doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 04. 007

基于残差结构的图卷积网络的药物靶点亲和力预测

金海峰，谭佳伟，刘铭
（长春工业大学数学与统计学院，吉林长春 130012）

【摘要】准确的药物靶点亲和力预测（DTA）能够缩短药物研发周期，节省人力和物力，加速药物研发过程。图神经网络（GNN）在药物靶点亲和力预测中得到了广泛应用，但现有的方法大多基于浅层 GNN。该文提出了一种基于残差结构的图卷积网络，残差结构的加入能够加深网络结构，借此构建一个具有 24 个图卷积层的深度图卷积网络，以此捕获药物分子的特征，学习有效的嵌入表达，并在两个基准药物靶点亲和力数据集上与几种先进的基于机器学习或深度学习的模型进行比较。结果表明，该文所提模型相较于其他基准模型有着更好的预测性能，验证了该文所提方法的有效性。

【关键词】药物靶点亲和力；图卷积网络；残差结构
【中图分类号】Q819　　【文献标志码】A　　文章编号：1674-1242（2023）04-0371-10

Prediction of Drug-Target Affinity Based on Residual Structure Graph Convolutional Network

JIN Haifeng, TAN Jiawei, LIU Ming
(College of Mathematics and Statistics, Changchun University of Technology, Changchun, Jilin 130012, China)

【**Abstract**】Accurate drug target affinity prediction (DTA) can shorten the drug development cycle, save manpower and material resources, and accelerate the drug development process. Graph Neural Networks (GNN) have been widely used in drug target affinity prediction, but most of the existing methods are based on shallow GNN. Therefore, a graph convolutional network based on the residual structure is proposed. The addition of the residual structure can deepen the network structure, thereby constructing a deep graph convolutional network with 24 graph convolutional layers to capture the characteristics of drug molecules, learn efficient embedding representations, and compare with several state-of-the-art machine learning or deep learning based models on two benchmark drug target affinity datasets. The results show that the proposed model has better predictive performance than other benchmark models, which verifies the effectiveness of the method proposed in this paper.

【**Key words**】Drug-Target Affinity; Graph Convolutional Network; Residual Structure

收稿日期：2023-10-15。
基金项目：吉林省发改委省预算内基本建设资金（2022C043-2），吉林省科技厅项目（20230204078YY）。
作者简介：金海峰（1999—），男，吉林省舒兰市人，硕士研究生，研究方向为计算智能与数据挖掘。
通信作者：刘铭，男，教授，博士研究生导师，邮箱（E-mail）：jlcclm@163.com。

0 引言

药物研发一直是医学领域的一项复杂、耗时的工作。为了研发疗效可靠且副作用较小的药物，研发人员通常需要进行大量的实验，从而消耗了大量的人力、物力和财力。而准确的药物靶点亲和力预测是研发新药和了解其副作用的重要一步，为药物研发起到了指导性作用，进而促进药物研发。近年来，利用机器学习等方法分析图结构的研究受到了越来越多的关注，因为图结构具有强大的表达能力，如药物分子结构、蛋白质相互作用网络等非欧式数据可以利用图结构进行有效的表示，且图神经网络（Graph Neural Networks，GNN）[1]能够对图结构的数据进行有效的处理和嵌入。因此，新兴的图神经网络被迅速应用于药物靶点亲和力预测，并被证明在加速药物研发、寻找重定位药物等研究中是有效的。

在传统的药物靶点亲和力预测方法中，通常使用蛋白质微阵列和亲和层析这两种生物实验方法。而为了确定对特定的蛋白质靶标有效且安全的药物分子，研发人员通常需要对数千种化合物进行测试。这一过程既费时又费资源。随着人工智能的快速发展，越来越多的行业开始使用基于机器学习和深度学习的方法[2]。对于药物靶点亲和力预测，基于机器学习和深度学习的方法因效率高、成本低等优点受到人们的广泛关注。对机器学习和深度学习中的模型而言，特征的提取是十分重要的，提取到生物特征向量后，可以将其用于训练机器学习模型或深度学习模型，如前馈神经网络、支持向量机、随机森林等基于核的机器学习模型。Wen 等[3]提出了 DeepDTI 模型，该模型选择了简单且常见的特征——扩展连通性指纹（Extended Connectivity Fingerprints，ECFP）和蛋白质序列描述符（Protein Sequence Descripor，PSC）分别作为药物分子和靶点蛋白质的表示，随后使用深度置信网络（Deep Belief Network，DBN）进行药物靶点亲和力预测。Cheng 等[4]首先提取药物分子的子结构并将其作为药物分子的嵌入表达，随后提取靶点蛋白质的结构域信息，以此作为靶点蛋白质的嵌入表达，然后将两者的张量积作为药物靶点相互作用对的特征，最终将其输入一个 Biased-SVM 进行分类训练。Rifaioglu 等[5]提出了一个系统——MDeePred，该系统提出了一种新的蛋白质特征化方法。在该方法中，多种类型的蛋白质特征（如序列、结构和物理化学性质等）被纳入多个二维向量中，借此来表示蛋白质序列描述符。然后将其输入一个混合深度神经网络中，以预测药物靶点的相互作用。Öztürk 等[6]提出了一个基于深度学习的模型——DeepDTA，该模型使用药物分子和靶点蛋白质的序列信息作为特征来训练并预测药物靶点相互作用的结合亲和力。具体方法是构建两个卷积神经网络（Convolutional Neural Networks，CNN）分别学习药物分子和靶点蛋白质的表达，获得两个序列的特征。然后将学习到的药物分子和靶点蛋白质的特征向量连接并输入一个多层感知机（Multilayer Perceptron，MLP）中进行药物靶点亲和力预测。在此基础上，Öztürk 等[7]又提出了一种集成 4 个文本信息源（蛋白质序列、配体 SMILES、蛋白质结构域和基序、最大共同亚结构词）来预测药物靶点结合亲和力的模型——WideDTA。该模型通过使用 4 个 CNN 将上述 4 个文本信息源编码为 4 个表示，使其在 KIBA 数据集上的表现优于 DeepDTA，具有统计学显著性。

虽然基于 CNN 的模型在药物靶点亲和力预测中取得了不错的成绩，但这些模型大多将药物分子表示为字符串序列，如果仅以字符串来表示药物分子，就忽略了它们的化学结构信息，可能导致模型在训练时无法学习到结构特征，导致预测时准确率降低。因为图结构能够很好地表现出数据的结构特征，同时随着 GNN 在许多领域的大放异彩，相继出现了许多将 GNN 应用于药物靶点亲和力预测模型。基于 GNN 的模型能够将药物分子表示为图，进而使用 GNN 进行药物靶点亲和力预测。Tsubaki 等[8]针对图和序列提出了一种端到端的化合物-蛋白质相互作用预测网络。他们使用 GNN 和 CNN 分别对药物分子的图与蛋白质序列进行学习，将它们表示为低维向量。实验表明，他们的方法在非平衡数据集上的性能显著优于其他方法。Nguyen 等[9]提出了 GraphDTA 模型，它将药物分子表示为图，并使用 GNN 来预测药物靶点亲和力。他们使用图卷积网络、图注意力网络等几种类型的 GNN 进行药物靶点亲和力预测并对结果进行评估。实验结果表明，相较于非深度学习模型，GNN 能够更好地预测药物靶点亲和力，而且优于其他基于深度学习的方法，其将药物表示为图，能够进一步提高预测精度。Jiang 等[10]利用分子和蛋白质的结构信息，分别构建了药物

分子和靶点蛋白质的两幅图，使用 GNN 获取其表示，并提出了 DGraphDTA 模型进行药物靶点亲和力预测。同时，为了提高模型的可解释性，他们还在模型中加入了注意力机制。

总之，无论是基于传统机器学习还是基于深度学习的模型，都对药物靶点相互作用、药物靶点亲和力预测提供了巨大的帮助，相较于传统的人工寻找药物分子的方法，它们极大地促进了药物研发，缩短了研发周期。但是，现有的基于 GNN 的方法大多基于浅层的 GNN，根据在图像识别领域使用 CNN 的经验，更深的网络结构对图像特征的捕捉可能更加丰富，进而模型的识别结果更加优秀。因此，具有深层结构的 GNN 对药物分子特征的捕获可能更加优秀，能够捕获药物分子的全局结构。据此，本文提出了一个基于残差结构的深度 GNN，用于获取药物分子的嵌入表达。使用一个多尺度的 CNN，用于获取靶点蛋白质的嵌入表达，随后将药物分子和靶点蛋白质的嵌入进行拼接，得到药物靶点的组合嵌入，将组合嵌入输入一个全连接层，进行药物靶点亲和力预测。

1　数据与方法

1.1　数据集

本文使用 Davis 和 KIBA 这两个数据集对提出的模型进行评估。这两个数据集是在药物-蛋白质亲和力预测中被广泛应用的基准数据集，数据集的整体情况如表 1 所示。在这两个数据集上对本文提出的模型进行评估，然后与其他药物靶点亲和力预测模型进行对比。

表 1　Davis 和 KIBA 数据集

Tab. 1　Davis and KIBA datasets

数据集	药物分子	靶点蛋白质	相互作用对
Davis	68 个	442 个	30 056 个
KIBA	2 111 个	229 个	118 254 个

Davis 数据集来自 Davis 等[11]的激酶抑制剂选择性综合分析中的实验数据，他们测试了 72 种激酶抑制剂与 442 种激酶的相互作用，这些激酶覆盖了超过 80% 的人类催化蛋白激酶组，最终形成的数据集包含 68 个药物分子和 442 个靶点蛋白质，以及这些药物分子与靶点蛋白质之间的 30 056 个相互作用对。其中，药物靶点亲和力预测的评价指标为解离常数 K_d，它反映的是药物分子对靶点蛋白质的亲和力大小，值越小，代表亲和力越强。其计算公式为

$$K_d = \frac{[E]\cdot[I']}{[EI']} \tag{1}$$

其含义为 50%的激酶 E 与激酶抑制剂 I 结合时对应的游离抑制剂的浓度。一般通过 SPR 等实验测定该浓度。

KIBA 数据集是由 Tang 等[12]对 3 种激酶抑制剂的大规模生化分析的靶点选择性进行系统评估，并进一步将这些标准化的生物活性分析与广泛使用的数据库 ChEMBL 和 STITCH 中报告的数据进行比较，将不同来源的药物靶点相互作用数据进行统一整合而得到的数据集。KIBA 原始数据集中包含 476 个靶点蛋白质、52 498 个化学分子和 246 088 个药物靶点相互作用对。在 KIBA 数据集中，药物与蛋白质的相互作用程度的评价指标为 KIBA score。其综合了 IC_{50}（半抑制浓度）、K_i（抑制常数）及 K_d 等信息，其中 IC_{50} 表示对指定的生物过程（或该过程中的某个组分，如激酶、受体、细胞等）抑制一半时所需的药物或激酶抑制剂的浓度，在药学中用于衡量表征拮抗剂进行体外实验时的拮抗能力，一般通过实验绘制量效曲线计算得到。K_i 反映的是激酶抑制剂对靶点的抑制强度，值越小，说明抑制能力越强，其计算公式为

$$K_i = \frac{IC_{50}}{1+[S]/K_m} \tag{2}$$

式中，IC_{50} 为半抑制浓度，[S]为实验底物浓度，K_m 为酶促反应速度达到最大反应速度一半时所对应的底物浓度，它是酶的特征常数。若底物浓度远小于 K_m 值，则 IC_{50} 等于 K_i。IC_{50} 与实验中所用酶的浓度、底物浓度有关，而 K_i 不受这些变量的影响。KIBA 数据集考虑了这 3 种主要生物活性类型的中位数，然后使用参数模型调整 K_i 和 K_d。其计算公式为

$$\begin{aligned} K_i\cdot\mathrm{adj} &= \frac{IC_{50}}{1+L_i(IC_{50}/K_i)} \\ K_d\cdot\mathrm{adj} &= \frac{IC_{50}}{1+L_d(IC_{50}/K_d)} \end{aligned} \tag{3}$$

式中，L_i 和 L_d 是基于模型调整 K_i 和 K_d 的超参数，据此决定 IC_{50} 的权重，这种调整方式的目的是利用 IC_{50} 中的信息来最大化 K_i 和 K_d 之间的一致性。最终，激酶抑制剂生物活性评分（KIBA score）的计算公式为

$$\mathrm{KIBA}=\begin{cases}K_i\cdot\mathrm{adj} & \text{if } IC_{50} \text{ and } K_i \text{ are present}\\ K_d\cdot\mathrm{adj} & \text{if } IC_{50} \text{ and } K_d \text{ are present}\\ (K_i\cdot\mathrm{adj}+K_d\cdot\mathrm{adj})/2 & \text{if } IC_{50},K_i \text{ and } K_d \text{ are present}\end{cases} \tag{4}$$

1.2　输入的表示

对于输入的药物分子，首先使用简化分子线性输入规范（Simplified Molecular Input Line entry System，SMILES）[13]进行表示。这是一种使用短 ASCII 字符串来明确描述分子结构的表示规范。随后将其处理为基于图的药物分子。对一个图来说，其可以表示为 $G=(V,E)$，其中 V 是顶点的集合，E 是边的集合。因此，在一个药物分子中，将其中的原子作为节点，化学键作为边，则其中 $v_i\in V$ 表示第 i 个原子，$e_{ij}\in E$ 表示第 i 个原子和第 j 个原子之间的化学键。然后使用 RDkit[14]将 SMILES 处理为具有节点特征和邻接矩阵的图结构。

对于输入的靶点蛋白质，将其表示为蛋白质序列，序列中的每个字符代表一个氨基酸。然后对蛋白质序列进行映射，将其中的每个字符映射为一个整数。例如，将丙氨酸（a）映射为 1，将半胱氨酸（C）映射为 2，等等。由此能够将蛋白质序列表示为一个整数序列，随后通过一个嵌入层将得到的整数序列映射为一个 128 维的向量。

1.3　基于残差结构的图卷积网络对药物分子的编码

在将药物分子处理为具有节点特征和邻接矩阵的图结构后，本文将残差结构引入图卷积网络（Graph Convolutional Neural Networks，GCN）[15-16]，设计了一个基于残差结构的图卷积网络（Residual Graph Convolutional Networks，RGCN），据此从处理好的药物分子图中学习特征。图 1 为本文提出的 RGCN 的网络架构。RGCN 中包含 4 个残差块，每个残差块后连接一个过渡层，最后经过一个全连接层得到输出。

在 ResNet[17]的启发下，本文将残差连接引入 GCN，设计了 RGCN。在 RGCN 中包含 4 个残差块，每个残差块中包含 3 个残差层，每个残差层中包含 2 层图卷积层。一个残差块内的具体结构如图 2 所示。残差连接的引入使每个残差层经过图卷积层后的输出都能参考前一步的输入，避免梯度消失的问题，同时使 RGCN 能够使用更多的图卷积层，进而达到更深层次的网络结构。由此，残差块内的计算过程为

$$\begin{aligned}x_i^{(1)}&=R(x_i^{(0)})+x_i^{(0)}\\ x_i^{(2)}&=R(x_i^{(1)})+x_i^{(1)}\\ x_i^{(3)}&=R(x_i^{(2)})+x_i^{(2)}\\ x_i^{(N)}&=R(x_i^{(N-1)})+x_i^{(N-1)}\end{aligned} \tag{5}$$

式中，R 表示残差层，$R(x_i^{(j)})$ 表示输入 $x_i^{(j)}$ 经过残差层内的图卷积层后得到的输出。同时，为了使 RGCN 能够有更深层的网络结构，在每两个残差块之间加入一个过渡层，过渡层能够将经过一个残差块计算得到的输出进行批量归一化，减少特征图的通道数，进而节省计算成本，获得整个药物分子图的特征向量。最后，经过 4 个残差块的计算，将得到的结果输入一个全连接层，就能得到药物分子的编码。

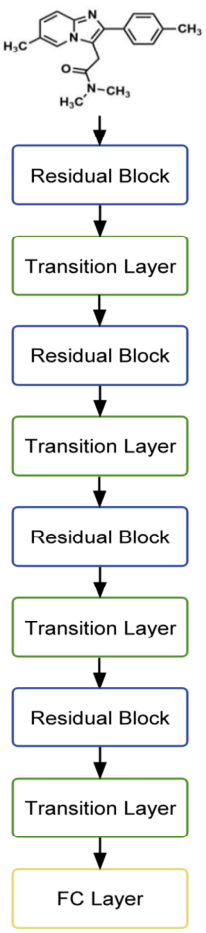


图 1　RGCN 的网络架构

Fig. 1　RGCN network architecture

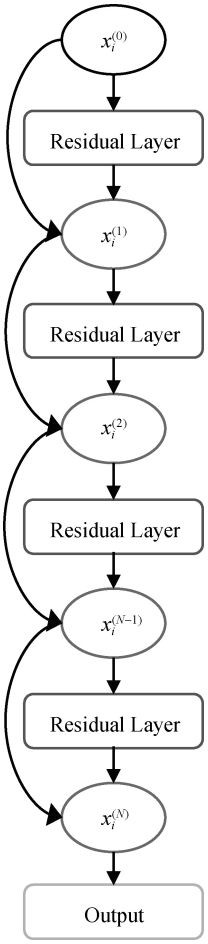


图 2　残差块内的结构

Fig. 2　Residual block internal structure

1.4　多尺度卷积神经网络对靶点蛋白质的编码

针对药物靶点，本文使用一个多尺度卷积神经网络（Multi-scale Convolutional Neural Networks，MCNN）[18]提取靶点蛋白质的特征，这一网络的具体结构如图 3 所示。

类似 Inception[19]的结构，本文使用一个具有 3 个分支的 CNN。这 3 个分支有着不同的感受野，借此捕捉靶点蛋白质不同尺度的特征。通过堆叠不同数量的 3×3 卷积层，让 3 个分支具有不同的感受野，具体为第一个分支使用 1 个 3×3 卷积层，第二个分支使用 2 个 3×3 卷积层，第三个分支使用 3 个 3×3 卷积层。同时，在每个卷积层后都连接一个 ReLu 激活函数，每个分支堆叠的卷积层后都连接一个最大池化层。在前面的操作中，已经将输入的蛋白质序列通过一个嵌入层[20]映射为 128 维的向量。随后使用 MCNN 将这一向量进行映射，将 3 个分支得到的结果进行拼接，得到最终的特征向量。最后将特征向量输入一个全连接层，得到靶点蛋白质的编码。

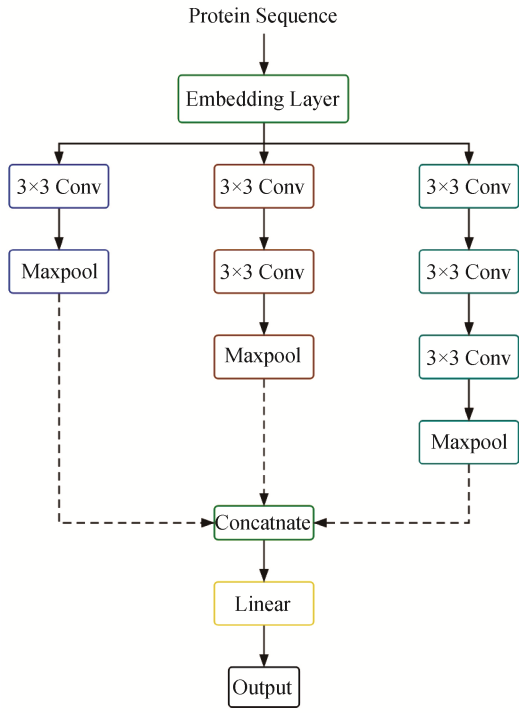


图 3　MCNN 的网络结构

Fig. 3　Multiscale convolutional neural network architecture

1.5　基于残差结构的图卷积网络的药物靶点亲和力预测

在使用 RGCN 和 MCNN 分别对药物分子与靶点蛋白质进行编码后，将得到的两个表示进行拼接，随后将其送入一个 MLP 中预测药物靶点亲和力得分，这一流程如图 4 所示。

本文所定义的 MLP 包含 3 个线性转换层，借此将拼接得到的特征向量映射为药物靶点亲和力得分。同时，在每个线性转换层之后连接一个 ReLu 激活函数，为网络增加非线性。随后在激活函数后引入 Dropout 层，设 dropout = 0.1（失活率为 0.1）。Dropout 层的加入使网络能够缓解过拟合现象，增加网络的鲁棒性。本文使用均方误差作为损失函数，具体公式为

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(P_i-Y_i)^2 \qquad (6)$$

式中，P_i 表示对第 i 个药物分子和第 i 个靶点蛋白质进行编码后，使用 MLP 预测的药物靶点亲和力得分；Y_i 表示数据集中对应的第 i 个药物靶点对之间的真实亲和力值；n 为样本总量。

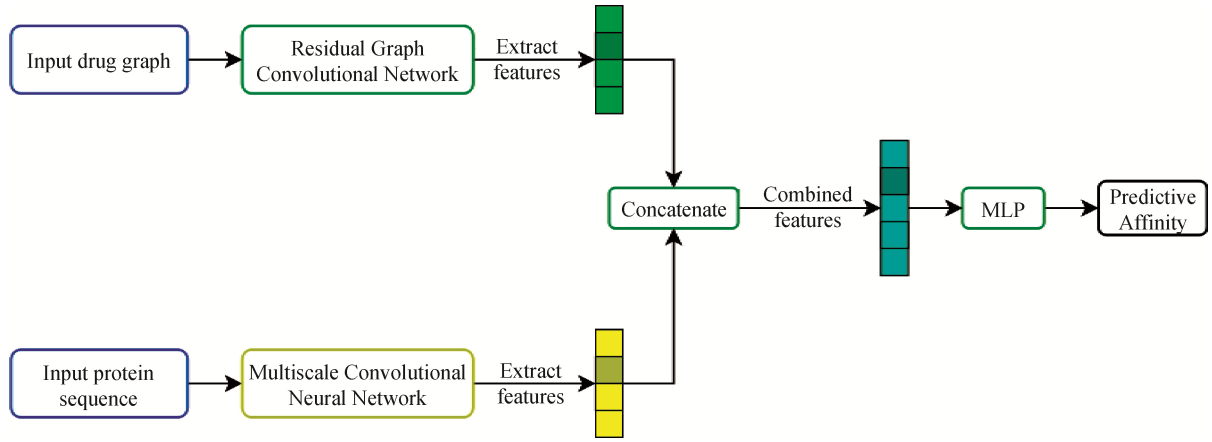


图 4　本文所提模型的药物靶点亲和力预测流程

Fig. 4　The drug target affinity prediction process of the model proposed in this paper

2　实验与结果

为了对本文提出的模型进行评估，将所提出的模型与 8 种经典的或先进的模型在 Davis 与 KIBA 这两个数据集上进行对比。在 8 个模型中，KronRLS[21]、SimBoost[22]、支持向量机（Support Vector Machine，SVM）及随机森林（Random Forest，RF）这 4 个模型是基于机器学习的药物靶点亲和力预测模型。其中，前两个模型都使用了相似性矩阵对药物分子和靶点蛋白质进行嵌入，然后使用机器学习算法进行预测；后两个模型都使用了 PSC 和 ECFP 分别对药物分子和靶点蛋白质进行嵌入，然后分别使用 SVM 和 RF 算法进行预测。DeepDTA 和 WideDTA 这两个模型基于深度学习，将 CNN 应用于药物分子和靶点蛋白质的嵌入，WideDTA 模型在此基础上还融合了蛋白质结构域和基序（Protein Domains and Motifs，PDM）及配体最大公共亚结构（Ligand Maximum Common Substructures，LMCS）信息作为辅助进行预测。GraphDTA 和 DeepAffinity 这两个模型则是基于 GNN 的模型。其中，GraphDTA 有 4 个版本。对于药物分子，分别使用 GCN、图注意力网络（Graph Attention Networks，GAT）、图同构网络（Graph Isomorphism Networks，GIN）及图注意力网络和图卷积网络的组合网络对药物分子进行嵌入；对于靶点蛋白质，这 4 个版本都使用 CNN 进行嵌入。DeepAffinity[23]有 5 个版本。对于药物分子，这 5 个版本分别使用循环神经网络（Rerrent Neural Networks，RNN）、GCN、GIN 等对药物分子进行嵌入；对于靶点蛋白质，这 5 个版本分别使用 RNN、CNN 和及分层递归神经网络（Hierarchical Recurrent Neural Networks，HRNN）等进行嵌入。

以上 8 个模型是本文进行比较研究的基线模型。对网络层数这一超参数的选择，本文引入的残差结构源于 ResNet，因此最初图卷积层个数为 34 层，旨在通过更深的网络结构捕获药物分子更多维度的信息，但此时模型存在过拟合现象。随后通过不断实验，利用引入 Dropout 层、降低网络层数、调节训练轮次等方法抑制过拟合现象。通过对比实验结果，最终确定图卷积层个数为 24 层。

同时，本文使用均方误差（Mean Square Error，MSE，该值越小代表模型性能越优）、一致性指数（Consistenly Index，CI，该值越大代表模型性能越优）和 r^{m^2} 指标（该值越大代表模型性能越优）[6, 24]作为模型的性能指标来评估模型的回归表现。其中，MSE 能够反映预测值和真实值之间的平均差异程度，详细解释请参阅 1.5 节；CI 值常用于回归问题，它能反映模型的排序能力，其计算公式为

$$\mathrm{CI}=\frac{1}{Z}\sum\nolimits_{b_i>b_j} h(y_i-y_j) \tag{7}$$

式中，y_i 代表具有更强药物靶点亲和力值 b_i 的样本的预测值；y_j 代表具有更小药物靶点亲和力值 b_j 的样本的预测值；Z 为标准化常量；$h(x)$ 为阶跃函数，其计算公式为

$$h(x)=\begin{cases}1 & x>0\\ 0.5 & x=0\\ 0 & x<0\end{cases} \tag{8}$$

r^{m^2} 指标能够评价模型的外部预测性能，当测试集的 r^{m^2} >0.5 时，代表该模型能够被接受，其计算公式[25]为

$$r^{m^2} = r^2(1-\sqrt{r^2-r_0^2}) \quad (9)$$

式中，r^2 和 r_0^2 分别是带截距的平方相关系数和不带截距的平方相关系数。本文所提模型与基线模型在 Davis 数据集上的结果对比如表 2 所示，在 KIBA 数据集上的结果对比如表 3 所示。两个表中 8 种基线模型的实验结果来自 DeepDTA 的实验结果及各模型的原始论文。同时，本文对 Davis 和 KIBA 原始数据集进行了重新划分，按照 8∶1∶1 的比例将两个数据集分别随机划分为训练集、验证集和测试集，使用训练集和验证集对模型进行训练并调优，最终在测试集上验证本文所提出的模型。

表 2　本文所提模型与基线模型在 Davis 数据集上的结果对比

Tab. 2　Comparison results of the proposed model and baselines on the Davis dataset

模型	药物分子	靶点蛋白质	MSE	CI	r^{m^2} 指标
DeepDTA	CNN	CNN	0.261	0.878	0.603
WideDTA	CNN+LMCS	CNN+PDM	0.262	0.886	—
GraphDTA	GCN	CNN	0.254	0.880	—
GraphDTA	GAT	CNN	0.232	0.892	—
GraphDTA	GIN	CNN	0.229	0.893	—
GraphDTA	GAT-GCN	CNN	0.245	0.881	—
DeepAffinity	RNN	RNN	0.253	0.900	—
DeepAffinity	GCN	RNN	0.260	0.881	—
DeepAffinity	GCN	CNN	0.657	0.737	—
DeepAffinity	GCN	HRNN	0.252	0.881	—
DeepAffinity	GIN	HRNN	0.436	0.822	—
KronRLS	PubChem Sim	S-W	0.379	0.871	0.407
SimBoost	PubChem Sim	S-W	0.282	0.872	0.655
RF	PSC	ECFP	0.359（0.003）	0.854（0.002）	0.549（0.005）
SVM	PSC	ECFP	0.383（0.002）	0.857（0.001）	0.513（0.003）
本文所提模型	**RGCN**	**MCNN**	**0.218（0.001）**	**0.903（0.001）**	**0.735（0.004）**

表 3　本文所提模型与基线模型在 KIBA 数据集上的结果对比

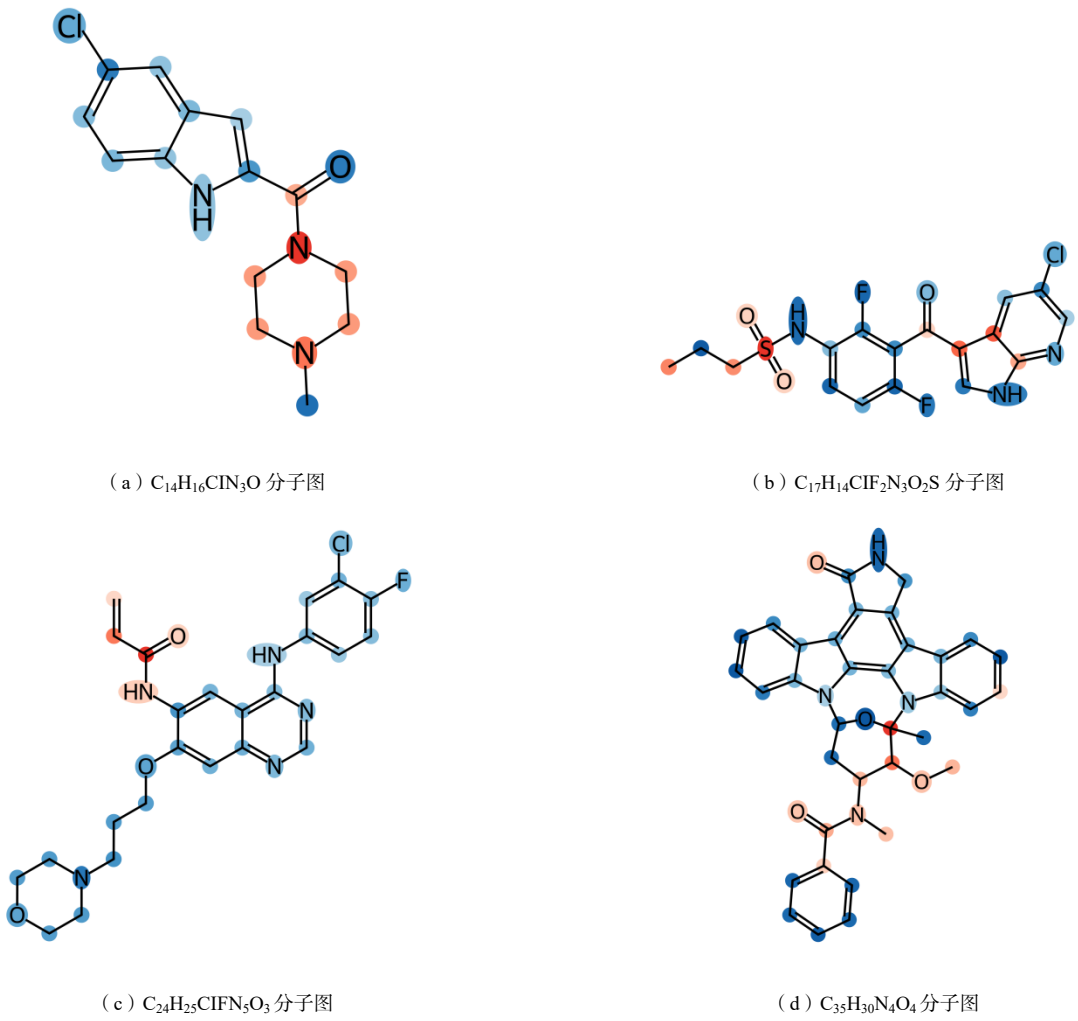
Tab. 3　Comparison results of the proposed model and baselines on the KIBA dataset

模型	药物分子	靶点蛋白质	MSE	CI	r^{m^2} 指标
DeepDTA	CNN	CNN	0.194	0.863	0.673
WideDTA	CNN+LMCS	CNN+PDM	0.179	0.875	—
GraphDTA	GCN	CNN	0.139	0.889	—
GraphDTA	GAT	CNN	0.179	0.866	—
GraphDTA	GIN	CNN	0.147	0.882	—
GraphDTA	GAT-GCN	CNN	0.139	0.891	—
DeepAffinity	RNN	RNN	0.188	0.842	—
DeepAffinity	GCN	RNN	0.288	0.797	—
DeepAffinity	GCN	CNN	0.680	0.576	—
DeepAffinity	GCN	HRNN	0.201	0.842	—
DeepAffinity	GIN	HRNN	0.445	0.689	—
KronRLS	PubChem Sim	S-W	0.411	0.782	0.342
SimBoost	PubChem Sim	S-W	0.222	0.836	0.629
RF	PSC	ECFP	0245（0.001）	0.837（0.000）	0.581（0.000）
SVM	PSC	ECFP	0.308（0.003）	0.799（0.001）	0.513（0.004）
本文所提模型	**RGCN**	**MCNN**	**0.150（0.003）**	**0.891（0.001）**	**0.791（0.001）**

由表 2 可见，本文所提模型在 Davis 数据集上相较于 KronRLS、SimBoost、RF、SVM 这 4 种基于机器学习的模型，在 3 种评价指标下准确率有比较明显的提升；同时相较于 DeepDTA、WideDTA 这两种基于深度学习的模型及 GraphDTA、DeepAffinity 这两种基于 GNN 的模型，本文所提模型的预测准确率也有一

定的提升。由表 3 可见，本文所提模型在 KIBA 数据集上同样有着优秀的表现，相较于 8 个基准模型，本文所提模型的 MSE、CI、r^{m^2} 得分都有明显提升。结果表明，本文所提模型能够对药物分子和靶点蛋白质进行有效的嵌入表达，并且拥有更加优秀的亲和力预测结果。

基于深度学习的模型通常被称为“黑盒”模型。虽然这类模型的预测效果很好，但是很难对预测结果进行追溯，明确模型的作用机理。因此，本文在 Davis 数据集上使用表 1 中所得结果的模型的权重，对药物分子的原子重要性进行了可视化[18]，所得的部分结果如图 5 所示。



（a）$C_{14}H_{16}ClN_3O$ 分子图　（b）$C_{17}H_{14}ClF_2N_3O_2S$ 分子图

（c）$C_{24}H_{25}ClFN_5O_3$ 分子图　（d）$C_{35}H_{30}N_4O_4$ 分子图

图 5　部分药物分子的原子重要性可视化

Fig. 5　Atom importance visualization for part of molecules

在图 5 中，原子呈蓝色，代表其在药物靶点相互作用中的重要程度较低，蓝色越深，重要程度越低。原子呈红色，代表其重要程度较高，红色越深，重要程度越高。如图 5（a）所示的分子的化学式为 $C_{14}H_{16}ClN_3O$，它是一种组胺 H4 受体拮抗剂，半数 IC_{50} 值为 4.5nM。它是第一个强效的且具有选择性的非咪唑类组胺 H4 受体拮抗剂，其 K_i 值为 4.5nM，与其他组胺受体相比，具有超过它们 1 000 倍的选择性[26]。同时，它与 H4 受体拮抗剂能够以非常高的亲和力结合，具有比其他组胺受体拮抗剂更高的选择性[27]。如图 5（b）所示的分子的化学式为 $C_{17}H_{14}ClF_2N_3O_3S$，它是一种 BRAF 激酶抑制剂，是通过结构导向的方法发现的一种 7-氮杂吲哚衍生物，能够抑制 B-RafV600E，IC_{50} 值为 13nM。B-RafV600E 是一种频繁发生的致癌蛋白激酶突变。与

其他激酶相比，该分子对 B-RafV600E 具有高选择性[28]。如图 5（c）所示的分子的化学式为 $C_{24}H_{25}ClFN_5O_3$，它是一种 3-氯代、4-氟代的 4-苯胺喹唑啉，是一种可口服、有效且不可逆的 Pan-erbB 酪氨酸激酶抑制剂，可在体外抑制 EGFR、HER2 和 HER4，IC_{50} 值分别为 0.8nM、19nM 和 7nM[29]。同时，其 C6 位点的丙烯酰胺侧链与 erbB 家族的半胱氨酸靠近，能迅速形成共价键，使 erB1、erB2 和 erB4 家族成员永久失去催化活性，有效抑制 erbB3 依赖的信号通路[30]。从图中可以看出，本文所提模型将 C6 位点的丙烯酰胺侧链标为红色，证明了模型的准确性。如图 5（d）所示的分子的化学式为 $C_{35}H_{30}N_4O_4$，它是一种 PKC 抑制剂，对大多数 PKC 亚型具有高效的抑制活性。同时，它对胞内 PKC 具有可逆的抑制活性，其 IC_{50} 值为 0.5μM。它还能抑制 PDGF/VEGF 和干细胞因子受体的自磷酸化。它通过诱导细胞周期 G2/M 停滞和细胞凋亡，从而抑制多种细胞系的细胞生长[31]。

3　讨论

本文提出了一种基于残差结构的深度图卷积网络（RGCN）来对药物分子进行嵌入学习，残差结构的引入使网络结构大大加深，利用一个具有 24 个图卷积层的 RGCN 来捕获药物分子的特征，同时利用一个深度卷积神经网络对靶点蛋白质进行嵌入学习，随后进行药物靶点亲和力预测。通过将本文所提模型在 Davis 和 KIBA 这两个数据集上进行实验，与几种经典的或当前表现较好的模型进行对比研究，证明本文所提模型在 MSE、CI、r^{m^2} 3 种评估标准下优于其他先进的方法，证明了本文所提模型的有效性和优越性。虽然本文所提模型在药物靶点亲和力预测中取得了明显的改进，但是也存在一定的局限性。本文使用的两个公开数据集的数据量相对较小，在更大规模数据集上的效果仍需进一步实验和调整。同时，对于靶点蛋白质的嵌入方式也可以进一步改进，以进一步提高药物靶点亲和力预测的准确率。因此，在未来的研究中可以在这两方面进一步优化模型。

参考文献

[1] SCARSELLI F, GORI M, TSOI A C, *et al*. The graph neural network model[J]. **IEEE transactions on neural networks,** 2008, 20(1): 61-80.

[2] HINTON G, DENG L, YU D, *et al*. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. **IEEE Signal Processing Magazine**, 2012, 29(6): 82-97.

[3] WEN M, ZHANG Z, NIU S, *et al*. Deep-learning-based drug–target interaction prediction[J]. **Journal of proteome research**, 2017, 16(4): 1401-1409.

[4] CHENG Z, ZHOU S, WANG Y, *et al*. Effectively identifying compound-protein interactions by learning from positive and unlabeled examples[J]. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2016, 15(6): 1832-1843.

[5] RIFAIOGLU A S, CETIN A R, CANSEN K D, *et al*. MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery[J]. **Bioinformatics**, 2021, 37(5): 693-704.

[6] ÖZTÜRK H, ÖZGÜR A, OZKIRIMLI E. DeepDTA: deep drug–target binding affinity prediction[J]. **Bioinformatics**, 2018, 34(17): i821-i829.

[7] ÖZTÜRK H, OZKIRIMLI E, ÖZGÜR A. WideDTA: prediction of drug-target binding affinity[J]. **arXiv preprint arXiv**, 1902. 04166, 2019.

[8] TSUBAKI M, TOMII K, SESE J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. **Bioinformatics**, 2019, 35(2): 309-318.

[9] NGUYEN T, LE H, QUINN T P, *et al*. GraphDTA: Predicting drug–target binding affinity with graph neural networks[J]. **Bioinformatics**, 2021, 37(8): 1140-1147.

[10] JIANG M, LI Z, ZHANG S, *et al*. Drug–target affinity prediction using graph neural network and contact maps[J]. **RSC advances**, 2020, 10(35): 20701-20712.

[11] DAVIS M I, HUNT J P, HERRGARD S, *et al*. Comprehensive analysis of kinase inhibitor selectivity[J]. **Nature biotechnology**, 2011, 29(11): 1046-1051.

[12] TANG J, SZWAJDA A, SHAKYAWAR S, *et al*. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis[J]. **Journal of Chemical Information and Modeling**, 2014, 54(3): 735-743.

[13] WEININGER D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. **Journal of chemical information and computer sciences**, 1988, 28(1): 31-36.

[14] LANDRUM G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling[J]. **Greg Landrum**, 2013, 8: 31.

[15] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. **arXiv preprint arXiv**: 1609. 02907, 2016.

[16] BRUNA J, ZAREMBA W, SZLAM A, *et al*. Spectral networks and locally connected networks on graphs[J]. **arXiv preprint arXiv**: 1312. 6203, 2013.

[17] HE K, ZHANG X, REN S, *et al*. Deep residual learning for image recognition: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016[C].

[18] YANG Z, ZHONG W, ZHAO L, *et al*. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction[J]. **Chemical science**, 2022, 13(3): 816-833.

[19] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al*. Rethinking the inception architecture for computer vision: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016[C].

[20] LEE I, KEUM J, NAM H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences[J]. **PLoS computational biology**, 2019, 15(6): e1007129.

[21] PAHIKKALA T, AIROLA A, PIETILÄ S, *et al*. Toward more realistic drug–target interaction predictions[J]. **Briefings in bioinformatics**, 2015, 16(2): 325-337.

[22] HE T, HEIDEMEYER M, BAN F, *et al*. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines[J]. **Journal of cheminformatics**, 2017, 9(1): 1-14.

[23] KARIMI M, WU D, WANG Z, *et al*. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks[J]. **Bioinformatics**, 2019, 35(18): 3329- 3338.

[24] PRATIM ROY P, PAUL S, MITRA I, *et al*. On two novel parameters for validation of predictive QSAR models[J]. **Molecules**, 2009, 14(5): 1660-1701.

[25] ROY K, CHAKRABORTY P, MITRA I, *et al*. Some case studies on application of " r^{m^2} " metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data[J]. **Journal of computational chemistry**, 2013, 34(12): 1071-1082.

[26] JABLONOWSKI J A, GRICE C A, CHAI W, *et al*. The first potent and selective non-imidazole human histamine H4 receptor antagonists[J]. **Journal of medicinal chemistry**, 2003, 46(19): 3957-3960.

[27] THURMOND R L, DESAI P J, DUNFORD P J, *et al*. A potent and selective histamine H4 receptor antagonist with anti-inflammatory properties[J]. **Journal of Pharmacology and Experimental Therapeutics**, 2004, 309(1): 404-413.

[28] TSAI J, LEE J T, WANG W, *et al*. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity[J]. **Proceedings of the National Academy of Sciences**, 2008, 105(8): 3041-3046.

[29] ARKIN M, MOASSER M M. HER2 directed small molecule antagonists[J]. **Current Opinion in Investigational Drugs (London, England: 2000)**, 2008, 9(12): 1264.

[30] CALVO E, TOLCHER A W, HAMMOND L A, *et al*. Administration of CI-1033, an irreversible pan-erbB tyrosine kinase inhibitor, is feasible on a 7-day on, 7-day off schedule: a phase I pharmacokinetic and food effect study[J]. **Clinical Cancer Research**, 2004, 10(21): 7112-7120.

[31] PANAGI I, JENNINGS E, ZENG J, *et al*. Salmonella effector SteE converts the mammalian serine/threonine kinase GSK3 into a tyrosine kinase to direct macrophage polarization[J]. **Cell host & microbe**, 2020, 27(1): 41-53.