Briefings in Bioinformatics, 22(5), 2021, 1-15

https://doi.org/10.1093/bib/bbab072
解决问题的协议

DeepDTAF：一种预测蛋白质-配体结合亲和力的深度学习方法

王凯丽†、周仁义†、李耀航和李敏

通讯作者:李敏，中南大学计算机科学与工程学院，长沙，410083。电子邮件：limin@mail.csu.edu.cn
王凯丽和周仁义对这项工作贡献相同，为第一作者。

摘要

配体和蛋白质之间的生物分子识别在药物发现和开发中起着重要作用。然而，通过实验来确定蛋白质与配体的结合亲和力是非常耗费时间和资源的。目前，人们提出了许多计算方法来预测结合亲和力，其中大多数方法通常需要蛋白质的三维结构，而这些结构并不常见。因此，我们非常需要能够充分利用序列级特征的新方法来预测蛋白质与配体的结合亲和力并加速药物发现过程。我们开发了一种新的深度学习方法，名为DeepDTAF，用于预测蛋白质-配体结合亲和力。DeepDTAF是通过整合局部和全局的特征来构建的。具体来说，蛋白质结合袋具有直接结合配体的一些特殊性质，首先被用作蛋白质与配体结合亲和力预测的局部输入特征。此外，扩张卷积被用来捕捉多尺度的长程相互作用。我们将DeepDTAF与最近的先进方法进行了比较，并分析了我们模型的不同部分的有效性，准确率的显著提高表明DeepDTAF是一个可靠的亲和力预测工具。资源代码和数据可在https://github.com/KailiWang1/DeepDTAF上找到。

关键词：蛋白质-配体结合亲和力；序列级特征；深度学习；局部和整体特征；蛋白质结合袋

简介

生物分子识别在许多生物过程中起着至关重要的作用，包括非自体蛋白的免疫靶向性，人类催化激酶的特异性[1]等。一般来说，蛋白质经常作为目标，需要与配体相互作用来调节药物发现中的进口生物功能[2，3]。以往的研究也表明，蛋白质与配体的相互作用在药物研发中是至关重要的。介导酶的催化作用[4]、信号转导[5]和其他生物分子功能。而它们的复杂性被破坏与一些疾病有关。结合亲和力可以提供关于蛋白质-配体相互作用强度的重要信息，通常用抑制常数K_i、解离常数K_d或半最大抑制浓度IC_{50}表示。亲和力的成功鉴定在药物发现和再利用的虚拟筛选中起着关键的作用。

王凯丽于2019年在中国华中师范大学物理科学与技术学院获得硕士学位。目前，她正在中国长沙的中南大学攻读计算机科学博士学位。她目前的研究兴趣包括生物信息学、深度学习和药物目标研究。

周仁义于2020年在中国长沙的中南大学获得计算机科学学士学位。他是中南大学计算机科学与工程学院的一名硕士。他目前的研究兴趣包括生物信息学和深度学习。

李耀杭是美国诺福克市老多明尼安大学计算机科学系的副教授。他目前的研究兴趣是计算生物学，蒙特卡洛方法，大数据分析和平行/分布/网格计算。

李敏分别于2001年、2004年和2008年在中国长沙的中南大学获得通信工程学士学位、计算机科学硕士和博士学位。她目前是中南大学计算机科学与工程学院的教授。她的主要研究兴趣包括生物信息学和系统生物学。

提交时间: 2020年11月17日; **接收时间(修订版):**2021年1月27日

的现有药物[6]。更具体地说，发现具有高亲和力的配体结合目标蛋白是早期药物研究的主要焦点[7]。在蛋白质与配体的相互作用中，结合口袋对其相互作用的特异性至关重要。以p38 MAP激酶蛋白为例，它可以形成一个异生结合口袋，通过直接结合化合物来调节蛋白功能。而这种结合可以诱发较大的封闭性变化，抑制激酶的活性，以治疗一些炎症[8]。因此，这些研究表明，局部和整体的结构特性对功能有重要贡献。

研究蛋白质-配体的相互作用机制对药物开发至关重要。然而，通过目前的实验方法从大规模的化学空间中确定结合配体仍然是一个挑战[9]，特别是对于结构未知的蛋白质或蛋白质-配体复合物。由于已知的蛋白质-配体复合物的结构不足[10]，以及实验中的时间或资源消耗，有必要开发一些计算方法来预测蛋白质-配体的结合亲和力。一些基于物理学的方法，如分子对接[11, 12]，分子动力学（MD）模拟[13]，已被广泛用于小分子与蛋白质相互作用的结合亲和力预测和虚拟筛选。这些方法对分子间的相互作用有很好的物理可预测性。然而，这些传统的基于结构的方法仍然存在巨大的计算资源消耗的挑战。一些基于相似性[14]或基于矩阵分解[15]的方法通过使用整个蛋白质或配体的全局相似性矩阵进行预测。这些方法的局限性在于忽略了每个分子中个别成分的详细特征。支持向量机（SVM）和随机森林（RF）算法[16, 17]用于预测蛋白质-配体的相互作用，主要集中在二元分类研究。随着数据的积累和人工智能的发展[18]，深度学习方法也开始在亲和力预测中流行，如Pafnucy[19]、DeepAtom[20]和拓扑神经网络TopologyNet[21]。这些基于结构的方法需要每个分子的原子的详细信息，同时限制了可用的高质量蛋白质-配体复合物结构。为了克服基于结构的限制，一些无结构的方法已经被提出来。MONN[22]是一种多目标模型，它对结合亲和力的预测具有更强的可解释性。但是，它需要花费很长的时间进行数据预处理。DeepDTA[23]和WideDTA[24]是依靠蛋白质序列和配体SMILES作为输入的模型，而忽略了更多的生理化学信息。

在本文中，我们开发了一种新的基于深度学习的方法，名为DeepDTAF，通过整合局部和全局特征来预测蛋白质与配体的结合亲和力。更具体地说，DeepDTAF包括三个独立的模块。
即整个蛋白质模块、局部口袋模块和配体SMILES模块。每个模块的输入是由序列的残基或化合物的SMILES字符串表示。序列中的残基信息不仅包含类型，还包含结构特性，即二级结构元素和物理化学特性。蛋白质模块和口袋模块分别用于提取全局和局部特征。扩张卷积和传统卷积被用来捕捉长程和短程相互作用。三个模块的卷积层和最大集合层的最终特征被串联在一起，并送入分类部分。此外，我们还对我们的模型进行了测试。并与其他竞争性模型进行了比较。这些结果表明，DeepDTAF是一个有用的工具，可以提供可靠的蛋白质-配体结合亲和力预测。

材料和方法

数据集

PDBbind数据库[25]包括一个用-logKi表示的实验验证的蛋白质-配体结合亲和力的集合。
-logKd或-logIC$_{50}$，来自蛋白质数据库[10]。在这里，我们像以前的计算方法一样，重点关注PDBbind数据库2016版中的三个数据集。通用集提供了9226个收集到的蛋白质-配体复合物。精炼集共包含4057个高质量的亲和力数据和复合物。而2016年的核心集通常被用作高质量的基准，其中包括用于评估各种对接方法的多元化结构和结合数据[26, 27]。为了确保这三个数据集之间没有数据重叠，2016年核心集中的290个蛋白质-配体复合物被从精炼集中删除。此外，为了方便模型比较，验证集中的85个蛋白配体复合物和训练集中的2个蛋白配体复合物也被删除，以保持与Pafnucy的一致性。最后，通用集包括9221个复合物，精炼集包括3685个复合物，2016年核心集包括290个复合物。这些数据集提供了蛋白质PDB文件、口袋PDB文件和配体SDF文件等。在此，我们根据PDB文件收集了蛋白质序列、口袋序列。并将SDF文件转换为SMILES字符串。在我们的模型中，考虑到一些蛋白质的三维结构尚不清楚，我们只使用一维序列数据来提供输入信息。此外，由于蛋白质序列、口袋序列和SMILES字符串的长度不同，为了建立一个有效的表示形式，有必要确保固定的长度。蛋白质序列、口袋序列和SMILES字符串的固定长度分别根据图1所示的分布情况选择。蛋白质、配体和口袋序列的最大长度分别为4720、472和125。因此，我们为蛋白质序列定义了固定的1000个字符，为SMILES字符串定义了150个字符，为口袋序列定义了63个字符，以覆盖这些数据集中大约90%的蛋白质、90%的配体和90%的口袋。长于固定字符的序列被截断，短于固定字符的序列被填充为0。

在这里，我们使用了Pafnucy[19]的相同方法来分割训练集和验证集。在精炼集中随机选择的数千个复合体被用作验证集。精炼集和普通集中剩下的11 906个复合物被用作训练集。此外，2016年的核心集被汇编为测试和评估我们的模型。2016年核心测试集中每个蛋白质序列的Smith-Waterman相似度[28]与训练集中99%的蛋白质对的任何序列最多只有60%[23]。此外，为了给一个更客观的评价，我们从蛋白质数据库[10]中收集了另外两个测试集。其中一个测试集包括105个数据（test105），每个蛋白质序列的Smith-Waterman相似度最多为训练集中任何序列的60%。另一个测试集有71个数据（test71），与训练集中的序列的一致性小于35%[29, 30]。而测试集则列在补充表S1中。2016年核心集、test105集、test71集和所有数据的亲和力值的分布情况见补充图S1。

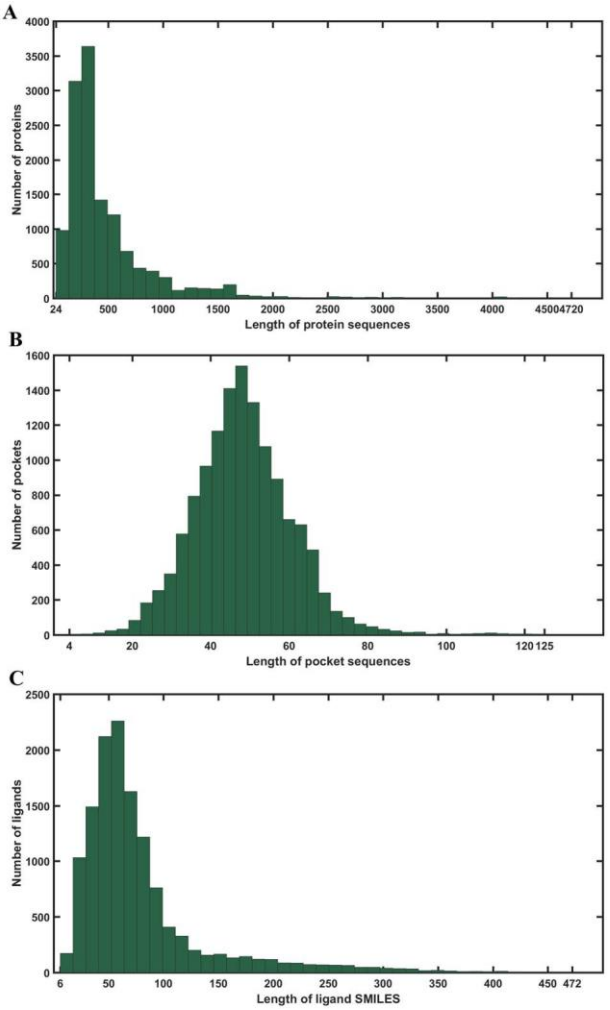

图1.研究中所有数据的长度统计。**(A)** 蛋白质序列的长度分布。最大和最小长度分别为4720和24。**(B)**口袋序列的长度分布。最大和最小的长度分别为472和6。**(C)** 配体SMILES的长度分布。最大和最小的长度分别是125和4。

输入表示

在本研究中，只有一维序列数据被用于标签编码，蛋白质、配体或其复合物的三维结构没有被包括在输入表示中。为了更有效地获得相互作用信息，我们将基于文本的输入信息分为三个部分，配体表示、蛋白质表示和口袋表示。在大多数以前的工作中，蛋白质序列和配体SMILES的输入表示被证明是有效预测蛋白质-配体结合亲和力的[23]。在这里，我们增加了额外的输入信息，可能是结构属性信息和结合口袋信息，这被证明有利于亲和力预测。输入信息的细节如下。

配体代表

流行的配体化学结构的一维表示法是基于原子、键、环等的简化分子输入线系统（SMILES）[31]。在这里，我们使用了Open Babel
[32]将所有的配体SDF文件转换为SMILES字符串。六十四个字符被用来表示配体SMILES字符串。每个字符都用一个特殊的整数来编码（例如，'H'。12, 'N':14, 'C':42, 'O':48, '（': 1，等等）。的例子是
编码SMILES字符串显示为[$C\,C\,C\,C\,(=O)\,C$] =
[42 42 42 42 1 40 48 31 42]

蛋白质代表

序列表示。氨基酸是蛋白质序列的组成部分。大多数蛋白质序列通常由20种不同类型的氨基酸组成。此外，非标准的残基也包括在一些蛋白质中。在此，我们使用21D单热载体来编码蛋白质序列中的21种不同类型的残基。

结构属性表示。考虑到实验解决的蛋白质结构仍然不足，没有模板的结构预测仍然是一个挑战，我们因此使用蛋白质结构特性作为替代特征。蛋白质的结构属性可以很好地提供更丰富的信息[33, 34]。在本研究中，结构属性包括二级结构元素（SSE）[35, 36]和理化特征。这里，我们使用SSPro程序[37]来预测每个序列的二级结构。八类二级结构状态包括*α-螺旋*（H）、隔离β 桥 的残基（B）、扩展链、参与β梯（E）、水基因结合的转折（T）、3_{10} 螺旋（G）、*π螺旋*（I）、弯曲（S）和线圈（C）[38]。我们使用一个8D的单射矢量来编码SSE。此外，根据侧链的结构提供了非极性、极性、酸性、碱性[39]，并根据每个残基的偶极和侧链体积提供了七个基团[40, 41]（见补充表S2），以描述物理化学特征。因此，11D向量被用来编码物理化学特征。合在一起，每个残基的19D向量被用来代表结构特性。

总之，我们为每个残基使用了一个40D的特征向量，通过整合序列和结构属性表示来描述全局的蛋白质特征。

袖珍代表

口袋通常是指在蛋白质内部或表面的一个结合腔，它具有一些特殊的物理化学和几何特性，可以直接结合小化合物。事实上，口袋的氨基酸定义了特定的物理化学特性，这些特性与口袋的形状和位置相结合，决定了蛋白质的功能。此外，蛋白质与配体的相互作用主要取决于配体与蛋白质口袋之间的结合[42]。而口袋是由一个不连续的序列组成的，包括蛋白质的一些关键氨基酸。因此，口袋被视为局部特征提取的整体。局部口袋特征是至关重要的，它首先被用作蛋白质-配体结合亲和力预测的输入信息。在此，通过整合序列表征和上节所述的结构属性表征，对口袋的每个残基使用40D特征向量来编码局部口袋特征。

模型

模型构建。在这项研究中，基于深度卷积神经网络的架构[43]被应用于DeepDTAF，用于预测结合亲和力。预测模型是通过以下程序来处理回归问题的（图2）。输入的特征已被描述在

上一节的内容。在这里，我们用嵌入层在三个模块中用128D密集向量来代表输入。嵌入层通过输入整数编码的输入，将稀疏向量转化为密集向量。因此，这些模块包括（1000，128）、（63, 128, 150, 128）维矩阵，分别用于蛋白质、口袋和配体。更具体地说，对于蛋白质模块，考虑到较长的蛋白质序列的长程相互作用，使用了具有五种不同扩展率的一维扩展卷积[44]。扩张卷积层之后是最大集合层，这与配体模块相同。然而，稀释卷积包括配体模块中的四个不同的稀释率。为了说明两个模块之间扩张卷积的不同，在程序中使用了扩张卷积A和B来区分它们。对于口袋模块，我们使用了三个一维的传统卷积，过滤器的数量不断增加。因此，卷积层由32、64、128个过滤器组成，过滤器的大小为3。 然后，最大池层随之而来。最后，三个模块的最大集合层的特征被串联在一起，并送入分类部分。分类部分由三个全连接（FC）层组成。第一个FC层有128个节点，第二个FC层有64个节点。每层后面都有一个速率为0.5的剔除层。剔除层随机地将一些隐藏单元的激活值设置为零，以防止过度拟合[45]。最后一个功能层之后是输出层。

在我们的架构中，卷积层和FC层都包括PRelu激活函数[46]，它被用来减少训练时间和避免过度拟合。而且PRelu克服了常用的激活函数的缺点。该函数表达式定义如下。

$$f(x) = \begin{matrix} x \ if \ x > 0 \\ ax \ 如果 x<0 \end{matrix} \tag{1}$$

其中a是一个可学习的参数。在这里，PRelu在所有的输入通道中使用一个单一的参数a。此外，为了使损失函数最小化，我们通过使用一个名为AdamW[47]的自适应矩优化器来优化模型中的参数，最大学习率为0.005，权重衰减为0.01，用于更新我们的模型权重。我们使用MSELoss作为损失函数，它创建了一个衡量平均平方误差的标准，以最小化训练期间目标和预测之间的差异。为了优化参数并确定我们的模型，在训练中使用了16和20个epochs的批处理量。最后，在验证集中使用误差最小的模型。

综上所述，我们建立了一个结合局部和全局特征的模型，以提取更丰富的相互作用信息，并使用扩张卷积取代传统的卷积，以扩大感受野并捕捉更多的长距离相互作用。

本地和全球特征

从生物学的角度来看，从局部序列中得出的关键氨基酸及其相互作用一般被认为是重要的。然而，在以前基于深度学习的模型研究中，从整个蛋白质序列中得到的全局特征被用于预测蛋白质与配体的结合亲和力，而局部特征往往被忽略[23]。事实上，在许多生物过程中，蛋白质口袋作为目标发挥着重要作用，直接结合小分子。例如，HIV-1蛋白酶的抑制剂结合口袋（PDB代码：1G2K）[48]被认为是抗病毒治疗的一个有吸引力的目标。在本研究中，1G2K和抑制剂的结合口袋之间的相互作用显示在图3和补充表S3。这里的相互作用是由LIGPLOT程序计算的[49]。此外，除了涉及结合袋的残基，序列中的其他残基也在蛋白质功能和与配体的一些长程相互作用中发挥重要作用。氨基酸的序列规定了一个蛋白质的结构和运动范围，这反过来又决定了它的功能。因此，代表整个蛋白质序列的全局特征在本研究中非常重要。总的来说，局部和全局特性对蛋白质功能和复杂的相互作用至关重要。在这里，通过整合蛋白质结合袋序列和整个蛋白质序列的局部和全局特征，构建了深度学习模型来捕捉不同输入位置的重要性。

扩张卷积

扩张卷积可以通过设置不同的扩张率来捕捉多尺度的上下文信息，并支持感受野的指数级扩张，与传统卷积相比，不会损失分辨率或覆盖范围。扩张卷积算子$*_l$被定义为。

$$(F \circledcirc_l k)(P) = \sum_{s+lt=P} F(s)k(t) \tag{2}$$

其中$F : Z^2 \rightarrow R$是离散函数，$k : \Omega_1 \rightarrow R$是离散的3×3滤波器sl是扩张率，$s$和$t$是元素向量的下标。应用指数级增长的滤波器的离散函数可以定义如下。

$$F_{i+1} = F_i *_{2^i} k_i \ 对于 i = 0, 1, \ldots, n-2. \tag{3}$$

这里，扩张卷积被用来捕获长距离的通过增加有效感受野的大小，蛋白质特征和配体SMILES的相互作用。蛋白质模块有5层，应用3×3卷积核，扩张率分别为1、2、4、8、16。配体模块有4层，应用3×3卷积核，扩张率为1、2、4、8。

评价指标

对于蛋白质-配体结合亲和力的预测，预测值与实验测量的亲和力值进行了比较。为了评估我们模型的性能，我们使用了均方误差（MAE）和均方根误差（RMSE）作为预测误差的衡量标准。对于预测值和实验测量的亲和力值之间的相关性，我们旨在用回归中的均方根误差（R）[50]和标准偏差（SD）[51]来评估它。回归中的SD定义如下。

$$SD = \sqrt{\frac{1}{N-1}} \sum_{i=1}^{N} (y_i - ap_i + b)^2 \tag{4}$$

其中N是蛋白质-配体复合物的数量，y_i和p_i是第i个复合物的实际和预测亲和力，a和b是实际和预测值之间的函数线的斜率和截距。作为另一个典型的指标，一致性指数（CI）[52, 53]是指按特定顺序随机选择的两个蛋白质-配体复合物的预测值和真实亲和力之间的概率。例如，CI被定义为

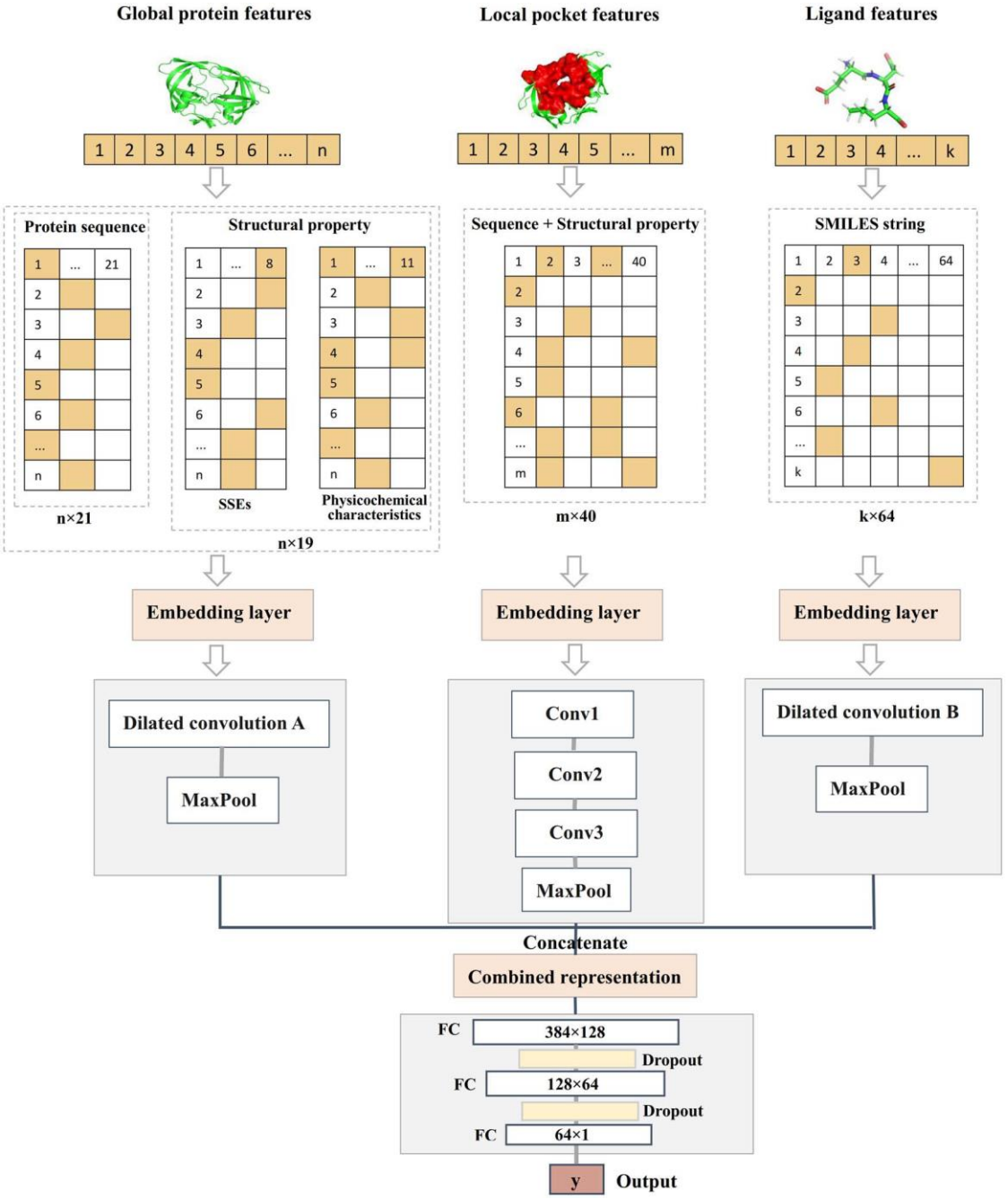

图2. DeepDTAF的结构。DeepDTAF首先将蛋白质、口袋和配体的一维序列转化为序列、结构属性信息或SMILES信息，然后将输入信息送入嵌入层和扩张或传统卷积层。最后，这些特征被串联起来，送入FC层进行结合亲和力预测。

如下：

$$CI = \frac{1}{Z} \sum_{i,j} h(p_i - p_j) \quad (5)$$

$$Z = \sum_{y>y} h(p_i - p_j)$$

其中p_i是较大的结合亲和力的预测值值y_i，p_j是较小亲和力值y_j的预测值。归一化常数Z是蛋白质-配体复合物的总数。而函数$h(u)$在$u>0$、$u=0$和$u<0$时分别等于1.0、0.5和0.0。CI值越大意味着该模型的预测性能越好。

结果和讨论

在这项研究中，我们引入了基于深度学习的模型来预测蛋白质与配体的结合亲和力。DeepDTAF模型纳入了最多三个基于文本的输入信息模型。

模块：全局蛋白质模块、局部口袋模块和配体SMILES模块。事实上，我们的模型是通过结合局部和全局特征构建的。此外，在我们的模型中还使用了扩张卷积来捕捉多尺度信息和更多的长程相互作用。结果表明，DeepDTAF的预测结果比DeepDTAF的预测结果更准确。

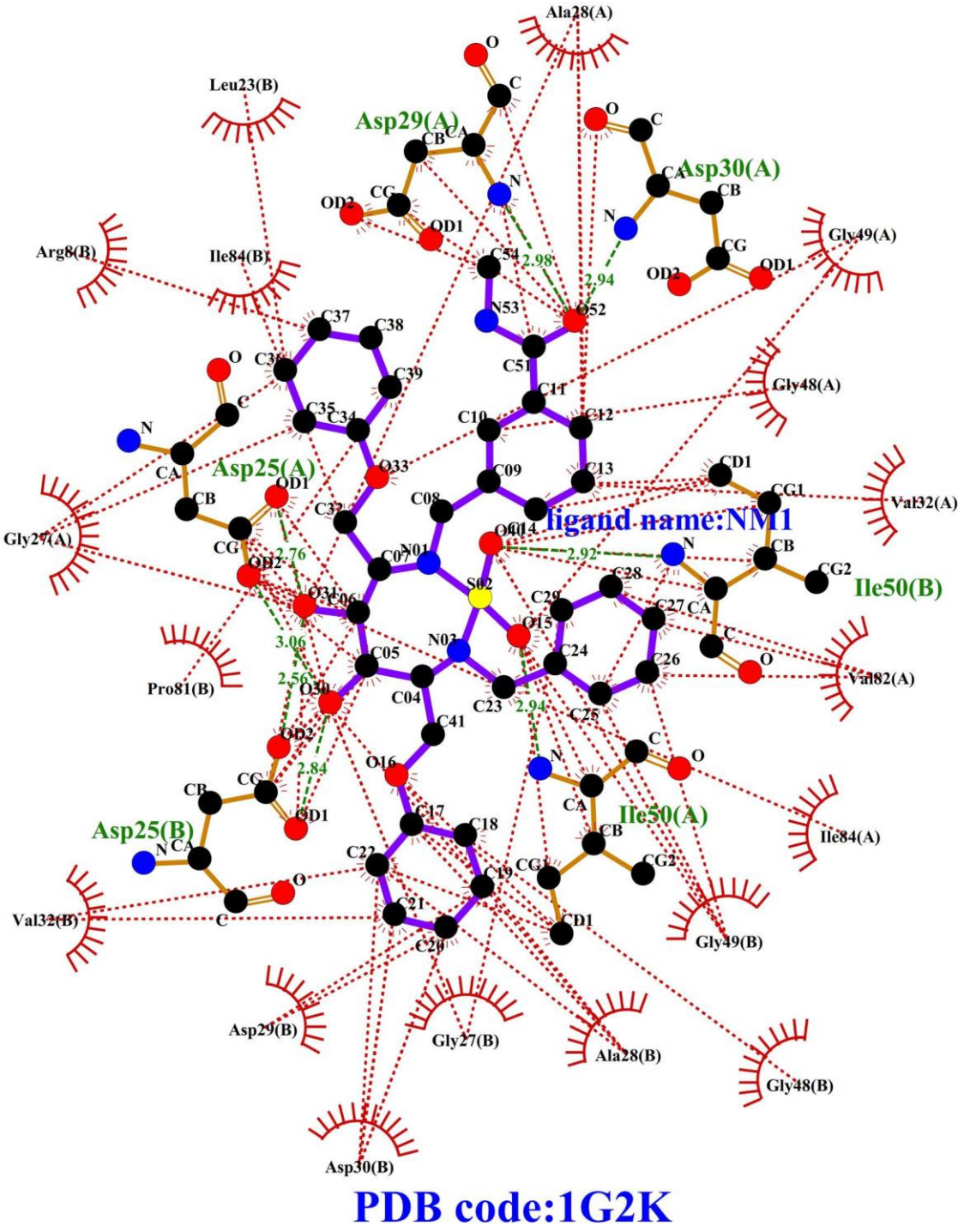

图3.1G2K的二维配体-口袋互动图。配体键和配体名称NM1的颜色为紫色。口袋的疏水残基和口袋残基与配体之间的疏水相互作用用红色表示。残基和配体之间的氢键用绿色虚线表示。氢键残基的颜色为黄色，名称为绿色。

其他竞争性方法。此外，我们还测试了我们的模型结构的不同部分的有效性。局部口袋特征，包含了更直接和详细的结合亲和力测定信息，被证明是该架构中最有用的信息。

与其他竞争方法的比较

在这里，我们进行了20个epochs的训练，选择了验证集中误差小的模型。此外，为了评估DeepDTAF在预测蛋白质-配体结合亲和力方面的性能，我们将DeepDTAF与三种最先进的

表1. DeepDTAF的性能

数据集	RMSE	MAE	R	ĀĀĀ	识别
测试	1.355	1.073	0.789	1.337	0.799
审定	1.367	1.054	0.747	1.365	0.779
培训	0.952	0.731	0.864	0.942	0.839

表2. DeepDTAF和其他竞争方法对2016年核心测试集的预测准确率

方法	RMSE	MAE	R	ĀĀĀ	识别和识别(CI)
DeepDTA	1.443	1.148	0.749	1.445	0.771
パーフナーション	1.418	1.129	0.775	1.375	0.789
深DTAF	1.355	1.073	0.789	1.337	0.799
拓扑网	3.713	3.151	0.173	2.142	0.555

深度学习模型，DeepDTA[23]，Pafnucy[19]和Topolog- yNet[21]。DeepDTA是一个传统的一维卷积神经网络预测模型，包括蛋白质的主要序列（一维代表）和配体的SMILES字符串，没有添加额外的输入信息。Pafnucy是一个深度神经网络模型，利用三维卷积来产生蛋白质-配体三维结构的特征图。我们重新计算了DeepDTA和Pafnucy模型对训练集的11 906个蛋白质-配体数据、验证集的1000个数据、2016年核心测试集的290个数据和测试105集的105个数据所预测的亲和力。下载的DeepDTA程序和Pafnucy程序被用来分析蛋白质-配体数据。TopologyNet是一个基于结构的模型，通过整合元素特定持久同源性（ESPH）方法和深度卷积神经网络。ESPH是通过一维拓扑不变量表示三维复杂几何的方法。我们通过使用TopologyNet计算了2016年核心测试集和测试105集的预测亲和力。TopologyNet是用于蛋白质-配体亲和力预测的在线软件（https://weilab.math.msu.e du/TDL/TDL-BP/index.php）。正如预期的那样，DeepDTAF可以在两个测试集中提供良好的性能，这两个测试集在模型训练和验证期间是未知的。如表1和图4所示，DeepDTAF在PDBbind数据库中的表现得到了评估。事实上，我们发现DeepDTAF在2016年的核心测试集上可以提供比其他竞争方法更好的性能（表2和图5）。DeepDTAF的RMSE（1.355）作为预测误差的指标之一，低于DeepDTA（RMSE=1.443）、Pafnucy（RMSE=1.418）、TopolopyNet（RMSE=3.713）。DeepDTAF中0.789的相关R提高了4.0%（DeepDTA）、1.4%（Pafnucy）和61.6%（TopologyNet），0.799的CI提高了2.8%（DeepDTA）、1.0%（Pafnucy）和24.4%（TopologyNet）。DeepDTAF的其他指标，如MAE和SD，也优于DeepDTA、Pafnucy和TopologyNet。此外，DeepDTAF与其他竞争性方法相比，在测试105集上更准确（表3和图6）。尽管有60%的序列相似性，但在test71数据集中，35%的序列相似性也被分析出来。表4和图7显示了RMSE、MAE、R、SD、CI的值。结果表明，我们的模型在亲和力预测方面表现更好。

当地口袋特征的影响

结合口袋拥有一些特殊的属性，用于直接结合配体以确定功能。更具体地说，一些结合点位于蛋白质的凹面。在生物分子识别过程中

小分子会与结合点结合，形成特殊的构象以发挥其功能。例如，乙酰胆碱酯酶（PDB代码：1H22）结合配体抑制剂经常被应用于阿尔茨海默病（AD）的治疗[54]。乙酰胆碱酯酶与抑制剂的晶体复合结构如图8所示。PyMOL被用来分析氢键，并可视化蛋白质-配体结构和假定的空腔。而在我们的测试集中，预测的蛋白质1H22的亲和力为9.18，这与实验测量的亲和力值9.10非常接近。简而言之，蛋白质结合袋是蛋白质-配体相互作用的关键，通常被用作疾病治疗的目标。

在这里，作为局部特征的蛋白质结合袋被认为是蛋白质-配体结合亲和力预测的至关重要的信息。因此，我们测试了局部口袋特征的效果。首先，我们在原始数据集的基础上训练了我们的模型，但去掉了局部口袋特征提取模块。我们的模型在2016年核心测试集上没有局部口袋特征的表现如表5所示，所有的评估指标都明显比原始DeepDTAF模型差。例如，与原始模型相比，它实现了高达5.7%的R下降。此外，没有全局蛋白特征的模型也被测试，结果是性能更差。综合来看，这些结果表明，我们可以通过结合局部口袋特征和全局蛋白质特征来获得更好的性能。而评估指标的大幅下降表明，局部口袋特征包括了对蛋白质-配体结合亲和力预测的极其重要的信息。

不同类型的结构特性的影响

在这项研究中，除了原始的蛋白质序列信息，蛋白质和口袋的结构属性信息也被用于模型中。结构特性包括SSEs和理化特性。蛋白质的SSEs和氨基酸的物理化学特性对功能的描述非常重要。为了研究不同类型的结构特性在DeepDTAF中的影响，我们通过去除SSEs和理化特性，分别进行了消减研究。如表5所示，结构特性，特别是物理化学特性对识别蛋白质-配体结合亲和力起着至关重要的作用。此外，为了验证固定输入长度的有效性，我们还将我们的模型中90%的长度截止点与80%的截止点、85%的截止点、95%的截止点和100%的截止点进行了比较。

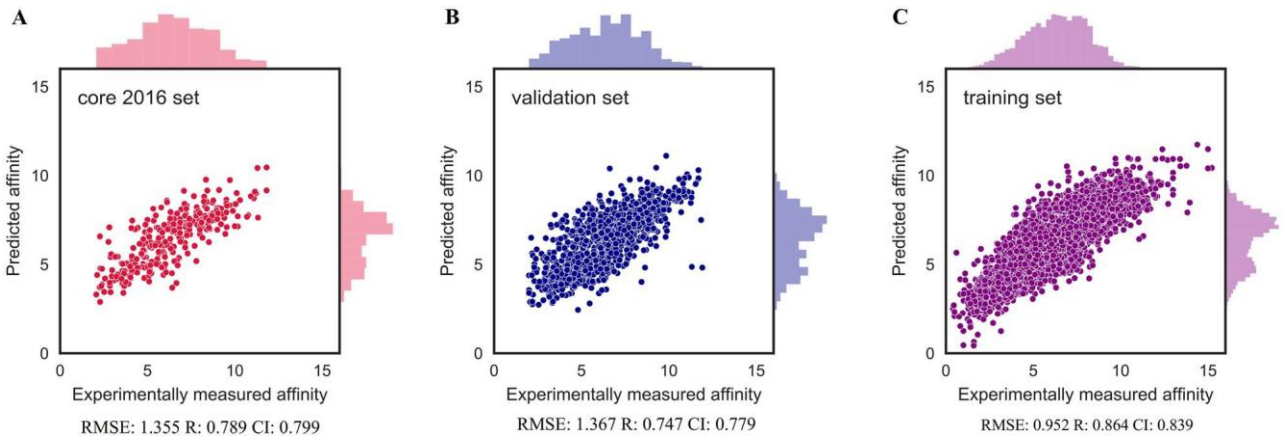

图4. DeepDTAF在2016年核心测试集（**A**）、验证集（**B**）和训练集（**C**）的预测亲和力分布。

表3. DeepDTAF和其他竞争性方法在测试105集上的预测准确率

方法	RMSE	MAE	R	ÃÃÃ	识别和识别(CI)
DeepDTA	1.425	1.134	0.652	1.432	0.738
帕夫努西	1.392	1.169	0.750	1.176	0.782
深DTAF	1.247	0.966	0.766	1.149	0.801
拓扑网	4.143	3.841	0.444	1.530	0.646

表4. DeepDTAF和其他竞争性方法在测试71集上的预测准确率

方法	RMSE	MAE	R	ÃÃÃ	识别和识别(CI)
DeepDTA	1.517	1.144	0.417	1.527	0.641
パーフナーショ ン	1.442	1.210	0.427	1.230	0.628
深DTAF	1.273	0.998	0.480	1.194	0.656
拓扑网	4.157	3.913	0.192	1.308	0.559

表5. DeepDTAF和DeepDTAF在测试集上无局部特征、理化特征、SSEs、扩张卷积的预测准确率

模型	RMSE	MAE	R	ÃÃÃ	识别
没有本地特征	1.518	1.268	0.732	1.482	0.767
没有物理化学特征	1.404	1.118	0.767	1.396	0.783
没有扩大卷积的情况下	1.403	1.134	0.775	1.376	0.788
不含SSE	1.338	1.112	0.781	1.360	0.790
深DTAF	1.355	1.073	0.789	1.337	0.799

我们的结果显示90%的截止值是最好的（补充图S2）。

预测的SSE和实际的SSE之间的比较

我们研究了预测二级结构的准确性和它对结果的影响。如原文章[37]所述，SSpro的准确率上升到92.9%。众所周知，二级结构的预测是相对成熟的。所以，每个序列的预测二级结构被作为我们模型的输入信息。此外，我们还使用DSSP程序[38]来生成真实的二级结构。我们应用真实的二级结构来替代我们模型中预测的二级结构，进行亲和力预测。图8显示了我们的模型和带有真实SSEs的模型在2016年核心测试集上的结果。带有真实SSEs的DeepDTAF的RMSE为1.340（DeepDTAF：1.355），CI为

有真实SSE的DeepDTAF是0.797（DeepDTAF：0.799），有真实SSE的Deep- DTAF的R是0.792（DeepDTAF：0.789）。此外，有真实SSE的DeepDTAF的MAE是1.059（DeepDTAF：1.073），有真实SSE的SD是1.329（DeepDTAF：1.337）。从结果中，我们可以得到，预测SSEs的模型和真实SSEs的模型的准确性是相似的。因此，在我们的模型中使用预测的SSEs作为输入特征是合理的（图9）。

扩张卷积的影响

扩张卷积被用来增加有效的接受文件大小和捕捉多尺度的上下文信息。与传统卷积相比，扩张卷积的优势在于它可以捕捉到氨基酸残基之间的多尺度长距离相互作用，从而获得长的上下文信息。

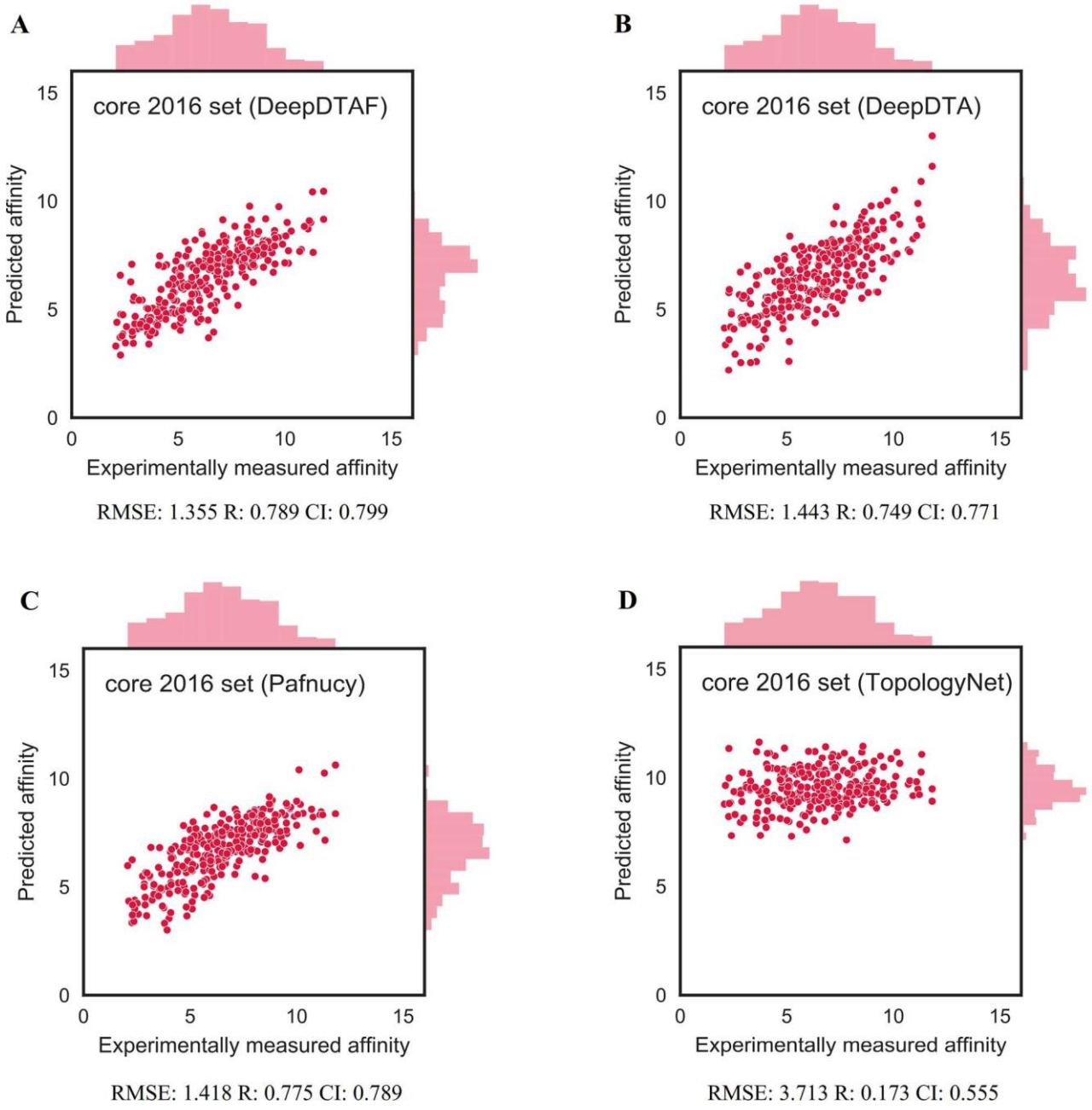

图5. DeepDTAF（**A**）、DeepDTA（**B**）、Pafnucy（**C**）和TopologyNet（**D**）在2016年核心测试集上预测结合亲和力的性能。

序。在本文中，我们在蛋白质和配体模块中使用了扩张卷积，以便更准确地预测。此外，为了证明扩张卷积的重要性，我们用传统卷积替换了扩张卷积的模型（表5）。结果表明，扩张卷积可以为预测提供更好的性能。

带有结合袋的亲和力分析

口袋在蛋白质-配体相互作用中起着至关重要的作用。在这里，除了我们的模型中使用的特征外，口袋体积和口袋氢键也在亲和力方面进行了分析。

预测。我们在2016年的核心测试集中随机选择了30个蛋白质进行口袋体积和氢键接受体的计算（图10）。值得注意的是，亲和力值（图10A）和口袋体积（图10C）之间存在相关关系，相关系数为0.692。而亲和力值（图10A）和口袋中氢键接受体的数量（图10D）之间的相关系数为0.667。结果表明，口袋体积和氢键受体是亲和力预测的有用信息。在未来，将考虑更多的口袋特征来优化我们的模型。同样，我们发现预测的亲和力值（图10B）和口袋体积（图10C）之间的相关系数为0.596。

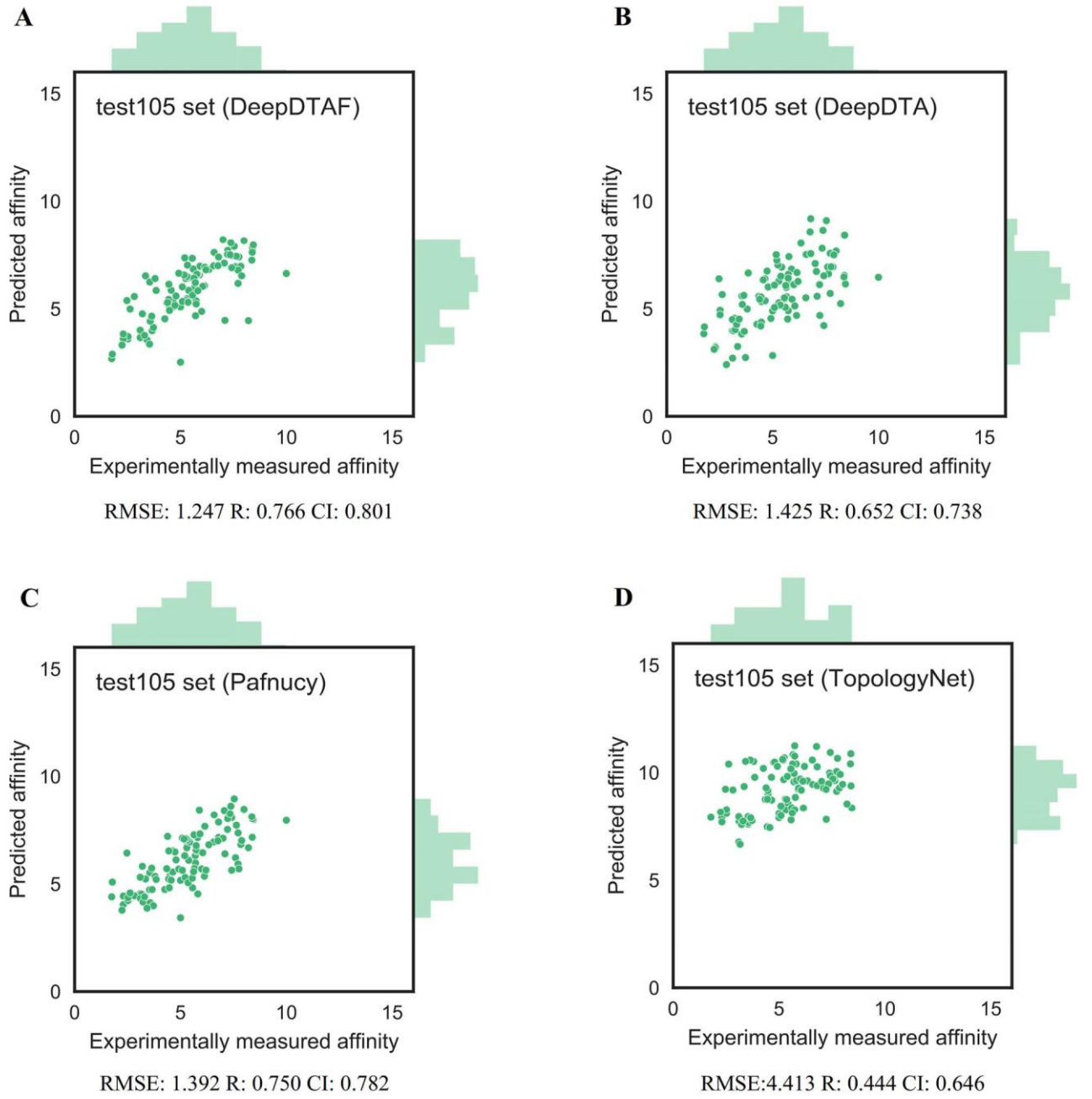

图6. DeepDTAF（**A**）、DeepDTA（**B**）、Pafnucy（**C**）和TopologyNet（**D**）在预测结合亲和力的测试105集上的表现。

而预测的亲和力值（图10B）和氢键接受体的数量（图10D）之间的相关系数为0.524。结果表明，虽然整个口袋的特征，如口袋体积和氢键接受体的数量，没有明确纳入我们的模型，但一些相关的信息仍然可以被DeepDTAF捕获。

总结

对于蛋白质-配体结合亲和力的预测，目前DeepDTA算法只使用蛋白质和配体的序列没有其他物理化学特征。Pafnucy和TopologyNet算法是基于蛋白质-配体复杂的三维结构。然而，这种方法仅限于已知的复杂结构。在这项研究中，我们开发了基于深度学习的DeepDTAF方法来预测结合亲和力。DeepDTAF在以下几个方面区别于其他竞争性算法。首先，我们整合了蛋白质的局部和全局特征以提取不同尺度的信息。第二，除了蛋白质序列特征，我们还增加了蛋白质的额外结构属性，即SSEs和物理化学特征，这些都具有更多的生物学意义。第三，扩展卷积被构建在整个

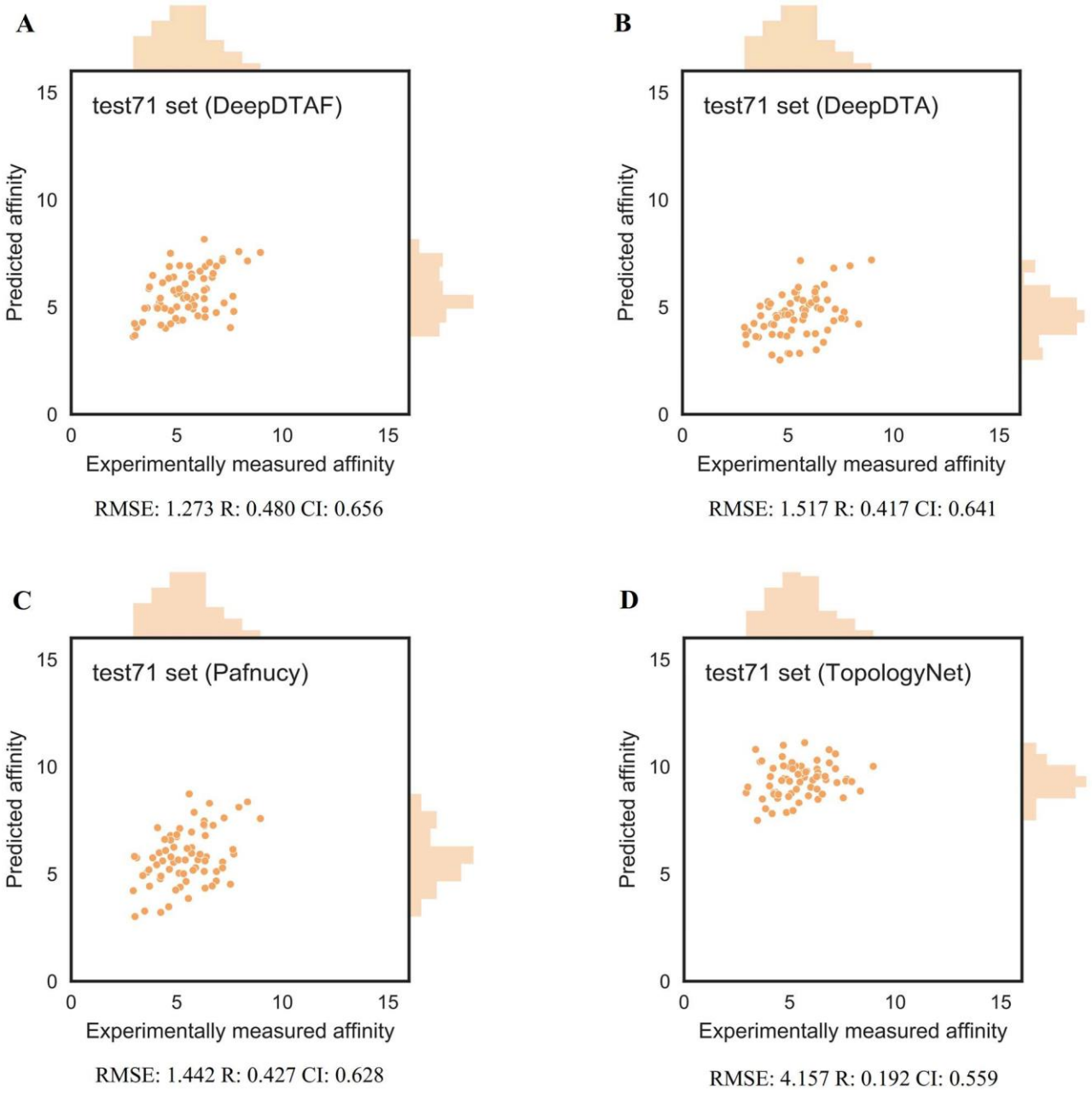

图7. DeepDTAF（**A**）、DeepDTA（**B**）、Pafnucy（**C**）和TopologyNet（**D**）在预测结合亲和力的测试71集上的表现。

蛋白质和配体模块来捕捉多尺度的长程相互作用。我们还测试了这些新特征的效果，结果表明它们对亲和力预测很有用。与其他竞争性方法相比，我们的模型在结合亲和力预测方面有更好的表现。

在蛋白质配体识别过程中，短程结合袋和长程异生效应可以为功能表征和蛋白质-配体相互作用提供有用信息。此外，残基的SSEs和理化特征对功能也有至关重要的影响。总的来说，DeepDTAF包括三个方面独立的模块，即整个蛋白质模块、局部口袋模块和配体SMILES模块。残基类型、SSEs和一些物理化学特征可以从1D序列中获得。因此，我们将它们作为蛋白质和口袋模块的输入信息。然后，扩张卷积被用来从蛋白质和配体模块中提取长程相互作用。而传统卷积法则用于从口袋模块中获取短程相互作用。最后，这三个模块一起被送入FC层以预测结合亲和力。尽管DeepDTAF被证明与其他竞争性方法相比有更好的结果，但它也有一些局限性。目前的架构依赖于以下类型的

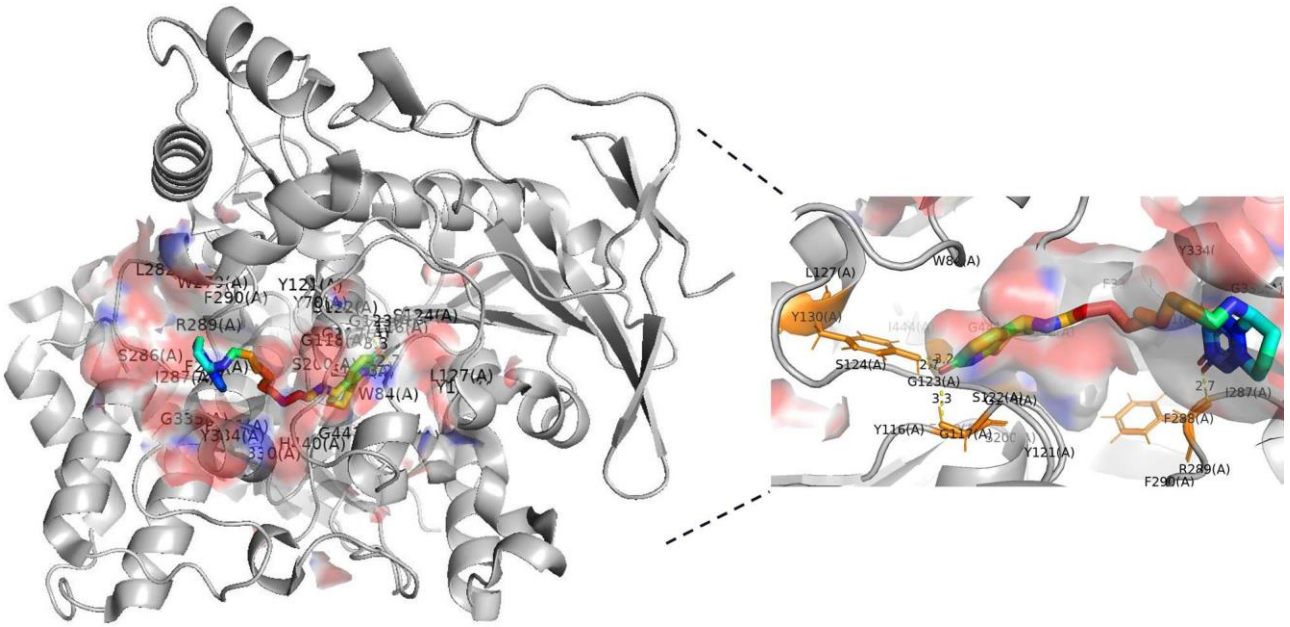

图8.乙酰胆碱酯酶（PDB代码：1H22）结合配体抑制剂E10的卡通图。在放大的图中显示了口袋和配体之间的氢键。配体用棍子表示，颜色为彩虹色，口袋用表面表示，颜色为混合色，而蛋白质的颜色为灰色。黄色的虚线表示配体和结合袋之间的氢键。与配体相互作用的口袋中的残基用橙色表示。

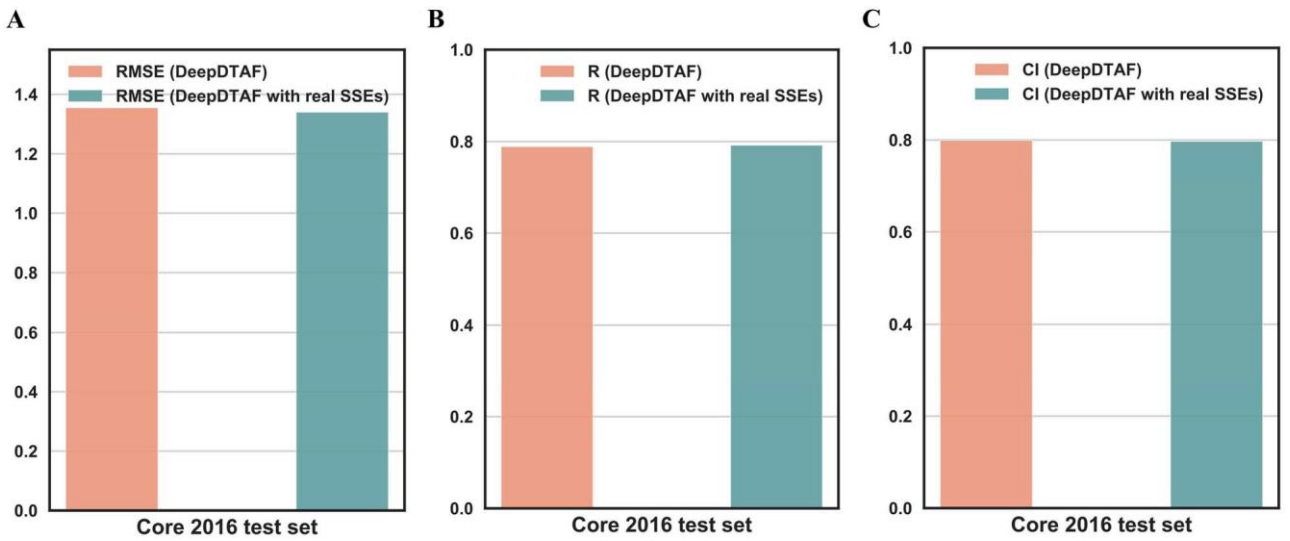

图9. DeepDTAF和DeepDTAF在2016年核心测试集上的RMSE（**A**）、CI（**B**）、R（**C**）与真实SSE的值。

训练数据，所以提供更大的训练数据可能更好。另一个原因是，我们的方法包括配体模块的单一信息。在未来，我们将优化配体模块以捕捉更多的重要特征。此外，口袋的形状和位置也应该被考虑，以提高预测效果。

在这项研究中，我们构建了一个新的深度学习架构--DeepDTAF，它通过结合局部和全局特征来捕捉短程和长程的相互作用。一些相关的结果表明，DeepDTAF是预测蛋白质-配体结合亲和力的可靠工具。

关键点

- 通过对局部和全局特征的整合，开发了一个新的基于深度学习的架构，并应用于蛋白质-配体结合亲和力的预测。
- 蛋白质结合袋首先被用作本地的模型中的输入特征，以预测蛋白质与配体的结合亲和力。
- DeepDTAF是一种有效的方法，它结合了扩张卷积与传统卷积相比，可以捕捉到多尺度的相互作用，用于蛋白质-配体结合亲和力预测。

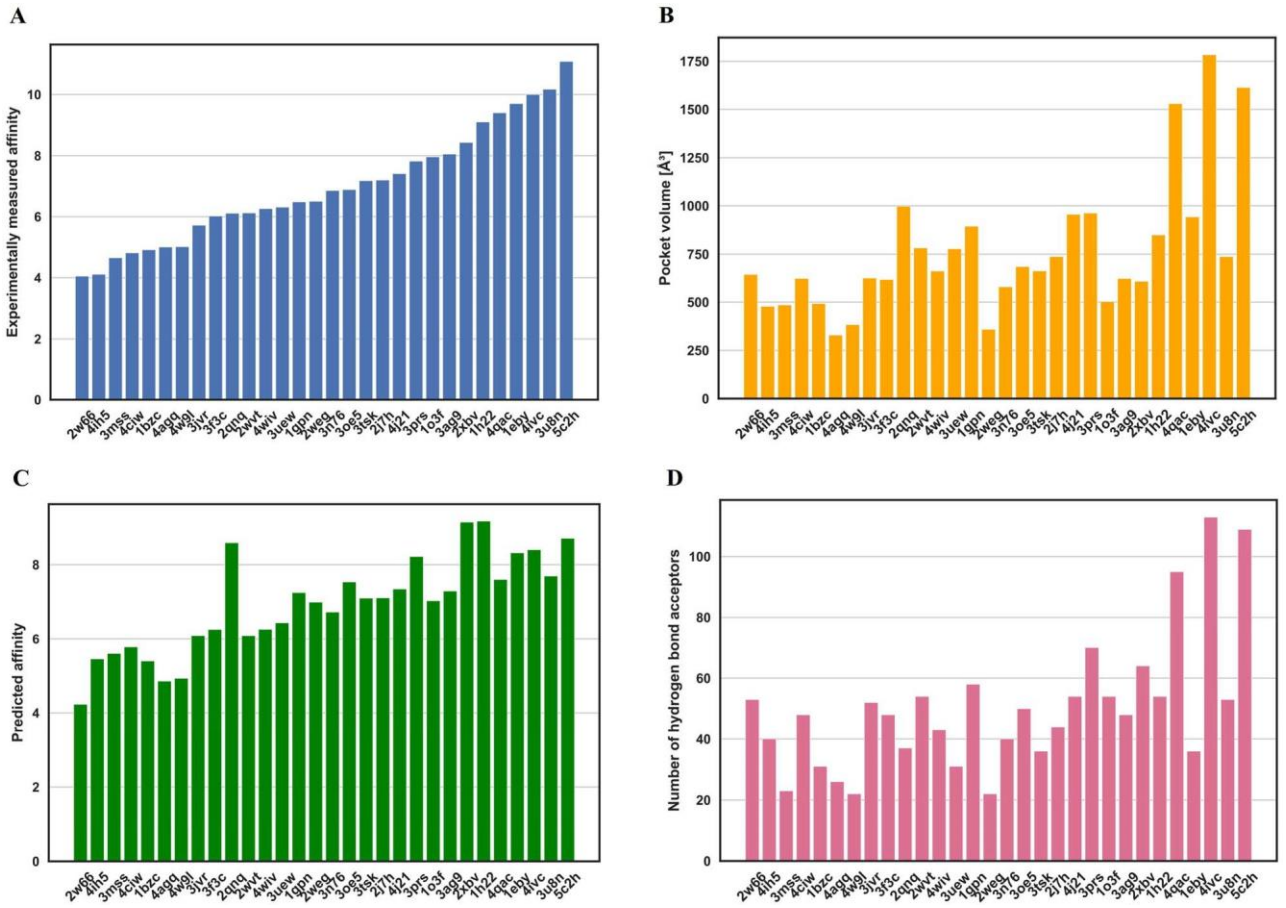

图10.2016年核心测试集中30个蛋白质的亲和力值（**A**和**B**）与口袋体积（**C**）和氢键接受者数量（**D**）之间的关系。预测的亲和力是由DeepDTAF生成的。

补充数据

补充数据可在*Briefings in Bioin- formatics*网上查阅。

资助

国家自然科学基金（资助号：61832019）；湖南省科技计划（2019CB1007）；湖南省学位与研究生教育改革项目（编号：2019JGYB051）；中南大学中央高校基本科研业务费（2282019SYLB004）。

参考文献

1. Gaestel M, Kotlyarov A, Kracht M. 瞄准炎症中先天免疫蛋白激酶信号。 *Nat Rev Drug Discov* 2009; **8**:480-99.
2. Pai MY, Lomenick B, Hwang H, *et al.* Drug affinity responsive target stability (DARTS) for small-molecule target identification.*Methods Mol Biol* 2015;**1263**:287-98.
3. Mutowo P, Bento AP, Dedman N, *et al.* A drug target slim: using gene ontology and gene ontology annotations to 在ChEMBL中导航蛋白质-配体目标空间。*J Biomed Semantics* 2016;**7**:59.
4. Wang W, Donini O, Reyes CM, *et al.* Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions.*Annu Rev Biophys Biomol Struct* 2001; **30**:211-43.
5. Nussinov R, Ma B. 双向信号转导中的蛋白质动力学和构象选择。 *BMC Biol* 2012; **10**:2.
6. Mofidifar S, Sohraby F, Bagheri M, *et al.* Repurposing existing drugs for new AMPK activators as a strategy to extend lifes- pan: a computer-aided drug discovery study.*Biogerontology* 2018; **19**:133-43.
7. Gilson MK, Zhou H-X.蛋白质-配体结合亲和力的计算。 *Annu Rev Biophys Biomol Struct* 2007; **36**:21-42.
8. Pargellis C, Tong L, Churchill L, *et al.* 利用一个新的异生结合点抑制p38 MAP激酶。 *Nat Struct Biol* 2002; **9**:268-72.
9. Inglese J, Auld DS.高通量筛选（HTS）技术：化学生物学中的应用。 *Wiley Encyclopedia of Chemical Biol* 2008;**1**:1-15.
10. Burley SK, Berman HM, Bhikadiya C, *et al.* RCSB蛋白质数据库：生物大分子结构使基础生物学的研究和教育成为可能。

生物医学、生物技术和能源。*核酸研究* 2019;**47**:D464-74.
11. Forli S, Huey R, Pique ME, *et al.* Computational protein-ligand docking and virtual drug screening with the AutoDock suite.*Nat Protoc* 2016; **11**:905-19.
12. Yan Y, Zhang D, Zhou P, *et al.* HDOCK：一个基于混合策略的蛋白质-蛋白质和蛋白质-DNA/RNA对接的网络服务器。*Nucleic Acids Res* 2017;**45**: W365-73.
13. Karplus M, McCammon JA.生物大分子的分子动力学模拟。*Nat Struct Biol* 2002; **9**: 646-52.
14. Cichonska A, Ravikumar B, Parri E, *et al.* Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors.*PLoS Comput Biol* 2017;**13**:e1005678.
15. Cobanoglu MC, Liu C, Hu F, *et al.* Predicting drug-target interactions using probabilistic matrix factorization.*J Chem Inf Model* 2013;**53**:3399-409.
16. Cao DS, Liu S, Xu QS, *et al.* Large-scale prediction of drug-target interactions using protein sequences and drug topological structures.*Anal Chim Acta* 2012; **752**:1-10.
17. Cao DS, Zhang LX, Tan GS, *et al.* Computational prediction of DrugTarget interactions using chemical, biological, and network features.*Mol Inform* 2014; **33**:669-81.
18. Meng X, Xiang J, Zheng R, *et al.* DPCMNE：通过多级网络嵌入从蛋白质-蛋白质相互作用网络中检测蛋白质复合体。*IEEE/ACM Trans Comput Biol Bioinform* 2021. doi: 10.1109/TCBB.2021.3050102.
19. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. 开发和评估用于蛋白质-配体结合亲和力预测的深度学习模型。*Bioinformatics* 2018;**34**:3666–74.
20. Rezaei M, Li Y, Li X, *et al.* Improveving the accuracy of protein- ligand binding affinity prediction by deep learning models: benchmark and model.*ChemRxiv* 2019. doi: 10.26434/chem- rxiv.9866912.v9866911.
21. Cang Z, Wei GW.Topology net: 基于拓扑结构的深度卷积和多任务神经网络用于生物分子特性预测。*PLoS Comput Biol* 2017;**13**:e1005690.
22. Li S, Wan F, Shu H, *et al.* MONN: 一个用于预测化合物-蛋白质相互作用和亲和力的多目标神经网络。*Cell Systems* 2020; **10**:308, e311-22.
23. Öztürk H, Özgür A, Ozkirimli E, *et al.* Deep drug-target binding affinity prediction.*Bioinformatics* 2018; **34**:i821-9.
24. Öztürk H, Ozkirimli E, Özgür A. Wide DTA: prediction of drug-target binding affinity.2019arXiv preprint arXiv:1902.04166.
25. Liu Z, Su M, Han L, *et al.* Forging the basis for developing protein-ligand interaction scoring functions.*Acc Chem Res* 2017; **50**:302-9.
26. Fine J, Konc J, Samudrala R, *et al.* CANDOCK: chemical atomic network-based hierarchical flexible docking algorithm using generalized statistical potentials.*J Chem Inf Model* 2020.
27. Yang J, Baek M, Seok C. Galaxy dock 3: 考虑了全部配体构象灵活性的蛋白质-配体对接。*J Comput Chem* 2019; **40**:2739-48.
28. Zhao M, Lee WP, Garrison EP, *et al.* SSW库：一个用于基因组应用的SIMD smith-waterman C/C++库。*PLoS ONE* 2013; **8**:e82138.
29. Johnson MS, Overington JPA.序列比较的结构基础，评分方法的评估。*J Mol Biol* 1993; **233**:716-88.
30. Ding CH, Dubchak I. 使用支持向量机和神经网络的多类蛋白质折叠识别。*Bioin- formatics* 2001; **17**:349-58.
31. Weininger D. SMILES，一种化学语言和信息系统。1.方法和编码规则介绍。*J Chem Inf Comput Sci* 1988;**28**:31-6.
32. O'Boyle NM, Banck M, James CA, *et al.* Open babel: an open chemical toolbox.*J Chem* 2011;**3**:33.
33. Wang S, Li W, Liu S, *et al.* Raptor X-property: a web server for protein structure property prediction.*Nucleic Acids Res* 2016; **44**:W430-5.
34. Cheng J, Randall AZ, Sweredoski MJ, *et al.* SCRATCH: a protein structure and structural feature prediction server.*Nucleic Acids Res* 2005; **33**:W72-6.
35. Ganapathiraju MK, Klein-Seetharaman J, Balakrishnan N, *et al.* 蛋白质二级结构的特征。*IEEE Signal Process Mag* 2004; **21**:78-87.
36. Zhang F, Shi W, Zhang J, *et al.* PROBselect: 通过动态预测器选择从蛋白质序列中准确预测蛋白质结合残基。*Bioinformatics* 2020;**36**:i735–44.
37. Magnan CN, Baldi P. SSpro/ACCpro 5：使用剖面图、机器学习和结构相似性对蛋白质二级结构和相对溶剂可及性进行几乎完美的预判。*生物信息学》* 2014;**30**:2592-7。
38. Kabsch W, Sander C. 蛋白质二级结构的字典：氢键和几何特征的模式识别。*Biopolymers* 1983; **22**:2577-637.
39. Bhushan R, Ali ITLC.氨基酸在新溶剂中的解析以及碱土金属的影响。*J Liq Chromatogr Relat Technol* 1987; **10**:3647-52.
40. Sun T, Zhou B, Lai L, *et al.* Sequased prediction of protein protein interaction using a deep-learning algorithm.*BMC Bioinformatics* 2017; **18**:277.
41. Shen J, Zhang J, Luo X, *et al.* Predicting protein-protein inter- action based only on sequences information.*Proc Natl Acad Sci U S A* 2007; **104**:4337-41.
42. Wang L, Berne BJ, Friesner RA.配体与蛋白质的结合--具有湿区和干区的结合袋。*国家科学院* 2010; **108**:1326-30.
43. Zeng M, Zhang F, Wu FX, *et al.* Protein-Protein interaction site prediction through combining local and global features with deep neural networks.*Bioinformatics* 2020;**36**:1114–20.
44. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions.2015; arXiv preprint arXiv:1511.07122.
45. Wu H.Gu X. towards dropout training for convolutional neural networks.*Neural Netw* 2015;**71**:1-10.
46. Wei Q, Wang W. 结合L1正则化和PRelu激活函数的深度卷积神经网络的图像检索研究。*IOP Conference Series:Earth and Environmental Science* 2017;**69**:012156.
47. Loshchilov I, Hutter F. Decoupled weight decay regularization.2017; arXiv preprint arXiv:1711.05101.
48. Schaal W, Karlsson A, Ahlsen G, *et al.* 对称和非对称环状磺酰胺HIV-1蛋白酶抑制剂的合成和比较分子场分析（CoMFA）。*J Med Chem* 2001; **44**(2):155-69.
49. Laskowski RA, Swindells MB.LigPlot+：多个配体的用于药物发现的蛋白质相互作用图。*J Chem Inf Model* 2011; **51**:2778-86.
50. Benesty J, Chen J, Huang Y, *et al.* Pearson correlation coefficient.*语音处理中的降噪* 2009;**2**:1-4.
51. Chesher D. 评估检测精度。*Clin Biochem Rev.* 2008;**29**:S23-6.

52. Gönen M, Heller G. 比例危害回归中的一致性概率和判别能力。*Biometrika* 2005; **92**:965-70.
53. Pahikkala T, Airola A, Pietila S, *et al.* Toward more realistic drug-target interaction predictions.*Brief Bioinform* 2015;**16**:325-37.
54. Wong DM, Greenblatt HM, Dvir H, *et al.* 乙酰胆碱酯酶与与Huperzine a有关的二价配体的复合：物种依赖的蛋白质-配体互补的实验证据。*J Am Chem Soc* 2003; **125**: 363-73.