# 4.新冠科研图谱

# 4.COVID-19 Research KG

## 4.1 简介

## 4.1 Introduction

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。

1. **Virus taxonomy KG v2.0**. The basic research data of virus families has been updated according to the latest ncbi website content, including lineage classification information of all viruses.

2.**SARS-CoV-2 gene-protein KG v1.0**. Based on the SARS-CoV-2 gene protein, host and other related information.

**3. Antiviral drug KG v1.0**. Construct a KG based on the relationship between antiviral drugs, viruses, virus-related proteins, the host and host proteins in the drugbank database.

4. **SARS-CoV-2 phylogeny KG v1.2**. Collating The short-term variation of SARS-CoV-2 virus strain on the nextstrain website and the geographical distribution data. (as of 2020.4.7).

5. **SARS-CoV-2 literature extraction KG v1.0**. Combined The NLP method and KG to automatically extract SARS-CoV-related knowledge from unstructured data such as professional literature and integrate them into related scientific research KG.(as of 2020.4.19)

Each KG has its own emphasis, providing data support for KG fusion and further research.

## 4.2 潜在应用

## 4.2 possible applications

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？

After supplementing and correlating other attributes and expanding the graph, the dataset can be used to try the following applications:

（1）Predict the biological classification of new viruses? New species discovered?

（2）Predict virus variability? High or low?

（3）Predict virus thermal stability?

（4）Predict the susceptible group and host of the virus?

（5）Predict the pathogenesis of the virus?

（6）Predict the symptoms that the virus can cause? Possible drugs that relieve symptoms?

（7）Potential treatment drugs? Old medicine for new use?

（8）Predict the spread of the virus? Type of transmission?

（9）Predict proteins that may interact with viral proteins and discover new protein targets?

（10）Drug development for new protein targets?

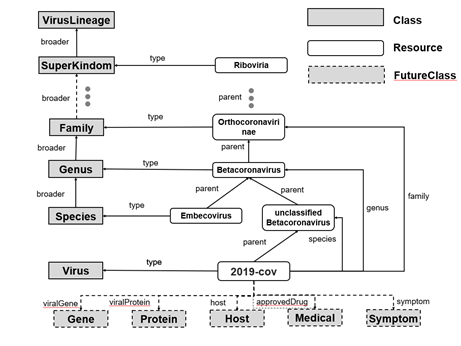
（11）Virus tracing, analysis and prediction of virus variation?

## 4.3 数据 schema

## 4.3 Data schema

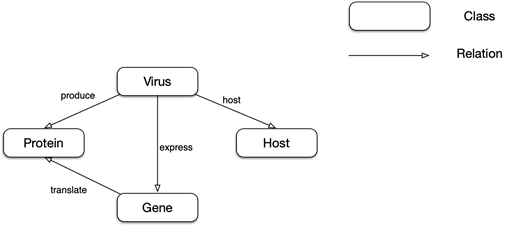
（1）病毒分类图谱

（1）Virus taxonomy KG



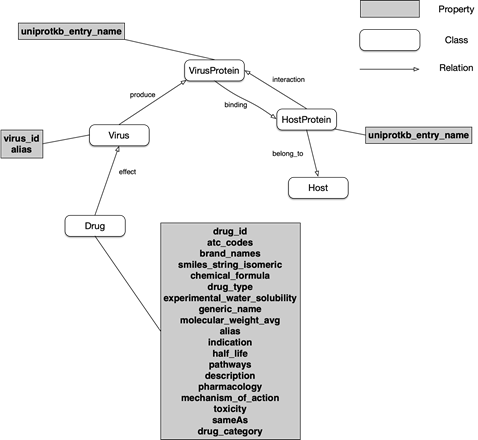
（2）新冠基本信息图谱

（2）**SARS-CoV-2 gene-protein KG**



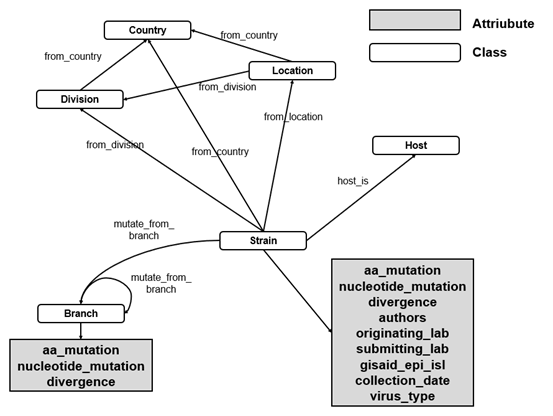
（3）抗病毒药物图谱

（3）**Antiviral drug KG**



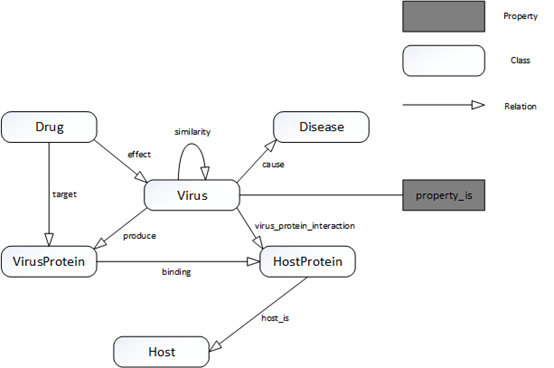
（4）新冠亲缘关系图谱

（4）SARS-CoV-2 phylogeny KG



（5）新冠文献抽取图谱

（5）SARS-CoV-2 literature extraction KG



## 4.4 图谱可视化

**4.4 KG Visualization**

（1）病毒分类图谱

（1）Virus taxonomy KG

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。

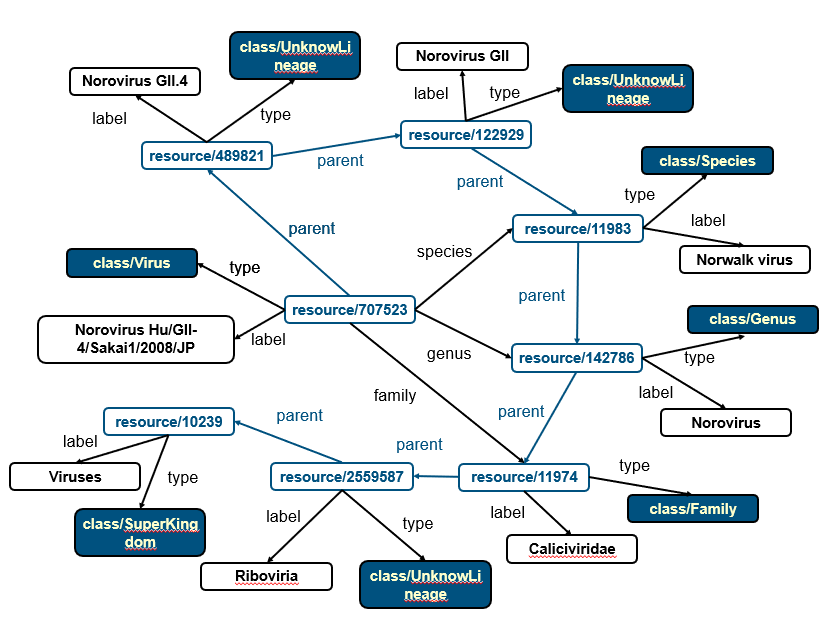
虚线中的关系和概念是未来工作中会添加的关系和连接的实体类型。比如病毒的基因实体、蛋白质实体、宿主实体等等。

The **parent** relationship is used to connect the hierarchical relationship between Resources. For example, the upper node of the COVID-19 virus in its family tree species is **unclassified Betacoronavirus**. The family tree can be traversed upward through the **parent** relationship.

**Species, genus, family** and other relationships are used to directly connect the virus to its genus and species node. In this way, you can directly find its genus without traversing the family tree.

**Broad** is used to connect different family concepts. For example, **Genus** is a broader family concept of Species.

The relationships and concepts represented by dotted line are the types of relationships and connected entities that will be added in future work. For example, **viral gene entities, protein entities, host entities, etc**.



（2）新冠基本信息图谱

（2）SARS-CoV-2 gene-protein KG

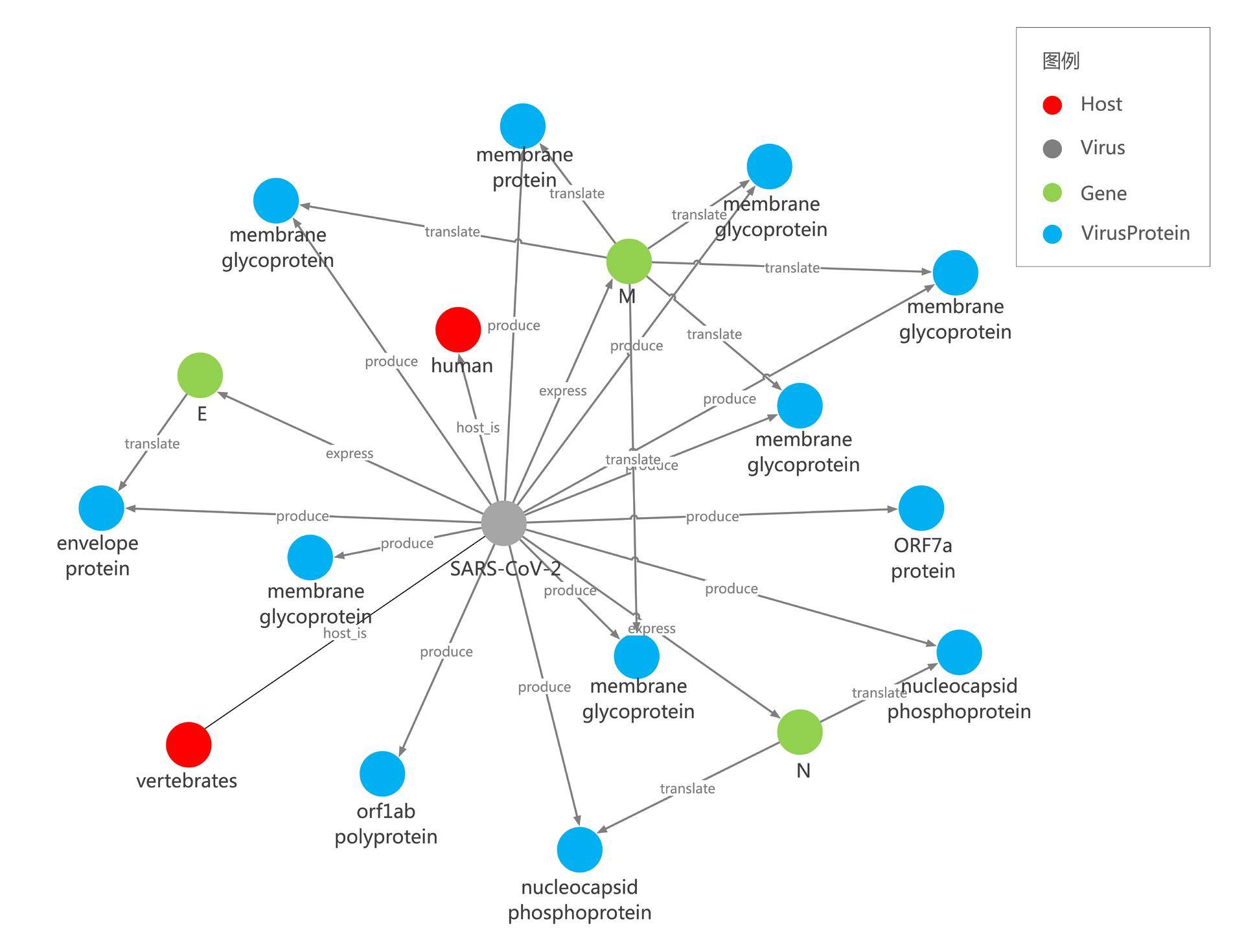
如下图可视化展示，express 表示 SARS-CoV-2 与其表达基因间的关系，例如 SARS-CoV-2 表达 membrane protein gene (M), Envelope protein gene(E) 等类型的基因；

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。

As shown in the figure below, **express** indicates the relationship between **SARS-CoV-2** and its expressed genes. For example, SARS-CoV-2 expresses genes such as **membrane protein gene (M)**, **Envelope protein gene (E)**. **Translate** indicates the relationship between genes and proteins The translation relationship, for example, the **M gene** of SARS-CoV2 can be translated into membrane protein; in addition, we use the **produce** relationship to indicate that SARS-CoV-2 can produce the corresponding protein, such as orf1ab polyprotein.

To further express the relationship between SARS-CoV-2 and the host, we use **host** to indicate that the host entities in the current map are mainly **human** and **vertebrates**.



（3）抗病毒药物图谱

（3）Antiviral drug KG

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。

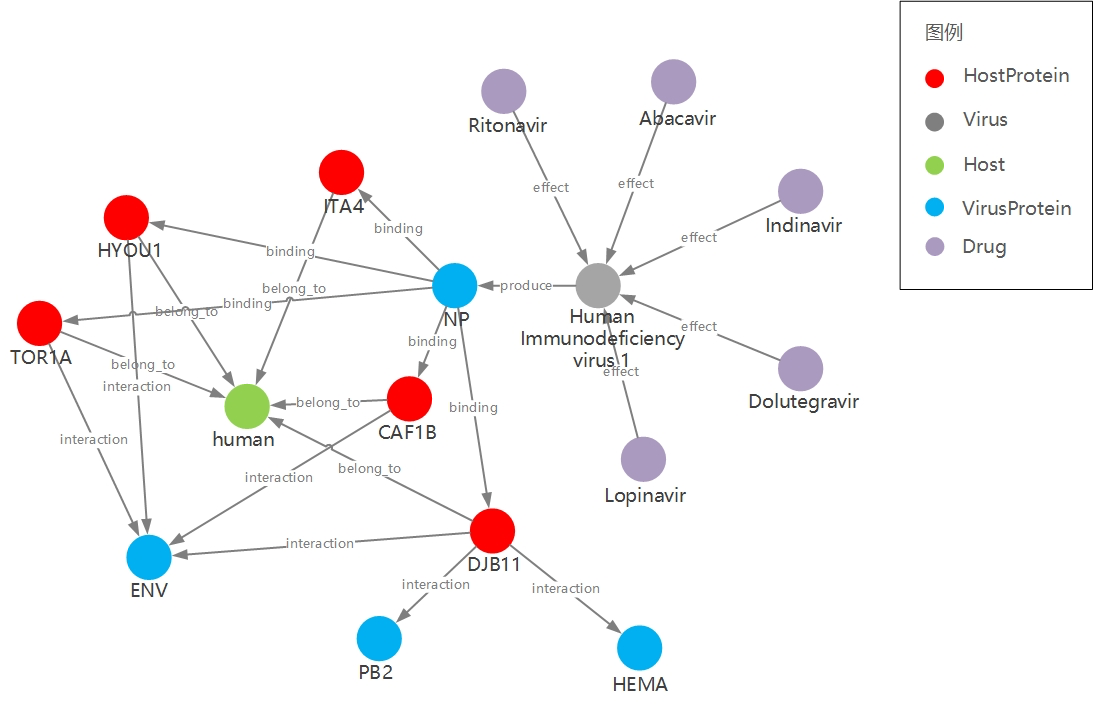
As shown in the figure below, the **effect** relationship is used to indicate that the antiviral drugs have a certain effect on the virus. For example, **Abacavir** and **Lopinavir** have antiviral effects on **Human immunodeficiency virus** **1.** Through this relationship, you can view the relationship between antiviral drugs and corresponding viruses

The **produce** relationship can be used to mine the **express** relationship between the virus and the protein it expresses, and can further mine the relationship between different viruses and the same type of protein, and then discover the indirect relationship between the two viruses. And then through the relationship between a virus and an antiviral drug it will be possible to find an indirect relationship between an antiviral drug and another virus, which could help find more drugs that may have an effect on a type of viral protein.

**Interaction** and **binding** relationships can be used to discover the interaction between viral proteins and host proteins, such as the binding relationship between **NP (nucleoprotein)** of **HIV1** and **Human HYOU1**. The target relationship between certain drugs and **HYOU1** can be added later, and then the role of these drugs and the **NP (nucleoprotein)** of **HIV1** can be reserched.

Through the **belong to** relationship, it can be clear which proteins belong to which type of host, and which proteins of this type of host have interaction with the virus.

In terms of attributes, we have further mined this information (product name, chemical formula, molecular weight, indication, etc.) from DrugBank as attributes of drug entities, so that researchers can more intuitively understand some important information of a drug. After that, we will further improve and supplement the important information of more key entities.



（4）新冠亲缘关系图谱

（4）SARS-CoV-2 phylogeny KG

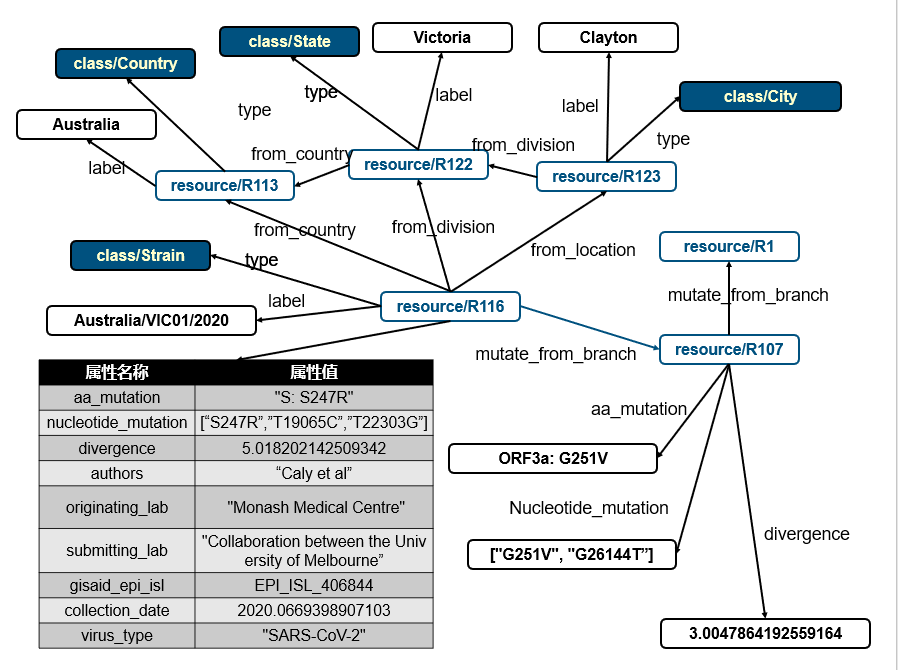
如图所示，图中有 Strain 毒株，Branch 分支，Country 国家，State 区域和 City 城市这些实体类型，其中地理位置与 Strain 毒株的连接通过 from\_country, from\_division 和 from\_location 来标识。Country，State，City 互相连接构成网络。

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。

As shown in the figure, there are entity types such as **Strain**, **Branch** , **Country**, **State** and **City**. The connection between geographic location and Strain is identified by **from\_country**, **from\_division** and **from\_location**. **Country**, **State**, **City** are connected to each other to form a network.

The attributes of the Strain include AA mutations, nucleotide mutations, the rate of difference from the previous branch, etc, which are shown in the figure. The variation data and the difference rate data are relative.

The **Branch** node indicated by **mutate\_from\_branch** is relative to the upper **Branch**, so the final Strain mutation should be the sum of mutations in the entire tree directory.

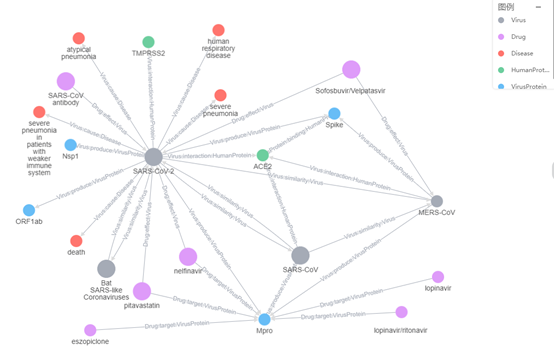


（5）新冠文献抽取图谱

（5）SARS-CoV-2 literature extraction KG

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。

This KG is constructed from viruses and various types of biomolecules/chemical molecules related to viruses, sorting out their relationships.Such as the relationship between viruses and disease symptoms, the types of viral proteins produced by viruses, the relationship between viral proteins and host proteins, the inhibitory relationship between drugs on viral proteins, etc. Various types of entity knowledge points, and the knowledge points relationship are included.



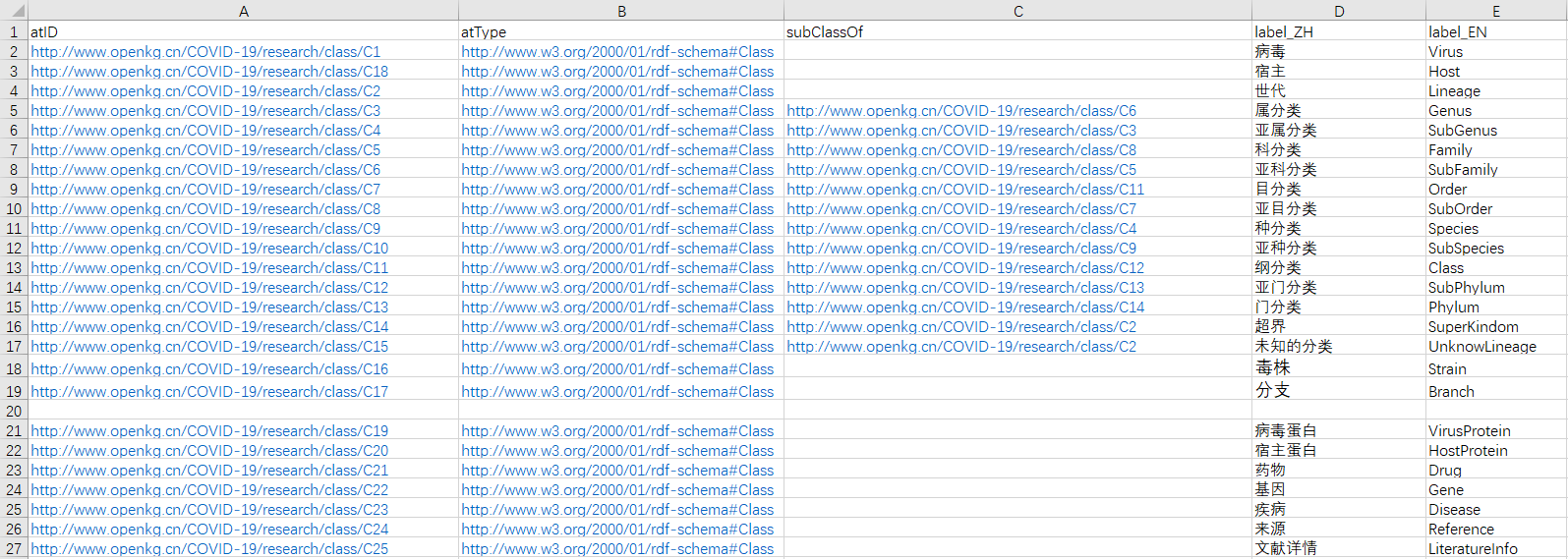
## 4.5 schema 简要说明及示例

## 4.5 Brief description and examples of schema

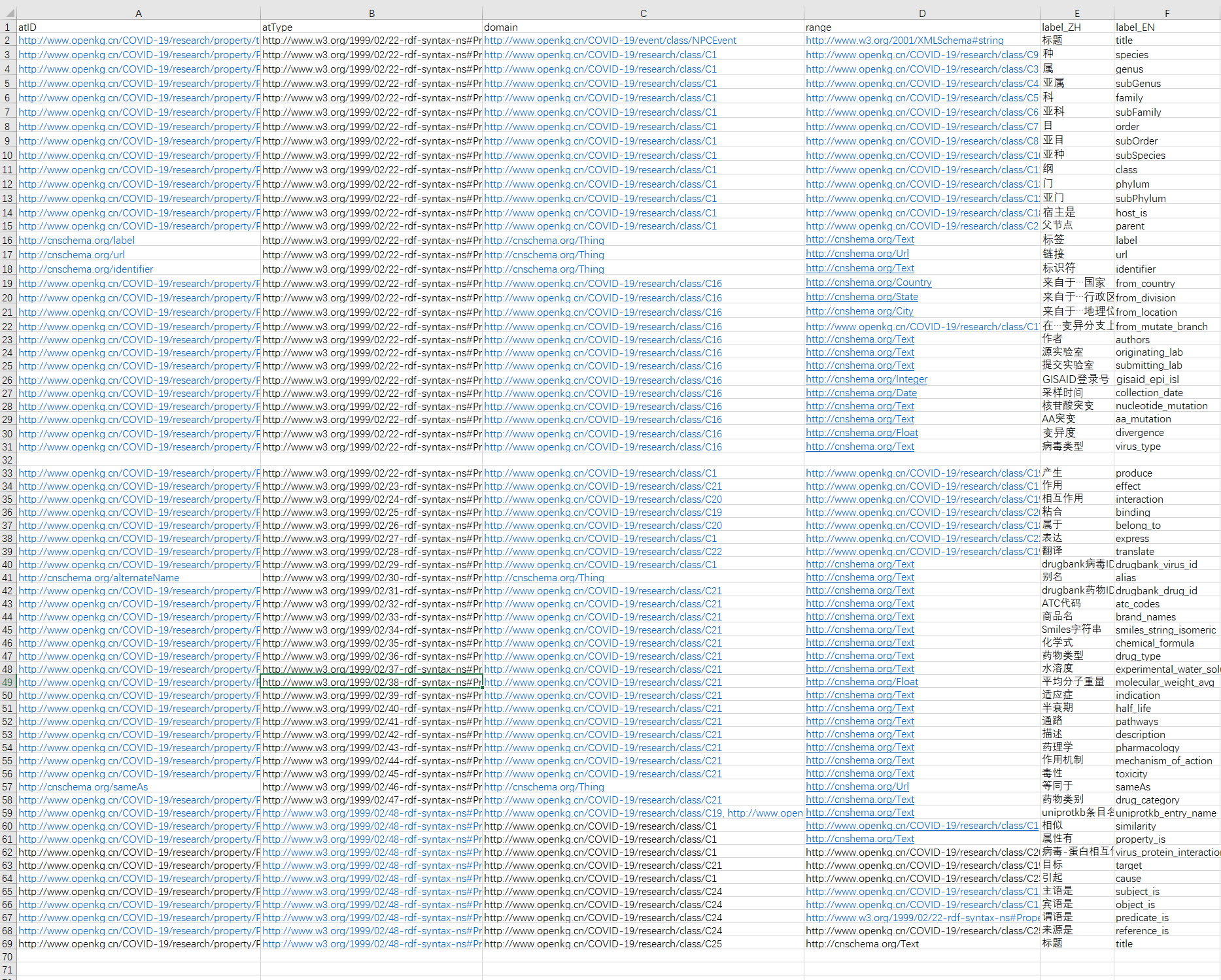
**总体 schema：(2020.04.19更新)**

**Overall schema: (updated on 2020.04.19)**

**Class:**

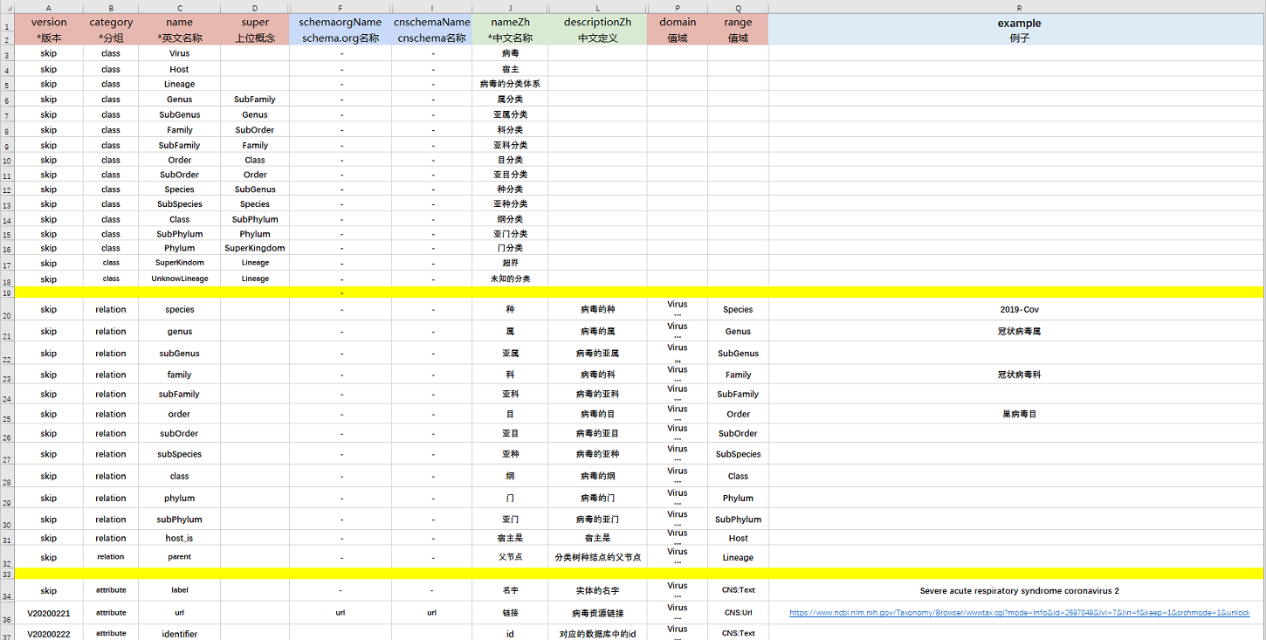


**Property**



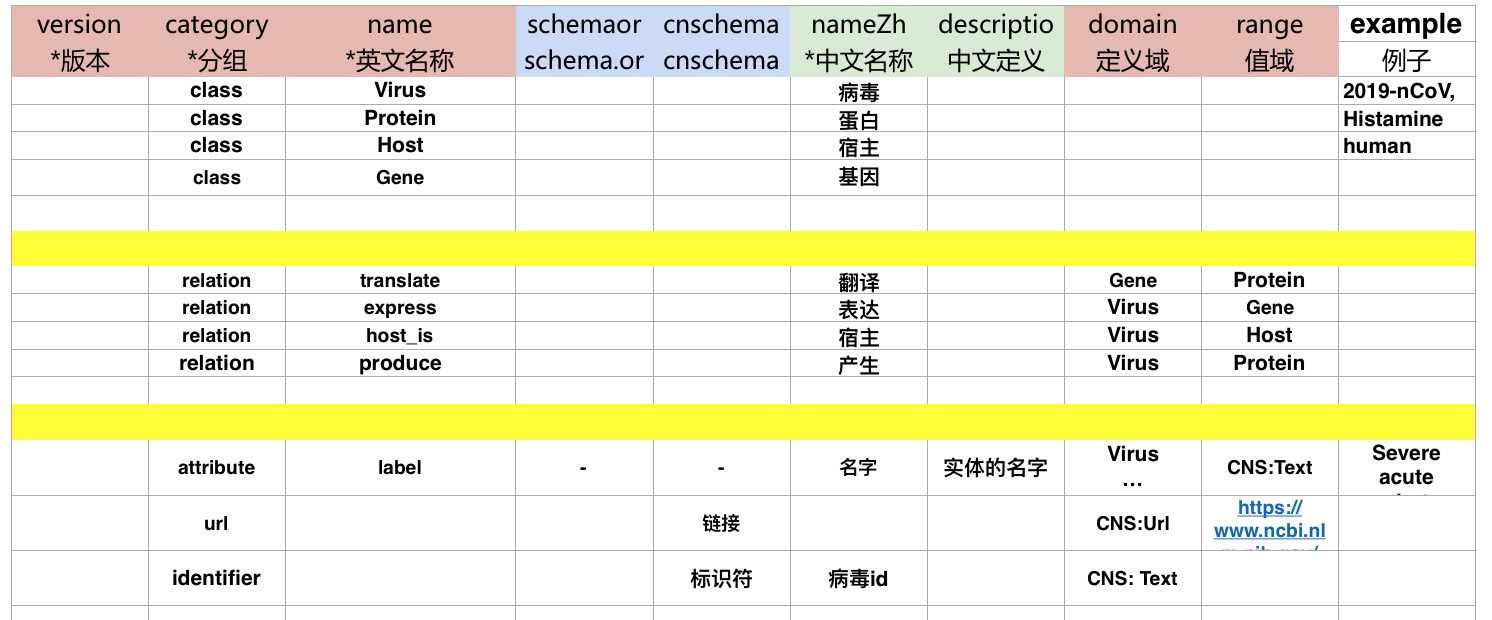
**（1）病毒分类图谱**

**（1）Virus taxonomy KG**



**（2）新冠基本信息图谱**

**（2）SARS-CoV-2 gene-protein KG**



**（3）抗病毒药物图谱**

**（3）Antiviral drug KG**



**（4）新冠亲缘关系图谱**

**（4）SARS-CoV-2  phylogeny  KG**



**（5）新冠文献抽取图谱**

**（5）SARS-CoV-2 literature extraction KG**



## 4.6 数据来源及规模

## 4.6 Data source and scale

**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   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

**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

**Data Sources:**

o  Taxonomy part of NCBI 2020.2.21

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

o   Gisaid database           2020.2.21

<https://www.gisaid.org>

o   Nextstrain Website           2020.4.07

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

o   DrugBankVersion 5      2019.08

o   VirHostNet2.0        2019.01

o   Virus Pathogen Database 2020.02

o   [www.biorxiv.org](http://www.biorxiv.org)      2020.02.26

**Data scale:**

* Virus taxonomy KG v2.0

o   Concept: 16

o   Examples: 205500

o   Object attributes: 15

o   Triples: 2140191

* SARS-CoV-2  gene-protein  KG v1.0

o   Concept: 4

o   Examples: 470

o   Object attributes: 7

o   Triples: 2096

* SARS-CoV-2  gene-protein  KG v1.0

o   Concept: 5

o   Examples: 7845

o   Object attributes: 27

o   Triples: 60662

* SARS-CoV-2  phylogeny  KG v1.2

o   Concept: 6

o   Examples: 7316

o   Object attributes:15

o   Triples: 78848

* SARS-CoV-2 literature extraction KG v1.0

o  Concept: 6

o  Examples: 95

o  Object attributes:9

o  Triples: 358

## 4.8 图谱规范

## 4.8 knowledge graph specifications

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

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

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 ，属性名称采用 URL 风格编码

Namespace: <http://www.openkg.cn/COVID-19/research/>

Property definition: <http://www.openkg.cn/COVID-19/research/property/> {property name}

Concept definition: <http://www.openkg.cn/COVID-19/research/class/> {concept identifier}

Entity definition: <http://www.openkg.cn/COVID-19/research/resource/> {entity identifier}

* + Entities and concept identifiers are encoded in Base64, and attribute names are encoded in URL style

## 4.9 作者介绍

## 4.9 Author introduction

**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。

**Zhejiang University:**

Zhuo Chen   chenzhuo98@zju.edu.cn   School of Computer Science

Yang Wu   yancy\_wu@outlook.com   School of Computer Science

Xing Shao   xin\_shao@zju.edu.cn    College of Pharmacy

Haihong Yang capriceyhh@zju.edu.cn School of Computer Science

Zhen Bi bz6046190@icloud.com School of Computer Science

Hongbing Ye yehb@zju.edu.cn School of Computer Science

Yin Fang fang\_yin941105@163.com School of Computer Science

Fan Yang yangfan9456@163.com School of Computer Science

Huajun Chen huajunsir@zju.edu.cn    School of Computer Science

**Huawei Cloud Team:**

Yi Zheng zhengyi29@huawei.com  Semantic Innovation Lab

Lei Zhang zhanglei359@huawei.com Medical Intelligent Agent

Peng Wang wangpeng231@huawei.com Semantic Innovation Lab

Dongcai Lu ludongcai@huawei.com Semantic Innovation Lab

Tao Zhang zhangtao202@huawei.com Semantic Innovation Lab

Denghui Liu liudenghui2@huawei.com Medical Intelligent Agent

Chi Xu xuchi15@huawei.com Medical Intelligent Agent

Jing Yuan nicholas.yuan@huawei.com Semantic Innovation Lab

Nan Qiao qiaonan3@huawei.com Medical Intelligent Agent

Baoxing Huai huaibaoxing@huawei.com Semantic Innovation Lab

* The Language and Speech Innovation Lab of Huawei CloudBU builds cutting-edge AI technology for enterprises of various industries. Our scope spans across the vast majority of natural language and speech processing, such as fundamental algorithms in NLP, knowledge graphs, conversational AI bots, and speech recognition, etc. We provide our products as cloud services and industrial solutions.
* EIHealth in Huawei Cloud is powered by the advantages of AI and big data technologies from HUAWEI CLOUD, EIHealth provides a professional AI R&D platform to accelerate AI researches and applications in genomics, drug discovery and medical imaging.