BioThings Explorer：面向生物医学 API 联邦知识图谱的查询引擎

Jackson Callaghan*[,1], Colleen H. Xu*[,1], Jiwen Xin*[,1], Marco Alvarado Cano[1], Anders Riutta[2], Eric Zhou[1], Rohan Juneja[1], Yao Yao[1], Madhumita Narayan[1], Kristina Hanspers[2], Ayushi Agrawal[2], Alexander R. Pico[2], Chunlei Wu**[,1], AndrewI Su**[,1]

1 整合结构与计算生物学系，斯克里普斯研究所 2 数据科学与生物技术，格拉德斯通研究所、加利福尼亚大学旧金山分校，旧金山，加利福尼亚州94158，美国。

* - 贡献相同 ** - 通讯作者

摘要

知识图谱是越来越常见的生物医学信息表示数据结构。这些知识图谱能够轻松表示异构类型的信息，并且已有许多用于查询和分析图谱的算法与工具。生物医学知识图谱已应用于多种领域，包括药物再利用、药物靶点识别、药物副作用预测以及临床决策支持。通常，知识图谱通过集中化并整合来自多个不同来源的数据来构建。本文介绍 BioThings Explorer，这是一款应用程序，可查询虚拟联邦知识图谱；该图谱由生物医学网络服务网络汇总的信息构建而成。BioThings Explorer 利用对各资源输入和输出进行的语义精确注释，并自动串联网络服务调用，以执行多步图查询。由于无需维护大型集中式知识图谱，BioThings Explorer 以轻量级应用程序的形式分发，可在查询时动态检索信息。更多信息请访问 https://explorer.biothings.io，并且代码可在 https://github.com/biothings/biothings_explorer。

引言

虽然可下载文件是共享生物医学数据最常见的方式，但应用程序编程接口（API）也是一种流行且强大的数据传播机制。通过 API 访问数据，相比下载本地数据副本，具有许多互补优势。API 通常允许用户查询感兴趣的特定数据子集。API 查询通常经过高度索引，因此能够高效地检索数据。最后，通过 API 访问数据，可以轻松地将最新数据整合到其他计算应用或工作流中。

虽然 API 在数据可访问性方面具有诸多优势，但这些优势并不会立即转化为高效的数据集成。API 通常遵循一些常见的架构和协议（例如表述性状态转移（REST）、JavaScript 对象表示法（JSON）输出），但在这一技术层面保持一致，并不能保证语法或语义上的互操作性。例如，API 可能会使用不同的标识符来指代同一个基因，使用不同的数据结构来表示基因属性，并使用不同的术语来描述生物医学实体之间的关系。

为实现数据互操作性，人们已经开展了一些定义和执行语义及句法标准的工作。这种方法的示例包括 GA4GH 联盟推出的 Beacon API[1] ，以及用于共享基因组特征注释的 DAS 规范[2]。这些工作依赖 API 开发者的积极参与，以便为其 API 端点采用社区标准。

在此，我们探讨将语义精确的 API 注释作为实现 API 互操作性的一种补充方法。我们将这项工作分为两个部分。首先，我们扩展了广泛使用的 OpenAPI 注释标准[3] ，以便从语义上描述 API 的输入和输出，并创建了一个用于组织这些 API 注释的注册表。其次，我们创建了一个名为 BioThings Explorer 的应用程序，用于使用 API 元数据，并执行跨越多个 API 的多跳图查询。总体而言，该架构使用户能够基于相互连接的生物医学 API 网络，对大型联邦知识图谱进行查询。这种联邦式设计为创建知识图谱提供了一种独特方法，可作为更常见的集中化和本地数据整合策略的补充。

语义标注 API 注册表

构建可互操作的 API 网络，第一步是以语义精确的方式为每个 API 添加注释。我们基于 OpenAPI 规范[3],构建了这套 API 注释系统；OpenAPI 规范是用于以人类和机器可读格式记录 API 元数据的事实标准。OpenAPI 描述了基本的 API 元数据（例如标题、描述、版本和联系信息），以及有关 API 端点运行方式的关键信息（例如服务器 URL、端点输入参数和端点响应架构）。

然而，这种基本的 OpenAPI 规范并不包含促进下游 API 互操作性所必需的关键领域特定信息。因此，我们定义了一个 OpenAPI扩展[4] ，用于捕获每个 API 端点语义精确的注释。这些注释包括生物医学实体的语义类型和标识符命名空间（它们既用于查询〔输入〕，也出现在响应中〔输出〕）、JSON 响应中输出标识符值的 JSON 路径，以及描述输入实体与输出实体之间关系的谓词（**补充图1**）。

我们还选择了一种策略，将任意 JSON 数据结构映射到一种既定的生物学数据模型。在这项工作中，我们将 API输出映射到 Biolink模型[5], ，这是一种由社区开发、经 NCATS Translator联盟[6]采纳并扩展的数据模型。

为了使用我们的 OpenAPI扩展为 API 添加注释并编目，我们借助了 SmartAPI 注册表[7,8]。SmartAPI 注册表目前收录了 33 个带有语义注释的 API。这些 API 注释集合可视为一个“元知识图谱”（meta-KG），其中节点表示生物医学实体的类型（基因、疾病、药物），边则表示描述两种生物医学实体类型之间关系的 API。SmartAPI 元知识图谱目前包含 22 个节点和 1041 条边。（完整的元知识图谱见**补充表1**，部分呈现见**图2**。）

使用 BioThings Explorer 实现 API 互操作性

创建我们的联邦式生物医学知识图谱的第二步，是开发 BioThings Explorer。它是一种引擎，可自主查询 SmartAPI 元知识图谱、查询带有注释的 API 以检索生物医学实体之间的关联，并整合这些 API 的响应。BioThings Explorer 的输入是一个查询图，其编码语法由 NCATSTranslator 联盟[6]定义。查询图的拓扑结构以及其节点和边上的属性共同定义了查询（**图1**）。

BioThings Explorer 分三个阶段执行查询：查询路径规划、查询路径执行以及集成与评分。

查询路径规划。对于查询图中的每条边，BioThings Explorer 都会查询 SmartAPI 注册表，以寻找提供这些关联类型的 API（**图 1**）。例如，在图 1 中，可以使用比较毒理基因组学数据库[9] 和 Monarch Initiative[5,10],提供的 Biolink API 查找疾病与基因之间的关联，而基因与化学物质之间的关联则可以使用 MyChem.Info[11,12]查找。能够满足原始查询的 API 调用序列称为“查询路径计划”。

查询路径执行。在此阶段，BioThings Explorer 根据上一阶段识别出的每个 API 的语义注释，以程序化且自主的方式执行每个查询路径计划中的每项查询。BioThings Explorer 会调用每个 API，利用 SmartAPI 注释构造语法正确且输入标识符适当的调用，并将 API 响应映射到 Biolink模型[5]。BioThings Explorer 还会执行 ID到对象的转换，从而便于将 API 调用从查询路径中的一个步骤串接到下一步。当查询路径计划中连续调用的 API 使用不同标识符来表示同一个生物医学实体时（例如，NCBI基因ID 与 Ensembl基因ID），这一步 ID 转换至关重要。此阶段的输出是查询路径每个步骤的一组边，它们表示从 API 中检索到的生物医学实体之间的关联。

集成与评分。在最后阶段，这些来自 API 查询的边集合会被组装成结果子图，每个结果子图的拓扑结构都与查询图相匹配。随后，使用基于归一化谷歌距离[13]的语义相似性度量为每个结果评分。

部署与使用。 BioThings Explorer 知识图谱完全由联邦式 API 网络构成。由于无需在本地组装和存储大型知识图谱，BioThings Explorer 是一款非常轻量的应用程序，几乎可以轻松部署在任何标准个人计算机上。每位用户都能创建自己的 BioThings Explorer 本地实例，从而消除与大型查询和/或高强度使用相关的集中式瓶颈。介绍安装流程的代码仓库位于 https://github.com/biothings/biothings_explorer。

对于不想创建 BioThings Explorer 本地实例的用户，我们还通过 NCATSTranslator Consortium 维护了一个供大众使用的社区实例（https://explorer.biothings.io/）。

讨论

整合来自多个不同来源的现有数据，是评估当前知识状况的关键步骤。目前已有许多工作致力于创建生物医学知识图谱：整合本地下载的数据，并使用通用数据模型 [14]<style id="2">–[18] 对其进行标准化。这些工作构建出规模庞大的集中式知识图谱，通常包含数百万个节点和数千万条边。

BioThings Explorer 提供了一种独特的数据集成策略，重点是通过对 API 进行语义注释来构建联邦知识图谱。它并不将所有数据汇集到一个庞大的集中式图数据库中，而是采用联邦式设计，让知识保留在各资源的 API 背后。查询时，通过动态执行 API 调用并对结果进行语义解析来检索数据。这种架构在功能上将数据传播（通过创建 API）与数据建模和数据集成（通过语义注释）分离开来。

这种方法有几项优势。首先，将互操作性的要求从代码实现转移到语义化 *API* 注释，大幅降低了参与我们 API 生态系统的门槛。其次，将这些职责分离到不同的层级，有助于提升系统整体的模块化程度。这些组件可以并行开发和演进，而且这两项职责甚至可以由不同的团队承担（例如，一个团队为另一个团队创建的 API 添加语义注释）。第三，这种设计便于采用迭代式的 API 注释方法。开发者和 API 注释人员可以先提供一组最精简的 API 元数据，之后再根据未来需求和使用场景进行扩展。

BioThings Explorer 的联邦式设计也存在一些显著局限。首先，我们在 SmartAPI 中用于对 API 添加语义注释的 OpenAPI 扩展，仅适用于遵循 REST 协议且以 JSON 格式提供输出的 API。其次，由于整个联邦知识图谱从未在单一位置实例化，因此，依赖将完整知识图谱载入内存的推理方法和评分方法无法用于 BioThings Explorer。

总而言之，我们认为 BioThings Explorer 可作为现有知识图谱构建方法的补充，并为从事生物医学研究的数据分析师和工具开发者提供强大且独特的功能。

可用性与实现

BioThings Explorer 是一个 NodeJS 应用程序。BioThings Explorer 项目的主要代码库位于 https://github.com/biothings/biothings_explorer，并链接和整合了其他代码库作为子模块。所有代码均根据 Apache 2.0 开源软件许可证发布。

致谢

这项工作的经费由美国国立卫生研究院下属的国家转化科学促进中心通过生物医学数据翻译器计划提供，资助编号为 OT2TR003427 和 OT2TR003445。

文献

1. Rambla, J. 等Beacon v2 和 Beacon网络：生物医学基因组学及其他领域中联邦式数据发现的“通用语言”。《人类突变》**43**, 791–799 (2022). 2. Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. 分布式注释系统。《*BMC*生物信息学》**2**, 7 (2001). 3. OpenAPI规范 3.1.0版。http://spec.openapis.org/oas/v3.1.0. 4. OpenAPI规范扩展。https://spec.openapis.org/oas/v3.1.0#specification-extensions. 5. Unni, D. R. 等Biolink模型：临床、生物医学和转化科学领域知识图谱的通用模式。《临床与转化科学》**15**, 1848–1855 (2022). 6. Fecho, K. 等通用生物医学数据翻译器的研发进展。《临床与转化科学》**15**, 1838–1847 (2022). 7. SmartAPI。SmartAPI。https://smart-api.info/. 8. Zaveri, A. 等smartAPI：迈向更智能的 Web API 网络。载于《计算机科学讲义》154–169 (2017). 9. Davis, A. P. 等比较毒理基因组学数据库（CTD）：2023年更新。《核酸研究》**51**, D1257–D1262 (2023). 10. Mungall, C. J. 等Monarch Initiative：集成式数据与分析平台

跨物种连接表型与基因型。《核酸研究》**45**, D712–D722 (2017). 11. Lelong, S. 等 BioThings SDK：用于在生物医学研究中构建高性能数据 API 的工具包。《生物信息学》(2022) doi:10.1093/bioinformatics/btac017. 12. MyChem.Info。*MyChem.Info*https://mychem.info/. 13. Cilibrasi, R. L. & Vitanyi, P. M. B. 谷歌相似度距离。《*IEEE*知识与数据工程汇刊》**19**, 370–383 (2007). 14. Wood, E. C. 等RTX-KG2：用于构建语义标准化知识图谱、服务于转化生物医学的系统。《*BMC*生物信息学》**23**, 400 (2022). 15. Fecho, K. 等一种针对真实世界环境健康观察提出机制假设的生物医学知识图谱系统：队列研究与信息学应用。《*JMIR*医学信息学》**9**, e26714 (2021). 16. Morris, J. H. 等可扩展的精准医学开放知识引擎（SPOKE）：一个包含海量生物医学信息的知识图谱。《生物信息学》**39**, (2023). 17. Himmelstein, D. S. 等系统整合生物医学知识，优先筛选可重新利用的药物。《*eLife*》**6**, (2017). 18. Mayers, M. 等用于自动确定药物机制优先级的知识网络的设计与应用。《生物信息学》**38**, 2880–2891 (2022).

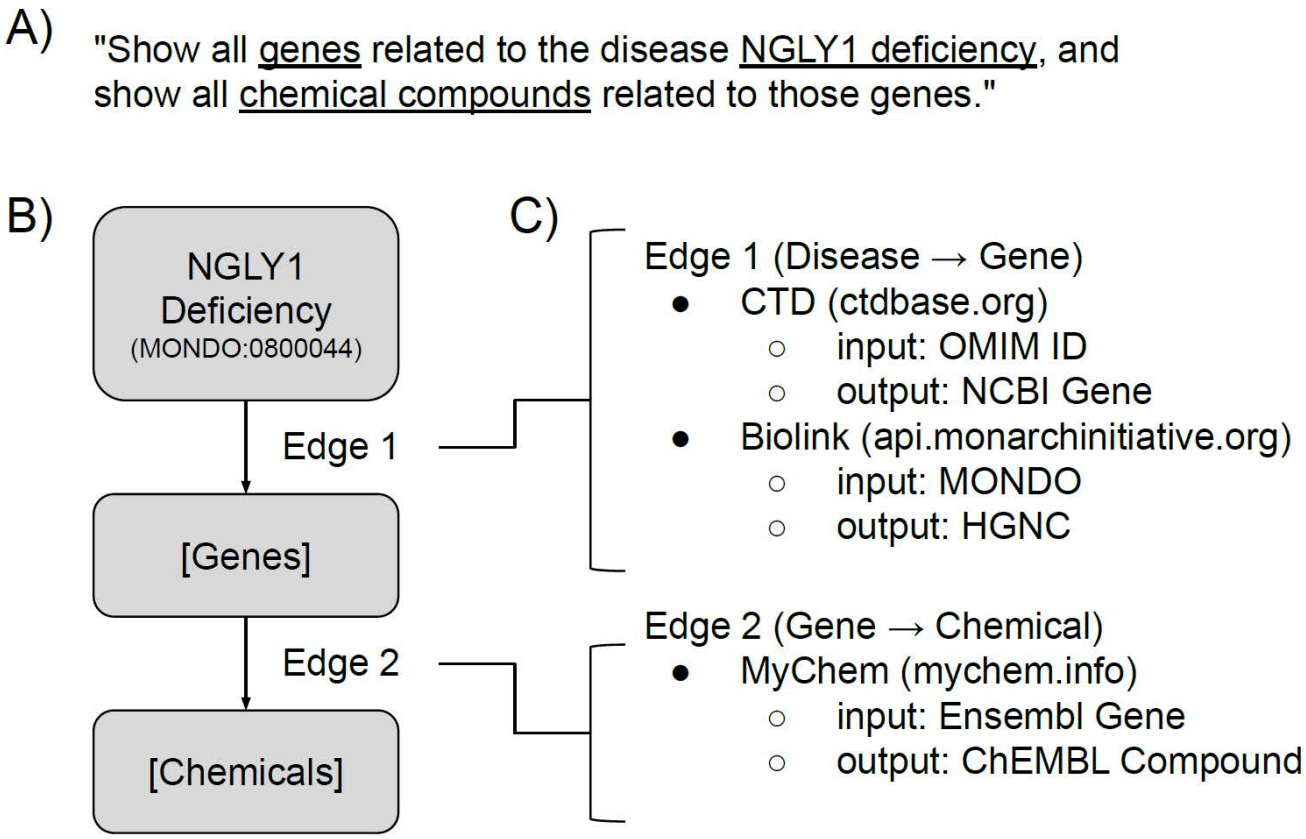


图 1. 解构： 一个 查询 在 BioThings Explorer 中 A）BioThings Explorer 可以回答的一个自由文本查询示例。B）同一查询的图表示。该图查询的确切语法由 Fecho 等人所述的 Translator Reasoner API（TRAPI）标准规定。[6] C）通过查询 SmartAPI 注册表中的元知识图谱，将图查询解构为多个 API 调用。

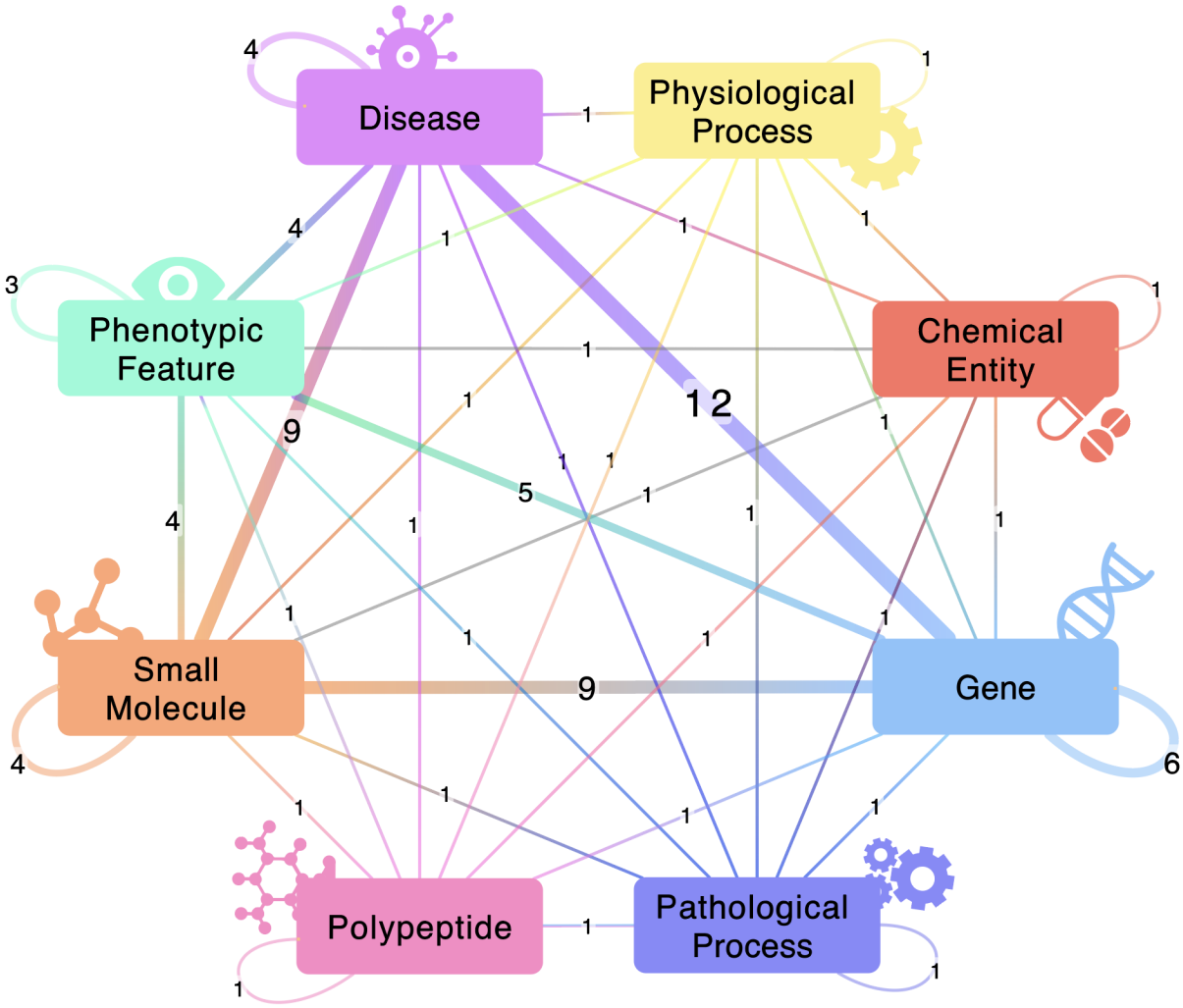


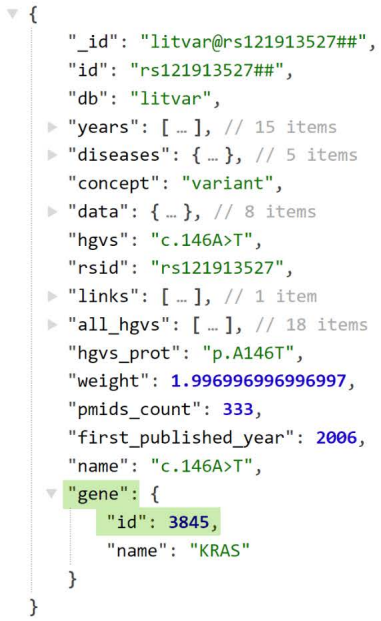
图2。BioThings Explorer 元知识图谱的可视化表示。该图中的节点是 BioThings Explorer 能够检索其关联关系的生物医学实体的语义类型（仅限最常见的前 10 种语义类型）。节点之间的边展示了通过 BioThings Explorer 可访问的语义 API 网络中存在的生物医学实体关联关系。边标签显示能够检索这些关联关系类型的 API 数量，边的宽度也表示该数量。

补充信息

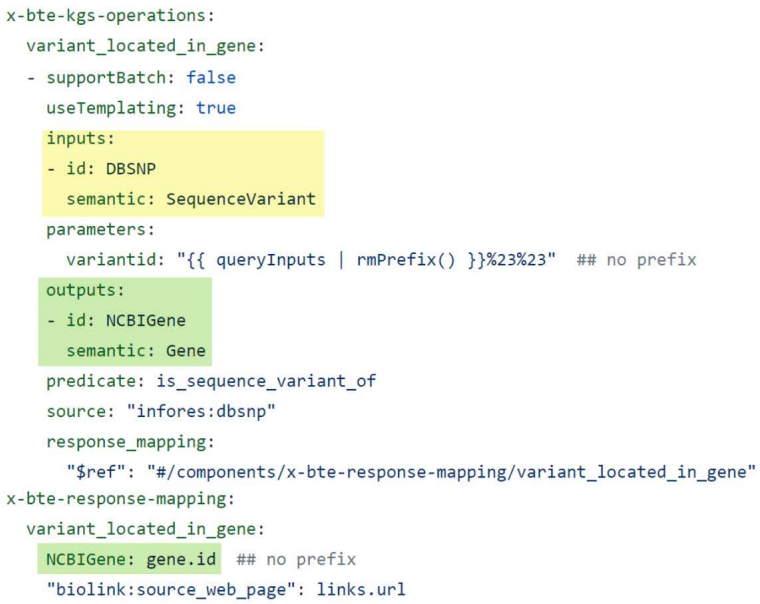
A)

https://www.ncbi.nlm.nih.gov/research/bionlp/litvar/api/v1/entity/litvar/rs121913527%23%23?format=json

B)



C)



补充图1。BioThings Explorer 使用对 OpenAPI 规范的扩展，为 API 添加语义注释。这些注释包括查询中使用（输入）和响应中返回（输出）的生物医学实体的语义类型和标识符命名空间、JSON 响应中输出标识符的 JSON 路径，以及描述输入实体与输出实体之间关系的谓词。此图展示了用于提供 LitVar 内容的 API 的这些扩展。A) LitVar API 可使用以下语法调用；注释说明，查询此 API 时应使用 DBSNP ID（黄色高亮显示）作为输入。B) API 返回一个 JSON 对象，其中包含与查询中指定的变异相关的基因（绿色高亮显示）。C)`x-kgs-operations` 部分的 SmartAPI 注释说明，LitVar API 检索的数据将“SequenceVariants”（使用 DBSNP ID 标识）与“Genes”（使用 NCBIGene ID 标识）相关联。SmartAPI 注释中的 `x-bte-response-mapping` 部分说明，相关基因的 NCBIGene ID 可在数据的 JSON 路径“gene.id”下找到。

补充表1：https://www.dropbox.com/s/ij09w88zwcuy4uv/supptable.tsv?dl=0