arXiv:2303.17657v1 [physics.data-an] 30 Mar 2023

利用机器学习改进 CMS 粒子流算法的进展

Farouk Mokhtar[1],Joosep Pata[2],Javier Duarte[1],Eric Wulff[3],Maurizio Pierini[3] 和 Jean-Roch Vlimant[4]
（代表 CMS 合作组）[1]加州大学圣迭戈分校，美国加利福尼亚州拉霍亚，邮编 CA92093[2]NICPB，爱沙尼亚塔林 R¨avala pst 10，邮编 10143[3]欧洲核子研究组织（CERN），瑞士日内瓦 23，邮编 CH 1211[4]加州理工学院，美国加利福尼亚州帕萨迪纳，邮编 CA 91125

电子邮箱：fmokhtar@ucsd.edu, joosep.pata@cern.ch, jduarte@ucsd.edu

摘要。 粒子流（PF）算法根据径迹和量能器簇推断粒子，对 CMS 实验在 CERN 大型强子对撞机上的事件重建至关重要。鉴于计划中的第二阶段运行条件将带来更高的堆积和更精细的探测器粒度，该算法一直是重点开发对象。近年来，CMS 开始探索机器学习粒子流（MLPF）算法；该算法采用图神经网络执行粒子流重建，具有直接针对所关注的物理量进行优化、能够灵活适应新条件，以及天然适用于异构加速器部署等潜在优势。我们将介绍 CMS 在改进 MLPF 重建实现方面取得的进展；此次首次使用生成器／模拟级粒子信息作为优化目标。此举为改善与所关注物理量相关的探测器响应铺平了道路。我们将介绍基于模拟的训练目标、基于事例的损失项相关进展与研究、模型超参数调优的细节，以及相对于当前 PF 算法的物理验证；验证采用喷注和缺失横动量分辨率等高层次物理量。我们发现，首次使用生成器／模拟级粒子信息训练的 MLPF 算法，在粒子和喷注重建性能方面总体上与基线粒子流算法相当，为通过增加训练统计量和调整模型来提升物理性能奠定了基础。

1. 引言

粒子流（PF）重建是一种全局事件重建方法，它结合 CMS 中不同子探测器（例如径迹探测器以及电磁和强子量能器）的信息来重建稳定粒子 [1]。机器学习粒子流（MLPF）算法是一种图神经网络（GNN），通过监督式机器学习（ML）训练，以执行粒子流重建 [2, 3, 4]。与基线规则式粒子流算法一样，MLPF 的输入是径迹和量能器簇（见图 1），输出则是稳定的粒子流候选粒子。MLPF 的优势包括可部署于异构计算加速器（例如图形处理器），并能根据新的实验条件重新优化算法。

在这项 项工作中，我们总结了 CMS 中 MLPF 的最新进展，其中包括

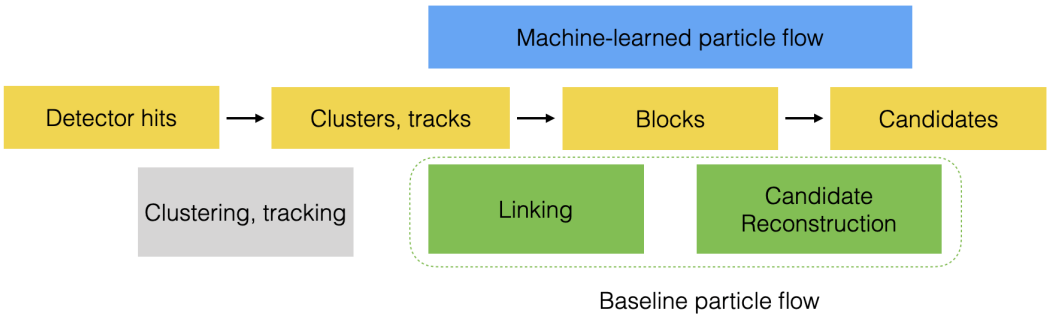


图 1. 所提出重建方案中 MLPF 的示意图 .

开发一种新的生成器/模拟层级训练目标，而不参考现有的粒子流算法。需要指出的是，这项工作属于当前和未来探测器重建算法开发这一更广泛背景的一部分，其中部分或全部算法将采用可训练、可微的机器学习 [5, 6, 7, 8, 9]。我们的实验结果表明，所得结果总体上与标准粒子流算法相当，在某些情况下甚至更好。完整结果见随附的探测器性能说明 [10]。

2. 粒子级训练目标定义

MLPF 的训练目标基于探测器模拟信息，旨在尽可能接近模拟从生成器接收的输入。MLPF 的真值粒子是 GEANT4 模拟树的根节点，包括模拟的粒子衰变及其与物质的相互作用，并通过以下真值定义算法确定。

真值算法以 GEANT4 模拟粒子树为输入，搜索其子粒子最早在径迹探测器或量能器[1]中留下可探测击中的粒子。确定这些衰变树根粒子后，我们首先通过移除具有重叠 GEANT4 模拟径迹标识符的粒子来消除重复计数。

现在已确定其衰变产物原则上会与探测器相互作用的根模拟粒子集合，接下来需要确定其中哪些粒子要作为粒子流粒子进行重建，以及采用何种粒度。模拟粒子按以下方式清理：

(i) 根据粒子流粒度合并粒子标签：所有带电强子归为同一带电强子类别，所有中性强子归为同一中性强子类别，依此类推。(ii) 几何上重叠、且仅在同一量能器簇中留下能量沉积的粒子无法分别重建，因此将其合并，并保留能量最高粒子的标签。(iii) 能量为 $p_{\mathrm{T}} < 1$吉电子伏特的电子或 μ 子会根据径迹和量能器中的能量沉积重新标记为带电强子或中性强子，以近似基线粒子流算法的行为。(iv) 为模拟基线粒子流算法的响应，径迹探测器接受度之外的粒子会根据其在相应量能器中的能量沉积，标记为 HF强子或 HF电磁

由此得到的模拟粒子集合称为 MLPF真值。粒子流与 MLPF 真值之间的比较见《探测器性能说明》 [10]的第 5–7 页。

[1] 我们使用 CMSSW 的 CaloParticle 和 TrackingParticle 模块遍历模拟 tr ee

3. 数据集

利用上述算法，我们生成用于 MLPF 真值验证和后续模型优化的数据集。我们使用官方 CMS 软件[2] 进行样本生成、模拟和基线粒子流重建，质心能量为 $\sqrt{s} = 14$太电子伏，运行条件对应于第三次运行。对于包含堆积的样本，我们采用平坦的堆积（PU）分布，其范围为 55–75，并按照泊松分布取值；堆积来自一个高统计量的最小偏置数据集[3]。训练和验证在表 1 所列全部样本的混合数据上进行。

物理过程	堆积配置	蒙特卡罗事件
顶夸克-反夸克对（顶夸克对）	均匀分布 55–75	100 千
量子色动力学 $\hat{}p_{\mathrm{T}} \in [15, 3000]$ 吉电子伏特	平坦分布 55-75	100 千
量子色动力学 $\hat{}p_{\mathrm{T}} \in [3000, 7000]$ 吉电子伏特	平坦分布 55–75	100 千
Z玻色子 $\rightarrow \tau\tau$ 全强子	均匀 55–75	100 千
单电子 均匀 $p_{\mathrm{T}} \in [1, 1000]$ 吉电子伏特	无堆积	10 k
单个 μ 对数均匀分布 $p_{\mathrm{T}} \in [0.1, 2000]$ 吉电子伏特	无堆积	10k
单个 π^0 均匀分布 $p_{\mathrm{T}} \in [0, 1000]$ 吉电子伏特	无堆积	10 k
单个 $\pi^{\pm}$ 均匀分布 $p_{\mathrm{T}} \in [0.7, 1000]$ 吉电子伏特	无堆积	10 k
单个 τ 平坦 $p_{\mathrm{T}} \in [1, 1000]$ 吉电子伏特	无堆积	10 k
单个 γ 平坦 $p_{\mathrm{T}} \in [1, 1000]$ 吉电子伏特	无堆积	10 k
单个质子平坦 $p_{\mathrm{T}} \in [0.7, 1000]$ 吉电子伏特	无堆积	10 k
单个中子，平坦分布 $p_{\mathrm{T}} \in [0.7, 1000]$ 吉电子伏特	无堆积	10 k

表 1. 用于优化 MLPF 模型的蒙特卡罗模拟样本。运动学量 $\hat{}p_{\mathrm{T}}$ 由出射生成器级部分子的标量和计算得出。

4. 事件损失扫描

由于粒子流的最终目标是实现全局事件重建，我们期望 MLPF 除了重建粒子流候选粒子外，还能高精度地重建事件级物理量，例如与喷注和缺失横动量相关的物理量。对于每个事件，使用反k_{T} 算法 [11, 12]，并将距离参数设为 0.4，从重建粒子或生成器级粒子中聚类得到喷注（AK4 喷注）。缺失横动量矢量 $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$通过对一个事件中所有粒子流候选粒子的横动量求矢量和再取负值得到，其大小记为 $p_{\mathrm{T}}^{\mathrm{miss}}$ [13]。逐粒子堆积效应识别（PUPPI）算法 [14, 15] 用于降低喷注和 $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$观测量 [13]对堆积的依赖性。

虽然基础的 MLPF 算法采用了逐粒子损失，但研究在损失函数中加入额外项是否能改善事件级重建，颇有意义。通常，将重建出的粒子聚类为喷注，再将所得的重建喷注与生成器级喷注进行比较，这种简单方法并不实用，因为目前尚无快速、可微且经过 GPU 优化的喷注聚类算法。我们测试了以下代理事件级损失项：

(i) 基线：仅使用基本的 MLPF 粒子分类和回归损失 (ii) 切片 Wasserstein 距离：在 $[\eta$,sin ϕ, cos $\phi]$ 空间中，计算以 p_{T} 加权的质量输运代价，以此衡量重建粒子集与目标粒子集之间的差异，并通过随机投影（切片）近似该度量 (iii) 生成器级喷注对数双曲余弦：假设重建粒子被聚类为与目标（MLPF 真值）粒子相同的喷注，计算有效的重建喷注 p_{T} 值，并将其与生成器级喷注 $[\eta,$ 值进行比较，

[2] CMSSW_12_3_0_pre6[3] /RelValMinBias_14太电子伏/CMSSW_12_2_0_pre2-122X_mcRun3_2021_realistic_v1_HighStat-v1/GEN-SIM

计算有效的重建喷注 p_{T} 值，并将其与生成器级喷注 p_{T} 值进行比较，使用对数双曲余弦损失

(iv) 缺失横动量：根据重建的粒子候选体计算重建的$p_{\mathrm{T}}^{\mathrm{miss}}$，并将其与目标$p_{\mathrm{T}}^{\mathrm{miss}}$进行比较，采用均方误差损失项

比较结果如图 2 所示。我们监测喷注响应分布的中位数和四分位距，并将其定义为两者的最优值均为零。我们发现，在训练过程中观察到的喷注级物理量并未因任何经过测试的损失函数而有所改善；其中基线表现最佳，即不在粒子级损失中添加额外的事件级损失项。

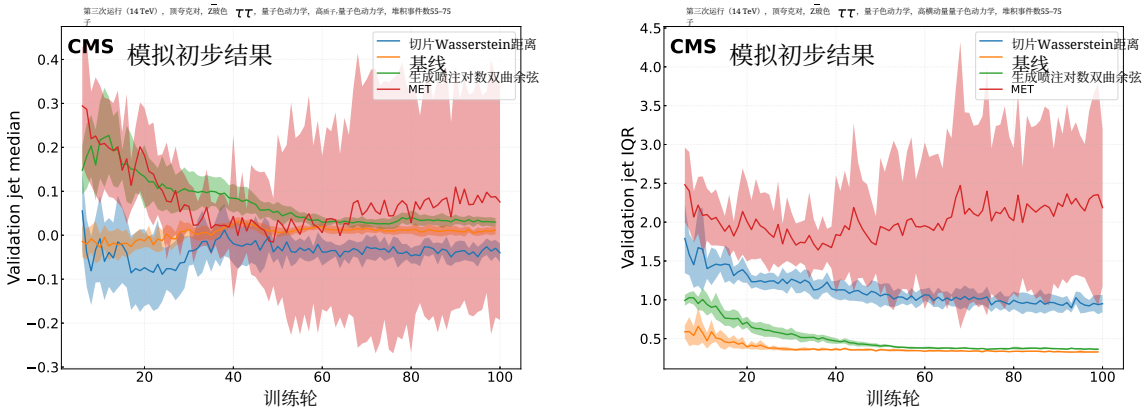


图 2. 采用不同事件级损失项时喷注响应中位数（左）和四分位距（右）的比较。实线表示 10 次配置相同的训练的均值，阴影区域表示这 10 次训练的均值上下各一个标准差。

5. CMSSW 中的事件级验证

MLPF 算法通过一个可选开关（默认关闭）集成到 CMS 软件框架（CMSSW）中，并使用 ONNXRUNTIME 进行推理。本次验证旨在真实集成环境中测试 MLPF，完全独立于机器学习训练数据格式和软件框架。我们通过在以下未用于训练的样本上进行验证，展示在 CMSSW 重建中启用 MLPF、取代标准粒子流 PF 的结果：横动量为 $\hat{}p_{\mathrm{T}} \in [15, 3000]$ 吉电子伏特的量子色动力学多喷注[4]（46 k 个事例）[5]，以及顶夸克-反夸克对（即顶夸克对，tt）（8 k 个事例）[6]。

以下结果是直接从 MINIAOD 事件数据中提取的事件级验证。我们还测试了 MLPF 在粒子级物理量上的性能，完整内容见《探测器性能说明》 [10]。

[4] 由强相互作用产生的喷注构成的标准模型事例背景称为量子色动力学（QCD）多喷注事例。[5] /RelValQCD_FlatPt_15_3000HS_14/CMSSW_12_3_0_pre6-PU_123X_mcRun3_2021_realistic_v11-v1/GEN-SIM-DIGI-RAW[6] /RelValTTbar_14TeV/CMSSW_12_3_0_pre6-PU_123X_mcRun3_2021_realistic_v11-v1/GEN-SIM-DIGI-RAW

5.1. 喷注

图 3 展示了粒子流和 MLPF 的生成器级与重建喷注分布的比较。粒子流与 MLPF 的喷注分布总体相容，凸显了纯机器学习方法 MLPF 的喷注重建性能；该方法是以生成器/模拟器粒子级目标进行训练的。

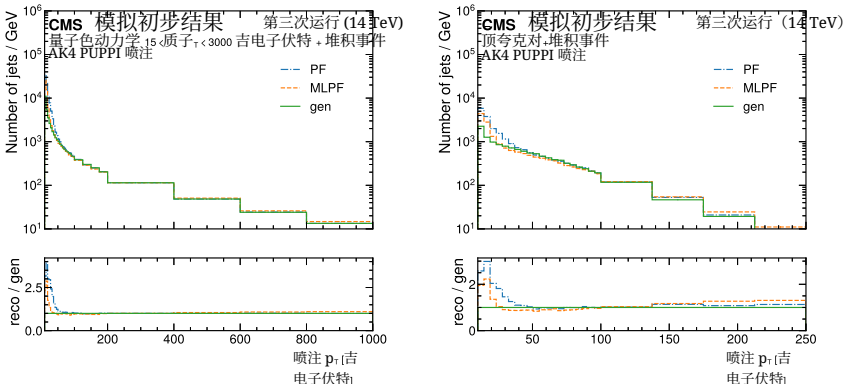


图 3. 在包含堆积的量子色动力学（左）和顶夸克对（右）样本中，使用 CMSSW 重建的 PUPPI 喷注 p_T 分布，以及粒子流和 MLPF 的重建结果与真实生成器级喷注分布的比较。

5.2. 缺失横动量

图 4 展示了 PF 和 MLPF 的生成器级 p_T^{miss}与重建后的 PUPPI p_T^{miss}分布的比较。特别是在量子色动力学样本中，我们观察到 PF 和 MLPF 都存在高-p_T 尾部，且 MLPF 中更为明显。目前，我们正通过改进数据样本并增加训练来减弱这一效应。

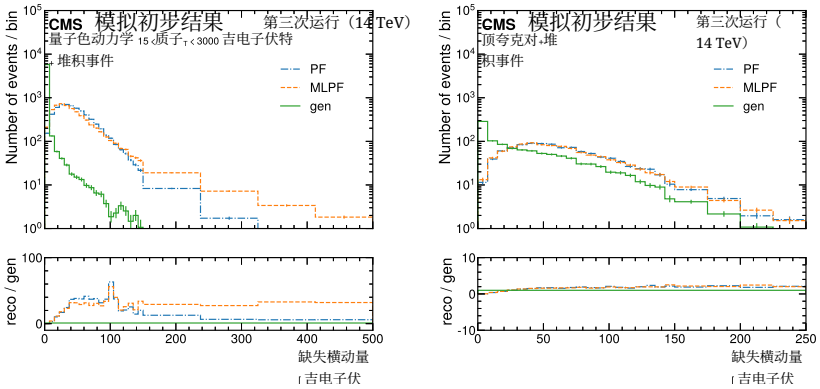


图 4. 在包含堆积的量子色动力学（左）和顶夸克对（右）样本中，使用 CMSSW 重建的 PUPPI p_T^{miss}分布，以及粒子流和 MLPF 的重建结果与真实生成器级 p_T^{miss}分布的比较。

6. 结论

我们定义了一种新的生成器/模拟器级机器学习粒子流（MLPF）训练目标，原则上可使基于 MLPF 的重建性能超越基准粒子流算法的重建性能。我们展示了 PF 与 MLPF 在粒子级别的比较，结果表明 MLPF 的性能与基于规则的粒子流（PF）相当。具体而言，我们发现中性强子 p_{T} 响应宽度改善了约 2 倍；对于能量为 $E > 10$吉电子伏特的量能器簇，MLPF 重建在假阳性率低于基线粒子流算法的情况下仍能达到完全效率。我们还在事件级别对性能进行了比较，直接在 CMS 软件框架（CMSSW）中进行；结果显示，两者在喷注方面的性能总体相当，而 $p_{\mathrm{T}}^{\mathrm{miss}}$性能受影响，MLPF 中的尾部略宽于 PF，我们计划在后续工作中解决这一问题。为了进一步改善事件级物理量的重建，我们尝试使用额外的事件级损失项；然而，我们发现这种情况下基线粒子级损失的表现最佳。

后续工作包括进一步完善训练数据集，以及改进机器学习（ML）建模与优化，从而进一步提升对复杂事件级变量的重建效果。探索将可解释人工智能技术应用于 MLPF 也很有价值，因为基于机器学习的算法普遍存在可解释性不足的问题。在此前的工作中，我们探索过这类技术，以深入了解模型的决策过程 [16]。如今我们已经有了定义明确的 MLPF 真值，重新开展这类研究很有意义。此外，鉴于新数据集，我们还计划在高性能计算资源上重新进行超参数调优 [17]。

7. 致谢

我们感谢 CMS 合作组的同事，尤其是粒子流、物理性能与数据集、离线与计算以及机器学习组的同事，特别感谢 Kenichi Hatakeyama、Lindsey Gray、Jan Kieseler、Danilo Piparo、Gregor Kasieczka、Salvatore Rappoccio、Kaori Maeshima、Kenneth Long 和 Juska Pekkanen 在本研究过程中提供的宝贵反馈。JP 获得了爱沙尼亚研究委员会 Mobilitas Pluss Grants 项目的 MOBTP187、PRG780 和 MOBTT86 资助。JD 和 FM 获得了美国能源部奖项 DE-SC0021187 和 DE-SC0021396（FAIR4HEP）以及美国国家科学基金会合作协议 OAC-2117997（A3D3）的资助。FM 还获得了加利福尼亚大学圣迭戈分校 HDSI 奖学金，以及通过美国国家科学基金会合作协议 OAC-1836650 提供的 IRIS-HEP 奖学金。EW 获得了 CoE RAISE 项目的资助，该项目获得了欧盟地平线 2020 研究与创新框架计划 H2020-INFRAEDI-2019-1 的资助，资助协议编号为 951733。图形处理器的使用部分由美国国家科学基金会项目 CNS-1730158、ACI-1540112、ACI-1541349、OAC-1826967、OAC-2112167、CNS-2100237、CNS-2120019、加利福尼亚大学校长办公室以及加利福尼亚大学圣迭戈分校加利福尼亚电信与信息技术研究院/高通研究院提供支持。感谢 CENIC 提供 100 Gbps 网络。

参考文献

[1] CMS 合作组。《利用CMS探测器进行粒子流重建和全局事件描述》。载于：*JINST*12.10 (2017)，P10003。 DOI: 10.1088/1748-0221/12/10/P10003。[2] Joosep Pata 等。《MLPF：利用图神经网络进行高效的机器学习粒子流重建》。载于：《欧洲物理学杂志*C*》81.5 (2021)，第381页。 DOI: 10.1140/epjc/s10052-021-09158-w。[3] Joosep Pata 等。《CMS中的粒子流重建机器学习》。载于：《物理学杂志：会议系列》2438.1 (2023)，第012100页。 DOI: 10.1088/1742-6596/2438/1/012100。

[4] CMS 合作组。CMS中的粒子流重建机器学习。CMS探测器性能说明 CMS-DP-2021-030。2021年。 URL: https : / / cds . cern . ch / record/2792320.[5] Bruno Alves、Felice Pantaleo 和 Marco Rovere。《CMS高粒度量能器探测器异构重建链中的聚类》。载于：《物理学杂志：会议系列》2438.1 (2023)，第012015页。 DOI: 10.1088/1742-6596/2438/1/012015.[6] Felice Pantaleo 和 Marco Rovere。《CMS高粒度量能器重建的迭代聚类框架》。载于：《物理学杂志：会议系列》2438.1 (2023)，第012096页。 DOI: 10.1088/ 1742-6596/2438/1/012096.[7] Saptaparna Bhattacharya 等。《CMS第二阶段高粒度量能器中基于图神经网络的端到端重建》。载于：《物理学杂志：会议系列》2438.1 (2023)，第012090页。 DOI: 10.1088/1742-6596/2438/1/012090.[8] Francesco Armando Di Bello 等。《迈向计算机视觉粒子流》。载于：《欧洲物理学杂志C》81.2 (2021)，第107页。 DOI: 10.1140/epjc/s10052-021-08897-0.[9] Francesco Armando Di Bello 等。《利用集合变换器和超图预测网络重建喷注中的粒子》。2022年12月。[10] CMS 合作组。利用机器学习改进CMS粒子流算法的进展。CMS探测器性能说明 CMS-DP-2022-061。2022年。 URL: http: //cds.cern.ch/record/2842375.[11] Matteo Cacciari、Gavin P. Salam 和 Gregory Soyez。《反-k_{T} 喷注聚类算法》。载于：《高能物理学杂志》04 (2008)，第063页。 DOI: 10.1088/1126-6708/2008/04/063.[12] Matteo Cacciari、Gavin P. Salam 和 Gregory Soyez。《FastJet用户手册》。载于：《欧洲物理学杂志C》72 (2012)，第1896页。 DOI: 10.1140/epjc/s10052-012-1896-2.[13] Albert M Sirunyan 等。《使用CMS探测器在 $\sqrt{s} = 13$太电子伏质子—质子碰撞中重建缺失横动量的性能》。载于：JINST 14 (2019)，P07004。 DOI: 10.1088/1748-0221/14/07/P07004.[14] Albert M Sirunyan 等。《13 TeV数据中的CMS堆积效应抑制》。载于：JINST 15 (2020)，P09018。 DOI: 10.1088/1748-0221/15/09/p09018.[15] Daniele Bertolini 等。《逐粒子堆积效应识别》。载于：《高能物理学杂志》10 (2014)，第059页。DOI: 10.1007/JHEP10(2014)059.[16] Faro uk Mokhtar 等。《阐释机器学习粒子流重建》。载于：第35届神经信息处理系统大会第四届机器学习与物理科学研讨会。2021年。[17] Eric Wulff、Maria Girone 和 Joosep Pata。《高性能计算系统上数据驱动人工智能模型的超参数优化》。载于：《物理学杂志：会议系列》2438.1 (2023)，第012092页。 DOI: 10.1088/1742-6596/2438/1/012092。