**CCKS 2022 技术评测任务书**

# 化学元素知识图谱构建及应用

## 评测背景

随着计算机软硬件相关技术的飞速发展，人们逐渐从信息时代进入智能时代。知识图谱作为承载底层海量知识并支持上层智能应用的重要载体，在智能时代中扮演了极其重要的角色。而受限于非结构化文本和结构化知识之间的巨大差异，自动化构造知识图谱以及利用知识图谱支持上层应用仍存在诸多挑战。因此对知识图谱的构建以及基于知识图谱的表示学习展开研究是十分有必要的。

知识图谱又可以划分成开放领域和特定领域。相对而言，开放领域知识图谱的特点是“广而浅”，即覆盖实体的范围广，但可能在某方面缺少深层次或不常见的知识；而特定领域知识图谱的特点是“专而深”，即只覆盖特定领域的实体信息，但会涵盖一些通常只有对应领域专业人士使用的术语、概念以及对应的知识。随着LOD、OpenKG等知识共享项目的展开，将包括开放领域和特定领域的若干知识库融合或链接在一起从而形成既全面又深入的“超级”知识图谱逐渐成为可能。

众所周知，药物研发的成本高、周期长、风险高。据美国塔夫茨大学2014年的研究，新药进入市场的平均成本约为26亿美元，从首次合成到进入临床试验的平均耗时为31.2个月，从一期临床到上市长达96.8个月。另一方面，随着全球迈入老龄化社会，对新药的需求也在逐年增加，到2024年全球医药市场总规模将超过11万亿。与之相反，制药公司每10亿美元投资所获得的上市新药数量却在逐年下降。如何通过新的技术手段，快速找到有潜力的候选药物，降低进入临床试验失败的风险，就成为药物研发领域最亟需解决的问题。

在计算方法出现之前，药物研发基本通过生物实验的方法来寻找药物，成本高昂且耗时长，随着计算化学和计算生物学的发展，也有通过传统机器学习方法辅助进行药物设计的，但这些方法或多或少在效果和效率层面有不足，以小分子为例，要找到一个候选药物，筛选（搜索）的数量级达到1060，传统计算方法很难高效完成。另一方面，随着AI技术的发展和普及，药物研发也逐渐进入到AI时代，擅长处理大数据的AI深度学习技术，就成为近年来大家关注的焦点，希望通过AI新技术提升药物研发效率，减少失败概率，降低药物研发成本。

化合物的性质预测的主要目的在于及时发现理化性质不达标的化合物，以降低候选化合物进入临床实验失败的风险，提升药物研发的成功率。传统的化合物性质预测分析一般采取实验方式，成本高昂且耗时长。很多的研究工作都证明了机器学习技术，特别是深度学习在化合物性质预测方面的巨大潜力，这些工作使用序列（SMILES表达式）或是图（原子为节点，化学键为边）来表示化合物，用序列建模或者图神经网络（GNN）去预测化合物的属性，以辅助进行药物研发，提升效率，降低成本。但是，这些方法往往只考虑了化合物分子的结构信息，而忽略了其中蕴含的化学领域知识。因此我们以化学元素周期表为核心构建了化学元素知识图谱，并于此针对知识图谱构建的关键技术及其核心应用提出评测任务。

## 任务描述

本任务围绕化学元素知识图谱的构建及化合物属性预测所展开。评测从化合物SMILES表示和初始的简单知识图谱开始，根据需要构建和扩充化学元素知识图谱，并**基于该知识图谱**进行化合物属性预测。评测本身不限制各参赛队伍使用的模型、算法和技术，但设计模型过程中**必须使用该化学元素知识图谱**。可以利用各种外部知识库扩充该化学元素知识图谱（例如引入官能团知识、wikipedia中的文本、图像信息等），可以利用各种序列/图算法模型、预训练手段等处理化合物分子，并进行化合物属性预测，共同促进知识图谱技术的发展。

#### 输入输出

* **输入**

triples.txt: 知识图谱的三元组，定义了实体(entities)、实体间的关系(relation)。

compound.csv: 化合物分子的SMILES表示。

* **输出**

newkg: 扩充后的任意形式(如：jsonld)表示的知识图谱。（选做）

property.json: 化合物分子的属性。分子属性预测任务包括二分类任务和回归任务，二分类任务需返回预测的标签，回归任务需返回预测的具体数值。

* **输入样例**

**triples.txt:**

AlkaliMetals isMetallicityOf Cs

AlkalineEarthMetals isMetallicityOf Be

Lanthanoids isMetallicityOf Yb

Period1 isPeriodOf H

Period2 isPeriodOf N

Solid isStateOf Mg

……

**compound.csv:**

Cc1cccc(N2CCN(C(=O)C34CC5CC(CC(C5)C3)C4)CC2)c1C

Cn1ccnc1SCC(=O)Nc1ccc(Oc2ccccc2)cc1

COc1cc2c(cc1NC(=O)CN1C(=O)NC3(CCc4ccccc43)C1=O)oc1ccccc12

……

* **输出样例**

**property.json:**

{

"task1": [0, 1, 1, ……],

"task2": [0, 1, 1, ……],

"task3": [-0.29, -5.82, 0.15, ……],

……

}

#### 知识图谱描述

在化学领域中，分子由若干个相同或不同的原子组成，而元素为具有相同核电荷数（即核内质子数）的同一类原子的总称。因此，本次测评希望将化学元素周期表转化为知识图谱，建模元素之间的微观联系，进而帮助分子的表示学习。

本次测评从元素周期表中提取了所有化学元素及其属性。每个元素包含超过15种属性，包括金属性、周期性、状态、质量、电负性、电子亲和性、熔点、沸点、电离能、半径、硬度、模量、密度、导率、热能和丰度。如下图展示，化学元素知识图谱描述了元素之间的微观联系及元素与属性之间存在的特定关系。

由于每个元素都有一些连续型属性，以数字的形态呈现（如半径、熔沸点等），此处我们对连续型属性进行了离散化，即对不同的数值进行分组，以建立不同元素之间的联系。



#### 分子SMILES数据处理

目前有多种分子处理库支持将化合物分子的SMILES表示转化为图表示，如RDKit、DGLLife, pysmiles和torchdrug等。通过这些库，可获取组成分子的具体原子和键类型，原子的编号及坐标信息，进而将化合物分子的SMILES表示转化为图结构，进行下一步的表示学习。当然，SMILES本身作为一种序列表示，也可以直接作为模型的输入。

例：

>>> from dgllife.utils import smiles\_to\_bigraph

>>> g = smiles\_to\_bigraph('CCO')

>>> print(g)

 DGLGraph(num\_nodes=3, num\_edges=4,

 ndata\_schemes={}

 edata\_schemes={})

#### 评价指标

本次评测任务分为二分类任务和回归任务。

1. 对于二分类任务，采用ROC-AUC来评估预测效果。相关定义如下：

ROC(Receiver Operating Characteristic)曲线，又称接收者操作特征曲线。该曲线最早应用于雷达信号检测领域，用于区分信号与噪声。后来人们将其用于评价模型的预测能力，ROC曲线是基于混淆矩阵得出的。一个二分类模型的阈值可能设定为高或低，每种阈值的设定会得出不同的FPR(伪阳性率)和TPR(真阳性率)，将同一模型每个阈值的(FPR, TPR)坐标都画在 ROC 空间里，就成为特定模型的ROC曲线。ROC曲线横坐标为FPR，纵坐标为TPR。AUC（Area Under Curve）被定义为ROC曲线下的面积。我们往往使用AUC值作为模型的评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

1. 对于回归任务，采用RMSE来评估预测效果。相关定义如下：

RMSE(Root Mean Square Error)是一种常用的测量数值之间差异的量度，它是观测值与真实值偏差的平方和观测点个数$n$比值的平方根。$RMSE=\sqrt{\frac{\sum\_{i=1}^{n}(\hat{y}\_{i}-y\_{i})^{2}}{n}}$, 其中$\hat{y}\_{i}$为预测值，$y\_{i}$为观测点的真实值，$n$为观测点个数。RMSE的值越小，说明预测模型描述实验数据具有更好的精确度。

#### 数据集

任务本身不限定方法类型，可以是无监督、半监督、有监督方法，因此不额外提供预训练相关数据集，参赛选手如有需要，可以自行搜索外部数据库作为预训练数据，但需要在技术报告中详细阐述技术路线。

在训练数据发布阶段，我们会发布**多组（每组代表一个属性预测任务）**标注好的数据（包括化合物分子的SMILES表示及其对应属性标签）作为训练集。同时发布不含标注结果的SMILES表示作为验证集，选手可以将自己生成的验证集答案提交，比赛系统会对答案进行评测，给出得分并进行排行。

在测试数据发布阶段，我们会发布验证集的标注结果（包括化合物分子的SMILES表示及其对应属性标签），同时发布不含标注结果的SMILES表示，作为测试，得到模型的最终测评结果。

## 任务提交

本次评测将采取刷榜方式，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（注意格式，如对于问答任务需要保留<>或""），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（每天提交不超过2次）。

最终提交文件要求：每一个参赛队需提交的材料如下。

1. 化合物属性预测测试集结果文件，用property.json命名
2. 扩展之后的化学元素知识图谱（选做）
3. 相关代码及说明
4. 方法描述技术文档（非评测论文，评测论文撰写要求见CCKS 2022官网）

以上四（或三）个文件需在任务提交截止日期前发送至邮箱kg\_tek@163.com。邮件的标题为：“CCKS-ElementKG-参赛队名称”，例如“CCKS-ElementKG-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property.json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。

## 时间安排

时间安排初定如下，后续如果有调整，以CCKS发布信息为准。

* 评测任务发布：4月6日
* 报名时间：4月6日—7月25日
* 训练及验证数据发布：4月25日
* 测试数据发布：7月25日
* 提交测试结果：7月31日
* 评测论文提交：8月12日
* CCKS会议日期(评测报告及颁奖)：8月25日—28日

## 评测规则

注意，以下通用规则适用于本评测任务。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。

1. 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，报名方式稍后在评测网站发布。
2. 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。
3. 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4人。
4. 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。
5. 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。允许使用外部数据，但该数据必须是公开的，并在提交最终结果时一并提交（如数据过大，需提供下载地址）。
6. 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。**
7. 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。

## 评测奖励

第一名：人民币5000元

第二名：人民币3000元

第三名：人民币2000元

创新奖：人民币5000元

备注：同时排名前三的队伍将获授精美参赛奖牌、证书。

## 组织者

张建 浙江大学 kg\_tek@163.com （联系人）

方尹 浙江大学 fangyin@zju.edu.cn （联系人）

张强 浙江大学 qiang.zhang.cs@zju.edu.cn