# 基于知识图谱嵌入的阿尔茨海默病药物重定位

摘要

TODO

**关键词：**药物重定位；阿尔茨海默病；知识图谱；知识图谱嵌入；知识图谱补全

## 引言

知识图谱（Knowledge Graph，KG）是一种基于图结构存储知识的数据库，主要用于网页搜索、问答系统、推荐任务[1]。知识中的具体事物和抽象概念在 KG 中被表示为实体（entity），实体之间的联系被表示为关系（relation），进而知识被表示成格式为（头实体，关系，尾实体）的三元组。KG是一个由大量的三元组组成的有向图结构，KG中的节点（node）代表上面的实体，边（edge）表示实体间的关系。

然而，许多KG都非常巨大，如药物再利用知识图谱（Drug Repurposing Knowledge Graph，DRKG）[2]包含97238个实体和5874261个三元组。如此巨大的KG，直接像关系数据库一样进行检索和多步推理非常耗时，无法达到现代应用的要求。而且知识图谱具有非常严重的长尾现象，有很多实体与其他实体间仅仅具有很少的关系，对于这些长尾实体，往往很难理解其含义，进而影响对其的推理。因此，如何表示KG并进而将其与深度学习联合是一个近些年热门的领域。

知识图谱嵌入（Knowledge Graph Embedding, KGE）是一种将实体和关系表示成低维稠密向量的技术，在这个向量空间内，语义相似的对象之间的距离很近。在过去几年中，研究人员提出了很多 KGE模型来学习实体和关系嵌入向量，包括但不限于TransE[3]、TransR[4]、RESCAL[5]、DistMult[6]、ComplEx[7]、RotatE[8]等。

阿尔茨海默病（Alzheimer’s disease，AD）是一种常见的神经退行性疾病，无法治愈且不可逆转[9]，其特征是伴有神经精神症状的渐进性严重痴呆[10]。中国阿尔茨海默病报告2021 显示我国 60 岁及以上人群中有 983 万例 AD 患者[11]，并且另一份研究报告称，我国 AD 患者的 2015 年治疗费用为 1 677.4 亿美元，到2050 年将高达 18871.8 亿美元[12]，这充分说明了AD给社会带来了巨大的经济负担。因此，AD 的治疗药物开发迫在眉睫。然而，早在 2015 年，开发一种新药就需要花费26 亿美元和10-12 年[13-14]。药物重定位是指将现有药物的适应症拓宽到其他疾病，从而能大大节省成本并缩短新药开发周期。

最近，研究人员提出了很多利用知识图谱进行药物重定位的方法。Zeng等[15]建立了一个1500万个三元组的综合知识图谱，包括药物、基因、疾病、蛋白质、基因4种实体以及它们之间的 39 种关系，然后利用RotatE模型学习实体和关系的表示，进而确定了41种针对 COVID-19 的治疗药物。Zhang等[16]提出了一种基于神经网络和文献发现的方法，首先利用PubMed和其他专注COVID-19的研究文献构建了一个生物医学知识图谱，然后利用多种 KGE 模型预测候选药物，并利用发现模式解释了KGE预测的合理性。目前也有研究人员利用KGE研究帕金森病的药物重定位，并取得了不错的效果[17]。

Wang等[14]提出了一种基于知识图谱的深度学习方法用于 AD 药物重定位。首先，利用 DistMult学习了构建的阳性药物靶点对知识图谱的实体和关系的嵌入表示，然后利用一个Conv-Conv模块来提取药物-靶点对的特征，提取到的特征传入到一个全连接网络进行二分类，最终通过载脂蛋白E作为靶点寻找治疗AD 的药物。Nian等[9]从文献中构建一个知识图谱，利用 TransE，DistMult和ComplEx 学习并预测有助于AD治疗或预防的候选物质，以研究 AD与化学物质，药物和膳食补充剂之间的关系，进而确定预防或延缓神经退行性进展的机会。

TODO

## 方法

我们首先利用多种KGE模型（TransE、DistMult、ComplEx和RotatE）在DRKG上学习实体和关系的嵌入向量，通过5种经典的知识图谱嵌入评价指标选择了RotatE模型作为我们最终的药物重定位模型；然后，在整个知识图谱上训练 RotatE 模型，并利用多种嵌入向量分析手段评价我们学习到的嵌入向量的质量；最终，通过链接预测技术找到10种治疗AD的候选药物。

### 数据

DRKG[2]是一个涉及基因、药物、疾病、生物过程、副作用和症状的综合生物知识图谱，包括来自DrugBank、Hetionet、GNBR、String、IntAct 和 DGIdb 等六个现有数据库的信息，以及从最近发表的 Covid19 出版物中收集的数据。它有属于13种实体类型的97238个实体；以及属于107种关系类型的5874261个三元组。此外它还提供了一个TransE\_l2的预训练KGE模型。

### 模型

在本小节中，我们首先介绍一些KG和KGE数学符号，这些符号将应用于本研究的其他部分。表1列出了文章后续部分的概念和定义。

**Table 1** The Notation and Definition of KG and KGE.

|  |  |  |
| --- | --- | --- |
| Name | Notation | Definition |
| Knowledge Graph | **T** | A set of triplets in the form **(h,r,t)** |
| Entity Set | **E** | The set of entities |
| Relation Set | **R** | The set of relations |
| Entity Vectors | **h, t** | The entity vectors for head entity **h** and tail entity **t** |
| Relation Vector | **r** | The vector for relation **r** |
| Scoring Function | **f(h, r, t)** | The scoring function for the triple **(h, r, t)** |
| Loss Function |  | The loss function of the model |

KG一般被标记为，是一组格式为三元组的集合，其中，，是KG中的实体集合，是 KG 中的关系集合。KGE模型旨在将KG建模为低维向量空间。在这个向量空间中，头实体和尾实体是一个单独的向量，每个关系是头实体和尾实体间的一个运算操作。

KGE是一个通过概率推断现有KG缺失关系进而补全KG的任务。近几年研究人员已经提出了很多KGE模型，它们都具有一个度量成立的概率的评分函数。表2列出了一些KGE模型的评分函数。

**Table 2** The brief introduction of some typical KGE models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Entity | Relation | Score Function | Complexity |
| TransE |  |  |  |  |
| TransR |  |  |  |  |
| RESCAL |  |  |  |  |
| DistMult |  |  |  |  |
| ComplEx |  |  |  |  |
| RotatE |  |  |  |  |

TransE是一个具有代表性的平移模型，它假设实体和关系属于同一向量空间，是向量空间的维度。关系被建模为实体向量的平移，如果三元组成立，那么，即应该是最近的实体向量；如果不成立，应该远离。

TransE只能建模1对1的关系类型，TransR通过从实体向量空间分离出关系的向量空间来解决这个问题。TransR通过学习一个关系投影矩阵将实体投影到不同的关系向量空间，进而捕获了不同关系的差异。然而TransR为了获得更强的表示能力，也将复杂度从提升为，导致训练时间急剧增加，因此我们并没有采取TransR作为药物重定位的候选模型。

RESCAL是一个双线性模型，它为每一种关系提供了一个矩阵来建模实体之间的交互进而捕获KG的潜在语义。该模型也像TransR一样，复杂度为，这也是它的扩展模型主要改进的方向，我们也没有使用它作为药物重定位的候选模型。

DistMult通过将换成对角方阵简化了RESCAL模型，使得复杂度变成了，但是该模型的表现能力相较于RESCAL模型也大大减弱，无法达到理想的结果。

由于DistMult使用的是对角方阵，因此仅仅能捕获对称关系。为了捕获反对称关系，ComplEx将向量空间从实数域扩展到复数域，极大的提升了模型的表现能力。

受到TransE和欧拉恒等式的启发，RotatE将头实体和尾实体映射到复数向量空间，即 ，将关系建模为从头部实体到尾实体的旋转。RotatE模型能够捕获对称、反对称、反转和组成四种类型关系，因此它最大限度的表示出KG的语义和结构信息。

在后续的实验中，我们都是使用最大间隔方法训练模型，这将最小化正确三元组的排名，损失函数如下：

其中，是正负例三元组得分的间隔距离。是正例三元组集合，是负三元组的集合，它是通过破坏原有三元组中的实体和关系得到的：

### 评估指标

KGE模型可以通过链接预测技术预测KG中缺失的三元组，即给定预测缺失的尾实体，或者给定预测缺失的头实体h。我们可以通过链接预测给出需要评估三元组的排名。为了评估 KGE 模型的性能，我们使用三种经典指标：正确实体评分函数的平均排名（Mean Rank，MR），正确实体评分函数的平均倒数排名（Mean Reciprocal Rank，MRR）和正确实体评分函数的前N的比例即前N命中率Hits@N（N = 1，3，10）。MRR和Hits@N都是越高越好，MR越低越好。

如果我们用和分别表示预测缺失头实体和尾实体三元组的排名，表示需要评估的三元组集合，那么MR具体的计算方法为：

MRR具体计算方法为：

Hits@N 被计算为

其中如果条件为真，等于1，否则等于0。

### 实验设置

我们将DRKG的三元组按照90%、5%、5%的比例划分为训练集、验证集和测试集，分别为5286834个，293713个和293714个。

我们综合上面5个指标的表现在验证集上利用网格搜索所有模型的超参数（TransE\_l1、TransE\_l2、DistMult、ComplEx和RotatE），所有模型的训练批次大小batch\_size和负采样大小neg\_sample\_size分别固定为4096和256，从{0.01, 0.05, 0.1}中选择学习率lr；由于RotatE 模型实体维度是超参数嵌入维度hidden\_dim的2倍，因此将其嵌入维度固定为200，从{200,400}中选择其他模型的嵌入维度hidden\_dime；从{6,12,18}中选择TransE\_l1、TransE\_l2和RotatE的超参数，从{50,125,200}中选择DistMult、ComplEx的超参数。

在药物重定位任务中，选择全部DRKG中的AD实体作为头实体（Disease::DOID:10652，Disease::MESH:C536599，Disease::MESH:D000544），选择DRKG中的治疗关系作为链接预测的关系（DRUGBANK::treats::Compound:Disease，GNBR::T::Compound:Disease，Hetionet::CtD::Compound:Disease），将Drugbank中所有被FDA批准的药物作为候选药物（分子量 250，共8104个），通过预测缺失尾实体的方法选择评分前50的药物作为初步AD的治疗药物。

## 结果

### KGE模型的对比

实验比较4种KGE模型在知识图谱补全任务中的性能。对于MR指标，TransE模型两种变体分别取得了最优结果60.83和次优结果62.64。对于MRR指标，ComplEx模型取得了最优结果0.621，RotatE模型次之0.614。对于Hits@1指标，ComplEx模型取得了最优结果0.537，RotatE模型次之0.515。对于Hits@3和Hits@10，RotatE模型取得了最优结果分别为0.681、0.780，ComplEx模型取得了次优结果分别为0.673、0.768。表3列出了KGE模型测试集的结果。

**Table 3** The evaluation results of the KGE model. The best results are in **bold** and the second best results are in underline.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MRR | MR | Hits@1 | Hits@3 | Hits@10 |
| TransE\_l1 | 0.530 | 62.64 | 0.412 | 0.606 | 0.740 |
| TransE\_l2 | 0.437 | **60.83** | 0.302 | 0.515 | 0.693 |
| DistMult | 0.484 | 105.55 | 0.401 | 0.515 | 0.643 |
| ComplEx | **0.621** | 112.74 | **0.537** | 0.673 | 0.768 |
| RotatE | 0.614 | 63.51 | 0.515 | **0.681** | **0.780** |

通过比较KGE模型的评估结果，我们能得出以下结论。DistMult模型受限于只能建模对称关系，因此各项指标都没有最优和次优结果。TransE 模型的 MR 指标达到了最优结果，但是受限于只能建模一对一的关系，无法在其他指标上达到理想的水平。对于MR 指标RotatE模型和ComplEx呈现出截然不同的结果，RotaE模型接近于TransE模型取得的最优结果，但是ComplEx模型取得了最差结果，这可能是因为RotaE相较于ComplEx多捕获了反转和组成两类关系。对于MRR和Hits@N两种指标，RotatE和ComplEx 模型各取得了2次最优和次优结果，且最优和次优结果也非常接近，充分说明将嵌入向量空间由实数域转换到复数域的必要性。考虑到MR结果，我们最终选择RotatE作为我们最终的药物重定位模型。

**Table 3** The optimal hyperparameters of the KGE model.

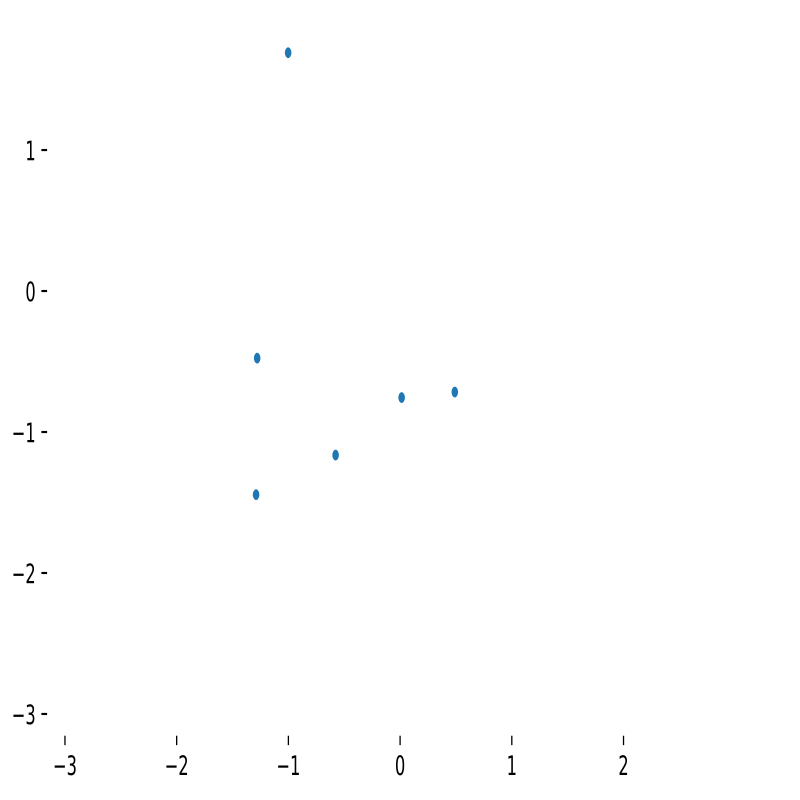
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | batch\_size | neg\_sample\_size | hidden\_dim |  | lr |
| TransE\_l1 | 4096 | 256 | 400 | 18 | 0.05 |
| TransE\_l2 | 4096 | 256 | 400 | 12 | 0.1 |
| DistMult | 4096 | 256 | 400 | 50 | 0.1 |
| ComplEx | 4096 | 256 | 400 | 50 | 0.1 |
| RotatE | 4096 | 256 | 200 | 18 | 0.05 |

表4列出了KGE的超参数，我们重新在整个DRKG上训练了RotatE模型，选择表4中RotatE的超参数作为模型的参数。

### RotatE模型的嵌入分析

由于 DRKG 结合了来自不同数据源的信息，我们希望验证使用知识图嵌入技术可以生成有意义的实体和关系嵌入。我们采用t-SNE将关系嵌入向量降维和可视化处理。由于实体的数量众多，且RotatE实体维度是400，直接利用t-SNE降维和可视化处理会引入很多噪声，因此我们首先使用PCA将其降维到30，然后再利用t-SNE将其将降维到2D空间并可视化处理。我们还使用余弦距离计算关系嵌入向量间的相似度，并输出了最相似的10对关系和相似度得分分布的直方图。

图1是关系2D空间的可视化图，可以发现关系广泛的分布在2D的空间中，即便来自相同源数据集的关系都没有出现聚集的现象，可以得出RotatE模型学习到了各个关系间差异的结论，这对于我们后面的药物重定位是很有意义的。



**Figure 1** Distribution of relation embeddings in 2D euclidean space

图2显示了基于嵌入的不同关系类型之间的成对余弦相似性的详细分布。可以发现绝大多数关系嵌入向量都有很小的相似度，而且最相似的也只有0.2。



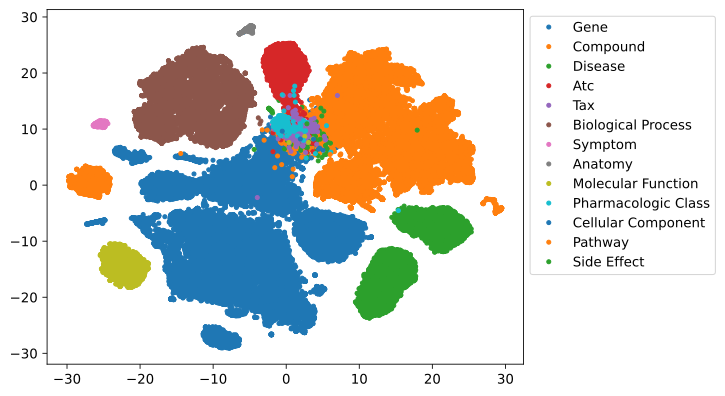
**Figure 2** Histogram of cosine similarity between relations

表5列出关系间余弦相似度最高的10对关系，可以发现所有关系对的相似得分都<0.25，与图2的结果相互得到印证。而且我们选择的三种治疗关系并不在其中，即RotatE模型能够很好的将治疗关系和其他类型的关系区分开，这对于我们的药物重定位非常重要。

**Table 5** Top 10 most similar relation pairs based on their cosine similarity.

|  |  |  |
| --- | --- | --- |
| Relation 1 | Relation 2 | Cosine score |
| Hetionet::DaG::Disease:Gene | INTACT::ASSOCIATION::Compound:Gene | 0.2335 |
| DRUGBANK::x-atc::Compound:Atc | GNBR::in\_tax::Gene:Tax | 0.2331 |
| GNBR::L::Gene:Disease | Hetionet::DpS::Disease:Symptom | 0.2269 |
| DGIDB::BLOCKER::Gene:Compound | GNBR::J::Gene:Disease | 0.2269 |
| DGIDB::AGONIST::Gene:Compound | INTACT::ASSOCIATION::Compound:Gene | 0.2226 |
| DRUGBANK::enzyme::Compound:Gene | GNBR::Ud::Gene:Disease | 0.2192 |
| Hetionet::DpS::Disease:Symptom | Hetionet::CbG::Compound:Gene | 0.2172 |
| GNBR::N::Compound:Gene | Hetionet::GcG::Gene:Gene | 0.2151 |
| GNBR::W::Gene:Gene | Hetionet::CdG::Compound:Gene | 0.2138 |
| GNBR::Z::Compound:Gene | GNBR::in\_tax::Gene:Tax | 0.2115 |

图3是实体2D空间的可视化图，不同的颜色表示不同的实体类型，观察到相同类别的实体正如我们期望的那样聚集到一起。



**Figure 3** Distribution of entity embeddings in 2D euclidean space

为了进一步分析RotatE模型学习到的关系嵌入向量的差异性，我们还分析了不同关系类型之间在链接预测时的相似程度，对于种子头实体，用链接预测找到关系下最可能成立的前10尾实体，然后对关系重复上述工作，并计算关系和的前十尾实体集合的 Jaccard相似系数（Jaccard similarity coefficient）。我们选择了100个随机头实体种子，然后报告了所有关系对的平均相似度，表6列出了基于链接预测相似度的前十关系对。Jaccard相似系数计算如下：

其中表示集合的基数，， 是两个关系对应的尾实体集合。

Table 6 Top 10 most similar relation pairs based on their link recommendation similarity

|  |  |  |
| --- | --- | --- |
| Relation 1 | Relation 2 | Jaccard score |
| INTACT::COLOCALIZATION::Gene:Gene | INTACT::DIRECT INTERACTION::Gene:Gene | 0.570 |
| Hetionet::GiG::Gene:Gene | INTACT::PHYSICAL ASSOCIATION::Gene:Gene | 0.566 |
| INTACT::PHYSICAL ASSOCIATION::Gene:Gene | INTACT::COLOCALIZATION::Gene:Gene | 0.564 |
| INTACT::PHYSICAL ASSOCIATION::Gene:Gene | INTACT::DIRECT INTERACTION::Gene:Gene | 0.558 |
| bioarx::DrugHumGen:Compound:Gene | DRUGBANK::target::Compound:Gene | 0.545 |
| GNBR::E::Compound:Gene | GNBR::E+::Compound:Gene | 0.534 |
| bioarx::DrugHumGen:Compound:Gene | Hetionet::CbG::Compound:Gene | 0.507 |
| Hetionet::GiG::Gene:Gene | INTACT::COLOCALIZATION::Gene:Gene | 0.505 |
| Hetionet::GiG::Gene:Gene | INTACT::DIRECT INTERACTION::Gene:Gene | 0.501 |
| GNBR::E+::Gene:Gene | GNBR::E::Gene:Gene | 0.490 |

可以发现我们的三种治疗关系依旧没有出现在前10的位置，更加说明RotatE模型能够将治疗关系能够与其他关系区分出来。

通过上面对RotatE模型学习到的实体和关系嵌入向量的分析结果，我们认为DRKG不同数据源的信息被药物重定位模型很好的整合到了一起，并且生成了有意义的实体和关系嵌入向量。

### AD 药物重定位

我们使用最终的RotatE模型链接预测进行 AD 药物重定位，将评分前50的作为初步的候选药物（由于RotatE模型的MR是63.51），其中前10种，只有第9名的药物不是DRKG已经存在的治疗AD的药物，证明该方法的有效性。然后我们排除了DRKG中已经存在治疗AD的药物，最终剩余了17种治疗药物。

我们将DRKG提供的TransE\_l2预训练模型重复上面的步骤，排除了DRKG中已经存在治疗AD的药物，最终剩余25种治疗药物，可以发现RotatE模型的效果要好于TransE模型。我们将上面两个模型得到的候选药物求交集，最终获得10种治疗AD的治疗药物，结果显示在Table 7。

Table 7 药物重定位得到10种候选药物

|  |  |  |
| --- | --- | --- |
| rank in RotatE | drug | literature support |
| 9 | Glutathione |  |
| 11 | Haloperidol |  |
| 13 | Capsaicin |  |
| 16 | Quercetin |  |
| 17 | Estradiol |  |
| 18 | Glucose |  |
| 21 | Adenosine |  |
| 31 | Cocaine |  |
| 39 | Paclitaxel |  |
| 41 | Cholesterol |  |

Table 7 中第一列是该药物在RotatE模型链接预测的排名。

## 结论

本文采用KGE 模型对AD进行了药物重定位。具体来说使用4种KGE模型在DRKG上学习实体和关系的嵌入向量，通过评价指标选择了RotatE模型作为最终的药物重定位模型。实验结果表明，RotatE模型能够预测出合理的AD治疗药物，能够为研究人员进一步探索AD的治疗药物提供研究思路。

## 参考文献

1. Lin YK,Shen SQ,Liu ZY,et al.Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).Berlin,Germany:Association for Computational Linguistics,2016:2124-2133.https://aclanthology.org/P16-1200.
2. Ioannidis VN,Song X,Manchanda S,et al.DRKG - Drug Repurposing Knowledge Graph for Covid-19[J].<https://github.com/gnn4dr/DRKG/,>2020.
3. Bordes A,Usunier N,Garcia-Duran A,et al.Translating embeddings for modeling multi-relational data[C]//Advances in Neural Information Processing Systems.Curran Associates, Inc.,2013,26.https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.
4. Lin YK,Liu ZY,Sun MS,et al.Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Palo Alto,California,USA:AAAI Press,2015,29(1).https://doi.org/10.1609/aaai.v29i1.9491.
5. Maximilian N,Volker T,Hans-Peter K.A Three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11).New York,NY,USA:ACM,2011:809-816.https://icml.cc/Conferences/2011/papers/438\_icmlpaper.pdf.
6. Yang BS,Yih S,He XD,et al.Embedding entities and relations for learning and inference in knowledge bases[C]//Proceedings of ICLR.2015.http://arxiv.org/abs/1412.6575.
7. Trouillon T,Welbl J,Riedel S,et al.Complex embeddings for simple link prediction[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning.JMLR.org,2016,48:2071-2080.https://arxiv.org/abs/1606.06357.
8. Sun ZQ,Deng ZH,Nie JY, et al. RotatE: knowledge graph embedding by relational rotation in complex space[C]//Proceedings of ICLR. 2019. https://openreview.net/forum?id=HkgEQnRqYQ.
9. Nian Y,Hu XY,Zhang R,et al.Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing[J].BMC Bioinformatics,2022,23(Suppl 6):407.https://doi.org/10.1186/s12859-022-04934-1.
10. Moya-Alvarado G,Gershoni-Emek N,Perlson E,et al.Neurodegeneration and Alzheimer's disease (AD).What can proteomics tell us about the Alzheimer's brain?[J].Molecular & Cellular Proteomics,2016,15(2):409-25.https://doi.org/10.1074/mcp.R115.053330.
11. Ren RJ,Yin P,Wang ZH,et al.China Alzheimer disease report 2021[J].Journal of Diagnostics Concepts & Practice(诊断学理论与实践),2021,20(04):317-337.https://doi.org/10.16150/j.1671-2870.2021.04.001.
12. Jia JP,Wei CB,Chen SQ,et al.The cost of Alzheimer's disease in China and re-estimation of costs worldwide[J].Alzheimer's & Dementia,2018,14(4):483-491.<https://doi.org/10.1016/j.jalz.2017.12.006>.
13. Avorn J.The $2.6 billion pill–methodologic and policy considerations[J].New England Journal of Medicine,2015,372(20):1877-1879.https://doi.org/10.1056/NEJMp1500848.
14. Wang SD,Du ZZ,Ding M.et al.KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer’s disease drug repositions[J].Applied Intelligence,2022,52(1): 846–857.https://doi.org/10.1007/s10489-021-02454-8.
15. Zeng XX,Song X,Ma TF,et al.Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning[J].Journal of proteome research,2020,19(11):4624-4636.https://doi.org/10.1021/acs.jproteome.0c00316.
16. Zhang R,Hristovski D,Schutte D,et al.Drug repurposing for COVID-19 via knowledge graph completion[J].Journal of Biomedical Informatics,2021,115(1):103696.https://doi.org/10.1016/j.jbi.2021.103696.
17. 李宗贤.基于知识图谱的帕金森病药物重定位[J].信息技术与信息化,2022,No.268(07):28-32.https://doi.org/10.3969/j.issn.1672-9528.2022.07.006.