用于计算药物重定位的平衡矩阵分解

Xinxing Yang、Genke Yang 和 Jian Chu

arXiv:2301.06448v1 [cs.CE] 16 Jan 2023

摘要——**计算药物重定位旨在发现已上市药物的新用途。然而，现有模型存在以下局限。首先，在现实世界中，只有少数疾病有明确的治疗药物。这导致经验证的药物-疾病关联（正样本）与未经验证的药物-疾病关联（负样本）比例失衡，从而扰乱模型的优化梯度。其次，现有的药物表示未考虑药物的行为信息，导致无法全面建模药物的潜在特征。为解决上述不足，我们提出一种用于计算药物重定位的嵌入行为信息的平衡矩阵分解（BMF）。具体而言，在 BMF模型中，我们提出一种新颖的平衡对比损失（BCL），以优化计算药物重定位中的类别不平衡问题。BCL通过最大化目标药物与正疾病之间的相似度，并将目标药物与负疾病之间的相似度降至间隔以下，从而优化模型参数。此外，我们设计了一种利用药物行为信息增强药物表示的方法。在三个计算药物重定位数据集上进行的综合实验验证了上述改进的有效性。通过与七种基准模型的实验对比，也证明了BMF模型的优越性。**

索引词—**计算药物重定位、矩阵分解、类别不平衡、行为信息、平衡对比损失。**

[1]

I. 引言

药RUG物开发是一个漫长、成本高昂且容易失败的过程 [1]。例如，在传统的小分子药物开发中，从零开始为一种疾病设计出成功上市的药物，需要经历三个阶段：药物发现、临床前研究和临床研究 [2], [3]。每个阶段又细分为许多更小的阶段，例如，药物发现包括靶点识别、先导化合物设计等过程。上述每个步骤都充满不确定性，并耗费大量时间和金钱 [4], [5]。此外，当前的药物发现策略无法应对突发疫情，例如当前的 COVID-19 疫情。因此，我们需要一种新的药物发现技术，以加快药物开发并降低研发成本 [6]。

计算药物重定位旨在发现已上市或处于临床阶段的药物在原有适应症之外的新用途 [7]。与传统的药物开发策略不同，后者需要从头设计新药；计算药物重定位方法只需预先掌握大量已知的药物-疾病关联、药物的分子结构以及有关疾病的医疗信息，随后即可通过数据挖掘算法或人工智能算法推断已知药物的新用途[8]。这项技术的优势在于，它能省去部分传统药物开发流程，从而节省时间和资金。此外，由于已上市药物已通过严格的临床试验并获得美国食品药品监督管理局的批准，因此它们后续成功上市的可能性较高，降低了风险 [9], [10]。

当前的计算药物重定位方法包括基于矩阵分解的模型和基于图的模型[11]–[16]。基于图的模型首先构建药物-疾病关联网络，随后使用图算法挖掘该网络中的信息，以推断潜在的药物-疾病关联 [17]–[19]。Luo 等人 [20] 提出了一种基于随机游走算法的新型计算药物重定位模型。该模型首先构建一个包含药物和疾病的异构网络，其中，药物节点之间的边表示这两种药物之间的相似度，疾病节点之间的边表示这两种疾病之间的相似度，而药物节点与疾病节点之间存在边表示二者之间存在治疗关系，没有边则表示二者之间不存在治疗关系。随后使用双随机游走算法挖掘潜在的药物-疾病关联。Chen 等人 [21] 借鉴了跨域推荐和推荐系统中的迁移学习，同时解决计算药物重定位和药物-靶标关联预测问题。这种方法的优势在于，通过同时考虑两个关联网络，即药物-疾病关联网络和药物-靶标关联网络，可以学习到信息更丰富的药物表示，从而增强其在预测阶段的表达能力。Cai 等人提出了一种 [22] 基于图卷积网络的模型。该模型首先利用药物的分子信息、疾病的医疗信息以及已知的药物-疾病关联构建异构网络。随后，他们使用图卷积神经网络和一种新颖的特征提取模块，挖掘该网络中药物和疾病的拓扑信息，以

本研究部分由中国国家研发重点研究计划（2020YFB1711204）（2019YFB1705702）资助。（通讯作者：Genke Yang；电子邮箱：gkyang@sjtu.edu.cn。）

X. Yang、G. Yang 和 J. Chu 任职于中国上海市上海交通大学自动化系，邮编 200240（电子邮箱：yangxinx- ing@sjtu.edu.cn；gkyang@sjtu.edu.cn；chujian@sjtu.edu.cn）。

[1]预印本。研究仍在进行中。

获取药物或疾病的表示。为增强药物或疾病表示的表达能力，他们采用注意力机制对这些表示进行整合。后续实验结果验证了该模型的有效性。Ge 等人 [23] 设计了一种计算药物重定位模型，用于发现可能治疗 COVID-19 的潜在药物。此次重定位考虑了 6255 种药物，包括已上市药物、研发中药物和实验性药物。该方法首先利用公开数据构建知识图谱，其中包含药物、病毒靶点和人类靶点等节点类型。随后，利用基于图的挖掘算法，将原有的 6,255 种药物筛选至 212 种。最后，通过文献筛选等方法，推断出可能与 COVID-19 存在治疗关系的药物。

基于矩阵分解的模型 [24]–[26] 分为两个主要步骤。第一步是通过编码技术将药物或疾病映射到相应的嵌入空间，并用稠密数值向量表示药物或疾病，这称为药物或疾病的潜在因子。因此，基于矩阵分解的模型也可称为基于潜在因子的模型。随后，使用相似度函数计算药物潜在因子与疾病潜在因子之间的兼容性。最经典的矩阵分解模型 [27] 使用两个低秩矩阵分别表示药物和疾病的潜在因子，并将内积运算用作相似度函数。通过标注数据与预测值之间的损失函数训练这两个低秩矩阵，使其能够预测新的药物-疾病关联。Chen 等人 [28] 认为，现有方法受到数据稀疏性的影响，导致预测性能下降，而且现有融合策略无法有效融合信息。因此，他们提出了一种新的计算药物重定位模型。该模型利用深度自编码器对药物和疾病的结构数据进行特征提取，以获得更好的潜在因子。然后，根据多种数据计算药物和疾病的相似性信息。最后，通过自适应融合策略融合多种数据，得到预测的药物-疾病关联概率。由于矩阵分解模型不满足三角不等式，因而无法有效捕捉药物和疾病更细粒度的偏好，Luo 等人 [29]提出了一种基于协同度量学习的计算药物重定位新方法，能够更直观地表征药物与疾病之间的关联。该方法利用度量学习挖掘药物和疾病各自的潜在因子，然后使用这些潜在因子计算药物与疾病之间存在治疗关联的概率。与 Luo 等人的工作类似，Yang 等人 [26] 也提出了一种神经度量分解方法，通过解决矩阵分解模型不满足三角不等式的问题来预测潜在的药物-疾病关联。不同于 Luo 等人采用度量学习的方法，他们将药物和疾病视为嵌入空间中的点，并使用改进的欧几里得距离表示药物与疾病之间的关系。在他们的文章中，他们展示了点和欧几里得距离等表达方式如何符合三角不等式。

然而，现有研究工作存在以下问题。首先，在现实世界中，只有少数疾病有明确的治疗药物，而大多数疾病没有明确的治疗药物。因此，计算药物重定位数据集中经验证的药物-疾病关联（正样本）与未经验证的药物-疾病关联（负样本）的比例失衡，约为 1:100。正样本与负样本数量不平衡，会导致模型学习到有关失衡比例的先验信息，从而影响模型学习更本质关联规律的能力 [31]。这使得模型在进行预测时倾向于将未知药物-疾病关联样本预测为负样本，降低了模型的鲁棒性。这一现象的原因在于，正负样本数量失衡时，模型训练其所包含的参数，负样本的梯度会主导损失函数的优化方向。此外，负采样技术会采样大量负样本，其中会包含冗余样本（无用信息），模型的优化梯度会被这些无信息样本淹没，从而降低模型性能。最后，由于正负样本数量失衡，高置信度样本相对较多，最终主导了损失函数中的梯度。然而，高置信度样本对模型性能的提升微乎其微，模型应着重提升损失函数中低置信度样本的梯度。

其次，现有研究将药物的小分子结构、化学式等信息作为药物特征，却没有考虑药物的行为信息，因此无法全面建模药物的内在特征。在推荐系统领域，许多研究已证明，利用用户行为信息可以增强其表征能力，从而提升模型的泛化能力。因此，如何对药物的行为进行建模，是提升模型性能的关键。顺应这一趋势，我们受到计算机视觉和推荐系统领域相关问题解决方案的启发 [30]–[32]。

在这项工作中，我们提出了一种嵌入行为信息的平衡矩阵分解（BMF），用于计算药物重定位，以解决上述不足。具体而言，在BMF模型中，我们提出了一种新颖的平衡对比损失（BCL），以优化计算药物重定位中的类别不平衡问题。BCL通过最大化药物潜在因子与疾病潜在因子之间的相似度（两者存在治疗关系），并将药物潜在因子与疾病潜在因子之间的相似度（两者不存在治疗关系）最小化至低于间隔，从而优化模型中的参数。该损失函数的新颖之处在于，首先，BCL通过设定最小间隔过滤掉无信息的负样本，使其梯度为零，从而加快模型的

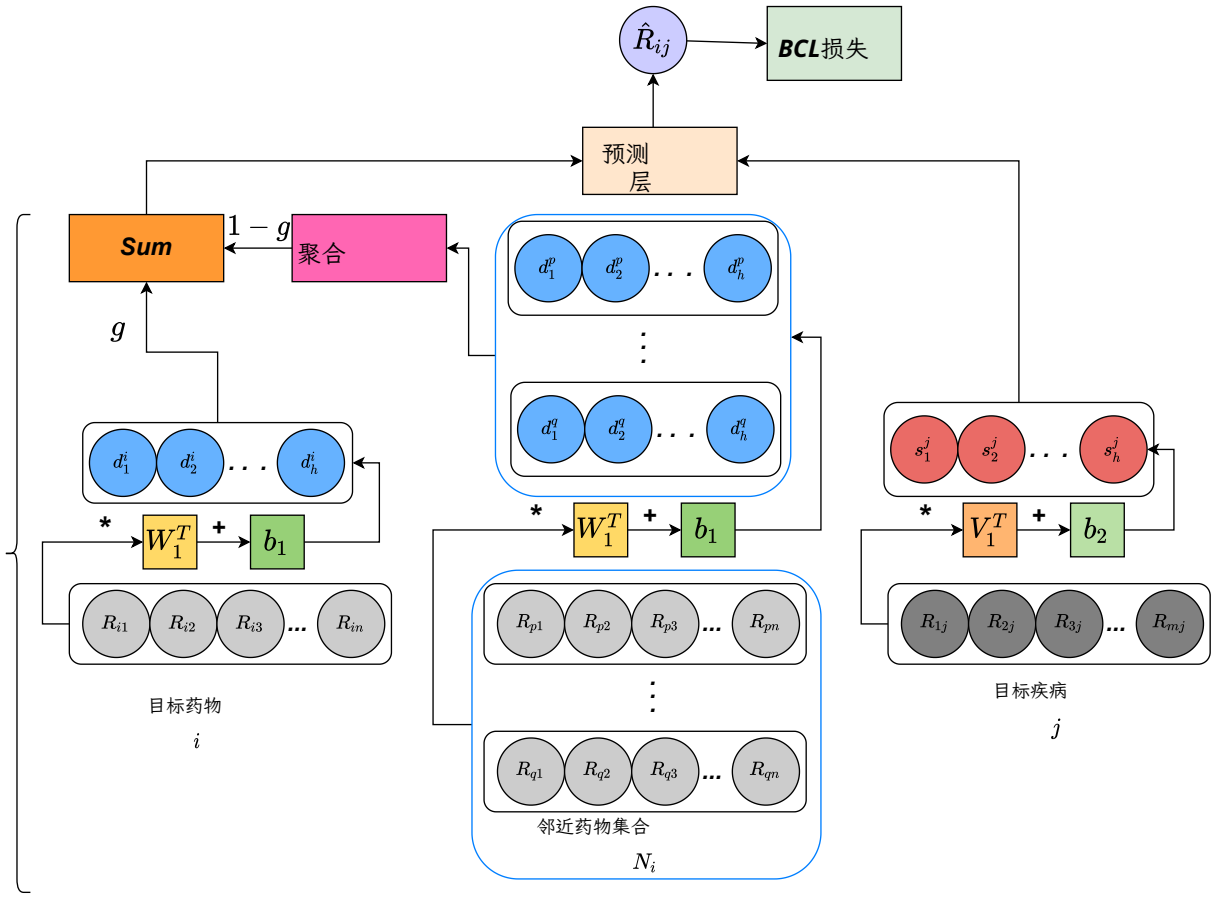


图 1. BMF模型的架构。

收敛，避免模型受到这些冗余样本的干扰。其次，BCL通过降低高置信度样本在损失函数中的权重来提升模型的泛化能力，使模型能够专注于提高低置信度样本的准确性。第三，BCL通过一个可控参数调整损失函数中正样本和负样本的权重，从而解决正负样本不平衡所引发的梯度问题。

此外，在 BMF模型中，我们设计了一种利用行为信息增强药物表示的方法。具体而言，在对目标药物 i 和目标疾病 j进行关联预测时，目标药物i 的表示会同时考虑药物和疾病两方面的行为信息。首先，用于挖掘药物 i 潜在因子的输入并非独热向量，而是该药物针对所有疾病的治疗记录向量（药物行为信息）。其次，所有与目标疾病 j 存在治疗关联的药物都被视为目标药物i的邻居，然后通过融合策略将目标药物 i 的表示与邻近药物的表示合并（疾病行为信息）。这样，多种行为信息便能嵌入目标药物 i的表示中。在三个真实数据集上开展的消融实验验证了上述策略的有效性。此外，BMF模型在三项评估指标上均优于当前最先进水平模型。

本工作的主要贡献如下：

1) 我们提出了一种嵌入行为信息的平衡矩阵分解（BMF），以增强药物表示，并将其应用于正负样本不平衡的计算药物重定位场景。2) 我们提出了一种新的平衡对比损失函数 BCL，用于解决计算药物重定位中的类别不平衡问题。具体而言，BCL 通过设置最小间隔来筛除无信息量的负样本，降低高置信度样本的损失权重，并调整正负样本的梯度权重，从而解决正负样本不平衡导致的准确率下降问题。3) 我们设计了一种利用行为信息增强药物表示的新方法。具体而言，在挖掘药物表示的过程中，我们还考虑了与药物相关的疾病信息以及邻近药物的表示信息，从而获得更好的药物表示。4) 通过在三个真实的计算药物重定位数据集上进行消融实验，我们验证了 BCL 和行为信息的有效性。我们还通过与六种流行的计算药物重定位模型进行对比，展示了 BMF模型的优越性能。

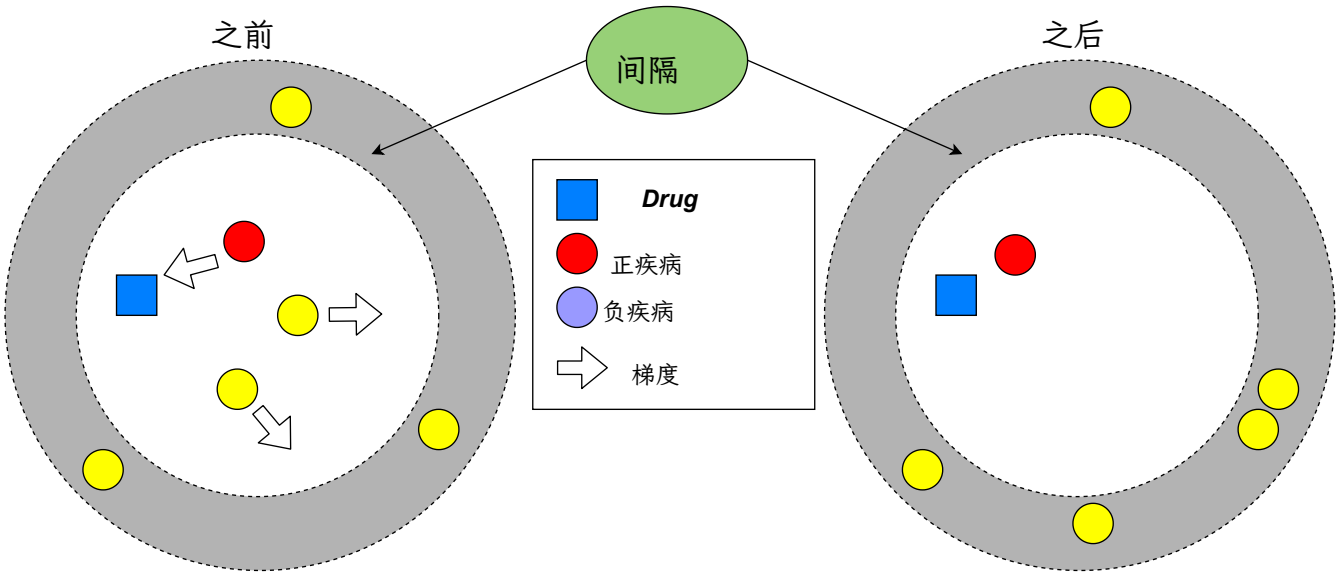


图2. BCL优化之后潜在因子（药物和疾病）的比较。

本文其余内容安排如下：第II节将介绍本文使用的数据集，第III节将介绍 BMF模型的详细内容。第IV节将介绍实验设置、消融实验和对比实验。第V节讨论本文的优缺点，并探讨未来的研究方向。

表I 本研究所用数据集的统计信息

数据集	药物	疾病	经验证的关联	稀疏度
Gottlieb	593	313	1933	98.95%
Cdataset	663	409	2532	99.06%
DNdataset	1490	4516	1008	99.98%

II. 数据集

我们首先介绍本研究所用数据集的统计信息。我们使用了三个公开可用的基准数据集：F数据集 [11], Cdataset [13] 和 DNdataset [20]。这些数据集在药物数量、疾病数量和经验证关联数量方面存在差异，因此能够全面验证模型的实际性能。如表I所示，F数据集、Cdataset 和 DNdataset 数据集中的药物数量分别为 593、663 和 1490，疾病数量分别为 313、493 和 4516，已知验证关联数量分别为 1933、2532 和 1008。

上述数据集中的所有药物均已获得美国食品药品监督管理局批准，其具体信息（包括化学结构式）可在 DrugBank 中查询 [33]。上述数据集中的所有疾病均可从 OMIM 数据库查询 [34]。值得注意的是，本文提出的方法不需要药物和疾病相似性信息[35], [36]。由于 DNdataset 中存在大量药物以及与药物之间尚无已知治疗关系的疾病，我们移除了对疾病没有任何治疗作用的药物，并且移除了没有任何药物可以治愈的疾病，最终在 DNdataset 中保留了 550 种药物和 360 种疾病。

III. 方法

本节将介绍本文提出的 BMF 模型，并讨论平衡对比损失函数和嵌入式行为信息。

A. 嵌入行为信息的平衡矩阵分解

在本研究中，我们提出了用于计算药物重定位的嵌入行为信息的平衡矩阵分解（BMF）。基于矩阵分解的方法使用独热编码对药物和疾病的潜在因子进行建模，然后使用相似度函数计算药物与疾病之间关联的概率。在训练阶段，模型使用预测数据与标注数据之间的误差来训练药物和疾病的潜在因子。在预测阶段，训练完成的药物和疾病潜在因子可以预测潜在药物与疾病之间是否存在关联。

图1展示了BMF模型的框架，该模型主要基于矩阵分解的思路进行扩展。从下到上，图中展示了挖掘目标药物 i 和目标疾病 j的潜在因子的过程，以及计算预测值 R_{ij} 的过程。BMF模型的创新之处在于，它在药物潜在因子的挖掘过程中嵌入了行为信息，并采用一种新颖的平衡对比损失来优化参数。

如公式 (1) 和 (2) 所示，对于药物 i 和疾病 j，BMF模型使用药物 i、$R_{i*} \in \mathbb{R}^{n\times 1}$ 的治疗行为以及疾病 $R_{*j} \in \mathbb{R}^{n\times 1}$ 的治疗行为作为各自的输入，而不是使用独热编码。其中，$R \in \mathbb{R}^{m\times n}$ 表示药物-疾病关联矩阵，m 表示药物的数量，n 表示疾病的数量，R_{i*} 是 R 的第 i 行，而 R_{*j} 是

R 的第 j 列。随后，将 R_{i*} 和 R_{*j} 输入单层神经网络，以提取药物 i 和疾病 j 的潜在因子。从本质上说，这一操作是将药物和疾病的行为信息嵌入其潜在因子中，以增强潜在因子的表达能力。

$$d^i = f(W_1^T R_{i*} + b_1) \tag{1}$$

$$s^j = f(V_1^T R_{*j} + b_2) \tag{2}$$

其中，$W_1^T \in \mathbb{R}^{h\times n}$ 和 $V_1^T \in \mathbb{R}^{h\times m}$ 是权重参数，h 是潜在因子向量的维度。$b_1 \in \mathbb{R}^{h\times 1}$和 $b_2 \in \mathbb{R}^{h\times 1}$ 是偏置参数。f 是激活函数，可以是 ReLU 或 Sigmod函数。$d^i \in \mathbb{R}^{h\times 1}$ 和$s^j \in \mathbb{R}^{h\times 1}$ 分别是药物 i 和疾病 j的潜在因子。

接下来，我们并非简单地计算药物 i 与疾病 j 的潜在因子之间的相似度，以此获得二者的兼容性。相反，目标药物i 的表示嵌入了目标疾病 j的行为信息。我们将所有与目标疾病 j 存在治疗关联的药物视为目标药物 i 的邻居集合，N_i。集合 N_i 包含 k 种疾病的潜在因子，k 的值可以手动设置。如公式（3）所示，我们首先使用平均池化聚合 N_i 中的所有潜在因子。其中，d^t 是邻近药物 t的潜在因子。o^i 是池化后包含行为信息的向量（目标疾病 j）。这种方法简单直接，但已被证明在多种场景下都有效。

$$o^i = \frac{1}{k}\sum_{t=1}^{k} d^t, \quad d^t \in N_i \tag{3}$$

然后，我们根据公式 (4) 将行为信息向量 o^i 与目标药物 i 的潜在因子结合起来。

$$h^i = gd^i + (1-g)o^i \tag{4}$$

其中，g 是手动定义的权重参数，其取值介于 $0-1$ 之间。通过这种方式，BMF模型成功地将多种行为信息（目标药物和目标疾病）嵌入目标药物的表示中 i (h^i)。最后，我们使用公式 (5) 预测药物 i 可以治疗疾病 j的概率。

$$\hat{R}_{ij} = g(W_2^T(h^i \odot s^j) + b_3) \tag{5}$$

其中，$W_2^T \in \mathbb{R}^{1\times h}$ 是权重参数，$\odot$ 表示^ 哈达玛积，R_{ij} 表示药物 i治疗疾病 j的概率。上述 BMF模型中包含的参数通过下一节所述的平衡对比损失进行优化。

B. 平衡对比损失

交叉熵损失函数是计算药物重定位领域中占主导地位的损失函数。但它存在类别不平衡问题，导致优化结果不佳。在本研究中，我们提出了平衡对比损失，以克服类别不平衡问题。给定经过验证的药物-疾病关联对（i，j），以及通过负采样技术采样得到的疾病集合 U，这些疾病与药物 i 不存在治疗关系。BCL 的数学公式如公式（6）所示。

$$\begin{aligned} BCL = &-\alpha(1-\hat{R}_{ij})^{\gamma}\log\hat{R}_{ij} \\ &+\sum_{u\in U}(1-\alpha)\hat{R}_{iu}^{\gamma}\max(0,-\log(1-\hat{R}_{iu}+c)) \end{aligned} \tag{6}$$

其中，R_{ij} 表示BCF模型的预测值，α 用于平衡损失函数中正负样本的权重比例。γ 是调整系数，用于降低高置信度样本的损失权重。c 是用于筛选信息量不足的负样本的间隔，其值通常设为 0 到 1 之间。具体而言，BCL通过最大化目标药物 i 与正疾病 j之间的相似度，并最小化目标药物 i 与某些负疾病之间的相似度，来优化BCL中的参数。

接下来，我们具体讨论 BCL 为什么能够解决类别不平衡问题。首先，对于已验证的药物-疾病关联（i，j），以往的工作会使用负采样技术挖掘大量负样本。随着样本数量增加，通常会产生许多冗余样本。这些样本不包含对模型训练有价值的信息，预测值通常也很低。然而，由于这类样本数量占优，其累积梯度会影响模型的优化方向，从而降低模型的泛化性能和收敛速度。为了解决这一问题，BCL 在损失函数中使用最小间隔 c 来筛除无用样本的梯度。当样本的预测值低于最小间隔 c 时，其损失函数将设为 0。这样，这些无用样本就不会干扰模型的优化方向。

其次，交叉熵损失函数对高置信度样本和低置信度样本一视同仁。正负样本的不平衡导致高置信度样本数量更多，它们占据了损失函数的大部分，而梯度方向则倾向于进一步提高高置信度样本的置信度。然而，这些高置信度样本对提升模型泛化性能的作用很小，模型应着重提高低置信度样本的置信度。具体而言，BCL 通过参数 γ 降低高置信度样本的损失权重，使模型聚焦于低置信度样本。值得注意的是，当 γ 的值为 1 时，BCL 等同于其他损失函数，即对高置信度样本和低置信度样本一视同仁。然而，当γ 的值为 2 时，如果 $\hat{r}_{iu}) = 0.01$, 0。01$_2$ 约等于 0.0001，而传统损失函数的损失权重为 1，损失权重便降低了近 1000 倍。这样一来，高置信度样本在损失函数中的权重占比降低，使模型能够专注于提升

低置信度样本的置信度，从而提升模型的泛化能力。

第三，传统损失函数将所有样本的权重都设为 1，但这种策略不适用于类别不平衡数据集。由于无治疗关系的药物-疾病关联数量过多，它们会主导优化梯度的方向，导致模型最终倾向于将待预测样本分类为无治疗关系的药物-疾病关联。在计算药物重定位问题中，我们应更加关注经验证的药物-疾病关联所带来的信息。因此，BCL 通过参数 α 调整损失函数中负样本的权重，使模型能够从不平衡数据中学习药物-疾病关联背后更本质的规律。

图2展示了训练前后潜在因子（药物和疾病）的变化。可以看出，BCL 最大化了具有治疗关联的药物与疾病之间的相似度。而且，最小间隔以下样本的梯度将被设为 0，因此不会干扰模型的收敛方向。算法1是 BMF模型的伪代码。

算法1BMF。

输入-1： 药物-疾病关联矩阵, R

输出： 药物 i 治疗疾病 j 的概率, $\hat{R}_{ij}$

对于 $(i,j) \in$ 采样的药物-疾病关联 **执行**

  利用行为信息挖掘潜在药物 i 的因子。

  $d^i \leftarrow f(W_1^T R_{i*} + b_1)$

  利用行为信息挖掘疾病的潜在因子 j。

  $s^j \leftarrow f(V_1^T R_{*j} + b_2)$

  聚合邻近药物的潜在因子。

  $o^i \leftarrow \frac{1}{k}\sum_{t=1}^{k} d^t, \quad d^t \in N_i$

  结合潜在因子。

  $h^i = gd^i + (1-g)o^i$

  正在计算预测值。

  $\hat{R}_{ij} = g(W_2^T(h^i \odot s^j) + b_3)$

  使用 BCL 优化 BMF模型中的参数。

  $BCL = -\alpha(1-\hat{R}_{ij})^\gamma \log \hat{R}_{ij}$

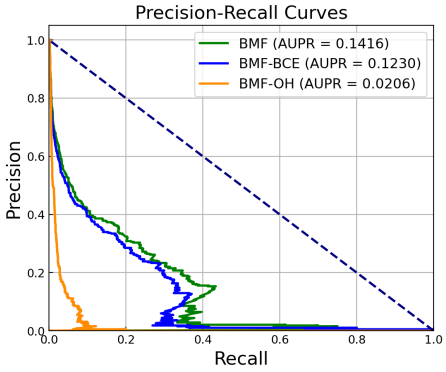
  $+\sum_{u\in U}(1-\alpha)\hat{R}_{iu}^\gamma \max(0, -\log(1-\hat{R}_{iu}+c))$

返回 $\hat{R}_{ij}$

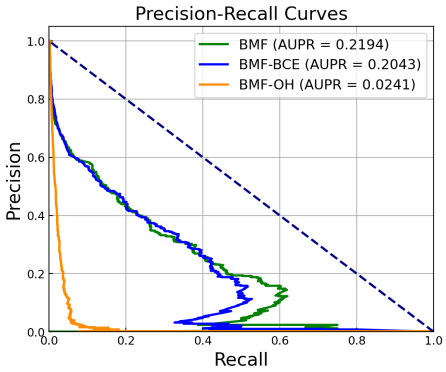
IV. 实验与讨论

本节讨论有关BMF模型的以下研究问题。

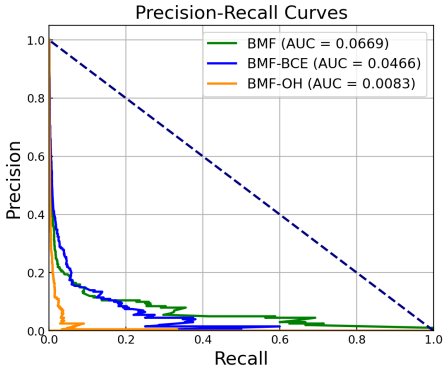
RQ1 将行为信息嵌入BMF模型以增强药物表示，是否有助于提升模型性能？**研究问题2** 与交叉熵损失函数相比，BCL能否提升模型在类别不平衡场景下的性能？**RQ3** BMF模型是否优于当前最先进模型?



(a) F数据集（精确率-召回率曲线）



(b) Cdataset（精确率-召回率曲线）



(c) DNdataset（精确率-召回率曲线）

图3. BMF、BMF-BCE和BMF-OH模型的精确率-召回率曲线。

A. 评估指标的设定

本研究使用三个公开可用的基准数据集评估 BMF模型。对于每个基准数据集，我们采用 5 折交叉验证来评估 BMF 模型及基准方法。所有模型均在预先划分的数据上进行训练和预测，以确保训练数据与测试数据的一致性，从而能够更公平地比较模型性能。我们使用三种常用的评估指标对所有模型的性能进行量化表示，分别是精确率-召回率曲线下面积（AUPR）、F1分数和命中率（HR）。

表II BMF模型和BMF-OH模型在三个数据集上的实验结果

数据集	Gottlieb			Cdataset			DNdataset		
	AUPR	F1分数	HR	AUPR	F1分数	HR	AUPR	F1分数	HR
BMF-OH	0.017	0.051	40%	0.019	0.044	46.4%	0.01	0.03	21.5%
BMF	0.166	0.247	60%	0.22	0.281	64.5%	0.057	0.115	33.9%

表III BMF模型和BMF-BCE模型在三个数据集上的实验结果

数据集	Gottlieb			Cdataset			DNdataset		
	AUPR	F1分数	HR	AUPR	F1分数	HR	AUPR	F1分数	HR
BMF-BCE	0.151	0.234	58.9%	0.214	0.275	63.2%	0.052	0.104	33%
BMF	0.166	0.247	60%	0.22	0.281	64.5%	0.057	0.115	33.9%

表 IV BMF模型与所有对比算法的实验结果

数据集	Gottlieb			Cdataset			DNdataset		
	AUPR	F1分数	HR	AUPR	F1分数	HR	AUPR	F1分数	HR
MF	0.033	0.01	22.1%	0.021	0.01	22.4%	0.005	0.003	10.5%
NCF	0.022	0.051	40.8%	0.023	0.046	45%	0.009	0.037	15.1%
MLP	0.021	0.05	40%	0.027	0.049	41.4%	0.007	0.038	14%
MetricF	0.039	0.073	43.7%	0.092	0.091	50.7%	0.006	0.028	21.9%
ADAE	0.021	0.63	37.6%	0.041	0.076	44.6%	0.039	0.054	25.4%
BMF-OH	0.017	0.051	40%	0.019	0.044	46.4%	0.01	0.03	21.5%
BMF-BCE	0.151	0.234	58.9%	0.214	0.275	63.2%	0.052	0.104	33%
BMF	0.166	0.247	60%	0.23	0.281	64.5%	0.057	0.115	33.9%

B. 超参数设置

BMF模型的超参数包括潜在因子的长度 h，该参数会影响药物或疾病潜在因子的表达能力；方程（3）中的邻域药物集合大小 N_i，该值越大，集合中包含的药物越多，信息越全面，但计算量也会增加；方程（4）中的参数 g，用于调节行为信息在药物表示中的权重。方程（6）中的参数 α、 γ 和 c，其中 α 用于调整损失函数中正样本和负样本的比例， γ 表示损失函数对高置信度样本的惩罚强度，c 表示最小筛选边界。上述超参数的取值范围分别为 [8,16,64,128,256,512]，[1,5,10,20], [0.1、0.3、0.5、0.7、0.9], [0.1、0.3、0.5、0.7、0.9], [1,2,3] 和 [0.001 、 0.01 、 0.1 、 0.2], 。

C. 行为信息的有效性（*RQ1*）

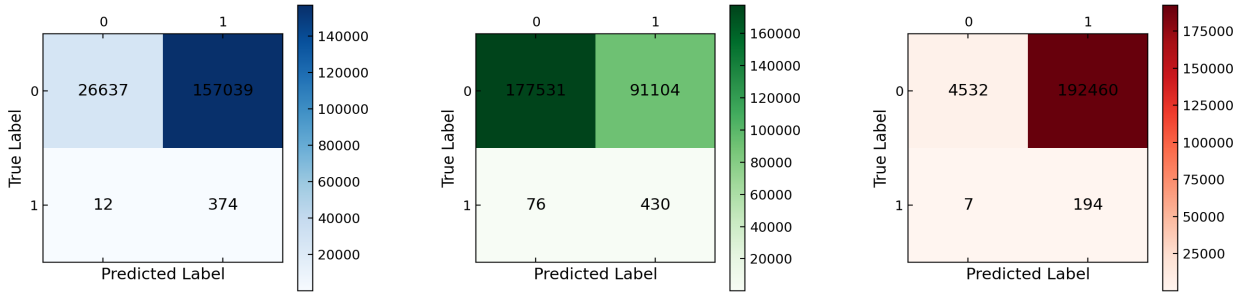
在本研究中，我们认为，将独热编码作为输入来挖掘药物的潜在因子，会降低模型的表达能力。独热编码仅包含药物的位置信息，缺乏对行为信息的建模。已有研究表明，行为信息能够有效提升推荐系统的模型准确率。因此，我们将药物的治疗行为作为输入来挖掘其潜在因子，并额外纳入邻近药物的信息。为验证上述行为信息能否优于传统的独热编码，我们通过实验比较了 BMF 模型与其变体 BMF-OH ； 后者以独热编码作为输入。

表II展示了BMF和BMF-OH在三个数据集上的实验结果。实验结果直观地表明，BMF模型在所有三个数据集上的表现均优于BMF-OH模型。具体而言，在F数据集中，BMF-OH的精确率-召回率曲线下面积为0.017，而BMF为0.166，提升了876%；BMF-OH的F1分数为0.051，而BMF为0.247，提升了384%；BMF-OH的命中率为40%，而BMF的命中率为60%，提升了49.9%。此外，在Cdataset和DNdataset上的平均提升幅度分别为402%和48.8%。

通过分析BMF模型和BMF-OH的实验结果，我们可以得出结论：对药物的行为信息进行建模，能够增强其潜在因子的表达能力，从而提高模型的预测准确率。

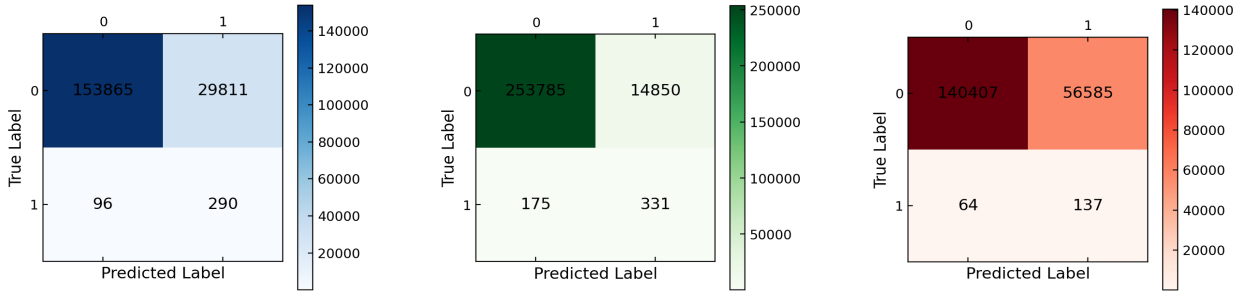
D. 平衡对比损失的有效性（研究问题*2*）

以往的损失函数无法有效解决计算药物重定位中的类别不平衡问题，因此我们在本研究中提出了平衡对比损失（BCL）。BCL 的优势在于，它更加重视低置信度样本，使用最小间隔筛除无信息样本，并动态调整损失函数中正样本和负样本的权重。为了



(a) F数据集 (b) Cdataset (c) DNdataset

图 4. 阈值为 = 0时 BMF 模型的混淆矩阵。1



(a) F数据集 (b) Cdataset (c) DNdataset

图5. 阈值为 = 0.3时BMF模型的混淆矩阵

研究 BCL 的有效性，我们设置了以下对比实验。该实验比较了 BMF 与其变体 BMF-BCE，二者的区别在于后者使用二元交叉熵损失函数。

表III展示了BMF和BMF-BCE在三个数据集上的实验结果。上述实验结果直观地表明，BMF模型在所有三个数据集上的表现均优于BMF-BCE模型。具体而言，在F数据集、C数据集和DNdataset上的性能提升分别为5.7%、3.9%和7.6%。

通过分析上述实验结果，我们可以轻松得出结论：与二元交叉熵损失函数相比，BCL能够更好地解决计算药物重定位中的类别不平衡问题。

E. 基线比较（*RQ3*）

以下是BMF模型与当前流行模型在三个公开数据集上的全面性能比较。为突出BMF模型的优势，我们对所有模型采用相同的数据集划分方法和评估指标。实验中所用基准模型的说明如下。

- 矩阵分解 [27]：矩阵分解模型（MF）使用药物和疾病的潜在因子，通过内积运算计算药物与疾病之间的治疗概率。
- 神经协同过滤 [37]：神经协同过滤模型（NCF）对药物与疾病的潜在因子进行哈达玛积运算，再将结果输入神经网络，以计算该药物与疾病之间的治疗概率。
- 多层感知机 [38]：多层感知机（MLP）由多个神经网络组成，能够有效捕捉药物与疾病之间的非线性关联。
- *MetricF* [39]：度量因子分解模型（MetricF）将药物和疾病的潜在因子视为欧几里得空间中的点，随后使用欧几里得距离表示药物与疾病之间的关系。
- *ADAE* [40]：附加去噪自编码器（ADAE）基于优质潜在因子能够还原原始输入这一理念，挖掘药物和疾病的潜在因子，并进一步考虑药物和疾病的相似性信息。随后，通过内积运算计算药物与疾病之间的治疗关系概率。

表 IV 展示了 BMF模型与基准模型在这三个数据集上的实验结果。我们可以得出以下几个明显结论。首先，我们发现BMF模型在 F数据集、Cdataset 和 DNdataset 上均取得了最佳预测结果。具体而言，BMF模型在 F数据集上的精确率-召回率曲线下面积、F1分数和命中率指标分别为 0.166、0.247 和 60%。BMF模型在 Cdataset 上的精确率-召回率曲线下面积、F1分数和命中率指标分别为 0.23、0.281 和 64.5%。BMF模型在 DNdataset 上的精确率-召回率曲线下面积、F1分数和命中率指标分别为

0.057、0.115 和 33.9%。

此外，作为基于潜在因子模型的基础，矩阵分解的预测性能最差。神经协同过滤的结果优于矩阵分解模型，因为它考虑了潜在因子向量中不同维度的权重。MetricF模型的实验结果也优于矩阵分解，这进一步验证了以点和距离进行表示优于以内积和向量进行表示，弥补了后者无法满足三角不等式的缺陷。

此外，图3展示了BMF、BMF-BCE和BMF-OH模型的精确率-召回率（P-R）曲线。我们发现，根据P-R曲线，BMF模型优于BMF-BCE模型。P-R曲线是二分类问题中主流的评估指标，能够公平地比较模型在类别不平衡数据集上的性能。因此，BMF模型和BMF-BCE模型在P-R曲线下的表现，验证了本文提出的平衡对比损失的有效性。另一方面，根据P-R曲线，BMF模型的性能优于BMF-OH模型。这说明了本文提出的行为信息的有效性。

图4和图5展示了 BMF 在不同阈值下的混淆矩阵。一方面，我们发现，当阈值为 0.1 时，BMF模型正确分类正样本的概率较高。但同时，它也更有可能将真实负样本预测为正样本。另一方面，当阈值为 0.3 时，BMF模型正确分类正样本的概率较低，但正确分类负样本的概率较高。因此，我们需要动态调整阈值，使模型的预测结果满足实际场景的需求。

经过上述讨论，我们可以发现，BMF模型在预测性能方面处于领先地位，这表明使用 BCL 能够更好地应对计算药物重定位中的类别不平衡问题，同时，对药物行为信息进行建模也能获得更好的药物表示。这些发现直接表明了本研究中这两项改进措施的有效性。

V. 结论

在本研究中，我们提出了一种用于计算药物重定位的嵌入行为信息的平衡矩阵分解（BMF），以解决上述不足。BMF模型利用行为信息增强药物表示，并提出了BCL损失函数，以改善类别不平衡问题下模型的优化方向，使模型参数能够更好地收敛。我们通过在三个真实数据集上进行消融实验，验证了这些创新的有效性。与主流基准模型的比较也证实了BMF模型的优越性。然而，作为一种基于矩阵分解的模型，BMF没有考虑邻域结构信息，因此其泛化能力受到限制。因此，在未来的工作中，我们将探索如何将邻域结构信息嵌入BMF模型。

参考文献

[1] M. Dickson 和 J. P. Gagnon，“新药发现与开发成本不断攀升的关键因素”，Nature Reviews Drug discovery，第 3 卷，第 5 期，第 417–429 页，2004 年。[2] N. A. Tamimi 和 P. Ellis，“药物开发：从概念到市场！”Nephron Clinical Practice，第 113 卷，131，2009 年 3，125 期，第 c – c 页。[3] T. T. Ashburn 和 K. B. Thor，“药物重定位：为现有药物识别并开发新用途”，Nature Reviews Drug Discovery，第 3 卷，第 8 期，第 673–683 页，2004 年。[4] N. Nosengo，“能让老药焕发新用吗？”Nature News，第 534 卷，第 7607 期，第 314 页，2016 年。[5] J.-L. E. Pritchard、T. A. O’Mara 和 D. M. Glubb，“通过遗传学提升药物重定位的前景”，Frontiers in Pharmacology，第 8 卷，第 896 页，2017 年。[6] M. Lotfi Shahreza、N. Ghadiri、S. R. Mousavi、J. Varshosaz 和 J. R. Green，“基于网络的药物重定位方法综述”，Briefings in Bioinformatics，第 19 卷，第 5 期，第 878–892 页，2018 年。[7] B. Padhy、Y. Gupta 等，“药物重定位：重新研究现有药物以寻找新的治疗适应症”，Journal of Postgraduate Medicine，第 57 卷，第 2 期，第 153 页，2011 年。[8] S. S. Sadeghi 和 M. R. Keyvanpour，“计算药物重定位的分析性综述”，IEEE/ACM Transactions on Computational Biology and Bioinformatics，第 18 卷，488，2019 年第 2 期，第 472–页。[9] J. K. Yella、S. Yaddanapudi、Y. Wang 和 A. G. Jegga，“计算药物重定位的趋势变化”，Pharmaceuticals，第 11 卷，第 2 期，第 57 页，2018 年。[10] H. Luo、M. Li、M. Yang、F.-X. Wu、Y. Li 和 J. Wang，“用于药物重定位的生物医学数据与计算模型：全面综述”，Briefings in Bioinformatics，第 22 卷，第 2 期，第 1604–1619 页，2021 年。[11] A. Gottlieb、G. Y. Stein、E. Ruppin 和 R. Sharan，“Predict：一种推断新药物适应症的方法及其在个性化医疗中的应用”，Molecular Systems Biology，第 7 卷，第 1 期，第 496 页，2011 年。[12] J. Yang、D. Zhang、L. Liu、G. Li、Y. Cai、Y. Zhang、H. Jin 和 X. Chen，“基于子结构与适应症之间关系的计算药物重定位”，Briefings in Bioinformatics，第 22 卷，第 4 期，第 bbaa348 页，2021 年。[13] H. Luo、M. Li、S. Wang、Q. Liu、Y. Li 和 J. Wang，“利用低秩矩阵近似和随机化算法进行计算药物重定位”，Bioinformatics，第 34 卷，第 11 期，第 1904–1912 页，2018 年。[14] M. Yang、G. Wu、Q. Zhao、Y. Li 和 J. Wang，“基于多相似性双线性矩阵分解的计算药物重定位”，Briefings in Bioinformatics，第 22 卷，第 4 期，第 bbaa267 页，2021 年。[15] X. Yang、Y. Liu、J. He 等，“用于计算药物重定位的附加神经矩阵分解模型”，BMC Bioinformatics，第 20 卷，第 1 期，第 1–11 页，2019 年。[16] J. He、X. Yang、Z. Gong 等，“用于计算药物重定位的混合注意力记忆网络”，BMC Bioinformatics，第 21 卷，第 1 期，第 1–17 页，2020 年。[17] W. Wang、S. Yang、X. Zhang 和 J. Li，“通过异构网络模型整合靶标信息进行药物重定位”，Bioinformatics，第 30 卷，第 20 期，第 2923–2930 页，2014 年。[18] X. Zeng、S. Zhu、X. Liu、Y. Zhou、R. Nussinov 和 F. Cheng，“deepdr：一种基于网络的深度学习计算机模拟药物重定位方法”，Bioinformatics，第 35 卷，第 24 期，第 5191–5198 页，2019 年。[19] Y. Meng、C. Lu、M. Jin、J. Xu、X. Zeng 和 J. Yang，“一种用于药物重定位的加权双线性神经协同过滤方法”，Briefings in Bioinformatics，2022 年。[20] H. Luo、J. Wang、M. Li、J. Luo、X. Peng、F.-X. Wu 和 Y. Pan，“基于综合相似性度量和双随机游走算法的药物重定位”，Bioinformatics，第 32 卷，第 17 期，第 2664–2671 页，2016 年。[21] H. Chen、F. Cheng 和 J. Li，“idrug：通过跨网络嵌入整合药物重定位与药物-靶标预测”，PLoS Computational Biology，第 16 卷，第 7 期，第 e1008040 页，2020 年。[22] L. Cai、C. Lu、J. Xu、Y. Meng、P. Wang、X. Fu、X. Zeng 和 Y. Su，“基于异构信息融合图卷积网络的药物重定位”，Briefings in Bioinformatics，第 22 卷，第 6 期，第 bbab319 页，2021 年。[23] Y. Ge、T. Tian、S. Huang、F. Wan、J. Li、S. Li、X. Wang、H. Yang、L. Hong、N. Wu 等，“一种综合药物重定位框架发现了靶向 COVID-19 的潜在治疗药物”，Signal Transduction and Targeted Therapy，第 6 卷，第 1 期，第 1–16 页，2021 年。[24] W. Zhang、H. Xu、X. Li、Q. Gao 和 L. Wang，“Drimc：一种采用贝叶斯归纳矩阵补全的改进型药物重定位方法”，Bioinformatics，第 36 卷，第 9 期，第 2839–2847 页，2020 年。

[25] Y. Yan、M. Yang、H. Zhao、G. Duan、X. Peng 和 J. Wang，“基于矩阵补全的多视图学习药物重定位”，*Briefings in Bioinformatics*，第 23 卷，第 3 期，第 bbac054 页，2022 年。[26] X. Yang 、 G. Yang 和 J. Chu，“用于计算药物重定位的神经度量分解”，*IEEE/ACM Transactions on Computational Biologyand Bioinformatics*，2022 年。[27] Y. Koren，“分解与邻域相遇：一种多方面的协同过滤模型”，发表于第 *14* 届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，2008 年，第 426–434 页。[28] P. Chen、T. Bao、X. Yu 和 Z. Liu，“一种基于深度自编码器和自适应融合的药物重定位算法”，*BMCBioinformatics*，第 22 卷，第 1 期，第 1–18 页，2021 年。[29] H. Luo、J. Wang、C. Yan、M. Li、F.-X. Wu 和 Y. Pan，“一种基于协同度量学习的新型药物重定位方法”，*IEEE/ACM Transactions on Computational Biology and Bioinformatics*，第 18 卷，第 2 期，第 463–471 页，2019 年。[30] K. Mao、J. Zhu、J. Wang、Q. Dai、Z. Dong、X. Xiao 和 X. He，“Simplex：一种简洁且强大的协同过滤基线方法”，发表于第 *30* 届 *ACM*信息与知识管理国际会议论文集，2021 年，第 1243–1252 页。[31] T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Dollár，“用于密集目标检测的焦点损失”，发表于 *IEEE*国际计算机视觉会议论文集，2017 年，第 2980–2988 页。[32] F. Wang 和 H. Liu，“理解对比损失的行为”，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021 年，第 2495–2504 页。[33] C. Knox、V. Law、T. Jewison、P. Liu、S. Ly、A. Frolkis、A. Pon、K. Banco、C. Mak、V. Neveu 等，“Drugbank 3.0：面向药物组学研究的综合资源”，《核酸研究》，第 39 卷，增刊 1，第 D1035–D1041 页，2010 年。[34] A. Hamosh、A. F. Scott、J. S. Amberger、C. A. Bocchini 和 V. A. McKusick，“人类孟德尔遗传在线数据库（OMIM）：人类基因与遗传疾病知识库”，《核酸研究》，第 33 卷，增刊 1，第 D514–D517 页，2005 年。[35] C. Steinbeck、Y. Han、S. Kuhn、O. Horlacher、E. Luttmann 和 E. Willighagen，“化学开发工具包（CDK）：一个用于化学信息学和生物信息学的开源 Java 库”，《化学信息与计算机科学杂志》，第 43 卷，第 2 期，第 493–500 页，2003 年。[36] D. Weininger，“SMILES，一种化学语言和信息系统。1. 方法学与编码规则简介”，《化学信息与计算机科学杂志》，第 28 卷，第 1 期，第 31–36 页，1988 年。[37] X. He 、 L. Liao 、 H. Zhang 、 L. Nie 、 X. Hu 和 T.-S. Chua，“神经协同过滤”，发表于第 *26* 届万维网国际会议论文集，2017 年，第 173–182 页。[38] Y. LeCun、Y. Bengio 和 G. Hinton，“深度学习”，《自然》，第 521 卷，第 7553 期，第 436–444 页，2015 年。[39] S. Zhang、L. Yao、B. Wu、X. Xu、X. Zhang 和 L. Zhu，“利用分解揭示用于推荐的度量向量空间”，*IEEE*工业信息学汇刊，第 16 卷，742，2019 年第 2 期，第 732–页。[40] X. Dong、L. Yu、Z. Wu、Y. Sun、L. Yuan 和 F. Zhang，“一种具有深层结构的推荐系统混合协同过滤模型”，发表于*AAAI*人工智能会议论文集，第 31 卷，第 1 期，2017 年。