150ns 下的实时事件级零样本学习：迈向用于粒子物理实验触发的基础模型

Leo Rozanov[1⋆]、Lauri Laatu[1],Benedikt Maier[1], Arianna Cox[1], Aadeesh Sharma[1],Zhiqiang Que[2] 和 Alex Tapper[1]

1 伦敦帝国理工学院，英国伦敦2 布里斯托大学，英国布里斯托⋆ 通讯作者

arXiv:2611.28355v1 [physics.ins-det] 20 Nov 2026

摘要

大型强子对撞机（LHC）实验的基于硬件的一级触发器，使用针对单一事例特征设计的算法来筛选碰撞；这些算法依据每个事例粗略且低延迟的视图进行判断。我们提出了一种适用于多项触发任务的学习式事例嵌入，并评估其向训练中未出现的过程迁移的能力。利用同一模拟碰撞的离线视图和一级触发视图组成的配对数据，我们通过监督式度量损失训练离线 Transformer，再逐个事例将其蒸馏到量化一级 Transformer 中；该模型的位宽在固定资源预算下学习得到。在 25 种标准模型过程中训练后，这种嵌入能将 28 种留出过程映射到与其物理相关过程相邻的位置。通过与从未参与编码器训练的背景碰撞进行最近邻筛选，可排除其中 99%。该方法保留了已见标准模型事件的中位数 90%、未见标准模型事件的 96%，以及 52 个超出标准模型基准样本的 88%；在使用相同事件评估的基线方法中，它在这三组上的得分均为最佳。针对已见过程的第二道门控可将高质量双轻子信号分离到无标签簇中；若提供位移径迹信息，也能分离长寿命信号。编码器包含 7 512 个参数，现场可编程门阵列上的延迟为 147 ns；它只需学习一次事例表示，便可据此生成多种筛选，为面向 LHC 基于硬件的触发器的基础模型铺平道路。

版权归作者所有。本作品投稿至 SciPost Physics。许可信息将在发表后公布。出版信息将在发表后公布。

收稿日期 接受日期 发表日期

目录

1 引言 2 2 数据集与实验设置 3 3 方法 4 4 结果 8 5 现场可编程门阵列实现 14

6 结论 14

A 过程列表和类别划分 15

B 训练详情 18

C 消融研究 18

D 附加诊断 21

参考文献 27

1 引言

大型强子对撞机（LHC）上的 ATLAS 和 CMS 实验会实时筛选碰撞：第一级触发器由定制硬件和现场可编程门阵列（FPGA）实现，随后由软件触发器进行筛选 [1,2]。在高亮度大型强子对撞机（HL-LHC）上，CMS 的一级触发（L1）将在 12.5 μ秒的延迟内，将 40 MHz 的束团交叉率降至 750 kHz [2]。任何单个算法可用的预算时间仅有几微秒，而在此阶段被拒绝的每个事例都会丢失。为适应 HL-LHC，CMS 将径迹探测器信息纳入 L1，并在关联触发器中将其与量能器和 μ子信息结合，执行简化的粒子流重建，并为每个粒子赋予堆积权重（PUPPI） [2,3]。随后，L1 算法便可处理重建粒子列表，而不再仅仅依赖量能器塔和 μ子短径迹。不过，该粒子列表是离线重建的一个降级版本，这意味着径迹追踪的角度范围受限、分辨率较粗，且阈值较高。

触发菜单由一组算法组成，每个算法都针对一种事例特征设计。迄今为止，机器学习触发器也遵循相同模式。CMS 的异常检测触发器 AXOL1TL 和 CICADA、 [4–7]，喷注标记器 [8] 以及 τ标识器，各自都有独立的模型、训练样本和现场可编程门阵列资源份额。另一种方案是使用单个编码器，将每个事例映射为一个向量，再通过廉价的下游操作从中得出多个判定。在离线分析中，这正是粒子物理学基础模型的基本思路 [9–12]：预先训练一次表示，并将其重复用于多项任务。我们将同一原则应用于构建事例级编码器，使其足够小巧，能够用于触发器，并可在无需重新训练的情况下重复使用。

我们通过零样本学习 [13]，将编码器应用于训练数据之外的数据。模型在一组已见类别上训练，并在未见类别上评估。我们采用类别留出意义上的这一术语，探究从已见过程中学到的表示是否仍适用于未见过程，以及同一编码能否支持多种触发任务，包括对标准模型（SM）和超出标准模型（BSM）过程的背景抑制与异常检测。

在 L1 进行此类嵌入学习的障碍，在于与离线视图相比，L1 数据的视图较为粗糙。COLLIDE-2V 数据集 [14] 同时提供了同一组模拟碰撞的离线视图和 L1视图。借助这一独特的数据集，我们能够通过对比目标 [16] 应用知识蒸馏 [15]，利用离线视图逐事例监督 L1视图。粒子物理学中最接近的先例是 RS3L [9],，它将同一喷注的两次重新模拟配对；而在这里，两个视图是对一次碰撞进行的两种探测器重建，并由容量不同的网络进行编码。

用于喷注标记的 Transformer 最近已在现场可编程门阵列上实现，延迟约为 100 ns（参考文献 [8] 及其中的参考文献），使用 hls4ml [17, 18] 和高粒度量化（HGQ） [19], , 该方法为每个权重和激活值学习位宽，并结合分布式算术优化 [20]。我们通过 Alkaid 编译器使用 HGQ 和分布式算术 [21],, 并将喷注级标记推进到事件级嵌入。

我们使用监督式三元组损失训练离线 Transformer，并逐个事例将其蒸馏到量化 L1 Transformer 中；其位宽在固定资源预算下进行学习（图 1）。随后，我们利用该嵌入且无需重新训练，剔除最小偏置碰撞——这类通用软相互作用几乎构成了全部输入速率——并通过两阶段筛选，选出不同于任何已见标准模型过程的事件，从而发现并标记超出标准模型过程。我们还提供了编码器的 FPGA 实现，使其能够在触发条件下运行。

2 数据集与实验设置

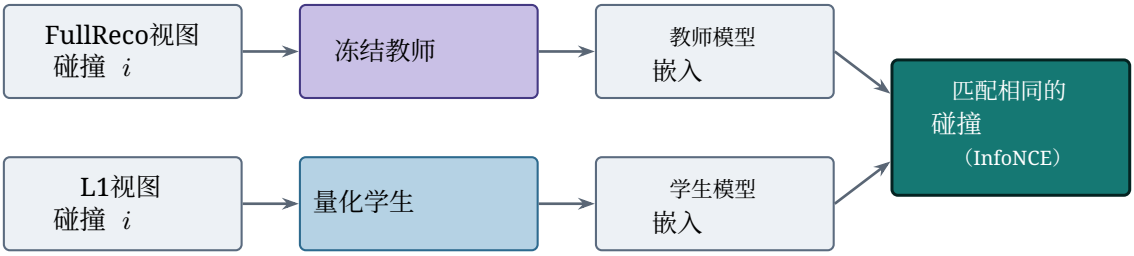
2.1 COLLIDE-2V 数据集

COLLIDE-2V [14] 是一个公开样本，包含约 7.5 亿个模拟质子–质子碰撞事件，由 MADGRAPH [22] 和 PYTHIA [23] 生成，并在高亮度大型强子对撞机条件下通过 DELPHES [24]处理；每次束团交叉平均发生 200 次同时的质子–质子相互作用（堆积）。每个事件以两种视图存储。离线视图（FullReco，FR）是通过经过调优的 DELPHES 卡获得的、类似 CMS 的完整重建。*L1* 视图由一级触发器重建的另一种参数化生成，其顶点、径迹和量能器性能有所降低，径迹重建限制在一定赝快度范围内，p_T 阈值更高，堆积权重也相应调整。在这两种视图中，我们都使用 PUPPI [3] 粒子候选列表，并按p_T降序存储。

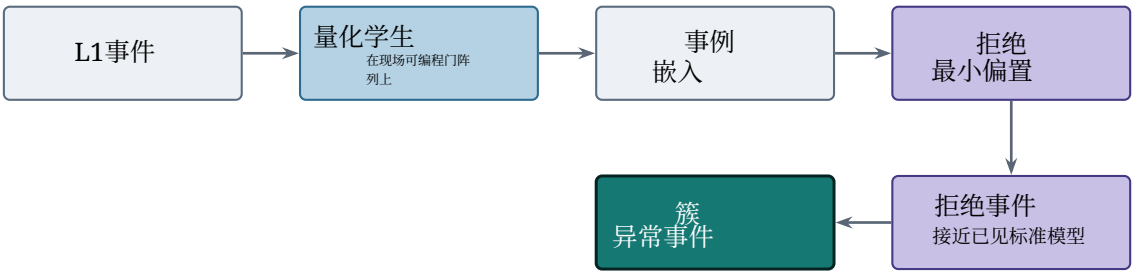
我们使用目录中的 53 种标准模型过程和 52 个超出标准模型（BSM）基准样本。标准模型样本涵盖希格斯玻色子、双希格斯玻色子、矢量玻色子、顶夸克和量子色动力学生成，$\Upsilon \to \ell\ell$以及最小偏置过程。超出标准模型基准样本包括隐藏谷双轻子、提示衰变和位移衰变的R宇称破坏超对称、类 SUEP 簇射、希格斯玻色子 $\to$ 双轻子衰变为底夸克、τ 轻子或光子、一个质量为 500 GeV 的$Z' \to$ 夸克¯夸克，以及参考文献 [25], 中的七个轻标量样本，其中一个是控制样本。若干样本具有位移衰变，其本征衰变长度为 1 至 100 毫米。完整样本清单及事件数见附录A。本文使用的标准模型样本每种包含 20 万至 150 万个事件（中位数为 141 万个，总计 7020 万个）；超出标准模型样本每种约包含 5 万个事件。各类别按相同权重抽样，而非按截面抽样，因此所有接受率和保留率均针对平衡的过程混合，而非物理触发率（第4.1节）。

2.2 输入与预处理

每个事例由其 PUPPI 候选粒子表示，每个粒子有 14 个特征：横向动量 p_T、赝快度 η、方位角 ϕ、电荷、质量、横向和纵向冲击参数 d_0 和 d_z 及其不确定度，以及分别表示带电强子、中性强子、光子、电子和 μ子的独热粒子类型标志。冲击参数d_0 对应真实粒子轨迹，不受分辨率效应或接受度限制的影响，因此在长寿命样本中，L1视图会保留 $|d_0|$达数厘米的带电候选粒子。基线径迹重建主要重建即时径迹；重建位移径迹则需要采用效率较低的扩展模式 [2, 26]。纵向参数d_z 是相对于 $z = 0$ 测量的，主要反映主顶点的位置。PUPPI权重为零的候选粒子（堆积候选粒子，占存储的L1列表的 95% 以上）以及 $p_T \leq 2$GeV 的候选粒子会被移除。教师模型读取排名前 256 的 FR粒子，而



(a) 冻结的 FullReco 教师模型为可训练的 HGQ L1 学生提供配对事例目标。



(b) 事件嵌入支持最小偏置拒绝、已见标准模型新颖性选择和群体聚类。

图 1：事件嵌入的配对视图训练（上）与使用（下）。在阶段 B 中，通过带有 HGQ 资源惩罚的对称 InfoNCE 损失，使学生模型针对每次碰撞的嵌入与冻结教师模型的嵌入相匹配。第 3.5 节中的门控和聚类均在嵌入上离线评估。

学生模型读取排名前 32 的 L1 粒子，较短的列表用零填充。

同一次碰撞的 FR 视图和 L1 视图构成一对；若任一视图经过筛选后为空，则丢弃该配对。约三分之一的最小偏置事件、37% 的 Drell–Yan 事件以及五分之一的 Υ 事件没有通过筛选的 L1 候选，因此这些过程仅由活动性较高的事件表示。

2.3 已见/未见类别划分与样本

标准模型过程被随机划分为25个已见类别和28个未见类别（表 4）。协议要求将最小偏置设为未见类别，以便第 3.5 节中的门控能够基于编码器从未见过的过程进行校准。此划分满足这一要求。两个训练阶段都仅使用已见类别，因此最小偏置事件仅在评估阶段作为参考和校准样本（第 3.5 节）使用。全部52个超出标准模型样本仅用于评估。

数据集中的每个文件按行划分为训练部分（前90%）和验证部分（后10%）。教师模型和学生模型使用配对训练行中相同的类别平衡子集进行训练，每个已见类别包含168990对事件（总计422万对），验证子集则每个类别包含18 340对。所有模型均在每个标准模型过程和超出标准模型过程的相同1 000个事件，以及10 000个最小偏置事件上进行评估。

3 方法

3.1 编码器架构

我们训练两个 Transformer 编码器 [27],：一个教师模型和一个学生模型，两者均作用于粒子集合（图 2，表 1）。每个粒子的特征经两个全连接层投影至模型宽度，

	教师模型	学生模型
输入视图	FullReco	L1
输入粒子	256	32
模型宽度	128	16
Transformer 模块	6	2
注意力	多头，4 个注意力头	Linformer，1 个头，$k = 2$
归一化	层归一化（残差后）	none
读出	全局平均池化	可学习类别标记
算术	浮点	HGQ定点
参数	623 296	7 512
EBOP（训练后）	–	4.93×10^5

表 1：教师模型和学生模型配置。位宽根据每个参数和激活值进行学习（第 3.4 节）。

随后经过带残差连接的注意力块和前馈块，并归约为单个事件向量。学生模型的 Linformer 投影作用于按 p_T 排序的粒子槽位列表，与参考文献 [8]中的做法相同。

该教师模型是一个浮点模型，其前馈块的宽度为 128，并使用 GELU激活。其事件向量是 256 个粒子槽位的全局平均值，随后通过线性投影映射到 128 维。该学生模型的宽度为 16，包含两个块，采用单头 Linformer注意力 [28]，将键和值投影到$k = 2$ 个序列行，使用一个 $16 \to 32 \to 16$前馈网络、ReLU激活且不进行归一化。一个可学习类别标记 [29, 30]经过量化后，添加到 32 个粒子标记之前，因此注意力堆栈作用于 33 个标记。最后一个块之后，该标记的输出作为事件向量，并通过线性投影映射到 128 维。类别标记读取取代了全局平均池化；对于固定长度的零填充输入，全局平均池化会将填充槽位纳入平均，并且需要在定点中进行除法。由于最终投影是从一个 16 维标记进行线性投影，学生模型的 128 维输出位于一个 16 维仿射子空间中。全部 7 512 个学生模型参数和激活值均使用 HGQ 进行量化（第 3.4 节）。训练后的学生模型达到 4.93×10^5有效位操作（EBOP），其 29 376 个学习得到的量化器的平均位宽为 2.4 位，量化器位宽中位数为 1.1 位。在每次损失计算和每次评估之前，两个事件向量都会归一化为单位长度（L2 归一化），因此所有距离都是单位球面上的弦距离，所有相似度都是余弦相似度。

3.2 阶段 A：教师模型的监督式度量学习

教师模型f_θ 使用三元组损失，在已见类别的离线视图上进行训练 [31]。对于锚点a、来自同一类别的正样本p 以及来自不同类别的负样本n ，

$$\mathcal{L}_{\mathrm{A}} = \max\bigl(0,\ \|\hat{z}_a - \hat{z}_p\|_2 - \|\hat{z}_a - \hat{z}_n\|_2 + \alpha\bigr), \qquad \hat{z} = \frac{f_\theta(x^{\mathrm{FR}})}{\|f_\theta(x^{\mathrm{FR}})\|_2}, \tag{1}$$

边距为 $\alpha = 0.5$。三元组从按类别分层的流中即时构造。锚点和正样本是从按类别划分的蓄水池中抽取的两个事例；负样本则从另一个类别中抽取，抽取概率与该类别缓冲的事例数成正比，不进行困难负样本挖掘。教师模型使用包含 128 个三元组的批次和 AdamW [32] 进行 50 个训练轮次，学习率为 10^{-5} (附录B)。

我们使用三元组损失训练教师模型，因为度量损失直接作用于嵌入空间中的距离，而学生模型继承的正是这种几何结构。在一项初步比较中，我们选取了五个

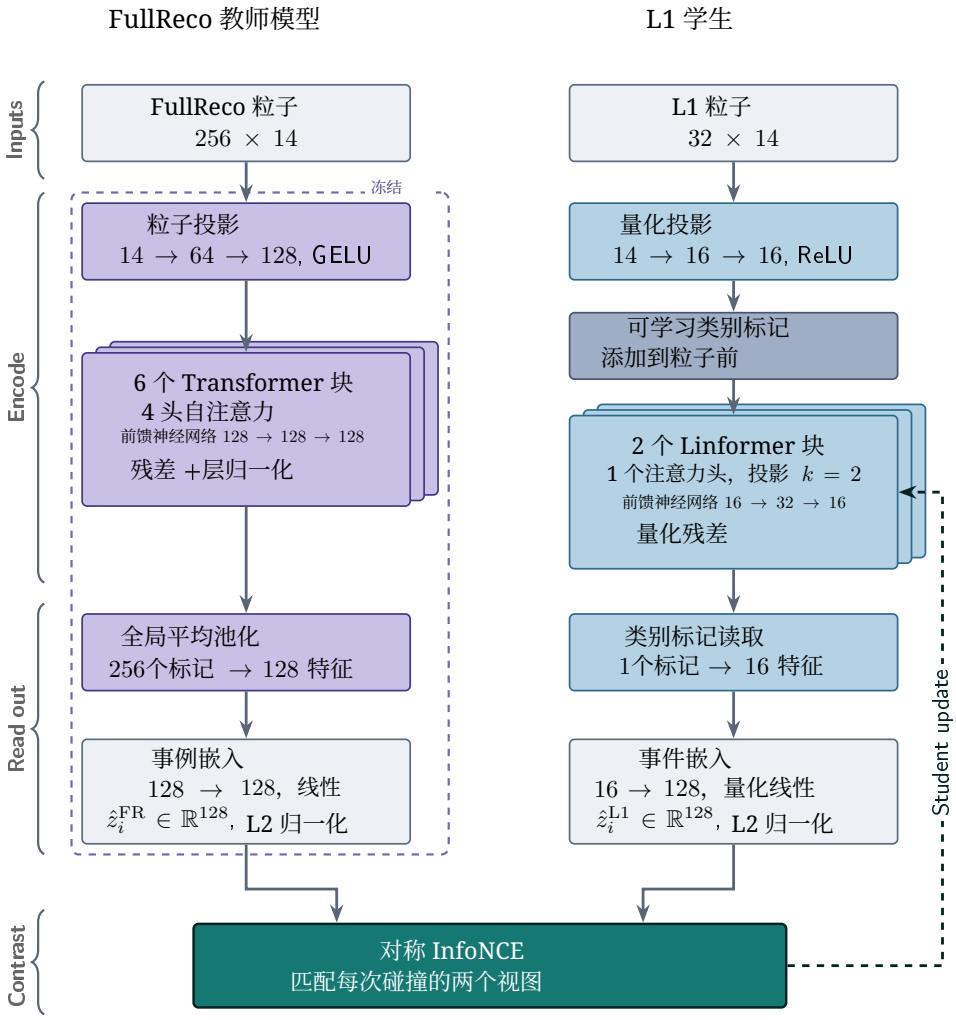


图2：编码器架构与蒸馏损失。FullReco 教师模型（虚线框）首先使用三元组损失进行训练，然后被冻结。对称的 In- foNCE 损失将每个 L1事件的学生模型嵌入拉向同一碰撞的教师模型嵌入，并使其远离批次中的其他事件。仅更新学生模型，并加入 HGQ 资源惩罚。

对于类别而言，它产生的类别区域比交叉熵分类器更加紧凑且排列更有序。我们没有进一步优化这一选择，因为本文重点在于蒸馏及其在现场可编程门阵列上的部署。

3.3 阶段 B：将事例配对蒸馏至一级触发

冻结教师模型后，学生模型g_ϕ 将每个 L1 事件嵌入，使其与配对的教师嵌入相匹配，并与批次中的其他嵌入区分开来。使用$s_{ij} = \hat{z}^{\mathrm{FR}}{}_i \cdot \hat{z}^{\mathrm{L1}}{}_j / \tau$，对称 InfoNCE 损失 [33] 为

$$\mathcal{L}_{\mathrm{B}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ij}}} + \log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ji}}}\right], \tag{2}$$

温度为 $\tau = 00.05$，批次大小为$N = 128$。阶段 B 不使用类别标签。负样本是批次中的其他事件，无论它们属于哪个类别。在 25 个均衡类别中，约有 4% 的负样本与锚点属于同一类别。因此，该损失要求学生模型区分单个事件，而不仅仅是类别。附录D.5表明，对于已见和未见过程，该方法都能做到这一点。每个学生模型最多训练 250 个训练轮次，使用 AdamW，学习率为 10^{-5}。

3.4 资源预算约束下的量化感知训练

学生模型使用 HGQ 进行训练 [19]，： HGQ 会为每个权重和激活值附加一个定点量化器，并通过梯度下降与权重一同学习位宽。训练目标为

$$\mathcal{L} = \mathcal{L}_{\mathrm{B}} + \beta \cdot \overline{\mathrm{EBOPs}}, \tag{3}$$

其中，$\overline{\mathrm{EBOP}}$ 是对模型 EBOP 的可微估计值，是 FPGA 逻辑资源用量的代理指标，计算时将每次乘法的两个操作数位宽相乘。系数 β 每个训练轮次调整一次，调整方式采用比例–积分控制器，推动模型达到 5×10^5 EBOP 的目标值；当模型超过目标值时增大 β，达到目标后则将其减半（附录 B）。报告所用的检查点，是满足预算的训练轮次中验证损失最低的那个。附录 C.2比较了两种注意力量化器配置。主结果学生模型采用第一种配置（附录 B）。

3.5 嵌入上的触发函数

所有决策均使用 L2 归一化的学生嵌入，以及与待评估事件相分离的带标签参考集。决策基于缓存的嵌入离线计算。

最小偏置门控。 事例的评分取其到最小偏置嵌入参考集中的 $k = 20$ 个最近邻的欧几里得距离的平均值，即最近邻异常值分数 [34,35]，分数越高，表示越不像最小偏置事件。阈值为 $(1-r)$分位数，取自独立校准集中的最小偏置事件分数，因此其中 r比例的事件能够通过。由于最小偏置是一个未见类别，编码器和归一化统计量均未使用任何最小偏置事件。门控使用来自训练中从未使用过的文件的 653 661 个最小偏置事件，随机进行十次划分，每次分为 50 000 个参考事件、50 000 个校准事件和 553 661 个测试事件；我们报告在测试事件上测得的接受率。信号保留率按每个过程 500 个事件测量。关于 k 的选择，详见附录 D。

已见标准模型新颖性门。 通过最小偏置门控的事件，将分别与 25 个已见类别进行评分。对于已见类别 c，设其参考集为 R_c，校准集为 C_c，令 $d_c(x)$ 表示从 x 到 R_c中第 k 个最近邻的距离，其中 $k = 10$。该分数按类别在对数尺度上进行标准化，

$$s_c(x) = \frac{\log d_c(x) - \mathrm{median}_{y\in C_c} \log d_c(y)}{\mathrm{IQR}_{y\in C_c} \log d_c(y)/1.349}, \qquad s(x) = \min_c s_c(x), \tag{4}$$

从而使宽分布和窄分布的已见类别具有可比性。得分 $s(x)$ 衡量x与最近的已见类别之间的距离，以该类别的离散程度为单位表示。s 的阈值设定为使最小偏置校准事件中有 1% 通过该阈值，与第一道门控的阈值分别设定。一个事例只有在其既远离最小偏置分布、又远离所有已见类别时，才能通过两道门控。该筛选在每个过程中选取 300 个事例进行评估；对于每个已见类别，参考集和校准集分别包含另外 400 个和 200 个事例，最小偏置样本则各包含 3 000 个事例。

聚类。 将通过筛选的样本使用在已见类别参考嵌入上拟合的主成分分析进行投影 [36]（32 个成分），并使用 HDBSCAN 进行聚类 [37]（最小簇大小为 20，最小样本数为 10）。这两个步骤均不使用标签。这些设置在分析开始前已固定，附录 D.4 研究了它们的影响。标签仅用于对结果进行评分，具体依据每个簇的过程组成以及附录 D.4中定义的逐样本得分。

3.6 基线方法

我们在匹配的最小偏置接受率下，对相同的 L1 事件比较五种评分：希格斯玻色子$_T$（候选体标量p_T 和）、候选体多重性、基于 18 个手工构建的事件特征计算的最近邻评分、基于学生模型的 32×14 输入计算的相同评分，以及自编码器重构误差。最近邻评分使用$k = 20$。希格斯玻色子$_T$、多重性和手工构建特征评分使用所有筛选出的候选体，而输入级分数和自编码器使用学生模型的前 32 个粒子。实现细节见附录D。

自编码器仅使用最小偏置事件进行训练。对已见标准模型、未见标准模型和超出标准模型事件进行评估时，不重新训练；已见与未见由嵌入的训练划分定义。每种评分的阈值均根据独立的最小偏置校准样本确定。离线 FullReco 教师模型单独列出，作为参考。我们还使用相同的数据、资源目标和初始化随机种子，在已见 L1 类别上采用阶段 A 三元组损失训练同一 HGQ 学生模型，但不使用教师模型。

4 结果

4.1 最小偏置拒绝

在 1% 的最小偏置接受率下，即 99% 的拒绝率下，嵌入保留了已见标准模型过程的中位数 90%、不包括最小偏置过程的未见标准模型过程的 96%，以及超出标准模型过程的 88%（表 2），实测测试接受率为 0.99%。尽管嵌入是在 25 个标准模型过程上训练的，且其阈值设定在它从未见过的最小偏置事件上，但它对未见标准模型过程和超出标准模型过程的保留率与对训练过程的保留率相当。未见过程略高的保留率反映了各组的构成，希格斯玻色子$_T$ 也显示了这一点。由于所有过程均以相同权重采样，这些是逐过程效率，而不是触发率。

在嵌入空间中，大多数过程都远离最小偏置（图3）。第十分位过程在三组中的事件保留率分别为 36%、57% 和 60%，而第 90 百分位

最小偏置接受率为 1% 时的得分	已见标准模型（%）	未见标准模型（%）	超出标准模型（%）
L1嵌入（本工作）	**90**	**96**	**88**
同一学生 , 无教师模型	86	95	85
手工构建特征、最近邻	78	91	82
H_{T}	86	**96**	75
自编码器	70	79	51
学生模型输入、最近邻	70	79	46
多重性	70	78	30

表2：在1%最小偏置接受率下，L1嵌入、未使用教师模型训练的同一学生模型，以及五种触发基线（第3.6节）的过程保留率中位数；所有方法均在大型最小偏置样本上完成校准。

则超过 99%。保留率最低的过程具有较低的粒子多重性，且末态以光子或中性粒子为主；对于这些过程，L1视图不包含径迹信息，例如 $\Upsilon \to \ell\ell$ (2%)、三光子产生（32%）以及 希格斯玻色子 → 双轻子 → 4γ （30–35%）（图 9）。

在三组数据中，嵌入的得分表现最佳（表2，图4）。它在已见标准模型过程和超出标准模型过程中分别比希格斯玻色子$_{\mathrm{T}}$ 高出4个百分点和13个百分点，在未见标准模型过程中则与其持平；与手工构建特征相比，它在标准模型过程中高出5至11个百分点，在超出标准模型过程中高出6个百分点。其保留率远高于自编码器、多重性，以及在学生模型输入上计算的相同距离，因此这种区分能力来自编码器，而非距离分数。在离线视图中，FullReco 教师模型对每组数据的保留率均达到94%至99%（图4），这说明L1视图以及压缩至7512个参数所带来的代价。

4.2 筛选与任何已见过程都不同的事件

第二个触发函数会选出既不像最小偏置事件、也不像任何已见标准模型过程的事件。两道门控共同放行了 0.23% 的最小偏置事件，而新颖性门几乎剔除了所有通过第一道门控的标准模型事件。没有任何已见过程保留了超过 2.7% 的事件（中位数为 1.0%）；未见标准模型过程也几乎同样有效地被剔除（中位数为 1.3%，最高为 3.3%）。因此，该门控能够拒绝类似已见过程的标准模型过程，而不仅仅是拒绝用于校准它的 25 个过程。

保留率最高的 20 个过程都是超出标准模型样本（图5），分为两组。长寿命隐藏谷和希格斯玻色子 → 双轻子样本的本征衰变长度为 10 或 100 mm，保留率为 12% 至 37%，是最高标准模型保留率的四到十一倍。这得益于冲击参数。推理时若将d_0 设为零，$c\tau = 100$ mm 样本的保留率便会降至标准模型水平。瞬发高质量双轻子样本中，质量为 500 和 1 000 GeV 的隐藏谷Z'模型，保留率最高可达 29%。强子型超出标准模型样本能够通过第一道门控，但与已见的顶夸克伴随、双希格斯玻色子及多喷注过程相似，因此第二道门控会按设计将其剔除。针对这些样本，可在相同嵌入上采用有监督选择。

保留下来的事件在没有标签的情况下按事例特征分组。在测试的全部 27 种聚类设置中，HDBSCAN 均找到一个包含 327 个事件的簇，其中 97% 来自瞬发高质量双轻子样本。长寿命事件构成第二个簇，包含 543 个事件，其中 80% 为长寿命事件，7% 为标准模型事件（附录 D.4）。共享相同事例特征的样本在 L1 看起来相似，因此这些簇区分的是事例特征，而非单个生成器样本。

两类样本群体都依赖于从离线视图进行的蒸馏（表 6，附录 C.1）。由 L1 视图教师模型蒸馏得到的同一学生模型，仅保留 8% 的长寿命事件和 0.3% 的高质量双μ子事件，相比之下，前者和后者分别为 24% 和 26%，且未形成任何簇。没有教师模型时，更多标准模型事件会通过新颖性门（某个过程的比例最高可达 6.3%），而这两类样本群体并不

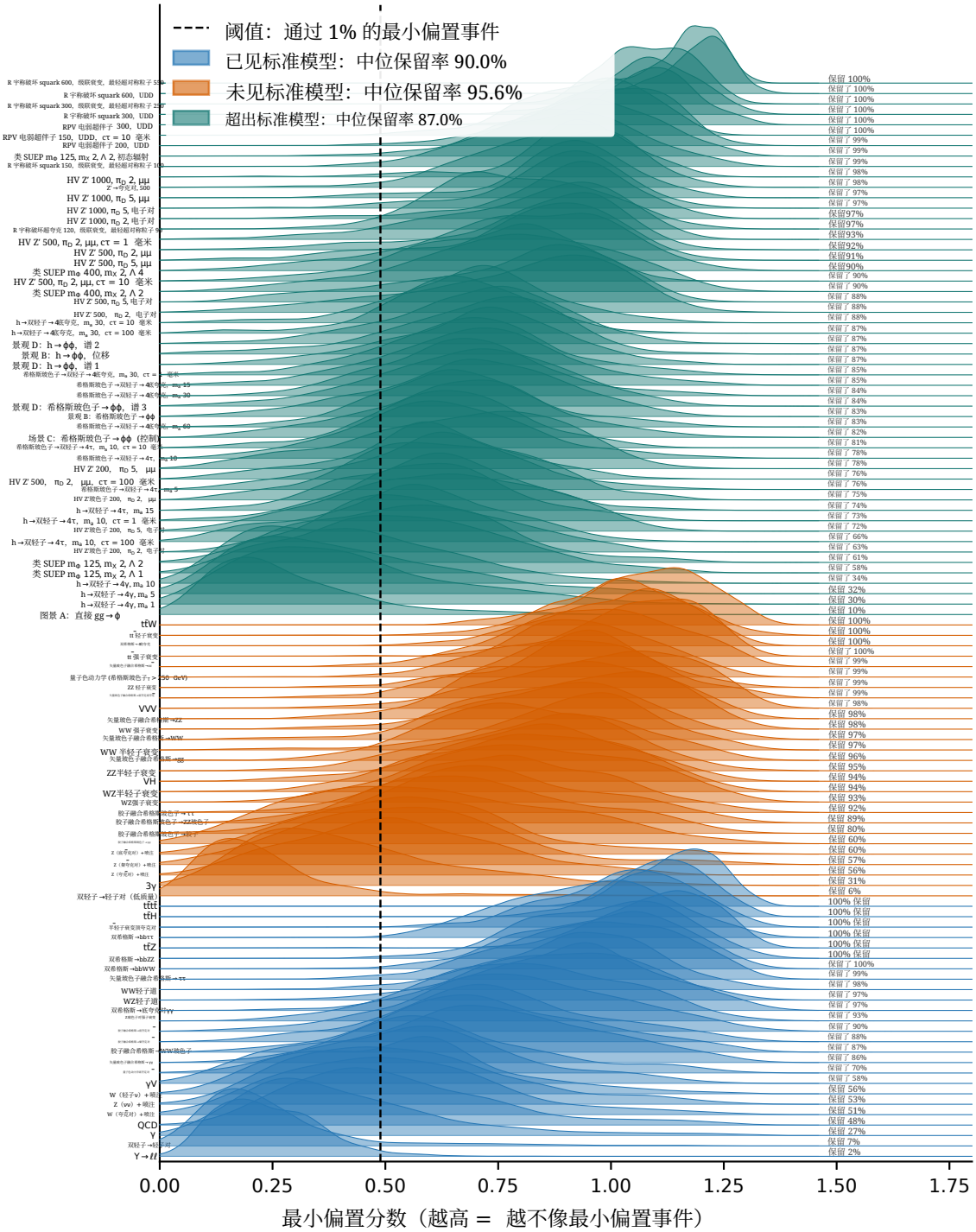


图3：每个已见标准模型（底部）、未见标准模型（中部）和超出标准模型（顶部）过程的最小偏置分数分布。该分数为到最近20个最小偏置参考事件的平均距离。图中展示了一个数据划分，每个过程包含500个事件。在每组中，按超过虚线阈值的事件比例对过程排序；该阈值对应通过筛选的最小偏置校准事件中1%的分位点。图例给出了该数据划分中各组的中位数。第4.1 节给出了各数据划分的平均值，图9给出了各过程的保留率。

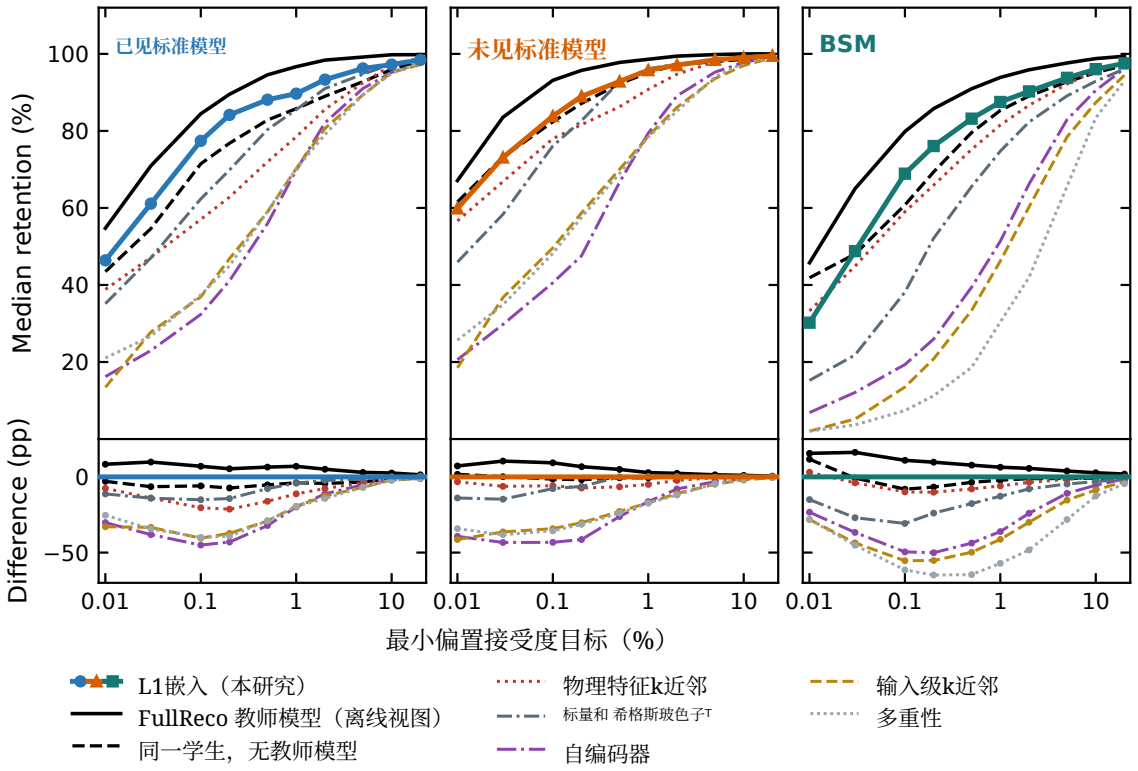


图4：L1嵌入、离线 FullReco 教师模型、未使用教师模型训练的同一学生模型，以及第3.6节定义的五种基线方法的比较。从左至右，各列分别展示已见标准模型、未见标准模型和超出标准模型过程。主图给出了第 3.5 节十个校准划分的平均中位保留率。底部各图展示了每个比较模型的保留率与相同接受率目标下 L1 嵌入保留率之差。正值表示比较模型表现更优。

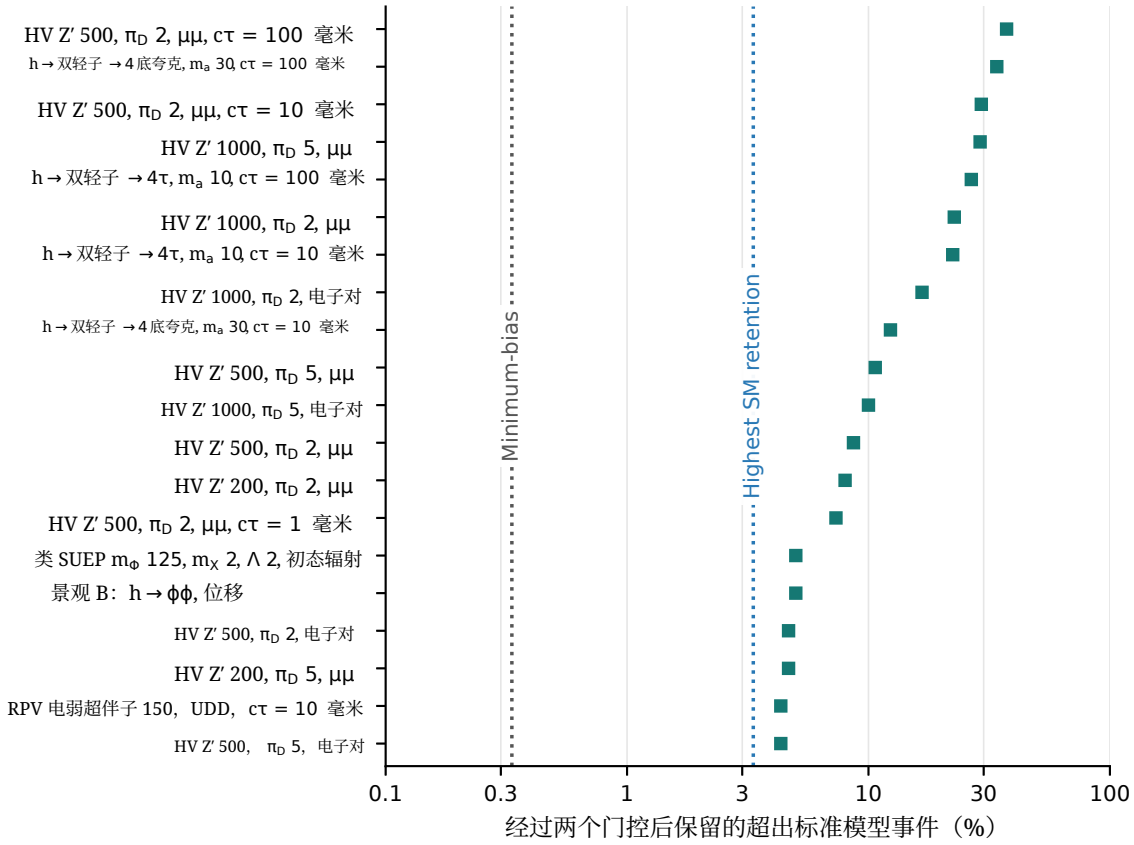


图 5：在每个过程的 300 个流事件中，经两道门控后保留率最高的 20 个超出标准模型过程。蓝色点划线表示已见标准模型过程和未见标准模型过程中最高的保留率；灰色点划线表示通过两道门控的最小偏置事件比例。

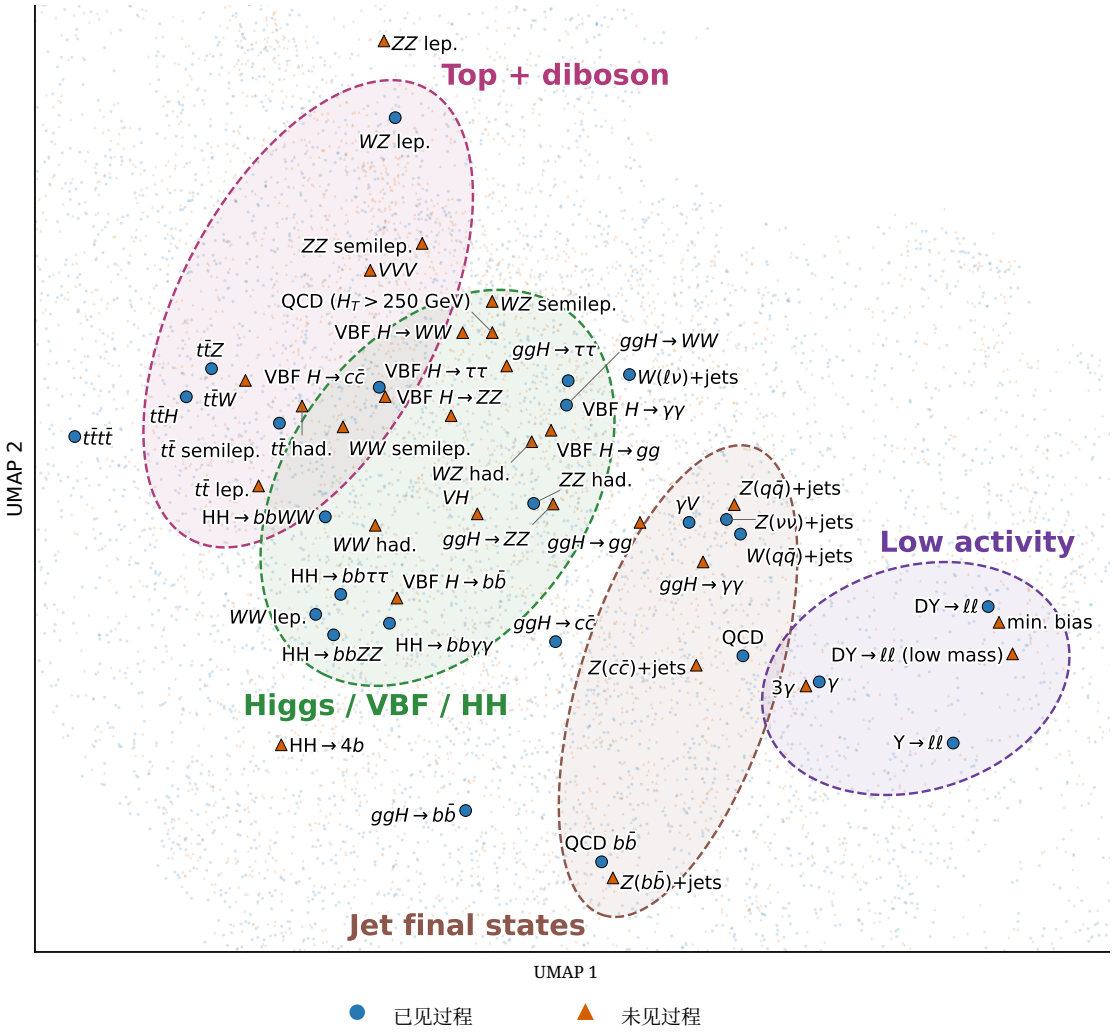


图 6：学生嵌入的 UMAP 可视化，其中每个事例都与其沿束流轴的镜像取平均，以消除其取向（附录 D.3）。浅色点表示事件，带标签的圆形和三角形标记分别表示已见和未见标准模型过程的中位数。

形成独立的簇。哪些事例特征能够保留下来，也取决于已见过程。采用第二次随机划分后，长寿命样本的保留率更高，而大多数即时双缪子事件都被移除（附录C.3）。

4.3 嵌入的结构

图 6 展示了学生模型嵌入中已见和未见标准模型过程的 UMAP [38] 可视化结果。顶夸克伴随产生和轻子双玻色子过程位于映射的一侧，胶子融合、矢量玻色子融合和双希格斯玻色子产生位于中心，而W/Z+喷注和量子色动力学过程则位于它们旁边。最小偏置样本位于角落中，与其他低活动度过程——Drell–Yan过程、 $\Upsilon \to \ell\ell$以及光子样本——彼此相邻。128维嵌入呈现出相同的模式。未见$ggH \to ZZ$的最近质心是已见的$ggH \to WW$；$t\bar{t}W$对应的最近质心是已见的$t\bar{t}Z$；而未见最小偏置样本的最近质心是已见的 Drell–Yan样本。

未见类别保留了已见类别局部分离性的 89%。一种带标签的参照 e

模型	延迟 (ns)	LUT (k)	FF	DSP	BRAM	II (clk)
Linformer	147	239.4	99 957	12	0	1

表3：VU13P 上编码器完成布局布线后的资源占用。LUT 数量以千为单位；FF、数字信号处理器和块随机存取存储器均为绝对数量。

最近邻探针（$k = 10$，每个类别使用 500 个参照事件和 500 个查询事件，并在各组内进行）在 25 个已见类别上达到 20.6%，在 28 个未见类别上达到 18.2%，而随机猜测水平分别为 4.0% 和 3.6%（图 7），约为 FullReco 教师模型 28.7% 和 24.6% 的四分之三。绝对准确率不高，因为许多已见过程仅在一种玻色子的衰变方式上有所不同，无法逐事例区分。尽管如此，嵌入仍会按物理内容对事件进行分组，并且对从未见过的过程，其表现几乎与训练过的过程一样好。

5 FPGA 实现

量化学生模型通过 Alkaid 转换为固件 [21]，，使用参考文献中引入的 Linformer 和多头注意力支持 [8]。常数矩阵—向量乘法通过分布式算术实现 [20]。目标器件是 AMD Virtex UltraScale+ VU13P，属于为 CMS 第二阶段关联器触发器预定的现场可编程门阵列系列 [2]。

表 3给出了使用 Vivado 2025.2.1 对`xcvu13p-flga2577-2-e`器件完成布局布线后的资源占用情况。该设计完全展开并采用流水线结构，全部布局于器件四个超逻辑区（SLR）中的一个区域内；这些区域是器件彼此独立的硅裸片。其延迟为 147 ns。由于平均位宽低于三位，算术运算几乎完全通过查找表实现。第 3.5节的距离分数在线下计算，而计算与少数几个质心的距离或使用学习型头部，则可以提供紧凑的固件门控。

6 结论

我们通过将离线 Transformer 蒸馏到量化 L1 Transformer 中，训练了一个用于一级触发的事件级嵌入。训练时，我们使用同一模拟碰撞的配对视图，使离线重建能够逐个事例地监督 L1 重建。该嵌入将未见标准模型过程映射到其物理上相关的过程附近，并保留了已见标准模型过程局部分离性的 89%。基于该嵌入计算的单一最近邻评分，在编码器从未见过的最小偏置碰撞上进行校准后，可剔除其中 99% 的事例。该评分保留了已见标准模型过程的中位数 90%、未见标准模型过程的 96% 和超出标准模型过程的 88%；在相同事例上与自编码器、输入级分数和运动学基线方法相比，它在这三类过程上的表现均为最佳。

针对已见过程设置的第二道门控几乎剔除了所有标准模型过程，同时保留了很大比例的即时高质量双轻子信号。借助一级触发器上的位移径迹撞击参数，它还保留了若干长寿命信号中至少十分之一的事件。剩余事件无需标签即可分为两个群体：纯度为 97% 的即时双轻子群体，以及纯度为 80% 的长寿命群体。两个群体都依赖于从离线视图进行的蒸馏。采用 L1 视图教师模型时，只有少量高质量双缪子事件能够保留；不使用教师模型时，则无法形成彼此独立的簇。该编码器可部署在一块用于一级触发器的现场可编程门阵列的一个超逻辑区中，延迟为 147 ns。

展望。 编码器使用部分可用数据进行训练，并在标准模型过程的两个数据划分上进行评估。通过使用更多数据和过程进行训练，并采用更多数据划分进行评估，可以了解向未见过程迁移的能力还能提升到何种程度。教师模型的三元组损失和学生模型的对比损失都未得到充分优化；其他目标函数、更大的教师模型以及其他聚类算法，或许能够得到更清晰的嵌入和纯度更高的群体。将门控逻辑部署到固件中，便可测量从粒子到决策的完整链路；最终还必须在碰撞数据上测试该方法，并以未经过任何触发选择而记录的零偏置事件作为参照。教师模型使用模拟过程标签进行训练，而在碰撞数据混合样本上采用弱监督 [39] 则有望降低对模拟的依赖。

致谢

我们感谢 COLLIDE-2V 的作者提供数据集，感谢 Oz Amram 提供 COLLIDE-2V 的超出标准模型样本，也感谢 HGQ、hls4ml、da4ml 和 Alkaid 的开发者提供工具和建议。我们感谢帝国理工学院研究计算服务（http://doi.org/10.14469/hpc/2232）提供的计算资源和支持。

资助信息 轻子对由科学与技术设施委员会（STFC）资助，资助编号为 ST/W000636/1。BM 感谢通过 STFC Ernest Ruther- ford Programme 获得的支持。LR、AC 和轻子对感谢 COST行动EPIGRAPHY（CA24153）提供的支持。

数据可用性 COLLIDE-2V 数据集 [14]已公开提供。

A 过程列表和类别划分

表格 4 和5 列出了第2.2节筛选前模拟样本中的标准模型过程和超出标准模型过程及其事件数。

已见标准模型（25）		未见标准模型（28）	
过程	事件	过程	事件
DYJetsToLL	499 661	HH_4b	1 499 060
HH_bbWW	1 220 458	VBFHWW	1 164 956
HH_bbZZ	1 238 527	VBFHZZ	1 161 110
双希格斯_底夸克伽马伽马	1 279 246	VBFHbb	1 265 403
HH_底夸克对τ轻子对	1 289 170	矢量玻色子融合产生的希格斯玻色子衰变至梨夸克对	1 136 392
量子色动力学_HT50toInf	1 158 265	矢量玻色子融合产生的希格斯玻色子衰变至胶子对	1 393 108
量子色动力学_HT50tobb	1 334 765	矢量玻色子伴随希格斯产生_包含	1 491 781
矢量玻色子聚变希格斯双光子	1 156 924	VVV_包含	1 220 691
矢量玻色子聚变希格斯双τ子	1 135 487	WW_强子	1 394 918
W+喷注至轻子和中微子	1 453 749	WW_半轻子	1 404 606
W玻色子喷注到夸克对	1 477 306	WZ_强子	1 329 343
WW玻色子_轻子衰变	1 441 092	WZ玻色子_半轻子衰变	1 352 558
WZ玻色子_轻子衰变	1 411 754	Z玻色子喷注到夸克对	1 486 935
ZJetsTovv	1 481 134	ZJetsTobb	1 486 973
ZZ_强子	1 324 962	ZJetsTocc	1 496 896
伽马	1 318 162	ZZ_轻子衰变	1 380 161
伽马_V	1 432 252	ZZ_半轻子衰变	1 344 927
胶子聚变产生希格斯玻色子并衰变为WW	1 498 590	胶子聚变产生希格斯玻色子并衰变为ZZ	1 498 562
胶子聚变产生希格斯玻色子并衰变为底夸克对	1 499 029	胶子聚变产生希格斯玻色子并衰变为双光子	1 428 695
胶子聚变产生希格斯玻色子并衰变为梨夸克对	1 458 757	胶子-胶子融合希格斯玻色子到胶子-胶子-胶子衰变	1 499 076
顶夸克对与希格斯玻色子_包含	1 410 706	胶子-胶子融合希格斯玻色子到 tau 轻子对衰变	1 488 506
ttZ_incl	1 431 410	最小偏置事例	1 500 000
tt_半轻子	1 462 009	tri_伽马	199 981
tttt_incl	1 470 000	ttW_incl	1 416 569
Y_至_轻子	1 500 000	tt_强子	1 464 185
		tt_轻子衰变	1 462 054
		DYJetsToLL _低质量*	999 365
		量子色动力学_HT250toInf*	880 072

表 4：53 个标准模型过程及其在已见过程和未见过程之间的随机划分（第 2.3 节）。事件数为此处所用样本在筛选前的数量，名称中省略了生成器后缀。 *在训练后添加，且对每个模型而言均为未见过程。

超出标准模型基准样本（52 个；每个约 50 000 个事件；仅用于评估）	
HVdilep_Zp1000_piD2_电子对	SUEP 类似_隐藏谷_mPhi125_mX2_Lam2_初态辐射
隐藏谷dilep_Zp1000_piD2_mumu	类似SUSY_隐藏谷_mPhi400_mX2_Lam2
隐藏谷双轻子_Zp1000_piD5_电子对	类似SUSY_隐藏谷_mPhi400_mX2_Lam4
隐藏谷双轻子_Zp1000_piD5_缪子对	Zprime_夸克对_m500
隐藏谷双轻子_Zp200_piD2_电子对	hToAA_4底夸克_ma15
隐藏谷双轻子_Zp200_piD2_mumu	hToAA_4底夸克_ma30
隐藏谷双轻子_Zp200_piD5_电子对	hToAA_4底夸克_ma30_ctau100毫米
隐藏谷双轻子_Zp200_piD5_mumu	hToAA_4底夸克_a粒子质量30_ctau10mm
隐藏谷双轻子_Zp500_piD2_电子对	hToAA_4底夸克_a粒子质量30_ctau1mm
隐藏谷双轻子_Zp500_piD2_mumu	hToAA_4底夸克_a粒子质量60
隐藏谷双轻子_Zp500_piD2_mumu_ctau100mm	hToAA_4伽马_ma1
隐藏谷双轻子_Zp500_piD2_缪子对_ctau10毫米	hToAA_4伽马_ma10
隐藏谷双轻子_Zp500_piD2_缪子对_ctau1毫米	hToAA_4伽马_ma5
隐藏谷双轻子_Zp500_piD5_电子对	hToAA_4tau_ma10
隐藏谷双轻子_Zp500_piD5_mumu	hToAA_4tau_a粒子质量10_ctau100毫米
RPV_ewkino150 _UDD_ctau10mm	hToAA _4tau_ma10_ctau10mm
RPV_电弱超对称粒子200 _UDD	hToAA _4tau_ma10_ctau1mm
RPV_电弱超对称粒子300 _UDD	hToAA _4tau_ma15
R宇称破坏_超夸克120_级联衰变_最轻超对称粒子90	hToAA_4τ子_a粒子质量5
R宇称破坏_超夸克150_级联衰变_最轻超对称粒子100	横向_LA_直接
R宇称破坏_超夸克300_UDD	横向_LB_希格斯玻色子
R宇称破坏_超夸克300_级联衰变_最轻超对称粒子250	背景格局_LB_希格斯玻色子_位移
R宇称破坏_超夸克600_UDD	背景格局_LC_控制（控制）
R宇称破坏_超夸克600_级联衰变_最轻超对称粒子550	背景格局_LD_谱1
SUEPlike_隐藏谷_mPhi125_mX2_Lam1	landscape_LD_谱2
SUEPlike_隐藏谷_mPhi125_mX2_Lam2	landscape_LD_谱3

表 5：52 个超出标准模型基准样本。HVdilep 样本是隐藏谷 $Z' \to$暗π介子 $\to$ 双轻子模型，标签由 Z' 质量（单位为 GeV）、暗π介子质量（单位为 GeV）和轻子味决定。RPV 样本是具有 UDD 耦合或级联衰变的 R 宇称破缺超对称模型，标签由超夸克或电弱微子的质量（单位为 GeV）决定。SUEPlike 样本模拟暗簇射产生的软非簇射能量模式，标签由参数 m_Φ、m_X 和 Λ决定。hToAA 样本是希格斯玻色子衰变为成对轻赝标量粒子的奇异衰变，轻赝标量的质量如所示（单位为 GeV）；Zprime_qq 是质量为 500 GeV 的 $Z' \to q\bar{q}$ 共振。后缀 ctau 表示位移样本的本征衰变长度。landscape 样本是参考文献 [25],中的轻标量模型，包括直接胶子融合产生（LA）、$h \to \phi\phi$ 的及时或位移衰变（LB）、具有另外三种质量谱的相同模型（LD），以及隐藏衰变通道关闭的控制样本（LC）。图中使用名称缩写，例如 HV Z' 500,π_D 2, $\mu\mu$, $c\tau = 100$mm 表示 HVdilep_Zp500_piD2_mumu_ctau100mm。

B 训练详情

预处理。 特征 p_T、d_0 和 d_z 经过带符号对数变换$(1 + |x|)$。随后，使用各视图训练样本的均值和标准差对连续特征进行标准化，而类型标志不进行缩放。经过筛选后，最小偏置事件平均有 3.9 个 L1 候选对象，超出标准模型事件有 19 个，标准模型事件有 26 个（标准模型中位数为 23）。由于 30% 的标准模型事件和 15% 的超出标准模型事件包含超过 32 个候选对象，学生模型平均读取 21 个（标准模型）和 16 个（超出标准模型）。

附录中的评估。 除非图注另有说明，附录 C 和 D 中的比较均采用第 2.3节的 10,000事件最小偏置样本，并将其划分 21 次，每次包含 4,000 个参考事件、3,000 个校准事件和 3,000 个测试事件。两阶段表格报告了这些测试事件上的最小偏置率。

教师模型（阶段 A）。 提前停止（在第 20 轮训练后启用，耐心值为 15 轮训练）从未触发，因此我们使用最后一轮训练后的模型。训练过程中，验证损失从 0.33 降至 0.23。

学生模型（阶段 B）。 我们按照第 1 表和第 3.3 节及3.4节所述训练表 10^{-8} 中的学生模型，在一块 NVIDIA H200 GPU 上每轮训练执行 33,006 步。每轮训练耗时 230 至 400 秒。HGQ 量化器从 8 位开始，权重采用 1 个整数位，激活值采用 6 个小数位；强度为 10^{-8} 的单调 L1 正则化项作用于位宽。权重在溢出时回绕，而注意力模块中的值则通过舍入实现饱和。 β 控制器作用于对数 EBOP 误差，比例增益为 4，积分增益为 0.1。EBOP 估计值起始于 7×10^6，并在 183 轮训练后达到目标。选定的检查点为第 216 轮训练，具有 4。93×10^5 EBOP。

注意力量化器。 附录 C.2 比较了注意力模块中两种量化器配置。在 KBI 中，数据通路量化器由符号位、总位数和整数位参数化，且整数位至少为 1 位；softmax 查找表固定为 5 位，其中整数位为 2 位。KIF 遵循参考文献 [8]。其数据通路量化器由符号位、整数位和小数位参数化，整数位至少为 0 位、小数位至少为 1 位，且 softmax 精度为学习得到。主结果学生模型采用 KBI。

消融实验。 每种教师模型配置均在同一个平衡子集上训练一次。小型教师模型有 4 个注意力头，前馈宽度为 16；提前停止会在训练 40 轮次后结束训练。读取 64 或 128 个粒子的学生模型采用相同的 EBOP 目标，其中 128 个粒子的配置在 β 上具有更高的上界。

C 消融研究

C.1 教师模型的贡献

表6 按照第4.1节的方案检验离线视图蒸馏的效果。该表比较了主结果学生模型、由大型 L1 视图教师模型蒸馏得到的同一学生模型，以及未使用教师模型、仅根据已见类别标签训练的同一学生模型。没有教师模型时，学生模型在最小偏置门处保留的已见标准模型和超出标准模型过程更少，其类别的可分性较差，通过第4.2节新颖性门的标准模型事件更多，而且长寿命粒子和双轻子粒子群未形成独立簇。L1 蒸馏学生模型的保留情况略

	FullReco 蒸馏	L1 蒸馏	无教师模型
最小偏置接受率为 *1%* 时的中位保留率（%）			
已见标准模型	90	**91**	86
未见标准模型	**96**	95	95
BSM	**88**	85	85
两阶段筛选（%）			
通过两道门的最小偏置事件	0.33	**0.00**	0.43
最高标准模型保留率	**3.3**	4.0	6.3
长寿命样本	**23.8**	8.2	13.1
高质量双缪子样本	**25.8**	0.3	2.0
最纯长寿命簇	**80**	–	23
最纯即时双轻子簇	**97**	–	15

表6：使用 FullReco 视图或 L1 视图，通过六个宽度为128的区块教师模型进行蒸馏，以及不使用教师模型训练得到的同一32粒子学生模型；训练样本为第3.5节所述的大型最小偏置样本。两阶段行的定义与表9相同，短横线表示未形成簇。

已见标准模型事件更多，且类别最容易区分。它通过两个门控的最小偏置事件最少。然而，它在门控处保留的超出标准模型事件较少，经过两个门控后保留的长寿命事件和双缪子事件也较少，而且没有形成任何簇。只有从 FullReco 教师模型蒸馏得到的学生模型，兼具较高的门控保留率、较低的标准模型事件经两个门控后的泄漏率，以及分别对应两类超出标准模型事件的明显簇。未使用教师模型训练的学生模型最终达到 2.3×10^5 EBOPs，低于其他模型的目标值。

C.2 教师模型视图与容量

我们测试从离线视图进行蒸馏，是否比仅使用 L1视图训练得到更好的 L1嵌入。我们改变教师模型的视图（FullReco 或 L1）及其容量。大型教师模型的宽度为 128，包含六个模块（第 3.1节），小型教师模型则采用与学生模型相同的宽度和深度，即浮点精度下宽度为 16、包含两个模块。学生模型架构、训练样本和资源预算保持不变。使用 L1 训练的教师模型读取前 128 个 L1 粒子。每个教师模型训练一次，并据此训练两个学生模型：一个使用主要的注意力量化器（KBI），另一个使用替代方案（KIF，附录 B）。表 7 给出了在包含 10 000 个事件的最小偏置样本上的结果。

对于大型教师模型，FullReco 和 L1 视图的已见标准模型保留率相差不超过 3 个百分点。在两种量化器配置下，FullReco 教师模型在未见标准模型和超出标准模型过程上的表现都更好，因此在测试泛化能力时，离线视图有所助益。每种配置仅由一个教师模型和一个学生模型表示，因此不应过度解读约 1 个百分点的差异。对于小型教师模型，结果排序则相反：标准模型上的差距最高可达 7 个百分点，超出标准模型过程上的差距为 3 至 9 个百分点。这可能是因为小型 FullReco 教师模型设定的目标较差，而该目标依赖于 L1 中不存在的信息。我们尚未直接检验这一解释。

采用相同教师模型和 EBOP 目标训练、读取 64 或 128 个粒子而非 32 个粒子的学生模型，使所有标准模型过程的中位保留率变化了 −4 至 +3 个百分点，但没有一致的提升；而且，并非所有 128 粒子学生模型都能在训练过程中达到目标。因此，我们始终采用实现成本最低的 32 粒子学生模型。

教师模型视图	教师模型规模	已见标准模型	未见标准模型	BSM	所有标准模型过程
KBI 注意力量化器					
FullReco	大型	89	**96**	**86**	**94**
L1	大型	**90**	95	82	93
FullReco	小	76	90	64	86
L1	小	83	93	73	90
*KIF*注意力量化器					
FullReco	大	**90**	**96**	**86**	**95**
L1	大	88	92	79	91
FullReco	小	78	91	70	86
L1	小	85	93	73	91

表7：由四个教师模型蒸馏得到的 32 粒子学生模型，在经校准的最小偏置接受率为 1% 时的过程保留率（%）；数值为对 10,000事件最小偏置样本进行 21 次划分后取平均所得。“所有标准模型过程”汇总了 52 个非最小偏差标准模型过程。

C.3 第二分割集

对标准模型过程进行了第二次随机划分，并用最小偏置作为未见组，使用各自的教师模型和学生模型，重复了整个流程。其中，25个已见类别中有14个也出现在主分割集中。表8比较了两个分割集在相同评估事件上的表现。

在最小偏置门控下，第二分割集重现了主分割集的结果。在1%的接受率下，各组的中位保留率均为81%至96%。已见组和未见组的顺序互换了，因为保留率最低的五个标准模型过程有四个在此分割集中属于未见组（$\Upsilon \to \ell\ell$，即两个Drell–Yan样本和三光子产生）。H_T无需训练，呈现出相同的逆转，因此这种排序源于各组的组成。与第3.6 节的基线方法相比，第4.1 节的结论不变。嵌入对每个组的保留率都高于H_T、自编码器、输入级分数和多重性。手工构建特征空间在超出标准模型过程上的表现最接近（79%，而嵌入为81%），在标准模型过程上的表现则较低。在第4.3节的带标签参考探测中，未见类别仍然保持局部可分。

两阶段筛选对分割集的依赖程度高于门控。它筛选出 0.7% 的最小偏置事件，以及高达 11% 的某个标准模型过程——矢量玻色子融合 希格斯玻色子 $\to \gamma\gamma$ （表 8）。与主分割集相比，它保留了更多长寿命样本，但剔除了大多数即时高质量双缪子事件；由此形成的簇，其长寿命样本纯度为 82%，即时双轻子纯度为 95%。因此，哪些超出标准模型信号看起来不同于所有已观测到的过程，取决于观测到了哪些标准模型过程。

	主分割集	第二分割集
最小偏置接受率为 *1%* 时的中位保留率（%）		
已见标准模型	89	96
未见标准模型	96	92
BSM	86	81
实测最小偏置接受率	0.99	1.00
两阶段筛选（%）		
通过两道门的最小偏置事件	0.33	0.73
最高标准模型保留率	3.3	11.0
长寿命样本	23.8	43.5
高质量双缪子样本	25.8	2.5
最纯长寿命簇	80	82
最纯即时双轻子簇	97	95

表8：主分割集和第二分割集中已见过程与未见过程的表现，均在相同的事件以及10,000事件最小偏置样本上进行评估。两阶段行的定义与表9相同。

D 附加诊断

基线实现。 18 个手工构建特征包括希格斯玻色子$_T$、多重性的对数，前四个p_T 值以及缺失的希格斯玻色子$_T$。此外，还包括五种粒子类型各自的多重性和标量p_T总和，以及带电强子最大的横向冲击参数。每个特征，或粒子输入基线的每个坐标，均根据最小偏置参考集进行标准化。粒子输入自编码器有 265 296 个浮点参数，层宽为

$$448 \to 256 \to 64 \to 16 \to 64 \to 256 \to 448.$$

该模型使用学生输入归一化和均方重构误差，在297,722个最小偏置事件上通过提前停止进行训练；这些事件来自未用于其校准或评估的文件。它是一种不受约束的稠密自编码器，灵感来自参考文献 [4,6]，且并未复现其中的触发器实现。

D.1 最小偏置门的变化

在包含 10000 个事例的最小偏置样本上，将门控分别设置为 $k = 5$、10、20 和 50 个近邻并重复测试，在 1% 接受率下，中位保留率的变化最多为 1.1 个百分点；测得的测试接受率为 0.97% 至 1.00%。

对于每个事例，取嵌入得分与希格斯玻色子$_T$ 对应的两个校准p值中的较小值，再像校准其他得分一样校准所得结果，即可将嵌入得分与希格斯玻色子$_T$ 结合起来。在 1% 接受率下，这种组合对已见标准模型、未见标准模型和超出标准模型过程的保留率分别为 89%、97% 和 86%（测得的接受率为 1.01%），与单独使用嵌入得分相比相差不到 1 个百分点。

一个具有以下层的自编码器 $18 \to 32 \to 8 \to 32 \to 18$ 在第 3.6 节的 18 个手工构建的事件特征上训练，并在相同的最小偏置事件上采用提前停止，能够以 1% 的接受率保留 50%、70% 和 49% 的已见标准模型、未见标准模型和超出标准模型过程。相同特征空间中的最近邻距离保留的比例更高（76%、89% 和 79%），因此，在事件级特征上，重构误差是较弱的离群值分数；第 4.1 节的粒子级输入也是如此。

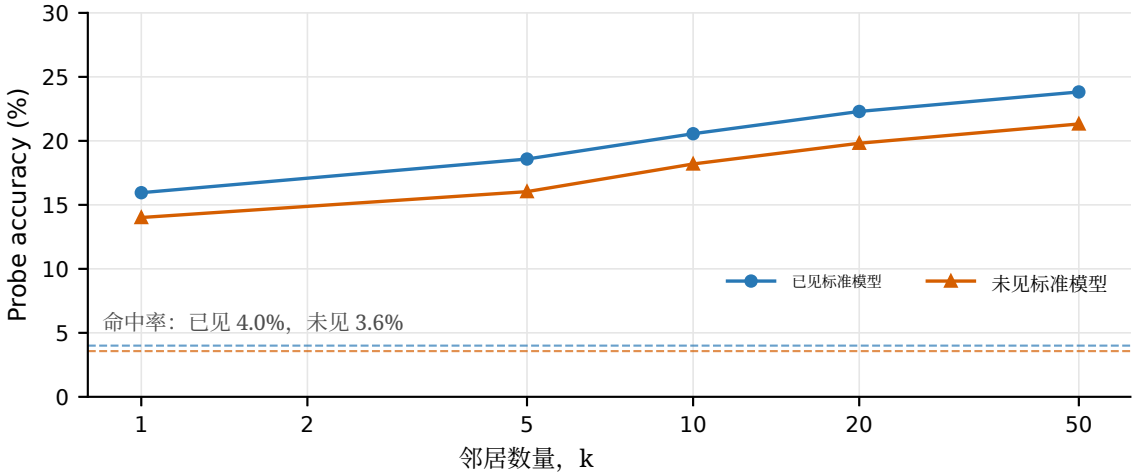


图 7：嵌入空间中基于已标注参考样本的最近邻准确率，分别在已见组和未见组内计算，并显示其随邻居数量k的变化。数据点表示对 21 次划分所得的平均准确率，每个类别包含 500 个参考事件和 500 个查询事件；虚线表示随机猜测水平。

D.2 嵌入的局部与全局结构

第 4.3 节的带标签参考探测表明，类别在局部上彼此分开，但并未在全局上分开（图 7）。其轮廓系数 [40] 略低于零， −0已见类别和未见类别的轮廓系数均为 .05，因此一个事例通常与最近的其他类别和自身类别距离相近。已见事件的前两个主成分解释了 28% 的方差。对每个过程取 200 个事件进行不使用标签的 HDBSCAN 聚类后，得到三个区域，将 87% 的事件归为噪声，且与真实标签的调整互信息为 0.06。因此，物理性质相似的过程在嵌入空间中相互重叠，并未形成独立簇。

D.3 嵌入中的事件取向

嵌入记录了事例沿束流轴的指向。对学生模型 128 维输出进行逻辑回归，可在 91.3% 的事例中恢复其 L1 候选对象的p_T加权平均值 η 的符号（随机猜测水平为 50%）；对教师模型中相同事例的嵌入进行回归，则可在 81.3% 的事例中恢复该符号。每种过程在这种反射下都具有对称性。尽管如此，取向在阶段 B 中仍然有用，因为它可以从一个批次的 128 个候选对象中排除约一半。对事例进行镜像处理（反转每个候选对象的 η 和d_z）会改变其嵌入；事例与其镜像之间的余弦相似度中位数为 0.40。

对每个事例及其镜像的归一化嵌入取平均可消除取向，此后每个过程族在映射上形成一个区域（图8，表9）。图6 使用这种对称化嵌入。所有触发器结果均使用训练完成后的嵌入，每个事例只需评估一次编码器。

以二分之一的概率独立镜像训练对中的每个视图，可消除编码取向的动机，且不增加推理成本。以这种方式训练的学生模型使用相同的教师模型、数据和随机种子，不再编码取向，其对某个事例及其镜像的嵌入也完全相同。在最小偏置门下，它在标准模型过程上的表现与主结果学生模型相当，但保留的超出标准模型事件更少，且经过两个门控后保留的即时高质量双μ子事件也少得多（表9）。因此，主要结果采用未进行数据增强的学生模型。

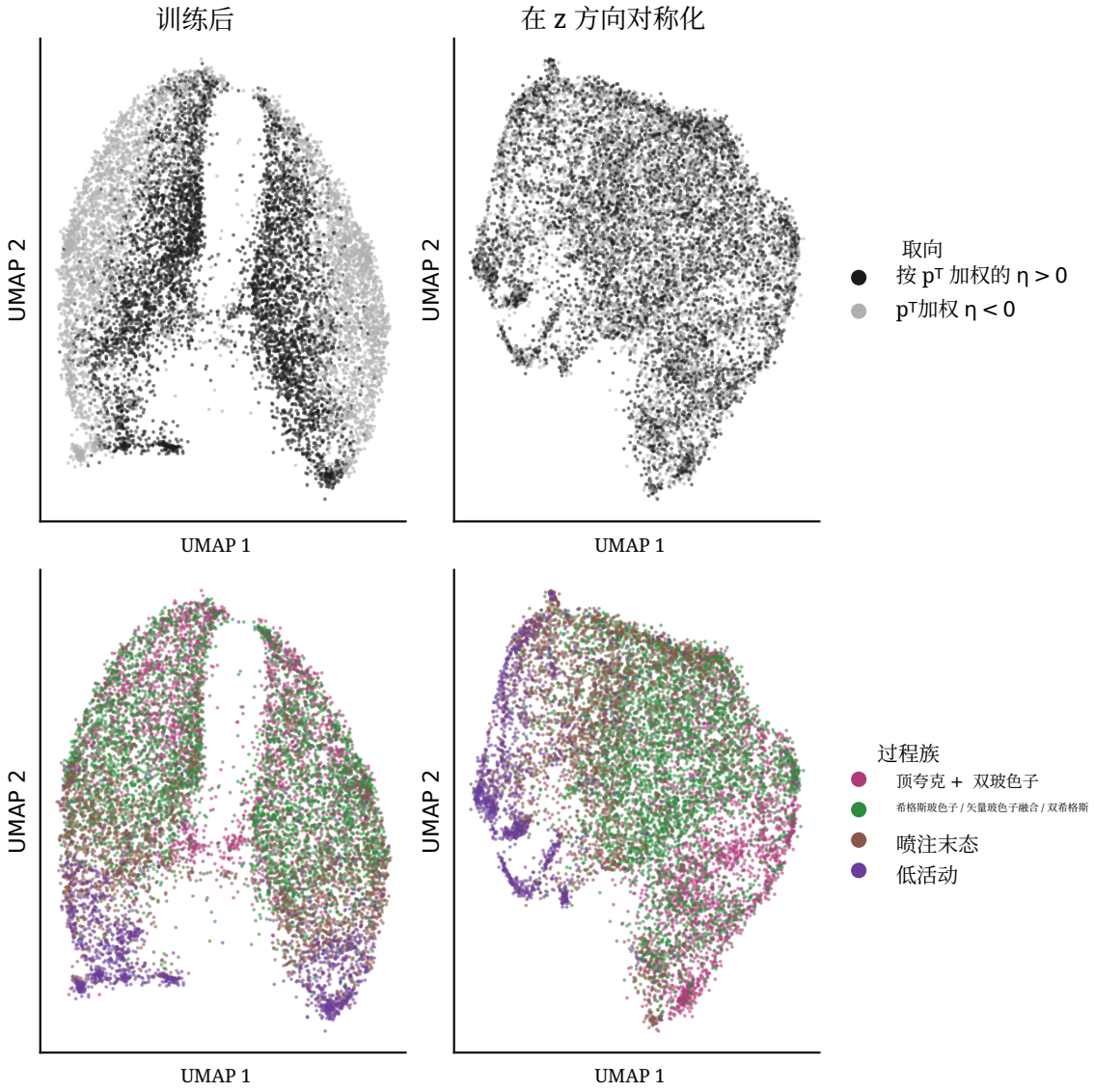


图 8：UMAP [38] 投影（30 个邻居，最小距离 0.1，在已见事例上拟合）展示了训练后的学生嵌入（左），以及将每个事例与其沿束流轴的镜像取平均后的学生嵌入（右）；图中包含图 6所示每个标准模型过程的 200 个事例。上排按取向着色，取向由其 L1 候选对象的 p_T 加权平均值的符号表示；下排则按过程族着色。训练后的映射按取向分裂为两个镜像叶瓣，每个叶瓣都包含所有过程族。对称化后，取向相互混合，每个过程族各自形成一个区域。

	训练后	平均后 训练	镜像于 训练
可解码的方向（%，随机猜测为 50%）	91.3	**50.5**	52.2
探针准确率（%）			
已见标准模型	20.6	**21.7**	21.1
未见标准模型	18.2	**19.8**	18.7
1% 最小偏置接受率下的中位保留率（%）			
已见标准模型	89	**91**	90
未见标准模型	**96**	**96**	**96**
BSM	86	**87**	84
两阶段筛选（%）			
通过两道门的最小偏置事件	0.33	0.30	**0.10**
最高标准模型保留率	**3.3**	5.0	7.0
长寿命样本	23.8	**25.0**	14.0
高质量双缪子样本	**25.8**	19.2	1.8
最纯长寿命簇	**80**	65	74
最纯即时双轻子簇	**97**	96	96

表9：对于按训练方式训练的主结果学生模型，每个事件均与其镜像取平均；同时给出使用随机镜像视图训练的学生模型（教师模型、数据和随机种子相同）在10,000事件最小偏置样本上的事件取向及其移除结果。长寿命样本的$c\tau = 10$ 或为100毫米。高质量双缪子样本是1 TeV下两个提示性隐藏谷$Z' \to \mu\mu$ 样本。簇行给出了在默认设置下，任一HDBSCAN簇中长寿命事件或提示性隐藏谷双轻子事件所占的最大比例。最高标准模型保留率是同时通过两个门控的任一标准模型过程所占的最大比例。

No.	超越标准模型基准	保留率（%）	簇大小	纯度	召回率	F_1
1	隐藏谷双轻子_Zp1000_piD5_缪子对	29	327	0.26	0.99	0.42
2	隐藏谷双轻子_Zp1000_piD2_缪子对	23	327	0.20	0.96	0.33

表10：发现候选项，即在通过两个门控后仍保留其300个流事件中至少10%的过程，以及与其最佳匹配的簇，其中 $F_1 \geq 0.3$。纯度、召回率和 F_1 的定义见附录 D.4。

D.4 聚类的评分与灵敏度

标签仅用于评估聚类。对于每个真实过程c，令n_c 表示经两个门控保留的该过程事件数。对于每个簇C，其纯度为$P = |c \cap C|/|C|$，召回率为$R = |c \cap C|/n_c$。将使下式最大化的簇计为该过程的对应簇

$$F_1 = \frac{2PR}{P+R}. \tag{D.1}$$

如果某个过程经过两个门控后至少保留了 10% 的事件，且其最佳簇的 $F_1 \geq 0.3$，则称其为发现候选项。这一描述性判据并不衡量统计显著性。两个提示产生的高质量双缪子样本符合候选项判据，其F_1 分别为 0.33 和 0.42，召回率均至少为 96%（表 10）。每个样本的纯度和 F_1 值都不高，因为多个样本具有相同的事例特征。

我们对主成分数（16、32 和 64）、最小簇大小（10、20 和 40）以及最小样本数（5、10 和 20）的所有组合，重复进行了聚类步骤。学生模型的 128 维输出位于一个 16 维仿射子空间中，因此使用 16 个或更多主成分即可保留几乎全部方差，主成分数几乎没有影响。两个提示产生的双缪子候选项在全部 27 种设置下均符合判据，而质量为 $c\tau = 100\text{mm}$ 的隐藏谷双缪子样本在 23 种设置下符合判据。在全部 27 种设置下，都会形成一个至少 90% 为提示产生的双轻子的簇；而至少 85% 为长寿命事例的簇仅在 3 种设置下形成。两个门控以及因此得到的图 5中的保留率均不受这些设置影响。

D.5 教师–学生配对对齐

教师模型和学生模型在每种标准模型过程的相同 1 000 个事例上进行评估，因此它们的嵌入按事例一一配对。同一事例的学生模型嵌入与教师模型嵌入之间的平均余弦相似度，在已见过程和未见过程中分别为 0.439 和 0.441。将每个学生事例与同一过程中的另一个教师事例重新配对后，这些平均值降至 0.232 和 0.224。此对照保留了过程身份，同时消除了事件身份，因此这一降幅衡量了阶段 B 所迁移的事件级信息。

我们还将每个学生事例配对的教师事例，与同一过程中的 128 个教师候选项进行排序比较，使得只有事件级信息能够发挥作用。在已见事例中，配对事例排名第一的比例为 39.0%；在未见事例中为 40.1%，而随机情况下的预期值为 0.8%。未见过程的匹配效果不逊于已见过程，这表明蒸馏在单个事例层面也具有泛化能力。

D.6 按进程保留

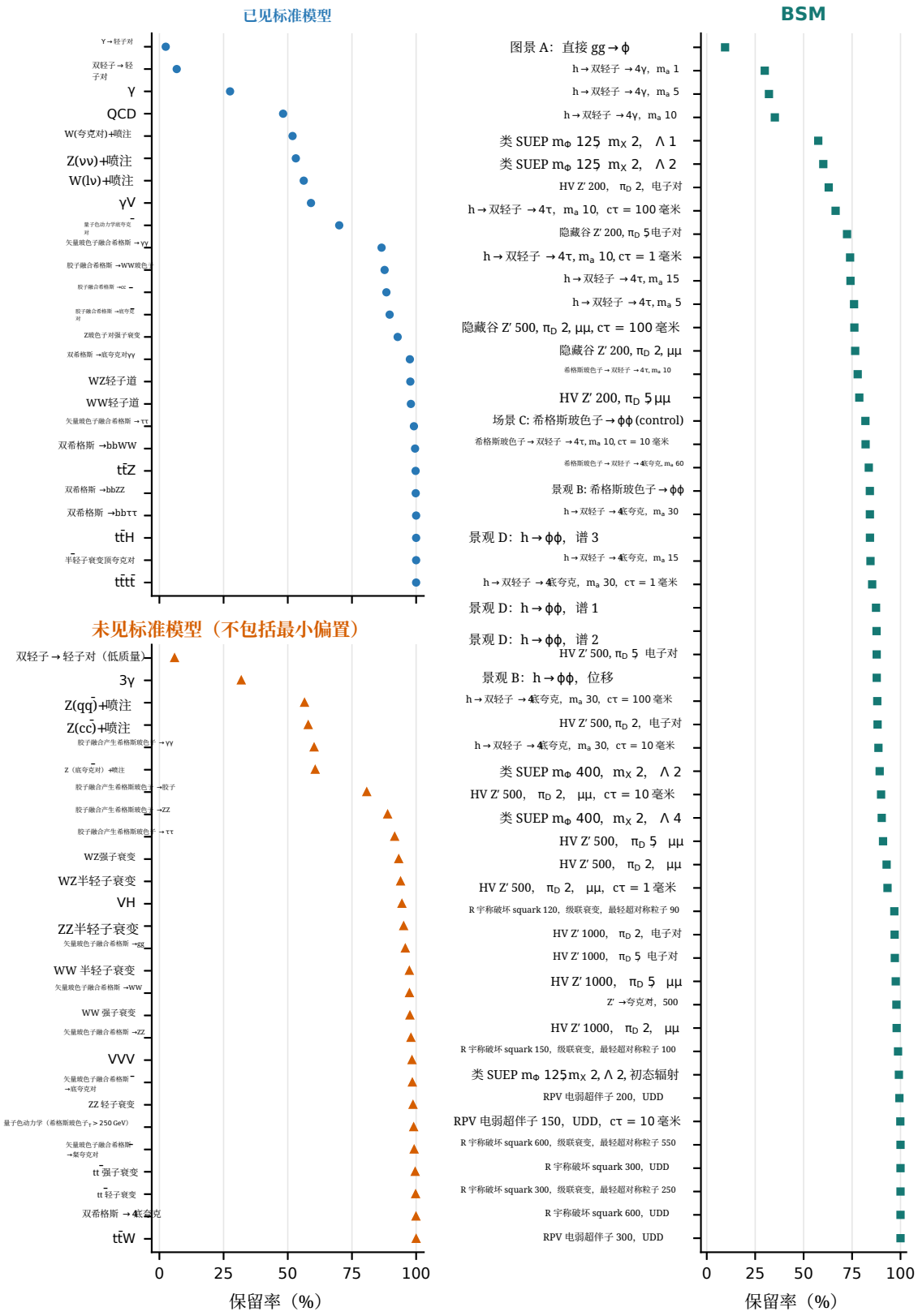


图 9：在目标最小偏置接受率为 1% 时，按进程保留率；该数值是对大型最小偏置样本的十个划分取平均所得。每组内的过程按保留率排序，名称中省略了生成器后缀。

参考文献

[1] ATLAS 合作组，《ATLAS 触发与数据采集系统第二阶段升级技术设计报告》，技术设计报告 CERN-LHCC-2017-020，ATLAS-TDR-029，CERN，日内瓦，doi:10.17181/CERN.2LBB.4IAL，https://cds.cern.ch/record/2285584(2017).[2] CMS 合作组，《CMS 一级触发器第二阶段升级》，技术设计报告 CERN-LHCC-2020-004，CMS-TDR-021，CERN，日内瓦，https://cds.cern.ch/record/2714892 (2020).[3] D. Bertolini、P. Harris、M. Low 和 N. Tran，《逐粒子堆积识别》，JHEP 10，059 (2014)，doi:10.1007/JHEP10(2014)059，arXiv:1407.6013.[4] E. Govorkova 等，《用于大型强子对撞机 40 MHz 实时无监督新物理检测的现场可编程门阵列自编码器》，Nature Mach. Intell. 4，154 (2022)，doi:10.1038/s42256-022-00441-3，arXiv:2108.03986.[5] N. Zipper，《第三次运行期间在 CMS 全局触发器测试机箱中测试用于异常检测的神经网络》，JINST 19，C03029 (2024)，doi:10.1088/1748-0221/19/03/C03029，代表 CMS 合作组，arXiv:2312.10009.[6] A. Gandrakota，《CMS 实验一级触发器上的实时异常检测》，PoS ICHEP2024，1025 (2025)，doi:10.22323/1.476.1025，代表 CMS 合作组，arXiv:2411.19506.[7] CMS 合作组，《大型强子对撞机 CMS 实验实时数据采集的无监督异常检测》，CMS 物理分析摘要 CMS-PAS-MLG-25-001，CERN，日内瓦，https://inspirehep.net/literature/3188808 (2026).[8] L. Laatu、C. Sun、A. Cox、A. Gandrakota、B. Maier、J. Ngadiuba、Z. Que、W. Luk、M. Spirop- ulu 和 A. Tapper，《用于现场可编程门阵列喷注标记的亚微秒级 Transformer》，FERMILAB-PUB-25-0779-CMS-LDRD (2025)，arXiv:2510.24784.[9] P. Harris、J. Krupa、M. Kagan、B. Maier 和 N. Woodward，《基于重新模拟的自监督学习：用于预训练物理基础模型》，Physical Review D 111，032010 (2025)，doi:10.1103/PhysRevD.111.032010，arXiv:2403.07066。

[10] J. Birk、A. Hallin 和 G. Kasieczka，*OmniJet-α*：首个面向粒子物理学的跨任务基础模型，Machine Learning: Science and Technology **5**，035031（2024年），DOI：10.1088/2632-2153/ad66ad，arXiv:2403.05618。

[11] T. Golling、L. Heinrich、M. Kagan、S. Klein、M. Leigh、M. Osadchy 和 J. A. Raine，集合上的掩码粒子建模：迈向高能物理自监督基础模型，Machine Learning: Science and Technology **5**，035074（2024年），DOI：10.1088/2632-2153/ad64a8，arXiv:2401.13537。

[12] V. Mikuni 和 B. Nachman，同时促进所有喷注物理任务的方法，Physical Review D **111**，054015（2025年），DOI：10.1103/PhysRevD.111.054015，arXiv:2502.14652。

[13] Y. Xian、C. H. Lampert、B. Schiele 和 Z. Akata，零样本学习——对优劣得失的全面评估，IEEE Transactions on Pattern Analysis and Machine Intelligence**41**(9)，2251（2019年），doi:10.1109/TPAMI.2018.2857768。

[14] E. A. Moreno、P. Ploner、R. Das、J. Ngadiuba、A. Gandrakota、B. Maier、T. Årrestad、M. Glowacki、S. Kuang、P. Harris、J. Duarte、D. Shih 等，COLLIDE-2V：用于低延迟机器学习的 7.5 亿个双视图大型强子对撞机事件数据集，发表于科学快速机器学习会议，苏黎世联邦理工学院，2025年9月3日；幻灯片见 CERN Next Generation Triggers 存储库，doi:10.17181/sbyss-as968，另于 CHEP 2026 发表，https://indico.cern.ch/event/1471803/contributions/6967164/ (2025)。[15] G. Hinton、O. Vinyals 和 J. Dean，《蒸馏神经网络中的知识》（2015），arXiv:1503.02531。[16] Y. Tian、D. Krishnan 和 P. Isola，《对比表示蒸馏》，载于国际学习表征会议（ICLR）（2020），arXiv:1910.10699。[17] J. Duarte 等，《用于粒子物理学的现场可编程门阵列深度神经网络快速推理》，JINST 13，P07027（2018），doi:10.1088/1748-0221/13/07/P07027，arXiv:1804.06913。[18] J.-F. Schulte 等，《hls4ml：用于可重构硬件深度学习加速的灵活开源平台》，ACM Trans. Reconfigurable Technol. Syst. 19，19（2026），doi:10.1145/3801979，arXiv:2512.01463。[19]C. Sun、Z. Que、T. K. Årrestad、V. Loncar、J. Ngadiuba、W. Luk 和 M. Spiropulu，《HGQ：用于现场可编程门阵列实时神经网络的高粒度量化》，载于 2026 年 ACM/SIGDA 国际现场可编程门阵列研讨会（FPGA '26）论文集，第 79 页，doi:10.1145/3748173.3779200（2026），arXiv:2405.00645。[20] C. Sun、Z. Que、V. Loncar、W. Luk 和 M. Spiropulu，《da4ml：用于现场可编程门阵列实时神经网络的分布式算术》，ACM Trans. Reconfigurable Technol. Syst. 19(1)，13（2026），doi:10.1145/3777387，arXiv:2507.04535。[21] C. Sun，《Alkaid：用于现场可编程门阵列低延迟静态数据流内核的编译器》，软件，版本 0.8.1，https://github.com/calad0i/alkaid（2026）。[22] J. Alwall、R. Frederix、S. Frixione、V. Hirschi、F. Maltoni、O. Mattelaer、H.-S. Shao、T. Stelzer、P. Torrielli 和 M. Zaro，《树级和次领头阶微分截面的自动计算及其与部分子簇射模拟的匹配》，JHEP 07，079（2014），doi:10.1007/JHEP07(2014)079，arXiv:1405.0301。[23] C. Bierlich 等，《PYTHIA 8.3 物理与使用综合指南》，SciPost Phys. Codebases 8（2022），doi:10.21468/SciPostPhysCodeb.8，arXiv:2203.11601。[24] J. de Favereau、C. Delaere、P. Demin、A. Giammanco、V. Lemaître、A. Mertens 和 M. Selvaggi，《DELPHES 3：用于通用对撞机实验快速模拟的模块化框架》，JHEP 02，057（2014），doi:10.1007/JHEP02(2014)057，arXiv:1307.6346。[25] R. T. D'Agnolo、M. Ettengruber 和 L.-T. Wang，《对撞机中的景观》，JHEP 09，256（2026），doi:10.1007/JHEP09(2026)256，arXiv:2512.18001。[26] A. Ryd 和 L. Skinnari，《高亮度大型强子对撞机的径迹触发器》，Ann. Rev. Nucl. Part. Sci. 70，171（2020），doi:10.1146/annurev-nucl-020420-093547，arXiv:2010.13557。[27] A. Vaswani 等，《注意力就是你所需要的一切》，载于神经信息处理系统进展第 30 卷（NeurIPS）（2017），arXiv:1706.03762。[28] S. Wang、B. Z. Li、M. Khabsa、H. Fang 和 H. Ma，《Linformer：具有线性复杂度的自注意力机制》（2020），arXiv:2006.04768。

[29] *J. Devlin*、*M.-W. Chang*、*K. Lee* 和 *K. Toutanova*，《*BERT*：用于语言理解的深度双向变换器预训练》，载于《*2019*年北美计算语言学协会分会会议论文集》（*NAACL-HLT*），第 *4171* 页，*doi:10.18653/v1/N19-1423*（*2019*年），*arXiv:1810.04805*。[30] *A. Dosovitskiy* 等，《一幅图像胜过 *16×16* 个词：用于大规模图像识别的 *Transformer*》，载于《国际学习表征会议》（*ICLR*）（*2021*年），*arXiv:2010.11929*。[31] *F. Schroff*、*D. Kalenichenko* 和 *J. Philbin*，《*FaceNet*：用于人脸识别和聚类的统一嵌入》，载于《*IEEE*计算机视觉与模式识别会议论文集》（*CVPR*），第 *815* 页，*doi:10.1109/CVPR.2015.7298682*（*2015*年），*arXiv:1503.03832*。[32] *I. Loshchilov* 和 *F. Hutter*，《解耦权重衰减正则化》，载于《国际学习表征会议》（*ICLR*）（*2019*年），*arXiv:1711.05101*。[33] *A. van den Oord*、*Y. Li* 和 *O. Vinyals*，《基于对比预测编码的表征学习》（*2018*年），*arXiv:1807.03748*。[34] *S. Ramaswamy*、*R. Rastogi* 和 *K. Shim*，《从大型数据集中挖掘离群点的高效算法》，载于《*2000*年*ACM SIGMOD*数据管理国际会议论文集》，第 *427* 页，*doi:10.1145/342009.335437*（*2000*年）。[35] *Y. Sun*、*Y. Ming*、*X. Zhu* 和 *Y. Li*，《利用深度近邻进行分布外检测》，载于《第*39*届国际机器学习会议论文集》（*ICML*），《机器学习研究论文集》第 *162* 卷，第 *20827* 页（*2022*年），*arXiv:2204.06507*。[36] *F. Pedregosa* 等，《*Scikit-learn*：*Python*中的机器学习》，《机器学习研究杂志》*12*，*2825*（*2011*年），*arXiv:1201.0490*。[37] *L. McInnes*、*J. Healy* 和 *S. Astels*，《*hdbscan*：基于层次密度的聚类》，《开放源代码软件杂志》*2(11)*，*205*（*2017*年），*doi:10.21105/joss.00205*。[38] *L. McInnes*、*J. Healy*、*N. Saul* 和 *L. Großberger*，《*UMAP*：均匀流形近似与投影》，《开放源代码软件杂志》*3(29)*，*861*（*2018*年），*doi:10.21105/joss.00861*。[39]*E. M. Metodiev*、*B. Nachman* 和 *J. Thaler*，《无标签分类：从高能物理中的混合样本学习》，*JHEP 10*，*174*（*2017*年），*doi:10.1007/JHEP10(2017)174*，*arXiv:1708.02949*。[40] *P. J. Rousseeuw*，《轮廓系数：用于解释和验证聚类分析的图形辅助工具》，《计算与应用数学杂志》*20*，*53*（*1987*年），*doi:10.1016/0377-0427(87)90125-7*。