基于知识图谱的多模态推理

摘要

目的：知识图谱(Knowledge Graph, KG)在在生物医学领域变得越来越重要。利用 KG 嵌入技术可以从已有知识中获取新的可靠知识，是一种很前沿方法。其中有些是添加各种附加的信息来辅助推理，即多模态推理。然而，基于现有生物医学知识图谱的工作很少集中在某种特定的疾病上。

结果：本次研究开发了一种特殊疾病知识图(SDKGs)的构建和多模式推理过程。作者构建了 SDKG-11，其中包括 5 种癌症疾病、6 种非癌症疾病、1 种联合癌症 5 和 1 种联合疾病 11 的 SDKG 集。目的是发现新的可靠知识，并为特定疾病领域提供普遍的预训练知识。SDKG-11 是通过提取三元组、构建标准实体集、链接实体、链接关系得到的。基于结构嵌入、类别嵌入和描述嵌入，作者利用逆超平面投影实现了 SDKGs 的多模态推理。多模态推理以实体预测任务作为评价协议，改进了 SDKGs 上已有的模型。作者通过人工校对预测的药物-基因、基因-疾病和疾病-药物对来验证模型在发现新知识方面的可靠性。利用嵌入结果作为生物分子相互作用的分类的初始化参数，证明了嵌入式模型的通用性。

可用性和实现：构建的 SDKG-11 和 TensorFlow 的实现可从 https://github.com/ZhuChaoY/SDKG-11 网址中获得。

一、介绍

知识图谱(Knowledge Graph, KG)是一种存储知识并展示某个领域动态发展规律的方式。知识图谱通过大量三元组(头部实体 head entity、关系实体 relation、尾部实体 tail entity)表示现实世界中的事实，表示为(h, r, t)。大型的集成知识图谱如 Freebase 和 DBpedia 一直在不断地扩展。它们已经成功地在许多应用中得到了运用，例如推荐系统和问答系统。

在生物医学领域，由于知识图谱的专业知识只有领域专家才能很好地理解，有关于知识图谱的应用正在变得越来越受欢迎。KG 在预测蛋白药物靶点和不良的药物反应中的影响作用均有令人信服的例子。三联体生物医学的知识

图谱可被专家或电子病历(EMR)和文献中手动填充。对于大型的知识图谱来说，前者属于劳动密集型，需要很大的劳动量；反观后者受益于自然语言处理的快速发展，正变得越来越高效。

大多数现有的生物医学知识图谱都专注于特定的子领域，例如药物银行(DrugBank)和蛋白质仓库(UniProt)。然而，这些子领域都是在实体层面上划分的，很少有知识图谱专注于某种特定的疾病。特定疾病知识图谱(Specific Disease Knowledge Graph, SDKG)主要聚焦于某一特定疾病的知识，可以在指导疾病的病因、治疗和预后方面发挥更专业的作用。近期，为应对新冠肺炎疫情COVID-19，已经为药物再利用构建了超过三个相关的特定疾病知识图谱。例如建立了一种关于慢性阻塞性肺病(COPD)的知识图谱来帮助确诊早期可治愈阶段的 COPD。还建立了关于黑素瘤(melanoma)的知识图谱用于支持精准医疗。考虑到很难从文献中获取全部的知识(如所有文献服务检索系统(PubMes)摘要)，将知识限制在其中的几种疾病上可以得到更加集中有效的信息。作者在本次研究中考虑了 11 种疾病，其中包含 5 种癌症(结肠癌 colon cancer，胆囊癌 gallbladder cancer，胃癌 gastric cancer，肝癌 liver cancer 和肺癌 lung cancer)和 6 种非癌性疾病(阿尔兹海默症 Alzheimer’s disease，慢性阻塞性肺病 COPD，冠心病 coronary heart disease，糖尿病 diabetes，心力衰竭 heart failure 和类风湿性关节炎 rheumatoid arthritis)，并将本次研究项目命名为 SDKG-11。这些疾病的发病率和死亡率相当高，严重地威胁着人们的生命。尤其是肺癌、结肠癌、肝癌和胃癌在 2020 年全球癌症死亡率中排名前四。

由于每天都会有新的生物医学的知识出现，几乎所有构造的有关于生物医学的知识图谱都是不完整的。除了前文所说的方法外，新的知识也可以通过现有的知识进行推理。近期，出现了一种嵌入式知识图谱(KGE)成为了知识图谱推理的一种典范。嵌入式知识图谱将实体和关系映射到低维的向量空间，并使用一些简单的数学计算来代替明确定义的推理过程，大大提高了计算效率。嵌入式知识图谱模型定义了得分函数(score function)f(h,r,t)来衡量三元组来测量三元组存在的概率。为了提高模型推理的准确性，一些模型会增强得分函数的表达能力，另一些多模态模型添加了额外的信息，例如类别和描述。

本次研究提出来一个完整 SDKG 模型的构建和多模态推理的过程。首先，作者从生物医学的相关文献中构建了一个原始的 SDKG 模型。其次，作者从专业的生物医学数据库中构建了标准的实体集，然后，作者通过实体连接和关系连接对原始的 SDKG 模型进行细化，得到了 SDKG-11 模型。最后，作者利用多模态 KGE 模型对 SDKG 模型进行了推理。为了验证推理结果的可靠性，作者对预测得到的药物与基因，基因与疾病，疾病与药物进行了人工校对。同时为了验证嵌入式结果的普遍性，作者将它们作为了生物分子相互作用分类的与训练知识。

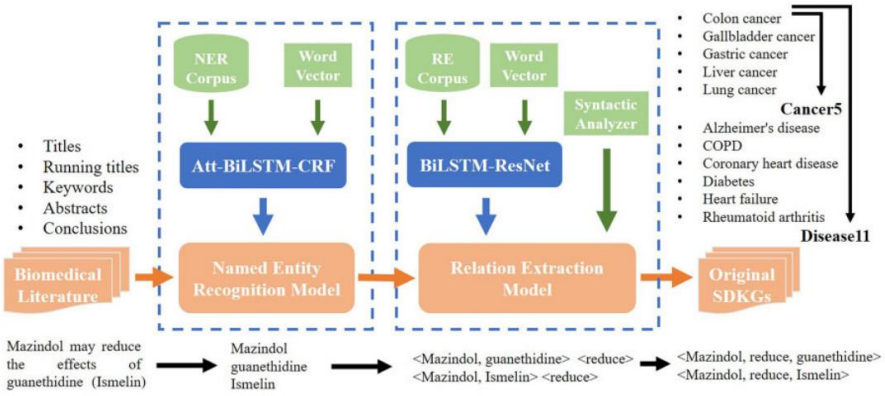
二、材料与方法

2.1 提取三元组

基于被选定的 11 种疾病的别名，作者从 1980-2020 年间发表的文献中索引文献的标题、主题词、关键词、摘要和结论中提取出三元组。其中作者只考虑了 2020 年影响因素大于等于 2.0 的期刊。

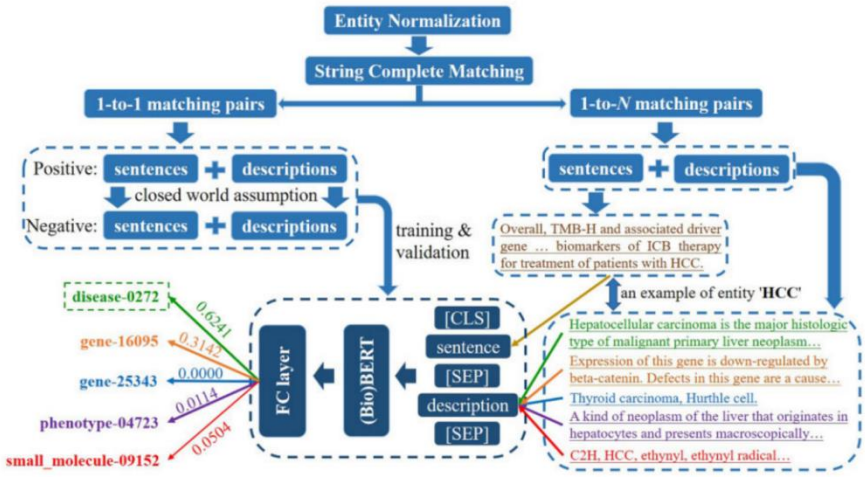
三元组的提取有两个主要的步骤：实体识别命名(Named Entity Recognition,NER)和关系提取(Relation Extraction,RE)。NER 是指从文本文献中识别生物医学实体，作者通过 Att-BiLSTM-CRF 模型来完成这一步骤。RE 提取了 NER 识别出的生物医学实体间的关系，作者通过双向长短期记忆网络(BiLSTM)和 ResNet 网络联合完成这一步骤。

作者将每种特定疾病所有提取出来的三元组整合成原始的 SDKG 模型，将 5 种癌症合并成 Cancer5 KG 模型，并构建了由 11 中疾病共同组成的 Disease11 KG 模型。除此之外，作者还记录了每个三元组的完整来源，方便后续的处理。下图是这个部分的流程图。



2.2 标准实体集

生物医学实体的名称很容易受到同义词和多义词的影响，这将增加 KG 模型的不可靠性和冗余性。针对同义词，“HCC”和“肝癌”都有指向实体“肝细胞癌”的可能性。构建包含所有同义词的标准实体集可以使这两个具有相同意义的实体指向唯一的结点。而针对于多义词，除了表示某种特定的疾病外，“HCC”还表示了两个基因，一个表现型和一个小分子，如下图。这将会通过下面 2.3.1 节中的实体链接来消除多义词的其他含义。



标准实体集中包含了 165062 个生物医学实体，见下表，都是从专门的数据库汇中提取出来的。其中，基因是从 NCBI-Gene 中提取的，miRNAs 是从 MirBase 中提取的，蛋白质是从蛋白质仓库(UniProt)中提取的，小分子是从 ChEBI 中提取的，药物是从药物银行(DrugBank)中提取的，表现型是从 HPO 中提取的，疾病是从 OMIM 库中提取的。作者用类似“gene-00001”这类序号对

所有实体进行编号。作者也提取了它们的功能注释(类别和描述)，以支持后续的多模态推理。

Table 1. Statistic of standard entity set

Entity type	Number	Number of category	Description coverage (%)
Gene	44 570	5	35.77
miRNA	4650	2	00.00
Protein	21 722	4	91.31
Small molecule	57 130	0	83.11
Drug	13 790	3	60.02
Phenotype	13 692	0	78.02
Disease	9508	1	85.01

Note: The last two columns indicate the number of category annotations and non-empty character percentage of description annotations for each entity type.

每一个实体类型都有特定的类别注释。例如，蛋白质的类别注释包括了它的状态(“蛋白质水平的实验证据”等)和它的基因本体注释。此外，实体类型(基因、疾病等)被视为一个特定的类别。详细的类别注释见附录 S2。

注释描述由按重要性顺序连接起来的多个文本内容组成。例如，通过摘要、临床特征、分子遗传性、图示和遗传文本，附加在某种疾病的注释描述上。但由于有些描述是空字符，所以作者使用有同义词进行拼接替代。详细的注释描述见附录 S3。

2.3 特殊疾病的 KG 模型构建

2.3.1 实体链接

从文献中提取出来的初始三元组应该要链接到标准实体集。首先，作者对标准实体集中的初始实体集与同义词进行实体标准化，其中包括筛选异常值、词干处理和标记重排序。之后，作者将初始实体通过该字符串完全匹配的原则连接到标准实体集上。但由于同义词可能会出现在多个实体集中，作者建立了一个端到端的实体消歧模型来选择一对多的映射中最适合的标准实体。

这个消歧模型使用了原始实体的来源语境和标准实体的注释描述作为输入，通过编码器和完全连接层(Fully Connected layer)。作者将所有的一对一的映射作为一对对正集合，并同时生成一个相等的负集合。最后，两个集合连接起来的结合体随机划分为训练集(90%)和测试集(20%)。

作者申请了使用了与训练语言模型 BERT 和它的生物医学版本 BioBERT 依次作为编码器。作为编码器，它们都通过 10 个周期进行微调，并在测试集中验证它们的准确性。该步骤的结构和示例，如下图表所示.有关于预训练语言模型和完整的训练细节以及更多信息，详见附录 S4 和 S5。

2.3.2 关系链接

生物医学的关系问题主要是噪声和同义现象，所以作者对所有原始关系进行关系标准化。除了筛选异常值和词干处理，作者还做了词性标注并只留下了名词、动词与副词。与实体不同的是，关系并没有那么多，并且缺乏一套匹配标准。因此，作者参考已经构建的关系层次结构，为频繁发生的关系手动构建了一个映射字典。例如，关系“相关(related)”和“相关(correlated)”都是由关系“关联(associate)”进行表示。

2.4 知识图谱嵌入

作者通过三个部分构建了多模态推理模型：结构嵌入(S)、类别嵌入(C)和描述嵌入(D)。对于每一个 SDKG 模型，作者将其随机划分为训练集(80%)、验证集(10%)和测试集(10%)，来确保所有的实体和关系都出现在训练集中。

2.4.1 结构嵌入

作者采用了三个代表性的模型分别来构造结构嵌入的部分，它们分别是：TransE、TransH 和 ConvKB。

TransE 认为，如果一个三元组存在，它的向量表示应该要符合：$h + r \approx t$。它的结构嵌入定义为：

$$S_{TransE}(h, r, t) = h + r - t。$$

尽管简单高效，但 TransE 并不能处理非一对一的关系。TransH 为每一个关系 r 都引入了一个超平面法向量 w，这样一来，每一个实体在面对外部关系的时候都有一个不同的向量进行表示。而 ConvKB 则是结合了 TransE 的原则(h+r-t)运用卷积运算使得模型的参数效率更高。它们的结构嵌入定义为：

$$S_{TransH}(h, r, t) = (h - W_r^T h w_r) + r - (t - W_r^T t w_r)$$

$$S_{ConvKB}(h,r,t) = ReLU([h;r;t] \times \Omega)$$

其中矩阵[h;r;t]是三元组的连接，*是卷积运算符，Ω 是过滤器的连接(初始化为 1x3 的向量[0.1,0.1,-0.1])，ReLU(x)=max(x,0)。

2.4.2 类别嵌入

作者首先随机初始化一个与结构嵌入具有相同维数的类别嵌入矩阵，这个矩阵将与结构嵌入联合学习。然后作者将实体 e 所有类别的嵌入向量的平均值作为其类别嵌入：

$$e_c = \frac{1}{|e^c|}\sum_{c \in e^c} c$$

其中e^c是 e 的类别集合。三元组的类别嵌入定义为：

$$C(h,r,t) = h_c - t_c$$

2.4.3 描述嵌入

作者使用 BioBERT 来将描述注释转换为可以计算的向量。实体 e 的描述嵌入定义为：

$$e_d = W_D{}^T BioBERT(e^d)$$

其中e^d是 e 的描述注释，W_d是权矩阵。作者通过 10 个微调周期提前训练所有实体的描述嵌入，并将其固定为特征输入。三元组的描述嵌入定义为：

$$D(h,r,t) = h_d - t_d$$

2.4.4 多模态学习

交叉嵌入式和超平面投影(HP)是两个两个传统的多模态的学习方法。TransE 的嵌入式交叉得分函数定义为：

$$f(h,r,t) = -\|h - r + t\|_2^2 - \sum_{M \in \{C,D\}} (\|h_M - r + t\|_2^2 + \|h - r + t_M\|_2^2)$$

并且 TransE 和 ConvKB 的公式与此类似。在 HP 方法中，类别嵌入和描述嵌入被视为两个法向量(如 Fig.3A)。然而，作者认为结构嵌入应当是多模态学习的核心部分，因为它包含了来自于文献的基本知识。因此，作者提出了逆超

平面投影(reverse-HP)，它将结构嵌入作为超平面(如 Fig.3B)。在一方面，作者想要将结构嵌入向量的模块最小化；另一方面，作者希望将类别嵌入和描述嵌入在结构超平面上的投影最大化，从而完全地提取注释的意义。本次研究的结果就是基于比较后的 reverse-HP(见 Fig.4)。

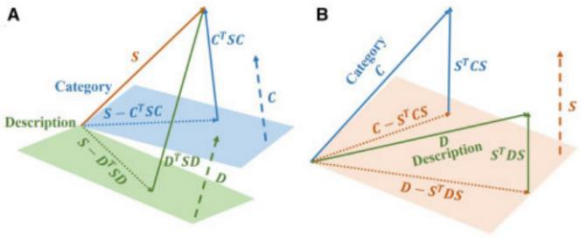


Fig. 3. The schematic of (A) hyperplane and (B) reverse-HP. Orange for structure, blue for category and green for description one. The dashed lines represent normal vectors, which should be normalized to unit normal vectors

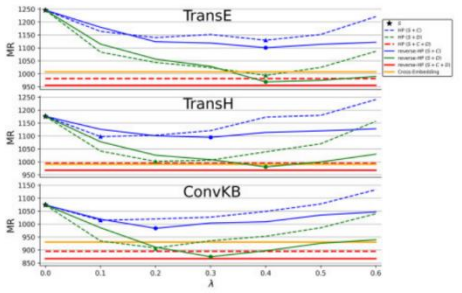


Fig. 4. Comparison of multimodal learning methods. Blue, green and red lines represent $S+C$, $S+D$ and $S+C+D$ configurations. HP and reverse-HP are denoted by dotted and solid lines, respectively. The yellow line represents cross-embedding. Triangles and dots represent the optimal λ^* for that annotation.

对于 TransE 和 TransH，reverse-HP 最后的得分函数定义为：

$$f(h,r,t) = \|S\|_2^2 + \sum_{M\in\{C,D\}} \lambda_M \left\|M - S_*{}^T M S_*\right\|_2^2$$

其中$S_* = S/\|S\|_2^2$，λ_C和λ_D是权重参数。损失函数定义为：

$$L = \sum_{(h,r,t)\in G^+,(h',r',t')\in G^-} \mathrm{ReLU}\left(\gamma - f(h,r,t) + f(h',r',t')\right)$$

其中γ是边际超参数，G^+表示正三重态集，G^-表示通过伯努利技巧(Bernoulli trick)人工生成的负三重态集。

对于 ConvKB，最后的得分函数定义为：

$$L = \sum_{(h,r,t)\in G^+\cup G^-} \log\left(1+\exp\left(-y_{brt}\cdot f(h,r,t)\right)\right)$$

其中当$(h,r,t)\in G^+$时，$y_{brt}=1$，其余情况$y_{brt}=-1$。

2.4.5 实验设置

作者使用 Adam 优化器来最小化训练集上的损失函数，在验证集上通过网格搜索策略找到最优超参数，在测试集上评估模型。超参数的搜索范围为:嵌入大小$k\in\{100,200\}$，边际距$\gamma\in\{0.2,0.6,1.0\}$，过滤器个数$nf\in\{10,20,30\}$，学习率$lr\in\{5\times10^{-4},10^{-3},5\times10^{-3}\}$，$\lambda\in\{0.0,0.1,0.2,0.3,0.4,0.5\}$。此外，作

者将批大小固定为训练集大小的 1/40，最大训练周期为 1000。所有嵌入都通过 Glorot 初始化进行初始化，边界为$(-\sqrt{6/k},\sqrt{6/k})$。

KGE 模型的评价协议是实体预测任务，即给定一个实体和一个关系，预测另一个实体。作者的评估指标是正确答案的平均排名(MR)。请注意，更低的 MR 值意味着性能更好，并且作者只考虑“过滤器”的设置。作者还评估了 PharmKG，这是一个专门用于生物医学数据挖掘的 KG 基准。由于 PharmKG 并没有实体类别和描述注释，所以作者用标准实体集注释重叠部分，并用空值覆盖其余部分。在实验中，作者考虑了以下四种配置:

1. S 表示仅使用结构嵌入($\lambda_c = 0$，$\lambda_D = 0$)。
2. S+C 代表使用结构和类别嵌入($\lambda_c \neq 0$，$\lambda_D = 0$)。
3. S+D 表示使用结构和描述嵌入($\lambda_c = 0$，$\lambda_D \neq 0$)。
4. S+C+D 同时代表使用三个嵌入($\lambda_c \neq 0$，$\lambda_D \neq 0$)。

2.5 验证

2.5.1 统计优势检验

为了检验所选模型和配置的优越性，作者对分数秩进行了多变量方差分析(ANOVA)。由于非正态秩分布，作者使用 R 包裹 WSR2，采用基于中位数和中位数估计的稳健方差分析。此外，从另一方面使用单件 Wilcoxon 检验作为事后 hoc 检验。

2.5.2 推断知识的可靠性

作者完成了药物与基因、基因与疾病和疾病与药物的 KG 模型。对于所有可能的配对，作者计算它们的得分(所有关系都被替换并保留最高分)通过每个 SDKG 上的训练得分函数。作者假设得分最高的推断项目(不在训练集中)作为可靠的新推断知识，其规模为训练集规模的 10%。然后，将它们与现有知识相结合，构建一个综合网络，如 Cytoscape 所示。

作者进一步将疾病基因预测结果与先进的基于网络的方法、LINE、Node2vec 和 HerGePred 进行比较。关联精度(Association Precision, AP)作为评价指标:

$$AP = \sum_{d \in D} |T(d) \cap P(d)| / \sum_{d \in D} |T(d)|$$

其中 D 是疾病测试集，T(d)表示疾病 d 的测试基因集，P(d)表示顶部 T(d)的预测试基因集。在临床意义方面，作者特别感兴趣的是新的推断疾病与药物对具有潜在的临床应用。作者使用 CoClust 对综合疾病-药物对进行共聚类。CoClust 是一个基于 K-means 聚类的 Python 包，用于 1-0 变量。作者将每个双边聚类的簇数设置为两个，因为疾病可以分为癌性和非癌性，而药物也是如此。

2.5.3 嵌入模型的通用性

作者使用嵌入结果作为生物分子相互作用分类任务的初始化参数，该任务的数据集提取自 Pathway Commons v12。作者的目标是通过两步来预测实体对中的交互：第一步判断实体对是否交互(手动生成相等数量的非交互实体对)，第二步预测交互实体之间具有哪种交互。最后，总体预测精度计算为:[Acc(Interacting 相互作用)× Acc(Step 2 步骤 2)+ Acc(Non-Interacting 非相互作用)]/2。

所有是实体的初始嵌入将取决于以下五种配置:NONE 表示 Glorot 随机初始化，P、P+C、P+D 和 P+C+D 分别表示通过 S、S+C、S+D 和 S+C+D 的预训练嵌入结果初始化配置。数据集描述、模型结构和训练细节见补充附录 S6。

三、结果

3.1 特殊疾病 KG

在实体消歧步骤中，BioBERT 和 BERT 都在 10 个微调周期内收敛，显示了预训练语言模型惊人的力量。它们在验证集中的准确率分别达到了 91.3%和 90.6%。因此，作者在下面的分析中使用了 BioBERT 实体消歧的结果(Table2)。

在关系链接步骤中，所有构建的 SDKG 都有 67 个关系，通过关系层次结构进行映射。

Table 2. Statistics of the specific disease KGs

Type	SDKG	Original		Constructed	
		#Triplet	#Entity	#Triplet	#Entity
Cancers	Colon cancer	360 695	95 837	53 858	8085
	Gallbladder cancer	36 865	13 286	4227	1585
	Gastric cancer	155 657	48 257	25 514	4854
	Liver cancer	515 371	142 564	76 723	10 186
	Lung cancer	383 582	117 338	59 262	8723
	Cancer5	1 387 710	301 460	197 009	15 258
Non-cancers	Alzheimer's disease	159 459	43 288	22 929	4427
	COPD	30 154	11 235	4981	1615
	Coronary heart disease	101 801	29 582	14 780	3332
	Diabetes	408 433	86 341	71 036	7886
	Heart failure	104 212	36 203	17 430	4457
	Rheumatoid arthritis	129 710	36 985	17 679	3725
Disease11		2 305 019	438 993	332 937	19 416

3.2 实体预测

从 MR 比较(Table3)以及稳健方差分析和事后 Wilcoxon 检验的统计分析(补充 Table S1)中，作者可以观察到:

在大多数情况下，ConvKB 都能达到最佳性能，且具有统计学显著性。

在 14 个 KGs 和 3 种结构嵌入算法中，35S+C+D 的配置的性能最佳(S 和 S+ C 为 0,S+D 为 8)，但 S+C+D 相对于 S+D 的优势并不具有统计学意义。

PharmKG 的提升相对较小，因为它仅仅只有部分注释。

Table 3. MR comparisons on each constructed SDKG under three structure embedding algorithms and four configurations

Model	TransE				TransH				ConvKB			
SDKG\configuration	*S*	*S*+*C*	*S*+*D*	*S*+*C*+*D*	*S*	*S*+*C*	*S*+*D*	*S*+*C*+*D*	*S*	*S*+*C*	*S*+*D*	*S*+*C*+*D*
Colon cancer	749	644	608	**606**	708	640	**576**	587	653	604	**549**	550
Gallbladder cancer	238	230	192	**189**	225	219	197	**185**	203	174	157	**154**
Gastric cancer	474	412	377	**376**	470	417	**378**	390	390	363	338	**333**
Liver cancer	912	788	745	**735**	869	769	711	**704**	802	720	629	**627**
Lung cancer	830	702	**669**	675	805	710	663	**645**	709	646	**564**	566
Cancer5	1157	1015	900	**892**	1105	1043	926	**914**	1045	957	820	**804**
Alzheimer's disease	433	382	354	**341**	388	360	321	**309**	352	318	293	**281**
COPD	222	194	185	**182**	210	181	**174**	**174**	175	155	145	**144**
Coronary heart disease	341	298	277	**267**	336	302	283	**270**	294	260	226	**217**
Diabetes	594	527	463	**462**	544	503	**460**	461	501	455	401	**399**
Heart failure	527	474	421	**419**	483	445	412	**411**	441	409	**348**	353
Rheumatoid arthritis	438	391	337	**336**	389	360	319	**318**	334	314	292	**291**
Disease11	1246	1101	969	**956**	1176	1095	981	**968**	1074	984	874	**867**
PharmKG	290	277	263	**255**	274	265	257	**254**	254	244	239	**236**

Note: The best configuration under each structure embedding algorithm is noted in bold. The best configuration for each SDKG is noted in underlining.

3.3 多模态学习的比较

Figure 4 展示了 HP 和反向 HP 分别通过改变 λ_C和 λ_D来改变性能(Disease11 的 MR)，当一个固定在 0 时，另一个从 0 变化到 0.6(作者假设 λ_C和 λ_D是独立的，因为它们是两部分的权重)。作者可以看到，HP 最初优于反向 HP，但随着 A 的增加，反向 HP 始终更好。与交叉嵌入相比，在适当的 λ_C^*和 λ_D^*下，HP 和反向 HP 的 S+C+D 配置都具有较好的性能。

3.4 新的知识推理

根据相应训练集规模的 10%为标准，作者最终通过 ConvKB(S+C+D)获得了可靠的药物与基因、基因与疾病和疾病与药物对的新的知识的推理(补充表 S2-S4)。这些结果可能是新的发现，也是潜在的研究方向。Table 4 列出了疾病的前 10 个新推断对。其中三种类型 10 对分别有 8、9 和 9 个文献证据，能够充分证明作者的模型在发现新知识方面的可靠性。而那些没有找到证据的配对可能是一些超出当前知识范围的发现。

Table 4. Validation of the top 10 drug–gene, gene–disease and disease–drug pairs

Pair type	Rank	Score	Head entity	Tail entity	PubMed evidence
Drug–gene	39	18.522	Interferon gamma	*IFNB1*	29 313 175
	51	17.990	Nerve growth factor	*NGF*	*Equivalent*
	54	17.874	Interferon alfa	*IL22*	30 976 912
	58	17.817	Interferon gamma	*CASP3*	22 785 177
	121	16.370	Insulin beef	*FGF21*	29 987 000
	122	16.086	Interferon kappa	*IFNG*	\
	140	16.086	Thrombin	*CD177*	\
	154	15.814	Docetaxel	*SERPINB3*	21 695 460
	159	15.534	Interferon gamma	*IFNG*	*Equivalent*
	175	15.375	Interleukin-7	*FOXP3*	32218828
Gene–disease	114	23.200	*MMP12*	Pulmonary fibrosis	33 065 600
	260	21.234	*HMGA2*	Bladder cancer	31 053 526
	452	19.706	*YAP1*	Squamous cell carcinoma	32 206 709
	565	19.159	*TIMP1*	Sarcoidosis	26 240 517
	573	19.098	*GPI*	Ovarian cancer	\
	596	19.015	*IRF8*	Autoimmune disease	30 285 234
	602	18.988	*TNFRSF25*	Lupus erythematosus	22 666 553
	645	18.789	*SATB1*	Squamous cell carcinoma	32 451 408
	654	18.755	*DLEC1*	Ovarian cancer	30 324 802
	662	18.726	*EPHA2*	Breast cancer	33 962 882
Disease–drug	37	16.939	Biliary cirrhosis	Interferon alfa	23 291 480
	128	14.023	Wilson disease	Iron	33 680 437
	151	13.373	Depressive disorder	Glutathione	26 706 022
	180	12.863	Obesity	Albumin human	22 230 555
	217	12.443	MODY	Insulin human	27 103 109
	272	11.815	MODY1	Insulin human	28 684 784
	274	11.803	Burkitt lymphoma	Fibronectin	19 625 084
	320	11.356	Pancreatic cancer	Insulin-like growth factor II	28 420 208
	356	11.026	Diabetes I	Vitamin D3	33 069 738
	383	10.744	Lamellar ichthyosis	Ethanol	\

Note: The Rank column also ranks the pairs in the training set. *Equivalent* means that drug and gene refer to the same biological concept, so we treat it as a correct prediction.

对于疾病基因预测任务(Table 5)，所有由 ConvKB 和其他三种先进的基于网络的方法构建了的疾病基因网络，由于域诱导的不完全性，SDKGs 具有相对

较小的 AP。然而，由于关系嵌入和多模态注释来补偿网络的不完全性，ConvKB 优于基于网络的方法。

Table 5. *AP* of disease–gene prediction results based on SDKGs

SDKG\methods	LINE	Node2vec	HerGePred	ConvKB
Colon cancer	0.0492	0.0431	0.0529	**0.0683**
Gastric cancer	0.0349	0.0403	0.0430	**0.1020**
Liver cancer	0.0368	0.0314	0.0378	**0.0650**
Lung cancer	0.0132	0.0286	0.0317	**0.0782**
Alzheimer's disease	0.0316	0.0150	0.0382	**0.1080**
COPD	0.0578	0.0751	0.0636	**0.0983**
Diabetes	0.0043	0.0281	**0.0583**	0.0540
Rheumatoid arthritis	0.0573	0.0466	0.0717	**0.1111**

Note: Bold number stands for the best method on that SDKG. We only list the SDKGs with $\sum_{d\in D}|T(d)| \geq 100$. ConvKB on $S+C+D$ configuration.

Figure 5 显示了包含通过推断边连接的所有实体的网络。每个网络中心的节点是疾病本身，大部分是由预期的已有知识(黑边)进行连接。然后作者的模型基于这一特定疾病领域的核心知识推断出新的知识(绿边)。

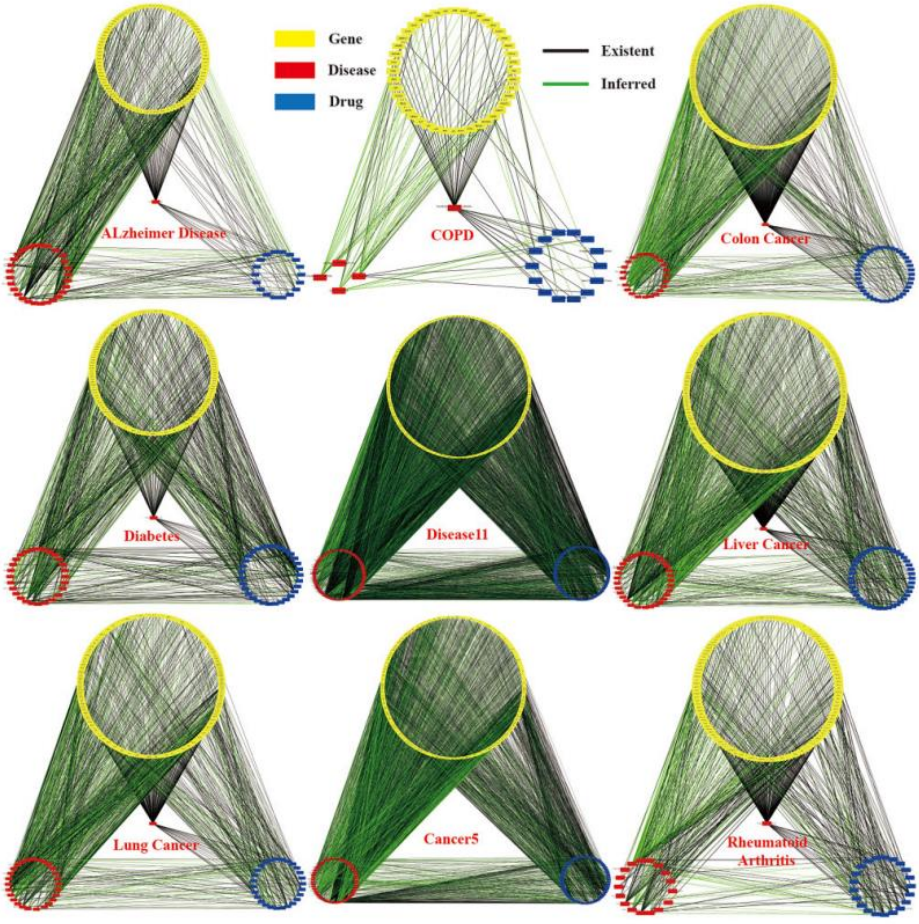


Fig. 5. Networks contain inferred (green edges) and existent (black edges) knowledge. Yellow, red and blue nodes denote gene, disease and drug entities, respectively

3.5 推断知识的应用

从 Didease11 中的药物与疾病部分(Fig.6A)作者可以观察到:在大多数情况下，抗癌药物是用来治疗癌症的，反之亦然。虽然，非抗癌药物也广泛用于癌症治疗，但反之则不然。大多新知识推断疾病与药物对是属于抗癌药物领域，用于治疗癌症。这意味着将一些抗癌药物扩展到更多类型的癌症将会是药物再利用的主流方向。将非抗癌药物在原领域进行再利用，用于治疗非癌性疾病，具有广泛的临床应用前景。然而，从现有的知识中很少能推断出非抗癌药物在癌症中的再利用。它们应该主要依赖于一些超越现有知识的颠覆性发现。

此外，作者聚焦{药物(drug)，基因(gene)，疾病(disease)}封闭三元组，来相互证实可靠性和系统性。所需的闭三元组由每个节点类型组成，并且至少包含一个新的推断边(补充 Table S5)。一方面，作者可以从三元组的另外两条边获得更多的支持证据;另一方面，这三元组本身是一个逻辑上的闭环，可以自然地形成一个命题，即“基因与疾病有关，药物通过影响基因(产物)对疾病产生作用”。以 Cancer5 为例，作者发现了一个以维生素 D3 为中心的封闭三元组集群(Fig.6B)。据预测，维生素 D3 可以预防或缓解乳腺癌，其效果可能是通过 IL 10 等基因起作用的。维生素 D3 对 IL10 基因的影响和 IL10 在乳腺癌中起重要作用都被研究过。虽然一些研究部分支持作者的预测，即维生素 D3 可能预防或缓解乳腺癌，但并没有直接的证据。

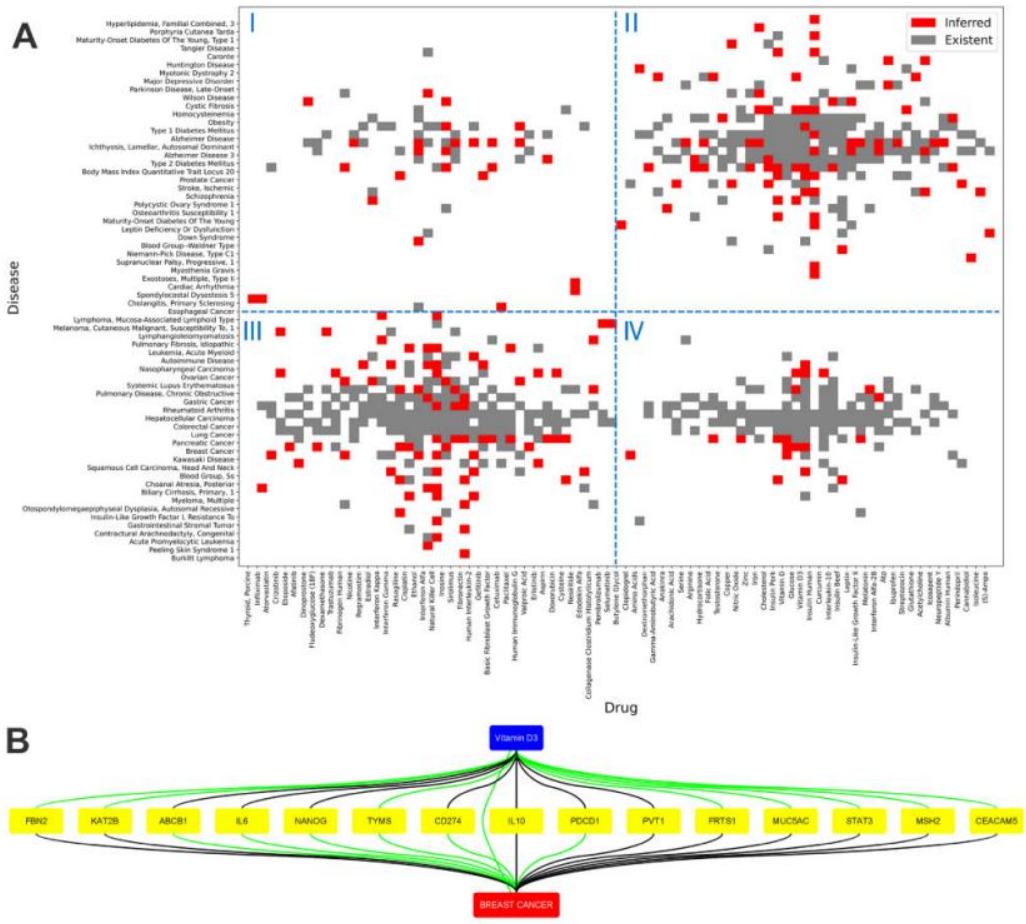


Fig. 6. (**A**) Co-clustering result for comprehensive disease–drug pairs on the KG combining all the 11 diseases. These pairs are automatically grouped into four distinct clusters (I: anticancer drugs ~ non-cancer diseases; II: non-anticancer drugs ~ non-cancer diseases; III: anticancer drugs ~ cancers; and IV: non-anticancer drugs ~ cancers). Red and gray nodes denote inferred and existent knowledge, respectively. (**B**) A cluster of 15 {drug, gene, disease} closed-triplets with vitamin D3 and breast cancer nodes have emerged in the Cancer5 after reasoning. The inferred and existent edges are in green and black, respectively

3.6 预训练生物分子相互作用分类

Table 6 展示了 Disease11 的 ConvKB 模型生成的生物分子相互作用预测结果，从中作者可以观察到：

Table 6. Accuracy of pre-trained biomolecular interaction classification

Step	Relation		Configuration				
	Name	Number	NONE	*P*	*P* + *C*	*P* + *D*	*P* + *C*+*D*
1	Interacting	727 318	0.920	0.903	0.921	**0.922**	**0.922**
	Non-interacting	727 318	0.819	0.865	0.889	0.890	**0.891**
	Binary categories	1 454 636	0.869	0.884	0.905	0.906	**0.907**
2	In-complex-with	51 173	0.656	0.649	0.677	0.685	**0.697**
	Catalysis-precedes	40 343	0.919	0.886	0.916	0.919	**0.934**
	Controls-expression-of	49 707	0.855	0.861	0.908	**0.920**	0.916
	Interacts-with	148,989	0.854	0.855	0.874	0.874	**0.878**
	Controls-state-change-of	55 257	0.758	0.771	0.846	0.865	**0.870**
	Controls-phosphorylation-of	7679	0.397	0.343	0.518	**0.600**	**0.600**
	Controls-production-of	11 187	0.991	0.991	0.991	**0.995**	0.989
	Controls-transport-of-chemical	2088	0.883	0.790	0.896	0.918	**0.959**
	Chemical-affects	333 375	0.997	**0.998**	**0.998**	**0.998**	**0.998**
	Consumption-controlled-by	10 929	0.783	0.706	0.766	0.767	**0.793**
	Used-to-produce	13 118	0.803	**0.878**	0.877	0.826	0.830
	Reacts-with	3473	0.406	0.266	0.290	0.449	**0.459**
	All the 12 interactions	727 318	0.894	0.895	0.916	0.920	**0.923**
Overall			0.821	0.837	0.866	0.869	**0.871**

Note: Bold number stands for the best configuration on prediction.

1. 尽管 NONE 和 P 配置具有相同的结构，但用结构嵌入初始化比随机初始化具有更好的预测精度。它有力地支持了预训练嵌入的通用性。

2. P+C+D 配置在所有步骤中预测精度最高。这表明多模态学习可以进一步提高生物分子相互作用分类的性能。

3. 在步骤 2 中，一些相互作用(controls-phosphorylation-of and reacts-with)的预测精度相当低，主要是由于它们的样本量相当小。

四、讨论

SDKG-11 是基于生物医学文献构建的，其构建过程几乎不需要人为干预就能产生大规模的原始三元组。生物医学文献作为一种高浓度的知识载体，几乎囊括了所有已经发现和正在研究的知识。因此，它应该是一个理想的三元组的提取源。对于生物医学知识的另一个主要来源，EMR 的整体数据质量较低，且其中有很多不可预测的噪声，并且在不同地区的差异也很大。然而，基于 EMR 的三元组提取最显著的优势之一是它更接近临床反应且更真实。因此，接下来作者希望将文献与 EMR 相结合，构建包含真实世界数据的综合 SDKGs。

在模型层面，作者对 TransE、TransH 和 ConvKB 作为结构嵌入部分进行了测试，实验结果表明，ConvKB 是最有效的结构嵌入部分。其主要原因是 ConvKB 既考虑了 TransE 的过渡特征，又利用了卷积神经网络的有效性。如果作者打算考虑更高级的结构嵌入，基于图神经网络的 KGE 模型将会是一个可靠的选择，因为它与 KG 的拟合结构更自然。KG 和图卷积网络的结合，以及 KG

和图自注意网络，两者都进行了研究。作者认为将多模态嵌入作为图节点的多个特征将会是个很有前景的尝试。

对于多模态学习，作者采用反向 HP，而不是交叉嵌入和 HP。交叉嵌入将结构嵌入与其他模态嵌入结合起来，但不能调整每个模态嵌入的权值。以前的 HP 模型只考虑描述超平面，并且它们的描述嵌入由主题模型和跳格模型生成。在这项工作中，权重参数代表每个注释的贡献程度，具有一定的可解释性。类别注释和描述注释都可以促进推理效果(Fig.4)，因为它们在一定程度上弥补了 SDKGs 的不完备性，为结构嵌入提供了新的知识来源。同时，作者还观察到描述注释比类别注释更好，因为前者会包含更多的信息。此外，在已有描述注释(S+D)的情况下，增加类别注释(S+C+D)也并不会有太大的改善。假设作者打算考虑更多的模态注释，图像注释(例如蛋白质和药物的结构示意图)是最有可能添加的注释。此外，关系标注和多组学也是非常有潜力的研究方向。

从知识相关性的角度来看，来自每种特定疾病文献的三元组更具有针对性。然而，从大数据的角度来考虑信息的完整性，作者建议可以将所有可用的信息合并到一个语料库中，因为从一个 SDKG 推断的一些知识可能已经存在于另一个 SDKG 中。接下来，作者将尝试从更一般的主题中提取知识，例如“癌症”或“疾病”来构建初始 KG。

五、总结

在本次研究中，作者提出了一个完整的特定疾病 KG 构建和多模态推理过程。作者构建了 SDKG-11, 一个 SDKG 集包括 5 种癌症、6 种非癌症疾病、1 种联合 Cancer5 和 1 种联合 Disease11。作者通过实体预测任务对作者的多模态 KGE 模型进行了评估，并在一些实例中进行了验证。然后，作者证明了学习嵌入在下游生物医学任务中的影响作用。所有这些都表明，作者所推理出的新知识是可靠的，作者的嵌入式学习是普遍的。这对于某些特定疾病领域的研究和临床工作人员具有一定的帮助。

解读：

随着人工智能技术的不断发展，多模态数据的应用越来越广泛。多模态数据包括文本、图像、语音等不同形式的信息，能够提供更全面和丰富的知识。然而，多模态数据的复杂性和异构性给其理解和应用带来了挑战。基于知识图谱的多模态推理能够有效地帮助机器理解和推理多模态数据。

基于知识图谱的多模态推理是指利用知识图谱中结构化的数据和多模态数据（如文本、图像、语音等）进行推理和解读。其中知识图谱是一种将实体、属性和关系以图形的形式表示的知识表示方法，能够帮助机器理解和推理知识。而多模态数据中则包含了不同形式的信息，它们之间可以互补，为模型提供更全面和丰富的知识。

在基于知识图谱的多模态推理中，我们首先需要处理大量的数据，对多模态数据进行处理和特征提取，将其转化为机器可理解的形式。对于文本数据，可以进行文本向量化或者自然语言处理等处理方式，将文本转化为机器可理解的形式。对于图像数据，可以进行图像特征提取，提取图像的特征向量。对于语音数据，可以进行声音信号处理，提取语音的特征。

然后再基于结构化的数据，构建知识图谱。我们可以定义实体、属性和关系，并将其以图的形式表示出来。实体可以表示人、物、地点等，属性可以表示实体的特征或者属性，关系可以表示实体之间的关联关系。知识图谱的构建可以通过手动构建、自动构建或者结合两者的方式进行。

然后需要将多模态数据与知识图谱进行关联。将多模态数据中的实体、属性和关系与知识图谱中的实体、属性和关系进行对应，建立起它们之间的关联关系。这样可以将多模态数据的信息与知识图谱的结构化数据进行融合，提供更全面和准确的知识。

最后，通过知识图谱中的结构化数据和关系，进行推理和解读。利用知识图谱中的关系和属性，结合多模态数据进行推理，得出结论或提供解释。

基于知识图谱的多模态推理可以应用于多个领域。例如在自然语言处理领域，我们可以利用知识图谱进行文本理解和语义分析；在计算机视觉领域，我们可以利用知识图谱进行图像识别和分类；在语音识别领域，我们可以利用知识图谱进行语音识别和语音合成。本文中所给的例子是关于生物医药领域。它

可以帮助机器更好地理解和推理多模态数据，提高多模态数据的理解能力和应用效果。在实际应用中具有非常广泛的应用前景。