Hans Journal of Computational Biology 计算生物学, 2024, 14(1), 1-11
Published Online March 2024 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2024.141001

Hans汉斯

GraphPLA：一种预测蛋白质－配体结合亲和力的图神经网络方法

郭岳松，刘立伟*

大连交通大学理学院，辽宁 大连

收稿日期：2024年4月22日；录用日期：2024年5月22日；发布日期：2024年5月30日

摘　要

蛋白质－配体结合亲和力是蛋白质与配体相互作用强度的一个重要指标。准确预测蛋白质－配体结合亲和力对于发现药物的新应用至关重要。迄今为止，已经开发了许多计算技术来预测结合亲和力；然而，这些技术中的一部分需要并不常见的蛋白质三维结构，一些方法还将配体表示为不是分子的适当表示的SMILES串。为了避免这些问题，本文开发了一种名为GraphPLA的新模型，使用具有直接结合配体的独特特征的蛋白质结合口袋作为局部输入特征。还使用扩展卷积来捕捉蛋白质的多尺度远程相互作用，图神经网络来学习配体的图表示。实验结果表明，GraphPLA的RMSE为1.388，MAE为1.118，R为0.795，SD为1.345，CI为0.796，优于目前最先进的预测方法，可以有效预测蛋白质－配体结合亲和力。

关键词

蛋白质－配体结合亲和力，图卷积神经网络，深度学习，蛋白质－配体结合口袋，图注意力网络

GraphPLA: A Graph Neural Networks Method to Predict Protein-Ligand Binding Affinity

Yuesong Guo, Liwei Liu*

College of Science, Dalian Jiaotong University, Dalian Liaoning

Received: Apr. 22nd, 2024; accepted: May 22nd, 2024; published: May 30th, 2024

Abstract

Protein-ligand binding affinity is an important indicator of the strength of protein-ligand interac-

*通讯作者。

文章引用：郭岳松，刘立伟．GraphPLA：一种预测蛋白质-配体结合亲和力的图神经网络方法[J]．计算生物学，2024，14(1)：1-11．DOI：10.12677/hjcb.2024.141001

tions. Accurately predicting protein-ligand binding affinity is crucial for discovering new applications for drugs. To date, many computational techniques have been developed to predict binding affinity. However, some of these technologies require uncommon protein three-dimensional structures, and some methods also represent ligands as SMILES strings that are not appropriate representations of molecules. To avoid these issues, this paper develops a new model called GraphPLA, which uses protein binding pockets with unique features of directly binding ligands as local input features. Extended convolution is also used to capture multi-scale remote interactions of proteins, and a graph neural network is used to learn the graph representation of ligands. The experimental results show that the RMSE of GraphPLA is 1.388, MAE is 1.118, R is 0.795, SD is 1.345, and CI is 0.796, which is superior to the most advanced prediction methods and can effectively predict protein-ligand binding affinity.

Keywords

Protein-Ligand Binding Affinity, Graph Convolutional Neural Network, Deep-Learning, Protein-Ligand Pocket, Graph Attention Network

1. 引言

研究配体与靶蛋白之间的相互作用是药物开发研究的一个重要目标。而蛋白质－配体相互作用的关键研究领域包括结合位点、结合模式和结合亲和力[1]。蛋白质－配体之间连接的强度体现在结合亲和力上，通常以抑制常数 K_i、解离常数 K_d、半最大抑制浓度 IC_{50} 等为特征。准确预测蛋白质－配体结合亲和力是研究生物分子作用机制、药物设计和再利用的基础。

以往的研究多基于分子力学模拟，如分子对接和分子动力学模拟。然而，由于该方法结构复杂，计算成本高，在预测精度和效率方面都面临着很大的挑战。一些基于相似性的或基于矩阵分解的方法可以通过使用整个蛋白质或配体的全局相似性矩阵来给出预测。这些方法的局限性在于忽略了每个分子中单个组分的详细特征。基于 SVM 和随机森林的蛋白质－配体相互作用模型的研究大多局限于二元系统。随着机器学习技术在过去几年的发展，它们已逐渐被用于预测蛋白质－配体结合亲和力。例如，Pafnucy [2]，DeepAtoms [3]，DEELIG [4]，TopologyNet [5]。其中，基于分子特征描述符的机器学习模型是目前研究的热点。如 DeepDTA [6]和 WideDTA [6]。然而，在特征选择和建模方面仍然存在许多困难。近年来，卷积神经网络(CNN)、递归神经网络(RNN)等技术通常用于预测蛋白质－配体结合亲和力[7] [8]。该方法可以在分子水平上了解蛋白质与配体的结构和相互作用模式，从而更好地掌握蛋白质与配体之间的关系和空间特征[9] [10]。此外注意力机制[11]、图神经网络等新技术也被引入到预测模型中，以提高预测精度[12]。

本文使用图神经网络创建了一个新的蛋白质－配体结合亲和力预测模型 GraphPLA。GraphPLA 利用具有直接结合配体独特性质的蛋白质结合口袋作为局部输入特征；利用扩展卷积捕捉蛋白质的多尺度远程相互作用；利用配体的结构信息，通过学习配体的图表达，获得配体之间的相互关系和空间特征。本文还将本文提出的算法与其他类似算法进行了比较和分析。结果表明，GraphPLA 是一种可靠的预测蛋白质－配体结合亲和力的模型。

2. 材料和方法

2.1. 数据准备

本文从 PDBbind 2016 数据库中提取了三组数据。一般集、精细集、核心集，其中分别有 9226、4057、290 个高质量亲和力数据和蛋白质－配体复合物。为了确保三组数据不重复，从精细集和核心集中去除核心集中的蛋白质－配体复合物。从常规集和精细集的样本中随机选择 1000 个样本作为验证集，核心集作为测试集。此外，为了便于与 Pafnucy 进行比较，从验证集删除了 85 个蛋白质－配体复合物，从训练集删除了 2 个蛋白质－配位体复合物。由于 PDBbind 数据库中的小分子在正常条件下是带电的，为了使用 RDKit 进行结构表征，移除 42 个未修饰的蛋白质－配体复合物。最终，共采集了 290 个测试样本、1000 个确认样本和 11864 个训练样本。在此基础上，本文还用现有的数据集 test105 (包括 105 个样本)和 test71 (包括 71 个样本)进行测试，使其更加客观。

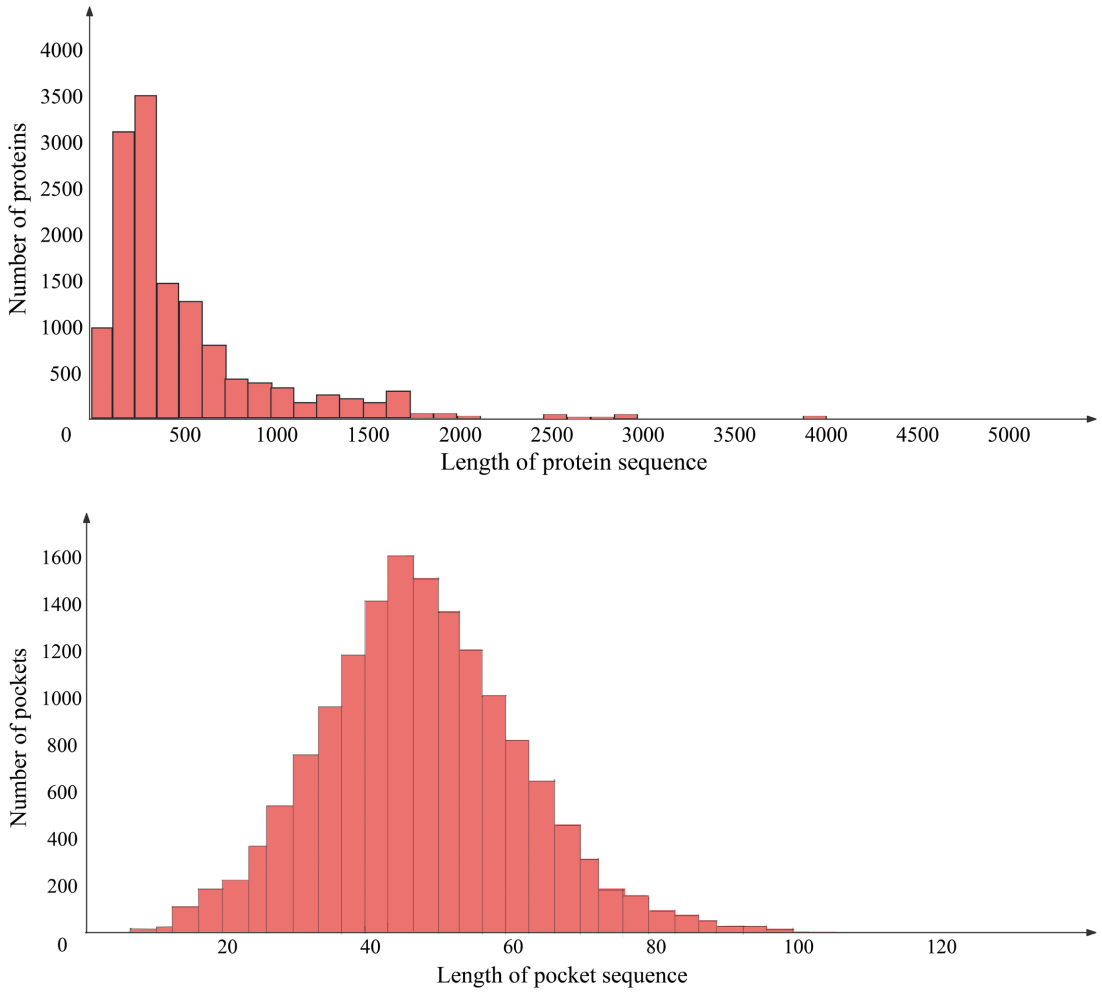

Figure 1. Length statistics for all study data
图 1. 所有研究数据的长度统计

根据图 1 所示的分布进行比较。为了覆盖 90%的蛋白质和 90%的口袋，为蛋白质定义了 1000 个固定长度，为口袋定义了 63 个固定长度。长度大于固定长度的被截断，长度小于固定长度的被补零。使用分子图描述配体的结构。基于文本的输入信息分为配体表示、蛋白质表示和口袋表示三部分。

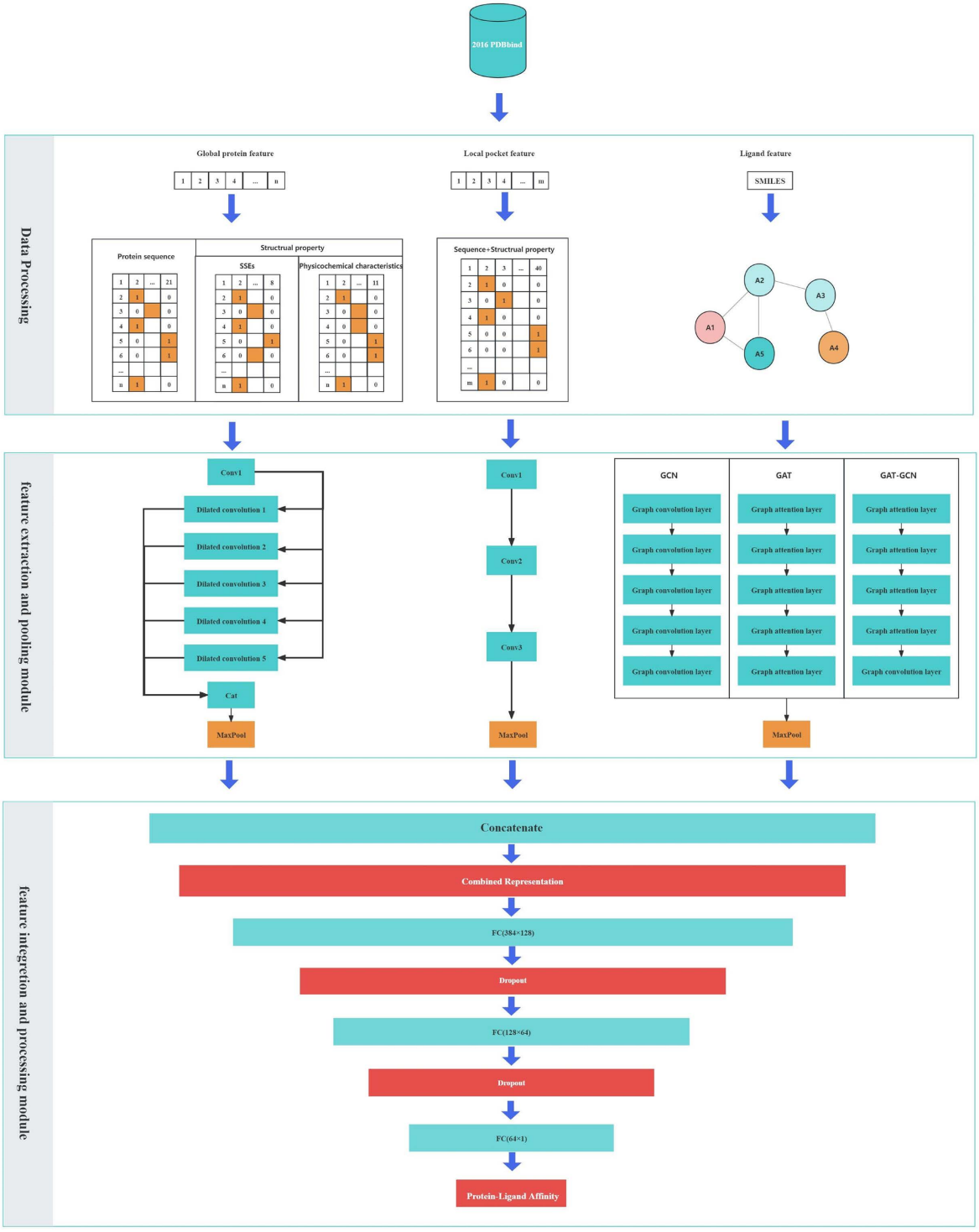

Figure 2. The GraphPLA architecture. Includes three modules: data processingmodule, feature extraction and pooling module, and feature integration and processing module
图 2. GraphPLA 架构。包括三个模块：数据处理模块、特征提取和池化模块以及特征集成和处理模块

2.2. 蛋白质表示(图 2)

本文对每个残基使用 40-D 特征向量，通过整合序列和结构属性表示来描述全局蛋白质特征。21-D

独热向量编码了 21 种不同蛋白质序列残基，由 20 种不同类型的氨基酸和非标准残基组成。此外，利用 19-D 向量用来表示结构属性，包括二级结构和理化性质。

本研究使用 SSPro 软件预测每个序列的二级结构。有八种不同的二级结构状态：α-helix (H)，extended strand，participates in β ladder (E)，residue in isolated β-bridge (B)，coil (C)，hydrogen bonded turn (T)，π-helix (I)，3_{10} helix (G)，and bend (S)。本文使用一个 8-D 独热向量来编码二级结构。在此基础上，将各种残留物分为非极性、极性、酸性和碱性，并对每种残留物的理化特性进行了检测。为此，提出了一种基于 11-D 向量的物理化学性质编码方法。

2.3. 口袋表示

口袋由几个不连续的序列组成，其中包含某些蛋白质的重要氨基酸，对蛋白质－配体的功能有重要影响。因此，将一个口袋作为整体进行局部特征提取。局部口袋特征是预测蛋白质－配体结合亲和力的重要输入信息。本文拟使用一个 40-D 特征向量，将前一节提出的序列表达式和结构属性表达式结合起来，对局部口袋特征进行编码。

2.4. 配体表示

SMILES 的开发目标是表示计算机可读的分子。本文将药物分子视为原子相互作用图。为了描述图中的节点，本文采用了一组 DeepChem 改编的原子特征。在本文中，每个节点是一个多维二元特征向量，表达五个信息：原子符号、相邻原子数、相邻氢原子数、原子隐式值以及原子是否为芳香结构。本文将 SMILES 代码转换为分子图，使用 RDKit 提取原子特征。

3. 模型

本文使用嵌入层来表示三个模块中的输入。这些模块分别由(1000, 128)维蛋白质矩阵，(63, 128)维口袋矩阵和配体图组成。在蛋白质模块中，使用具有 5 个不同扩展率的一维扩展卷积来描述较长的蛋白质序列，之后是最大池化层。在配体模块中，使用分子图并试验了三种不同的图神经网络模型，包括 GCN [13]、GAT [14]和 GAT-GCN [15]。之后是最大池化层。在口袋模块中，使用了三个一维的传统卷积，卷积层由 32、64、128 个滤波器组成，滤波器的大小为 3。然后是最大池化层。最后，将三个模块的最大池化层的特征连接在一起，并提供给分类部分。

分类部分由三个 FC 层组成。第一层具有 128 个节点，而第二层有 64 个节点。每一层都有一个速率为 0.5 的脱落层。最后一个 FC 层之后为输出层。

PReLU 激活函数存在于该架构的 FC 层、蛋白质卷积层和口袋部分，目的是缩短训练时间，避免过拟合。这个函数的表达式定义为：

$$f(x)=\begin{cases}x & \text{if } x>0\\ ax & \text{if } x<0\end{cases}$$

其中 a 表示可学习的参数。对于配体部分，本文采用 ELU 激活函数，它不会引起梯度饱和。这个函数的表达式定义为：

$$f(x)=\begin{cases}x & \text{if } x>0\\ a\left(e^{x}-1\right) & \text{if } x\le 0\end{cases}$$

其中 a 是超参数。设置为 1.0。

本文使用 MSELoss 作为损失函数，为了最小化损失函数，使用一种名为 AdamW 的优化器对模型中的参数进行了优化。

综上所述，本文提出了一种结合局部、全局特征和分子图的模型。

3.1. 基于 GCN 的图表示学习

本文拟采用 GCN [13]对药物的图形表示进行建模。形式上，用 $G=(V,E)$ 表示给定药物的曲线图，其中 V 是 N 个节点的集合，每个节点由一个 C 维向量表示，E 是用邻接矩阵 a 表示的边的集合。多层图卷积网络(GCN)以一个节点特征矩阵 $X\in R^{N\times C}$ ($N=|V|$，C：每个节点的特征数)和邻接矩阵 $A\in R^{N\times N}$。然后产生一个节点级输出 $Z\in R^{N\times F}$ (F：每个节点输出特征的数量)。GCN 层与层之间的传播方式是：

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

其中，$\tilde{A}=A+I_N$ 为带有自连接的无向图的邻接矩阵，$\tilde{D}_{ii}=\sum_i\tilde{A}_{ii}$，其中，$H^{(l)}\in R^{N\times C}$ 是第 l 层的激活矩阵，$H^{(0)}=X$，σ 为激活函数，W 为可学习参数。

可以近似地使用分层卷积操作：

$$Z=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$$

其中 $\Theta\in R^{C\times F}$ (F：滤波器或滤波器映射的数量)是滤波器参数的矩阵，$Z\in R^{N\times F}$ 是卷积信号矩阵。而在预处理步骤中计算出 $\hat{A}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ 后，我们可将其简化为

$$Z=f(X,A)=\mathrm{softmax}\left(\hat{A}\,\mathrm{Re}\,LU\left(\hat{A}XW^{(0)}\right)W^{(1)}\right)$$

其中 $W^{(0)}\in R^{C\times H}$ 是具有 H 个特征图的隐藏层的隐藏权重矩阵的输入，$W^{(1)}\in R^{H\times F}$ 是一个隐藏的输出权重矩阵。softmax 激活函数定义为：

$$\mathrm{softmax}(x_i)=\frac{1}{Z}\exp(x_i)$$

其中 $Z=\sum_i\exp(x_i)$。

本文提出了一种新的 GCN 算法，该算法使用 5 个相邻的 GCN 层，由 ReLU 函数激活，然后增加一个全局最大池化层来获得图形表示向量。

3.2. 基于 GAT 的图表示学习

图注意力层是图注意力网络(GAT) [14]体系结构的构建块。图注意力层的输入是一个节点特征向量集：

$$\boldsymbol{h}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N\},\boldsymbol{h}_i\in R^F$$

其中 N 为节点个数，F 为节点特征的个数，矩阵 $\boldsymbol{h}$ 的大小是 $N\times F$，代表了所有节点的特征，而 R 只代表了某一个节点的特征，所以它的大小为 $F\times 1$。图注意力层的输出是一个新的节点特征向量集：

$$\boldsymbol{h}'=\{\boldsymbol{h}_1',\boldsymbol{h}_2',\cdots,\boldsymbol{h}_N'\},\boldsymbol{h}_i'\in R^F$$

然后 GAT 层使用加权矩阵 $\boldsymbol{W}$ 对每个节点执行线性变换，对于图中的每个输入节点 i，i 与其一阶邻居之间的注意系数计算为：

$$e_{ij}=a\left(\boldsymbol{W}\boldsymbol{h}_i,\boldsymbol{W}\boldsymbol{h}_j\right)$$

这个数字表示节点 j 对节点 i 的重要性。为了使得注意力系数更容易计算和便于比较，本文引入了 softmax 对所有的 i 的相邻节点 j 进行正则化：

$$\alpha_{ij}=\mathrm{softmax}(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})}$$

然后通过使用 softmax 算法对这些注意力系数进行归一化来确定节点的输出特征：

$$\boldsymbol{h}_1' = \sigma\left(\sum_{j\in N(i)} \alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\right)$$

归一化注意力系数由 α_{ij} 表示，非线性激活函数由 $\sigma(.)$ 表示。

本文中基于 GAT 的图学习架构包括五个 GAT 层，由 ELU 函数激活。前四层使用多头注意力，将头的数量设置为 10，输出特征的数量与输入特征的数量相对应；第五层的输出特征设为 128。

3.3. GAT-GCN 组合图神经网络

本文还对 GAT-GCN 组合模型进行了研究[15]。该模型从 GAT 层开始，将图像作为输入传递给后续的 GCN 层。ELU 函数激活每个 GAT 层，并且通过 ReLU 函数触发 GCN 层。在此基础上，将 GCN 层的整体最大池化层与整体平均池化层结合，得到最终的图向量。

3.4. 局部和全局特征

在生物学研究中，DNA 片段产生的关键氨基酸及其相互作用具有重要意义[6]。研究表明，蛋白质的局部和全局特征是决定其功能和相互作用的关键因素。因此，本文构建了一个深度学习模型，通过整合蛋白质结合口袋序列和整个蛋白质序列的局部和全局特征来捕获不同输入位置的重要性。

3.5. 扩张卷积

与传统卷积相比，扩展卷积可以通过设置不同的扩展率来捕获多尺度上下文信息，并且在不损失分辨率和覆盖范围的情况下支持感受野的指数扩展。扩展卷积算子 $*_l$ 定义为

$$(F*_l k)(P) = \sum_{s+lt=P} F(s)k(t)$$

这里 $F: Z^2 \to R$ 是一个离散函数，$k: \Omega_1 \to R$ 是膨胀率。元素向量的下标为 s 和 t。应用膨胀率呈指数增长的滤波器的离散函数可以定义为：

$$F_{i+1} = F_i *_{2^i} k_i \quad for\, i = 0,1,\cdots,n-2$$

在此基础上，通过增加有效感受野大小，使用扩展卷积来捕获蛋白质信息的远程相互作用。蛋白质模块有 5 层，使用 3 × 3 卷积核，膨胀率为 1，2，4，8，16。

3.6. 评估指标

为了评估模型的性能，本文使用均方误差(MAE)和均方根误差(RMSE)作为预测误差的度量。RMSE 定义如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

MSE 定义如下：

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

其中，$y_i - \hat{y}_i$ 为第 i 个复合物的实际亲和力 - 预测亲和力。对于预测亲和力与实验测量亲和力之间的相关性，用皮尔逊相关系数(R)和标准差(SD)来评估其相关性。回归中的 R 定义如下：

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

其中，x_i 表示第 i 个复合物的实际亲和力，y_i 表示第 i 个复合物的预测亲和力。SD 定义如下：

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left[y_i - (ap_i + b)\right]^2}$$

N 表示蛋白质－配体复合物的数量，y_i 和 p_i 表示第 i 个复合物的实际亲和力和预测亲和力，其中 a 和 b 分别表示实际结果和预期结果之间的关系曲线的斜率和截距。一致性指数(CI)是指两个随机选择的蛋白质－配体复合物按照特定顺序的预测亲和力值与真实亲和力值之间的概率。例如：

$$CI = \frac{1}{Z}\sum_{y_i > y_j} h\left(P_i - P_j\right)$$

p_i 为结合亲和力 y_i 的最大预测值，p_j 为结合亲和性 y_i 的最小预测值。蛋白质－配体复合物的总数由标准化常数 Z 表示。当 $u > 0$，$u = 0$，$u < 0$ 时，$h(u)$ 分别为 1.0，0.5，0.0。较大的 CI 表明该模型具有良好的预测能力。

4. 结果与讨论

本文使用多个公共数据库对 GraphPLA 算法进行测试，并与其他算法进行比较分析。研究发现，GraphPLA 可以有效地预测蛋白质－配体结合亲和力。

Table 1. Manifestation of GraphPLA
表 1. GraphPLA 的表现

Datasets	RMSE	MAE	R	SD	CI
Test	1.388	1.118	0.795	1.345	0.796
Validation	1.366	1.106	0.805	1.322	0.799
Training	1.015	0.775	0.841	1.107	0.806

Table 2. Manifestation of GraphPLA and other competing methods on the core 2016 test set
表 2. GraphPLA 和其他方法在 2016 测试集上的表现

Methods	RMSE	MAE	R	SD	CI
GraphPLA GAT-GCN	1.388	1.118	0.795	1.345	0.796
GraphPLA GCN	1.392	1.121	0.795	1.351	0.792
GraphPLA GAT	1.396	1.126	0.796	1.348	0.789
DeepDTA	1.443	1.148	0.749	1.445	0.771
Pafnucy	1.418	1.129	0.775	1.375	0.789
TopologyNet	3.713	3.151	0.173	2.142	0.555

4.1. 与竞争方法的比较

本文拟使用现有的三种深度学习方法 DeepDTA [6]、Pafnucy [2]和 TopologyNet [5]与 GraphPLA 进行比较。表 2 比较了三种不同的 GraphPLA 模型与 PDBbind 数据集上现有模型的性能，表现最好的变体是 GAT-GCN。因此，在之后的统计分析中都集中于 GAT-GCN。

如表 1 所示，在 PDBbind 数据库上对 GraphPLA 的性能进行了评估。在表 2 中，GraphPLA 的表现优于其他同类算法。与 DeepDTA、Pafnucy 和 TopologyNet 相比，GraphPLA 的 RMSE(1.388)更低。GraphPLA

的相关 R 为 0.795，比 DeepDTA 增加 6.1%，比 Pafnucy 增加 2.5%，比 TopologyNet 增加 61.6%，CI 为 0.796。与 DeepDTA、Pafnucy、TopologyNet 算法相比，该算法在 MAE、SD 等性能指标上也有明显提升。如表 3 所示，对于 test 105 集，GraphPLA 在准确性方面优于其他可比较的算法。如表 4 所示，还对 test 71 组进行了测试。结果表明，GraphPLA 可以更准确地预测亲和力。

Table 3. Accuracy of GraphPLA's predictions and those of other competing techniques on the test 105 set
表 3. GraphPLA 和其他方法在 test 105 集上的表现

Methods	RMSE	MAE	R	SD	CI
GraphPLA	1.134	0.881	0.802	1.116	0.804
DeepDTA	1.425	1.134	0.652	1.432	0.738
Pafnucy	1.392	1.169	0.750	1.176	0.782
TopologyNet	4.143	3.841	0.444	1.530	0.646

Table 4. Accuracy of GraphPLA's predictions and those of other competing techniques on the test 71 set
表 4. GraphPLA 和其他方法在 test 71 集上的表现

Methods	RMSE	MAE	R	SD	CI
GraphPLA	1.118	0.867	0.807	1.103	0.807
DeepDTA	1.144	1.144	0.417	1.527	0.641
Pafnucy	1.442	1.210	0.427	1.230	0.628
TopologyNet	4.157	3.913	0.192	1.308	0.559

Table 5. Accuracy of GraphPLA's predictions and those of other competing techniques on the 2016 test set
表 5. GraphPLA 和其他方法在 2016 测试集上的表现

Models	RMSE	MAE	R	SD	CI
Without local features	1.511	1.266	0.736	1.488	0.764
Without physicochemical characters	1.404	1.186	0.762	1.396	0.779
Without dilated convolution	1.403	1.154	0.771	1.374	0.782
Without SSEs	1.402	1.117	0.784	1.365	0.789
GraphPLA	1.388	1.118	0.795	1.345	0.796

4.2. 局部口袋特征的影响

作为局部特征的蛋白质结合口袋是蛋白质－配体结合亲和力预测的重要信息。因此，去掉局部口袋特征提取模块后训练模型。没有局部口袋特征的模型在 2016 测试集上的表现如表 5 所示。结果表明，该模型在 5 项评价指标上均低于原模型。此外，本文还尝试使用不具有全局蛋白质特性的模型，结果表明性能较差。综上所述，局部口袋特征和全局蛋白质特征包含了对蛋白质－配体结合亲和力预测极其重要的信息。

4.3. 不同类型结构特性的影响

为了研究不同类型的结构性质对 GraphPLA 的影响，本文分别通过去除二级结构和物理化学特性进行了实验。从表 5 可以看出，在识别蛋白质－配体结合亲和力方面，结构特性，尤其是理化性质是非常重要的。

4.4. 扩张卷积的影响

与传统卷积相比，扩展卷积可以捕捉长序列蛋白质中氨基酸残基之间的多尺度长程相互作用。如表5所示，通过用常规卷积代替扩张卷积，验证其对模型的意义。总体而言，改进后的卷积算法具有较好的预测能力。

5. 结论

本文创建了一种基于图神经网络的绑定亲和预测技术，称为GraphPLA。使用具有直接结合配体的独特性质的蛋白质结合口袋作为局部输入特征，使用多图神经网络来学习配体的图表示，并使用扩展卷积来捕获蛋白质之间的多尺度远程相互作用。本文还对这些新的特征进行了测试，结果表明，它们可用于亲和力预测。与其他竞争方法相比，我们的模型可以更好地预测结合亲和力。

虽然已经证明了GraphPLA具有更优越的性能，但它也不是没有限制。配体模块使用图神经网络提取特征，与某些复杂的卷积神经网络相比，需要更大的训练数据来展示图神经网络的优势。另一个问题是，图神经网络和卷积神经网络在模型中不能很好地结合，导致模型运行速度不理想。未来我们将尝试编写并完善新的卷积神经网络，使其能与图神经网络更好地结合，达到加快模型运行速度的效果。此外，还需要考虑获取更多的口袋信息，以提高预测结果。未来，我们将致力于收集更高质量的蛋白质配体数据集，探索如何获得更多有用的口袋信息，以及如何更好地融合和构建模型，以提高GraphPLA的预测性能。

基金项目

海南省计算科学与应用重点实验室开放课题(JSKX202102)。

参考文献

[1] 宋益东, 袁乾沐, 杨跃东. 深度学习在蛋白质功能预测中的应用[J]. 合成生物学, 2023, 4(3): 488-506.

[2] Stepniewska-Dziubinska, M.M., Zielenkiewicz, P. and Siedlecki, P. (2018) Development and Evaluation of a Deep Learning Model for Protein-Ligand Binding Affinity Prediction. *Bioinformatics*, **34**, 3666-3674. https://doi.org/10.1093/bioinformatics/bty374

[3] Rezaei, M., Li, Y., Li, X., *et al*. (2019) Improving the Accuracy of Protein-Ligand Binding Affinity Prediction by Deep Learning Models: Benchmark and Model. *ChemRxiv*. https://doi.org/10.26434/chemrxiv.9866912

[4] Ahmed, A., Mam, B. and Sowdhamini, R. (2021) DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinformatics and Biology Insights*, **15**. https://doi.org/10.1177/11779322211030364

[5] Cang, Z. and Wei, G.W. (2017) TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLOS Computational Biology*, **13**, e1005690. https://doi.org/10.1371/journal.pcbi.1005690

[6] Öztürk, H., Özgür, A. and Ozkirimli, E. (2018) DeepDTA: Deep Drug-Target Binding Affinity Prediction. *Bioinformatics*, **34**, i821-i829. https://doi.org/10.1093/bioinformatics/bty593

[7] Liu, L., Zhang, Q., Wei, Y., *et al*. (2023) A Biological Feature and Heterogeneous Network Representation Learning-Based Framework for Drug-Target Interaction Prediction. *Molecules*, **28**, Article 6546. https://doi.org/10.3390/molecules28186546

[8] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.

[9] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755-780.

[10] 焦李成, 杨淑媛, 刘芳, 等. 神经网络七十年: 回顾与展望[J]. 计算机学报, 2016, 39(8): 1697-1716.

[11] Wang, T., Sun, J. and Zhao, Q. (2023) Investigating Cardiotoxicity Related with hERG Channel Blockers Using Molecular Fingerprints and Graph Attention Mechanism. *Computers in Biology and Medicine*, **153**, Article 106464. https://doi.org/10.1016/j.compbiomed.2022.106464

[12] 吴博, 梁循, 张树森, 等. 图神经网络前沿进展与应用[J]. 计算机学报, 2022, 45(1): 35-68.

[13] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.

[14] Veličković, P., Cucurull, G., Casanova, A., *et al.* (2017) Graph Attention Networks. arXiv preprint arXiv:1710.10903

[15] Nguyen, T., Le, H., Quinn, T.P., *et al*. (2021) GraphDTA: Predicting Drug-Target Binding Affinity with Graph Neural Networks. *Bioinformatics*, **37**, 1140-1147. https://doi.org/10.1093/bioinformatics/btaa921