# 1表型-药物-分子多层次知识图谱的链接预测

## 1.1任务描述

过去的三十多年里，基于靶点的药物发现（Target-based Drug Discovery, TDD）是药物发现的主要方法。由于靶点明确，使用TDD很容易设计药物筛选试验，因此在过去有大量的成功案例。使用TDD方法来开发药物，药物发现的时间、过程和所需资源都是可预测的。近年来，随着细胞表型筛选技术，如诱导多能干细胞（iPS）技术、CRISPR-Cas技术、类器官和成像分析技术的飞速发展，表型药物发现（Phenotypic Drug Discovery, PDD）重新回到了人们的视线。在药物发现的研发中，研究者需要对各种药物做大量生物实验，从而确定药物治疗效果，但是此步骤需花费大量时间。机器学习是AI领域的一个强大的分支，它通过复杂的数据决策方案加速研究。在过去的10年里，AI方法学和机器学习算法已经成功地应用于药物发现的各个阶段，并逐步融入到大型制药公司的药物发现策略中。在大数据和人工智能时代，我们可以收集大量的临床和实验数据，利用AI算法对药物和基因的靶向作用、药物和表型的治疗作用等进行预测。其中，基于知识图谱是解决该问题的一种途径，通过构建表型-药物-分子多层次知识图谱，利用大规模图算法以及知识推理来预测疾病、症状、药物、基因、副作用等之间关系。表型-药物-分子多层次知识图谱依据表型（疾病、症状）、药物、基因等及其之间的关系来构建的知识图谱，其包含了大量实体及其关系数据，可以为致病机理和药理作用机制的研究提供一定支持。这些问题对于进一步了解疾病的分子机制、发掘潜在的药物治疗机理具有重大的实际意义和学术价值。

本任务基于表型-药物-分子多层次知识图谱进行关系预测，如药物和基因/蛋白的靶向作用、药物和表型的治疗作用、蛋白间的交互作用等。评测依据知识图谱模式（Schema）和知识图谱的实体、实体之间的关系，预测新的两个实体的关系。评测本身不限制各参赛队伍使用的模型、算法和技术。可以利用各种图算法模型，预训练等手段进行图谱上的关系预测，共同促进知识图谱技术的发展。

## 1.2输入输出

* **输入**

schema.json：知识图谱的图谱模式（Schema），即定义了知识图谱的实体类型（Entity）、和实体间的关系（Relationship）。

entities.json：实体列表，即药物、基因、表型（疾病和症状）、基因通路等实体。

relationships.json：实体与实体之间的关系列表，即药物-基因，药物-疾病，药物-症状、疾病-症状、疾病-基因、症状-基因、蛋白-蛋白的关系等。

link\_prediction.json：待评测的头实体或者尾实体和关系组成的对。

* **输出**

result.txt：针对每个部分缺失的实体关系对所预测的对应top 10头或尾缺失实体队列集合（按可能性从高到低排序）。**选手需要从队列中删除训练集已有链接的实体数据，防止占用并浪费队列位置。**

* **输入样例**
* schema.json：

{

 "entity\_type": ["disease", "drug", "gene/protein", ……],

 "relationships": [["disease", "associated\_with", "symptom"], ["disease", "disease\_mapped\_to\_gene", "gene/protein"], ["drug", "treats", "disease"], ……],

}

* entities.json：

{

 "disease": ["C0008626", "C3805375", "C1561850", "C0733682", ...],

"gene/protein": ["SEPT4", "MARCH2", "SEPT8", …],

…

}

* relaionships.json

{

 "relationships": [["DB00855","treats","C0020649"], ["DB00855","treats","C0037274"],

 …

]

}

* link\_prediction.json

{

 "relationships": [["DB01623","treats","?"], ["PSMC3","interacts\_with","?"],

 …

]

}

* **输出样例**

link\_prediction.json:

{

"results":

[

[A1, B1, C1, D1, …],

[A2, B2, C2, D2, …],

……

]

}

## 1.3知识图谱描述

该任务中的知识图谱是依据表型（疾病、症状）、药物、基因等及其之间的关系来构建的知识图谱，其包含了大量实体及其关系数据，可以为致病机理和药理作用机制的研究提供一定支持。对疾病、症状、药物、基因、副作用等之间关系的预测是医学研究的重要问题，这些问题对于进一步了解疾病的分子机制、发掘潜在的药物治疗机理具有重大的实际意义和学术价值。

本任务中的知识图谱共包含如下7种关系类型：

associated\_with关系可用来表示疾病与症状之间的联系，通过已有的大量疾病和症状间的关系，可以发掘潜在的疾病与症状间的某种联系，有助于完善我们对疾病症状的了解。

disease\_mapped\_to\_gene表示疾病与基因的关系。通过分析大量有关联的疾病与基因的关系，可以发现潜在存在联系的疾病与基因，这对于疾病预防和治疗有着重要意义。

treats关系是药物和疾病的对应关系。作为治疗疾病的重要方法，有效的药物一直是人们不断追求的。很多药物的潜在作用并没有被发现，而通过分析我们提供的数据将有助于解决这一问题。

targets关系表示的是药物与基因的关系。研究药物与基因的关系，是为了更加精准地找到治疗之策，通过研发针对某种基因的特效药物对治疗疾病有着重大帮助。

interacts\_with是基因与基因之间的关系。很多情况下，疾病不仅仅与一种基因有关，通常情况下都是多个基因共同对某种疾病产生影响。所以，深入发掘基因之间的关系有助于发现更多的致病基因，进而为治疗疾病提供更充分的依据。

annotates和pathway\_has\_gene\_element分别表示的是基因通路、go本体与基因的关系。通路富集分析和go分析是两种重要的基因富集方法，可以找到具有某种共同特性的基因，从而进一步发掘这些基因表示的信息。

## 1.4评价指标

本次评测任务借鉴MRR (Mean reciprocal rank，平均倒数排名) 来评估构建效果，我们考虑多种关系预测：药物-基因关系预测、蛋白-蛋白交互预测，疾病-蛋白关系预测等，这些共同构成全局的关系预测。相关的定义如下：

(每轮最多一次命中。)

MRR (Mean reciprocal rank，平均倒数排名)：对于一个query，若第一个正确答案排在第n位，

则MRR得分就是 $1/n$。

$$MMR=\frac{1}{|Q|}\sum\_{i=1}^{|Q|}\frac{1}{rank\_{i}}$$

其中，Q为样本query的集合（*link\_prediction.json*），|Q|表示query的个数。

$$\frac{1}{rank\_{i}}=\left\{\begin{array}{c}\frac{1}{i} (目标实体在第 i个结果中命中)\\0 (目标实体在所有结果中未命中)\end{array}\right.$$

## 1.5数据集

**本任务不允许使用公开的外部知识库。**

本任务目前总共包含80000左右实体，1200000左右的三元组。

数据集来源于正规网站的结构化数据，训练集与测试集的比例约9 : 1 。

数据的构建过程如下：

（1）首先进行数据抽取，数据的主要来源是：DO, HPO, MalaCards,OrphaNet, HSDN ,UMLS，DisGeNet，Disease-Connect等；（2）接着将这些数据表示成三元组的形式。经过讨论并咨询医学专业老师，合理确定每种头实体和尾实体的关系类型；（3）最后将不同关系类型的数据整合起来，去重后统计每种关系类型的数目。

## 1.6基线方法

本次评测任务提供两个基线方法：TransE和TuckER。数据集在基线方法上的性能结果如下：

|  |  |
| --- | --- |
| 方法 | MRR |
| TransE | 0.091 |
| TuckER | 0.237 |

# 2任务提交

本次评测额将采取刷榜方式，各任务验证集发布后，允许参赛队伍多次向平台提交结果，文件命名参考具体任务说明，并以队伍名作为前缀。格式与任务描述中的示例输出相同，排名实时更新。参赛队伍可在评测集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况；

测试集发布后，允许参赛队伍多次提交测试集结果文件。

最终提交文件要求：每一个参赛队需提交的材料如下。

问答任务测试集结果文件，用result.txt命名（UTF-8格式）

相关代码及说明

方法描述文档（非评测论文，评测论文撰写要求见CCKS 2020官网）

以上三个文件需在任务提交截止日期前发送至邮箱yangkuo@bjtu.edu.cn。邮件的标题为：“CCKS-OpenKG-参赛队名称”，例如“CCKS-OpenKG-火箭队”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，程序应当可以运行且所得结果与result.txt相符。如果方法使用了额外资源，要求说明并提供资源文件或地址。

本次评测将依托biendata平台（https://biendata.com/）展开，请有意向的参赛队伍关注平台上的竞赛列表。

# 3时间安排

评测任务发布：4月1日

报名时间：4月1日—7月15日

训练及验证数据发布：4月20日

测试数据发布：7月15日

提交测试结果：7月20日

评测论文提交：8月5日

CCKS会议日期(评测报告及颁奖)：8月18日—21日

# 4评测规则

注意，以下通用规则适用于本评测当中的所有子任务。在此基础上，参赛选手还需遵循具体任务的特有规则（参见章节2）。如有冲突，以后者为准。

参赛选手需要提交“参赛队名，队长信息（姓名，邮箱，联系电话），参赛单位名称”等信息，发送至邮箱yangkuo@bjtu.edu.cn，并在评测网站报名。

报名截止到测试数据集发布，在测试数据集发布之后，未报名的选手/队伍不能再报名或提交。

每支队伍需指定一名队长，队伍名称不超过15个字符，队伍成员不超过5人。

每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消所有相关队伍的参赛资格。

允许使用开源代码或工具，但不允许使用任何未公开发布或需要授权的代码或工具。

参赛选手最终需要提交可运行的代码和方法描述文档，若在排行榜上的结果无法复现，将取消参赛资格。

欢迎国内外在校生及社会在职人士参加。比赛组织方成员不可参赛。

# 5评测奖励

专委会为每个任务前三名设立奖金并颁发证书,任务第一名5k，第二名3k，第三名2k，技术创新奖5k，同时各排名靠前队伍将获授精美参赛奖牌、证书。

# 6组织者

周雪忠 北京交通大学 xzzhou@bjtu.edu.cn

杨扩 北京交通大学 yangkuo@bjtu.edu.cn

贾婷 北京交通大学 tingjia@bjtu.edu.cn

夏佳楠 北京交通大学 xiajn@bjtu.edu.cn

董汉阳 北京交通大学 20125154@bjtu.edu.cn