山西大学学报(自然科学版)47(5):982—992,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024032

非线性子空间驱动下的耐药性预测方法

董云云[1],张源榕[1],龚怡丰[2],白玉洁[1],常云青[1],杨炳乾[1],杨紫婷[1],徐双[3],强彦[3]

(1.太原理工大学 软件学院,山西 太原 030024;
2.太原理工大学 电子信息与光学工程学院,山西 太原 030024;
3.太原理工大学 计算机科学与技术学院(大数据学院),山西 太原 030024)

摘 要:癌症的耐药性预测任务已经成为精准医学领域前瞻性研究方向之一。针对现有耐药性预测方法难以深度表征药物和细胞系之间协同关系的问题,提出一种非线性子空间驱动下的耐药性预测方法NLS-DRP(Nonlinear Subspace-Driven Drug Resistance Prediction)。NLS-DRP包括Cell分支、Drug分支和协同融合三个关键学习模块,分别用于构建非线性子空间提取细胞系特征,拆分药物结构提取子序列特征,设计非线性协同空间融合细胞系和药物特征;最后,通过融合三个模块的特征,实现细胞系对药物的耐药性预测。在癌症药物敏感性基因组学和癌症细胞系百科全书两个公开数据集上进行实验,结果表明所提NLS-DRP模型显著优于对比的基准方法,取得了0.945 8的皮尔逊相关系数(PCC)值和0.924 2的斯皮尔曼相关系数(SCC)值,验证了本文方法的有效性。

关键词:图神经网络;特征融合;非线性子空间;智能用药
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)05-0982-11

Nonlinear Subspace-Driven Drug Resistance Prediction

DONG Yunyun[1], ZHANG Yuanrong[1], GONG Yifeng[2], BAI Yujie[1], CHANG Yunqing[1], YANG Bingqian[1], YANG Ziting[1], XU Shuang[3], QIANG Yan[3]

(1. *School of Software, Taiyuan University of Technology, Taiyuan* 030024*, China;*
2. *College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan* 030024*, China;*
3. *College of Computer Science and Technology* (*College of Data Science*)*, Taiyuan University of Technology, Taiyuan* 030024*, China*)

Abstract: The task of predicting drug resistance in cancer has emerged as a prospective research direction in the field of precision medicine. To address the challenge of limited representation of the synergistic relationship between drugs and cell lines in existing resistance prediction methods, this paper proposes a nonlinear subspace collaborative learning model, named NLS-DRP (Nonlinear Subspace-Driven Drug Resistance Prediction). The NLS-DRP consists of three key learning modules: the Cell branch, the Drug branch, and the Collaborative Fusion module. These modules are used to construct nonlinear subspaces for extracting cell line features, decompose drug structures to extract subsequence features, and design a nonlinear collaborative space for the fusion of cell line and drug features, respectively. Finally, by integrating the features from the three modules, the resistance of cell lines to drugs is predicted. Experiments conducted on two public datasets, the Genomics of Drug Sensitivity in Cancer(GDSC) and the Cancer Cell Line Encyclopedia (CCLE), demonstrate that the proposed NLS-DRP model significantly outperforms the benchmark methods, achieving a Pearson Correlation Coefficient (PCC) value of 0.945 8 and a Spearman's Correlation Coefficient (SCC) value of 0.924 2,

收稿日期:2023-11-29;**接受日期**:2024-03-04
基金项目:国家自然科学基金(62306206;62102280);山西省重点研发计划项目(202101010101007;202102020101001);山西省基础研究计划资助项目(202203021212207;20210302124167)
作者简介:董云云(1988-),女,山西晋城人,博士,讲师,研究方向为医学图像处理,影像基因智能计算。E-mail:dongyunyun@tyut.edu.com
引文格式:董云云,张源榕,龚怡丰,等.非线性子空间驱动下的耐药性预测方法[J].山西大学学报(自然科学版),2024,47(5):982-992. DOI:10.13451/j.sxu.ns.2024032

thereby confirming the effectiveness of the method presented in our paper.

Key words: graph neural network; feature fusion; nonlinear subspace; intelligent medication

0 引言

精准医疗旨在有效整合临床数据、基因组学以及其他组学等多组学数据,以发现具有预后生物标志物的信息,为患者制定个性化治疗方案[1-2]。然而,肿瘤内部和肿瘤间的异质性导致不同患者对同一药物的反应差异较大,这成为制约精准医疗惠及患者的一大难题[3]。耐药性预测通过对病原体基因组序列和表达等因素进行分析和建模,协助医生制定最优化的治疗方案,提供重要的用药反馈和参考,已成为精准医学领域前瞻性研究方向之一。耐药性预测涉及基因组学、转录组学和蛋白质组学等多个领域[4],传统的预测方法主要包含统计学方法[5]以及机器学习算法[6-9]等。传统方法的主要缺点是依赖于特定的实验数据,难以处理多个药物间的相互作用。由于数据样本有限和特征维度高的问题,此类方法易产生过拟合现象,导致预测精度有限。

随着人工智能技术的迅速发展,基于深度学习的耐药性预测方法克服了传统方法的局限性。CDRscan模型[10]作为早期采用深度学习方法预测药物反应的模型,基于基因组图谱和药物结构图谱来预测细胞系对抗癌药物的敏感程度。DeepDSC模型[11]将细胞系的基因特征和药物的化学信息进行整合,用于预测耐药性。SWnet模型[12]将基因表达、基因突变和化合物的化学结构集成在多任务卷积架构中,实现了耐药性预测任务。上述研究将深度学习引入了耐药性预测任务中,然而忽略了药物本身的拓扑结构,DeepCDR[13]是首个将图卷积神经网络用于预测癌症药物反应的深度模型,它整合了癌细胞的多组学特征并探索药物的内在化学结构,提出了用于识别特定癌症类型中相关潜在基因的探索策略。tCNNs模型[14]引入了双卷积神经网络,预测药物与癌细胞系之间的相互作用。MOLI[15]将体细胞突变、拷贝数畸变和基因表达数据作为输入,整合起来进行药物耐药性反应预测。DeepTTA[16]基于Transformer在药物特征提取中通过神经网络对转录组特征建模进行耐药性预测。

以上研究通过分析细胞系的基因表达数据和药物的分子结构信息,预测不同细胞系对特定抗癌药物的敏感性。然而上述方法主要是对药物分子和细胞系结构分别进行建模,提取药物和细胞系特征进行耐药性预测任务,尚未考虑到“细胞系-药物”关联对药物反应的影响,导致模型无法从数据中学习到关联规律。除此之外,传统的耐药性预测任务通常只通过不同的分支结构从整体的细胞系和药物中提取特征,忽略了细胞系和药物中的子结构特征。

癌症患者病情的发展与耐药性之间存在着紧密的依赖关系,因此药物与细胞系之间的相互作用对耐药性预测任务至关重要。考虑到药物分子结构和细胞系拓扑结构的复杂关联以及细胞系和药物中的子结构,本文提出了非线性子空间驱动下的耐药性预测(Nonlinear Subspace-Driven Drug Resistance Prediction, NLS-DRP)方法。NLS-DRP模型基于药物分子和细胞系的独特结构构建了Cell分支和Drug分支,提出非线性子空间来提取细胞系的潜在特征,引入频繁连续子序列来提取药物子序列特征;同时,设计非线性协同空间融合表征算法捕获两者潜在关联,融合细胞系和药物的特征信息;最后,利用相关线性函数进行耐药性预测任务。

本文主要工作如下:

(1)结合药物分子表达和细胞系拓扑结构间潜在关联,提出非线性子空间驱动下的耐药性预测模型,用于预测细胞系中的药物反应;

(2)构建非线性子空间驱动下的Cell分支、提取药物子序列特征的Drug分支和非线性协同空间驱动下的特征学习网络,实现“细胞系-药物”特征空间映射和协同融合;

(3)与五种先进的算法进行了对比实验。实验结果表明,NLS-DRP模型在癌症药物敏感性基因组学(Genomics of Drug Sensitivity in Cancer, GDSC)和癌症细胞系百科全书(Cancer

Cell Line Encyclopedia,CCLE)数据集上取得了更精准的预测结果。

1 非线性子空间驱动下的耐药性预测模型

考虑到药物分子表达和细胞系拓扑结构通常具有复杂的关联,传统的特征提取方法往往忽略了细胞系和药物之间的相互作用,导致其潜在链接丢失,无法捕获药物和细胞系之间的关联规律。本文提出了端到端的非线性子空间驱动下的耐药性预测方法,其总体框架如图1所示。NLS-DRP模型由非线性子空间映射下的Cell特征提取,Drug子序列特征提取和特征融合模块组成。

1.1 非线性子空间映射下的Cell分支

本文构建了非线性子空间映射下的Cell分支,如图1(a)所示。首先,基于细胞系中独特的通路对细胞系进行结构化表示,构建细胞系非线性子空间(Nonliner Subspace);其次,提出了一种基于细胞系处理的非线性子空间编码器(Nonliner Subspace Encoder),用于提取非线性子空间中基因的相互作用;最后,将非线性子空间编码器得到的特征表示输入Transformer得到Cell空间表征。

1.1.1 细胞系非线性子空间构建

在模型训练之前对细胞系C进行结构化表示,利用细胞系中独特的通路,将细胞系划分为多个子细胞系$C(C_1,C_2,C_3,\cdots,C_n)$。本文基于细胞系中的基因-基因相互作用,构建非线性空间,将每个子细胞系中特有的结构映射为非线性子空间,在非线性子空间中基于基因之间的关联性构建同构图$G(X,A)$,将基因映射为节点v;将基因之间的关联表示为边e。其中同构图G中,X是基因的特征表示,其每一行对应于一种基因的特征;A是图G的邻接矩阵,表示基因-基因相互作用,代表图中节点的总连通性。

1.1.2 非线性子空间特征提取

为了提取非线性子空间的全局信息,本文构建了非线性子空间编码器来捕获细胞系特征和基因-基因相互作用,对节点特征和网络拓扑结构建模,提取前一层的节点信息和节点间

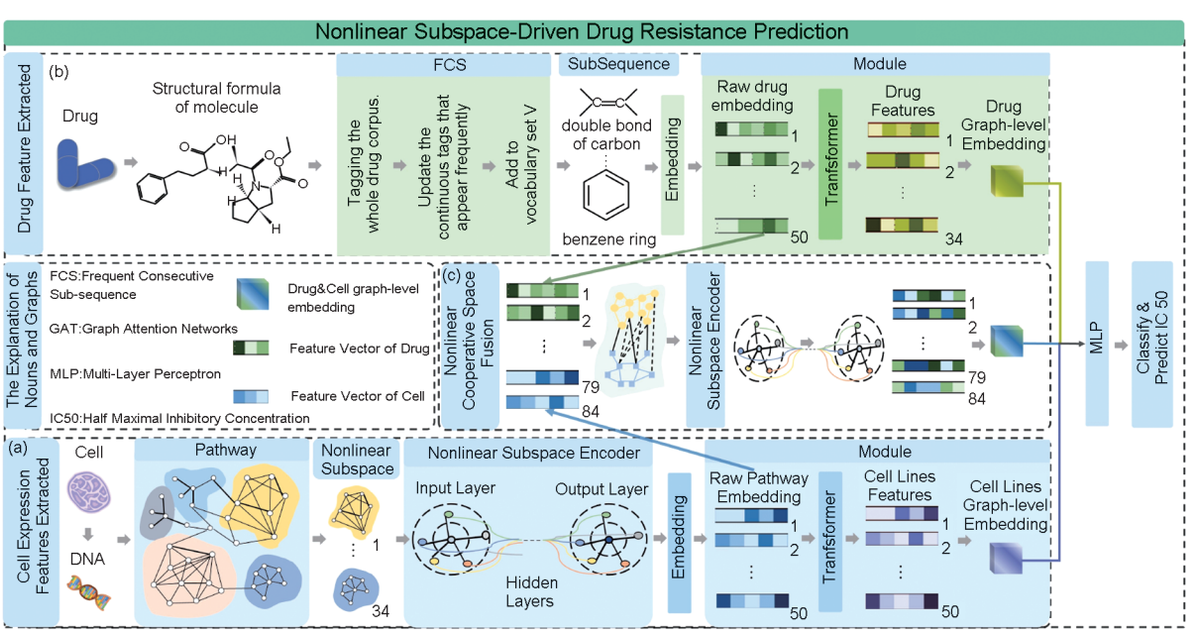


图1 NLS-DRP模型结构示意图

(a)分支为Cell分支,通过非线性子空间提取细胞系特征;(b)分支为Drug分支,通过SMILES药物分子的子序列提取药物特征;(c)分支为非线性协同空间,通过非线性子空间编码器进行特征融合,挖掘细胞系和药物的相互作用

Fig. 1 Schematic diagram of the structure of the NLS-DRP model

(a) The branch is the Cell branch, which extracts cell line features through nonlinear subspace; (b) the branch is the Drug branch, which extracts drug features through the subsequences of SMILES drug molecules; (c) the branch is nonlinear collaboration space, feature fusion through nonlinear subspace encoder, mining the interaction between cell lines and drugs

的链接,对于顶点v_i,逐个计算它的邻居v_j与其之间的相似系数e_{ij}:

$$\| e_{ij} = LeakyReLU\left(\gamma^T[Wh_i \| Wh_j]\right), \quad (1)$$

其中$\|$表示向量拼接操作,W,γ分别为不同的权重参数,激活函数设计为$LeakyReLU$,h_i为当前节点v_i的特征表示,h_j为邻居节点v_j的特征表示。为了更合理地分配权重,对每个节点的表示进行编码,将中心节点与邻居节点计算出来的相关度进行softmax归一化处理,得到归一化注意分数α_{ij}:

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in S_i}\exp(e_{ik})}, \quad (2)$$

其中S是图中节点个数。本文使用归一化注意分数α_{ij}反映邻居节点信息的重要程度,利用消息传递机制对相邻节点的特征进行传播,通过对邻居节点h_j的聚合来传播节点的信息,从而更新每个中心节点h_i的状态。如公式(3)所示,得到更新后的特征表示$\vec{h}_i$:

$$\vec{h}_i = \sigma\left(\sum_{j\in S_i}\alpha_{ij}\Psi h_j\right), \quad (3)$$

其中Ψ是可学习的权重参数,σ是一种非线性激活函数,S是图中节点个数。非线性子空间的特征可表示为$E_c = \sum_{i=1}^{N}\vec{h}_i$,$N$为非线性子空间的个数。

1.1.3 Cell空间表征

本文使用由多头自注意力层和全连接前馈网络组成的Transformer编码器对非线性子空间捕获到的特征进行编码,利用多头自注意力机制对特征进行交互,融合多个子空间的特征表示,挖掘细胞系内多个非线性子空间的一致性特征,多头自注意力层的输出按公式(4)输入到全连接前馈层中得到Cell空间表征F_C:

$$F_C = \max\left(0, \mathrm{Attention}(E_c)W_1 + b_1\right)W_2 + b_2, \quad (4)$$

其中W_1, W_2是可学习参数,b_1, b_2为偏置。

1.2 Drug子序列特征提取分支

本文构建了一个Drug子序列特征提取分支,以药物分子的特殊结构为基础,对药物信息进行特征提取,如图1(b)所示。首先,对药物分子结构进行子序列提取;其次,针对提取的子序列提出了内容和位置embedding;最后,将经过内容和位置embedding得到的子序列表征输入到Transformer编码器进行融合编码得到Drug空间表征。

算法1 频繁连续子序列挖掘(FCS)

输入:A为包含所有初始SMILES标记的集合
D为包含已经标记的药物的集合
ξ为指定频率阈值
Max为A的最大尺寸
输出:D为更新后的标记药物,A为更新后的标记词汇表
for $t = 1\cdots$ Max do
  扫描D获得子序列(M,N)的频率θ
  if 频率$\theta > \xi$
    将D中的(M,N)替换为(MN)
    将(MN)加入A
end for
Return D,A

1.2.1 频繁连续子序列挖掘

本文引入了频繁连续子序列挖掘(Frequent continuous subsequence mining,FCS)[17]方法,基于简化分子线性输入规范(Simplified molecular input line entry system,SMILES)药物分子结构将药物分解为一组明确的子结构序列。FCS如算法1所示。本研究将包含初始SMILS标记的集合记作A,将已经标记的药物集合记为D,ξ为指定频率阈值。FSC算法对于A中的每个标记进行迭代处理,在每次迭代中,算法扫描D,寻找频率最高的连续标记对(M,N),当(M,N)的频率高于指定的阈值ξ,在D中将所有(M,N)的出现替换为组合标记(MN),将新的组合标记(MN)添加到A中;如果(M,N)的频率低于阈值ξ,迭代停止。最终FCS算法输出更新后的D和A。

FCS算法旨在识别和组合在药物SMILES序列中频繁出现的子序列。通过这种方式,它揭示了药物分子的重要亚结构特征。FCS算法首先将药物分子序列分解为亚结构,即较小的、有意义的序列片段。对输入的SMILES字符串进行分子解析和标准化处理,确保分子结构的一致性。通过对药物分子的遍历,提取由相邻原子和键组成的子结构,构建子结构库。这种分解基于识别序列中重复出现的模式或子序列,这些子序列可能对理解药物与蛋白质之间的相互作用具有重要意义。为了提取有意义的子结构,FCS算法通过搜索子结构库中的组合,根据预设的评价准则选择最具代表性的子

结构序列,将子结构表示为子序列。算法通过设置一个频率阈值ξ来确定子序列的重要性。只有那些出现频率高于此阈值的子序列才会被考虑。此外,算法专注于连续子序列,即那些在序列中连续出现的部分。通过利用大量的未标记数据,FCS算法能够识别难以发现的重要子结构。这种方法提升了子结构挖掘的质量并增强了算法在耐药性方面的预测能力。

1.2.2 基于内容嵌入和位置编码的药物子序列提取方法

本文提出了一种挖掘药物子序列的内容嵌入和位置编码的方法,用于处理SMILES药物分子结构中相邻子结构之间复杂的化学关系,该方法将每个子结构C_d嵌入到一个潜在特征向量$\boldsymbol{E}_d$中。

具体而言,将子序列结构C_d映射为一个矩阵$\boldsymbol{M}$,$\boldsymbol{M}$的任意一列j代表一种药物序列的子结构。内容嵌入模块通过可学习的字典查找矩阵$\boldsymbol{W}_{\text{cont}}^{d}\in\mathbb{R}^{\vartheta\times l}$将每个药物子序列的化学特征$\boldsymbol{M}_j^d$编码为一个具有丰富化学信息的向量表示,以捕获其结构信息,其中,ϑ为每个子结构潜在嵌入的大小,l为药物子结构的总数量;位置编码模块通过查找字典$\boldsymbol{W}_{\text{pos}}^{d}\in\mathbb{R}^{\vartheta\times\Theta_d}$,为每个子结构分配一个位置编码向量,表示其在分子中的相对位置,捕获当前特征在整个子序列中的相对位置信息,其中,$\mathbb{I}_j^d\in\mathbb{R}^{\Theta_d}$是一个One-Hot向量,$\Theta_d$是药物子序列的最大长度。任意药物$E_j^d$内容嵌入和位置编码的结果如公式(5)所示:

$$E_j^d=\boldsymbol{W}_{\text{cont}}^{d}M_j^d+\boldsymbol{W}_{\text{pos}}^{d}\mathbb{I}_j^d。\tag{5}$$

药物子序列的特征可表示为$E_D=\sum_{j=1}^{K}E_j^d$,其中K为非线性子空间的个数。

综合内容嵌入和位置编码,能够有效地处理SMILES药物分子结构中的复杂化学关系,更准确地捕捉药物分子的结构特征,为后续的药物分子分析和预测任务提供更有力的特征表达。

1.2.3 药物特征融合

为了进一步融合药物特征并深入理解药物分子的结构特征,将特征向量$\boldsymbol{E}_d$传入Transformer编码器。通过利用Transformer的多头自注意机制,学习不同药物子序列特征的相关性,增强了模型对药物特征的提取和理解能力,获得了药物结构的特征表示F_D:

$$F_D=\max\left(0,\text{Attention}(\boldsymbol{E}_d)\mu_1+\beta_1\right)\mu_2+\beta_2,\tag{6}$$

其中μ_1,μ_2是可学习参数,β_1,β_2为偏置。

1.3 非线性协同空间特征融合

基于细胞系和药物之间的相互作用,本文提出了一个非线性协同空间,如图1(c)所示,用于对细胞系和药物之间的相互作用建模,该模块将经过非线性子空间编码器编码的细胞系特征和经过子序列嵌入模块得到的药物特征共同输入协同空间中,融合得到细胞系-药物特征。

考虑到基因与药物间的复杂特征及其关联,本文将细胞系特征和药物子序列特征映射到非线性协同空间内,构建异构图$G'(X',\boldsymbol{A}')$。其中,X'是基因节点和药物节点特征之和,$X'=E_C\cup E_D$,$\boldsymbol{A}'$是图G'的邻接矩阵,代表节点间的关联性。将非线性子空间编码器获取的细胞系特征映射为基因节点V_{Gene};将经过内容嵌入和位置编码后的药物子序列特征映射为药物节点V_{Drug},将基因间相互作用,药物间相互作用和基因和药物之间的相互作用分别映射为图G'中的链接E_{Gene}、E_{Drug}和$E_{\text{Gene-Drug}}$。其中$V=V_{\text{Drug}}\cup V_{\text{Gene}}$,$E=E_{\text{Gene-Drug}}$。将构建的异构图输入到非线性子空间编码器中,如算法2所示。在算法2中,使用随机矩阵初始化异构图的节点嵌入表示,对于每一个节点V_i,计算其与邻居节点V_j之间的注意力分数,使用注意力分数更新节点V_i嵌入表示E_U。

本文通过将细胞系和药物的特征在协同空间内融合,能够更全面地捕捉它们之间的复杂相互作用。通过基因和药物之间的相互作用建立不同节点之间的链接,从而更好地理解它们之间的复杂关系。通过非线性子空间编码器学习不同节点的嵌入表示,捕捉节点之间的复杂关系,自适应地分配注意力权重来捕捉节点之间的重要性和关联性,通过融合不同类型的特征,挖掘异构图的局部和全局上下文信息,实现不同类型节点之间的信息交互和传播,从而提高预测和推理性能。

非线性协同空间的特征表示为$F_U=\sum_{k=1}^{N}E_U$,其中N为异构图中节点的个数。

综上,本文将药物特征、细胞系特征和细胞

算法2 非线性协同空间特征融合

输入:基因特征集合E_C,药物特征集合E_D,基因和药物的相互作用矩阵A'

输出:节点嵌入表示E_U

#构建异构图G

$X'=E_C \bigcup E_D$

$G'=(X',A')$

#非线性协同空间的特征提取

for each $v \in G'$ do

计算节点v的相似系数e_{ij}

计算节点v的归一化注意分数α_{ij}

更新节点v的嵌入表示E_U

end for

Return E_U

系-药物融合特征输入到多层感知机(Multilayer Perceptron,MLP)中,利用MLP的非线性建模能力,从多种特征F_C,F_D,F_U中提取相关信息,计算IC_{50}值。具体地,通过将F_C,F_D,F_U的特征输入到MLP中,利用MLP的多个隐藏层和非线性激活函数,从特征中提取相关信息,学习特征之间的复杂关系和权重分配,充分利用不同特征之间的非线性关系,提取特征的高阶表示,建立药物特征、细胞系特征和药物-基因融合特征与IC_{50}值之间的关联,进行耐药性预测。

2 实验及结果

2.1 数据集介绍

在本研究中,采用Genomics of Drug Sensitivity in Cancer2(GDSC2)、Cancer Cell Line Encyclopedia(CCLE)[18]数据集进行实验。

GDSC2数据库包含了大量关于不同癌症细胞系对多种抗癌药物反应的数据。这些数据涵盖了近700个癌症细胞系对约251种抗癌药物的反应,总计近7.5万个药物敏感性实验数据。GDSC提供了各种癌症细胞系的全基因组数据,包括基因表达、拷贝数变异(Copy Number Variation,CNV)、基因突变等,这些数据记录了细胞系对不同药物的敏感性。

CCLE收集了36种癌症细胞系的基因组学数据,这些细胞系来源于21种不同的人体组织,记录了不同癌症细胞系对药物的反应,用于衡量细胞系对药物的敏感性。CCLE数据集包含491个同时具备基因表达、突变、拷贝数变异以及药物反应数据的癌症细胞系。

本文从CCLE数据集中提取了基因组变异、基因表达、蛋白质表达等数据,用于识别与癌症发展相关的关键基因和信号通路;利用GDSC2记录的Pubchem ID从Pubchem数据库中提取SMILES描述符。其中,基因突变的数据反映了遗传变异,这些变异可能与癌症发展和药物敏感性相关,可用于识别癌症驱动基因和潜在的治疗靶点;基因表达数据显示了细胞内不同基因的活跃程度,通常通过RNA测序获得,可以揭示细胞系的生物学特性和药物响应;拷贝数变异是指基因组DNA片段数量的变化,可能导致基因剂量的改变,CNV在许多癌症类型中普遍存在,与癌症的发展和药物敏感性有关;而SMILES字符串通常可以反映出药物分子结构的特征,通过将SMILES表示的药物分子特征与癌症细胞系的特征相结合,模型可以理解药物与生物靶标的相互作用。

最终本文根据癌症类型从KEGG(Kyoto Encyclopedia of Genes and Genomes)[19]通路数据库中筛选数据,剔除掉重复路径和代谢路径后,最终选定34条与癌症相关的通路,包括170种药物,580个细胞系,共计98 600个IC_{50}值。由于部分细胞系缺失,最终参与计算的IC_{50}值为82 833个。

2.2 实验细节

本文使用RTX Nvidia 4090 GPU,基于PyTorch框架进行训练和测试。将原始数据集按80%、10%、10%的比例划分为训练集、测试集和验证集。采用Adam优化器来更新模型参数。实验中将Batchsize设置为128;Learning rate设置为1×10^{-4};Dropout设置为0.1;实验中使用了6层的Transformer编码器,我们为Transformer编码器设置了16个注意力头,并将隐藏层的维度设置为256;此外本文还构建了非线性子空间编码器,我们将每个非线性子空间编码器的编码器层数设置为3层,使用8个注意力头,本文将第一个非线性子空间编码器的隐藏层数设置为8,第二个的隐藏层数设置为256。

为了验证本文所提出模型的有效性,将实验划分为分类和回归任务。回归任务采用

MLP层进行预测,使用均方根误差作为损失函数来衡量预测结果与真实值之间的差异;分类任务采用交叉熵作为损失函数来度量预测结果与真实标签之间的差异。

2.3 评价指标

对于回归预测模型,使用均方根误差(Root Mean Squared Error,RMSE)来评价精度水平;为了衡量IC_{50}的真实值与预测值的线性相关关系,使用皮尔逊相关系数(Pearson Correlation Coefficient,PCC)、斯皮尔曼相关系数(Spearman Rank Correlation Coefficient,SCC)作为回归指标。

RMSE测量预测值与真实值之间的差值,其中n为数据个数,o_i为第i个样本的基真值,y_i为第i个样本的预测值,如式(7)所示。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i}^{n}(o_i-y_i)^2}。\tag{7}$$

PCC用于衡量两个变量之间的关联强度,如式(8)所示。

$$PCC=\frac{\sum_{i}^{n}(o_i-y_i)^2}{\delta_O\delta_Y}。\tag{8}$$

SCC是一个非参数的、与分布无关的秩统计参数,通常被认为是排序变量之间的皮尔逊线性相关系数。SCC相关系数如式(9)所示。

$$SCC=1-\frac{6\sum_{i=1}^{n}(O_i-Y_i)^2}{n(n^2-1)}。\tag{9}$$

2.4 实验结果

本文与近期流行的耐药性预测深度模型MOLI、tCNNS、CDRscan、DeepCDR和DeepTTA等进行了对比,结果如表1所示。相对于其他方法,NLS-DRP方法显示出更高的PCC、SCC和AUC(Area Under Curve)值,同时具有最低的RMSE值。与已有的深度方法相比,本文提出的NLS-DRP方法专注于提取药物的非线性子空间特征和细胞系的子序列特征,能够聚焦于捕捉药物和基因之间内在结构信息,具备更好地预测性能。

本文利用NLS-DRP模型预测未知细胞系的耐药性。为此使用了包含561个细胞系和238种药物的已知药物-细胞系相互作用来训练NLS-DRP模型,并将该模型应用于预测GDSC2数据库中缺失的药物-细胞系。图2展示了按药物分组的GDSC2数据库中预测的IC_{50}值的分布情况,药物根据其在细胞系中的平均预测IC_{50}值进行了排序。该结果图直观展示了前10个“敏感性”药物以及最后10个“耐药性”药物。在这里,IC_{50}值较低表示药物的治疗效果较好,被称为“敏感性”,IC_{50}值较高的药物则被归类为“耐药性”。与预期一致,实验证明在多种细胞系实验和癌症治疗中bortezomib是最有效的药物之一[20]。

2.5 可视化结果

2.5.1 回归结果

图3的可视化结果呈现了NLS-DRP模型在药物反应回归任务中的表现。通过散点图的形式对真实值与预测值的分布情况进行可视化,结果显示,所提出的NLS-DRP模型对于大量随机样本都表现出了良好的预测性能。

2.5.2 分类结果

本节对提出的NLS-DRP模型和对比模型在药物反应中进行了分类任务的性能测试,采用受试者工作特征曲线(Receiver Operating Characteristic, ROC)和精确率-召回率曲线(Precision-Recall, PR)作为评价指标,如图4所示。在本节中,根据Iorio等[21]提供的每种药物的阈值,对IC_{50}进行了二值化处理。尽管数据

表1 NLS-DRP与其他模型实验结果的对比

Table 1 Comparison of experimental results between NLS-DRP and other models

	RMSE↓	PCC↑	SCC↑	AUC↑
MOLI[15]	2.282 3±0.000 8	0.813 7±0.000 7	0.782 1±0.000 5	0.699±0.000 7
tCNNS[14]	1.782 1±0.000 6	0.885 5±0.000 8	0.862 9±0.000 6	0.755±0.000 6
CDRscan[10]	1.982 6±0.000 5	0.871 1±0.000 5	0.852 3±0.000 3	0.746±0.000 4
DeepCDR[13]	1.058 3±0.000 6	0.923 5±0.000 6	0.903 7±0.000 4	0.841±0.000 5
DeepTTA[16]	0.952 1±0.000 7	0.941 0±0.000 3	0.914 3±0.000 4	0.835±0.000 6
NLS-DRP	0.912 6±0.000 5	0.945 8±0.000 3	0.924 2±0.000 7	0.862±0.000 3

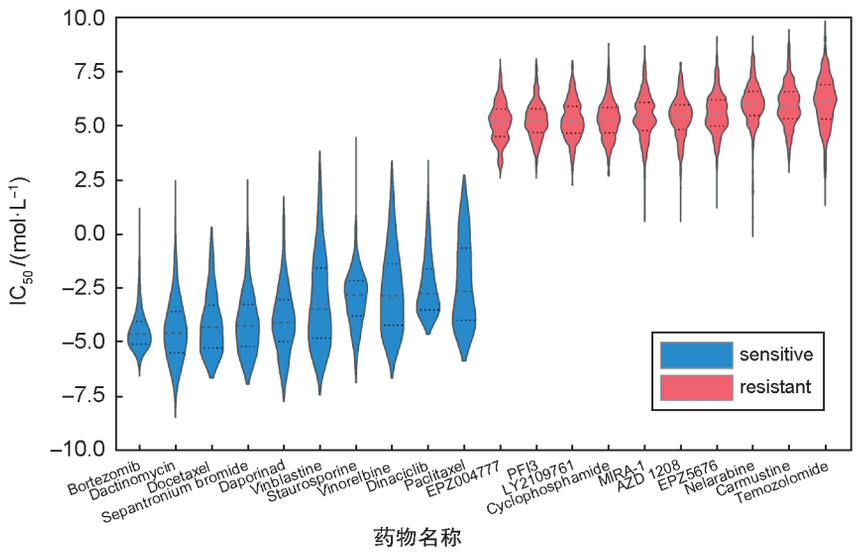
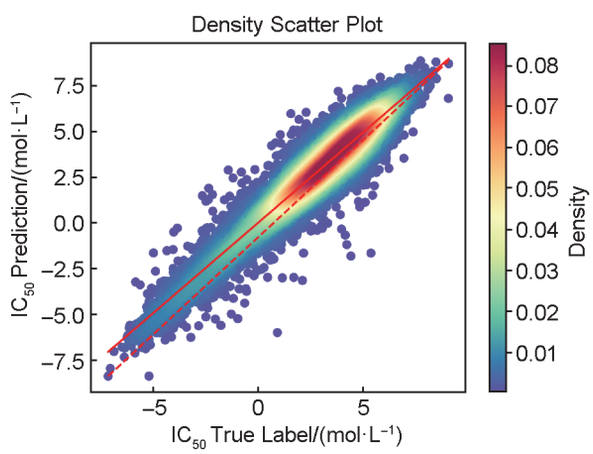


图2 按药物分组的GDSC2数据库中缺失数据的预测IC_{50}值

每个小提琴图代表了在所有细胞系中对特定药物的反应情况。蓝色和红色的小提琴图分别对应着疗效最佳和最差的前10种药物

Fig. 2 Predicted IC_{50} values for missing data in the GDSC2 database grouped by drug, with each violin plot representing the response to a specific drug across all cell lines. The blue and red violin plots correspond to the top 10 drugs with the best and worst efficacy, respectively



注:横坐标代表实际的IC_{50}值,纵坐标则表示预测的IC_{50}值,颜色从浅至深表示密度的增加程度。红色实线表示基于数据进行线性拟合的预测结果;而红色虚线则表示在理想情况下,预测结果与真实结果完全一致。

图3 NLS-DRP模型在回归任务中的可视化图

Fig. 3 Visualization of the NLS-DRP model in the regression task

集存在较为严重的不平衡(约1∶7),但NLS-DRP在AUC和PR分数上明显优于其他四种方法,分别达到0.862和0.532。

2.6 消融实验

本研究致力于将药物结构和基因编码进行非线性空间协同的深度模型的构建。为此,提出了一种构建非线性子空间的方法,以提取基因信息;构建了非线性协同空间用于融合基因-药物特征;设计了非线性子空间编码器用于提取信息表征。本节中,对非线性子空间、非线性协同空间的构建、非线性子空间编码的有效性进行了评估。

2.6.1 非线性子空间的消融实验

在本次消融实验中,对非线性子空间的构建对模型的影响进行了评估。具体地,将细胞系中的基因信息映射为图结构,而非通过癌症通路对细胞系进行划分,并评估了该方法与本文方法(基于癌症通路构建非线性子空间)的性能差异,对比结果如表2所示。结果显示,相较于本文提出的癌症通路构建非线性子空间的方法,直接映射细胞系结构进入编码器导致模型的RMSE增加,同时PCC和SCC降低。进一步验证了基于癌症通路的非线性子空间构建方法对于准确提取细胞系特征的重要性。通过使用癌症通路来划分细胞系,能够更好地捕捉基因之间的相互作用和关联性,从而提高编码器的性能。

2.6.2 非线性协同空间的消融实验

在本次消融实验中,对非线性协同空间在模型中的影响进行了评估。探究了将细胞系特征和药物子序列特征直接输入Transformer模型

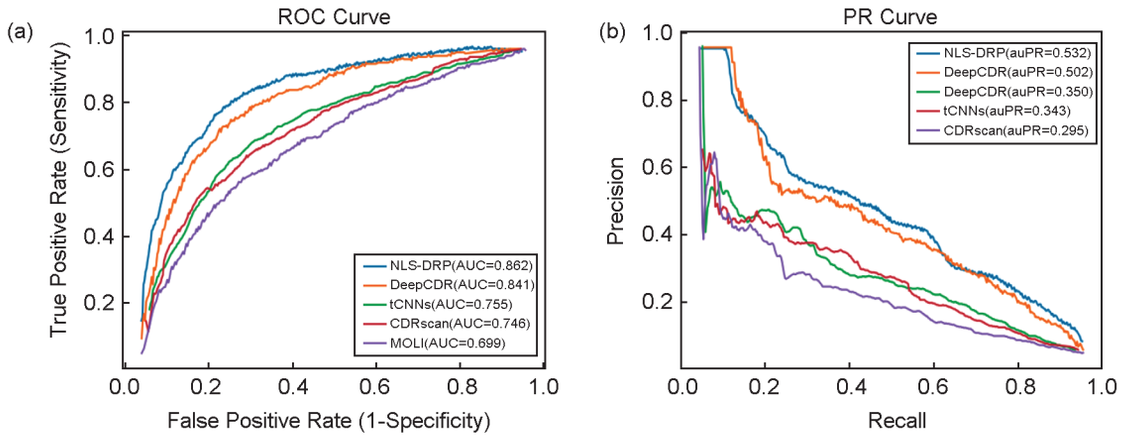


图4 NLS-DRP模型和对比模型在分类任务中结果分析

(a)为ROC曲线,其中横坐标代表假阳性率,纵坐标则表示真阳性率;(b)为PR曲线,其中横坐标代表召回率(Recall),纵坐标代表精度(Precision)

Fig. 4 NLS-DRP model and comparative models' results in classification tasks

(a) the ROC curve, where the abscissa represents the false positive rate, and the ordinate represents the true positive rate; (b) the PR curve, where the abscissa represents the recall rate, and the ordinate represents the precision

表2 非线性子空间和非线性协同空间的消融实验结果,本文对非线性子空间和非线性协同空间在模型中的作用做了不同的消融实验已验证其效果

Table 2 Results of ablation experiments on nonlinear subspace and nonlinear co-subspace. This paper has conducted different ablation experiments on the role of nonlinear subspace and nonlinear collaborative space in the model to verify their effects

非线性子空间	非线性协同空间	RMSE↓	PCC↑	SCC↑
×	√	0.946 5±0.000 9	0.924 3±0.000 8	0.909 6±0.000 3
√	×	0.959 0±0.000 6	0.917 2±0.000 6	0.891 3±0.000 6
×	×	0.985 8±0.000 4	0.893 7±0.000 1	0.882 9±0.000 4
√	√	**0.912 6±0.000 5**	**0.945 8±0.000 3**	**0.924 2±0.000 7**

的方式,而非构建协同空间。对比结果如表2所示。

结果显示,与本文提出的模型相比,对比方式导致了模型的RMSE增加,同时PCC和SCC减少。相比之下,本文提出的模型采用了非线性协同空间来处理细胞系特征和药物子序列特征,能够更好地捕捉细胞系和药物之间的复杂关系,并提取出具有更丰富表征能力的特征,提高模型的整体性能。

2.6.3 非线性子空间编码器的消融实验

在本研究中,构建了提取非线性子空间特征的编码器,模型构建过程考虑了几种常用的模型,包括图神经网络(Graph Neural Networks, GNN)[22],图卷积网络(Graph Convolutional Network, GCN)[23]和图注意力网络(Graph Attention Network, GAT)[24],本次消融实验探索使用不同的编码器对模型整体性能的影响。此外,为了证明非线性子空间编码器对模型的有效性,消融实验中去除了非线性子空间,将高维数据的特征直接嵌入到低维的空间当中,以此来对比非线性子空间在模型中的作用。对比结果如图5所示。结果显示,直接将特征嵌入到低维空间中的性能均低于使用非线性子空间编码器,证明了本文中提出的非线性子空间编码器的有效性。而在不同的非线性子空间的对比中,相比于GNN,使用GAT作为非线性子空间编码器导致RMSE降低了3.324%,同时PCC和SCC分别提升了1.079%和3.691%,结果表明GAT网络在本次实验中对非线性子空间的编码具有优势。GAT利用注意力机制有效地捕捉了节点之间的重要关系,更准确地学习和表示非线性子空间的特征。通过使用GAT网络作为编码器,模型能够更好地提取和表达特征信息,从而在整体上实现更好的性能。

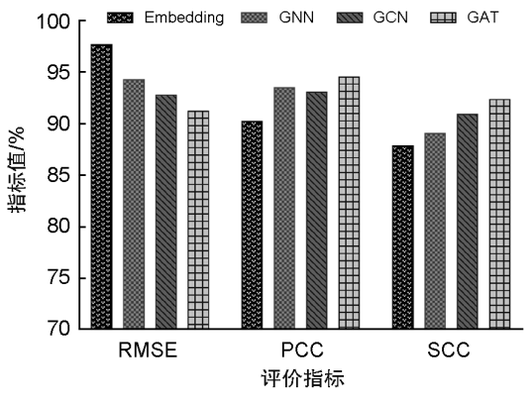


图5 不同编码器的性能对比,本文使用低维嵌入、GNN、GCN和GAT分别作为非线性子空间编码器以验证不同非线性子空间编码器的效果

Fig. 5 Performance comparison of different encoders. This paper uses embedding, GNN, GCN and GAT as nonlinear subspace encoders, respectively, to verify the effects of different nonlinear subspace encoders

2.6.4 非线性子空间编码器时间复杂度分析

本文将基于图的非线性拓扑结构映射到非线性子空间内,将每个子细胞系中特有的结构映射为非线性子空间,在非线性子空间中基于基因之间的关联性构建同构图$G(X,A)$。在本文中,将非线性子空间中映射的图结构输入到非线性子空间编码器中,使用$|V|$表示节点数,$|E|$表示边的数量,使用F表示原始的特征维度,F'表示非线性子空间输出的特征维度。

非线性子空间编码器的时间复杂度可以分为对每个节点的特征映射和对注意力头的注意力计算两部分。对于任意一个节点而言,非线性子空间编码器节点的原始输入维度映射到输出维度,则对单独节点而言,映射复杂度为$O(F\times F')$。而对于输入的非线性子空间而言,非线性子空间内的所有节点都需要进行 特征维度的映射,则计算复杂度为$O(|V|\times F\times F')$。在计算图中的注意力权重时,非线性子空间编码器计算图中每一条边的注意力系数,将其特征维度映射为一个实数,则对于单个注意力头而言,计算复杂度为$O(E\times F')$。综上所述,K个注意力头的非线性子空间编码器的计算复杂度为$O(|V|FF'+KEF')$。

3 结论

本文提出了一种名为NLS-DRP的端到端耐药性预测模型,这是首个将药物结构和基因编码进行非线性空间协同的深度模型。NLS-DRP中提出非线性子空间提取细胞系的潜在特征,引入频繁连续子序列来提取药物子序列特征,设计非线性协同空间融合表征算法捕获两者潜在关联。与现有的预测模型相比,NLS-DRP模型在多个数据集上取得较好的耐药性预测性能。然而NLS-DRP模型尚未考虑生成任务,后续的研究将结合生成任务,通过将预测融入生成过程,为药物研发带来更多可能性。

参考文献:

[1] BHINDER B, GILVARY C, MADHUKAR N S, *et al*. Artificial Intelligence in Cancer Research and Precision Medicine[J]. *Cancer Discov*, 2021, **11**(4): 900-915. DOI: 10.1158/2159-8290.CD-21-0090.

[2] LOOMANS-KROPP H A, UMAR A. Cancer Prevention and Screening: The Next Step in the Era of Precision Medicine[J]. *NPJ Precis Oncol*, 2019, **3**: 3. DOI: 10.1038/s41698-018-0075-9.

[3] WANG X, ZHANG H Y, CHEN X Z. Drug Resistance and Combating Drug Resistance in Cancer[J]. *Cancer Drug Resist*, 2019, **2**(2): 141-160. DOI: 10.20517/cdr.2019.10.

[4] BAPTISTA D, FERREIRA P G, ROCHA M. Deep Learning for Drug Response Prediction in Cancer[J]. *Brief Bioinform*, 2021, **22**(1): 360-379. DOI: 10.1093/bib/bbz171.

[5] LIU C Y, WEI D, XIANG J, *et al*. An Improved Anticancer Drug-response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression[J]. *Mol Ther Nucleic Acids*, 2020, **21**: 676-686. DOI: 10.1016/j.omtn.2020.07.003.

[6] GAO Y N, LYU Q, LUO P, *et al*. Applications of Machine Learning to Predict Cisplatin Resistance in Lung Cancer[J]. *Int J Gen Med*, 2021, **14**: 5911-5925. DOI: 10.2147/IJGM.S329644.

[7] ZHANG H X, CHI M, SU D Q, *et al*. A Random Forest-based Metabolic Risk Model to Assess the Prognosis and Metabolism-related Drug Targets in Ovarian Cancer[J]. *Comput Biol Med*, 2023, **153**: 106432. DOI: 10.1016/j.compbiomed.2022.106432.

[8] YUAN S, CHEN Y C, TSAI C H, *et al*. Feature Selection Translates Drug Response Predictors from Cell Lines to Patients[J]. *Front Genet*, 2023, **14**: 1217414. DOI: 10.3389/fgene.2023.1217414.

[9] SHARMA A, RANI R. Ensembled Machine Learning Framework for Drug Sensitivity Prediction[J]. *IET Syst Biol*, 2020, **14**(1): 39-46. DOI: 10.1049/iet-syb.2018.5094.

[10] CHANG Y, PARK H, YANG H J, *et al*. Cancer Drug

Response Profile Scan (CDRscan): A Deep Learning Model that Predicts Drug Effectiveness from Cancer Genomic Signature[J]. *Sci Rep*, 2018, **8**: 8857. DOI: 10.1038/s41598-018-27214-6.

[11] LI M, WANG Y K, ZHENG R Q, *et al*. DeepDSC: a Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2021, **18**(2): 575−582. DOI: 10.1109/TCBB.2019.2919581.

[12] ZUO Z R, WANG P L, CHEN X W, *et al*. SWnet: a Deep Learning Model for Drug Response Prediction from Cancer Genomic Signatures and Compound Chemical Structures[J]. *BMC Bioinformatics*, 2021, **22**(1): 434. DOI: 10.1186/s12859-021-04352-9.

[13] LIU Q, HU Z Q, JIANG R, *et al*. DeepCDR: a Hybrid Graph Convolutional Network for Predicting Cancer Drug Response[J]. *Bioinformatics*, 2020, **36**(Suppl_2): i911−i918. DOI: 10.1093/bioinformatics/btaa822.

[14] LIU P F, LI H J, LI S, *et al*. Improving Prediction of Phenotypic Drug Response on Cancer Cell Lines Using Deep Convolutional Network[J]. *BMC Bioinformatics*, 2019, **20**(1): 408. DOI: 10.1186/s12859-019-2910-6.

[15] SHARIFI-NOGHABI H, ZOLOTAREVA O, COLLINS C C, *et al*. MOLI: Multi-omics Late Integration with Deep Neural Networks for Drug Response Prediction [J]. *Bioinformatics*, 2019, **35**(14): i501−i509. DOI: 10.1093/bioinformatics/btz318.

[16] JIANG L K, JIANG C Z, YU X Y, *et al*. DeepTTA: a Transformer-based Model for Predicting Cancer Drug Response[J]. *Brief Bioinform*, 2022, **23**(3): bbac100. DOI: 10.1093/bib/bbac100.

[17] HUANG K X, XIAO C, GLASS L M, *et al*. MolTrans: Molecular Interaction Transformer for Drug-target Interaction Prediction[J]. *Bioinformatics*, 2021, **37**(6): 830−836. DOI: 10.1093/bioinformatics/btaa880.

[18] BARRETINA J, CAPONIGRO G, STRANSKY N, *et al*. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity[J]. *Nature*, 2012, **483**(7391): 603−607. DOI: 10.1038/nature11003.

[19] KANEHISA M, GOTO S. KEGG: Kyoto Encyclopedia of Genes and Genomes[J]. *Nucleic Acids Res*, 2000, **28**(1): 27−30. DOI: 10.1093/nar/28.1.27.

[20] KOZALAK G, BÜTÜN İ, TOYRAN E, *et al*. Review on Bortezomib Resistance in Multiple Myeloma and Potential Role of Emerging Technologies[J]. *Pharmaceuticals*, 2023, **16**(1): 111. DOI: 10.3390/ph16010111.

[21] IORIO F, KNIJNENBURG T A, VIS D J, *et al*. A Landscape of Pharmacogenomic Interactions in Cancer[J]. *Cell*, 2016, **166**(3): 740−754. DOI: 10.1016/j.cell.2016.06.017.

[22] ZHOU J, CUI G Q, HU S D, *et al*. Graph Neural Networks: A Review of Methods and Applications[J]. *AI Open*, 2020, **1**: 57−81. DOI: 10.1016/j.aiopen.2021.01.001.

[23] ZHANG S, TONG H H, XU J J, *et al*. Graph Convolutional Networks: A Comprehensive Review[J]. *Comput Soc Netw*, 2019, **6**(1): 11. DOI: 10.1186/s40649-019-0069-y.

[24] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, *et al*. Graph Attention Networks[EB/OL]. arXiv Preprint: 1710.10903, 2017. https://arxiv.org/abs/1710.10903.