格点 QCD 工作流的溯源

Tanja Auge，[1] Gunnar Bali，[2] Meike Klettke，[1] Bertram Ludäscher，[3] Wolfgang Söldner，[2] Simon Weishäupl，[2] Tilo Wettig[2]{firstname.lastname}@ur.de ludaesch@illinois.edu[1]雷根斯堡大学，计算机科学与数据科学学院，德国[2]雷根斯堡大学，物理系，德国[3]伊利诺伊大学厄巴纳-香槟分校，信息科学学院，美国

arXiv:2303.12640v1 [hep-lat] 22 Mar 2023

摘要

我们提出了一种适用于格点量子色动力学（QCD）计算通用工作流的溯源模型，这类计算是粒子物理学研究的重要组成部分。这些计算在全球规模最大的超级计算机上进行，生成并分析的数据量达到数拍字节。在格点量子色动力学社区中，针对工作流的一个部分——所谓的配置生成 ——已经有一项自定义元数据标准（QCDml），其中包含某些溯源信息。

本文遵循 W3C PROV 标准，构建了一个溯源模型，涵盖格点量子色动力学工作流中的生成部分和所谓的测量部分。我们展示了该模型的适用性，并说明如何利用该模型回答一些与溯源相关的研究问题。不过，格点量子色动力学学术共同体中的许多重要溯源问题，都需要扩展这一溯源模型。为此，我们提出一种多层分层溯源方法，将前瞻性和回顾性要素结合起来。

关键词

工作流溯源，W3C PROV，格点量子色动力学

1 引言

溯源通常指“任何描述最终产品生产过程的信息，而最终产品可以是从一份数据到一个实物的任何东西” [17]。这也是我们在应用——格点 QCD 工作流——中面临的挑战。下面简要介绍这一应用。从根本上说，我们对自然的理解建立在基本粒子及其相互作用的标准模型之上。标准模型以量子场论的形式构建，其中包括量子色动力学 （QCD）。目前，许多实验和理论研究正致力于寻找标准模型之外的物理。这些探索需要精确度极高的 QCD 计算。此类计算首选的工具是在时空格点上对 QCD 进行数值模拟（格点量子色动力学）。格点量子色动力学发展至今已有四十多年，如今已成为一个成熟的领域，全球有数百名研究人员参与其中。与实验类似，实验中先收集数据，之后再进行分析，格点量子色动力学计划也分为三个部分：生成、测量和分析 [4]。第一部分中，使用马尔可夫链蒙特卡罗方法 生成所谓的规范场构型 系综，然后将其存储到磁盘。第二部分中，在这些配置上计算与特定物理研究计划相关的所谓关联函数 ，并将所得数据也存储到磁盘。第三部分中，将关联函数组合起来，得到感兴趣的可观测量。前两个部分的计算量非常大，需要使用全球最大的超级计算机。目前，某个合作组的典型数据规模约为一拍字节的系综和数拍字节的派生数据 [4]。

早在二十年前，格点量子色动力学学术共同体就发起了国际格点数据网格（ILDG） [20, 22, 23, 27] ，旨在共享构型系综（遵循如今所称的*FAIR* 原则——FAIR 是**可**发现、**可**访问、**可**互操作和**可**重用的缩写，其中可重用性涵盖溯源方面 [14, 37, 42]）。对于格点量子色动力学计划的生成部分，ILDG元数据标准（QCDml[27, 41]）也包含跟踪信息，但迄今尚未形成完整的溯源概念。本文提出了一个基于 W3C PROV 标准（该标准在创建 ILDG元数据标准时尚不存在）的生成部分溯源模型。此外，我们将模型扩展至格点量子色动力学计划的测量部分，而此前学术共同体尚未就这一部分开展溯源方面的协作。我们认为，这一工作是朝着在科研数据管理中落实 FAIR 原则迈出的重要一步。

为说明溯源在格点量子色动力学中的重要性，我们来讨论一个实际发生过的例子。一组配置存储在一家外部研究机构。在存储期间，由于文件系统问题，发生了静默数据损坏。在发现数据损坏之前，可能已经基于损坏的构型进行了测量。在这种情况下，溯源可以将不正确的测量结果追溯（上溯）到损坏的构型。反过来，溯源也可以识别可能受到损坏构型影响的测量结果（下游结果）。

另一个重要方面是科学论文中发表结果的可重复性和可复现性（定义见 [1]）。一方面，我们需要溯源信息来确定哪些数据实际与已发表结果相关。另一方面，我们还需要相关工作流的溯源信息。这将使我们（或其他人）能够以透明的方式重现已发表的结果。详情请参阅 [2, 3]。

要将溯源集成到格点 QCD 工作流中，我们需要一个可持续的溯源模型。在分析了其他物理问题（例如 [21, 40]）之后，我们决定采用 W3C PROV 模型 [15]。该模型定义了“一种数据模型、序列化格式和相关定义，用于支持在万维网上交换溯源信息”。在 [21], 中，作者“为开发天文/数据处理流水线的自动化溯源生成工具奠定了基础”，但没有使用 W3C PROV 模型；而 [40] 则描述了一种 W3C PROV 模型，用于使软件 BACARDI 具备溯源能力。BACARDI 提供了一个数据库，其中包含有关围绕地球运行的活动和非活动天体的信息。该模型使用 Python 库 prov [9] 实现。

研究人员自然倾向采用的溯源模型，是从高层次的概念工作流模型出发，再对细节进行特化或实例化，以回答溯源问题。因此，为了构建我们的溯源模型，我们首先描述一个通用的格点量子色动力学工作流（见下文图 2）。接下来，我们将更详细地分析一个具体的格点量子色动力学工作流，并为其建立一个初步的 W3C PROV 模型（图 3）。最后，我们提出一个多层模型（图 4），并设想将其实现为 W3C PROV 扩展。

本文的结构如下。首先，我们介绍溯源和 W3C PROV 标准的背景信息（第2节）。接着，我们介绍用例（格点量子色动力学），并建立基于 W3C PROV 的溯源模型（第3节）。随后，我们讨论扩展溯源模型的新思路（第4节）。最后，我们进行总结并展望未来（第5节）。

贡献。我们的主要贡献如下。

- 我们展示了溯源建模在格点量子色动力学这一数据密集型科学领域中的适用性。
- 我们展示了面向格点量子色动力学工作流的 W3C PROV 模型如何用于回答常见的溯源相关研究问题，例如**Q1**至**Q5**（第3节）。
- 我们阐明了扩展模型的必要性，以支持希望利用溯源提高研究工作流透明度，并收集和使用更详细溯源信息的研究人员。

2 背景与相关工作

溯源信息涵盖生产过程中涉及的实体、活动和代理的元数据。在科学工作流中，这些溯源信息通常会被记录 [24]。在

对于格点量子色动力学，我们处理的是一个复杂的工作流，它利用超级计算机生成并分析大量数据。为了回答溯源相关问题（我们将在第3节详细介绍），我们采用 W3C PROV 模型来描述相应的溯源信息。该模型于十年前开发，定义了一个“用于溯源的核心数据模型，用来构建对生成某项数据或事物所涉及的实体、人员和过程的表示，并可据此评估其质量、可靠性或可信度” [13]。该模型区分三个核心概念：实体、活动和代理。实体是数据或制品，也可以派生自 其他实体。活动可以生成 或使用 实体。代理可以执行或控制活动，也可以生成实体。工作流溯源可以用图来描述，其中节点

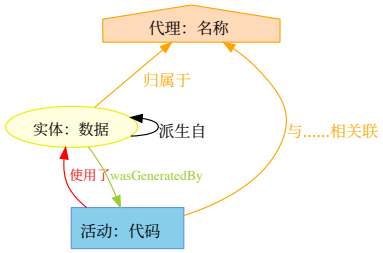


图 1：PROV 核心概念与关系，基于 [13]。

是实体、活动或代理。它们之间的关系由图中不同的边类型描述（图 1）。

W3C PROV 标准的核心是 PROV 数据模型（PROV-DM [32]）。它定义了用于表达和交换溯源信息的概念，并通过一系列相关规范予以实现，例如面向人类阅读的溯源表示（PROV-N）、PROV 本体（PROV-O）以及 XML 模式（PROV-XML）；详情请参见 [13]。W3C PROV 模型的直接前身是开放溯源模型（OPM [31]）。

在科学工作流领域，溯源是对一组结果之派生过程的记录。溯源有两种不同形式 [44]：W3C PROV 模型和 OPM 都记录回溯性溯源，即关于过去工作流执行和数据派生的信息。相比之下，前瞻性溯源记录工作流的结构，可视为未来工作流执行的配方。工作流图也可以作为对过去发生情况的高层次“总结”；也就是说，尽管这类图具有前瞻性，但仍可用于在更高的概念层面描述过去的工作流执行。由于 W3C PROV 模型（OPM 也类似）旨在为回溯性溯源提供一个最小化模型，因此，涉及前瞻性溯源或混合溯源元素的问题无法直接在该模型中得到解答 [25, 38]。

为了解决这些局限，多年来，人们对回溯性溯源标准进行了多项扩展：ProvONE [6] 通过增加一个工作流（即前瞻性溯源）层来扩展 W3C PROV 模型，并将该层与回溯性层关联，从而支持科学工作流应用中结合两种溯源类型的混合查询。ProvONE“旨在提供理解和分析基于科学工作流的计算实验所需的基本信息” [6]。类似地，其前身 D-PROV [30], 以及相关的工作流开放溯源模型OPMW [12], 为科学家提供了词汇表和关系结构，以回答混合溯源问题。Wf4Ever 研究对象模型包含用于工作流执行溯源的词汇表 [39]。人们已开发出许多用于捕获计算溯源的工具，例如， [36] 对脚本溯源工具进行了综述，[5] 则介绍了一种用于捕获工作流溯源的工具。此前也有人采用将回溯性溯源与前瞻性溯源相结合的思路来支持混合查询，例如参见 [43] 或 [35]。后者将基于注释的前瞻性溯源建模工具 YesWorkflow [28] 与一种从 Python 脚本中捕获细粒度回顾性溯源的工具 [34]相结合。

3 格点量子色动力学的溯源

在本节中，我们将使用格点量子色动力学学术共同体熟悉的词汇来构建一个溯源模型。下方图 3 展示了该模型的 W3C PROV 表示。

3.1 格点量子色动力学计算工作流

如引言所述，通用的格点量子色动力学工作流由三个主要部分组成。第一部分（生成）通过蒙特卡洛技术生成所谓的规范场配置，通常采用 HMC 算法 [10]。在此过程结束时，会得到一定数量的配置，每个配置都由固定数量的复数组成。这些配置会存储在磁盘上，以供后续分析。第二部分（测量）根据这些配置计算关联函数及其他派生数据（统称为测量数据）。这些关联函数包含有关物理可观测量的信息。第三部分（分析）根据关联函数计算可观测量，其中包括对配置取平均、外推至特定极限以及其他活动。随后，可以将这些计算结果与实验结果进行比较。

工作流的前两个部分如图 2 中的灰色方框所示。这里，我们不考虑第三个部分，原因有二：前两个部分的计算密集程度高得多；第三个部分取决于所关注的可观测量，因此其通用性远不如前两个部分。我们将分析部分溯源模型的构建留待未来工作。

生成部分的输入是模拟参数，这些参数控制 HMC 算法（算法参数）和物理细节（物理参数）。请注意，物理参数也会用于测量部分。通常，这些参数可在生成部分产生的配置元数据中获取。其他参数（测量参数）也会用于测量部分。

我们简要讨论一下图 2 中标记为“数据管理”的活动。典型的格点量子色动力学计算会生成海量数据 [4]，因此数据管理已变得至关重要 [23]。收集和处理元数据显然是一项必不可少的任务。此外，为确保正确性，还需要进行数据验证，而这会用到元数据。另外，由于生成——
由于生成配置的计算成本非常高，因此备份数据以防止数据丢失非常重要。为符合科学标准，通常需要将数据归档一段时间。请注意，数据管理对于配置和测量数据同样重要，也就是说，图 2 中的测量部分也包含数据管理，只是没有明确显示出来。

3.2 纳入溯源

为了更好地支持研究人员开展计算科学工作，数据管理系统和网络基础设施应能够回答如下问题。

Q1 哪些数据集受到错误或缺陷的影响？**Q2** 修改参数会如何影响数据集？**Q3** 谁参与了数据生成？**Q4** 重复执行某个工作流需要哪些代码和专家？

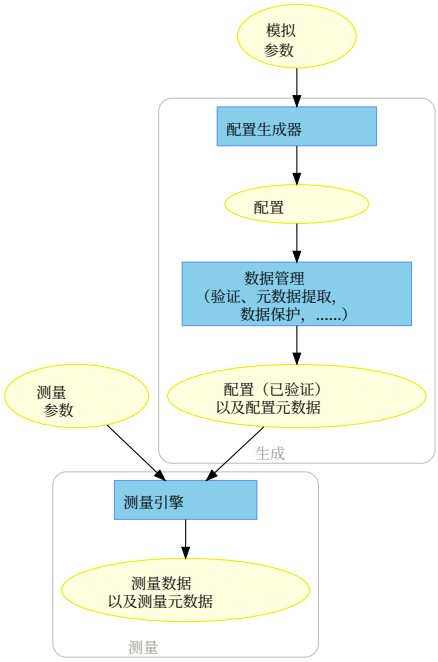


图 2：通用格点量子色动力学工作流前两部分的表示。

Q5 要（重新）生成某个结果，需要哪些数据/参数？这些问题只是计算科学家经常需要解答的几个溯源相关问题示例，与 [40] 中提出的问题类似。它们始终以实体、活动或代理为中心，例如用于缺陷检测、质量保证、流程验证、监控、统计分析、开发者评估或信息收集。[1] 由于问题种类繁多，解答上述问题需要不同的溯源要素。问题**Q1** 和**Q2** 主要涉及数据层面，而问题**Q3** 至**Q5** 则与生成数据的底层工作流关系更为密切。

我们遵循 [13] 中介绍的 W3C PROV 标准，为研究组实施的特定格点量子色动力学工作流定义了一个溯源模型，见图 3。蓝色代码框中显示的四项活动读取两组输入参数、管理数据并生成最终的 HDF5 文件。黄色框中显示的七个实体 代表输入参数、数据和元数据。该模型还包含三个代理 （橙色框），它们执行这些活动，并负责输入参数。在我们的案例中，代理 Alice 和 Bob 分别定义输入参数并执行前两项活动：openQCD 和元数据提取与验证。在此工作流中，会生成已验证数据和元数据。随后，Charly 使用这些数据、元数据以及测量参数，执行测量部分的活动，并生成实际的 HDF5 文件。

图 2 和图 3 之间有两个主要区别。首先，图 2 中的数据管理活动还包括备份和归档；为了简化起见，我们在图 3 中省略了这些内容。

[1]有关其他计算科学溯源问题，请参阅例如 [29]和 [35]。

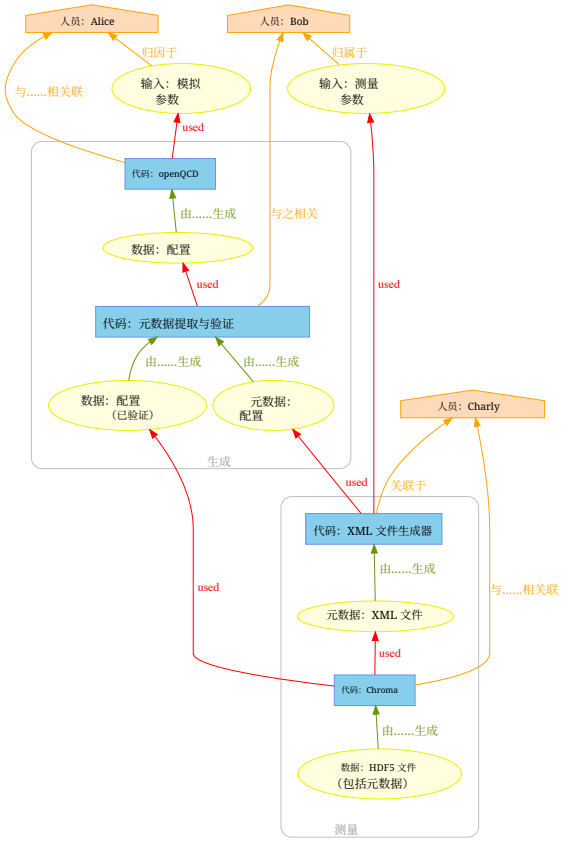


图 3：作为 W3C PROV 模型的格点量子色动力学工作流。

其次，图 2 中的测量子工作流由单个活动组成，而图 3 中对应的子工作流则包含第二个活动。这是因为在测量引擎的实际实现中，还需要一个中间步骤，即生成合适的 XML 输入文件。总之，图 3 中的溯源图可以看作是对图 2 中更抽象的工作流版本的细化。

仔细观察后，我们发现 PROV 图不包含形式为 $(D_1) \overset{w}{\rightsquigarrow} (D_2)$ 的 *w*-边（源自），也就是说，一个数据实体 D_2 源自另一个实体 D_1。相反，图 3 包含由形式为

$$(D_1) \overset{u}{\rightsquigarrow} [A] \overset{g}{\rightsquigarrow} (D_2),$$

也就是说，某个过程（活动） A 使用了数据实体 D_1，并生成了数据实体 D_2。在许多应用中（包括我们的应用），*used* 和 *was-GeneratedBy* 关系对于显式建模数据实体流入和流出处理步骤（活动）的过程至关重要，从而支持强大的溯源分析查询。尽管该标准并未假设 *used-wasGeneratedBy* 链总是意味着存在 *wasDerivedFrom* 关系，但这在实践中往往如此，我们的格点量子色动力学溯源模型也是如此。为使图示清晰，我们在图 3 中省略了这些 *wasDerivedFrom* 边。模型的实现可以使用自定义派生规则按需构造这些边；有关这些不同关系之间相互作用的更多细节，请参见 [8] 和 [33]。

图 3 中的溯源图包含特定的实例级信息，例如代理 Alice、Bob 和 Charly 的名称，以及架构层面的抽象标识符，例如 XML 文件。在许多情况下，添加此类实例级信息就能满足格点量子色动力学学术共同体的需求。例如，计算通常由同一批人执行，因此可以将这些个人／具体代理视为“系统的一部分”。

图 3 列出了在特定研究组和时间段内定期担任代理的人员姓名。更通用的设置会将代理标记为 Person 类型。使用人员姓名，可以直接从溯源图中回答问题 **Q3**。相比之下，问题 **Q4** 涉及工作流层面的活动和代理，需要前瞻性溯源。**Q5** 是一个类似的实体层面问题。

与 **Q3** 类似，问题 **Q1** 是一个回溯性溯源问题。例如，我们可能会发现输出数据不正确。在这种情况下，必须从结果追溯到来源，追踪数据派生链（上游 传播）。再比如，我们可能会发现某项活动或实体中存在错误，例如前文提到的配置中的静默数据损坏。此时需要修正出错的活动或实体，并重新计算或修正后续实体（下游 传播）。这两个例子都需要沿着溯源图中的派生链进行依赖追踪。

问题 **Q2** 可以有不同的解读。如果我们只关注概念层面的依赖结构（图 3），那么问题 **Q2** 只需要前瞻性溯源。或者，如果我们关注某个参数变更对先前工作流运行中数据集的影响，则需要回溯性溯源。

总而言之，问题 **Q1** 和 **Q3** 需要回溯性溯源，问题**Q4** 和 **Q5** 对应前向溯源查询，而问题 **Q2** 则结合了这两类溯源。

4 迈向分层溯源

Alice、Bob 和 Charly 每次执行工作流，都会生成一个与图 3 所示类似的溯源图。由于所有情况下整体溯源模型的结构都保持不变，因此图 3 所示的溯源图实际上是一个溯源模板图；也就是说，每次工作流运行（执行）都会生成各自的溯源实例图，其中模式级元素（例如 data:HDF5 文件）会被对具体实例对象（例如 X251r000n1000_run3.hd5）的引用所取代。反过来，图 3 中的溯源模板也可以看作是图 2 中工作流图的一种专门化形式；图 2 描述了学术共同体所使用的格点 QCD 工作流的一般形式。例如，工作流中的通用步骤配置生成器和测量引擎分别专门化为代码：*openQCD* [26] 和代码：*Chroma*[11]，它们是我们研究组的物理学家所使用的具体工具。

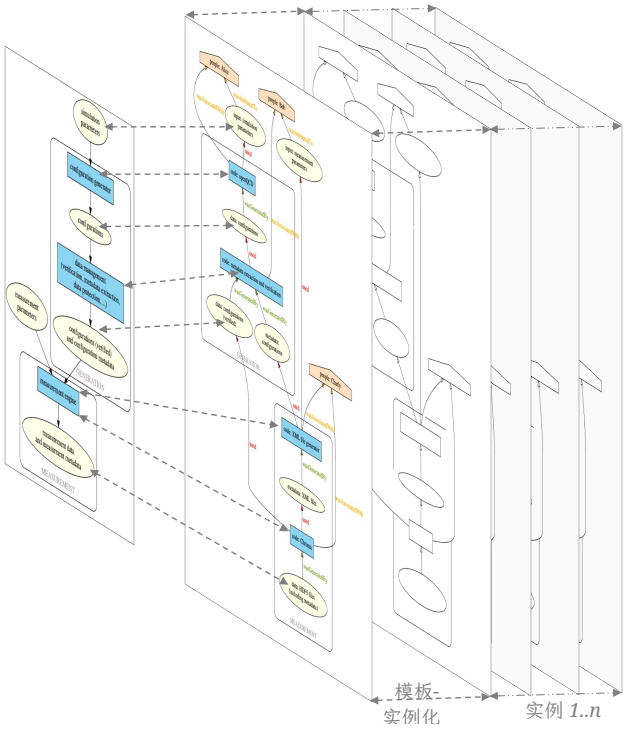


图 4：所提出的混合溯源模型的层次：概念工作流的元素（左）自然映射到 W3C PROV 模板模型中的元素（中）。每次工作流执行都会创建兼容 PROV 的回溯性溯源图（右），即中间溯源模板的实例1、**……**，***n***。

4.1 溯源模板与实例

为满足研究人员的实际需求，同时采用标准模型以促进数据交换和透明度，我们提议扩展 W3C PROV 模型，使其同时包含实例级溯源图，以及与之关联的模板级溯源图。模板图与实例图之间的关系十分自然：后者可以看作前者的同构副本，其中模式级元素已被对象标识符取代。这样，模板图便可作为众多实例图的概览或摘要。

我们进一步提议在此扩展中加入一个工作流层。在由此形成的多层溯源模型中（图 4），学术共同体层面的工作流图可以专门化为溯源模板（适用于各个研究小组），并在工作流运行执行时将其实例化。这些实例随后会包含所有输入参数的具体值、包含配置的数据文件的名称和时间戳、代码的版本号或 Git 哈希值、编译器版本和标志、执行编译的人员姓名、运行计算的机器详情等。

正如我们对第3.2节中从**Q1** 到**Q5** 的溯源相关查询所做的初步分析所示，要回答所有这类问题需要结合回溯性溯源元素（由 W3C PROV 标准提供）和前瞻性元素（由工作流或溯源模板图提供）。以这种方式扩展我们的模型后，它应满足以下期望条件：

(1) 学术共同体层面的工作流结构应能够关联到溯源模板图（研究组层面）。(2) 数百或数千次运行的实例级溯源图（参数设置各不相同等）应能自动关联到溯源模板。(3) 领域感知溯源模型应允许用户区分不同类型的数据，例如使用命名空间。

拟议的多层模型（图 4）可以通过不同方式实现。我们希望汇聚两个学术共同体，即格点量子色动力学社区的同事和溯源研究社区的同仁，共同开发适用的 W3C PROV 扩展。我们目前计划探索并评估现有标准，尤其是 W3C PROV 和资源描述框架 [7]。后者还将使我们能够将多层溯源模型嵌入更通用的知识图谱/关联开放数据框架中，再次借助现有标准、工具和命名空间。借助这些工具，格点量子色动力学社区可以决定在对不同类型的语义进行建模和形式化方面希望推进到何种程度，例如创建受控词表和/或 OWL 形式化本体，并由学术共同体达成共识。我们向溯源研究社区寻求协助，以便在模型构建工作中抢占先机。既然我们并非首批发现需要混合溯源模型的人（参见 [5, 6, 12, 39, 44] 以及许多其他文献），我们希望在现有工作的基础上继续推进，而不是重新发明轮子。

4.2 实现与评估计划

为了评估所提出模型的有效性和实用性，我们计划为格点量子色动力学社区实现一个原型。由于现有量子色动力学工作流已经在日志文件中记录了大量溯源信息，我们将首先开发一个基于 Python 的溯源信息采集工具。然后，需要将采集到的信息映射到适合存储并实现我们模型的存储库（例如关系型数据库或图数据库）。其他溯源信息可从 HDF5 数据集的属性和数据空间对象，以及文件名和文件夹名中获取。如 [29], 所述，这些溯源信息将根据适用的约定提取。最后，我们模型所需但尚无法通过采集获取的溯源信息，则必须通过其他方式记录，例如使用轻量级溯源记录器、对代码进行插桩，或将其他信息写入日志文件。

我们的开发工作将参考关于如何从脚本中收集和管理溯源的调研 [36]，以及使用能够捕获并整合前瞻性和回溯性溯源信息的工具的实践经验 [28, 34, 35, 43]。为以符合 W3C PROV 标准的形式导出可互操作的溯源信息，我们将使用 Python prov 库 [18]。我们还将探索将扩展后的溯源模型转换为符合万维网联盟标准的

知识图谱。这样我们就能利用基于流行的三元组存储和图查询语言（例如 SPARQL [16] 或 Cypher [19]）的其他标准和工具。

5 总结与未来工作

格点量子色动力学是粒子物理学的一个重要领域，它会生成并分析海量数据。我们提出了一种基于万维网联盟 PROV 标准的格点 QCD 工作流溯源模型。从一个通用工作流（图 2）出发，我们推导出一个 PROV 模板（图 3），借此能够回答若干典型的溯源相关问题。为回答更全面的溯源问题（例如，**Q1** 至**Q5**），我们提出了一种分层模型，以提供所需信息。该模型由工作流层、溯源模板层和实例层组成（图 4）。

未来的工作中，我们计划将溯源模型应用于整个格点 QCD 工作流的第三部分（分析）。如前所述，这一部分的通用性较低，更贴合特定合作组的物理研究计划。

我们在第 3.1 节中指出，数据管理对格点 QCD 工作流的所有部分都很重要。因此，我们希望在格点量子色动力学社区在 ILDG 和 PUNCH4NFDI 中所做工作的基础上，制定更全面的数据管理标准，并将溯源元数据纳入其中。

我们计划与格点量子色动力学和溯源社群合作，完善初步模型提案，随后实现一个原型，以便评估其有效性和实用性。

致谢

我们感谢 Sara Collins、Christoph Lehner、Nils Meyer 和 Stefan Solbrig 富有启发性的讨论。本研究部分得到德国研究联合会（DFG）“NFDI 39/1”（PUNCH4NFDI）项目资助。

参考文献

[1] ACM。2020年。工件评审与徽章认证——当前版本。https://www.acm.org/publications/policies/artifact-review-and-badging-current [2] Tanja Auge、Moritz Hanzig 和 AndreasHeuer。2022年。ProSAPipeline：溯源征服追逐。载于*ADBIS*（短文）（*CCIS*，第 *1652* 卷）。施普林格，89–98。[3] Tanja Auge 和 Andreas Heuer。2019年。ProSA——使用 CHASE 进行溯源管理。载于*ADBIS*（*LNCS*，第 *11695* 卷）。施普林格，357–372。[4] Gunnar Bali 等。2022年。格点规范场系综与数据管理。科学论文集（*PoS*）LATTICE（ 2022 203 2212 10138 ），。arXiv: 。[5] Flavio Costa 等。使用 PROV 捕获和查询工作流运行时溯源：一种实用方法。载于 *EDBT/ICDT* 研讨会。ACM，282–289。[6] Víctor Cuevas-Vicenttín 等。2016年。ProvONE：面向科学工作流溯源的 PROV 扩展数据模型。http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html。[7] Richard Cyganiak、David Wood 和 Markus Lanthaler。2014年。RDF 1.1 概念与抽象语法。https://www.w3.org/TR/rdf11-concepts/ [8] Saumen C. Dey 等。2012年。将 Datalog 作为溯源查询与推理的通用语言。载于*TaPP*。USENIX 协会。[9] DLR。2022年。将溯源感知集成到空间碎片处理系统 *BACARDI* 中。https://github.com/DLR-SC/bacardi-prov-model [10] Simon Duane 等。1987年。混合蒙特卡罗。物理评论快报 *B*195（1987），216–222。[11] Robert G. Edwards 和 Bálint Joó。2005年。用于格点量子色动力学的 Chroma 软件系统。核物理 *B* 增刊 140 2005 832 0409003 （），。arXiv:hep-lat/ [12] Daniel Garijo、Yolanda Gil 和 Oscar Corcho。2014年。OPMW-PROV：工作流开放溯源模型。https://www.opmw.org [13] Yolanda Gil 等。2013年。PROV 模型入门。https://www.w3.org/TR/prov- primer/ [14] GoFair。2023年。FAIR 原则。https://www.go-fair.org [15] Paul Groth 和 Luc Moreau。2013年。PROV-Overview。https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/ [16] Steve Harris、Andy Seaborne 和 Eric Prud’hommeaux。2013年。SPARQL 1.1 查询语言。https://www.w3.org/TR/sparql11-query/ [17] Melanie Herschel、Ralf Diestelkämper 和 Houssem Ben Lahmar。2017年。溯源综述：为何需要？形式为何？源自何处？VLDB J. 26，6（2017），881–906。[18] Trung Dong Huynh。2020年。prov 2.0.0。https://pypi.org/project/prov/ [19] Neo4j Inc. 2021年。什么是 openCypher? https://opencypher.org [20] A.C. Irving 等。2004年。国际格点数据网格建设进展。核物理 B 增刊 129 2004 159 163 0309029 （），–。[21] 2021 arXiv:hep-lat/ Michael A. C. Johnson 等。天文流水线溯源：用例评估。载于 TaPP。USENIX 协会。[22] Bálint Joó 和 C. M. Maynard。2006年。国际格点数据网格建设进展。PoS LAT 2006 2006 028 10 22323 1 032 0028 （），。[23] 2022 https://doi.org/ 。/ 。Frithjof Karsch、Hubert Simma 和 Tomoteru Yoshie。国际格点数据网格——迈向 FAIR 数据。载于第 39 届国际格点场论研讨会。arXiv: 2212 08392 [ ]。[24] 2022 hep-lat Meike Klettke 和 Uta Störl。数据工程面向数据科学的四代演进。Datenbank-Spektrum 22，1（2022），59–66。[25] Chunhyeok Lim 等。2010年。在科学工作流环境中进行前瞻性与回溯性溯源收集。载于 SCC。IEEE 计算机学会，449–456。[26] Martin Lüscher 和 Stefan Schaefer。2013年。具有开放边界条件和扭质量重加权的格点量子色动力学。Comput. Phys. Commun. 184（2013），519 528 10 1016 2012 10 003 1206 2809 [ ]–。[27] 2005 https://doi.org/ 。/j.cpc. 。. arXiv: 。hep-lat C. M. Maynard 和 D. Pleiter。QCDml：建设国际格点数据网格的第一个里程碑。核物理 B 增刊 140（2005），213–221。https://doi.org/ 10 1016 2004 11 116 0409055 。/j.nuclphysbps. 。. arXiv:hep-lat/ [28] 2015 Timothy McPhillips 等。YesWorkflow：一种以用户为中心、独立于语言的工具，用于从脚本中恢复工作流信息。国际数字策展期刊 10, 1 2015 298 313 10 2218 10 1 370 （），–。[29] 2015 https://doi.org/ 。/ijdc.v i 。 Timothy M. McPhillips 等。无需运行时溯源记录器的回溯性溯源。载于 TaPP。USENIX 协会。[30] Paolo Missier 等。2013年。D-PROV：以工作流结构扩展 PROV 溯源模型。载于 TaPP。USENIX 协会。[31] Luc Moreau 等。2011年。开放溯源模型核心规范（v1.1）。Future Gener. Comput. Syst. 27，6（2011），743–756。[32] Luc Moreau 等。2013年。PROV-DM：PROV 数据模型。https://www.w3.org/TR/2013/REC-prov-dm-20130430/ [33] Luc Moreau 等。2015年。PROV 的设计依据。J. Web Semant. 35（2015），235–257。[34] Leonardo Murta 等。2014年。noWorkflow：捕获和分析脚本溯源。载于 IPAW（LNCS，第 8628 卷）。施普林格，71–83。[35] João Felipe Pimentel 等。2016年。Yin & Yang：展示 noWorkflow 与 YesWorkflow 互补的溯源能力。载于 IPAW（LNCS，第 9672 卷）。施普林格，161–165。[36] João Felipe Pimentel 等。2019年。脚本溯源的收集、管理与分析综述。ACM Comput. Surv. 52，3（2019），47:1–47:38。[37] PUNCH4NFDI。2023年。NFDI（国家研究数据基础设施）中的一个联盟。https://www.punch4nfdi.de [38] Yogesh Simmhan、Paul Groth 和 Luc Moreau。2011年。专刊：使用开放溯源模型实现互操作性的第三次溯源挑战。Future Gener. Comput. Syst. 27，6（2011），737–742。[39] Stian Soiland-Reyes 等。2013年。Wf4ever 研究对象模型。http://wf4ever.github.io/ro/ [40] Martin Stoffers 等。2022年。将溯源感知集成到空间碎片处理系统 BACARDI 中。载于 2022年 IEEE 航空航天会议（AERO）。1 12 10 1109 53065 2022 9843783 –。https://doi.org/ 。/AERO 。. [41] 2004 2013 ILDG 元数据工作组。–。QCDml 标准规范。https://www2.ccs.tsukuba.ac.jp/ILDG/ [42] Mark D. Wilkinson 等。2016年。科学数据管理与维护的 FAIR 指导原则。Scientific Data 3（2016），160018。[43] Qian Zhang 等。2017年。利用混合溯源揭示脚本输出的详细数据沿袭。国际数字策展期刊（IJDC）12，2（2017），390–408。[44] Yong Zhao、Michael Wilde 和 Ian T. Foster。2006年。应用虚拟数据溯源模型。载于 IPAW（LNCS，第 4145 卷）。施普林格，148–161。