nature machine intelligence

文章

https://doi.org/10.1038/s42256-022-00605-1

具有域自适应功能的可解释双线性注意力网络改善了药物 - 靶点预测

收到时间：2022 年 8 月 1 日

收稿日期：2022 年 12 月 22 日

在线发布时间：2023 年 2 月 2 日

检查更新

Peizhen Bai [1], Filip Miljković [2] ✉, Bino John [3] & Haiping Lu [1] ✉

预测药物 - 靶点相互作用是药物发现的关键。近期的基于深度学习的方法表现出色，但仍有两个挑战存在：如何明确地建模和学习药物与靶点之间的局部相互作用以获得更好的预测和解释，以及如何优化对新型药物 - 靶点对的预测的泛化性能。在此，我们提出了 DrugBAN，这是一个具有域自适应功能的深度双线性注意力网络（BAN）框架，用于明确地学习药物与靶点之间的成对局部相互作用，并对分布外数据做出响应。DrugBAN 基于药物分子图和靶点蛋白质序列进行预测，采用条件域对抗学习在不同分布中对齐学习到的相互作用表示，以便在新的药物 - 靶点对上实现更好的泛化。在三个基准数据集上的实验表明，在域内和跨域设置下，DrugBAN 与五个最先进的基准模型相比实现了最佳的整体性能。此外，对学习到的双线性注意力图的可视化提供了来自预测结果的可解释见解。

药物-靶点相互作用（DTI）预测是药物发现过程中的一个重要步骤[1–3]。传统的体外实验生物医学测量虽然可靠，但成本高昂，开发周期耗时，难以应用于大规模数据4。相比之下，通过计算方法识别高置信度的DTI对可以大大缩小化合物候选物的搜索范围，并深入了解药物组合中潜在副作用的原因。因此，计算方法在过去几年中越来越受到关注，并取得了很大进展5,6。

对于计算化学方法，传统的基于结构的虚拟筛选和基于配体的虚拟筛选方法因其相对有效的性能而得到了广泛研究。然而，基于结构的虚拟筛选需要分子对接模拟，如果目标蛋白质的三维（3D）结构未知，则不适用。此外，基于配体的虚拟筛选基于同一蛋白质的已知活性分子来预测新的活性分子，但当已知活性分子的数量不足时，其性能较差

最近，基于深度学习的方法在计算 DTI 预测方面迅速发展，因为它们在其他领域取得了成功，能够在相对较短的时间内进行大规模验证。其中许多方法是从化学生物学的角度构建的，将化学空间、基因组空间和相互作用信息整合到一个统一的端到端框架中。由于具有可用三维结构的生物靶点数量有限，许多基于深度学习的模型将药物和蛋白质的线性或二维（2D）结构信息作为输入。它们将 DTI 预测视为二元分类任务，并通过将输入输入不同的深度编码和解码模块进行预测，例如

[1] Department of Computer Science, University of Sheffield, Sheffield, UK. [2]Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden. [3]Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, USA. ✉e-mail: h.lu@sheffield.ac.uk

如深度神经网络（DNN）[11,12]、图神经网络（GNN）[9,13–15]或Transformer架构16[17]。随着深度学习技术的进步，此类模型能够从大规模的药物 - 蛋白质相互作用（DTI）数据中自动学习药物和蛋白质的数据驱动表示，而不仅仅使用预先定义的描述符。

尽管这些有前景的发展令人鼓舞，但现有的基于深度学习方法仍面临两个挑战。第一个挑战是药物和蛋白质局部结构之间相互作用的明确学习。药物-蛋白质相互作用本质上是由药物化合物中重要分子亚结构和蛋白质序列中的结合位点之间的相互作用决定的。然而，许多先前的模型使用单独的编码器来学习全局表示，而没有明确学习局部相互作用[9]。因此，药物和蛋白质的表示首先是为整个结构学习的，而互信息只是在黑箱解码模块中隐性地学习。药物和目标之间的相互作用与它们的关键亚结构特别相关；因此，单独的全局表示学习往往会限制建模能力和预测性能。此外，如果没有明确学习局部相互作用，即使预测准确，预测结果也很难解释。

第二个挑战是跨域泛化预测性能，超越已学习的分布。由于化学和基因组空间的广阔区域，在实际应用中需要预测的药物-靶点对通常是未见的，并且与训练数据中的任何对都不同。它们具有不同的分布，因此需要跨域建模[21,22]。一个强大的模型应该能够将学习到的知识转移到只有未标记数据的新域。在这种情况下，我们需要通过学习可转移的表示来对齐分布并提高跨域泛化性能；例如，从“源”到“目标”。据我们所知，这是药物发现中一个未充分探索的方向[23]。

为了应对这些挑战，我们提出了一种基于可解释的双线性注意力网络的模型（DrugBAN）用于药物-靶点相互作用（DTI）预测，如图1a所示。DrugBAN是一个深度学习框架，用于明确学习药物与靶点之间的局部相互作用，并用于跨域学习可转移表示的域适应。具体来说，我们首先使用图卷积网络（GCNs）和卷积神经网络（CNNs）将局部结构编码为二维分子图和一维（1D）蛋白质序列。然后将编码的局部表示输入到由双线性注意力网络组成的成对交互模块中，以学习局部交互表示，如图1b所示。最后，通过全连接层对局部联合交互表示进行解码，以进行DTI预测。通过这种方式，我们可以利用成对的双线性注意力图来可视化每个子结构对最终预测结果的贡献，从而提高可解释性。对于跨域预测，我们应用条件域对抗网络27（CDAN）将源域中学习到的知识转移到目标域，以增强跨域泛化能力，如图1c所示。我们在药物发现的域内和跨域环境中与五种最先进的DTI预测方法进行了全面的性能比较。结果显示，与最先进的方法相比，我们的方法实现了最佳的整体性能，同时为预测结果提供了可解释的见解。

总之，DrugBAN 在三个主要方面与以往的工作有所不同。首先，它通过双线性注意力机制捕捉药物与靶点之间的成对局部相互作用。其次，它通过对抗域适应方法增强跨域泛化能力。它通过双线性注意力权重给出可解释的预测，而非黑箱结果。

结果

问题表述

在药物-蛋白质相互作用预测中，任务是确定一对药物化合物和目标蛋白质是否会发生相互作用。对于目标蛋白质，我们将每个蛋白质序列表示为PPPPPa$_1$，…, a_n），其中每个标记a_i表示 23 种氨基酸之一。对于药物化合物，大多数现有的基于深度学习的方法通过简化的分子输入线性输入规范（SMILES）来表示输入，这是一种描述药物分子中化学原子及化学键标记信息的一维序列。SMILES 格式使得许多经典的深度学习架构能够对药物信息进行编码。然而，鉴于一维序列并非分子的自然表示形式，一些重要的药物结构信息可能会丢失，从而降低模型的预测性能。我们的模型将输入的 SMILES 转换为相应的二维分子图。具体而言，药物分子图被定义为GGPPGG，G)，其中GG是顶点（原子）集合，G是边（化学键）集合。

给定一个蛋白质序列PP和一个药物分子图GG，药物-蛋白质相互作用（DTI）预测旨在学习一个模型⊠，将联合特征表示空间PPPGG映射到一个交互概率分数$p \in [0, 1]$。补充表3提供了本文中常用的符号。

DrugBAN 框架

所提出的 DrugBAN 框架如图 1a 所示。给定输入的药物 - 靶点对，我们首先分别使用独立的 GCN 和 1D 卷积神经网络（1D CNN）模块来对分子图和蛋白质序列信息进行编码。然后，我们使用双线性注意力网络模块来学习编码的药物和蛋白质表示之间的局部交互。双线性注意力网络由双线性注意力步骤和双线性池化步骤组成，以生成联合表示，如图 1b 所示。其次，一个全连接分类层学习一个预测分数，表示交互的概率。为了提高模型在跨域药物 - 靶点对上的泛化性能，我们进一步将 CDAN 嵌入到框架中，以适应表示，从而更好地对齐源和目标分布，如图 1c 所示。

评估策略和指标

我们分别在三个公开数据集上研究分类性能：BindingDB[29]、BioSNAP[30]和Human[16,31]，并留出测试集（“未知”）用于评估。我们在域内和跨域设置中使用了两种不同的分割策略。对于域内评估，每个实验数据集按照 7:1:2 的比例随机分为训练集、验证集和测试集。对于跨域评估，我们提出了一种基于聚类的对分割策略来构建跨域场景。我们在大规模的 BindingDB 和 BioSNAP 数据集上进行跨域评估。对于每个数据集，我们首先分别使用单链算法通过 ECFP4（扩展连接指纹，最多四个键）[32] 指纹和伪氨基酸组成（PSC）[33] 对药物和蛋白质进行聚类。然后，我们从聚类结果中随机选择 60%的药物簇和 60%的蛋白质簇，并将所选药物和蛋白质之间的所有药物 - 靶点对视为源域数据。剩余簇中药物和蛋白质之间的所有对都被视为目标域数据。聚类的实现细节在补充信息第1节中提供。在基于聚类的对偶分裂策略下，源域和目标域是非重叠的，具有不同的分布。遵循域适应的一般设置，我们使用所有标记的源域数据和80%未标记的目标域数据作为训练集，剩余的20%标记的目标域数据作为测试集。跨域评估比域内随机分割更具挑战性，但在现实世界的药物发现中能更好地衡量模型的泛化能力。为了进行更全面的研究，我们报告了在不同蛋白质家族、未见过的药物或靶点以及具有高比例缺失数据的蛋白质家族上的额外实验（分别在补充信息第4-6节）。

受试者工作特征曲线下面积（AUROC）和精确率 - 召回率曲线下面积（AUPRC）分别为

https://doi.org/10.1038/s42256-022-00605-1

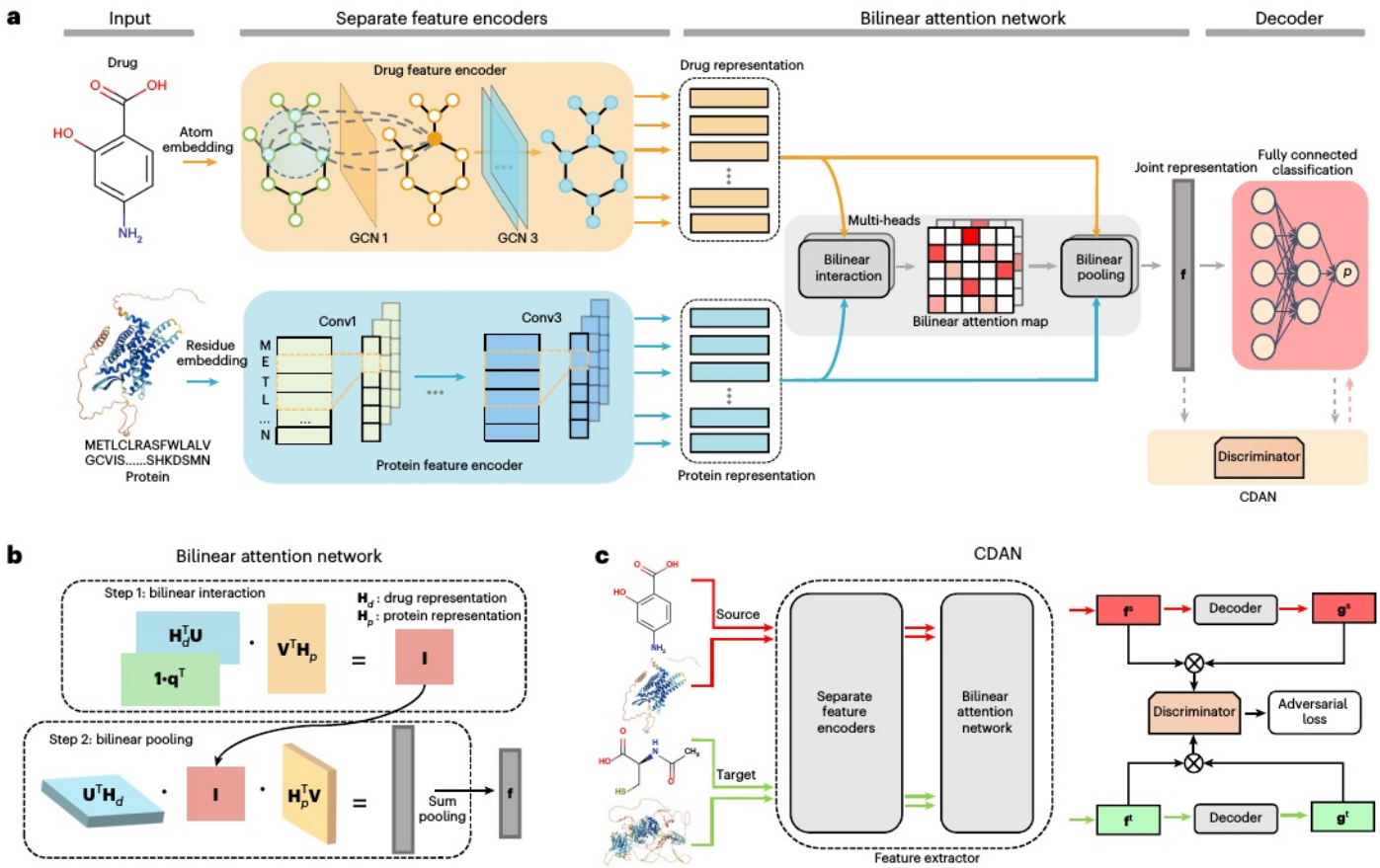

图 1 | DrugBAN 框架概述。**a**，输入的药物分子和蛋白质序列分别通过 GCN 和 1D CNN 进行编码。编码的药物表示的每一行是药物分子中相邻原子的聚合表示，编码的蛋白质表示的每一行是蛋白质序列中的子序列表示。药物和蛋白质的表示被输入到双线性注意力网络中，以学习它们成对的位置交互。联合表示 **f** 由全连接解码器模块进行解码，以预测药物-蛋白质相互作用（DTI）概率 p。如果预测任务是跨域的，则采用 CDAN[27] 模块来对齐源域和目标域中学习到的表示。**b**，双线性注意力网络架构。$\mathbf{H}_d$ 和 $\mathbf{H}_p$ 是经过编码的药物和蛋白质表示。在步骤 1 中，通过变换矩阵 **U** 和 **V** 进行低秩双线性交互建模来获得双线性注意力图矩阵 **I**，以测量子结构级别的交互强度[59]。然后在步骤 2 中，通过共享变换矩阵 **U** 和 V 进行双线性池化来利用 **I** 生成联合表示 **f**。**c**，CDAN 是一种域适应技术，用于减少不同数据分布之间的域转移。我们使用 CDAN 将源域和目标域的联合表示 **f** 和 softmax 对数 **g** 嵌入到由判别器生成的联合条件表示中，判别器是一个两层全连接网络，用于最小化域分类误差以区分目标域和源域。

用作评估模型分类性能的主要指标。此外，我们还报告了在最佳 F1 分数阈值下的准确率、灵敏度和特异度。对于每个数据集分割，我们使用不同的随机种子进行了五次独立运行。表现最佳的模型是在验证集上具有最佳 AUROC 的模型。然后，我们在测试集上对模型进行评估，以报告性能指标。

域内性能比较

在此，我们在随机分割设置下将 DrugBAN 与五个基准进行比较：支持向量机（SVM）[34]、随机森林（RF）[35]、DeepConv-DTI[11]、GraphDTA[13] 和 MolTrans[17]。这是域内场景，因此我们使用不嵌入 CDAN 模块的原始 DrugBAN。表 1 展示了在 BindingDB 和 BioSNAP 数据集上的比较结果。在 AUROC、AUPRC 和准确率方面，DrugBAN 始终优于基准，同时在敏感性和特异性方面的表现也具有竞争力。结果表明，在域内 DTI 预测中，数据驱动的表示学习能够捕获比预先定义的描述符特征更重要的信息。此外，DrugBAN 可以通过其成对交互模块捕获交互模式，进一步提高预测性能。

人类数据集上的域内结果如图2所示。在随机分割下，基于深度学习的模型都实现了相似且令人期待的性能（AUROC>0.98）。然而，正如参考文献16所指出的，人类数据集存在一些隐藏的配体偏差，导致正确预测仅基于药物特征，而不是相互作用模式。高准确性可能是由于偏差和过拟合，而不是模型在现实世界中的前瞻性预测性能。因此，我们进一步采用冷对分割策略来评估模型，以减轻由于数据偏差导致的性能评估的过度乐观。冷对分割策略确保在训练期间未观察到所有测试药物和蛋白质，因此对测试数据的预测不能仅仅依赖于已知药物或蛋白质的特征。我们随机将5%和10%的DTI对分别分配到验证集和测试集中，并从训练集中删除它们相关的药物和蛋白质。图2表明，从随机分割到冷对分割，所有模型的性能都有显著下降，尤其是对于支持向量机（SVM）和随机森林（RF）。然而，我们可以看到 DrugBAN 仍然在与其他最先进的深度学习基准的对比中表现最佳。

跨域性能比较

在随机分割下的域内分类是一个更简单的任务，且在实际应用中重要性较低。因此，接下来，我们研究更具现实性和挑战性的跨域 DTI 预测，其中训练数据

https://doi.org/10.1038/s42256-022-00605-1

表 1 | 在 BindingDB 和 BioSNAP 数据集上的域内性能比较（随机分割）（五次随机运行的统计数据）

数据集	方法	曲线下面积（AUC）	平均精度均值（AUPRC）	精度	灵敏度	特异性
绑定数据库	SVM[34]	0.939 ± 0.001	0.928 ± 0.002	0.825 ± 0.004	0.781 ± 0.014	0.886 ± 0.012
	射频频谱	0.942 ± 0.011	0.921 ± 0.016	0.880 ± 0.012	0.875 ± 0.023	0.892 ± 0.020
	深度卷积 - 弥散张量成像[11]	0.945 ± 0.002	0.925 ± 0.005	0.882 ± 0.007	0.873 ± 0.018	0.894 ± 0.009
	GraphDTA[13]	0.951 ± 0.002	0.934 ± 0.002	0.888 ± 0.005	0.882 ± 0.012	0.897 ± 0.008
	MolTrans[17]	0.952 ± 0.002	0.936 ± 0.001	0.887 ± 0.006	0.877 ± 0.016	0.902 ± 0.009
	禁毒行动	**0.960** ± 0.001	**0.948** ± 0.002	**0.904** ± 0.004	**0.900** ± 0.008	**0.908** ± 0.004
生物SNAP	SVM[34]	0.862 ± 0.007	0.864 ± 0.004	0.777 ± 0.011	0.711 ± 0.042	0.841 ± 0.028
	RF[35]	0.860 ± 0.005	0.886 ± 0.005	0.804 ± 0.005	**0.823** ± 0.032	0.786 ± 0.025
	深度卷积 - 弥散张量成像[11]	0.886 ± 0.006	0.890 ± 0.006	0.805 ± 0.009	0.760 ± 0.029	0.851 ± 0.013
	GraphDTA[13]	0.887 ± 0.008	0.890 ± 0.007	0.800 ± 0.007	0.745 ± 0.032	**0.854** ± 0.025
	MolTrans[17]	0.895 ± 0.004	0.897 ± 0.005	0.825 ± 0.010	0.818 ± 0.031	0.831 ± 0.013
	禁毒行动	**0.903** ± 0.005	**0.902** ± 0.004	**0.834** ± 0.008	0.820 ± 0.021	0.847 ± 0.010

结果以均值±标准差的形式呈现。每个数据集和指标的最佳结果以粗体标记，次优结果以下划线标记。

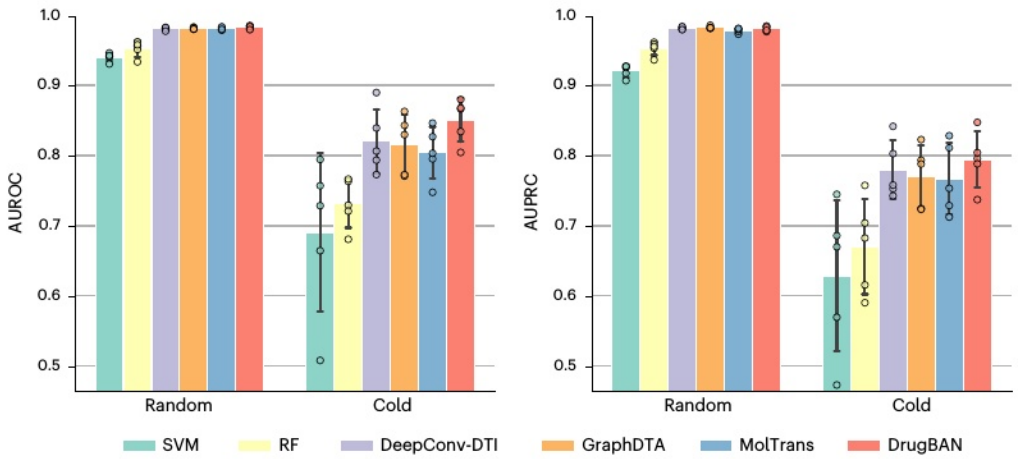

图 2 | 在人类数据集上使用随机分割和冷对分割进行域内性能比较（五次随机运行的统计数据）。垂直条表示均值，黑色线条是误差条，表明标准偏差。这些点表示模型每次随机运行的性能得分。补充表 2 提供了人类数据集的数据统计信息。

并且测试数据具有不同的分布。为了模拟这种情况，原始数据通过基于聚类的成对分割被分为源域和目标域。我们开启 DrugBAN 的 CDAN 模块（即我们使用$DrugBAN_{CDAN}$)来研究跨域预测中的知识可迁移性）。

图3展示了基于聚类的配对分割在BindingDB和BioSNAP数据集上的性能评估。与之前的域内预测结果相比，由于训练和测试数据之间的信息重叠减少，所有DTI模型的性能显著下降。在这种情况下，原始的DrugBAN仍然在整体上优于其他最先进的模型。具体来说，在BioSNAP和BindingDB数据集上，它的 AUROC 分别比 MolTrans 高出 2.9% 和 7.4%。结果表明，DrugBAN在域内和跨域设置下都是一种稳健的方法。有趣的是，在BindingDB数据集上，RF模型表现良好，甚至始终优于其他深度学习基准模型（DeepConv、GraphDTA和MolTrans）。结果表明，在跨域设置下，深度学习方法并不总是优于浅层机器学习方法。

最近，由于其跨域转移知识的能力，域适应技术受到了越来越多的关注，但它们主要应用于计算机视觉和自然语言处理问题。我们将原始的 DrugBAN 与 CDAN 相结合，以解决跨域 DTI 预测问题。如图 3 所示，引入域适应模块后，$DrugBAN_{CDAN}$ 的性能有了显著的提升。在 BioSNAP 数据集上，它在 AUROC 和 AUPRC 方面的表现分别比原始的 DrugBAN 高出 4.6%和 16.9%。通过最小化跨域的分布差异，CDAN 可以有效地增强 DrugBAN 的泛化能力，并提供更可靠的结果。

这些结果证明了 DrugBAN 在跨领域泛化预测性能方面的强大能力。

消融实验

在此，我们对双线性注意力和域适应模块对 DrugBAN 的影响进行了消融研究。结果如表 2 所示。为了验证双线性注意力的有效性，我们研究了 DrugBAN 的三个变体，它们在药物和蛋白质的联合表示计算方面有所不同：单边药物

https://doi.org/10.1038/s42256-022-00605-1

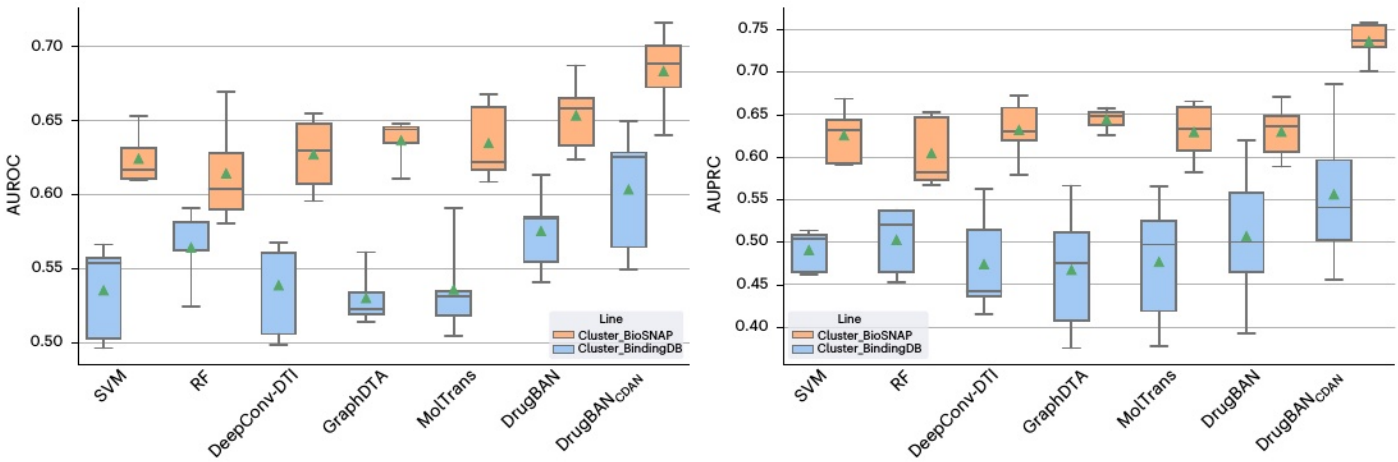

图 3 | 基于聚类的对偶分割在 BindingDB 和 BioSNAP 数据集上的跨域性能比较（五次随机运行的统计数据）。箱线图以中位数为中心线，以均值为绿色三角形。最小值和较低百分位数表示最差的情况和第二差的分数。最大值和上百分位数表明了最佳和第二最佳的分数。补充表 2 提供了 BindingDB 和 BioSNAP 数据集的数据统计信息。

表 2 | 在 BindingDB 和 BioSNAP 数据集上进行的 AUROC 消融研究，采用随机分割和基于聚类的分割策略（五次随机运行的统计数据）

消融测试	BindingDB$_{random}$	BioSNAP$_{random}$	BindingDB$_{cluster}$	BioSNAP$_{cluster}$
线性连接 2、11、13	0.949 ± 0.002	0.887 ± 0.007	不适用	不适用
单侧目标注意力 14	0.950 ± 0.002	0.890 ± 0.005	不适用	不适用
单侧药物注意力 14	0.953 ± 0.002	0.892 ± 0.004	不适用	不适用
禁毒行动	**0.960** ± 0.001	**0.903** ± 0.005	0.575 ± 0.025	0.654 ± 0.023
MolTrans$_{CDAN}$	不适用	不适用	0.575 ± 0.038	0.656 ± 0.028
DrugBAN$_{DANN}$	不适用	不适用	0.592 ± 0.042	0.667 ± 0.030
DrugBAN$_{CDAN}$	不适用	不适用	**0.604** ± 0.039	**0.684** ± 0.026

结果以均值±标准差的形式呈现。前四个模型展示了我们的双线性注意力模块的有效性，而后三个模型展示了DrugBAN$_{CDAN}$在跨域预测中的强大作用。每个数据集的最佳 AUROC 结果以粗体标记，次优结果以下划线标记。NA，不适用于本研究。

注意，单边蛋白质注意力和线性连接。单边注意力等同于参考文献[14]中引入的神经注意力机制，用于捕获药物向量表示和蛋白质子序列矩阵表示之间的联合表示。我们将DrugBAN中的双线性注意力替换为单边注意力，以生成两个变体。线性连接是在最大池化层之后对药物和蛋白质向量表示进行简单的向量连接。如表2的前四行所示，结果表明双线性注意力是捕获相互作用信息以进行药物-蛋白质相互作用预测的最有效方法。为了检验CDAN的效果，我们研究了两个变体：带有域对抗神经网络（DANN）[36]的DrugBAN（即DrugBAN$_{DANN}$）和带有CDAN的MolTrans（即MolTrans$_{CDAN}$）。DANN是另一种不考虑分类分布的对抗域适应技术。表2的后四行表明，DrugBAN$_{CDAN}$在跨域预测中仍然实现了最佳性能提升。

具有双线性注意力可视化的可解释性

DrugBAN 的另一个优势在于能够提供对药物设计工作至关重要的分子层面见解和解读，利用双线性注意力图的组成部分来可视化每个子结构对最终预测结果的贡献。在此，我们研究了来自蛋白质数据库（PDB）[40]的共晶体配体的前三位预测（PDB 编号：6QL2（参考文献[37]）、5W8L（参考文献[38]）和 4N6H（参考文献[39]））。只有分辨率大于 2.5 埃、与人类蛋白质靶点相对应的 X 射线结构才会进入筛选阶段。此外，共晶配体需要具有 pIC$_{50}$≤100 纳摩尔，并且不能是训练集的一部分。可视化结果如图 4a 所示，同时还有来自相应 X 射线结构的配体 - 蛋白质相互作用图。对于每个分子，我们在双线性注意力图中将其权重最高的 20%原子涂成橙色。

对于 PDB 结构 6QL2（与人类碳酸酐酶 2 结合的乙氧唑酰胺），我们的模型正确地解释了磺酰胺区域对于配体 - 蛋白质结合至关重要（磺酰胺的氧原子作为氢键受体与 Leu198 和 Thr199 的主链相互作用，氨基作为氢键供体与 His94 和 Thr199 的侧链相互作用）。相反，乙氧唑酰胺的乙氧基被错误地预测为与蛋白质形成特定的相互作用，尽管其暴露于溶剂中可能会促进进一步的结合（蓝色突出显示）。此外，苯并硫唑骨架与 Leu198 形成芳香 - H 相互作用，但我们的可解释模型仅部分突出显示。值得一提的是，尽管乙氧唑酰胺的前 20%的相互作用原子仅对应三个突出显示的原子，但它们都表明了不同的配体 - 蛋白质相互作用位点，这与 X 射线结构相符。

在结构 5W8L（9YA 配体与人类左旋乳酸脱氢酶 A 结合）中，可解释性特征再次突出了配体 - 蛋白质结合的重要相互作用模式。例如，磺酰胺基团再次被表明会形成特定的

https://doi.org/10.1038/s42256-022-00605-1

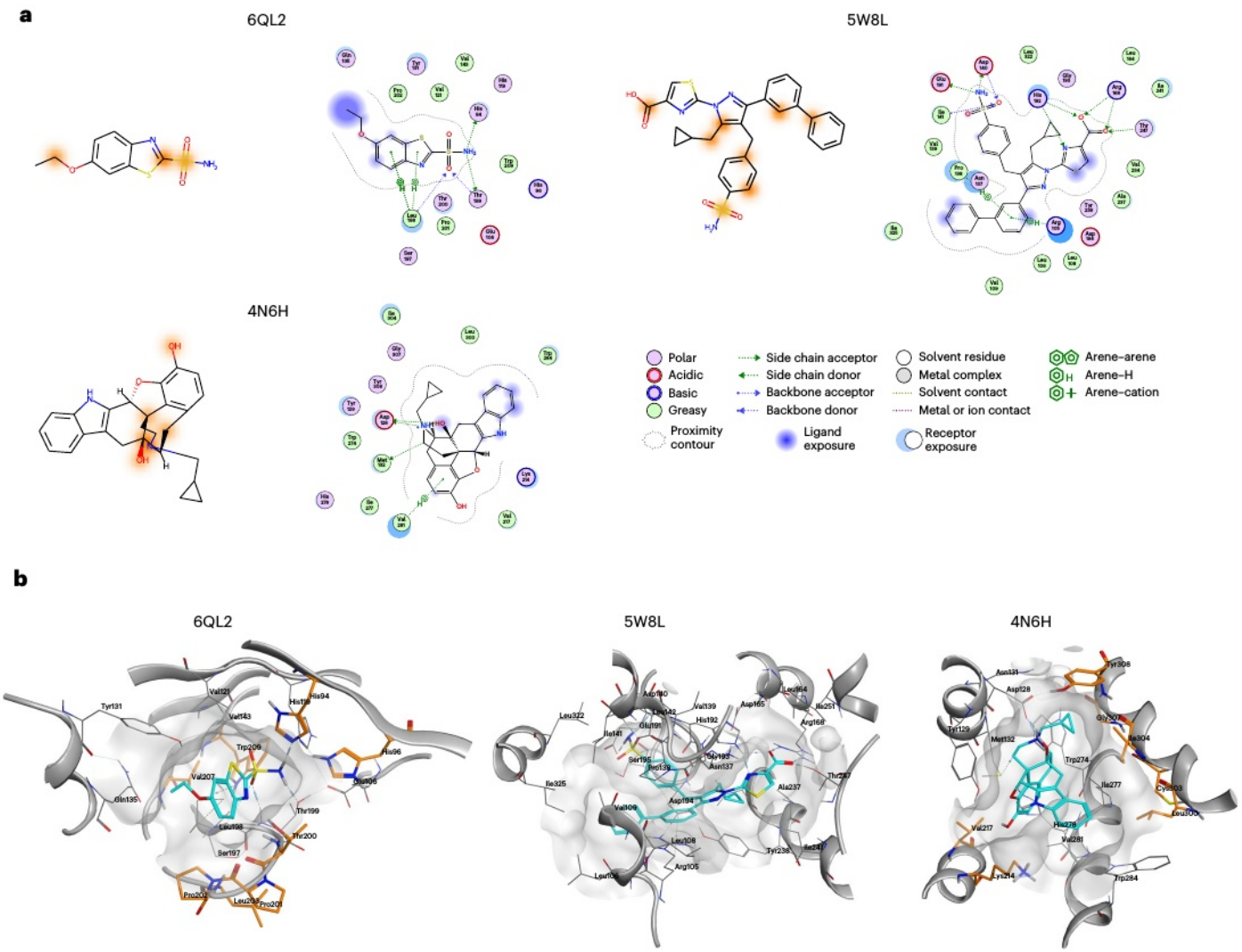

图4 | 用于可解释性研究的配体和结合位点的可视化。a，共晶配体的可解释性。每个面板的左侧显示了配体的二维结构，其中突出显示的原子（橙色）被预测为有助于蛋白质结合。所有结构均使用RDKit[56]进行可视化。此外，还提供了这些配体的相应晶体结构中的配体-蛋白质相互作用图（每个面板的右侧）。**b**，结合位点结构的可解释性。提供了配体 - 蛋白质结合口袋，突出显示正确预测的氨基酸残基（橙色），这些残基围绕着相应的配体（青色）。其余的氨基酸残基、二级结构元件和表面图被涂成灰色。所有配体 - 蛋白质相互作用图以及 X 射线结构的三维表示均使用分子操作环境（MOE）软件进行可视化。

与蛋白质的相互作用（氨基作为氢键供体与天冬氨酸 140 和谷氨酸 191 的侧链相互作用，磺酰胺氧作为氢键受体与天冬氨酸 140 和异亮氨酸 141 的主链相互作用）。同样，我们注意到羧酸基团也有部分突出显示（在 5W8L 中，羧酸氧原子作为氢键受体与精氨酸 168、组氨酸 192 和苏氨酸 247 的侧链相互作用）。此外，模型正确预测了联苯环参与配体 - 蛋白质结合（在 5W8L 中，与精氨酸 105 和天冬酰胺 137 的芳香环 - 氢相互作用）。尽管 9YA（与 5W8L 结合）比（与 6QL2 结合的）乙氧唑酰胺大得多且复杂得多，但该模型对大多数实验证实的相互作用显示出良好的可解释性潜力。

在第三个例子中，人类δ型阿片类受体与 EJ4 配体的 4N6H X 射线复合物中，EJ4 的主要相互作用官能团再次被正确突出显示。在此，脂肪环复合物的一个羟基和一个相邻的三级胺（均作为氢键供体与天冬氨酸 128 的侧链相互作用）被正确被解读为形成特定的相互作用。然而，苯酚基团被错误地预测为参与蛋白质结合。

至于更具挑战性的蛋白质序列可解释性，总体结果不如配体可解释性。尽管许多被预测可能参与配体结合的氨基酸残基实际上与相应的化合物相距甚远，但一些构成结合位点的氨基酸残基被正确预测（图4b）。例如，在6QL2复合物中，以下残基被突出显示：His94、His96、Thr200、Pro201、Pro202、Leu203、Val207和Trp209。其中，只有His94与乙氧唑酰胺形成特定的相互作用。在5W8L中，构成配体-蛋白质结合位点的残基没有被突出显示。然而，在4N6H中，结合位点内有几个残基被正确预测：Lys214、Val217、Leu300、Cys303、Ile304、Gly307和Tyr308。不幸的是，这些残基都没有与配体形成特定的相互作用。鉴于这些结果，预计蛋白质序列的可解释性会降低，因为

一维蛋白质序列（在我们的模型中用作蛋白质信息输入）不一定能表明结合口袋的三维构型和位置。然而，来自主要蛋白质序列的结果令人鼓舞，足以安全地假设将三维蛋白质信息进一步纳入建模框架最终将提高药物 - 靶点相互作用网络的模型可解释性。

此外，由于 DrugBAN 所提供的可解释性是从 DTI 数据本身自适应地学习到的，这种解释有可能发现一些尚未探索的局部相互作用的隐藏知识，并且能够帮助药物猎手改善给定支架的结合特性，或者降低化合物的脱靶风险。

结论

在这项工作中，我们提出了 DrugBAN，这是一个用于药物-药物相互作用（DTI）预测的端到端双线性注意力深度学习框架。我们将 CDAN（一个对抗性域适应网络）集成到建模过程中，以增强跨域泛化能力。与其他最先进的 DTI 模型和传统机器学习模型相比，实验结果表明，DrugBAN 在域内和跨域设置中都始终实现了 DTI 预测性能的提升。此外，通过将注意力权重映射到蛋白质亚序列和药物化合物原子，我们的模型能够为解释相互作用的本质提供生物学见解。所提出的想法具有普遍性，可以扩展到其他相互作用预测问题，例如药物-药物相互作用和蛋白质-蛋白质相互作用的预测。

这项工作重点研究基于化学生物学的药物靶点相互作用（DTI），其输入为 1D 蛋白质序列和 2D 分子图谱。鉴于高精度的 3D 结构蛋白质数量仅占已知蛋白质序列的一小部分，本研究未考虑利用此类结构信息进行建模。然而，DeepMind 的 AlphaFold41 在蛋白质 3D 结构预测方面取得了巨大进展，最近从 100 万种物种中生成了 20 亿个蛋白质 3D 结构预测。这种进展为在基于化学生物学的 DTI 预测中利用 3D 结构信息打开了大门。遵循成对局部交互学习和域自适应思想，我们相信在未来的工作中，进一步拓展我们的想法到复杂的三维结构上，能够带来更好的性能和可解释性。最后，本研究分别对不同的数据集进行了研究；将数据集整合与 DrugBAN 相结合将是另一个有趣的未来研究方向。

方法

双线性注意力网络

这是一种基于注意力的模型，最初是为了解决视觉问答（VQA）问题而提出的。给定一张图像和相关的自然语言问题，VQA 系统旨在提供一个文本-图像匹配的答案。因此，VQA 可以被视为一个多模态学习任务，类似于 DTI 预测。双线性注意力网络使用双线性注意力图优雅地扩展单注意力网络以适应多模态学习，通过考虑每对多模态输入通道（即图像区域和问题词的成对）来学习交互表示。与直接在多模态数据上使用单一注意力机制相比，双线性注意力网络可以提供更丰富的联合信息，同时将计算成本保持在相同规模。考虑到 VQA 和 DTI 问题之间的相似性，我们为 DTI 预测设计了一个受双线性注意力网络启发的成对交互模块。

领域适应

这些方法训练的模型能够减少源域和目标域之间的域分布变化，主要是在计算机视觉领域开发的。早期的域适应方法倾向于重新加权样本重要性或在浅层特征空间中学习不变特征表示，使用源域的有标记数据和目标域的无标记数据。最近，深度域适应方法将适应模块嵌入各种深度架构中，以学习可转移表示[43,44]。特别是，参考文献27提出了一种新颖的深度域适应方法CDAN，该方法将对抗网络与多元条件结合，用于可转移表示学习。通过将分类器预测信息引入对抗学习中，CDAN可以有效地对齐不同域中的数据分布。我们将CDAN作为适应模块嵌入DrugBAN中，以提高模型性能，用于跨域DTI预测。

DrugBAN 架构

CNN 用于蛋白质序列。蛋白质特征编码器由三个连续的一维卷积层组成，将输入的蛋白质序列转换为潜在特征空间中的矩阵表示。矩阵的每一行表示蛋白质中一个子序列的表示。借鉴词嵌入的概念，我们首先将所有氨基酸初始化为一个可学习的嵌入矩阵$Ep \in \mathbb{R}^{23\times D_p}$，其中23是氨基酸类型的数量，$D_p$是潜在空间的维度。通过查找Ep，每个蛋白质序列PP都可以初始化为相应的特征矩阵$Xp \in \mathbb{R}^{\Theta_p\times D_p}$。在这里，$\Theta_p$是蛋白质序列的最大允许长度，用于对齐不同长度的蛋白质序列并进行批量训练。遵循先前的研究[2,14,17]，超过最大允许长度的蛋白质序列会被截断，而长度较短的蛋白质序列会用零填充。

CNN 阻断式蛋白质编码器从蛋白质特征矩阵 X_p 中提取局部残基模式。在此，蛋白质序列被视为重叠的 3 个氨基酸，例如 METLCL... DSMN → MET、ETL、TLC...、DSM、DLK。第一个卷积层用于捕获核大小为 3 的 3 个残基级别的特征。然后接下来的两层继续扩大感受野，并学习局部蛋白质片段的更多抽象特征。蛋白质编码器的描述如下：

$$\mathbf{H}_p^{(l+1)} = \sigma(\mathrm{CNN}(\mathbf{W}_c^{(l)}, \mathbf{b}_c^{(l)}, \mathbf{H}_p^{(l)})), \quad (1)$$

其中 $\mathbf{W}_c^{(l)}$ 和 $\mathbf{b}_c$ 是可学习的权重矩阵（滤波器）和偏差，第 l 个卷积*神经网络*（CNN）层中的向量。$\mathbf{H}_p^{(l)}$ 是第 l 个隐藏蛋白质表示并且 $\mathbf{H}_p^{(0)} = \mathbf{P}\mathbf{X}_p$，$\sigma(\cdot)$ 表示一个非线性激活函数，在我们的实验中使用了 ReLU（·）。

用于分子图的 **GCN**。对于药物化合物，我们将每个 SMILES 字符串转换为其二维分子图GG。为了表示GG中的节点信息，我们首先通过其化学性质初始化每个原子节点，如 DGL-LifeSci[45] 包中所实现的那样。每个原子都表示为一个 74 维的整数向量，描述八条信息：原子类型、原子度、隐含氢的数量、形式电荷、自由基电子的数量、原子杂化、总氢的数量以及该原子是否芳香。类似于上述蛋白质序列的最大允许长度设置，我们设定了一个最大允许的节点数量Θ_d。节点数量较少的分子将包含用零填充的虚拟节点。因此，每个图的节点特征矩阵表示为 $\mathbf{Md} \in \mathbb{R}^{\Theta_d\times 74}$。此外，我们使用一个简单的线性变换来定义 $\mathbf{X}_d = \mathbf{P}\mathbf{W}_0\mathbf{M}_d$，从而得到一个实值密集矩阵 $\mathbf{Xd} \in \mathbb{R}^{\Theta_d\times D_d}$ 作为输入特征。

我们使用一个三层 GCN 模块来有效地学习药物化合物的图表示。GCN 将卷积算子泛化到不规则域。具体而言，我们通过聚合其相应的邻域原子集来更新原子特征向量，这些原子由化学键连接。这种传播机制会自动捕获分子的子结构信息。我们保留节点级别的药物表示，以便后续明确地学习与蛋白质片段的局部相互作用。药物编码器如下所示

$$\mathbf{H}_d^{(l+1)} = \sigma(\mathrm{GCN}(\tilde{\mathbf{A}}, \mathbf{W}_g^{(l)}, \mathbf{b}_g^{(l)}, \mathbf{H}_p^{(l)})), \quad (2)$$

其中 **W(l** $_g$ 和 $\mathbf{b}_g$ 是特定层的可学习权重矩阵并且 GCN 的偏差向量，**A 是**在分子图 G 中添加自连接的邻接矩阵，$\mathbf{H}^{(l)}{}_d$ 是第 *l* 个隐藏节点表示为 **H(0)** $_d$ P **Xd** 。

成对交互学习。我们应用双线性注意力网络模块来捕获药物和蛋白质之间的成对局部交互。它由两层组成：用于捕获成对注意力权重的双线性交互图，以及用于从交互图上提取药物 - 靶点联合表示的双线性池化层。

在第三层中，分别的卷积神经网络（CNN）和图卷积网络（GCN）编码器生成隐藏的蛋白质和药物表示 $\mathbf{H}^{(3)}$ p P{h1p， **h2p**, ..., **h**Mp } 以及 $\mathbf{Hd}^{(3)}$ P{$\mathbf{h}_{1_d}$,$\mathbf{h}_{2_d}$ 等等，其中 *M* 和 *N* 分别表示蛋白质中编码子结构的数量和药物中原子的数量。我们使用这些隐藏表示来构建双线性交互图，以获得一个单一的标量配对交互矩阵 **I** ∈☒N×M：

$$\mathbf{I} = ((\mathbf{1}\cdot\mathbf{q}^\top)\circ\sigma((\mathbf{H}_d^{(3)})^\top\mathbf{U}))\cdot\sigma(\mathbf{V}^\top\mathbf{H}_p^{(3)}), \quad (3)$$

其中，**U** ∈☒Dd×K 和 **V** ∈☒Dp×K 是用于药物和蛋白质表示的可学习权重矩阵，**q** ∈☒K 是可学习的权重向量，**1** ∈☒N 是固定的全1向量，☒ 表示哈达玛（元素级）积。**I** 中的元素表示各自药物-靶标亚结构对的相互作用强度，映射到潜在的结合位点和分子亚结构。为了直观地理解双线性相互作用，方程（3）中的元素 $\mathbf{I}_{ij}$ 也可以写成

$$\mathbf{I}_{ij} = \mathbf{q}^\top(\sigma(\mathbf{U}^\top\mathbf{h}_d^i)\circ\sigma(\mathbf{V}^\top\mathbf{h}_p^j)), \quad (4)$$

H(3) 的第 *i* 列在哪里**隐藏**?

^{d} 并且$\mathbf{h}_{pj}$ 是 ***H*(3)** $_p$ 的第 j 列，具体来说具体而言，表示药物和蛋白质的第 *i* 和第 j 个子结构表示。因此，我们可以看到双线性交互作用是首先将表示 **hid** 和 $\mathbf{h}_{pj}$ 映射到公共特征空间，使用权重矩阵 **U** 和 **V**，然后在哈达玛积和向量 **q** 的权重上学习交互作用。通过这种方式，成对交互作用为子结构对预测结果的贡献提供了可解释性。

为了获得联合表示 **f'**∈☒**K**，我们在交互图 **I 上引入了一个双线性池化层**。具体而言，**f**is 的第 k 个元素计算如下

$$\begin{aligned}\mathbf{f}'_k &= \sigma((\mathbf{H}_d^{(3)})^\top\mathbf{U})_k^\top\cdot\mathbf{I}\cdot\sigma((\mathbf{H}_p^{(3)})^\top\mathbf{V})_k \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{I}_{ij}(\mathbf{h}_d^i)^\top(\mathbf{U}_k\mathbf{V}_k^\top)\mathbf{h}_p^j,\end{aligned} \quad (5)$$

其中，$\mathbf{U}_k$ 和 $\mathbf{V}_k$ 表示权重矩阵 **U** 和 **V** 的第 k 列。值得注意的是，在这一层没有新的可学习参数。权重矩阵 **U** 和 **V** 与前面的交互图层共享，以减少参数数量并缓解过拟合。此外，我们对联合表示向量进行求和池化，以获得一个紧凑的特征图：

$$\mathbf{f} = \mathrm{SumPool}(\mathbf{f}', s), \quad (6)$$

其中，SumPool（☒） 函数是一个具有步长 *s* 的一维非重叠求和池化操作。它将 **f'**∈☒**K** 的维度降低到 **f** ∈☒K/s。此外，我们可以通过计算多个双线性交互图将单个成对交互扩展为多头形式。最终的联合表示向量是各个头的总和。由于权重矩阵 **U** 和 **V** 是共享的，每个额外的头只增加一个新的权重向量 **q**，这是参数高效的。在我们的实验中，多头交互比单头交互表现更好。

因此，利用新颖的双线性注意力机制，模型能够明确地学习药物与蛋白质之间的成对局部相互作用。这种交互模块的灵感和改编来自于参考文献[26,25]，其中为视觉问答（VQA）问题设计了两个双线性模型。为了计算交互概率，我们将联合表示 **f** 输入到解码器中，解码器是一个全连接分类层，后面跟着一个 sigmoid 函数：

$$p = \mathrm{Sigmoid}(\mathbf{W}_o\mathbf{f} + \mathbf{b}_o), \quad (7)$$

其中$\mathbf{W}_o$ 和 $\mathbf{b}_o$ 是可学习的权重矩阵和偏差向量。

最后，我们通过反向传播共同优化所有可学习的参数。训练目标是使交叉熵损失最小化，如下所示：

$$\mathcal{L} = -\sum_i(y_i\log(p_i) + (1-y_i)\log(1-p_i)) + \frac{\lambda}{2}\|\Theta\|_2^2, \quad (8)$$

其中，**Θ 是**上述所有可学习的权重矩阵和偏差向量的集合，y_i 是第 *i* 个药物 - 靶点对的真实标签，p_i 是该模型对其的输出概率，λ 是用于 L2 正则化的超参数。

跨域适应以提高泛化能力。机器学习模型往往在来自相同分布的相似数据（即域内数据）上表现良好，但在具有不同分布的异构数据（即跨域数据）上表现较差。提高模型在跨域 DTI 预测中的性能是一个关键挑战。在我们的框架中，我们嵌入 CDAN 以增强从具有足够标记数据的源域到只有无标记数据的靶域的泛化能力。

给定一个源域 SS_s P{Px $_{si}$， y_{si})}$_{i=1}$ 纳秒

关于N_s 标记的药物 - 靶点对和目标域 SS_t P{x_i }$_{j=1}$of N_t 未标记的药物 - 靶点对 时间

我们使用 CDAN 来对齐它们的分布，并提高跨域的预测性能。图 1c 展示了在我们的框架中 CDAN 的工作流程，包括三个关键组成部分：特征提取器 *F*(☒)，解码器 G(☒)和域判别器 *D*(☒)。我们使用 F（☒） 来表示单独的特征编码器和双线性注意力网络一起生成输入域数据的联合表示；即，**fsi** PFPx $_{fi}$)and **ftj** PFPx $_{tj}$)。接下来，我们使用上述的全连接分类层，并使用 softmax 函数作为 *G*(☒) 来获得分类器预测 **gsi** PGPf $_{fi}$) ∈☒ 2 和 **gtj** PGPf $_{ft}$) ∈☒2。此外，我们应用一个多元映射将联合表示 **f** 和分类器预测 **g** 嵌入到联合条件表示 **h** ∈☒2K/s 中，h 被定义为这两个向量的外积的展平：

$$\mathbf{h} = \mathrm{FLATTEN}(\mathbf{f}\otimes\mathbf{g}), \quad (9)$$

其中⊗ 表示外积。

多元映射捕捉了两个独立分布之间的乘法交互作用[46,47]。遵循 CDAN 机制，我们通过将域判别器 *D*(☒) 的条件设定为 **h** 来同时校准源域和目标域的联合表示和预测分类分布。域判别器 *D*(☒) 由三层全连接网络组成，用于学习区分联合条件表示 **h** 是来自源域还是目标域。相反，特征提取器 *F*(☒) 和解码器 *G*(☒) 被训练以最小化具有源域标签信息的源域交叉熵损失☒，同时生成难以区分的表示 **h** 来迷惑域判别器 *D*(☒)。因此，我们可以在跨域建模中定义这两个损失：

$$\mathcal{L}_s(F,G) = \mathbb{E}_{(x_i^s,y_i^s)\sim\mathcal{S}_s}\mathcal{L}(G(F(x_i^s)),y_i^s), \quad (10)$$

$$\mathcal{L}_{adv}(F,G,D) = \mathbb{E}_{x_i^t\sim\mathcal{S}_t}\log(1-D(\mathbf{f}_i^t,\mathbf{g}_i^t)) + \mathbb{E}_{x_j^s\sim\mathcal{S}_s}\log(D(\mathbf{f}_j^s,\mathbf{g}_j^s)), \quad (11)$$

其中， 是在有标签的源域上的交叉熵损失，Lado 是用于域判别的对抗损失。优化问题被表述为一个极大极小范例：

$$\max_{D} \min_{F,G} \mathcal{L}_s(F, G) - \omega \mathcal{L}_{adv}(F, G, D), \quad (12)$$

其中 w > 0 是一个超参数，用于对 Aadv 进行加权。通过在 Aad 上引入对抗训练，我们的框架能够减少源域和目标域之间的数据分布差异，从而在跨域预测中提高泛化能力。

实验设置

数据集。我们在三个公开的DTI数据集上评估了DrugBAN和五个最先进的基准模型：BindingDB、BioSNAP和Human。BindingDB数据集是一个网络可访问的数据库，包含实验验证的结合亲和力，主要关注小类药物分子和蛋白质的相互作用。我们使用在我们早期工作中构建的低偏差版本的BindingDB数据集（参考文献*），并采用补充信息第2节中描述的减少偏差的预处理步骤。BioSNAP数据集由DrugBank数据库创建，包含4,510种药物和2,181种蛋白质。它是一个平衡的数据集，包含经过验证的正向相互作用，以及从未见过的对中随机获得的等数量负样本。Human数据集由参考文献3构建，包括通过计算筛选方法获得的高可信负样本。遵循先前的研究l4.1620，我们还使用了包含相同数量正负样本的平衡版本Human数据集。为了减轻隐藏数据偏差的影响l6，我们在Human数据集上使用额外的冷对分裂进行性能评估。补充表 2 展示了这三个数据集的统计数据。

实现。DrugBAN 是在 Python 3.8 和 Py Torch 1.7.1（参考文献 5）中实现的，同时还使用了来自 DGL 0.7.1（参考文献 2）和 DGL-ifeSci 的函数。
0.2.8（参考文献 4）、Scikit-learn 1.0.2（参考文献 3）、Numpy 1.20.2（参考文献 5）、Pandas
1.2.4（参考文献55）和RDKit 2021.03.2（参考文献59）。批量大小设置为64，使用Adam优化器，学习率为Se-5。对于所有数据集，我们允许模型最多运行100个epoch。在验证集上给出最佳AUROC分数的epoch选择最佳性能模型，然后使用该模型评估测试集上的最终性能。蛋白质特征编码器由三个IDCNN层组成，迭代次数为[128, 128, 128]，核大小为[3, 6, 9]。药物特征编码器由三个GCN层组成，隐藏维度为[128, 128, 128]。蛋白质设定的最大允许序列长度为1,200，药物分子的最大允许原子数为290。在双线性注意力模块中，我们只使用两个注意力头以提供更好的可解释性。潜在嵌入大小k设置为78，求和池化窗口大小为3。全连接解码器中的隐藏神经元数量为512。我们的模型性能对超参数设置不敏感。配置细节和敏感性分析在补充信息第3节中提供。我们在补充信息第 7 节中还展示了一项可扩展性研究。

基准测试。我们将 DrugBAN 在 DTI 预测方面的性能与以下五个模型的性能进行了比较。首先和第二，两种浅层机器学习方法，即支持向量机（SVM）和随机森林（RF），应用于连接的手印指纹 ECFP4 和 PSC 特征。第三，“DeepConv-DTI”，它使用卷积神经网络（CNN）和一个全局最大池化层来提取蛋白质序列中的局部模式，并使用全连接网络对药物指纹 ECFP4 进行编码。第四，“GraphDTA”，它使用图神经网络对药物分子图进行编码，并使用卷积神经网络对蛋白质序列进行编码来对 DTI 进行建模。将学习到的药物和蛋白质表示向量通过简单的连接进行组合。为了将 Graph DTA 从原始回归任务适应到二元分类任务，我们遵循早期文献中的步骤，在其最后一个全连接层中添加 Sigmoid 函数，然后使用交叉熵损失对其参数进行优化。第五，“MolTrans”，一个深度学习模型，它适应了 Transformer 架构来编码药物和蛋白质信息，并使用基于卷积神经网络的交互模块来学习亚结构相互作用。对于上述深度 DTI 模型，我们遵循其原始论文中推荐的模型超参数设置。

报告摘要

关于研究设计的更多信息可在本文所链接的《自然》系列报告摘要中获取。

数据可用性

本研究中使用的实验数据可在https://github.com/peizhenbai/DrugBAN/tree/main/datasets 网站上获取。本研究中使用的所有数据均来自公共资源。BindingDB 的来源可在https://www.bindingdb.org/bind/index.jsp 网站上找到；BioSNAP1730 的来源可在https://github.com/kexinhuang12345/MolTrans/tree/master/dataset/BIOSNAP/full_data 网站上找到，而之前研究中使用的“人类”来源可在 https://github.com/ifanchen-simm/transformerCPI/blob/master/Human%2CC.elegans/dataset/human_ 网站上找到。
data.txt. 来自 PDB 的共晶配体可通过搜索其 PDB 编号在 https://www.rcsb.org 网站上获取。

代码可用性

DrugBAN 的源代码和实现细节在 GitHub 库（https://github.com/peizhenbai/DrugBAN）和 CodeOcean 胶囊（https://doi.org/10.24433/CO.3558316.v1）中均可免费获取。该代码也在 Zenodo 存档（https://doi.org/10.5281/zenodo.7231657）。

参考文献

1. Luo, Y. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning
from heterogeneous information. Nat. Commun. 8, 1-13 (2017).
2. Öztürk, H., Olmez, E. O. & Özgür, A. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 34, 1821-i829 (2018).
3. Yamanishi, Y., Araki, M., Gutteridge, A, Honda, W. & Kanehisa, M. Prediction of drug-target interaction networks from the
integration of chemical and genomic spaces. Bioinformatics 24, 1232 - i240 (2008).
4. Zitnik, M. et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf. Fusion 50, 71-91 (2019).
5. Bagherian, M. et al. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Brief. Bioinform. 22, 247-269 (2021).
6. Wen, M. et al. Deep-learning-based drug-target interaction prediction. J. Proteome Res. 16, 1401-1409 (2017).
7. Sieg, J, Flachsenberg, F.& Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. J. Chem. Inf. Model. 59, 947-961 (2019).
☒. Lim, S. et al. A review on compound-protein interaction prediction methods: data, format, representation and model. Comput.
Struct. Biotechnol. J. 19,1541-1556 (2021).
☒. Gao, . Y. et al. Interpretable drug target prediction using deep neural representation. In Int. Joint Conf. on Artificial Intelligence (IJCA) 3371-3377 (2018).
10. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat. Rev. Genet. 5, 262-275 (2004).
11. Lee, I, Keum, J. & Nam, H. DeepConv-DT: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLos Comput. Biol. 15, 1007129 (2019).

12. Hinnerichs, T.& Hoehndorf, R. DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug-target interactions. Bioinformatics 37, 4835-4843 (2021).

13. Nguyen, T. et al. GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics 37, 1140-1147 (2021).

14. Tsubaki, M., Tomil, K. & Sese, J. Compound protein interaction prediction with end to end learning of neural networks for graphs and sequences. Bioinformatics 35, 309-318 (2019).

15. Feng, Q. Dueva, E., Cherkasov, A. & Ester, M. PADME: a deep learning-based framework for drug-target interaction prediction.

Preprint at arxiv https://arxiv.org/b/807.09741 (2018).

16. Chen, L. et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with

self-attention mechanism and label reversal experiments.

Bioinformatics 36, 4406-4414 (2020).

17. Huang, K., Xiao, C., Glass, L. & Sun, J. MolTrans: molecular interaction transformer for drug-target interaction prediction.

Bioinformatics 37, 830-836 (2021).

18. Schenone, M., Dancik, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. Nat. Chem. Biol. 9, 232-40 (2013).

19. Öztürk, H., Ozkirimli, E. & Özgür, A. WideDTA: prediction of drug-target binding affinity. Preprint at arxiv https://ariv.org/

abs/1902.04166 (2019).

20. Zheng, S. Li, Y, Chen, S., Xu, J. & Yang, Y. Predicting drug-protein interaction using quasi-visual question answering system. Nat. Mach. Intell. 2, 134-140 (2020).

21. Abbasi, K. et al. DeepCDA: deep cross-domain compound-protein affinity prediction through lstm and convolutional neural networks. Bioinformatics 36, 4633-4642 (2020).

22. Kao, P.-Y., Kao, S.-M. Huang, N.L.& Lin, Y-C. Toward drug-target interaction prediction via ensemble modeling and transfer

learning. In IEEE Int. Conf. on Boinformatics and Biomedicine

(BIBM) 2384-2391 (2021).

23. Abbasi, K., Razzaghi, P, Poso, A., Ghanbari-Ara, S. & Masoudi-Nejad, A. Deep learning in drug target interaction

prediction: current and future perspectives. Curr. Med. Chem. 28, 2100-2113 (2021).

24. Kipf, T. & Welling, M. Semi-supervised classification with graph convolutional networks. In Int. Conf. on Learning Representations

(ICLR, 2017).

25. Yu, Z. Yu, J, Xiang, C., Fan, J. & Ta, D. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. 29,

5947-5959 (2018).

26. Kim J. H. Jun, J. & Zhang, B. T. Bilinear attention networks.In Advances in Neural Information Processing Systems (NeuriPS,

2018).

27. Long, M., Cao, Z. Wang, J. & Jordan, M. I. Conditional adversarial domain adaptation. In Advances in Neural Information Processing Systems (NeurlPS, 2018).

28. Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules.

J. Chem. Inf. Comput. Sci. 28, 31-36 (1988).

29. Liu, T., Lin, Y., Wen, X., Jorissen, R. N.& Gilson, M. K. BindingDB: a web-accessible database of experimentally determined

protein-ligand binding affinities. Nucleic Acids Res. 35, D198-D201 (2007).

30. Zitnik, M., Sosic, R., Maheshwari, S. & Leskovec, J. BioSNAP datasets: Stanford biomedical network dataset collection. https://snap.stanford.edu/biodata (2018).

31. Liu, H., Sun, J., Guan, J, Zheng, J. & Zhou, S. Improving compound-protein interaction prediction by building up highly credible negative samples. Bioinformatics 31, i22-229 (2015).

32. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742-754 (2010).

33. Cao, D., Xu, Q. & Liang, Y. Propy: a tool to generate various modes of chou's pseaac. Bioinformatics 29, 960-962 (2013).

34. Cortes, C. & Vapnik, V. Support-vector networks. Mach. Learn. 20, 273-297 (1995).

35. Ho, T.K. Random decision forests. In Int. Conf. on Document Analysis and Recognition, vol. 1, 278-282 (1995).

36. Ganin, Y. et al. Domain-adversarial training of neural networks. In J. Mach. Learn. Res. 17, 1-35 (2016).

37. Kazokaite, J. et al. Engineered carbonic anhydrase vi-mimic enzyme switched the structure and affinties of inhibitors. Sci. Rep. 9,1-17 (2019).

38. Rai, G. et al. Discovery and optimization of potent, cell-active pyrazole-based inhibitors of lactate dehydrogenase (Idh). J. Med. Chem. 60, 9184-9204 (2017).

39. Fenalti, G. et al. Molecular control of -opioid receptor signalling. Nature 506, 191-196 (2014).

40. Berman, H. M. et al. The protein data bank. Nucleic Acids Res. 28, 235-242 (2000).

41. Jumper, J. M. et al. Highly accurate protein structure prediction with alphafold. Nature 596, 583-589 (2021).

42. Pan, S. J. & Yang, Q. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345-1359 (2010).

43. Gong, B. Grauman, K. & Sha, F. Connecting the dots with landmarks: discriminatively learning domain-invariant features for unsupervised domain adaptation. In Int. Conf. on Machine Learning (ICML) 222-230 (2013).

44. Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. & Schölkopf, B. Correcting sample selection bias by unlabeled data. In Advances in Neural Information Processing Systems (NIPS) 601-608 (2006).

45. Li, M. et al. DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. ACS Omega 6, 27233-27238 (2021).

46. Song, L., Huang, J., Smola, A. & Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Int. Conf. on Machine Learning (ICML) 961-968 (2009).

47. Song, L. & Dai, B. Robust low rank kernel embeddings of multivariate ditributions. In Advances in Neural Information Processing Systems (NIPS) 3228-3236 (2013).

48. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045-D1053 (2016).

49. Bai, P. et al. Hierarchical clustering split for low-bias evaluation of drug-target interaction predicton. In IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM) 641-644 (2021).

50. Wishart, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 36, D901-D906 (2008).

51. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems (NeuriPS, 2019).

52. Wang, M. t al. Deep graph library: a graph-centric, highly-performant package for graph neural networks. Preprint at arXiv https://arxiv.or/abs/1909.01315 (2019).

53. Pedregosa, F. et al. Sciki-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825-2830 (2011).

54. Harris, C. R. et al. Array programming with numpy. Nature 585, 357-362 (2020).

55. The pandas development team. pandas-dev/pandas: Pandas 1.2.4. Zenodo https://doi.org/10.5281/zenodo.4681666(2021).

56. Landrum, G. et al. RDKit: open-source cheminformatics. https//github.com/rdkit/rdkit (2006).
57. Bai, P. Milikovic, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. CodeOcean https://doi.org/10.24433/C0.3558316.M1(2022).
58. Bai, P, Milikovi⌧, F, John, B. & Lu, H. peizhenba/drugban: V1.2.0. Zenodo https://doi.org/0.5281/zenodo.7231657 (2022).
59. Kim, J.-H. et al. Hadamard product for low-rank milinear pooling. In Int. Conf. on Learning Representations (ICLR, 2017).

致谢

我们感谢周先生、刘先生和施博布女士对这项工作的有益建议和讨论。P.B. 得到了谢菲尔德大学工程学院研究奖学金（拨款编号：169426530）的支持。

作者贡献

P.B. F.M、B.J. 和 H. 构思并设计了这项工作。P.B. 在 B.J. 和 H.L 的指导下开发模型并进行了实验。F.M. 和 P.B. 分析数据并进行了方法比较。F.M. 为材料和分析工具做出了贡献。所有作者都为撰写论文做出了贡献。

利益冲突

作者声明不存在利益冲突。

附加信息

补充信息 在线版包含补充材料，可在 https://doi.org/10.1038/s42256-022-00605-1 获取。

通信和材料请求应寄给陆海平。

同行评审信息 《自然·机器智能》感谢匿名评审员对本研究工作的同行评审所做的贡献。

自然研究

通讯作者：陆海平
上次由作者更新：2022 年 12 月 14 日

报告摘要

自然研究希望提高我们所发表的研究工作的可重复性。此表格为报告的一致性和透明度提供了结构。有关自然研究政策的更多信息，请参阅我们的编辑政策以及编辑政策检查表。

统计数据

对于所有的统计分析，请确认以下术语在图中列出：样本、偏差、主要测试、假设检验。

N/A 已确认

- [ ] [x] 每个实验组/条件的确切样本量（n），以离散数字和测量单位的形式给出
- [ ] [x] 关于测量是从不同的样本中获取，还是对同一样本进行了重复测量的一份声明
- [x] [ ] 所使用的统计检验以及它们是单侧检验还是双侧检验
  对于常见的测试，应仅以名称来描述；更复杂的测试技术应在“方法”部分进行描述。
- [x] [ ] 对所有测试协变量的描述
- [x] [ ] 对任何假设或校正的描述，例如正态性检验以及针对多重比较的校正
- [ ] [x] 对统计参数的完整描述，包括集中趋势（例如均值或其他基本估计值，如回归系数）和变异（例如标准偏差），或相关的不确定性估计值（例如置信区间）
- [x] [ ] 对于零假设检验，检验统计量（例如 F、t、ri 与置信区间、效应量、自由度和 P 值），只要合适，就给出 P 值的精确值。
- [x] [ ] 对于贝叶斯分析，有关先验选择和马尔可夫链蒙特卡罗设置的信息
- [x] [ ] 对于分层和复杂的设计，确定测试的适当级别并充分报告结果
- [x] [ ] 效应大小的估计值（例如科恩的 d 值、皮尔逊的 ）i，表明我们所评估的

我们关于博艾斯统计数据的网页收集包含了大量的信息。

软件与代码

有关计算机代码可用性的政策信息

数据收集	Python 3.8、Pandas 1.2.4 以及 RDKit 2021.03.2
数据分析	Python 3.8、PyTorch 1.7.1、DGL 0.7.1、DGLlife 0.2.8、Sciki-lean 1.2、NumPy 1.20.2、Pandas 1.2.4、DKit 2021.03.2 以及 MOE 2020.09 。 我们还将在 GitHub（https://github.com/peizhenbai/DrugBAN）和 Zenodo（https://doi.org/10.5281/zenodo.7231657）上提供本研究的源代码。

对于利用自定义算法或软件的手稿，这些算法或软件对研究至关重要，但尚未在已发表的文献中有所描述，必须向编辑和审稿人提供软件。我们强烈鼓励将代码存放在社区存储库（例如 Gitub）中，有关如何提交软件补充信息的《自然研究指南》请参阅。

数据

有关数据可用性的政策信息

所有手稿都必须包含一份数据。 可用性声明。本声明应在适用的情况下提供以下信息：

公开可用数据集的访问代码、唯一标识符或网络链接

一份具有相关原始数据的图形列表

- 关于数据可用性的任何限制的描述

本研究中使用的实验数据可在我们的公共存储库 https://github.com/peizhenbai/DrugBAN/ree/main/datasets 中获取。所有数据集均来自公共资源。BindingDB 的来源是 https://www.bindingd.or/bind/index.jsp 。BiosNAP 的来源是 https://github.com/exinhuang12345/MolTrans/tree/master/dataset/BIOSNAP/full_data 。之前研究中使用的人类数据集可在 http://github.com/fanchen-simm/ransformerCP/blob/master/Human%2CC.elegans/dataset/human_data.txt 中获取。来自蛋白质数据库（PDB）的共晶体配体可在 https://w.rcsb.org 上获取。

特定领域的报告

请选择最适合您研究的那一项。如果您不确定，请在做出选择之前阅读相应的部分。

生命科学　　行为与社会科学　　生态、进化与环境科学

有关包含所有章节的文档参考副本，请参阅 aturcomdcmens/r-reortin-sumar-flat.odf 。

生命科学研究设计

所有研究都必须披露这些要点，即便披露的内容是负面的。

样本规模

我们研究了三个公开数据集：BindingDB、BioSNAP 和 Human，其样本规模分别为 49k、27k 和 6.7k。确定这些样本规模时考虑了以下三个方面：i）我们研究了该领域的发展，并选择与大多数最先进工作相同的样本规模；ii）我们选择声誉高且广泛使用的公开数据集；iii）我们使用的交互数据经过了实验验证。通过这种方式，我们选择的样本规模足以与最先进的工作进行公平的性能评估。

数据排除

如上所述选择数据集之后，在本研究中没有排除任何数据。

复制

为了验证我们实验结果的可重复性，我们在每个实验中进行了五次独立运行，并报告了平均值。
并且利用标准偏差来对重复实验进行定量评估。源代码和数据可在我们的公共 GitHub 上获取。
供其他研究人员复制的资料库。

随机化

我们使用两种分割策略将数据样本随机分配到实验组（分割组）中。第一种分割策略仅仅是随机分割。
它将随机划分的药物 - 靶点对（数据样本）分为训练集、验证集和测试集。第二种划分策略是基于聚类的。
配对拆分，用于评估在分布外数据上的预测性能。这种第二策略首先将原始数据聚类为簇。
然后将这些簇随机划分为不同的集合。因此，这两种策略均基于随机样本分配。

令人目眩神迷

在数据收集和分析期间，我们对分组分配情况一无所知。分组分配过程是通过计算机脚本进行的。
无需任何人工干预。

关于特定材料、系统和方法的报告

我们要求作者提供许多研究中使用的某些类型的材料、实验系统和方法的相关信息。在此，请表明列出的每种材料、系统或方法是否与您的研究相关。如果您不确定某项列表是否适用于我们的研究，请在选择回答之前阅读相应的部分。

材料与实验系统

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

方法

N/A	参与研究
☐	染色质免疫沉淀测序
☒	流式细胞术
☒	基于磁共振成像（MRI）的神经影像学