nature machine intelligence

文章

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

具有领域自适应功能的可解释双线性注意力网络提升了药物 - 靶点预测效果

按照作者提供的格式，未经
编辑

目录

1 基于聚类的成对拆分策略

正如正文所述，我们分别对 BindingDB 和 BioSNAP 数据集中的药物化合物和靶蛋白进行聚类，以进行跨域性能评估。具体而言，我们选择单连接聚类，这是一种自底向上的层次聚类方法，以确保不同聚类中样本之间的距离始终大于预先定义的距离，即最小距离阈值 γ。这种特性可以防止聚类过于接近，有助于生成跨域场景。

我们使用二值化的 ECFP4 特征来表示药物化合物，使用积分 PSC 特征来表示目标蛋白。为了准确测量成对距离，我们分别在 ECFP4 和 PSC 上使用杰卡德距离和余弦距离。在药物和蛋白质聚类中，我们均选择 γ= 0.5，因为这一选择可以防止聚类过大，并确保不同类别的样本相互分离。对于 BindingDB 数据集，我们获得了 2780 个药物聚类和 1693 个蛋白质聚类；对于 BioSNAP 数据集，我们获得了 2387 个药物聚类和 1978 个蛋白质聚类。表 1 展示了聚类结果中十个最大聚类的样本数量。结果表明，在药物聚类方面，BindingDB 的聚类分布比 BioSNAP 更加均衡。此外，在两个数据集中，蛋白质聚类结果倾向于生成许多仅包含少量蛋白质的小聚类，这表明蛋白质之间的平均相似度低于药物之间的平均相似度。我们从聚类结果中随机选取 60% 的药物聚类和 60% 的蛋白质聚类，并将与它们相关的所有药物 - 目标对视为源域数据。剩余聚类中的相关配对被视为源域数据。我们使用不同的随机种子进行五次独立的基于聚类的配对拆分，以用于下游模型的训练和评估。基于聚类的配对拆分通过考虑药物或蛋白质之间的相似性，能够定量地构建跨域任务。

表 1.基于聚类的成对拆分在 BindingDB 和 BioSNAP 数据集中生成的十个最大簇的大小。

数据集	对象	#1	#2	#3	#4	#5	#6	#7	#8	#9	# 10
BindingDB	药物	598	460	304	290	253	250	203	202	198	158
BioSNAP	药物	294	267	75	68	36	35	28	26	24	24
BindingDB	蛋白质	17	15	15	12	10	10	10	9	9	8
BioSNAP	蛋白质	8	8	8	6	5	4	4	4	4	4

2 数据集统计信息、符号说明及预处理步骤

表 2 展示了实验数据集的统计信息，表 3 列出了本文所使用的符号及其说明。BioSNAP 数据集由 Huang 等人（2021 年）创建 1，Human 数据集由 Liu 等人（2015 年）创建 2。对于 BindingDB 数据集，我们依照之前研究中的降偏预处理步骤 4 从 BindingDB 数据库源 3 中创建了一个低偏倚版本：i）我们仅将 IC50 小于 100 nM 的药物 - 靶点对视为阳性，仅将 IC50 大于 10,000 nM 的药物 - 靶点对视为阴性，以 100 倍的差异来减少类别标签噪声。这些 IC50 阈值是根据之前的研究 5,6 选定的。ii）我们移除了所有药物仅有一种类型配对（阳性或阴性）的药物 - 靶点对，以改善药物层面的配对类别平衡，并减少仅基于药物特征就能做出正确预测的潜在配体偏差。

表 2.实验数据集统计

数据集	# 药物	# 蛋白质	# 互动
BindingDB[4] 数据库	14,643	2,623	49,199
BioSNAP[1]	4,510	2,181	27,464
人类[2]	2,726	2,001	6,728

3 超参数设置与敏感性分析

表 4 展示了实验中所用模型超参数及其取值的列表。由于我们的模型性能对这些参数不敏感，对于超参数设置，我们在所有实验数据集（BindingDB、BioSNAP 和 Human）上使用相同的超参数。

图 1 展示了在 BindingDB 验证集上，采用不同超参数选择时的学习曲线，其中包括 4 个双线性嵌入大小、学习率和注意力头。这表明性能差异不大，通常在 30 到 40 个周期之间收敛。

表 3.符号和描述

符号表示法	描述
对于每个 \(p \in \mathbb{R}^{23 \times D}	蛋白质氨基酸嵌入矩阵
f 属于实数集 R 的 K 除以 s 次幂	药物-靶点联合表示
F(⊠),G(⊠),D(⊠)	在 CDAN 中的特征提取器、解码器和领域判别器
g 属于 R^2	通过 softmax 函数输出交互概率
$H_p^{(l)}$,$H_D^{(l)}$	在第 L 层卷积*神经网络(图卷积网络)中的*蛋白质（药物）的隐藏表示
I 属于 R 的 N×M 维空间	药物与蛋白质亚结构之间的成对相互作用矩阵
记 \(M_d \in \mathbb{R}^{\Theta_d \times 74}\)	根据其化学性质生成药物节点特征矩阵
P∈⊠1	通过 Sigmoid 函数输出交互概率
⊠, ⊠	蛋白质氨基酸序列，药物二维分子图
q 属于实数集的 K 次幂	双线性变换的权重向量
U∈R^(D×d×K)	用于编码药物表示的权重矩阵
V 属于 R 的 Dp×K 维实数空间	编码蛋白质表示的权重矩阵
Wc，bc	蛋白质卷积神经网络编码器的权重矩阵和偏差
Wg，bg	药物 GCN 编码器的权重矩阵和偏差
哇，好。	解码器的权重矩阵和偏差
\(X_p \in \mathbb{R}^{\Theta_p \times D_p}\)	潜在蛋白质基质表示层
对于所有的 \(d \in \mathbb{R}^{\Theta_d \times D_d}	潜在药物矩阵表示。

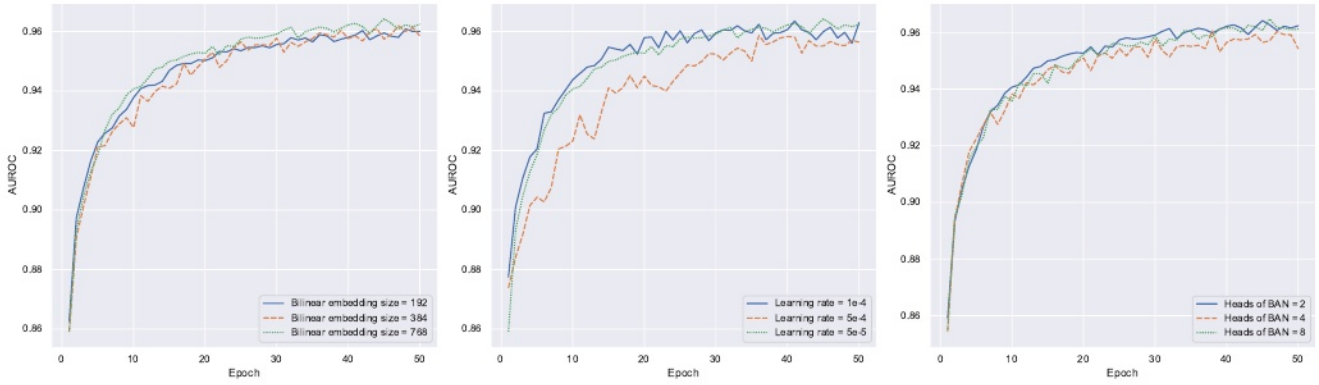

图 1. 在 BindingDB 验证集上采用不同超参数选择时的学习曲线。

4 不同蛋白质家族间的性能比较

我们开展实验以研究 DrugBAN 在不同蛋白质家族中的性能表现。依照先前的研究[1,7]，我们选取了四个主要的蛋白质家族：酶、G 蛋白偶联受体（GPCRs）、离子通道和核激素受体（NHRs）。我们分别从 BindingDB 和 BioSNAP 中随机抽取一个同域测试集，并使用 GtoPdb 数据库（https://www.guidetopharmacology.org/targets.jsp）将其中的蛋白质映射到这四个蛋白质家族。表 5 展示了测试集中每个蛋白质家族的相互作用数量。图 2 显示了在不同蛋白质家族下性能（AUROC 和 AUPRC）仅有轻微变化。

5 对未知药物/靶点的性能比较

为了研究 DrugBAN 以及其他深度学习基准模型在未见过的药物/靶点上的表现，我们在 BindingDB 和 BioSNAP 上进行了额外的实验。对于每个数据集，我们随机选取 20% 的药物/靶点蛋白。然后，我们对与这些药物/靶点蛋白相关的所有药物靶点相互作用（DTI）对进行预测性能评估（70% 作为测试集用于评估，30% 作为验证集用于确定提前停止），其余的对作为训练集用于模型优化。每个未见过的设置都有五次。

表 4.药物BAN超参数配置

模块	超参数	价值
优化器	学习率	5e-5
小批量	批量大小	64
三层卷积神经网络蛋白质编码器	初始氨基酸嵌入	128
	滤波器数量	[128, 128, 128]
	内核大小	[3, 6, 9] （此部分为数字列表，无需翻译）
三层 GCN 药物编码器	初始原子嵌入	128
	隐藏节点维度	[128, 128, 128]
双线性交互注意力	双线性注意力头	2
	双线性嵌入尺寸	768
	求和池化窗口大小	3
全连接解码器	隐藏神经元数量	512
判别器	隐藏神经元数量	256

表 5.测试集中主要蛋白质家族的相互作用数量。

数据集	# 酶	# G 蛋白偶联受体（GPCR）	# 离子通道	# 国家人权机构
BindingDB	5,277	472	440	144
BioSNAP	1,956	536	510	103

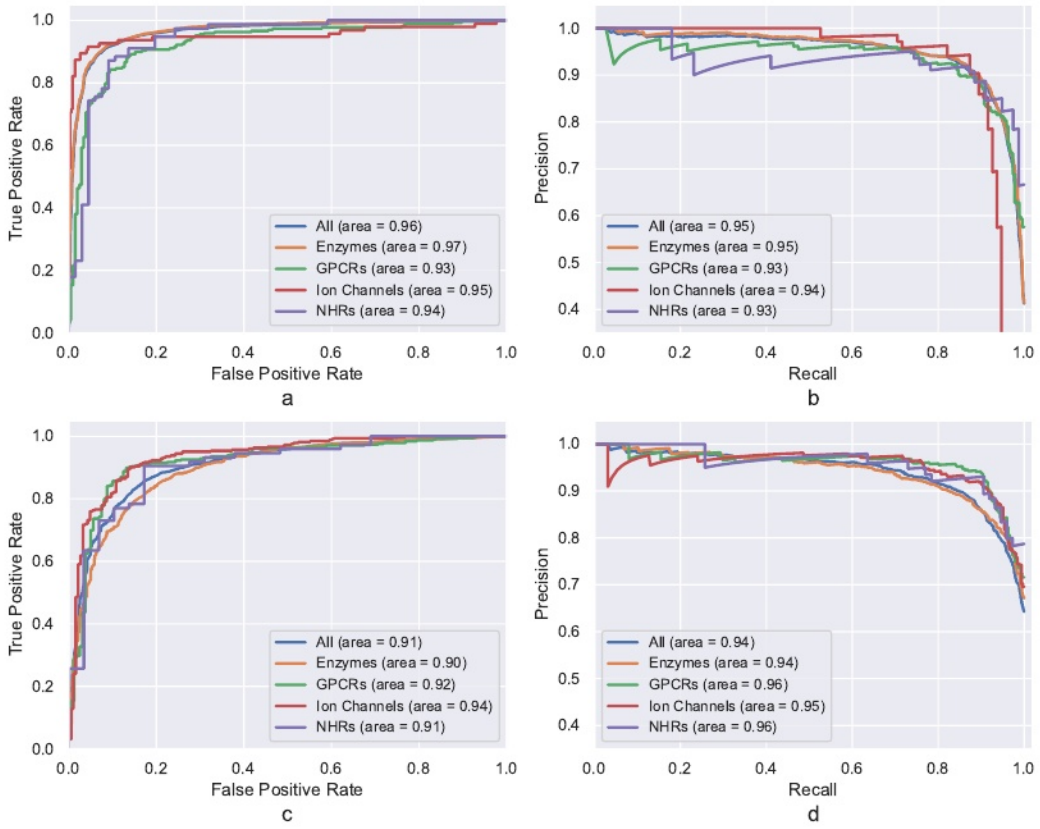

图 2.DrugBAN 在不同蛋白质家族上的**表现**。（**a**）在 BindingDB 数据集上的 AUROC 曲线。（**b**）在 BindingDB 数据集上的 AUPRC 曲线。（**c**）在 BioSNAP 数据集上的 AUROC 曲线。（**d**）在 BioSNAP 数据集上的 AUPRC 曲线。

表 6. 在 BindingDB 和 BioSNAP 数据集上采用随机划分、未见药物和未见靶点设置时的性能（五次随机运行的平均 AUROC）比较（最佳、次佳）。

设置	深度卷积药物-靶点相互作用预测模型[2]	DeepDTA[9]	MolTrans[1]	药物禁用令
		BindingDB		
随机拆分	0.945±0.002	0.951±0.002	0.952±0.002	**0.960±0.001**
未见药物	0.943±0.004	0.950±0.004	0.945±0.004	**0.959±0.002**
未见目标	0.627±0.070	0.670±0.023	0.661±0.037	**0.692±0.038**
		BioSNAP		
随机拆分	0.886±0.006	0.887±0.008	0.895±0.004	**0.903±0.005**
未见药物	0.856±0.005	0.858±0.007	0.856±0.008	**0.886±0.005**
未见目标	0.692±0.017	0.704±0.010	**0.714±0.014**	0.710±0.016

表 6 展示了在测试集上的 AUROC 结果，其中包括通常随机划分的结果以作对比。DrugBAN 在六种设置中的五种中表现最佳，而在 BioSNAP 的未见目标设置中，其表现也极具竞争力。

需要指出的是，在 BindingDB 上，所有方法在未见药物设置下的模型性能相较于随机划分设置仅略有下降。这是因为药物靶点相互作用数据集中存在大量高度相似的分子，而简单的未见药物设置无法区分它们。在我们之前的研究中，基于聚类的划分策略是更好的策略，能够缓解这一问题，从而形成更具挑战性的跨域任务。

6 缺失数据比例较高时的性能比较

表 7. 在 BindingDB 和 BioSNAP 数据集上具有高缺失数据比例情况下的性能比较（五次随机运行的平均 AUROC 值）（最佳、次佳）

缺失（%）	深度卷积药物-靶点相互作用预测模型[2]	DeepDTA[9]	MolTrans[1]	药物禁用令
		BindingDB		
95	0.773±0.005	0.831±0.002	0.846±0.004	**0.856±0.003**
90	0.840±0.002	0.867±0.002	0.874±0.003	**0.887±0.004**
80	0.877±0.002	0.897±0.003	0.905±0.001	**0.920±0.003**
70	0.890±0.005	0.916±0.002	0.923±0.001	**0.934±0.001**
		BioSNAP		
95	0.710±0.005	0.768±0.005	0.767±0.006	**0.770±0.008**
90	0.781±0.003	0.798±0.003	0.800±0.004	**0.802±0.003**
80	0.816±0.003	0.829±0.003	0.835±0.001	**0.836±0.002**
70	0.839±0.002	0.851±0.002	0.853±0.002	**0.860±0.003**

我们开展实验以阐明所提出的模型在 BindingDB 和 BioSNAP 数据集中缺失数据比例较高时的表现情况。依照 MolTrans1 中的缺失数据设置，我们仅使用每个数据集的 5%、10%、20% 和 30% 来训练 DrugBAN 和深度学习基线模型，并在剩余数据（90% 作为测试集，10% 作为验证集以确定提前停止时机）上评估预测性能。表 7 展示了所获得的结果，表明 DrugBAN 在所有设置下均表现最佳。尤其在较大的数据集（BindingDB）上，其改进效果更为显著。

7 可扩展性

我们从三个不同角度研究了 DrugBAN 的可扩展性：模型优化时间、数据加载时间和 GPU 内存使用情况。我们使用表 4 中的默认超参数配置，并采用单个 Nvidia V100 GPU 在 100 个周期内训练模型。图 3a 展示了针对来自 BindingDB 数据集的 4,919（10%）至 49,199（100%）个药物 - 靶点相互作用（DTI）对，模型优化时间和数据加载时间的变化情况。我们通过实验观察到，DrugBAN 的优化时间（红线）几乎与 DTI 对的数量呈线性增长关系。对于 49,199 个 DTI 对，完成优化大约需要两小时。数据加载过程（蓝线）所花费的时间比模型优化时间更长。不过，由于数据加载可以在 CPU 上进行，我们可以通过并行使用多个加载工作进程（子进程）来加速这一过程。图 3b 展示了数据加载时间随工作进程数量的变化情况，仅增加两个工作进程，加载时间就显著减少。

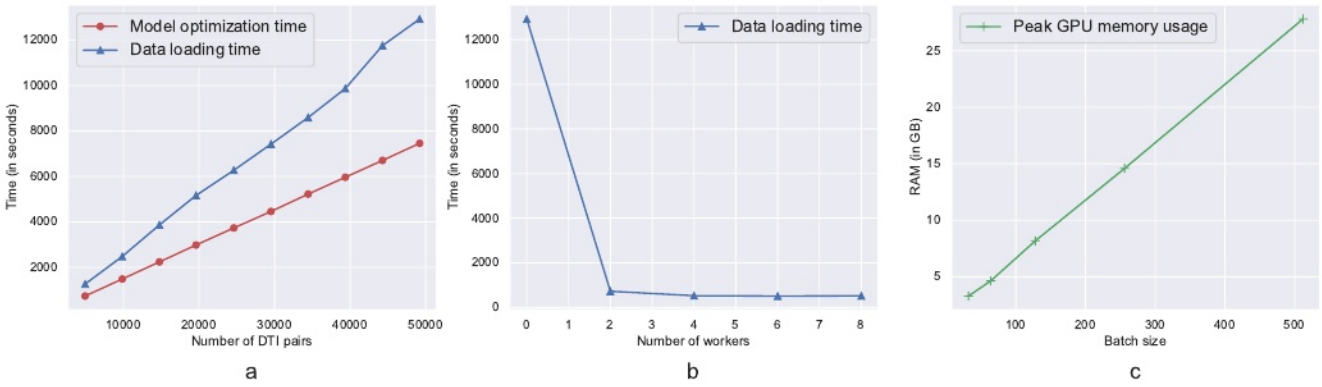

图 3.DrugBAN 在 BindingDB 数据集上的**可扩展性** （a）模型优化和数据加载时间几乎与药物靶点相互作用（DTI）对的数量呈线性增长。 （b）随着工作进程数量的增加，数据加载时间显著减少。 （c）峰值 GPU 内存使用量随批量大小呈线性增长。

工人们补充道。图 3c 展示了峰值 GPU 内存使用量与批处理大小的关系。我们发现，DrugBAN 在默认批处理大小为 64 时仅占用 4.63GB 内存，这非常高效。与优化时间类似，内存使用量也随批处理大小呈线性增长。这项研究证明了 DrugBAN 的可扩展性。

参考文献

1. Huang, K., Xiao, C., Glass, L. & Sun, J. MolTrans: Molecular interaction transformer for drug−target interaction prediction. *Bioinformatics* **37**, 830 −836 (2021).

2. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound−protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **31**, i221 −i229 (2015).

3. Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045−D1053 (2016).

4. Bai, P. *et al.* Hierarchical clustering split for low-bias evaluation of drug-target interaction prediction. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 641−644 (2021).

5. Gao, K. Y. *et al.* Interpretable drug target prediction using deep neural representation. In *IJCAI*, 3371−3377 (2018).

6. Wang, Z., Liang, L., Yin, Z. & Lin, J. Improving chemical similarity ensemble approach in target prediction. *Journal of cheminformatics* **8**, 1−10 (2016).

7. 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**, i232 −i240 (2008).

8. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology* **15** (2019).

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