中图分类号：R97；TQ311.5　　文献标志码：A　　文章编号：1006－4931(2024)14－0047－07
doi:10.3969/j.issn.1006－4931.2024.14.012

基于集成深度学习框架的新型冠状病毒感染治疗药物活性预测*

许　强[1]，罗杰斯[2]，杨　明[1]，张永林[1△]

（1. 川北医学院附属医院，四川 南充　637000；2. 西南医科大学，四川 泸州　646000）

摘要：目的　建立预测新型冠状病毒感染治疗药物活性的集成深度学习框架。**方法**　采用卷积神经网络(CNN)和递归神经网络(RNN)从简化分子线性输入规范(SMILES)字符串序列信息中筛选出代表性的特征标识，以深度神经网络(DNN)从离散特征信息中提取更高级别的抽象特征，均以网格筛选法生成 1 个主框架模型和 7 个离散特征模型的最优结构，构成 8 种架构的 127 种可能组合。通过准确率(ACC)、F、召回率(Recall)、精确度(PRE)和马修斯相关系数(MCC)5 个标准指标评估模型的预测性能。建立和维护最终框架。**结果**　最终建立了 1 个以 BiLSTM 为集成深度学习框架的核心架构和 4 个不同的离散特征模型组成的集成深度学习模型，训练集 ACC 为 72.84%，F 为 69.70，Recall 为 72.21%，PRE 为 68.03，MCC 为 0.456 9；测试集中成功预测了 23 种可能对新型冠状病毒感染有治疗作用的药物。**结论**　集成深度学习框架相较于单个模型具有更强的预测能力，该研究为新型冠状病毒感染治疗药物的筛选提供了新的选择。

关键词：集成深度学习框架；新型冠状病毒感染；药物活性；神经网络；自动生物序列

Prediction of Drug Activity for COVID－19 Based on Ensemble Deep Learning Framework

XU Qiang[1], *LUO Jiesi*[2], *YANG Ming*[1], *ZHANG Yonglin*[1]

(1. Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, China 637000; 2. Southwest Medical University, Luzhou, Sichuan, China 646000)

Abstract: Objective　To establish an ensemble deep learning framework for predicting the activity of drugs for Corona Virus Disease 2019 (COVID－19). **Methods**　Convolutional neural network (CNN) and recursive neural network (RNN) were used to screen the representative feature identifiers from the simplified molecular input line entry system (SMILES) sequence. Deep neural

*基金项目：西南特色中药资源国家重点实验室开放基金[SKLTCM2022028]；川北医学院校级科研发展计划项目[CBY22－QNA38]。

第一作者：许强，男，大学本科，药师，研究方向为药物设计，(电子信箱)516900753@qq.com。

△通信作者：张永林，男，硕士研究生，主管药师，研究方向为临床药学，(电子信箱)zyonglin2021@163.com。

办公厅印发《关于深化审评审批制度改革鼓励药品医疗器械创新的意见》[A/OL].(2017－10－08)[2021－06－18]. https://www.gov.cn/zhengce/2017－10/08/content_5230105.htm.

[7] NEWMAN MEJ, GIRVAN M. Finding and evaluating community structure in networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69(2 Pt2): 026113.

[8] BLONDEL VD, GUILLAUME JL, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): 10008.

[9] YANG H, LEE HJ. Long－Term Collaboration Network Based on Clinical Trials. gov Database in the Pharmaceutical Industry[J]. Sustainability, 2018, 10(2): 322.

[10] DENG JL, SITOU K, ZHANG YP, et al. Analyzing the Chinese landscape in anti－diabetic drug research: Leading knowledge production institutions and thematic communities[J]. Chin Med, 2016, 11: 13.

[11] YOU H, NI JY, BARBER M, et al. China's landscape in oncology drug research: Perspectives from research collaboration networks[J]. Chinese Journal of Cancer Research, 2015, 27(2): 138－147.

[12] Word Health Organization. International Classification of Diseases 11th Revision[EB/OL].(2021－05－10)[2021－06－17]. https://icd.who.int/browse/2021－05/mms/en.

[13] JACOMY M, VENTURINI T, HEYMANN S, et al. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software[J]. PLoS One, 2014, 9(6): e98679.

[14] 潘幸梅，谢林利，马　攀，等. 某院药物临床试验质量控制存在的问题及改进措施[J]. 中国药业，2023，32(8)：1－4.

[15] 国家食品药品监督管理总局. 国家食品药品监督管理总局关于开展药物临床试验数据自查核查工作的公告[A/OL].(2015－07－22)[2021－06－17]. https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20150722173601172.html.

[16] 国家市场监督管理总局. 药品注册管理办法[A/OL].(2020－01－22)[2021－06－17]. https://www.gov.cn/zhengce/zhengceku/2020－04/01/content_5498012.htm.

[17] 张百红，岳红云. 抗肿瘤靶向药物的分类[J]. 现代肿瘤医学，2017，25(2)：299－303.

[18] 解　影，刘　浩. 表皮生长因子受体酪氨酸激酶抑制剂在肺癌以外的应用进展[J]. 中国药理学通报，2021，37(4)：467－472.

[19] 陈春燕，单慧亭，伊力亚斯·买买提艾力，等. 新型抗肿瘤药物药品不良反应及危险因素分析[J]. 中国药业，2023，32(1)：107－110.

[20] 苏　娴，姚珠星，王海学，等. 2020 年中国药物临床试验进展分析[J]. 中国食品药品监管，2021(10)：14－20.

（收稿日期：2023－07－11；修回日期：2023－12－20）

network (DNN) was used to extract higher - level abstract features from discrete feature information. The optimal structure of one main framework model and seven discrete feature models was generated by the grid search method, forming 127 possible combinations of eight architectures. The predictive performance of model was evaluated by the accuracy (ACC), F, Recall, precision (PRE) and Matthews correlation coefficient (MCC). The final framework was established and maintained. **Results** An ensemble deep learning model with BiLSTM as the core architecture and consisting of four different discrete feature models was ultimately established. The ACC of the training set was 72.84%, the F was 69.70, the Recall was 72.21%, the PRE was 68.03, and the MCC was 0.456 9. Twenty - three drugs that might be effective against COVID - 19 were successfully predicted in the test set. **Conclusion** The ensemble deep learning framework has better predictive performance than a singular model, this study provides a new choice for the screening of the drugs for COVID - 19.

Key words: ensemble deep learning framework; Corona Virus Disease 2019; drug activity; neural network; autoBioSeqpy

新型冠状病毒(简称新冠病毒,SARS - CoV - 2)感染具有高传染性和致病性[1],长时间感染可能对心血管系统、神经系统等产生不良影响[2-4],故迫切需要开发快速、高效的治疗方法。虽然在短期内开发针对新冠病毒的新药似乎不切实际,但可利用已批准的药物作为治疗疾病的有效策略[5]。有体外试验显示,一些已知药物如氯喹[6-7]、洛匹那韦[8]、伊维菌素[9]可抑制新冠病毒的活性。药物再利用技术可快速发现潜在治疗药物[10-11]。本研究中采用的计算方法分别基于结构和配体,前者依赖于新冠病毒蛋白结构数据[12-15],后者依赖于分子生物活性数据[16]。机器学习和深度学习在药物发现过程中得到了广泛应用,用于预测活性和分子分类[17-19]。其中深度学习在药物设计方面表现出色,常用于药物再利用和对抗新冠病毒[20-22]。为提高预测性能,本研究中提出了一种集成深度学习框架,结合简化分子线性输入规范(SMILES)字符串和离散信息,以预测新冠病毒感染治疗药物活性[23]。现报道如下。

1 资料与方法

1.1 数据收集

目前,新冠病毒感染治疗药物的研究仍处于发展阶段,因此药物信息相对受限。通过查阅文献筛选出与其治疗相关的有效或有潜在疗效的药物,作为训练集的正样本化合物来源。检索DrugBank数据库(http://www.drugbank.ca/),随机筛选出相同数量的负样本化合物,该数据库包含了详细的药物结构、药品靶点、药物作用、药物相互作用信息[24]。正、负样本化合物间有显著差异($P<0.05$)。此外,需构建1个用于测试集成框架性能的独立测试数据集。该数据集由新冠病毒感染治疗的正、负样本化合物组成,并剔除了在训练集中已存在的化合物。本研究中所建模型预测数据集包括训练集和测试集两部分,前者含正、负样本各209个,后者含正、负样本各50个。检索PubChem数据库[25]得到以SMILES格式表示的药物化学结构,并通过DeepChem[26]开源库生成分子描述符。

1.2 特征表示方法

One - hot编码:SMILES是一种表示化学分子结构的字符串(由一系列原子、键、括号等符号组成)编码方式。使用one - hot编码可将SMILES字符串表示为1个二维矩阵,每1行表示1个字符,每1列表示1个可能的字符类型。在向量中,只有1个元素被设置为1,表示该元素,其他元素均被设置为0。如对1个包含原子符号、键符号和括号的SMILES字符串,可将每个字符表示为一个向量,如下所示,原子符号"C":[1 0 0 0 0 … 0],原子符号"N":[0 1 0 0 0 … 0],键符号"-":[0 0 1 0 0 … 0],左括号"(":[0 0 0 1 0 … 0],右括号")":[0 0 0 0 1 … 0]等。SMILES字符串共有74种字符,因此它是一个$74\times L$维度的矩阵,L表示字符串的长度。

DeepChem库:其为开源的Python库,用于分子机器学习和药物发现领域,可处理分子数据,生成分子描述符,构建分子模型,并进行分子属性预测、分子筛选、分子设计等任务。DeepChem支持多种分子描述符生成方法,包括基于分子结构的描述符(如MACCSKeysFingerprint, CircularFingerprint等),基于分子图谱的描述符(如MolGraphConvFeaturizer_edge, MolGraphConvFeaturizer_node等),基于分子物理化学性质的描述符(如MordredDescriptors, ChemicalFeaturizer等)等。同时,DeepChem还支持多种深度学习和传统机器学习算法,包括神经网络、支持向量机、随机森林、K最近邻等。

MACCS Keys算法:该算法是一种分子指纹算法。MACCSKeysFingerprint是该算法生成的一种指纹特征,可用于计算化合物间的相似性和差异性,通常用于计算机辅助药物设计和化学信息学研究等领域。该指纹由166个预定义的二进制比特位构成,每个比特位表示分子中的1个结构特征,如化学键、环、官能团等,因此该指纹的维度为166。每个分子对应1个长度为167的二进制向量,向量中的每个比特位表示该分子是否存在对应的结构特征。

MolGraphConvFeaturizer_edge算法:该算法是一种用于计算分子图形特征的算法,可将分子结构表示为1个带权无向图,其中每个原子均为图的节点,每个化学键均为图的1条边,每条边均带有1个权重值,用于表示2个相邻原子间的化学性质。具体来说,该算法可

计算出每个原子和化学键的一系列特征，如原子的电荷及化学键的类型、键长、键角等，这些特征构成分子的特征向量表示。该算法生成的特征向量维度取决于所使用的参数设置，通常包括原子特征、化学键特征、图形特征等多个方面。

MolGraphConvFeaturizer_node 算法：该算法是一种用于计算分子图形特征的算法，可将分子结构表示为 1 个带权无向图，其中每个原子均为图的节点，每个化学键均为图的 1 条边，每个节点（即原子）均带有 1 个特征向量，用于表示该原子的化学性质。该算法可计算出每个原子的一系列特征，如原子的电荷、原子类型、杂化状态、价电子数、是否为环中的原子等，这些特征构成了分子的节点特征向量表示。生成的节点特征向量维度取决于所使用的参数设置，通常包括原子特征、连接边特征、图形特征等多个方面。

CircularFingerprint 算法：该算法是一种分子指纹算法，通常用于计算分子间的相似性和差异性。它可将分子的结构表示为一组二进制比特位的向量，其中每个比特位表示分子中的 1 个环系统。具体来说，该算法首先通过指定半径和位移来定义一系列圆形子结构，然后对每个子结构计算其哈希值，并将哈希值映射到 1 个固定长度的二进制比特位上，从而生成分子的指纹表示。生成的指纹向量维度通常为固定长度，如默认设置下为 2 048 位。每个分子都对应一个长度为 2 048 的二进制向量，向量中的每个比特位表示分子中是否存在对应的圆形子结构，若存在则对应比特位的值为 1，否则为 0。

MordredDescriptors 算法：该算法是一种分子描述符算法，用于计算分子结构的数值特征，如物理性质、化学性质等。该算法是一个基于 Python 的开源库，可计算 1 828 种不同的分子描述符，包括结构、电荷、能量、热力学性质、拓扑、质子接受 / 给予能力、极性等多个方面特征。生成的分子描述符维度数量取决于所选择的描述符种类和参数设置。通常每个分子可表示为 1 个由多个数值特征组成的向量，每个特征对应 1 个维度。

PubChemFingerprint 算法：该算法是一种分子指纹算法，用于计算分子结构的二进制指纹特征，可用于快速比较分子结构的相似性和差异性。该算法基于 PubChem 数据库中的分子数据，利用分子中的化学信息和结构信息计算指纹特征，包括分子中出现的化学基团、键类型、环结构等信息。生成的指纹特征维度数量为 881，每一位均为二进制的 0 或 1，表示该分子是否具有对应的化学特征。

SmilesToImage 算法：该算法是一种将 SMILES 字符串转换成分子结构图片的算法。该算法可将 SMILES 字符串转换为对应的分子结构图像，方便进行可视化展示或其他进一步分析。生成的图像通常为平面（2D）或立体（3D）图像，具体维度取决于所选择的图像生成工具、图像大小、像素密度等因素。

1.3 深度学习架构

深度学习是目前较成功的机器学习形式，其采用由多个顺序层组成的复合神经网络架构，能对输入数据进行训练，以实现预测任务。深度学习的理念在于，简单层的叠加能进行端到端的学习，自动地发掘原始数据中更高层次的表征，这种方法在模拟各种关系时非常强大和灵活[27]。为了支持构建高度灵活的模型架构，研究者们提出了各种类型的网络层，包括卷积层、池化层、递归层、激活层、全连接层等。其中卷积层使用卷积操作来融合相似的特征，并通过卷积核来传输。卷积神经网络（CNN）层使用卷积操作来融合彼此相近的特征，并通过内核来传输，公式如下。

$$convolution(X)_{ik} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{mn}^{k} X_{i+m,n} \tag{1}$$

其中，X 为层输入，i 和 k 分别为输出位置和滤波核的索引[28]。卷积滤波器是一种由权重矩阵组成的算子，它通过窗口大小 M 和 N 及输入通道来定义。此外，卷积操作可通过改变填充大小和扩张大小来适应更广泛的信息融合。对于连续数据，CNN 可利用每个基点的上下文信息来进行特征融合，从而让每个基点可根据其邻位基点的不同而产生不同的输出。虽然使用 K - mer 或滑动窗口操作也能实现类似的效果，但使用 CNN 可减少稀疏性，为进一步的数据处理提供可调节的通道数量。在卷积层之后，通常会添加池化层，其功能是通过计算区域内特征的最大值或平均值对层的输入进行降采样。在 RNN 系列中，每个层使用序列的每个单元来更新隐藏状态，以便从上下文中学习和推理。在递归层中，张量被用来表示隐藏状态，每个序列单元均被编码为 1 个或多个全连接层，以更新隐藏单元。例如，长短时记忆（LSTM）层包含隐藏状态和单元状态，在每次迭代中都会更新，公式如下。

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \tag{2}$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \tag{3}$$

$$\widetilde{C}_t = \tanh(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C) \tag{4}$$

$$C_t = f_t C_{t-1} + i_t \times \widetilde{C}_t \tag{5}$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \tag{6}$$

$$h_t = o_t \times \tanh(C_t) \tag{7}$$

其中，c_t 和 h_t 分别是细胞状态和隐藏状态[29-31]。而门控递归单元（GRU）层只更新隐藏状态，公式如下。

$$r_t = \sigma(W_{rh}h_{t-1} + W_{rx}x_t + b_r) \tag{8}$$

$$\widetilde{h}_t = \varphi_h(W_{hh}(r_t \times h_{t-1}) + W_{hx}x_t + b_h) \tag{9}$$

$$z_t = \sigma(W_{zh}h_{t-1} + W_{zx}x_t + b_z) \tag{10}$$

$$h_t = (1 - z_t) \times \widetilde{h}_t + z_t \times h_{t-1} \tag{11}$$

此外，双向操作，允许RNN层从头和尾2个方向开始学习1个序列，因为RNN处理序列没有预定方向。密集层，也被称为全连接层，是最简单的层类型，每个输入均与每个输出相连。激活函数的作用是在输入－输出关系中引入非线性。经常使用的激活函数包括sigmoid和rectified linear unit（ReLU），公式如下。

$$\mathrm{sigmoid}(x) = 1/(1 + e^{-x}) \quad (12)$$

$$\mathrm{ReLU}(x)\begin{cases} x & (if\ x \geqslant 0) \\ 0 & (if\ x < 0) \end{cases} \quad (13)$$

通常，ReLU被用于模型的非线性增益，而sigmoid被用于二元分类模型的输出层。

本研究中提出了4种不同的神经网络结构，包括CNN、双向长短期记忆网络（BiLSTM）、双向门控递归单元网络（BiGRU）和深度神经网络（DNN）。上述模型以SMILES字符串的74个字符作为输入，并输出了该药物是否对SARS－CoV－2具有活性的预测概率。其中CNN，BiLSTM，BiGRU模型的输入使用了SMILES字符串的One－hot编码，并产生1个范围在[0，1]之间的分数，代表正类（T或1）和负类（F或0）。进行了网格搜索，并详细测试了卷积层（1，2），核大小（3，5，7，9，11），过滤器数量（50，150，250），LSTM层（1，2），LSTM层单元数（64，128，256），GRU层（1，2）及GRU层单元数（32，64，128，256）等超参数，以选择最佳模型。第4种DNN结构使用DeepChem的特征作为输入，输出1个概率分数。为了获得每个模型的最佳性能结果，采用网格搜索方法建立密集连接神经网络来确定超参数的最佳组合。

1.4 集成框架设计

为了有效提高基础模型的性能，本研究中开发了深度学习集成框架。此框架中输入数据集的类型不同，采用了不同的构建方式。对于序列数据（即SMILES字符串），采用One－hot编码方式进行编码。编码后的矩阵用于CNN或递归神经网络（RNN）层处理，从中提取包括CNN层的局部信息和RNN层的全局信息。对于离散特征组，如基于DeepChem的描述符，采用DNN层进行处理。首先使用DNN层作为预测的投影层，单独对每个离散特征组进行预测；然后在特征维度上进行串联操作，将所有顺序和离散特征的组合用于预测。因此，可认为CNN和RNN是集成深度学习框架的核心架构，而DNN的不同描述符组合被用作附加信息，并串联到核心架构的最后一个全连接层。本研究中采用7种特征组合（包括CircularFingerprint，MACCSKeysFingerprint，MolGraphConvFeaturizer_edge，MolGraphConvFeaturizer_node，MordredDescriptors，PubChemFingerprint，SmilesToImage），共生成127个DNN模型，用于串联核心架构。如使用SmilesToImage和MACCSKeysFingerprint对核心架构进行串联时，3个输出会在最后一个全连接层之前串联到一起。首先使用相同的训练数据集比较了3种序列级深度学习模型（CNN，BiLSTM，BiGRU）的性能，以找到最佳的核心架构；之后，选择最佳核心架构来串联不同离散特征组合的DNN模型，为减少采样和模型拟合中随机因素的影响，每个训练程序重复5次；最终进行了635个模型训练程序，其中50个用于单一模型，585个用于集成深度学习框架。

1.5 实践

使用深度学习工具autoBioSeqpy[23]和Keras[32]后端来完成设计、训练和评估上述单一深度学习模型和集成深度学习框架，训练过程是在运行Windows 11的工作站上进行的，该工作站配备了带有CUDA 12.1的NVIDIA GeForce RTX 3070 GPU。

1.6 评估标准

为了衡量集成深度学习框架的预测性能，使用5个标准指标，即准确率（ACC）、精确度（PRE）、F值、召回率（Recall）和马修斯相关系数（MCC）。上述指标定义如下。TP，TN，FP，FN表示二元分类中真阳性、真阴性、假阳性和假阴性的数量。公式如下。

$$\mathrm{ACC} = (TP + TN)/(TP + FP + TN + FN) \quad (14)$$

$$\mathrm{PRE} = TP/(TP + FP) \quad (15)$$

$$\mathrm{F-value} = 2 \times [TP/(2TP + FP + FN)] \quad (16)$$

$$\mathrm{Recall} = TP/(TP + FN) \quad (17)$$

$$\mathrm{MCC} = [(TP \times TN) - (FN \times FP)]/ \quad (18)$$
$$[(TP + FN) \times (TN + FP) \times (TP + FP) \times TN + FN]^{0.5}$$

1.7 autoBioSeqpy概述

自动生物序列分类程序是一款基于Keras的深度学习软件，旨在快速、便捷地开发、训练和分析用于生物序列分类的深度学习模型架构[23]。与其他工具或库相比，其最大的优势为操作简单，特别适合对深度学习技术了解有限或不具备专业知识的非专业人员使用。用户无须编程，只需准备输入数据集和模型模板即可，此外进行指令操作即可自动执行整个工作流程（即文件读取、数据编码、参数初始化、模型训练、评估、可视化等）。

2 结果

2.1 模型评估

使用相同的训练集，对CNN，BiLSTM，BiGRU深度学习模型进行了不同超参数组合的测试，每一组合重复训练5次，取平均值，结果见表1和表2。可见，3种模型的不同超参数组合的预测性能均表现良好。CNN模型的超参数组合的层数为2、过滤器数量为250、卷积核大小为9时，ACC为67.16%、F为63.43、MCC为0.3532，高于其他组合。在3种模型中，CNN模型得分最高（F值为63.43，Recall为67.85%）。RNN的2种模型中，

BiLSTM 模型的性能表现最好，ACC（70.45%）、PRE（77.01%）和 MCC（0.425 4）均最高时的超参数组合为 LSTM 层数为 2，隐藏层单元数为 128，BiGRU 模型最高（LSTM 层数为 2，隐藏层单元数为 64）的 F（62.82）和 Recall（61.30）高于 BiLSTM 模型（见表 2）。以 ACC 和 MCC 为比较对象，BiLSTM 模型的预测性能优于其他模型（见表 3），最终以 BiLSTM 模型作为集成框架的核心架构。

表 1　CNN 最优超参数优化结果

Tab. 1　Optimization of the optimal hyperparameters for CNN

项目	值	结构	ACC(%)	F	Recall(%)	PRE(%)	MCC
过滤器数量	50	CNN	59.40	58.31	63.00	56.24	0.207 7
	150		62.99	59.90	62.01	62.31	0.286 8
	250		67.16	63.43	67.66	61.55	0.353 2
卷积核大小	3	CNN	63.28	62.64	67.85	59.37	0.275 8
	5		62.39	59.13	54.10	65.87	0.253 5
	7		66.27	61.20	59.04	64.91	0.322 8
	9		67.16	63.43	67.66	61.55	0.353 2
	11		64.48	60.42	62.14	61.31	0.303 3
卷积层数量	1	CNN	60.30	60.06	61.44	59.79	0.208 3
	2		67.16	63.43	67.66	61.55	0.353 2

表 2　RNN 最优超参数优化结果

Tab. 2　Optimization of the optimal hyperparameters for RNN

项目	值	结构	ACC(%)	F	Recall(%)	PRE(%)	MCC
隐藏层单元数	64	BiLSTM	61.79	47.32	52.00	44.95	0.238 6
	128		70.45	60.57	54.57	77.01	0.425 4
	256		58.81	43.25	42.13	45.37	0.179 2
	32	BiGRU	63.28	57.77	58.00	61.90	0.274 1
	64		68.66	62.82	61.30	65.69	0.366 5
	128		63.28	53.98	51.32	68.27	0.266 6
	256		65.07	56.50	49.97	70.00	0.318 0
LSTM 层数	1	BiLSTM	65.37	45.80	44.65	49.98	0.257 1
	2		70.45	60.57	54.57	77.01	0.425 4
	1	BiGRU	60.30	55.65	53.82	59.99	0.213 5
	2		68.66	62.82	61.30	65.69	0.366 5

表 3　CNN，BiGRU，BiLSTM 最优超参数的 ACC 和 MCC 比较

Tab. 3　Comparison of the ACC and MCC for the optimal hyperparameters of CNN，BiGRU and BiLSTM

结构	ACC(%)	F	Recall(%)	PRE(%)	MCC
CNN	67.16	63.43	67.66	61.55	0.353 2
BiGRU	68.66	62.82	61.30	65.69	0.366 5
BiLSTM	70.45	60.57	54.57	77.01	0.425 4

2.2　性能评估

获得最优层数的特征描述符见表 4，其中最优层数 MolGraphConvFeaturizer_edge，MordredDescriptors，PubChemFingerprint 均为 5，MACCSKeysFingerprint 为 4，SmilesToImage 和 CircularFingerprint 均为 3，MolGraphConvFeaturizer_node 为 2。结果显示，特征描述符 MolGraphConvFeaturizer_node 的性能优于其他，其 ACC（71.34%）、F（65.31）、Recall（61.97%）和 MCC（0.412 8）均最高，MordredDescriptors 的 PRE（75.91%）最高，但其他指标略低。

表 4　7 种特征描述符的最优层数的预测性能

Tab. 4　Prediction performance of the optimal number of layers for seven types of feature descriptors

描述符	结构	ACC(%)	F	Recall(%)	PRE(%)	MCC
MACCSKeysFingerprint	DNN	59.10	54.15	54.31	54.12	0.172 4
MolGraphConvFeaturizer_edge	DNN	68.36	62.98	55.77	74.93	0.371 8
MolGraphConvFeaturizer_node	DNN	71.34	65.31	61.97	70.00	0.412 8
MordredDescriptors	DNN	60.60	40.97	32.00	75.91	0.228 9
PubChemFingerprint	DNN	63.88	59.20	61.80	57.79	0.264 7
SmilesToImage	DNN	64.18	61.47	59.92	65.58	0.299 8
CircularFingerprint	DNN	55.82	45.73	40.49	53.99	0.107 7

2.3　集成框架预测性能提高

上述比较了 3 个序列级深度学习模型的预测性能，最后以 BiLSTM 模型为核心架构。对 7 种特征描述符进行组合训练，生成 127 种 DNN 模型，用于串联核心架构。最后进行了 635 个模型训练程序，重复训练 5 次，取平均值。以 ACC 为比较对象，其中以 4 个特征描述符的组合最优，即 MACCSKeysFingerprint，SmilesToImage，MolGraphConvFeaturizer_edge，MolGraphConvFeaturizer_node。这 4 个特征描述符组合的 DNN 层的集成框架的 5 次预测性能见表 5，其平均 ACC 达 72.84%，平均 MCC 达 0.456 9。将该集成框架的 ACC 和 MCC 与上述 BiLSTM 模型和 DNN 模型的最高值比较，结果见图 1。可见，集成框架的 ACC 和 MCC 均高于其他单个模型，表明集成框架进一步提高了预测性能。

表 5　最佳组合的集成深度学习框架在训练集上的性能测试结果

Tab. 5　Results of performance test of the optimal combination of ensembled deep learning frameworks on the training set

次序	ACC(%)	F	Recall(%)	PRE(%)	MCC	次序	ACC(%)	F	Recall(%)	PRE(%)	MCC
1	82.09	81.25	81.25	81.25	0.641 1	4	74.63	73.02	71.88	74.19	0.491 1
2	70.15	65.52	79.17	55.88	0.424 7	5	68.66	65.57	66.67	64.52	0.368 4
3	68.66	63.16	62.07	64.29	0.359 2	$\bar{X}$	72.84	69.70	72.21	68.03	0.456 9

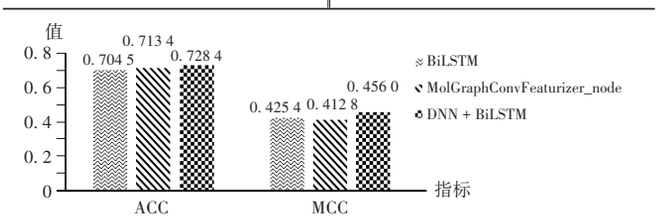

图 1　集成深度学习框架与单个模型的 ACC 和 MCC 比较

Fig. 1　Comparison of ACC and MCC for ensembled deep learning framework versus singular model

2.4 测试集下集成深度学习框架的性能

使用测试集对集成深度学习框架进行测试，每个程序重复5次。结果最高的ACC为72.84%，MCC为0.456 9。表明在测试集下的预测性能表现同样较好。正样本中有23种药物的预测概率大于0.5（见表6），这些药物可能对新冠病毒感染有较强的治疗作用。

表6 测试集中预测概率大于0.5的药物

Tab. 6 Drugs with predicted probabilities greater than 0.5 in the test set

种类	药物
呼吸系统用药	bromhexine hydrochloride
激素类	methylprednisone, dexamethasone
抗病毒药	viracept, molnupiravir, sofosbuvir, tmc－310911
抗凝剂	cyclosporine
抗炎剂	icatibant, apremilast, indomethacin, andrographolide, defibrotide
心血管系统用药	fostamatinib disodium, fostamatinib
抗肿瘤药	cyproterone acetate
其他类	danoprevir sodium, losmapimod, vidofludimus calcium, brensocatib, zavegepant, pam2csk4, tridecactide

3 讨论

新冠病毒的特效药研究需进行大量实验证明才能用于临床，而该病的突发特点导致特效药品短缺的情况发生，因此药物再利用成为普遍方法。以往的研究主要使用传统的机器学习方法预测药物的活性，常采用的特征编码方法包括分子指纹和SMILES特征等序列相关特征。虽然这些特征编码方法考虑了与药物分子相关的许多特征，但忽略了药物分子中最关键的结构组成信息。本研究旨在弥补这个缺陷，计算了与药物分子SMILES相关的各种衍生特征，从不同角度勾勒出药物分子的结构信息。研究结果表明，药物分子的保守性是一个高度稳健的特征，能有效区分活性和非活性。此外，在算法方面，采用了新颖的集成深度学习技术，不同于简单地合并所有模型，系统地探索了所有可行的模型组合方法，以确定最优组合模型。进行了1 015次训练迭代，尝试了635种不同的模型组合，最终确定表现最佳的模型架构。测试数据集中，成功预测出了23种可能对新冠病毒感染具有活性的药物。其中，bromhexine hydrochloride能有效改善病毒引起的呼吸道症状[33]，methylprednisone和dexamethasone可有效改善其引发的炎症问题[34]。viracept, molnupiravir, sofosbuvir, tmc－310911作为目前少数几种在治疗新冠病毒方面表现出特效的药物，能有效降低轻症转为重症的概率[35－38]。上述研究间接证明了本研究中最终确定的模型具有可推广性。

虽然集成模型在药物活性方面表现出了竞争力，但还存在需要改进的领域。一个主要限制是模型对训练集的质量和代表性的依赖。如果数据集不完整或存在偏差，就会影响模型在真实世界场景中的泛化能力。因此，获取多样化和平衡的数据集仍为严峻挑战。此外，本研究中集成模型主要依赖于从PSSM矩阵中获取的衍生信息。虽然这些信息有价值，但可能无法捕捉药物分子的所有功能或结构特征。而实际药物分子数据中，新冠病毒感染治疗药物的频率低于非新冠病毒感染治疗药物，这导致解决类别不平衡问题具有挑战性，且该模型在罕见新冠病毒感染治疗药物方面的表现可能不太可靠。

本研究仍有很大改进空间。为了提高预测能力，本研究的目标是探索表征新冠病毒感染治疗药物分子序列的更多特征，并开发更先进的深度学习算法来挖掘隐藏信息。此外，新冠病毒感染治疗药物在治疗过程中的错综复杂的相互作用和协调过程仍是未来研究的重点。总之，虽然集成模型已显示出竞争力，但其存在一定局限性。后期将致力于推进对新冠病毒活性药物分子的理解，并改进预测模型，以便在未来的实际应用中发挥作用。

参考文献

[1] WANG L, WANG Y, YE D, et al. Review of the 2019 novel coronavirus (SARS－CoV－2) based on current evidence[J]. International Journal of Antimicrobial Agents, 2020, 55(6): 105948.

[2] RAMAN B, BLUEMKE DA, LÜSCHER TF, et al. Long COVID: post－acute sequelae of COVID－19 with a cardiovascular focus[J]. European Heart Journal, 2022, 43(11): 1157－1172.

[3] DAVIS HE, MCCORKELL L, VOGOL JM, et al. Long COVID: major findings, mechanisms and recommendations[J]. Nature Reviews Microbiology, 2023, 21(3): 133－146.

[4] LI Q, GUAN X, WU P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus－Infected Pneumonia[J]. New England Journal of Medicine, 2020, 382(13): 1199－1207.

[5] PUSHPAKOM S, IORIO F, EYERS PA, et al. Drug repurposing: progress, challenges and recommendations[J]. Nature Reviews Drug Discovery, 2019, 18(1): 41－58.

[6] WANG M, CAO R, ZHANG L, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019－nCoV) in vitro[J]. Cell Research, 2020, 30(3): 269－271.

[7] GAO J, TIAN Z, YANG X. Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID－19 associated pneumonia in clinical studies[J]. Bioscience Trends, 2020, 14(1): 72－73.

[8] CAO B, WANG Y, WEN D, et al. A Trial of Lopinavir－Ritonavir in Adults Hospitalized with Severe Covid－19[J].

New England Journal of Medicine, 2020, 382(19): 1787 - 1799.
[9] CALY L, DRUCE JD, CATTON MG, et al. The FDA - approved drug ivermectin inhibits the replication of SARS - CoV - 2 *in vitro*[J]. Antiviral Research, 2020, 178: 104787.
[10] LI J, ZHENG S, CHEN B, et al. A survey of current trends in computational drug repositioning[J]. Briefings in Bioinformatics, 2016, 17(1): 2 - 12.
[11] DOTOLO S, MARABOTTI A, FACCHIANO A, et al. A review on drug repurposing applicable to COVID - 19[J]. Briefings in Bioinformatics, 2021, 22(2): 726 - 741.
[12] CHELLAPANDI P, SARANYA S. Genomics insights of SARS - CoV - 2 (COVID - 19) into target - based drug discovery[J]. Medicinal Chemistry Research, 2020, 29(10): 1777 - 1791.
[13] TREZZA A, IOVINELLI D, SANTUCCI A, et al. An integrated drug repurposing strategy for the rapid identification of potential SARS - CoV - 2 viral inhibitors [J]. Scientific Reports, 2020, 10(1): 13866.
[14] ZHANG H, SRARVANAN KM, YANG Y, et al. Deep Learning Based Drug Screening for Novel Coronavirus 2019 - nCov[J]. Interdisciplinary Sciences, 2020, 12(3): 368 - 376.
[15] ZHAI T, ZHANG F, HAIDER S, et al. An Integrated Computational and Experimental Approach to Identifying Inhibitors for SARS - CoV - 2 3CL Protease[J]. Frontiers in Molecular Biosciences, 2021, 8: 661424.
[16] YU W, MACKERELL AD JR. Computer - Aided Drug Design Methods[J]. Methods in Molecular Biology, 2017, 1520: 85 - 106.
[17] ALTAE - TRAN H, RAMSUNDAR B, PAPPU AS, et al. Low Data Drug Discovery with One - Shot Learning [J]. ACS Central Science, 2017, 3(4): 283 - 293.
[18] LO YC, RENSI SE, TORNG W, et al. Machine learning in chemoinformatics and drug discovery[J]. Drug Discovery Today, 2018, 23(8): 1538 - 1546.
[19] ZENG X, ZHU S, LIU X, et al. deepDR: a network - based deep learning approach to *in silico* drug repositioning [J]. Bioinformatics, 2019, 35(24): 5191 - 5198.
[20] NEVES BJ, BRAGA RC, ALVES VM, et al. Deep Learning - driven research for drug discovery: Tackling Malaria[J]. PLoS Computational Biology, 2020, 16(2): e1007025.
[21] MARAGAKIS P, NISONOFF H, COLE B, et al. A Deep - Learning View of Chemical Space Designed to Facilitate Drug Discovery[J]. Journal of Chemical Information and Modeling, 2020, 60(10): 4487 - 4496.
[22] ISERT C, ATZ K, SCHNEIDER G. Structure - based drug design with geometric deep learning [J]. Current Opinion in Structural Biology, 2023, 79: 102548.
[23] JING R, LI Y, XUE L, et al. autoBioSeqpy: A Deep Learning Tool for the Classification of Biological Sequences[J]. Journal of Chemical Information and Modeling, 2020, 60(8): 3755 - 3764.
[24] WISHART DS, FEUNANG YD, GUO AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J]. Nucleic Acids Research, 2018, 46(D1): D1074 - D1082.
[25] KIM S, CHEN J, CHENG T, et al. PubChem 2023 update[J]. Nucleic Acids Research, 2023, 51(D1): D1373 - D1380.
[26] DeepChem: Deep - learning models for Drug Discovery and Quantum Chemistry [DB/OL]. [2023 - 05 - 12]. http://github.com/deepchem/deepchem.
[27] WAINBERG M, MERICO D, DELONG A, et al. Deep learning in biomedicine[J]. Nature Biotechnology, 2018, 36(9): 829 - 838.
[28] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436 - 444.
[29] HOCHREITER S, SCHMIDHUBER J. Long short - term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.
[30] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the Properties of Neural Machine Translation: Encoder - Decoder Approaches [DB]. [2023 - 05 - 12]. http://aclanthology.org/W14 - 4012/.
[31] JIN XB, YANG A, SU T, et al. Multi - Channel Fusion Classification Method Based on Time - Series Data [J]. Sensors (Basel), 2021, 21(13): 4391.
[32] CHOLLET F. Keras, Deep learning library for theano and tensorflow [DB]. [2023 - 05 - 12]. https://github.com/keras - team/keras.
[33] MAGGIO R, CORSINI GU. Repurposing the mucolytic cough suppressant and TMPRSS2 protease inhibitor bromhexine for the prevention and management of SARS - CoV - 2 infection[J]. Pharmacological Research, 2020, 157: 104837.
[34] ETREMOV DO, BELOBORODOV VB. The role and place of pathogenetic therapy with glucocorticosteroid hormones in the treatment of patients with novel coronavirus infection (COVID - 19)[J]. Terapevticheskii Arkhiv, 2021, 93(11): 1395 - 1400.
[35] MUSARRAT F, CHOULJENKO V, DAHAL A, et al. The anti - HIV drug nelfinavir mesylate (Viracept) is a potent inhibitor of cell fusion caused by the SARSCoV - 2 spike (S) glycoprotein warranting further evaluation as an antiviral against COVID - 19 infections [J]. Journal of Medical Virology, 2020, 92(10): 2087 - 2095.
[36] JAYK BERNAL A, GOMES DA SILVA MM, MUSUNGAIE DB, et al. Molnupiravir for Oral Treatment of Covid - 19 in Nonhospitalized Patients[J]. New England Journal of Medicine, 2022, 386(6): 509 - 520.
[37] HEO YA, DEEKS ED. Sofosbuvir/Velpatasvir/Voxilaprevir: A Review in Chronic Hepatitis C[J]. Drugs, 2018, 78(5): 577 - 587.
[38] KERETEU S, BHUJBAL SP, CHO SJ. Rational approach toward COVID - 19 main protease inhibitors via molecular docking, molecular dynamics simulation and free energy calculation[J]. Scientific Reports, 2020, 10(1): 17716.

（收稿日期：2023 - 06 - 07；修回日期：2023 - 12 - 29）