# 4.新冠科研图谱

## 4.1 简介

1.病毒分类图谱 v2.0，根据最新 ncbi 网站内容更新了病毒族系相关的科研基础数据，包括了所有病毒的 lineage 分类信息。

2.新冠基本信息图谱 v1.0, 从新冠病毒的基因蛋白，宿主等相关信息出发构图。

3.抗病毒药物图谱 v1.0，依据 drugbank 数据库中抗病毒药物、病毒、病毒相关蛋白和宿主及宿主蛋白间的关系构建图谱。

4.新冠亲缘关系图谱 v1.2，整理 nextstrain 网站上短期内 SARS-CoV-2病毒株的变异方向与地理位置分布数据，截止 2020.4.7。

5.新冠文献抽取图谱 v1.0, 为了帮助研究人员方便快捷地查询最新研究信息，将 NLP 方法和知识图谱结合，自动化从专业文献等非结构化数据中抽提SARS-CoV-相关的知识点，整合为相关科研知识图谱。发布日期2020.4.19

**多个图谱各有侧重，为后续的图谱融合以及进一步研究提供数据支撑。**

## 4.2 概要介绍

SARS-CoV-2 是最新爆发的一种严重威胁人类健康的新型冠状病毒，它具有快速传染的特点，但是目前针对该病毒的已知信息很少，例如不了解其致病机制、传播机制、抑制机制等。

美国国家生物技术信息中心（NCBI）提供了大量有关生物信息和数据的在线资源，包括GenBank核酸序列数据库以及相关已发表生命科学期刊引用和摘要的PubMed数据库。在第一阶段的发布版本中，我们从病毒的生物学分类角度出发，以 NCBI中Taxonomy 板块下数据为基础，构建了一个所有病毒的族系网络——病毒分类图谱。其中不同的病毒通过生物学分类方法（例如界，门，纲，目，科，属，种等）进行链接和。该图谱具有较大的规模，实例的数量达20w以上，三元组的数量更是达到了190w+。但是其在分类及族系关系以外未引入其他信息，具有一定的局限。

我们通过与医药、生物学相关科研人员的交流及合作，从专业的角度了解病毒差异的来源以及区分/联系不同病毒的重要属性；同时，通过与华为云语音语义创新 Lab 及华为云医疗智能体团队开展合作，从不同科研内容角度展开多方面的工作，期望通过双方深入的交流及合作，共同构建并完善用于科研用途的知识图谱。当前工作建立在不断完善 schema 的基础上, 通过梳理病毒、基因、蛋白、药物等相关概念，围绕新型冠状病毒、抗病毒药物等展开工作。从病毒的基因，蛋白，宿主以及核苷酸序列等相关信息出发，基于 NCBI 数据库中新冠病毒相关具体数据，构建l 新冠基本信息图谱。

drugbank是一个基于Web的数据库，其中包含有关药物，其作用机理，相互作用和靶标的全面分子信息。药物对病毒的作用主要来自相关蛋白或人类免疫系统的抑制。我们从抗病毒药物、Human Protein、Virus Protein、宿主等信息出发，基于 DrugBank 等数据库的数据，构建了抗病毒药物图谱。

GISAID是目前全球最大的流感及新型冠状病毒数据平台，是由全世界一组权威的医学科学家组建，该组织致力于改善流感数据的共享。考虑到病毒流行学是重要的科研方向，我们从该角度出发，以 Gisaid 数据库所提供的实时新冠病毒 sequence 数据为基础，结合 nextstrain 对于新冠病毒基因组流行病学的分析及相关算法, 整理出了一个短期病毒突变的连接网络——病毒亲缘关系图谱（持续更新）。其中包含了地理，时间，变异度，突变的基因蛋白等信息，为后续的相关研究(病毒溯源)及图谱的融合提供支撑。

近几个月，随着大量研究人员对该病毒的多方位深入科学研究，研究者对于该病毒的了解程度逐渐加深。但是由于这些研究成果大多以专业文献形式发表于各种学术期刊，其中涵盖的关键知识并未被系统地梳理，不便于研究人员利用这些最新的研究发现。因此，为了帮助研究人员更加方便快捷地查询最新的研究信息，我们希望通过 NLP 和知识图谱结合的方法，自动从专业文献中抽提与 SARS-CoV-2 相关的知识点，并利用知识图谱整合各知识点涵盖的重要信息，整合为与 SARS-CoV-2 有关的文献抽取科研知识图谱（持续更新）。

## 4.3 潜在应用

在后续对于其他属性进行补充与关联，扩充图谱之后，数据集可用于尝试以下应用：

（1）预测新病毒的生物学分类？新物种发现？

（2）预测病毒变异性？高/低？

（3）预测病毒热稳定性？

（4）预测病毒的易感群体、宿主？

（5）预测病毒的致病部位？

（6）预测病毒可导致的症状？可缓解症状的药物？

（7）潜在治疗的药物？老药新用？

（8）预测病毒的传播途径？传播种类?

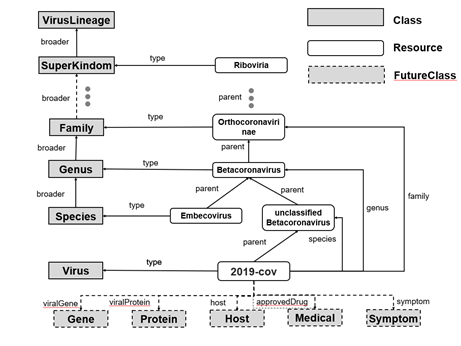
（9）预测可能与病毒蛋白相互作用的蛋白，发现新的蛋白靶点？

（10）针对新的蛋白靶点，进行药物开发？

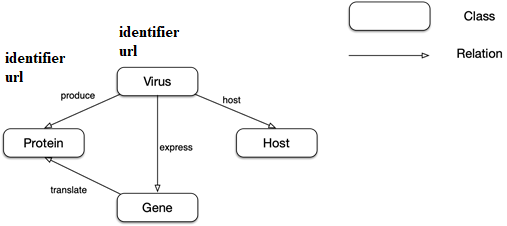
（11）病毒溯源，病毒变异的分析与预测？

## 4.4 数据 schema

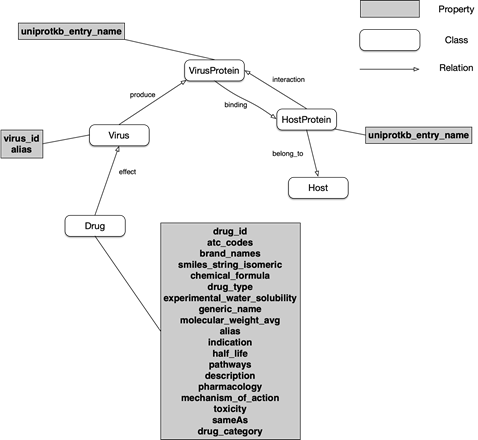
（1）病毒分类图谱



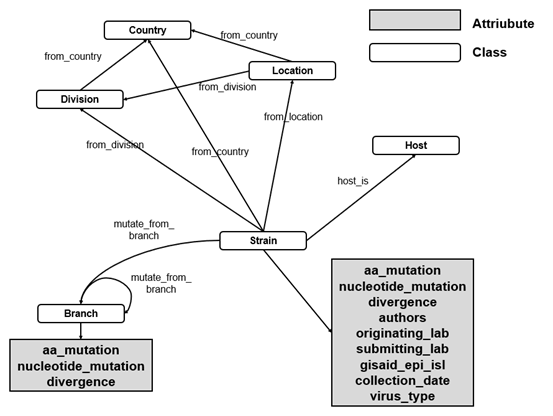
（2）新冠基本信息图谱



（3）抗病毒药物图谱



（4）新冠亲缘关系图谱



（5）新冠文献抽取图谱

## 

## 4.5 图谱可视化

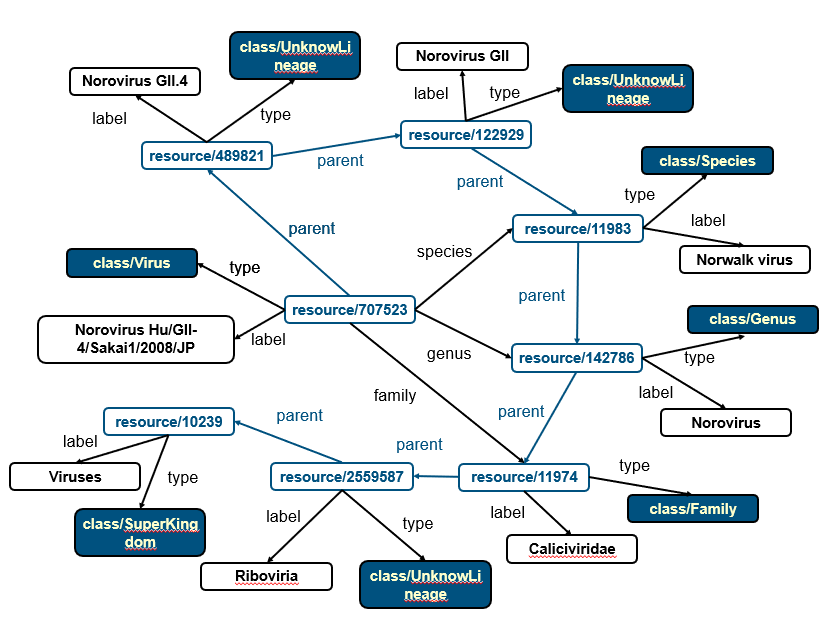
（1）病毒分类图谱

parent 关系用于连接 Resource 之间的层级关系，比如 COVID-19 病毒在其家族树种的上层结点是 unclassified Betacoronavirus。通过 parent 关系可向上遍历家族树。

species、genus、family 等关系用于直接连接病毒与其属类别、种类别结点。这样即可不通过遍历家族树直接找到其属。

broader 用于连接不同的家族概念，比如 Genus 属就是 Species 种的一个更宽泛的家族概念。

虚线中的关系和概念是未来工作中会添加的关系和连接的实体类型。比如病毒的基因实体、蛋白质实体、宿主实体等等。

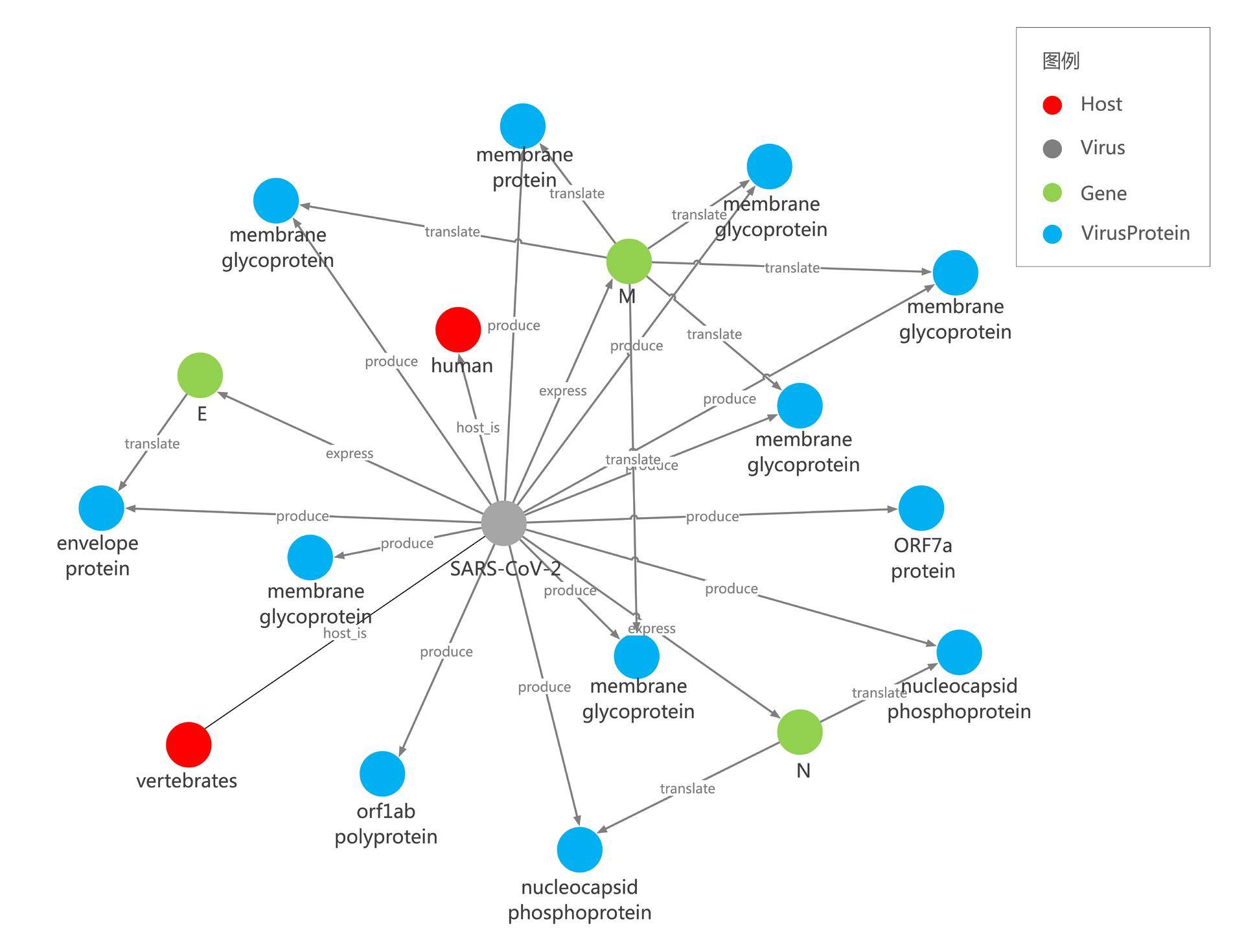


（2）新冠基本信息图谱

如下图可视化展示，express 表示 SARS-CoV-2 与其表达基因间的关系，例如 SARS-CoV-2 表达 membrane protein gene (M), Envelope protein gene(E) 等类型的基因；

translate 表示基因与蛋白间的翻译关系，例如 SARS-CoV2 的 M 基因可翻译出 membrane protein；另外，我们又用 produce 关系表示 SARS-CoV-2 可产生对应的蛋白，例如产生 orf1ab polyprotein。

为了进一步表示 SARS-CoV-2 和宿主的关系，我们使用 host 表示，目前图谱中的宿主实体主要是 human 和 vertebrates。



（3）抗病毒药物图谱

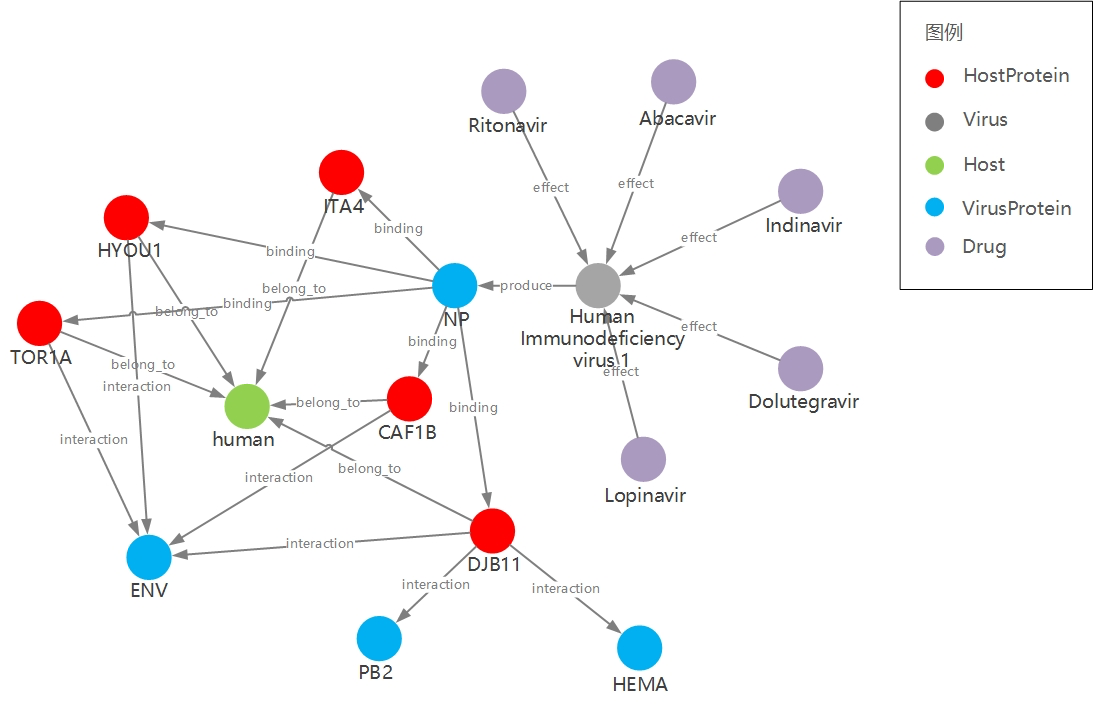
如下图可视化展示，effect 关系用于表示抗病毒药物对于病毒有一定效果，例如 Abacavir 和 Lopinavir 对 Human  immunodeficiency virus 1 有抗病毒效果，通过本关系可查看抗病毒药物和对应病毒间的联系；

produce 关系可用于挖掘病毒与其所表达蛋白间的 express 关系，并可进一步挖掘不同病毒与同一类型蛋白间的关系，进而发现两个病毒间的间接关系；再通过某病毒与某抗病毒药物间关系，又可发现某抗病毒药物和另一病毒的间接关系；可以帮助发现对于某一类型病毒蛋白可能有作用的更多药物。

Interaction 和 binding 关系可以用于发现病毒蛋白与宿主蛋白间的相互作用，比如 HIV1 的 NP（nucleoprotein）和 Human 的 HYOU1 和间的 binding 关系；之后可再补充某些药物和 HYOU1 间的 target 关系，进而可以研究这些药物和 HIV1 的 NP（nucleoprotein）间的作用.

通过 belong to 关系，可以明确哪些蛋白属于哪类宿主，这类宿主的哪些蛋白与病毒有相互作用关系等。

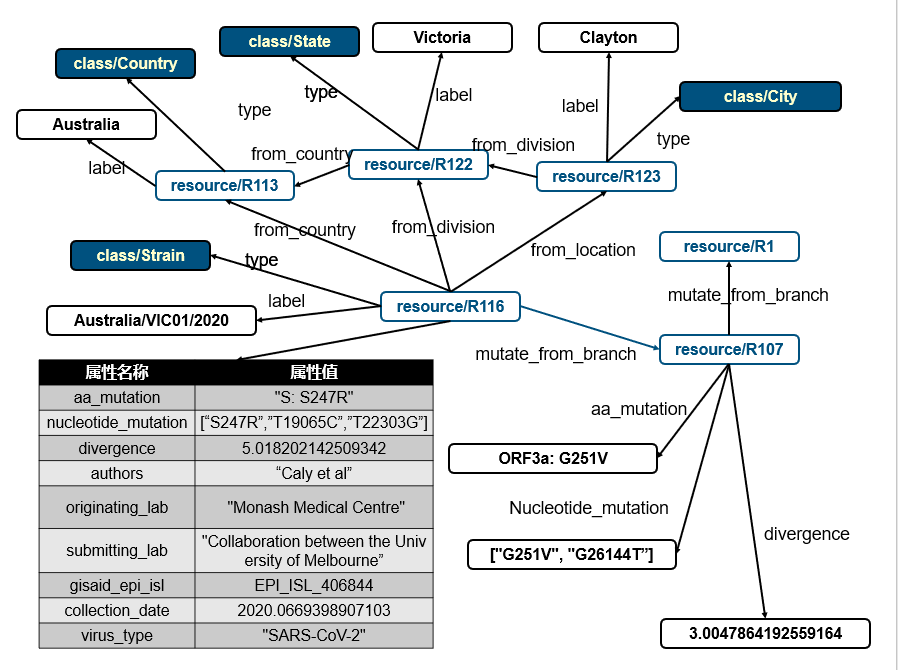
在属性方面，我们又进一步从 DrugBank 挖掘了这些信息（商品名、化学式，分子重量，indication 等），作为 drug 实体的属性，使得研究者可以更直观的了解某 drug 的一些重要信息。之后我们会进一步完善补充更多关键实体的重要信息。



（4）新冠亲缘关系图谱

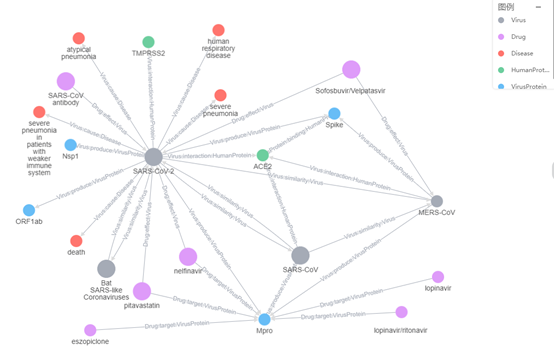
如图所示，图中有 Strain 毒株，Branch 分支，Country 国家，State 区域和 City 城市这些实体类型，其中地理位置与 Strain 毒株的连接通过 from\_country, from\_division 和 from\_location 来标识。Country，State，City 互相连接构成网络。

对于 Strain 毒株的属性，包括了 AA 变异，核苷酸突变，与上一分支的差异率等等信息，展示在图中的表格中。其中变异数据和差异率数据都是相对 mutate\_from\_branch 指示的 Branch 结点来说的，而 Branch 也相对上层的 Branch，因此最终的 Strain 突变应当是整个树目录突变的总和。



（5）新冠文献抽取图谱

从病毒及与病毒相关的多种类型生物分子/化学分子出发，梳理其关系，构建SARS-CoV-2知识图谱。例如病毒和疾病症状的关系、病毒所产生的病毒蛋白种类、病毒蛋白和宿主蛋白之间的关系、药物对病毒蛋白的抑制关系等等，多种类型的实体知识点，以及知识点之间的关系。

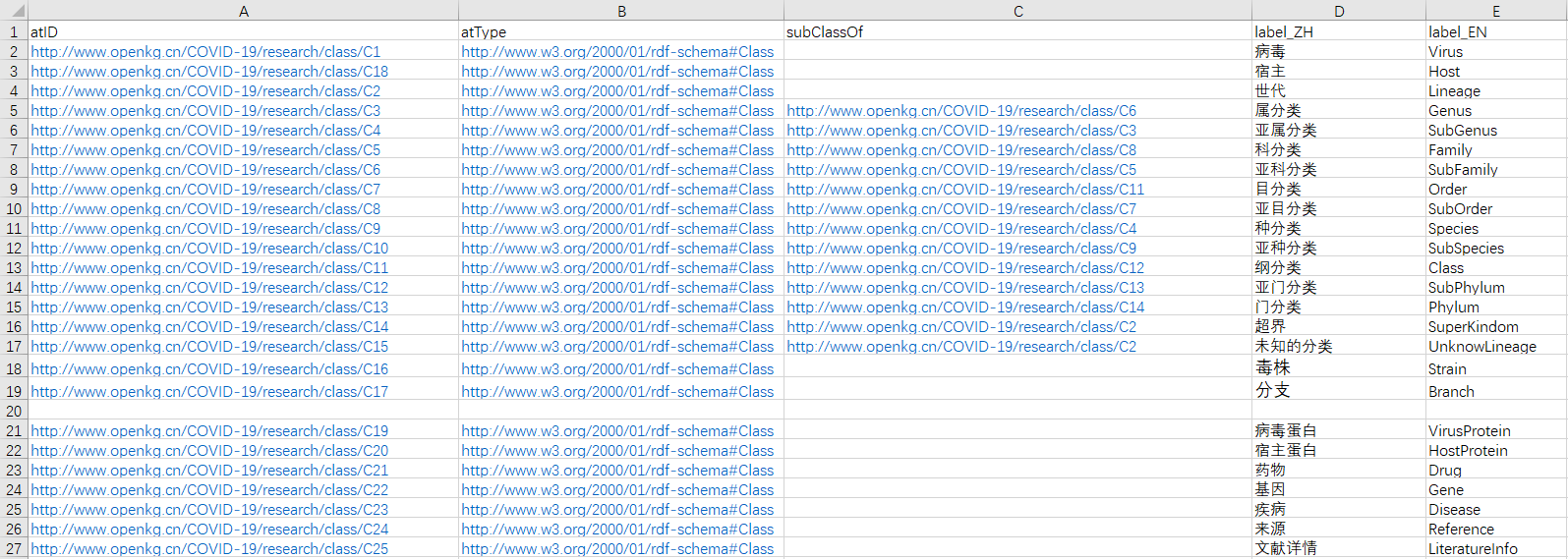


## 

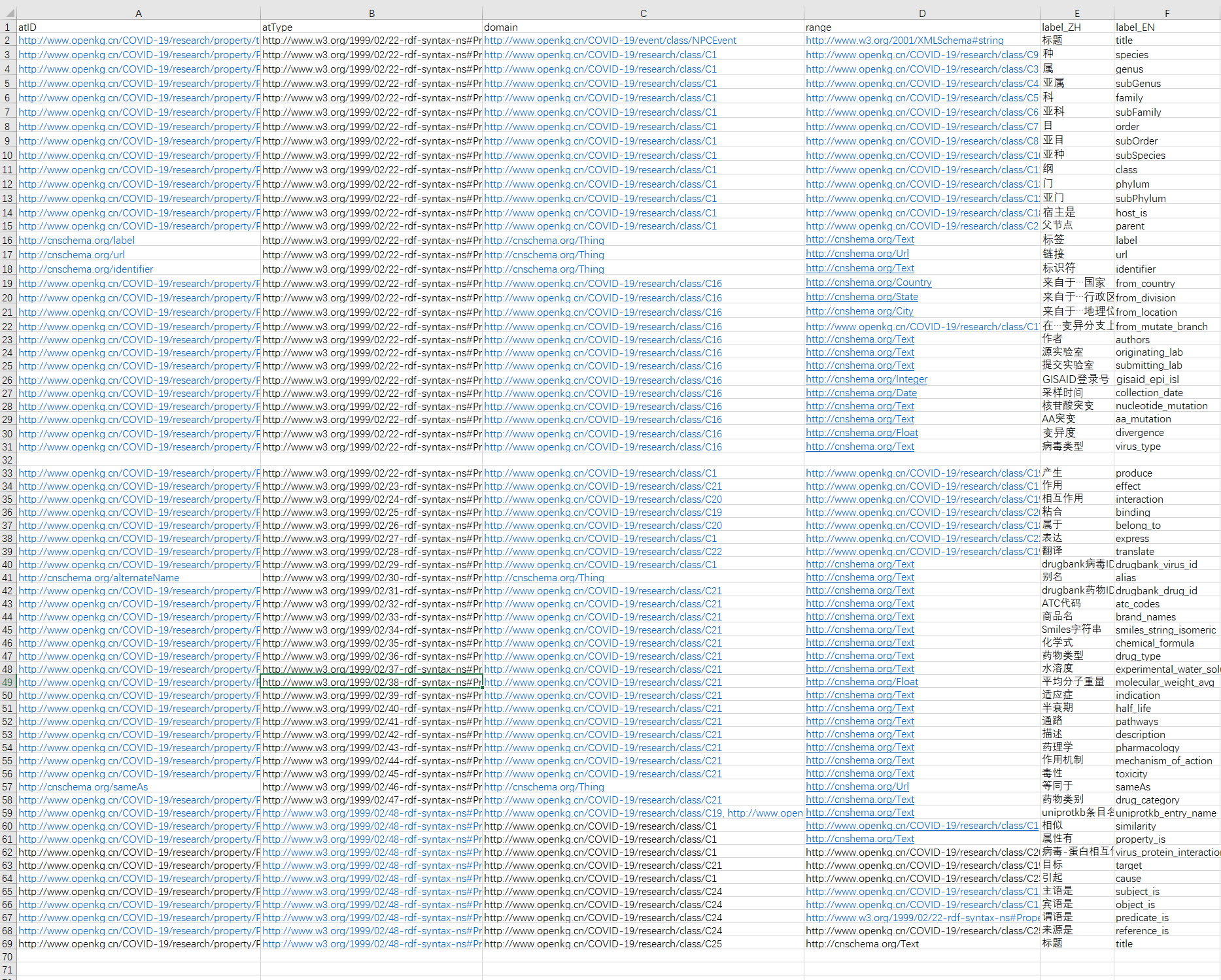
## 4.6schema 简要说明及示例

**总体 schema：(2020.04.19更新)**

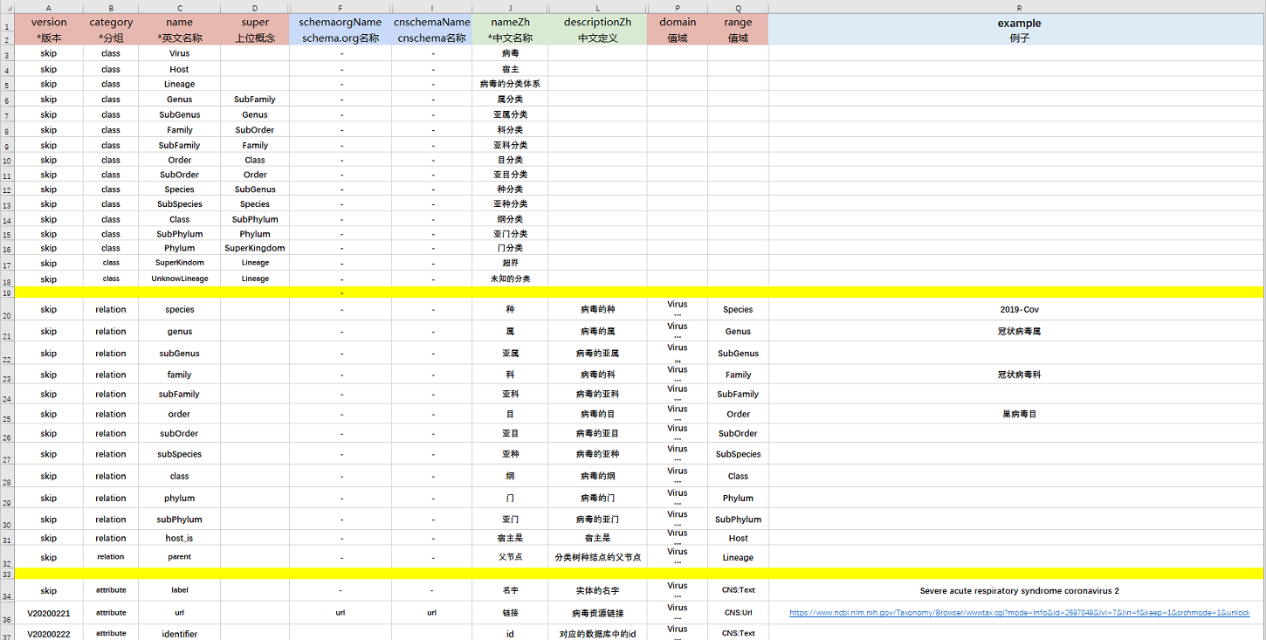
Class:



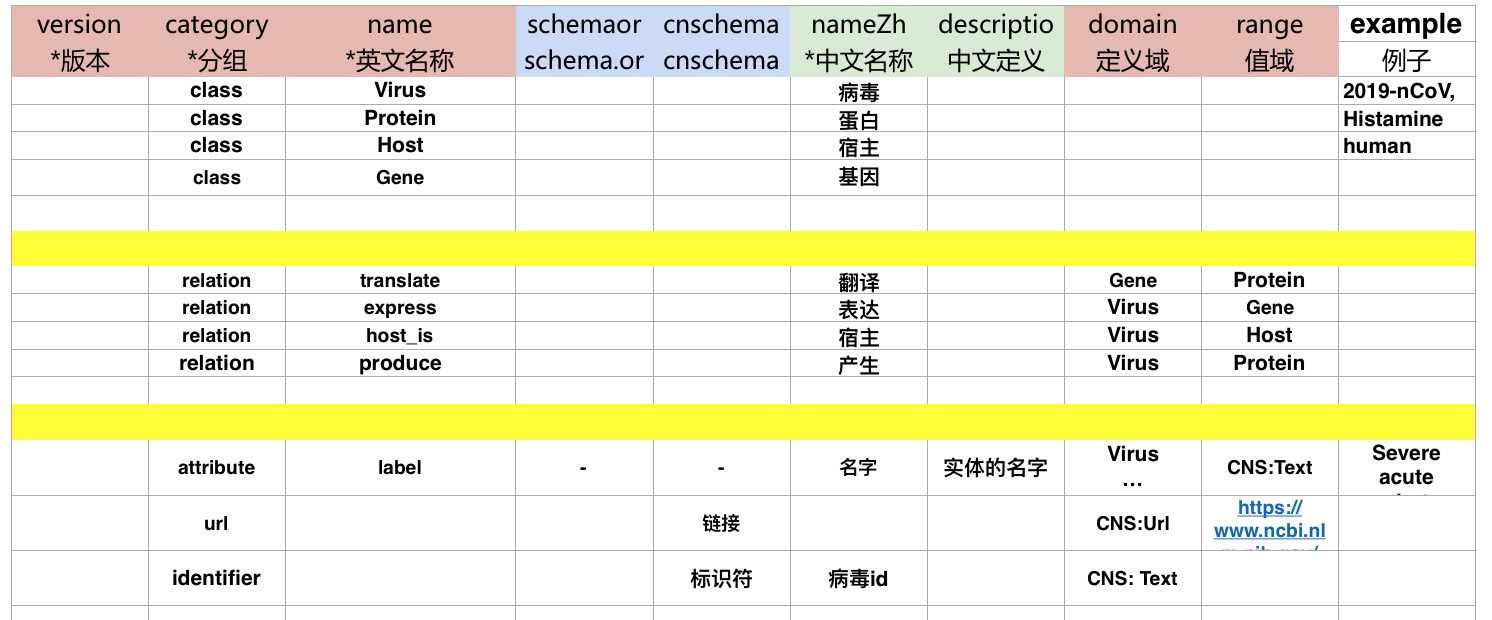
Property



（1）病毒分类图谱



（2）新冠基本信息图谱



（3）抗病毒药物图谱



**（4）新冠亲缘关系图谱**



（5）新冠文献抽取图谱



## 4.7 数据来源及规模

**数据来源：**

o   NCBI 中的 Taxonomy 板块    截止至 2020.2.21

<https://www.ncbi.nlm.nih.gov/Taxonomy>

o   全球流感数据库           截止至 2020.2.21

<https://www.gisaid.org>

o   Nextstrain 网站           截止至 2020.4.07

<https://github.com/nextstrain/ncov>

o   DrugBankVersion 5      截止至 2019.08

o   NCBI Taxonomy数据   截止至 2020.2.20

o   VirHostNet2.0        截止至 2019.01

o   Virus Pathogen Database 截止至 2020.02

o   [www.biorxiv.org](http://www.biorxiv.org)      截止2020.02.26 (39篇关于SARS-nCoV-2的论文)

**数据规模：**

* 病毒分类图谱 v2.0

o   概念：16 个

o   实例：205500 个

o   对象属性：15 个

o   三元组数目：2140191 个

* 新冠基本信息图谱 v1.0

o   概念：4 个

o   实例：470 个

o   对象属性：7 个

o   三元组数目：2096 个

* 抗病毒药物图谱 v1.0

o   概念：5 个

o   实例：7845 个

o   对象属性：27 个

o   三元组数目：60662 个

* 新冠亲缘关系图谱 v1.2

o   概念：6 个

o   实例：7316 个

o   对象属性：15 个

o   三元组数目：78848 个

* 新冠文献抽取图谱 v1.0

o   概念：6 个

o   实例：95 个

o   对象属性：9 个

o   三元组数目：358 个

## 4.8 规范

新冠科研图谱采用统一规范，定义如下。

命名空间： <http://www.openkg.cn/COVID-19/research/>

属性定义： <http://www.openkg.cn/COVID-19/research/property/>

概念定义： <http://www.openkg.cn/COVID-19/research/class/> {概念标识}

实体定义： <http://www.openkg.cn/COVID-19/research/resource/> {实体标识}

* + 概念及实体标识采用 Base64 编码

## 4.9 作者介绍

**浙江大学：**

陈卓   chenzhuo98@zju.edu.cn   浙江大学计算机学院

吴杨   yancy\_wu@outlook.com   浙江大学计算机学院

邵鑫   xin\_shao@zju.edu.cn    浙江大学药学院

杨海宏 capriceyhh@zju.edu.cn 浙江大学计算机学院

毕祯 bz6046190@icloud.com 浙江大学计算机学院

叶宏斌 yehb@zju.edu.cn 浙江大学计算机学院

方尹 fang\_yin941105@163.com 浙江大学计算机学院

杨帆 yangfan9456@163.com 浙江大学计算机学院

陈华钧 huajunsir@zju.edu.cn    浙江大学计算机学院

**华为云团队：**

郑毅 zhengyi29@huawei.com  华为云语音语义创新 Lab

张雷 zhanglei359@huawei.com 华为云医疗智能体

王鹏 wangpeng231@huawei.com 华为云语音语义创新 Lab

卢栋才 ludongcai@huawei.com 华为云语音语义创新Lab

章涛 zhangtao202@huawei.com 华为云语音语义创新 Lab

刘登辉 liudenghui2@huawei.com 华为云医疗智能体

徐迟 xuchi15@huawei.com 华为云医疗智能体

袁晶 nicholas.yuan@huawei.com 华为云语音语义创新 Lab

乔楠 qiaonan3@huawei.com 华为云医疗智能体

怀宝兴 huaibaoxing@huawei.com 华为云语音语义创新 Lab

* 华为云语音语义创新 Lab：面向行业客户构建业界领先的语音语义技术，基于华为云平台提供自然语言处理、知识图谱、对话机器人、语音识别等云服务、解决方案及核心技术支持。
* 华为云医疗智能体：医疗智能体平台基于华为云 AI 和大数据技术优势，为基因组、医疗影像、药物研发三个领域提供专业 AI 研发平台。