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。伴随着深度学习的发展，各种类型的 KG 已经被建立，如 WikiData[2]、Freebase[3]、DBpedia[4]、YAGO[5]及 WordNet[6]等经典知识图谱。

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）[7]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实值向量的技术，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，包括但不限于 TransE[8]、TransR[9]、RESCAL[10]、DistMult[11]、ComplEx[12]、RotatE[13]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[14]，其特征是伴有神经精神症状的渐进性严重痴呆[15]。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[16]。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[17]，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[18-19]。药物重定位是指将现有药物的适应症拓宽到其他疾病，从而大大节省成本并缩短新药开发周期。药物重定位利用了同一个分子的代谢途径可以导致不同疾病的事实，因此一些药物可以治疗不同的疾病[7]。

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 等人[20]建立了一个 1500 万个三元组的综合知识图谱，包括药物、疾病、蛋白质、基因、代谢途径和表达等多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 [21]等人提出了一种基于神经网络和文献发现的方法。首先利用 PubMed 和其他以 COVID-19 为重点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[22]。

Wang[23]等人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，然后利用一个 Conv-Conv 模块来提取药物-靶点对的特征，之后通过一个全连接的神经网络进行药物靶点计算，最终通过载脂蛋白 E 寻找治疗 AD 的药物。Nian[24]等人通过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 治疗或预防的候选者，以研究 AD 与化学品，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，通过对比 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的质量；最终，通过链接预测技术找到 10 种候选药物。

数据

DRKG 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。它包含属于 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。

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。将使用小写字母表示向量，大写字母表示矩阵，小写斜体表示下表概念。Table 1 显示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	Δ	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	$\mathcal{L}$	The loss function of the model

KG 一般被标记为 **Δ**，是一组格式为 **(h, r, t)** 三元组的集合，其中 **h**，**t** ∈ **E**，**t** ∈ **R**，**E** 是 KG 中的实体集合，**R** 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 **h** 和尾实体 **t** 是一个单独的向量，每个关系 **r** 是头实体 **h** 和尾实体 **t** 间的一个运算操作。

KGE 是通过概率推断现有 KG 的缺失的关系补全 KG 的任务。近几年研究人员已经提出了多种 KGE 模型，它们都具有评分函数。评分函数度量 **(h, r, t)** 成立的概率，Table 2 列出了一些 KGE 模型的评分函数。

Table 2 KGE 模型的评分函数

Models	Entity	Relation	Score Function	Complexity
TransE	$\mathrm{h,t} \in \mathbb{R}^d$	$\mathrm{r} \in \mathbb{R}^d$	$-\|\|\mathrm{h} + \mathrm{r} - \mathrm{t}\|\|_{1/2}$	$\mathrm{O(d)}$
TransR	$\mathrm{h,t} \in \mathbb{R}^d$	$\mathrm{r} \in \mathbb{R}^k, M_r \in \mathrm{R^{k\times d}}$	$-\|\|\mathrm{M_r h} + \mathrm{r} - \mathrm{M_r t}\|\|_2^2$	$\mathrm{O(d^2)}$
RESCAL	$\mathrm{h,t} \in \mathbb{R}^d$	$M_r \in \mathrm{R^{d\times d}}$	$h^T M_r t$	$\mathrm{O(d^2)}$
DistMult	$\mathrm{h,t} \in \mathbb{R}^d$	$\mathrm{r} \in \mathbb{R}^d$	$h^T diag(r) t$	$\mathrm{O(d)}$
ComplEx	$\mathrm{h,t} \in \mathbb{C}^d$	$\mathrm{r} \in \mathbb{C}^d$	$\mathrm{Real(h^T diag(r)\bar{t})}$	$\mathrm{O(d)}$
RotatE	$\mathrm{h,t} \in \mathbb{C}^d$	$\mathrm{r} \in \mathbb{C}^d, \|r_i\| = 1$	$-\|\|\mathrm{h} \circ \mathrm{r} - \mathrm{t}\|\|^2$	$\mathrm{O(d)}$

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 $\mathbb{R}^d$，d 是向量空间的维度。关系 r 被建模为实体向量的平移，如果三元组 $(\mathrm{h,r,t})$ 成立，那么 $\mathrm{h} + \mathrm{r} \approx \mathrm{t}$，即 t 应该是 $\mathrm{h} + \mathrm{r}$ 最近的实体向量；如果不成立，$\mathrm{h} + \mathrm{r}$ 应该远离 t 。

TransE 只能建模 1 对 1 的关系类型，TransR 通过从实体向量空间 $\mathrm{h,t} \in \mathbb{R}^d$ 分离出关系的向量空间 $\mathrm{r} \in \mathbb{R}^k$ 来解决这个问题。TransR 通过学习一个关系投影矩阵 $M_r \in \mathrm{R^{k\times d}}$ 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 $\mathrm{O(d)}$ 提升为 $\mathrm{O(d^2)}$，导致训练时间急剧增加，因此我们并没有采取 TransR 作为我们的药物重定位的候选模型。

RESCAL 是一个双线性模型，它通过将实体向量的交互来捕获 KG 的潜在语义，并将每个关系表示为一个矩阵，该矩阵对实体之间的交互进行建模。该模型也像 TransR 一样，复杂度为 $\mathrm{O(d^2)}$，这也为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 M_r 换成对角方阵简化了 RESCAL 模型，使得复杂度变成了 O(d)，但是该模型的表现能力相较于 RESCAL 模型也大大减弱。

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。

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 $h, t \in \mathbb{C}^d$，将关系建模为从头部实体 h 到尾部实体 t 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。

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：

$$\mathcal{L} = \sum_{(h,r,t) \in D^+} \sum_{(h',r,t') \in D^-} max(0, \gamma - f(h,r,t) + f(h',r,t'))$$

其中，D^+ 和 D^- 分别是正例三元组与负三元组的集合；γ 是正负例三元组得分的间隔距离。

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 $(h, r, ?)$ 预测缺失的尾实体 t，或者给定 $(?, r, t)$ 预测缺失的头实体 h。我们可以通过链接预测给出测试集中三元组的排名。为了评估 KGE 模型的性能，我们使用三种经典指标：平均倒数排名（mean reciprocal rank，MRR），平均排名（mean rank，MR）和 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$rank_h$和$rank_t$分别表示预测缺失头实体和尾实体的排名，D 表示需要评估的三元组集合，那么 MR 被表示为

$$\mathrm{MR} = \frac{1}{2|\mathrm{D}|} \sum_{(h, r, t) \in \mathrm{D}} \mathrm{rank_h} + \mathrm{rank_t}$$

MRR 被计算为

$$\mathrm{MRR} = \frac{1}{2|\mathrm{D}|} \sum_{(h, r, t) \in \mathrm{D}} \frac{1}{\mathrm{rank_h}} + \frac{1}{\mathrm{rank_t}}$$

Hits@N 被计算为

$$\text{Hits@N} = \frac{1}{2|D|}\sum_{(h,r,t)\in D} I[\text{rank}_h \le N] + I[\text{rank}_t \le N]$$

其中如果条件为真，$I[*]$ 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、验证机和测试集，分别为 5286834 个，293713 个和 293714 个。

我们综合上面 5 个指标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，学习率 lr 从{0.01, 0.05, 0.1}中选择；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，其他模型从{200,400}中选择；TransE_l1、TransE_l2 和 RotatE 的 γ 从{6,12,18}中选择，DistMult、ComplEx 的 γ 从{50,125,200}中选择。

在药物重定位任务中，为预测 AD 的治疗药物，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 中的治疗作为链接预测的关系（DRUGBANK::treats::Compound:Disease，GNBR::T::Compound:Disease，Hetionet::CtD::Compound:Disease），将所有的 Drugbank 中 FDA 批准的药物作为候选药物（分子量>250），通过链接预测缺失尾实体的方法选择评分前 50 的药物。

结果

KGE 模型的对比

Table 3 和 Table 4 显示 KGE 模型的最优超参数和对应测试集的结果。

Table 3 KGE 模型的最优超参数

Model	batch_size	neg_sample_size	hidden_dim	γ	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

Table 4 KGE 模型测试集结果

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**

其中粗体表示最优得分，下划线表示次优得分。

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。但是对于 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，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。

我们重新在整个 DRKG 上训练了 RotatE 模型，超参数是 Table 4 中的超参数。

RotatE 模型的嵌入分析

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 对和相似度得分分布的直方图。

图 1 是关系的 2 维可视化图，可以发现关系广泛的分布在 2D 的空间中，即便来自相同数据集的关系都没有聚集，可以返现 RotatE 模型学习到了各个关系间的差异，这对于我们后面的药物重定位是很有意义的。

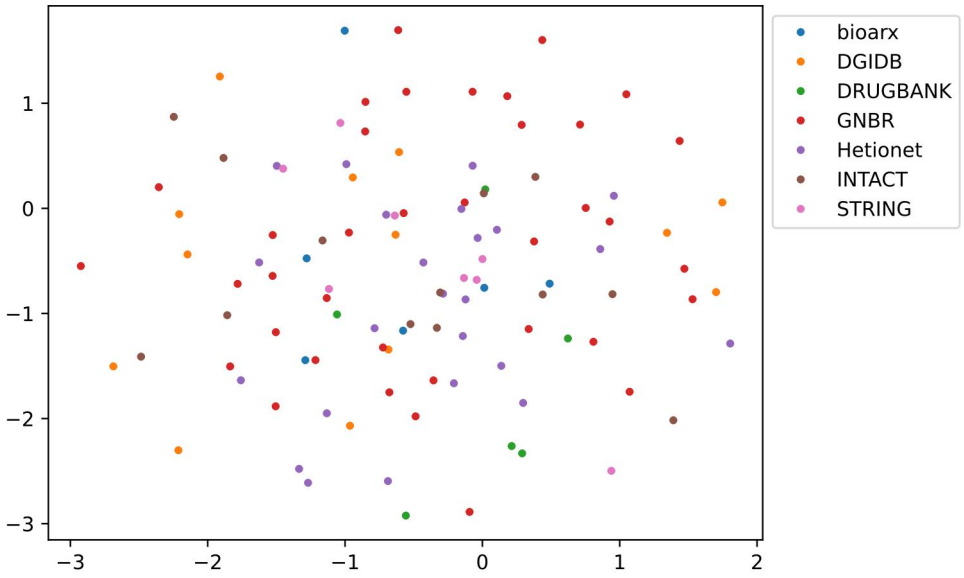

图 1 关系嵌入的 2 维空间图

图 2 是关系间余弦相似度得分的分布图，可以发现绝大多数关系都有很小的相似度。

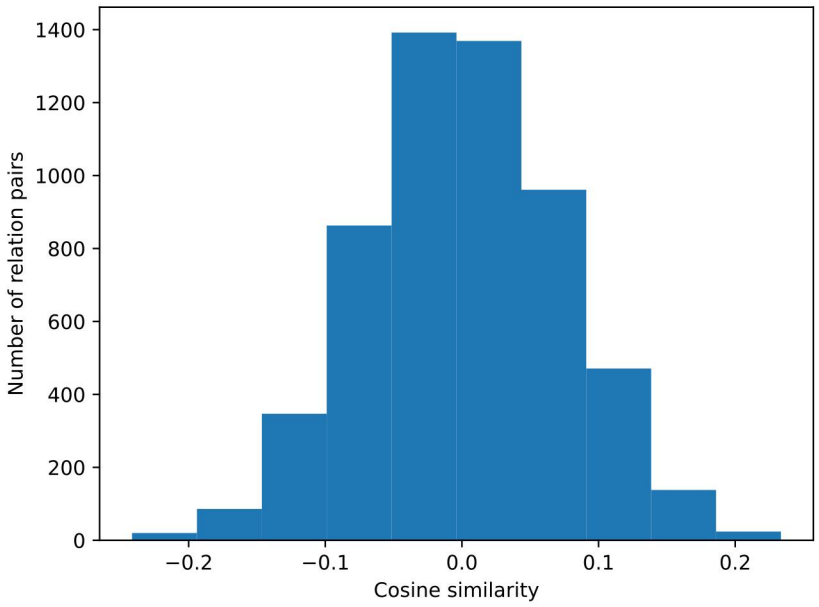

图 2 关系间余弦相似度得分的分布图

Table 5 展示关系间相似度最高的 10 对关系，可以发现相似得分非常的低，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 关系间余弦相似度得分最高的 10 组

Relation type 1	Relation type 2	Cosine score
Hetionet::DaG::Disease:Gene	INTACT::ASSOCIATION::Compound:Gene	0.23346853
DRUGBANK::x-atc::Compound:Atc	GNBR::in_tax::Gene:Tax	0.23311995
GNBR::L::Gene:Disease	Hetionet::DpS::Disease:Symptom	0.22690767
DGIDB::BLOCKER::Gene:Compound	GNBR::J::Gene:Disease	0.2268842
DGIDB::AGONIST::Gene:Compound	INTACT::ASSOCIATION::Compound:Gene	0.22258793
DRUGBANK::enzyme::Compound:Gene	GNBR::Ud::Gene:Disease	0.21916792
Hetionet::DpS::Disease:Symptom	Hetionet::CbG::Compound:Gene	0.21718228
GNBR::N::Compound:Gene	Hetionet::GcG::Gene:Gene	0.21505778
GNBR::W::Gene:Gene	Hetionet::CdG::Compound:Gene	0.21384156
GNBR::Z::Compound:Gene	GNBR::in_tax::Gene:Tax	0.21150547

参考文献

[1] Lin Y K, Shen S Q, Liu Z Y, et al. Neural relation extraction with selective attention over instances[C]. Proceedings of ACL, 2124–2133, 2016.

[2] Vrandecicd D, Krotzsch M. WikiData: a free collaborative knowledgebase[C]. Communications of the ACM, 2014, 57(10): 78-85.

[3] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Proceedings of KDD, 1247-1250, 2008.

[4] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data[C]. Proceedings of ISWC, 722-735, 2007.

[5] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61.

[6] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[7] Ioannidis VN,Song X,Manchanda S,et al. DRKG - Drug Repurposing Knowledge Graph for Covid-19[J]. https://github.com/gnn4dr/DRKG/, 2020.

[8] Bordes A, Usunier N, Garcia Duran A, et al. Translating embeddings for modeling multi-relational data[C]. Proceedings of NIPS, 2787-2795, 2013.

[9] Lin Y K, Liu Z Y, Sun M S, et al. Learning entity and relation embeddings for knowledge graph completion[C]. Proceedings of AAAI, 2181-2187, 2015.

[10] Nickel M, Tresp V, Kriegel H P. A Three-way model for collective learning on multi-relational data[C]. Proceedings of ICML, 809-816, 2011.

[11] Yang B S, Yih W T, He X D, et al. Embedding entities and relations for learning and inference in knowledge bases[C]. Proceedings of ICLR,1-13, 2015.

[12] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]. Proceedings of ICML, 2071-2080, 2016.

[13] Sun Z Q, Deng Z H, Nie J Y, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[C]. Proceedings of ICLR, 2019.

[14] Nian Y, Hu X, 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 Sep 30;23(Suppl 6):407.

[15] 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]. Mol Cell Proteomics. 2016 Feb;15(2):409-25.

[16] 任汝静,殷鹏,王志会, et al.中国阿尔茨海默病报告 2021[J].诊断学理论与实践,2021,20(04):317-337.

[17] Jia, J., Wei, C., Chen, S., et al. The cost of Alzheimer's disease in China and re-estimation of costs worldwide[J]. Alzheimer's & Dementia, 2018, 14: 483-491.

https://doi.org/10.1016/j.jalz.2017.12.006

[18] Avorn J. The $2.6 billion pill–Methodologic and policy considerations[J]. New England Journal of Medicine, 2015, 372(20): 1877-1879.

[19] Wang, S., Du, Z., 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]. Appl Intell 2022, 52: 846–857.

[20] Zeng X, Song X, Ma T, et al. Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning[J]. J Proteome Res. 2020 Nov 6;19(11):4624-4636 .

[21] Rui Zhang and Dimitar Hristovski and Dalton Schutte and Andrej Kastrin and Marcelo Fiszman and Halil Kilicoglu. Drug repurposing for COVID-19 via knowledge graph completion[J]. Journal of Biomedical Informatics. 2021

[22] 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(07):28-32.

[23] Wang, S., Du, Z., 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]. Appl Intell 2022, 52, 846–857.

[24] Nian Y, Hu X, 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 Sep 30;23(Suppl 6):407.