 *Tc* 值小于 0.25，该 *S. aureus* 建模数据集具有高度的化学结构多样性。

2.1.2 模型的分类性能　如方法所述，本研究通过随机采样获得了 10 组训练集和测试集，随后针对每一组训练-测试集的数据，采用 Text-CNN 算法建立抗 *S. aureus* 活性预测模型。表 1 和图 3B 展示了所建立的 10 个模型的超参数和在测试集上的性能指标。所有模型的 MCC 范围为 0.485 ~0.582，AUC 值范围为 0.863 ~0.881，表明各模型的分类性能相差不大。其中，基于第 4 对训练-测试集所建立的名为 Text-CNN3 的分类模型性能最好，其 AUC 值为 0.881。建立该模型所采用的超参数为：批量大小为 32，无效神经元的

比例为 0.1,学习率为 0.001。



A – 基于 Morgan2 指纹的两两结构相似度(Tanimoto 系数,*Tc*)分布;B – 10 个 Text-CNN 模型在测试集上的评估分类性能:马修斯相关系数(MCC)和受试者工作特征曲线下面积(ROC AUC)。

A – distribution of pairwise structural similarity (Tanimoto coefficient, *Tc*) based on the Morgan2 fingerprints; B – the Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (ROC AUC) of 10 text-CNN models evaluated on test sets.

图 3 金黄色葡萄球菌建模数据集及模型的分类性能

Fig. 3 The *S. aureus* modeling set and the classification performance of the models

表 1 每个模型的超参数和在测试集上的性能指标

Tab. 1 The hyperparameters of each model and classification performance on the test set

Models	Hyperparameters			SE	SP	MCC	AUC
	Batch size	Dropout rate	Learning rate				
Text-CNN0	64	0.1	0.000 1	0.849	0.732	0.564	0.878
Text-CNN1	64	0.1	0.000 1	0.953	0.439	0.485	0.863
Text-CNN2	64	0.1	0.000 1	0.889	0.640	0.545	0.867
Text-CNN3	32	0.1	0.000 1	0.900	0.652	0.573	0.881
Text-CNN4	32	0.1	0.000 1	0.833	0.751	0.561	0.879
Text-CNN5	128	0.1	0.001	0.835	0.742	0.559	0.865
Text-CNN6	64	0.1	0.000 1	0.903	0.635	0.564	0.871
Text-CNN7	32	0.1	0.000 1	0.881	0.694	0.582	0.878
Text-CNN8	32	0.1	0.000 1	0.905	0.616	0.551	0.868
Text-CNN9	32	0.1	0.000 1	0.879	0.673	0.557	0.876

注:SE – 灵敏度;SP – 特异性;MCC 马修斯相关系数;AUC – 曲线下面积。

Note: SE – sensitivity; SP – specificity; MCC – matthews correlation coefficient; AUC – areunder curve.

2.1.3 最佳性能模型 Text-CNN3 的分析 表 2 展示了 Y-随机化检验过程中所建立的 3 个 Text-CNN 模型的分类性能。其 AUC 值分别为 0.509、0.478 和 0.506,MCC 值分别为 – 0.009、– 0.016 和 0.000。MCC 和 AUC 两个指标显著低于 Text-CNN3 模型(AUC:0.881; MCC:0.573),证明基于 Y-随机化后的数据集训练得到的模型无法准确预测抗菌活性。因此,模型 Text-CNN3 的预测性能具有鲁棒性。

表 2 Y-随机化检验中建立的模型的性能和超参数

Tab. 2 Performance and hyperparameters of three models trained in Y-randomization tests

Models	Hyperparameters			SE	SP	MCC	AUC
	batch size	dropout rate	learning rate				
M1	64	0.3	0.001	1.000	0.000	– 0.009	0.509
M2	32	0.1	0.001	0.999	0.000	– 0.016	0.478
M3	64	0.2	0.01	1.000	0.000	0.000	0.506

本研究还对 Text-CNN3 模型的适用域进行了研究。结果显示(图 4),AUC 值随着测试子集内化合物间相似度值 *Tc* 阈值的减小而减小。说明测试集内化合物越多样,模型性能越差。为使得 AUC 值大于 0.70(即较好的分类性能),化合物库内的分子与训练集内分子的相似度建议在 0.80 以上。

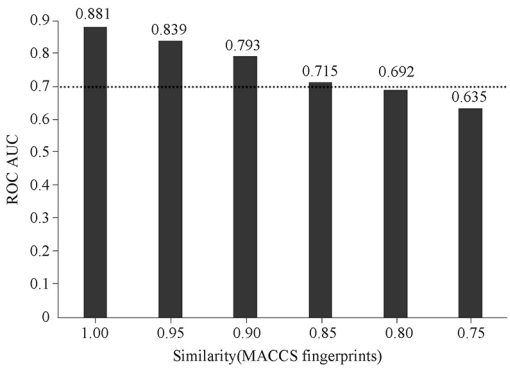


图 4 Text-CNN3 模型的 ROC AUC 值随训练化合物与测试化合物之间最大结构相似度的变化

Fig. 4 The corresponding ROC AUCs of the Text-CNN3 model with the maximum structural similarity of test compounds to the training compounds.

2.2 虚拟筛选和体外抗 *S. aureus* 活性评价

采用 Text-CNN3 模型对药物所内部化合物库(435 个)进行预测(图 5),保留预测为有活性的化合物 137 个。为保证结构新颖性,保留其与已知抗 *S. aureus* 化合物的相似度值 *Tc* 小于 0.75 的化合物

36 个。最后,基于 FCFP_6 指纹将其聚成 10 簇,根据结构多样性和合成可行性选取 9 个化合物(表 3)。经过 PubChem 查询,结果显示上述化合物的抗菌活性未见报道。购买上述化合物并进行体外抗菌活性评价,发现化合物 **Y5**(7S)-4-氨基-7-甲基-5,6,7,8-四氢-[1]苯甲硫醇[2,3-d][1,3]噻嗪-2-硫酮和 **Y7**(5Z)-3-[(4-乙基苯胺基)甲基]-5-[(4-碘苯基)亚甲基]-1,3-噻唑烷-2,4-二酮对 *S. aureus*(ATCC29213)有较好的抑制活性,MIC 分别为 8 和 4 μg·mL^{-1}。

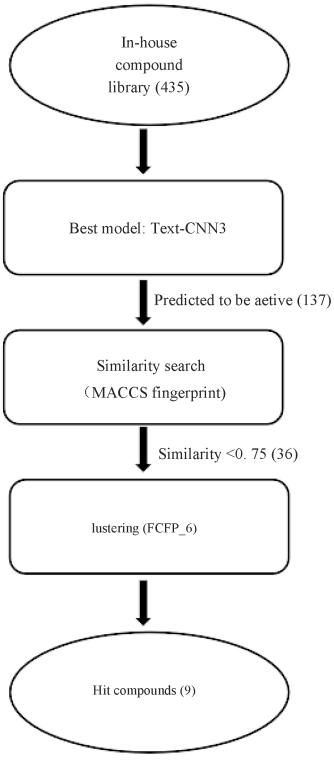


图 5 抗金黄色葡萄球菌化合物的虚拟筛选流程

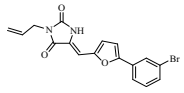
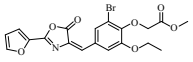
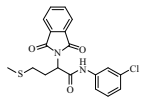
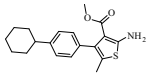
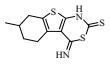
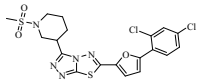
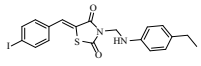
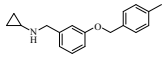
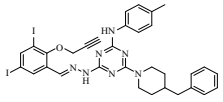
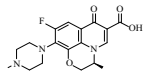
Fig. 5 The workflow of virtual screening for anti- *S. aureus* compounds

3 结论与讨论

本研究通过收集和处理标注有 *S. aureus* 活性数据的化合物,建立了可用于机器学习建模的数据集,含 26 327 个 *S. aureus* 相关化合物。随后通过随机采样建立了 10 对训练集和测试集。本文首次采用基于化合物的 SMILES 表示及原子级分词方法,以 Text-CNN 算法训练了 10 个抗 *S. aureus* 活性预测模型。经模型性能评估,Text-CNN3 模型性能最好,且该模型通过了 Y 随机化测试和适用域分析,可以用于虚拟筛选。在此基础上,本研究开展了基于 Text-CNN3 模型的虚拟筛选,经过结构相似度比对和聚类分析,选定 9 个结构多样的化合物进行体外抗 *S. aureus* 活性评价,发现化合物 **Y5** 和 **Y7** 可以有效抑制 *S. aureus*(MIC:4 ~ 8 μg·mL^{-1}),是新结构类型的抗菌药物苗头化合物。

表 3 潜在的苗头化合物的结构及体外抗 *S. aureus*(ATCC29213)活性

Tab. 3 The structures of potential hit compounds and *in vitro* anti-*S. aureus*(ATCC29213) activity

ID	Chemical structure	MIC /μg·mL^{-1}
Y1		>32
Y2		>32
Y3		>32
Y4		>32
Y5		8
Y6		>32
Y7		4
Y8		>32
Y9		>32
Levofloxacin		0.015

本研究通过理论研究和药物发现实践,验证了 Text-CNN 算法用于抗菌药物发现的价值。但本研究尚存在一定的局限性:一方面,化合物 **Y5** 和 **Y7** 的抗菌活性与左氧氟沙星等上市药物还有较大差距,需进一步通过化学结构改造提高其抗菌活性。另一方面,Text-CNN3 的性能指标还有提升空间,后续应注重发展更为新颖的方法提升其预测性能。

REFERENCES

[1] CHIN C Y, TIPTON K A, FAROKHYFAR M, *et al*. A high-frequency phenotypic switch links bacterial virulence and environmental survival in Acinetobacter baumannii [J]. *Nat Microbiol*, 2018, 3(5): 563-569.

[2] DAVID L, BRATA A M, MOGOSAN C, *et al*. Artificial intelligence and antibiotic discovery [J]. *Antibiot Chemother (Basel)*, 2021, 10(11): 1376. DOI: 10.3390/antibiotics10111376.

[3] DING L, YANG Y, ZHENG C, *et al*. Activities of eravacycline, tedizolid, norvancomycin, nemonoxacin, ceftaroline, and comparators against 1 871 staphylococcus and 1 068 enterococcus species isolates from China: updated report of the CHINET study 2019 [J]. *Microbiol Spectr*, 2022, 10(6): e0171522. DOI: 10.1128/spectrum.01715-22.

[4] DIMASI J A, GRABOWSKI H G, HANSEN R W. Innovation in the pharmaceutical industry: new estimates of R&D costs [J]. *J Health Econ*, 2016, 47: 20-33.

[5] BUTLER M S, HENDERSON I R, CAPON R J, *et al*. Antibiotics in the clinical pipeline as of december 2022 [J]. *J Antibiot (Tokyo)*, 2023, 76(8): 431-473.

[6] WANG L, LE X, LI L, *et al*. Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches [J]. *J Chem Inf Model*, 2014, 54(11): 3186-3197.

[7] STOKES J M, YANG K, SWANSON K, *et al*. A deep learning approach to antibiotic discovery [J]. *Cell*, 2020, 180(4): 688-702.

[8] BADURA A, KRYSINSKI J, NOWACZYK A, *et al*. Application of artificial neural networks to prediction of new substances with antimicrobial activity against Escherichia coli [J]. *J Appl Microbiol*, 2021, 130(1): 40-49.

[9] IVANENKOV Y A, ZHAVORONKOV A, YAMIDANOV R S, *et al*. Identification of novel antibacterials using machine learning techniques [J]. *Front Pharmacol*, 2019, 10: 913. DOI: 10.3389/fphar.2019.00913.

[10] KIM Y. Convolutional neural networks for sentence classification [C]. Qatar: Association for Computational Linguistics, 2014: 1746-1751.

[11] QIN T, GAO X, LEI L, *et al*. Machine learning-and structure-based discovery of a novel chemotype as FXR agonists for potential treatment of nonalcoholic fatty liver disease [J]. *Eur J Med Chem*, 2023, 252: 115307. DOI: 10.1016/j.ejmech.2023.115307.

[12] ZHANG Y, WALLACE B. A sensitivity analysis of(and practitioners′ guide to) convolutional neural networks for sentence classification [C]. China Taiwan: Asian Federation of Natural Language Processing, 2017: 253-263.

[13] MATTHEWS B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme [J]. *Biochim Biophys Acta*, 1975, 405(2): 442-451.

[14] LIAN X, XIA Z, LI X, *et al*. Anti-MRSA drug discovery by ligand-based virtual screening and biological evaluation [J]. *Bioorg Chem*, 2021, 114: 105042. DOI: 10.1016/j.bioorg.2021.105042.

[15] NIE X. Computational prediction studies of the biological activity of aurora kinase inhibitors [D]. Beijing: University of Chemical Technology, 2013.

[16] TANG H, WANG X S, HUANG X P, *et al*. Novel inhibitors of human histone deacetylase(HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation [J]. *J Chem Inf Model*, 2009, 49(2): 461-476.

[17] VOGT M, BAJORATH J. Modeling Tanimoto Similarity Value Distributions and Predicting Search Results [J]. *Mol Inform*, 2017, 36(7). DOI: 10.1002/minf.201600131.

[18] CERETO-MASSAGUÉ A, GUASCH L, VALLS C, *et al*. DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets [J]. *Bioinformatics*, 2012, 28(12): 1661-1662.

[19] GATICA E A, CAVASOTTO C N. Ligand and decoy sets for docking to G protein-coupled receptors [J]. *J Chem Inf Model*, 2012, 52(1): 1-6.

[20] XIA J, REID T E, WU S, *et al*. Maximal unbiased benchmarking data sets for human chemokine receptors and comparative analysis [J]. *J Chem Inf Model*, 2018, 58(5): 1104-1120.

[21] PASUPA K, KUDISTHALERT W. Virtual screening by a new clustering-based weighted similarity extreme learning machine approach [J]. *PLoS One*, 2018, 13(4): e0195478. DOI: 10.1371/journal.pone.0195478.

[22] XUE W, LI X, MA G, *et al*. *N*-thiadiazole-4-hydroxy-2-quinolone-3-carboxamides bearing heteroaromatic rings as novel antibacterial agents: design, synthesis, biological evaluation and target identification [J]. *Eur J Med Chem*, 2020, 188: 112022. DOI: 10.1016/j.ejmech.2019.112022.

（收稿日期:2023-07-20）