基于半监督域适应的多级注意力网络药物-靶标预测

谢舟三[1]，图世奎1*，徐磊1,2*
[1]上海交通大学计算机科学与工程系，上海200240[2]广东省智能科学技术研究院，广东珠海519031
{waduhek, tusishikui, leixu}@sjtu.edu.cn

摘要

药物-靶标相互作用(DTI)的预测是药物发现的关键步骤，深度学习方法在各种DTI数据集上显示出巨大的前景。然而，现有方法仍然面临着几个挑战，包括标记数据有限、隐藏偏差问题以及对域外数据缺乏泛化能力。这些挑战阻碍了模型学习真正有信息量的交互特征的能力，导致捷径学习和对新颖药物-靶标对的预测性能较差。为了解决这些问题，我们提出了MlanDTI，一种用于DTI预测的半监督域自适应多层次注意网络(Mlan)。我们利用两个预训练BERT模型来获取双向表示，这些表示丰富了来自未标记数据的信息。然后，我们引入了一种多层注意机制，使模型能够在不同层次上学习领域不变的DTI。此外，我们提出了一种简单而有效的半监督伪标记方法，以进一步增强我们的模型在跨域场景中的预测能力。在四个数据集上的实验表明，MlanDTI在域内设置下的性能优于其他方法，在跨域设置下的性能优于所有其他方法。源代码可从https://github.com/CMACH508/MlanDTI获得。

介绍

药物发现和开发的过程具有高成本和时间密集型的特点。将一种一流的药物推向市场通常需要几十年的时间和高达数十亿美元的巨额投资。预测药物-靶标相互作用(DTIs)是药物发现和药物再利用的重要任务(Paul et al. 2010)，在生物医学领域具有重要价值(Agamah et al. 2020;Ezzat et al. 2019)。虽然高通量筛选、蛋白质组学和基因组学等传统技术仍然流行，但由于涉及的化学空间巨大，它们受到时间和成本的限制(Broach, Thorner et al. 1996;Bakheet和Doig 2009)。

为了加快药物发现过程并降低成本，已经开发了虚拟筛选(VS)技术来帮助硅(Rifaioglu et al. 2019)。分子对接和分子模拟在药物发现方面取得了巨大成功(Cheng et al. 2012)，但由于计算资源密集和依赖于3D结构数据的可用性，它们受到了限制。包括机器学习方法在内的方法(Faulon et al. 2008;Wang et al. 2021;Meng et al. 2017)在预测已知药物靶标对的DTI方面表现良好，而当应用于未知结构时，它们的性能往往会下降。

近年来，随着大量标记DTI数据的积累，许多端到端深度学习方法被用于预测DTI。从输入数据的角度来看，DTI预测模型可以分为三类。第一类是基于序列的模型，其中药物被表示为简化分子输入线输入系统(SMILES)或扩展连接指纹(ECFP)，蛋白质被视为氨基酸序列。这些模型通常使用1D-CNN (Ozt¨urk, ¨Ozg¨ur, ¨and Ozkirimli 2018;Lee、Keum和Nam 2019;赵等2022;Bai et al. 2023)或变压器架构(Chen et al. 2020;Huang et al. 2022)。其次，药物分子可以表示为图形(Nguyen et al. 2021; Tsubaki、Tomii和Sese 2019;Huang et al. 2022)或图像(Qian, Wu, and Zhang 2022)。同样，蛋白质距离图可以作为其3D结构信息的2D抽象(Zheng et al. 2020)，从而可以使用图神经网络(GNNs) (Scarselli et al. 2008)、图卷积网络(GCNs) (Kipf and Welling 2016)和卷积神经网络(cnn)。第三，与直接使用整个3D数据作为输入(Wallach, Dzamba, and Heifets 2015)相比，结合3D结构数据(如蛋白质口袋(Yazdani-Jahromi et al. 2022)或分子动力学模拟数据(Wu et al. 2022)无疑提高了模型性能并降低了计算复杂度;Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2018)。尽管如此，他们仍然受到3D结构数据可用性有限的限制。

尽管有这些显著的发展，深度学习方法仍然面临一些挑战。第一个挑战是有限的标记数据的限制。之前的工作主要集中在利用可用的标记数据，并学习数千个标记药物-靶标对(Ozt¨urk, ¨Ozg¨ur, and Ozkirimli 2018;Lee、Keum和Nam 2019;椿、富美、Sese 2019;阮等人。

*Corresponding authors

2021;黄等2022;赵等2022;Bai et al. 2023)。然而，这些方法往往忽略了大量未标记的生物医学数据，这阻碍了模型充分利用药物和蛋白质的化学结构和相互作用。因此，模型很难提取真正有信息量的特征，导致泛化能力有限。

第二个挑战是隐藏的偏差和捷径学习。在DUD-E和MUV数据集(Sieg, Flachsenberg和Rarey)上已经报道了隐藏偏差的问题

2019)。已经观察到，在DUD-E数据集(Chen等人，2019)和其他数据集(Chen等人，2019)上训练的模型。2020)，在进行预测时往往主要依赖于药物模式，而不是捕捉药物和靶标之间的综合相互作用。这导致了理论建模和实际应用之间的差距。我们进一步确定了这一问题的两个主要原因:1)数据集中药物分子的种类和数量比蛋白质更大;2)与蛋白质相比，药物分子的特征提取具有固有的便利性。这些因素导致了捷径学习(shortcut learning)，即模型往往优先从更大、更容易学习的药物分子数据中学习特征，而不是专注于蛋白质的特征。因此，模型很难有效地捕捉药物和蛋白质之间的相互作用特征。

第三个挑战在于模型对域外数据的泛化和预测能力，这与前两个挑战密切相关。开发一流药物通常涉及到使用新化合物预测与全新靶点的相互作用，这些化合物的分布可能与模型所训练的数据有显著差异。因此，模型需要具备跨域泛化能力(Abbasi et al. 2020;Bai等2023;Kao et al. 2021)。目前，大多数模型都是在有限的标记数据上进行训练的，无法解决捷径学习的问题，导致预测全新药物和蛋白质之间相互作用的能力有限。

为了解决这三个挑战，我们提出了一种用于DTI预测的半监督域自适应多级关注网络MlanDTI。我们利用两个预训练的BERT模型从数百万未标记的数据中获得蛋白质和SMILES(药物)序列的双向嵌入。受最小均方误差重构(Lmser)网络的启发(徐 1993;Huang, Tu, and徐2022)，然后我们设计了一种具有多层次注意机制的变压器，以药物和蛋白质嵌入作为输入。它能够以减少隐藏偏差的方式联合提取药物和靶标特征，并促进多层次相互作用的学习。此外，我们融入了一种简单而有效的半监督伪标记方法，以进一步增强我们的模型在跨域场景中的预测能力。在四个数据集上的实验表明，MlanDTI在域内设置下的性能优于其他方法，在跨域设置下的性能优于所有其他方法。

主要贡献如下三方面:

· 为了利用大量未标记的生物医学数据，我们采用了两个预训练的BERT模型来获取代表性具有更好的鲁棒性和泛化能力的语句。我们观察到，由BERT模型获得的表征显著提高了伪标记的准确性。

· 我们提出了一种新的多级注意力机制，通过允许模型在学习过程中动态地关注蛋白质和药物的不同方面，从而实现有效的特征提取。注意力机制缓解了捷径学习问题，并减少了隐藏偏差对预测的影响。

· 我们提出了一种简单而有效的伪标签域适应方法，显著降低了伪标签的噪声。

相关工作

利用额外的数据

DTI预测的关键之一是如何表示药物分子和蛋白质，使模型能够学习有用的特征。从3D结构信息中学习(Wallach, Dzamba, and Heifets 2015;Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2018)无疑是最直接的方法，但受限于高计算成本和模型复杂度。另一种间接方法是提供包含3D结构信息的额外数据，例如分子动力学模拟(Wu等人，2022)和蛋白质口袋数据(Yazdani-Jahromi等人，2022)。虽然上述方法受到有限数量3D结构数据可用性的限制，但相比之下，Moltrans (Huang等人，2021)利用大量未标记的蛋白质和药物序列，通过使用频繁连续子序列(FCS)算法提取高质量的子结构，并使用变压器增强表征。然而，FCS在从序列数据中全面提取信息的能力上存在一定的局限性，未标记数据的利用数量也不足。在本文中，我们利用在大量未标记数据上学习的两个预训练BERT模型(Devlin et al. 2018)来获得具有强大泛化能力的蛋白质和药物序列的丰富表示。

学习交流

蛋白质和药物是两种根本不同类型的数据，DTI预测的任务要求模型学习它们的相互作用特征。最简单的方法是将特征连接起来(Ozt¨urk, ¨Ozg¨ur, ¨and Ozkirimli 2018;Lee、Keum和Nam 2019;郑等人2020;Nguyen et al. 2021)，并通过全连接网络(FCN)传递它们以获得预测结果。另一种方法(Qian, Wu, and Zhang 2022)是将特征映射重叠，并使用CNN提取交互特征。然而，这些方法缺乏可解释性，忽略了交互的内在结构。最近，注意力机制已经被证明在捕获蛋白质和药物之间复杂的相互作用方面是有效的。多头注意力(Bian et al. 2023;Chen等人，2020)和其他注意力变化(Bai等人，2023;Zhao et al. 2022)在DTI预测中得到了广泛的应用。然而，(Chen等人。2020)发现，在一些数据集中隐藏的偏见

Led模型主要依赖于药物模式而不是相互作用进行预测。我们进一步观察到，这个问题在现有的模型中普遍存在。为了解决这个问题，我们提出了一种多级注意力机制。

DTI预测中的领域泛化

在之前的作品中(Huang等人，2021;Yazdani-Jahromi等人，2022;Zhao et al. 2022)，模型泛化性的评估往往是通过将数据集划分为“未见过的药物”或“未见过的蛋白质”场景进行的，其中药物或蛋白质只存在于测试集中。然而，这样的评估仍然属于域内设置，不同于现实世界的应用。目前，DTI预测领域泛化的研究比较有限。DrugBAN (Bai等人，2023)利用条件域对抗网络(CDAN)将学习到的知识从源域转移到目标域，从而提高了模型在跨域设置中的性能，从而解决了这一挑战。在这里，我们利用伪标记技术来缓解目标域和源域之间的分布差异。通过集成辅助分类器和BERT模型强大的表示能力，我们的方法显著提高了伪标记的准确性。在跨域设置下，我们的方法表现出优于DrugBAN的显著性能。

方法

问题公式化

DTI预测的任务是确定药物化合物和靶蛋白是否会相互作用。对于药物化合物，大多数现有的深度学习方法使用SMILES字符串来表示药物。具体来说，一种药物表示为D = (D_1，…，d_m)，其中d_i是SMILES符号，具有原子等化学含义，m是长度。至于目标蛋白质，每个蛋白质序列表示为T = (a_1，…，a_n)，其中a_i对应23种氨基酸中的一种，n是蛋白质序列的长度。

给定一个药物SMILES序列D和一个蛋白质序列T，目标是训练一个模型，通过映射联合特征表示空间D × T来分配一个相互作用概率得分P∈[0,1]。

所提出的框架

图1描述了MlanDTI的概述。首先，通过预先训练的BERT模型，即ChemBERTa-2 (Ahmad et al. 2022)和ProtTrans (Elnaggar et al. 2021)，将药物和靶标序列编码到载体嵌入中。随后，这些嵌入通过具有多级注意力模块的改进transformer架构的编码器和解码器进行传递，以提取交互特征。该分类器包括双线性关注模块和最大池化层，然后是用于预测的FCN。对于跨域预测，我们使用直接接受BERT输出的辅助分类器。它有助于从BERT表示中学习隐式分布信息，从而提高伪标签的准确性。在标记源域数据上训练两个分类器后，模型在未标记的目标域数据上进行预测，从而获得伪标签。伪标签学习过程包括学习高置信度的伪标签和最小化冲突预测。

蛋白质序列的编码器我们通过采用类似于TransformerCPI的转换器的修改来构建编码器(Chen et al. 2020)。我们没有使用自注意力模块，而是使用1D-CNN和GLU(门控线性单元)(Dauphin et al. 2017)作为替代。隐藏层h_0，…，h_L在编码器中计算为:

$$h_i(\mathbf{X}_T) = (\mathbf{X}_T\mathbf{W}_{i1} + \mathbf{s}) \otimes \sigma(\mathbf{X}_T\mathbf{W}_{i2} + \mathbf{t}), \quad (1)$$

其中XT∈Rn×m^1为第hi层的输入，Wi1∈$Rk_{\times m1\times m2}$, s∈Rm2, Wi2∈$Rk_{\times m1\times m2}$, t∈Rm2为参数，n为输入序列长度，k为patch大小，m_1，m_2为输入和隐藏向量的维数，σ为sigmoid函数，⊗为元素积。

由于蛋白质序列的长度可能在数千甚至数万的范围内，transformer中的self-attention模块以O(n^2)的时间和空间复杂度构成了显著的计算和记忆负担，并且在处理小数据集时容易出现过拟合。Eq.(1)的上述修改减轻了长蛋白质序列的计算和存储负担，并补救了小数据集上的过拟合。

多水平交叉注意对于DTI预测任务，模型最关键的能力是学习药物与靶点之间的相互作用模式。它涉及到在共享特征子空间中将蛋白质的特征与药物的特征对齐。然而，从蛋白质中提取多层次特征比从药物中提取特征更具挑战性，因为蛋白质序列明显很长，具有复杂的多层次结构，而药物往往是化学小分子。这种差异导致了DTI模型中的隐藏偏差(另一个是固有的数据集偏差)。将蛋白质特征与药物特征对齐也需要一个多层次的过程，但模型可能无法很好地捕获蛋白质的多层次特征，并有效地将其与药物特征对齐。因此，现有的模型倾向于通过依赖药物分子的特征来学习一条捷径来预测药物-靶标相互作用。

在早期文献(Xu 1993)中，提出Lmser网络通过在编码器和解码器之间的每个层上建立双向跳过连接来增强表示学习。在深度CNN实现中首次证明了它在图像处理方面的鲁棒性和有效性(Huang, Tu, and Xu 2022;Xu 2019)，然后开发了Lmser-transformer，通过在原始变压器上添加分层连接来改进分子表示学习(Qian et al. 2022)。在这些工作的启发下，我们提出了一种多层次的交叉注意力机制来解决这个问题，如图1(a)所示。在vanilla transformer中，编码器使用编码器最后一层的蛋白质特征作为解码器交叉注意力层的键和值，将它们与解码器中的药物特征对齐。然而，从编码器的输出中获得的蛋白质特征并不能完全捕捉到多层次结构的表达

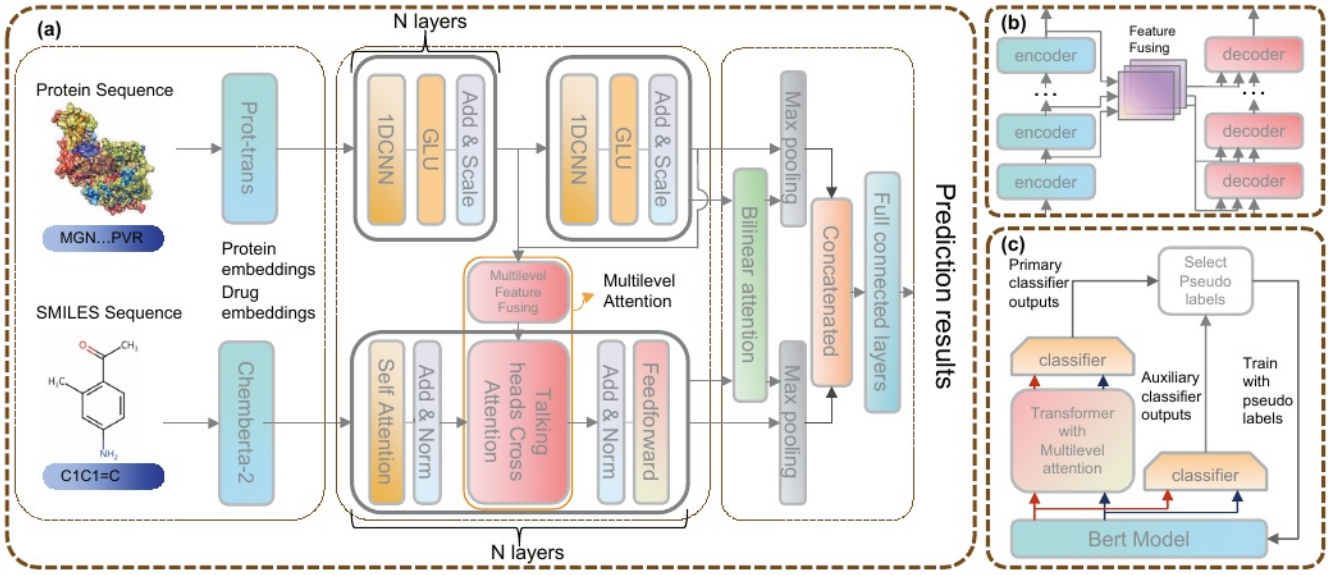

图1:(a) MlanDTI的总体框架，它由两个预训练的BERT模型组成，将SMILES和氨基酸序列转换为向量嵌入。编码器和解码器通过多级注意模块连接，最终输出通过具有双线性注意模块和最大池化层的分类器进行处理，然后送入FCN生成预测结果。(b)多水平关注的细节。(c).使用辅助分类器进行伪标注训练。

蛋白质的信息，它们在解码器中不与不同层次的药物特征对齐。

本文通过两个步骤开发多级注意力机制:1)多级特征融合步骤和2)交叉注意力特征对齐步骤。假设每个编码器层的蛋白质特征矩阵是T_0，…，$T_n \in R^{m\times d}$，其中n为transformer层数，m为蛋白质序列大小，d为向量维度。对于ℓ-th解码器层，我们将前面ℓ层的蛋白质特征矩阵串联起来，形成T_{cat_ℓ}=[T0，…， Tℓ]∈Rℓ×m×d。然后，我们通过应用融合矩阵$F_\ell \in R^{\ell\times 1}$进行跨层特征聚合。这就得到了多层次的融合蛋白质特征矩阵Tℓ' = Fℓ T Tcatℓ。总而言之，我们将所有Tℓ'计算为:

$$\mathrm{diag}(T'_0, ..., T'_n) = \mathbf{F} \cdot \mathrm{diag}(T_{cat_0}, ..., T_{cat_n}), \quad (2)$$

其中F是一个可学习的对角矩阵，每个对角元素都是F_ℓ from每一层，即ℓ= 0，…， n.然后，ℓ-th层的多级交叉注意力机制的query、key和value分别计算为

$$Q = D_\ell W_q, \quad K = T'_\ell W_k, \quad V = T'_\ell W_v, \quad (3)$$

式中$D_\ell$is为通过自关注模块的药物特征矩阵，T_ℓ'the为Eq. (2)给出的多层次蛋白质特征矩阵。

为了增强注意力头对多层次交互的提取能力，我们纳入了talking-heads注意力机制(Shazeer et al. 2020)进行特征对齐。transformer中多头注意力的这种变化引入了两个额外的线性投影。这些投影分别转换了注意力logits和注意力权重，允许信息在不同的注意力头之间流动，并提高了模型的整体性能，即:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(P_\ell \frac{QK^T}{\sqrt{d_k}}\right) P_w V, \quad (4)$$

其中Q, K, V由Eq.(3)给出，$P_\ell \in R^{h^k\times h^k}$, $Pw \in R^{h^k\times h^v}$是两个附加的线性投影。h_k表示键和查询的注意力头数量，h_v表示值的注意力头数量，它们可以选择性地在大小上有所不同。

下面简要总结了所提出的多层次注意力机制的优点。

- 鼓励多层次特征学习:通过融合蛋白质特征，衍生出药物特征与相关特征相互作用，从而捕获多级相互作用特征，从而更全面地理解药物-靶标相互作用。
- 缓解隐藏偏差和减少过拟合:多级注意力鼓励模型更多地关注分层交互特征，模型变得不太容易出现只关注特定模式可能出现的偏倚表示，因此模型不太可能过度拟合数据的噪声模式。
- 提高泛化能力:多级注意力使模型能够学习域不变的交互特征。这些表示表现出鲁棒性并增强了跨不同数据域的可迁移性。

分类器分类器由hyperattentionDTI的双线性注意力模块组成(Zhao et al. 2022)

来进一步提取双向交互特征。随后，我们利用每一层都有一个多层FCN，后面跟着一个leaky ReLU激活函数(He et al. 2015)来组合这些特征并生成预测结果。由于这是一个二分类问题，我们利用二值交叉熵损失函数来训练模型。

$$\mathcal{L}_{CE} = -[y \log \hat{y} + (1-y)\log(1-\hat{y})], \quad (5)$$

其中y是ground truth标签，y☒是分类器的输出。

伪标签(Lee et al. 2013)是一种半监督学习(SSL)方法，它利用在标记的源领域数据上训练的模型为未标记的目标领域数据生成伪标签。通过将这些伪标签合并到训练过程中，该模型可以自适应目标域，这特别适用于标记数据有限而未标记数据大量的DTI预测。然而，在其他领域，基于伪标签的SSL方法往往由于存在噪声伪标签而导致模型性能不佳(Rizve et al. 2021)。在这里，我们提出了一种简单而有效的方法，可以显著提高生成标签的准确性，并减少噪声标签对模型的影响。

我们的方法由两个步骤组成。第一步，我们执行高置信度伪标签的选择和学习。为了实现这一点，我们引入了一个用于协同训练的辅助分类器，它本质上是前面提到的分类器，但直接将BERT表示作为输入。设P1 = {P1 (i)} Ni=1, P0 = {P0}i=1, P1,aux = (i) N {p1,aux(i)}Ni=1, P0,aux ={P0,aux (i)}i=1 Nbe模型和辅助分类器对目标域数据Xt ={x(i)}Ni=1的概率输出，其中P0 (i)， P0,aux (i)为样本不相互作用的概率$x^{(i)}$，p1 (i)， p1, aux(i)为样本相互作用的概率。我们没有选择阈值，因为我们观察到阈值可能导致伪标签不平衡，而是对(P1 + P1,aux)和(P0+P0,aux)进行降序排序，并根据概率选择前M个正、负样本对来分配伪标签:

$$Y_1 = \{\hat{y}_1^{(i)} = 1 | p_1^{(i)} + p_{1,aux}^{(i)} \in \text{top}_M(\mathbf{P}_1 + \mathbf{P}_{1,aux})\}, \quad (6)$$

Y0 = {y☒(i)0 = 0|p(i)0 + p0,aux (i)∈topM (p0 + p0,aux)}，(7)

其中，y_1，y_0分别表示正样本和负样本的伪标签集合，M是所选样本的数量，随着迭代次数的增加而增长。

辅助分类器专注于学习BERT表示中目标域和源域数据之间的潜在关系，而主模型优先学习领域不变的DT交互特征。这导致了本质上的分类器差异，使两个分类器上具有高置信度的伪标签具有更高的准确率。在生成伪标签后，我们使用交叉熵损失来训练模型，即:

$$\mathcal{L}_{pseudo} = -\frac{1}{2M}\sum_{i=1}^{2M}[y \log \hat{y}^{(i)} + (1-y)\log(1-\hat{y}^{(i)})]. \quad (8)$$

第二步是对冲突预测进行惩罚。设X_d为两个分类器表现出冲突分类的样本集，即:

$$X_d = \{x^{(i)} | x^{(i)} \in X_t, \text{argmax}\,\mathbf{p}^{(i)} \neq \text{argmax}\,\mathbf{p}_{aux}^{(i)}\}, \quad (9)$$

p (i) = (p (i) 0, p (i) 1), paux (i) = (p0,辅助(我),p1,辅助(我))。

我们从X_d中随机选择一个大小为M'的子集$X_{d'}$，其中M'的值随着模型迭代次数的增加而增加。我们利用修正的二元交叉熵损失来增加两个分类器之间冲突样本的预测不确定性，即:

$$\mathcal{L}_{conf} = -\frac{1}{M'}\sum_{i=1}^{M'}[y \log 0.5 + (1-y)\log(1-0.5)]. \quad (10)$$

这两个步骤都使模型能够获得噪声减少的伪标签进行训练，从而增强了模型在目标域的性能。

实验

数据集

我们在人类数据集、秀丽隐杆线虫数据集(Tsubaki, Tomii, and Sese 2019)、binding数据库数据集(Liu et al. 2007)和Biosnap数据集(Huang et al. 2021)上评估了我们的模型。具体来说，我们对BindingDB和Biosnap数据集进行了域内和跨域测试。对于域内评估，我们将数据集随机分为训练集、验证集和测试集，在较小的人类和秀丽隐杆线虫数据集中比例为8:1:1，在较大的BindingDB和Biosnap数据集中比例为7:1:2。我们还对BindingDB和Biosnap数据集进行了冷对分割实验。我们随机选择70%的药物/蛋白质，收集所有相关的DT对作为训练集。随后，将剩余30%的DT对分成3:7的比例，作为验证集和测试集。这确保了测试集中的所有药物和蛋白质都是无法建模的。

对于跨域评估，我们遵循了DrugBAN中使用的基于聚类的分裂策略。我们应用ECFP4和PSC算法分别对药物和蛋白质进行聚类。然后，我们随机选择了60%的药物和蛋白质簇，并使用属于这些簇的所有药物-蛋白质对作为源域数据。剩余40%的聚类中的药物-蛋白质对作为靶域数据。这种数据划分确保了目标域和源域数据来自不相交的分布，使评估更具挑战性，并能够真实评估模型预测未知蛋白质和分子相互作用的能力。

对于域适应设置，我们使用所有标记的源域数据和80%的未标记的目标域数据作为训练集。这80%的目标域数据也被用作验证集，而来自目标域的剩余20%的带标签数据作为测试集。

基线和实现细节

我们将我们提出的方法与8种基线方法进行了比较:支持向量机

methods	human			C.elegans			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
SVM	0.910	–	**0.967**	0.894	–	0.801	0.939	0.928	0.787	0.862	0.864	0.762
RF	0.940	–	0.878	0.902	–	0.832	0.942	0.921	0.858	0.860	0.886	0.808
GraphDTA	0.960	0.959	0.897	0.974	0.975	0.919	0.951	0.934	0.867	0.887	0.890	0.789
DeepConvDTI	0.967	0.964	0.922	0.983	0.985	0.944	0.945	0.925	0.859	0.886	0.890	0.797
MolTrans	0.974	0.976	0.944	0.982	0.985	**0.966**	0.952	0.936	0.865	0.895	0.897	0.824
TransformerCPI	0.973	0.975	0.920	0.988	0.986	0.952	0.943	0.925	0.855	0.889	0.893	0.798
HyperAttDTI	0.984	0.984	0.946	0.989	0.990	0.958	**0.959**	**0.948**	**0.887**	0.901	0.902	0.838
DrugBAN	0.981	0.983	0.940	0.986	0.988	0.949	**0.959**	0.947	0.881	0.903	0.902	0.832
Ours	**0.988**	**0.990**	0.961	**0.990**	**0.992**	0.962	0.945	0.926	0.857	**0.909**	**0.912**	**0.841**

表1:提出的模型和基线在4个数据集(5个随机运行)上的结果，Metric: AUROC (AUC)， AUPRC (AUPR)， F1-score (F1)，最好的结果用粗体表示。“-”表示这个指标没有结果。

methods	cold						cross-domain					
	BindingDB			BioSNAP			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
Moltrans	0.595	0.522	0.511	0.672	0.697	0.437	0.537	0.476	0.389	0.632	0.635	0.401
TransformerCPI	0.656	0.594	0.566	0.680	0.708	0.523	0.568	0.450	0.410	0.656	0.693	0.432
HyperAttDTI	0.661	0.598	0.582	0.732	0.760	0.539	0.545	0.462	0.376	0.654	0.685	0.395
DrugBAN	0.655	**0.600**	0.542	0.651	0.667	0.449	0.578	0.471	0.484	0.608	0.606	0.438
DrugBAN$_{CDAN}$	NA	NA	NA	NA	NA	NA	0.616	0.512	0.426	0.673	0.706	0.542
Ours	**0.671**	0.594	**0.601**	**0.782**	**0.801**	**0.653**	0.657	0.537	0.489	0.728	0.759	0.604
Ours (with PL)	NA	NA	NA	NA	NA	NA	**0.687**	**0.579**	**0.564**	**0.749**	**0.770**	**0.629**

表2:BindingDB和BioSNAP数据集的域内(冷对分裂:未见的药物和蛋白质)和跨域(基于聚类的分裂)比较(随机运行5次)。1)下划线值解释:我们选择0.5的阈值(与MolTrans相同)来计算DrugBAN的f1分数。这是为了保证公平的比较，避免DrugBAN中阈值过低导致分类无效。更多信息见附录。2) NA，不适用于本研究。3)括号内的“with PL”是指我们的方法包含了伪标签模块。

(SVM) (Cortes and Vapnik 1995)、随机森林(RF) (Ho 1995)、GraphDTA (Nguyen et al. 2021)、deepconvn - dti (Lee, Keum, and Nam 2019)、MolTrans (Huang et al. 2021)、TransformerCPI (Chen et al. 2020)、HyperattentionDTI (Zhao et al. 2022)和DrugBAN (Bai et al. 2023)。这些基线包含了经典的机器学习方法和当前最先进的深度学习方法，确保了全面的比较。所有深度学习方法都被采用，其默认配置由各自的作者提供。我们提出的方法在PyTorch中实现，利用初始学习率为0.001的Adam优化器。附录中提供了详细的超参数设置。

域间的实验

表1显示了人类和秀丽隐杆线虫数据集的比较。这两个数据集相对较小，正负样本平衡，使我们能够在相同的分布内评估模型的预测能力。我们的方法在AUROC和AUPRC方面优于所有深度学习基线，并且在f1分数方面也表现出竞争力。

我们还对较大的数据集BindingDB和BioSNAP进行了比较。在随机分割测试中，我们的模型在BioS-NAP数据集上实现了最先进的性能，但它在BindingDB数据集上的性能并不是特别具有竞争力。这种差异是由于BindingDB数据集中存在隐藏的偏差问题。

BindingDB数据集包含14643种药物和2623种蛋白质，与其他数据集(BioSNAP: 4510 / 2181，人类:2726 / 2001，秀丽隐杆线虫:1767 / 1876)相比，这导致药物与蛋白质的比例极度不平衡。与其他三个数据集相比，在BindingDB数据集上，深度学习模型甚至难以超越传统的机器学习方法(AUC: RF 0.942, deepconvo - dti 0.945)。先前的研究(Bai等人，2023)也报道了与随机分割相比，未见药物设置下BindingDB数据集的性能下降最小。这一现象归因于数据集中存在大量高度相似的分子，这使得天真的未见药物设置难以区分它们。高度相似的药物样本数量过多，导致基线模型倾向于学习药物模式，而不是药物-靶标相互作用进行预测。因此，深度学习和机器学习方法表现出相似的性能水平。然而，这种捷径学习方法违背了DTI预测的初衷，在实际应用中不能被认为是可靠的。

然而，我们的模型更侧重于学习蛋白质和药物之间的多级相互作用。在表2的冷分裂设置中，模型只能学习药物-靶标

Ablation	BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1
	0.687	0.579	0.564	0.749	0.770	0.629
-BERT	0.573	0.455	0.413	0.648	0.671	0.499
-MLA	0.628	0.511	0.523	0.731	0.753	0.585
-PL	0.657	0.537	0.489	0.728	0.759	0.604
-Aux Cls	0.626	0.486	0.503	0.739	0.776	0.633

表3:BindingDB和BioSNAP数据集的消融研究(跨域，随机5次运行)

相互作用特征，由于缺乏足够相似的药物和蛋白质分子作为参考。我们的模型在BindingDB数据集上优于其他基线，而在更平衡的BioSNAP数据集上，我们的模型与基线相比取得了更好的性能。

总的来说，BindingDB数据集上隐藏的偏差问题所带来的挑战突出了我们的模型捕获多层次药物-靶标相互作用的能力的重要性，这使得它在其他基线难以保持有效性的情况下表现良好。

跨域实验

表2给出了跨域设置下BindingDB和BioSNAP数据集上的模型性能比较。与域内设置相比，由于数据分布的差异，大多数模型都经历了性能的显著下降。特别是，对于BindingDB数据集，基于聚类的策略确保在训练集和测试集之间没有相似的药物或蛋白质，从而防止模型依赖于药物模式。这打破了在域内场景中观察到的虚假的高性能错觉，一些模型甚至没有显示出比随机猜测(AUC: 0.5)更好的性能。在所有基线中，利用条件域对抗网络(CDAN)进行域适应的DrugBAN$_{CDAN}$取得了最好的性能。然而，DrugBAN$_{CDAN}$并没有超过我们的无伪标记的vanilla模型，而我们的有伪标记的模型明显优于所有最先进的模型，包括带域适应模块的DrugBAN。具体来说，我们的模型在BindingDB和BioS-NAP数据集上分别优于DrugBAN$_{CDAN}$11.52% 和 11.29% (AUROC)。

消融研究

我们在BindingDB和BioSNAP数据集的跨域设置下进行了表3中的消融研究，以分析我们提出的模型中模块的有效性。

我们用TransformerCPI (Chen et al. 2020)中使用的Word2Vec和GCN代替BERT，以获得药物和蛋白质的嵌入。如表3所示，模型的性能经历了显著的下降。这一结果可以归因于辅助分类器无法通过表示有效地捕获源域和目标域之间的隐式关系。因此，在图2中，伪标签的准确率表现出显著下降，引入了a

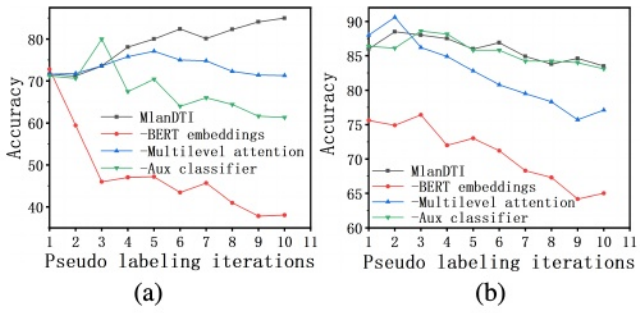

图2:在(a) BindingDB数据集(b) BioSNAP数据集上伪标记精度的消融实验

大量的噪声伪标签，恶化了模型的性能。

多级注意力的有效性我们用原始的Transformer多头注意力替换了多级注意力(MLA)机制。然而，在两个数据集上，该模型都表现出不同程度的性能下降。随着训练迭代次数的增加，观察到伪标签的准确率显著下降。事实证明，多级注意力机制能够更好地捕捉域不变的药物-靶标相互作用特征，从而增强模型在目标域的表现。

伪标记和辅助分类器的有效性伪标记(PL)被证明可以有效地提高模型在目标域内的性能。同时，辅助分类器有助于减少这些伪标签内的噪声。这种影响在BindingDB数据集中尤其明显，它在域分布中显示出很大的差异。辅助分类器的缺失加剧了伪标签中存在的噪声，导致伪标签方法在增强模型性能方面的不足。

结论

在本文中，我们提出了MlanDTI，这是一种半监督域自适应多层次注意网络，它利用大量未标记数据从预训练的BERT模型中获得丰富的药物和蛋白质的双向表示。此外，我们引入了多级注意力机制来捕获不同层次和深度的蛋白质和药物之间的域不变的相互作用特征。最后，我们加入了一种简单而有效的伪标记方法，以进一步增强我们模型的泛化能力。我们的模型展示了优秀的域泛化能力，使其非常适合预测药物开发中新药和靶点之间的相互作用。通过与最先进的模型进行全面比较，建立了比之前方法的实质性性能优势。

致谢

本研究得到国家自然科学基金项目(62172273)和上海市科技重大专项(2021SHZDZX0102)资助。涂世奎、徐磊为共同通讯作者。

参考文献

Abbasi, K.; Razzaghi, P.; Poso, A.; Amanlou, M.; Ghasemi, J. B.; and Masoudi-Nejad, A. 2020. DeepCDA: deep cross-domain compound−protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, 36(17): 4633−4642.

Agamah, F. E.; Mazandu, G. K.; Hassan, R.; Bope, C. D.; Thomford, N. E.; Ghansah, A.; and Chimusa, E. R. 2020. Computational/in silico methods in drug target and lead prediction. *Briefings in bioinformatics*, 21(5): 1663−1675.

Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; and Ramsundar, B. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Bai, P.; Miljkovi´c, F.; John, B.; and Lu, H. 2023. Interpretable bilinear attention network with domain adaptation improves drug−target prediction. *Nature Machine Intelli-gence*, 5(2): 126−136.

Bakheet, T. M.; and Doig, A. J. 2009. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4): 451−457.

Bian, J.; Zhang, X.; Zhang, X.; Xu, D.; and Wang, G. 2023. MCANet: shared-weight-based MultiheadCrossAt-tention network for drug−target interaction prediction. *Brief-ings in Bioinformatics*, 24(2): bbad082.

Broach, J. R.; Thorner, J.; et al. 1996. High-throughput screening for drug discovery. *Nature*, 384(6604): 14−16.

Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; and Kurtzman, T. 2019. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, 14(8): e0220113.

Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; and Zheng, M. 2020. TransformerCPI: improving compound−protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36 (16): 4406−4414.

Cheng, F.; Zhou, Y.; Li, J.; Li, W.; Liu, G.; and Tang, Y. 2012. Prediction of chemical−protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*, 8(9): 2373−2384.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273−297.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, 933−941.

PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112−7127.

Ezzat, A.; Wu, M.; Li, X.-L.; and Kwoh, C.-K. 2019. Computational prediction of drug−target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, 20(4): 1337−1357.

Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; and Sapra, R. 2008. Genome scale enzyme−metabolite and drug−target interaction predictions using the signature molecular descrip-tor. *Bioinformatics*, 24(2): 225−233.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on im-agenet classification. In *Proceedings of the IEEE interna-tional conference on computer vision*, 1026−1034.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278−282. IEEE.

Huang, K.; Xiao, C.; Glass, L. M.; and Sun, J. 2021. MolTrans: molecular interaction transformer for drug−target interaction prediction. *Bioinformatics*, 37(6): 830−836.

Huang, L.; Lin, J.; Liu, R.; Zheng, Z.; Meng, L.; Chen, X.; Li, X.; and Wong, K.-C. 2022. CoaDTI: multi-modal co-attention based framework for drug−target interaction anno-tation. *Briefings in Bioinformatics*, 23(6): bbac446.

Huang, W.; Tu, S.; and Xu, L. 2022. Deep CNN based Lmser and strengths of two built-in dualities. *Neural Processing Letters*, 54(5): 3565−3581.

Kao, P.-Y.; Kao, S.-M.; Huang, N.-L.; and Lin, Y.-C. 2021. Toward drug-target interaction prediction via en-semble modeling and transfer learning. In *2021 IEEE In-ternational Conference on Bioinformatics and Biomedicine (BIBM)*, 2384−2391. IEEE.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and effi-cient semi-supervised learning method for deep neural net-works. In *Workshop on challenges in representation learn-ing, ICML*, volume 3, 896. Atlanta.

Lee, I.; Keum, J.; and Nam, H. 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; and Gilson, M. K. 2007. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic acids research*, 35(suppl 1): D198−D201.

Meng, F.-R.; You, Z.-H.; Chen, X.; Zhou, Y.; and An, J.-Y. 2017. Prediction of drug−target interaction networks

from the integration of protein sequences and drug chemical structures. *Molecules*, 22(7): 1119.

Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; and Venkatesh, S. 2021. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8): 1140–1147.

Ozt¨urk, ¨ H.; Ozg¨ur, ¨ A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; and Schacht, A. L. 2010. How to improve R&D productivity: the pharmaceutical industry' s grand challenge. *Nature reviews Drug discov-ery*, 9(3): 203–214.

Qian, H.; Lin, C.; Zhao, D.; Tu, S.; and Xu, L. 2022. AlphaDrug: protein target specific de novo molecular generation. *PNAS nexus*, 1(4): pgac227.

Qian, Y.; Wu, J.; and Zhang, Q. 2022. CAT-CPI: Combining CNN and transformer to learn compound image features for predicting compound-protein interactions. *Frontiers in Molecular Biosciences*, 9: 963912.

Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; and Do□gan, T. 2019. Recent applications of deep learning and machine intelligence on in silico drug discov-ery: methods, tools and databases. *Briefings in bioinformat-ics*, 20(5): 1878–1912.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80. Shazeer, N.; Lan, Z.; Cheng, Y.; Ding, N.; and Hou, L. 2020. Talking-heads attention. *arXiv preprint arXiv:2003.02436*. Sieg, J.; Flachsenberg, F.; and Rarey, M. 2019. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *Journal of chemical in-formation and modeling*, 59(3): 947–961. 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(21): 3666–3674.

Tsubaki, M.; Tomii, K.; and Sese, J. 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2): 309–318.

Wallach, I.; Dzamba, M.; and Heifets, A. 2015. Atom-Net: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Wang, X.-r.; Cao, T.-t.; Jia, C. M.; Tian, X.-m.; and Wang, Y. 2021. Quantitative prediction model for affinity of drug– target interactions based on molecular vibrations and overall system of ligand-receptor. *BMC bioinformatics*, 22 (1): 1– 18.

Wu, F.; Jin, S.; Jiang, Y.; Jin, X.; Tang, B.; Niu, Z.; Liu, X.; Zhang, Q.; Zeng, X.; and Li, S. Z. 2022. Pre-Training of Equivariant Graph Matching Networks with Conformation Flexibility for Drug Binding. *Advanced Science*, 9 (33): 2203796.

Xu, L. 1993. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural networks*, 6 (5): 627–648.

Xu, L. 2019. An overview and perspectives on bidirectional intelligence: Lmser duality, double IA harmony, and causal computation. *IEEE/CAA Journal of Automatica Sinica*, 6(4): 865–893.

Yazdani-Jahromi, M.; Yousefi, N.; Tayebi, A.; Kolanthai, E.; Neal, C. J.; Seal, S.; and Garibay, O. O. 2022. AttentionSit-eDTI: an interpretable graph-based model for drug-target in-teraction prediction using NLP sentence-level relation clas-sification. *Briefings in Bioinformatics*, 23(4): bbac272.

Zhao, Q.; Zhao, H.; Zheng, K.; and Wang, J. 2022. HyperAt-tentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3): 655–662.

Zheng, S.; Li, Y.; Chen, S.; Xu, J.; and Yang, Y. 2020. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2 (2): 134– 140.