DSGAT：通过图注意力网络预测药物副作用的频率

Xianyu Xu, Ling Yue, Bingchun Li, Ying Liu, Yuan Wang, Wenjuan Zhang and Lin Wang

摘要

药物风险-效益评价的一个关键问题是确定药物副作用的发生频率。随机对照试验是获取不良反应频率的常规方法，但费时费力。因此，有必要用计算方法来引导轨迹。现有的药物副作用频率预测方法主要集中在建立药物-副作用相互作用图。这些方法固有的缺点是，它们的性能与交互的密度密切相关，但交互的密度是高度稀疏的。更重要的是，对于训练数据中没有出现的冷启动药物，这种方法无法学习到药物的偏好嵌入，因为相互作用图中没有与药物的链接。在这项工作中，我们提出了一种新的预测药物副作用频率的方法——DSGAT，即使用药物分子图代替常用的相互作用图。这导致了使用图注意力网络学习冷启动药物嵌入的能力。提出了一种新的损失函数，即加权 ε 不敏感损失函数，可以缓解稀疏性问题。在一个基准数据集上的实验结果表明，DSGAT 对冷启动药物产生了显著的改善。

关键词：副作用频率，化学结构，冷启动，深度学习，图形注意力网络

1 引言

药物风险-收益评价是指患者使用药物后所能获得的疗效与所承担的风险之间的评价，比如瑞德西韦、洛匹那韦-利托那韦治疗 COVID-19 的收益-风险评价。这项评价的核心问题就是确定药物副作用的发生频率。目前频率采集的标准方法是随机对照试验，即将研究对象随机分组，在不同组中实施不同的干预措施，对比不同的效果。但是由于时间的限制和样本量有限，一些副作用在临床试验中并没有被发现，直到上市多年之后才出现，因此，药物的副作用仍然是医疗卫生中疾病和死亡的主要原因。与此同时，副作用是导致药物退市的主要因素，导致药物研发失败，损失巨额资金。而目前的一些计算方法大多数只是预测给定药物中某些副作用是否存在，而不能预测副作用发生的频率；其他有一些方法可以预测已有相关信息的药物的副作用频率，但是无法对缺乏信息的新药物进行预测。因此现有的方法并不能很好应用于药物风险-收益评价。为了充分实现药物风险-收益评价，需要一种既能预测现有药物的副作用频率，也能预测新药物的副作用频率的方法。

2 相关工作

Galeano 等人提出了第一个利用非负矩阵分解[1]预测药物副作用频率的计算方法。他们的方法仅使用药物副作用频率矩阵，因此不能用于预测冷启动药物的副作用频率。最近，Zhao 等人提出了一种从多视图数据中预测药物副作用频率的图注意模型 MGPred，包括相似度、可用药物副作用频率和词嵌入[2]。他们的方法学习药物嵌入和药物-副作用相互作用图的副作用嵌入。尽管他们的方法以较低的 RMSE 预测了已知的药物副作用频率，但在实际使用中经常会出现假阳性，即未知的药物副作用关联往往被预测为频繁类别。

3 本文方法

3.1 本文方法概述

本文提出一种基于图注意网络 (GAT) 的药物副作用频率预测深度学习模型 DSGAT[3]。由于现有的药物副作用频率预测方法主要集中在建立药物-副作用相互作用图上。这些方法的固有缺点就是，它们的性能与交互的密度密切相关，但是实际上交互密度很稀疏；此外，对于新药物，由于作用图上没有新药物的信息，所以无法对新药物进行预测。

DSGAT 模型使用药物的分子图代替常用的药物-副作用相互作用图，使得可以使用 GAT 网络学习到新药物的 embedding。此外，作者还构建了副作用之间的关联图，GAT 通过这个关联图也可以学习到副作用的 embedding。最后作者提出了一种新的损失函数，加权 ε-不敏感损失函数，缓解了药物-副作用频率矩阵 M 稀疏性的问题。模型流程图如图 1所示：

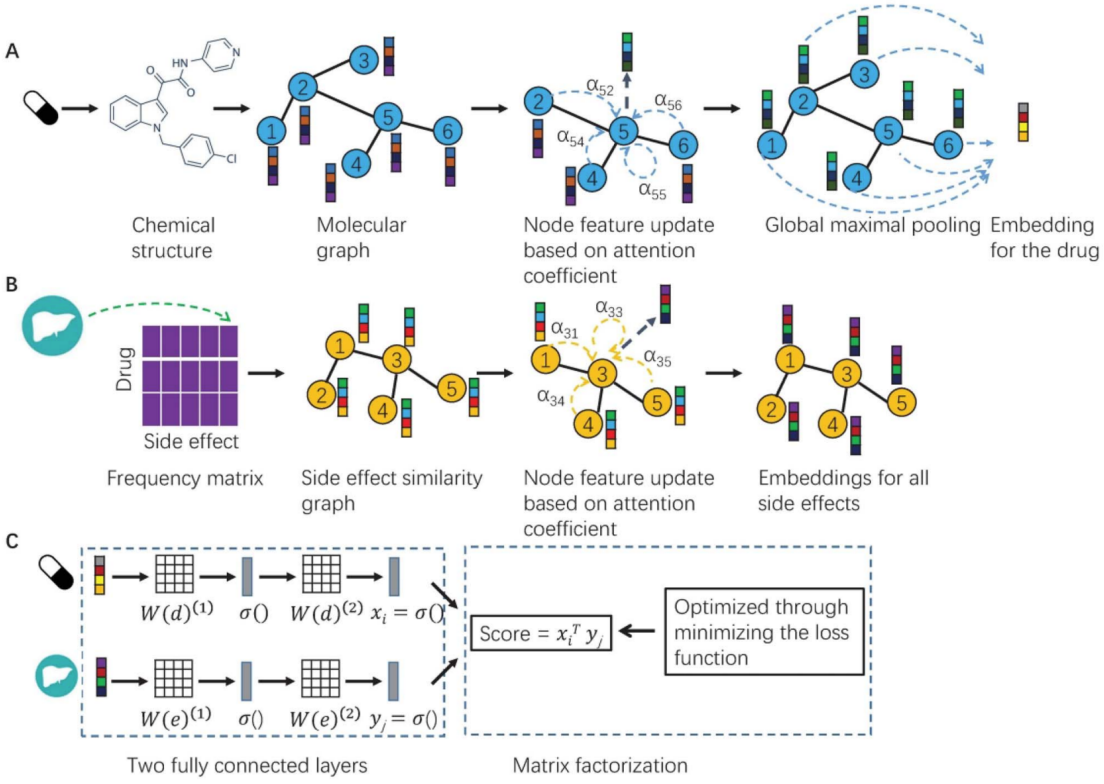

图 1: 模型 DSGAT 的流程图。(**A**) 对于每种药物，根据化学结构，使用 RDKit 包获得相应的分子图谱。结点和边分别代表化学原子和化学键。化合物中每个原子的性质被编码为一个多热矢量，它表示化学和拓扑属性。然后，将三层 GAT 网络应用于分子图，得到所有原子的嵌入。在此基础上，利用全局最大池化方法得到每种药物的嵌入。(**B**) 我们根据任何两个副作用对应频率谱的余弦相似度来测量它们之间的相似度，并使用 KNN 来构造副作用图。图的每个结点 (副作用) 都由 MedDRA 术语层次结构的前两个描述层编码。然后，将三层 GAT 网络应用于副作用图，得到所有副作用的嵌入。(**C**) 一旦获得任何一对药物和副作用的嵌入，我们通过分别添加两个完全连接的层将这些嵌入投影到一个共享空间。然后，我们使用矩阵分解作为解码器，即使用这些投影的内积作为预测得分，通过最小化损失函数来优化。

3.2 数据集

本文使用 Galeano 等人[1]和 Zhao 等人[2]中使用的基准数据集来验证作者的药物副作用频率预测方法的有效性。它包括 750 种药物和 994 种副作用以及 37 071 个已知频率项，这些项来自 SIDER 数据库 4.1 版本。将药物副作用的频率项分为 5 类，用整数编码: 非常罕见 (频率 = 1)、罕见 (频率 = 2)、不频繁 (频率 = 3)、频繁 (频率 = 4) 和非常频繁 (频率 = 5)。在频率项中，非常罕见、罕见、不频繁、频

繁和非常频繁项所占百分比分别为 3.21%、11.29%、26.92%、47.46% 和 11.12%。作者用评级矩阵 M 表示药物和副作用之间的频率，其中非零评级值表示特定药物副作用对的已知频率，否则为 0。评级矩阵 M 极其稀疏，非零元素仅占 4.97%。

3.3 图的表示

3.3.1 药物分子图

使用 (H_i,A_i) 这两个矩阵来表示药物 d_i 的分子图。其中，H_i 是特征矩阵，表示的是药物 d_i 中每一个原子的特征（使用多热向量表示，一共 109 维），H_i 是一个 $k_i \times 109$ 维的矩阵。k_i 是药物 d_i 的原子数。A_i 是邻接矩阵，表示的是药物 d_i 中原子之间的连接情况（化学键连接），A_i 是一个 $k_i \times k_i$ 维的矩阵。

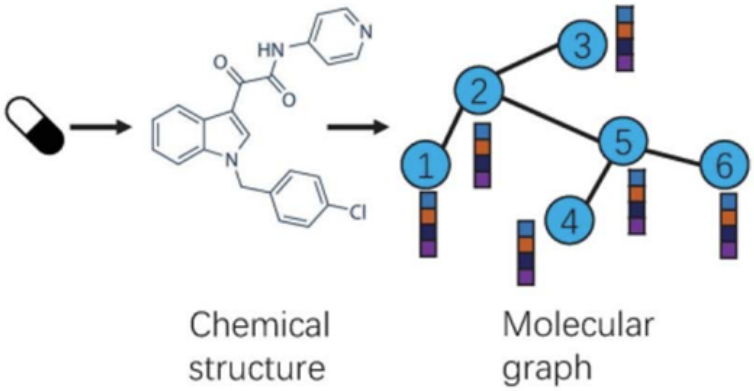

图 2: 药物分子图

3.3.2 副作用关联图

由于不同的副作用之间的连接情况不像药物的分子结构一样可以得知，所以需要先构建出不同副作用的邻接矩阵 A_e。作者使用余弦相似度来衡量不同副作用之间的距离，再结合 KNN，得出不同副作用的邻居，进而得到邻接矩阵 A_e。使用 (H_e ,A_e) 表示副作用关联图。其中，H_e 是特征矩阵，副作用的特征是选取了 MedDRA 术语层次结构的前两个描述级别进行表示。所有副作用一共有 243 个描述，所以，特征是一个 243 维的多热向量。H_e 是一个 $k_e \times 243$ 维的矩阵，k_e 是副作用的个数。A_e 是一个邻接矩阵，是一个 $k_e \times k_e$ 维的矩阵。

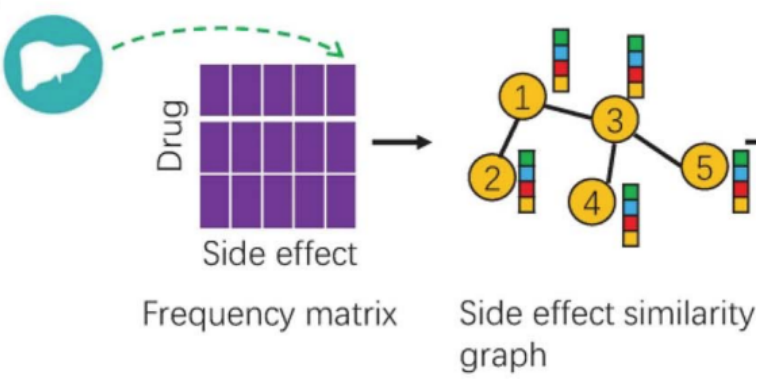

图 3: 副作用关联图

3.4 特征提取模块

作者使用 GAT 模型作为特征提取模块，GAT 网络提取药物 embedding 和副作用 embedding。GAT 模型在 (H_i,A_i) 药物分子图上应用，每一个原子都得到了综合邻居原子特征的新特征，把所有原子的新特征利用全局最大池化法进行融合，得到药物 d_i 的 embedding；GAT 模型在 (H_e,A_e) 副作用关联图上应用，每一个副作用都得到综合邻居副作用的新特征。

单层 GAT 只考虑使用邻居的信息，而多层 GAT 往往会带来更多的噪声。作者从 1,2,3,4 中选择 GAT 层数。作者根据 AUC 值来选取层数，当 GAT 层数设置为 3 时，可以获得较好的 AUC 值。除此之外，三层 GAT 层中，第一第二层采用了 10 头注意力机制，激活函数选用 ReLu 函数。

3.5 解码模块

当使用三层的 GAT 模型分别获得任何一对药物和副作用的 embedding 后，通过分别添加两个全连接层，将这些 embedding 投影到共享空间。然后，本文使用矩阵分解作为解码器，即使用这些投影的内积作为预测得分，通过最小化损失函数进行优化。

3.6 损失函数定义

本文提出了一个新的损失函数，叫做加权 ε 不敏感损失函数。作者采用了最小化偏好矩阵与标签矩阵之差的 Frobenius 范数的损失函数。与药物副作用对的数量相比，已知频率的数量要少得多，这使得频率矩阵 M 极其稀疏。

对于我们使用的数据集，只有 37071 个发现的药物副作用频率对，涵盖 750 种已知药物、9 种未知药物和 994 种副作用。为了克服极其稀疏这一点，对于未知的药物副作用关联，作者使用固定变量 ε= 0.5 控制预测评分和标签之间的差值。此外，已知的副作用关联在提高预测性能方面比未知的关联更值得信赖和更为重要，因此作者设置了一个调整权重 α= 0.03，以减少未知的药物副作用关联带来的影响。综上所述，我们可以得到加权 ε-不敏感损失函数如下:

$$\left\|I^{\Omega} \circ (S-M)\right\|_F^2 + \alpha \left\|I^0 \circ (S-\varepsilon A)\right\|_F^2$$

其中 I^{Ω} 和 I^0 是区分评级矩阵 M 中已知和未知项的映射函数，即如果 M_{ij}>0，则 I_{ij}=1，否则 I_{ij}=0；M_{ij}=0 时 I_{ij}^{O}=1，否则 I_{ij}^{O}=0, S 为预测得分矩阵，A 为全为 1 的矩阵，$|\cdot|_F$ 是 Frobenius 范数，∘表示两个矩阵的哈达玛乘积。

4 复现细节

4.1 与已有开源代码对比

本次复现的论文有提供源代码，本人在复现的过程中，参考了作者源代码进行复现。具体包括参考 Net.py 搭建了改进的 RGCN 模型；参考了 ics_750+9.py 实现了 750 种已知药物和 9 种未知药物的 994 种副作用频率预测，以及药物-副作用关联性和药物-副作用频率预测性能指标的生成。与源代码相比，本人在特征提取上做了一些改进，具体就是把药物特征提取使用的 GAT 模型更换成 RGCN 模型，通过药物-副作用关联性能指标和药物-副作用频率预测性能指标来分析改进后的优势。

4.2 实验环境搭建

- **python**：version 3.8

- **pytorch**：version 1.7.0+cu101
- **torch-geometric**：1.6.0
- **RDKit**：2022.9.1
- **networkx**：2.8.8

4.3 界面分析与使用说明

执行代码

- **ics_750+9.py**：759 种药物与 994 种副作用频率预测的执行程序

运行参数

- **model**：选取的模型
- **epoch**：epoch 的数量，默认值为 3000
- **lr**：学习率
- **tenfold**：使用十折交叉验证，默认是 True
- **执行命令事例**：python ics_750+9.py –tenfold –save model –model 1 –epoch 3000 –lr 0.0001

数据文件

- **750+9_frequency.mat**：750 种已知药物 +9 种未知药物与 994 种副作用的初始副作用频率矩阵
- **drug_SMILES_759.csv**：759 种药物的 SMILES 表示文件，也就是 759 种药物的化学表示
- **side_effect_label_750.mat**：994 种副作用的特征矩阵
- **blind_mask_mat.mat**：用于十折交叉验证进行数据选择的矩阵

4.4 创新点

在本次复现工作中，主要的工作体现在对药物特征提取上的思考和改进。本次复现的论文使用的是 GAT 模型对药物化学结构图中的每一个原子进行特征提取，主要的目的是想充分的考虑每一个原子与邻居原子（有化学键相连的）的特征联系。

但是，药物原子之间的化学键它不是单一的，化学键也有很多种，比如共和键，非共和键等。因此，药物的原子之间的关系他再只是简单的相连关系了，需要把原子之间的关系加进去一起考虑。所以我采用了关系图卷积神经网络充当特征提取器。此外，在 RGCN 模型参数的选取上，进行了多参数的对比实验，选取了 MAP 值、AUC 值、Spearman 系数、RMSE 值和 MAE 值综合表现最好的数据，具体的对比结果下一章呈现。

5 实验结果分析

模型对药物-副作用预测的性能主要从两方面来评判：识别药物-副作用关联的性能和预测药物-副作用频率的性能。在关联评价方面，采用 PR 曲线下面积 (AUPR)、所有药物的平均 AUPR 也被定义

为平均精密度 (MAP)、ROC 曲线下面积 (AUC)、归一化折损累计增益 (NDCG@N)、精度 Precision@N 和标准召回率 Recall@N 五个指标进行性能评价。因为我们通常对一些排名靠前的副作用感兴趣，所以我们使用 NDCG@N，precision@N 和 recall@N 来评估前 N 个推荐的表现，在我们的实验中，我们为 NDCG 设置 N = 10，为精确度和召回设置 N =1,15。在频率预测方面，我们使用均方根误差 (RMSE) 和平均绝对误差 (MAE) 作为评价指标。

实验采用十折交叉验证验证把数据划分成数据集和测试集。具体来说，所有药物被随机分为 10 个几乎相同大小的组。在交叉检验中，以其中一个子集的副作用频率作为测试集，其余 9 个子集的副作用频率构成训练集。

结果主要展示 RGCN 模型参数选择实验的结果和使用 RGCN 模型改进后的与 DSGAT 模型的对比实验的结果。RGCN 模型参数选择实验中参数主要有药物原子之间的关系数目和 RGCN 的层数。我们选择了 {4,5} 的关系数 r 和 {2,3} 的 RGCN 层数 l。

表 1: **RGCN 模型参数的选择**

(r,l)	MAP	AUC	NDCG	P@1	P@15	R@1	R@15	Spearman	RMSE	MAE
(4,2)	0.429	0.834	0.796	0.778	0.552	0.026	0.304	0.324	1.893	1.762
(4,3)	0.422	0.826	0.821	0.804	0.548	0.029	0.292	0.291	2.026	1.893
(5,2)	0.439	**0.895**	0.818	0.778	0.570	0.024	0.314	**0.374**	**1.672**	**1.332**
(5,3)	**0.443**	0.848	0.833	0.889	0.562	0.038	0.317	0.302	1.989	1.694

经过参数选择实验的结果对比，可以得出一个结论：关系数 r 的选取与药物-副作用的关联性能有关，当关系数较少时，提取出来的药物特征并不一定能够很好的代表药物，从而导致关联性能略差；而 RGCN 的层数 l 与药物-副作用频率预测性能有关，当层数过多时，预测的精度反而越差。综上比较分析后，决定选取关系数为 5，RGCN 层数为 2 的 RGCN 模型作为药物特征的提取器。

分别使用原文中的 GAT 模型作为特征提取器和关系数为 5 层数为 2 的 RGCN 模型作为特征提取器对同一批数据做药物-副作用频率预测实验，通过性能指标来检验改进后的优势，结果如表 2 所示。

表 2: **GAT 模型与 RGCN 模型作药物特征提取器的预测性能**

model	MAP	AUC	NDCG	P@1	P@15	R@1	R@15	Spearman	RMSE	MAE
GAT	0.409	0.824	0.819	0.667	0.548	0.021	0.323	**0.384**	1.714	1.334
RGCN	**0.439**	**0.895**	0.818	0.778	0.570	0.024	0.314	0.374	**1.672**	**1.332**

从上面表格的数据可以看出，总体上改进后的模型，在药物-副作用关联性能和药物-副作用预测精度性能上表现都比原来的使用 GAT 做特征提取器的模型更好，这也进一步说明药物原子之间的关系不是单一的连接关系，考虑药物特征时需要把药物原子之间的关系也考虑进去，才能更好的获得药物的特征表示，进而对后续的药物-副作用预测产生更好的影响。

6 总结与展望

在这篇文章中，作者提出了一种基于编码器-解码器框架的药物副作用频率预测方法 DSGAT。首先，用每种药物的天然分子图来表示每种药物，并应用三层 GAT 网络来获得药物的 embedding。其次，利用药物副作用频率矩阵，通过计算不同副作用之间的余弦相似度，我们获得副作用相似度，然后使

用每个副作用的前 k 个邻居来构建副作用图。随后，在副作用图上使用三层 GAT 网络进行副作用嵌入。最后，通过两层全连接网络将药物和副作用嵌入投影到共享空间中，并采用矩阵分解作为解码器。在分析整个模型的框架时，我发现药物分子图中每一个药物原子之间的联系不应该被简单的看作是单一的相连关系，而需要考虑更加复杂的关系，因此使用了 RGCN 模型作为药物的特征提取器，为了获得更好的结果，对 RGCN 模型的参数进行了选择，最后通过对比，得出使用 RGCN 模型作为特征提取器是可以在一定程度上提升药物-副作用关联性能和药物-副作用预测精度性能的。

DSGAT 能正确预测真实的副作用频率，有助于指导药物的风险-效益评价。目前，这种方法仅使用药物的化学结构，而整合异质数据，如药物靶点[4]、治疗指征[5]、微扰转录组数据[6]和微扰蛋白质组数据[7]，可能进一步提高预测性能。此外，DSGAT 中学习得到的 embedding 对药物和副作用的作用还没有完全研究，可以在未来研究它的生物学解释。

参考文献

[1] GALEANO D. Predicting the frequencies of drug side effects[J]. Nat Commun, 2020, 11(1): 4575.

[2] ZHAO H. A novel graph attention model for predicting frequencies of drug–side effects from multi-view data[J]. Brief Bioinform, 2021, 22(6): bbab239.

[3] XU X. DSGAT: predicting frequencies of drug side effects by graph attention networks[J]. Briefings in Bioinformatics, 2022, 23(2): bbab586.

[4] DS W. Drugbank 5.0: a major update to the drugbank database for 2018[J]. Nucleic Acids Res, 2018, 46(D1): D1074-82.

[5] AP D. Comparative toxi-cogenomics database (CTD): update 2021[J]. Nucleic Acids Res, 2021, 49(D1): D1138-43.

[6] A S. A next generation connectivity map: L1000 platform and the first 1,000,000 pro-files[J]. Cell, 2017, 171(6): 1437-52.e17.

[7] W Z. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines[J]. Cancer Cell, 2020, 38(6): 829-843.e4.