



本科生毕业论文

题目： CMS 实验中大动量 $H \rightarrow WW$
的标记技术开发及物理应用

姓 名： 付大为

学 号： 1800011105

院 系： 物理学院

专 业： 物理学

指导教师： 冒亚军 教授

李强 长聘副教授

二〇二二年五月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则一旦引起有碍作者著作权之问题，将可能承担法律责任。

摘要

2012 年在 LHC 的 ATLAS 实验和 CMS 实验上发现的希格斯玻色子（也被称为“上帝粒子”）填补了粒子物理标准模型的最后一块拼图。统一三大相互作用的标准模型成为了物理学目前最精确最成功的理论，但仍然存在着尚待解决的物理难题，包括：标准模型仍未统一引力、未发现的暗物质、无法完全用标准模型解释的正反物质不对称性。

在大型强子对撞机的后希格斯粒子时代有两个主要研究方向：一是对标准模型的精确检验（包括测量希格斯粒子的属性），二是搜寻 TeV 能标处可能存在的新物理（如扩展额外维模型预言的三玻色子共振态）。在即将开始的 CMS 实验的 RUN 3 阶段和 2027 年升级的高亮度 LHC 的未来背景下，粒子物理学家们将手握更多的统计量去尝试推动物理学前沿发展。

本文将着眼于大动量希格斯粒子到双 W 玻色子的衰变过程，这是因为：一，CMS 实验对大动量希格斯粒子分析还有很多空白等待填充，对大动量希格斯粒子的测量有利于在更高能量区域精确检验标准模型的高阶圈图修正，同时也是对高能标物理（有效理论的高阶算子等）的间接搜寻；二，大动量希格斯粒子会导致衰变末态形成合并喷注，具有独特的喷注子结构和全新相空间，利用深度学习开发的标记器能在这种场景下执行传统标记器难以执行的任务，是神经网络在物理领域的绝佳应用；三，可以通过此类场景搜寻其他可能的非标准模型 $X \rightarrow WW$ 共振态。在 2022 年美国 CDF 实验组的 W 玻色子质量反常的实验结果背景下，有利于补充更多 W 玻色子相关超出标准模型的证据或否决部分新物理模型，缩小对超出标准模型探索的范围，提高探索效率。

本文在对 WW 共振态的研究中，针对 WW 的全强子衰变和半轻衰变场景，开发了 CMS 实验中首个针对 $H \rightarrow WW$ 喷注的多分类标记器，并且利用了质量去相关 (Mass-Decorrelation) 技术，从而为研究大动量希格斯 $H \rightarrow WW$ 以及搜寻类希格斯的 $X \rightarrow WW$ 共振态提供了广阔前景。该标记器基于 ParticleNet 深度神经网络，创新性地使用点云表示粒子对象，并通过边卷积 (EdgeConv) 的图神经网络技术实现了网络中粒子的交换对称性，体现了对传统深度学习标记器的喷注图像表示和粒子列表表示的物理优越性，从而充分挖掘了深度神经网络的标记潜力。我们开发出的 $H \rightarrow WW$ 标记器比先前最佳的 DeepAK8 标记器在 $H \rightarrow WW \rightarrow 4q$ 道的标记性能提升了接近 50%，是一次卓有成效的喷注技术革新，经由作者在 CMS 国际合作组的 JMAR(喷注及丢失横动量算法及重建组) 官方会议上汇报，获得了广泛关注。同时，该技术也已经在 $H \rightarrow WW$ 等相关分析上已经得到了初步应用，展示了良好的应用前景。在未来的研究中，该标记器有望获得更多更好的性能改善，助推我们研究感兴趣的大动量 $H \rightarrow WW$ 衰变道，同时也迈出了向未来更大规模的通用喷注标记器的重要一步，有望大大提高 CMS 实验在 RUN 3 阶段和未来高亮度 LHC 上的测量精确性和有利于探索更多新奇的超出标准模型物理。

关键词：大动量希格斯粒子， $X \rightarrow WW$ 共振态，喷注标记技术，超出标准模型

Development and Physics Application of Tagging Technique for Boosted Higgs decaying to WW in CMS Experiment

Dawei Fu (Physics)

Directed by Prof.Yajun Mao and Prof.Qiang Li

ABSTRACT

The discovery of Higgs boson, namely "God Particle", by the ATLAS and CMS experiments on the LHC in 2012, completed the Standard Model of particle physics, a successful theory describing all the known fundamental particles and three kinds of interactions apart from gravitation. However, there still remain several key questions to be answered: combining the SM with gravity, explaining dark matter source, and understanding matter vs. anti-matter asymmetry.

During the post Higgs-discovery era on LHC, there are two main research motivations: One is for the precise measurement of Standard Model including measuring the properties of Higgs boson. The other is to search possible new physics model at TeV energy-scale like Tri-boson resonances predicted by extra-dimension model and extended gauge symmetry. Under the background of upcoming RUN 3 phase of CMS experiment and futural High-Luminosity LHC upgraded in 2027, particle physicists will try to promote the frontier of physics with more statistics.

This article will focus on the WW production process of boosted Higgs, with the reasons as follows: 1. There are lots of unexplored boosted Higgs analysis in CMS experiment, therefore, both the test on loop-level diagram correction of Standard Model in high pt region and the indirect search for new physics on high energy-scale (e.g., higher-order operator of effective theory) will benefit a lot from boosted Higgs precise measurement; 2. Boosted Higgs will result in merged jet in final states and thus have unique jet-substructure and new phase-space. In such scenario, our tagger developed by deep learning can execute the task where traditional taggers usually failed, which indicates a great application of deep learning technique in physics; 3. Exploring the similar scenario, we will be able to search possible BSM (Beyond-Standard-Model) $X \rightarrow WW$ resonances. Under the background of overweighted W boson result, this motivation will help to supplement more W-related BSM evidence or

veto some new physics model to shrink the BSM area waiting to be explored so as to increase exploring efficiency.

In the study of WW resonances in this article, we focused on the all-hadronic and semi-leptonic decays from WW and exploited the first multi-class tagger for $H \rightarrow WW$ jet in CMS experiment. Additionally, the tagger is designed in Mass-Decorrelation version to be used in non-Higgs WW resonances' scenery. Our tagger is based on ParticleNet, creatively using cloud points to represent particle objects. And also it implements permutation-invariance with EdgeConvolution technique in Graph-Neural-Network so that reflects the physics supremacy vs. the jet-image representation and the particle-list presentation in traditional taggers, which dug out the full potential in deep tagging. Our tagger performs about 50% better than the previous best tagger DeepAK8 on $H \rightarrow WW \rightarrow 4q$ tagging task, which indicates a great revolution in jet-tagging technique and has gained continuous attention from JMAR meeting in CMS experiment. Having been preliminarily applied the tagger on $H \rightarrow WW$ related analysis, the tagger is believed to be improved more in the future and help in $H \rightarrow WW$ analysis of our interest. More important, it's a great step towards futural lager classification-scale tagger which will be used in RUN 3 phase of CMS experiment and HL-LHC for more precise measurement and more fascinated BSM models.

KEY WORDS: Boosted Higgs, $X \rightarrow WW$ resonance, Jet-tagging technique, Beyond Standard-Model

目录

第一章 引言	1
1.1 标准模型	1
1.1.1 标准模型的粒子组成	2
1.1.2 标准模型的相互作用	3
1.2 超出标准模型的迹象	6
1.2.1 b 夸克衰变中的轻子普适性异常	6
1.2.2 μ 子 g-2 实验结果与标准模型的偏差	8
1.2.3 超重的 W 玻色子	10
1.3 LHC 上的 CMS 实验	11
1.3.1 大型强子对撞机 (LHC)	11
1.3.2 紧凑缪子螺线管实验 (CMS)	12
第二章 大动量希格斯粒子和 $X \rightarrow WW$ 共振态的物理和研究动机	15
2.1 希格斯粒子的产生和衰变	15
2.1.1 希格斯粒子的产生	15
2.1.2 希格斯粒子的衰变	15
2.2 大动量希格斯粒子的物理特性和研究动机	16
2.3 $X \rightarrow WW$ 的物理背景及搜寻动机	18
2.3.1 标准模型的 $H \rightarrow WW$	18
2.3.2 超标准模型的 $X \rightarrow WW$	19
第三章 CMS 实验的重建与标记技术介绍	23
3.1 重建事例时缓解顶点堆积的 PUPPI 算法	23
3.1.1 顶点堆积 (pile-up)	23
3.1.2 PUPPI 算法	23
3.2 重建喷注的 anti-kT 算法	27
3.3 喷注标记算法历史发展 ^[25]	29
3.3.1 基于理论的高级变量算法	29
3.3.2 基于机器学习的高级变量算法	31
3.3.3 基于深度学习的初级变量算法	32
第四章 用于喷注标记的 ParticleNet 深度神经网络	33

4.1 喷注表示方式	33
4.2 边卷积 (EdgeConv)	34
4.3 ParticleNet 网络架构	36
第五章 开发 H→WW 质量去相关多分类标记器	39
5.1 质量去相关技术	39
5.2 分类标签	40
5.2.1 信号分类标签	40
5.2.2 本底分类标签	41
5.3 数据集	41
5.3.1 训练集和验证集	41
5.3.2 测试集	42
5.4 标注器设置	42
5.4.1 预挑选条件	42
5.4.2 重加权设置	43
5.4.3 神经网络输入	43
5.4.4 神经网络输出	45
5.5 标记器在测试集上效果	45
5.6 在分析中的初步应用效果和前景	48
第六章 总结和展望	51
参考文献	53
附录 A 关键代码	57
A.1 训练样本的喷注标签部分代码	57
致谢	65
北京大学学位论文原创性声明和使用授权说明	69

第一章 引言

高能物理的主流理论框架是粒子物理标准模型，这是经过狄拉克、温伯格、费曼、朝永振一郎、汤川秀树等伟大物理学家一步步搭建起来，人类迄今为止最精确最普适用的理论模型，是物理学界的不朽杰作。但是，它仍然有很多不足之处，例如：没有把引力相互作用统一进来，无法解释暗物质之谜，无法完全解释宇宙中为什么正物质比反物质多那么多……，这些都是留待我们新一代物理学者去攻克的问题。而实验是理论的基础，高能物理实验，也就成了人类突破未知的前哨站，其中最具有代表性的是欧洲大型强子对撞机（LHC）上的一系列粒子对撞实验，紧凑缪子螺线管（CMS）实验便是其中最大的实验之一。

最近两年先后出现了 μ 子磁矩g-2实验结果与标准模型异常偏差^[1]，弱相互作用传播子W玻色子质量超出标准模型预言^[2]等显著冲击标准模型的实验结果，极大地震惊和鼓舞了高能物理学界乃至整个科学界对新物理的期待。

同时，大型强子对撞机的CMS实验一方面对 μ 子探测比其他实验有极大的优势，而且尚未完成大型强子对撞机第二轮运转时的对W质量的数据分析测量，这两点都为潜在的突破性发现提供了极佳条件。

作者参与了LHC上的CMS实验组，瞄准“上帝粒子”希格斯粒子的WW衰变过程，开发了CMS实验上首个对H→WW喷注的多分类标记器，同时也可用于XtoWW共振态的搜寻，有助于实现对大动量希格斯粒子的精确测量，同时搜寻可能存在的新物理迹象。

1.1 标准模型

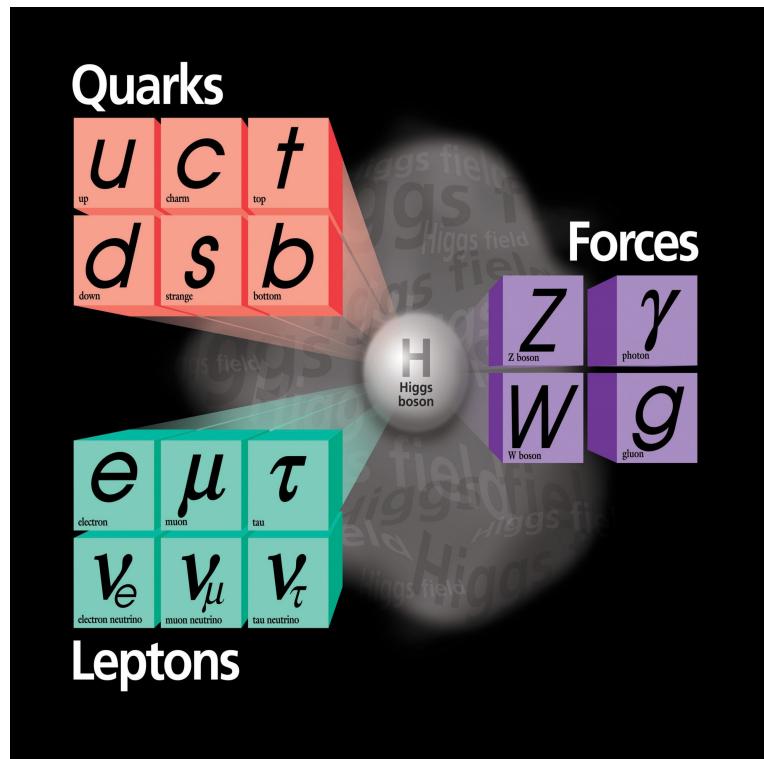
标准模型是当前粒子物理学（也称作高能物理）中得到广泛认可的理论框架，主要包括两方面的内容：第一，给出了构成我们大千世界的基本粒子，包括组成物质的费米子和负责传递各种相互作用的玻色子；第二，它统一了四种基本相互作用中的三种：电磁相互作用、弱相互作用和强相互作用，其中电磁相互作用和弱相互作用在标准模型中通过电弱统一理论进行描述，强相互作用通过量子色动力学进行描述。并且预言了赋予基本粒子质量的希格斯机制和希格斯粒子（也叫“上帝粒子”）。

标准模型的建立是在整个20世纪下半叶，通过世界各地许多科学家的工作分阶段发展起来的，并且在1970年代中期的夸克存在实验后逐渐确定整个框架。从那时起，陶子^[3]（1975）、顶夸克^[4]（1995）和希格斯玻色子^[5]（2012）的发现进一步证明了标准模型的正确性。此外，标准模型也非常准确地预测了弱中性电流以及W和Z玻色子的

各种特性。

1.1.1 标准模型的粒子组成

图 1.1 标准模型框架下的基本粒子



标准模型共有 61 种基本粒子（见表1.1），包含费米子及玻色子——费米子为拥有半奇数的自旋并遵守泡利不相容原理的粒子；玻色子则拥有整数自旋并且不遵守泡利不相容原理。简单来说，费米子就是组成物质的粒子而玻色子则负责传递各种作用力。

基本粒子中所有费米子自旋都为 $\frac{1}{2}$ ，包括三代夸克及其反粒子，三代轻子及其反粒子，正反粒子具有相同的质量和相反的电荷。基本粒子中 W, Z, 光子传播电弱相互作用，自旋都为 1；胶子传播强相互作用，自旋也为 1；希格斯粒子自旋为 0，通过 Yukawa 相互作用与粒子耦合并赋予它们质量。

值得一提的是，所有下型轻子（也就是中微子，包括电子中微子 ν_e , 缪子中微子 ν_μ , 陶子中微子 ν_τ ）电荷为 0 且无色荷，所以不能参与电磁相互作用和强相互作用，只能参与难以直接探测的弱相互作用，只能通过量能器的能量沉积得到击中信息，并且很多时候根本探测不到，所以在实验中也经常被称为“消失的中性粒子”。

表 1.1 基本粒子

名称	自旋类型	同位旋数 (上下型)	世代数	电荷种类 (正反粒子)	色荷种类	总计
夸克	半整数	2	3	2	3	36
轻子	半整数	2	3	2	/	12
胶子	整数	1	1	1	8	8
W	整数	1	1	2	/	2
Z	整数	1	1	1	/	1
光子	整数	1	1	1	/	1
希格斯	整数	1	1	1	/	1
总计						61

1.1.2 标准模型的相互作用

1.1.2.1 强相互作用

强相互作用由 $SU(3)_c$ 群的量子色动力学 (QCD) 描述，该群的生成元是 8 个线性无关矩阵 $T^a = \frac{\lambda^a}{2}$ ，其中 λ^a 是 Gell-Mann 矩阵， a 表示着 8 个色自由度，为了保证规范不变性，需要引入协变微分

$$(D_\mu)_{ij} = \partial_\mu \delta_{ij} - ig (T_a)_{ij} \mathcal{A}_\mu^a, \quad (1.1)$$

然后将拉氏量改写为

$$\mathcal{L}_{\text{QCD}} = \bar{\psi}_i (i\gamma^\mu (D_\mu)_{ij} - m \delta_{ij}) \psi_j - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}, \quad (1.2)$$

这里的 G 表示为规范不变胶子场强度张量

$$G_{\mu\nu}^a = \partial_\mu \mathcal{A}_\nu^a - \partial_\nu \mathcal{A}_\mu^a + g f^{abc} \mathcal{A}_\mu^b \mathcal{A}_\nu^c, \quad (1.3)$$

其中 $\mathcal{A}_\mu^a(x)$ 是胶子场。

根据量子场论的规则和相关的费曼图，上述理论产生了三种基本相互作用：一个夸克可以发射（或吸收）一个胶子，一个胶子可以发射（或吸收）一个胶子，以及两个胶子可以直接互动。这与 QED 形成对比，在 QED 中只发生第一种相互作用，因为光子没有电荷。

使用上述拉氏量的详细计算表明，介子中夸克与其反夸克之间的有效势包含一项与夸克和反夸克之间的距离成比例增加的项 ($\propto r$)，它表示粒子与其反粒子在远距离相互作用的某种“刚度”，类似于橡皮筋的熵弹性。这导致夸克被限制在强子内部，即介子和核子，具有特征半径 $R_c \sim 1 \text{ fm} (= 10^{-15} \text{ m})$ 。

在一对正反夸克体系的能标（距离尺度）下，它们之间的强相互作用势能可以用

如下式子表示

$$V(r) = -\frac{4\alpha_s(r)}{3r} + kr, \quad (1.4)$$

这里 $\alpha_s(r)$ 是 QCD 耦合常数，是距离 r (亦即能标) 的函数，会随着能量的增大 (距离的减小) 而减小，这也被叫做“渐近自由”。

当距离增加时，势能也会随之增加，同时产生吸引力，所以我们看到正反夸克对不可能无限远离 (因为势能不可能无限大)，而往往被束缚在有限的距离内，这种现象就被称为“夸克禁闭”。同样的，当出现单个夸克时，我们可以认为这就相当于正反夸克无限远离，所以粒子物理实验中是不会有夸克单独射出，而是会从夸克海中生成新的正反夸克对与单夸克结合形成束缚态，从而降低单夸克的强相互作用势能，这个过程中还往往伴随着强子化过程形成喷注 (高能粒子束流)。

1.1.2.2 电弱相互作用

在粒子物理学中，电弱相互作用是电磁相互作用与弱相互作用的统一描述，而这两种作用都属于自然界中四种已知基本相互作用。在粒子物理的 [GeV] 及以下能标中，电磁作用与弱作用存在很大的差异，然而在能标超过 W 的不变质量，即至少在 100[GeV] 的能标下，这两种作用力会统一成的电弱相互作用。

数学上是用一个 $SU(2) \otimes U(1)$ 的规范群统一描述电磁作用及弱作用。当中对应的零质量规范玻色子分别是三个来自 $SU(2)$ 弱同位旋的 W 玻色子 (W^+ , W_0 和 W) 以及一个来自 $U(1)$ 弱超荷的 B_0 玻色子。

在标准模型里 W 和 Z_0 玻色子和光子是经由 $SU(2) \otimes U(1)_Y$ 的电弱对称性自发对称破缺成 $U(1)_{em}$ 所产生的，此一过程称作希格斯机制 (见希格斯玻色子)。 $U(1)_Y$ 和 $U(1)_{em}$ 都属于 $U(1)$ 群，但两者不同； $U(1)_{em}$ 的生成元是电荷 $Q = Y/2 + T^3$ ，而其中 Y 是 $U(1)_Y$ (叫弱超荷) 的生成元， T^3 (弱同位旋的一个分量) 则是 $SU(2)$ 的其中一个生成元。

属于 $SU(2) \otimes U(1)_Y$ 自由费米子场的拉氏量如下

$$\mathcal{L} = i\bar{\psi}\gamma^\mu\partial_\mu\psi \quad (1.5)$$

对于场在 $SU(2) \otimes U(1)_Y$ 的规范变换

$$\psi \rightarrow \psi' = e^{igT^a\Lambda^a(x)} e^{\frac{i}{2}g'Y\zeta(x)} \psi, \quad (1.6)$$

为了满足场在 $SU(2) \otimes U(1)_Y$ 规范变换下的局域不变性，我们必须得引入协变微商 \mathcal{D}_μ 以代替 ∂_μ ：

$$\mathcal{D}_\mu = \partial_\mu - igT^a W_\mu^a - i\frac{g'}{2}YB_\mu, \quad (1.7)$$

其中 g 为 $SU(2)_L$ 作用的耦合常数, g' 为 $U(1)_Y$ 作用的耦合常数, $T^a = \frac{\sigma^a}{2}$ 是同位旋算符 (σ^a 是泡利矩阵), a 是欧式指标可取 1,2,3 (度量矩阵为 3 阶单位阵), μ, ν 是四维指标。(这里也可以看出关系式 $Q = T^3 + \frac{Y}{2}$, 粒子电荷等于同位旋第三分量加上超荷一半)

从而将带电弱相互作用的费米子场拉氏量改写为

$$\mathcal{L} = i\bar{\Psi}\gamma^\mu \mathcal{D}_\mu \Psi - \frac{1}{4}W_a^{\mu\nu}W_{\mu\nu}^a - \frac{1}{4}B^{\mu\nu}B_{\mu\nu}, \quad (1.8)$$

其中有三个带电的无质量玻色子 W^1, W^2, W^3 和一个无质量的中性玻色子 B , 由于对称性自发破缺, 将会出现我们实验中观测到的有质量正负 W 玻色子, 可表示为

$$W^\pm = \frac{1}{\sqrt{2}}(W^1 \mp iW^2) \quad (1.9)$$

相应地, Z 和 γ 光子可表示为

$$\begin{pmatrix} \gamma \\ Z \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} B \\ W_3 \end{pmatrix}, \quad (1.10)$$

其中 θ_W 是电弱混合角, 也被称为温伯格角。

1.1.2.3 引入质量耦合的 Higgs 机制

上面提到电弱统一中本来是无质量的 W^1, W^2 可通过自发对称破缺变为有质量的 W^\pm , 这就是通过希格斯机制实现的质量赋予。

因为电弱统一是 $SU(2) \otimes U(1)$ 理论, 讨论起来较为复杂, 所以我们从 $U(1)$ 群的希格斯机制开始讨论。

因为我们知道自旋为 0 质量为 m 的粒子的拉氏量为

$$\mathcal{L} = (\partial_\alpha \phi)^*(\partial^\alpha \phi) - m^2 \phi^* \phi - V(\phi^* \phi), \quad (1.11)$$

反过来, 如果一个拉氏量中有 ϕ 的二阶项, 那么它的系数就可以认为是场对应粒子的不变质量。

现在让我们考虑一个无质量标量场 ϕ , 并将该场的拉氏量写为

$$\mathcal{L} = \partial_\mu \bar{\phi} \partial^\mu \phi + \mu^2 \bar{\phi} \phi - \lambda (\bar{\phi} \phi)^2, \quad (1.12)$$

其中 λ, μ 都大于 0, 希格斯势为 $V = \lambda \bar{\psi} \psi)^2 - \mu^2 \bar{\psi} \psi$, 该拉氏量满足 $U(1)$ 的局部对称变换。显然这个类二次函数的最低值在 $\bar{\phi} \phi \frac{|v|^2}{2} = \frac{-\mu^2}{2\lambda}$ 处 (其中 $v = \frac{\mu}{\sqrt{\lambda}}$ 也就是量子场论中的真空态), 在 ϕ 二维复平面上考虑 V 的高度就可以形成一个旋转对称墨西哥草帽的形状。

我们现在将原始的 ϕ 写成 $\phi = \phi_1 + i\phi_2 = (\varphi_1 + v + i\varphi_2)/\sqrt{2}$, 其中 $v = \frac{\mu}{\sqrt{\lambda}}$ 当成真空态, $(\varphi_1 + i\varphi_2)/\sqrt{2}$ 才当成我们现在要考虑的 $U(1)$ 相互作用的粒子, 我们知道 $U(1)$ 相互作用的协变微分可以写为 $\mathcal{D}_\mu = \partial_\mu + iqA_\mu$, 把以上结果代入 $U(1)$ 相互作用的拉氏量有

$$\begin{aligned}\mathcal{L} = & \frac{1}{2}[(\partial_\alpha + iqA_\alpha)(\varphi_1 + v + i\varphi_2)]^*[(\partial^\alpha + iqA^\alpha)(\varphi_1 + v + i\varphi_2)] - \frac{1}{4}F_{\alpha\beta}F^{\alpha\beta} \\ & + \frac{\mu^2}{2}(\varphi_1 + v + i\varphi_2)^*(\varphi_1 + v + i\varphi_2) - \frac{\lambda}{4}[(\varphi_1 + v + i\varphi_2)^*(\varphi_1 + v + i\varphi_2)]^2,\end{aligned}\quad (1.13)$$

化简后有

$$\mathcal{L} = \frac{1}{2}(\partial_\alpha\varphi_1)(\partial^\alpha\varphi_1) - \mu^2\varphi_1^2 + \frac{1}{2}(\partial_\alpha\varphi_2)(\partial^\alpha\varphi_2) - \frac{1}{4}F_{\alpha\beta}F^{\alpha\beta} + \frac{1}{2}q^2v^2A_\alpha A^\alpha + \mathcal{L}_{int} \quad (1.14)$$

又因为 $U(1)$ 对称性要求拉氏量在 $\phi \rightarrow e^{i\theta(x)}\phi$ 旋转变换下不变, 于是我们在局部变换中设定相位 $\theta = -\arctan(\phi_2/\phi_1)$, 并令 $\phi \rightarrow \phi' = e^{i\theta}\phi = (\phi_1 \cos \theta - \phi_2 \sin \theta) + i(\phi_1 \sin \theta + \phi_2 \cos \theta)$ 即可得到

$$\mathcal{L} = \frac{1}{2}(\partial_\alpha\varphi_1)(\partial^\alpha\varphi_1) - \mu^2\varphi_1^2 - \frac{1}{4}F_{\alpha\beta}F^{\alpha\beta} + \frac{1}{2}q^2v^2A_\alpha A^\alpha + \mathcal{L}_{int}, \quad (1.15)$$

其中 $F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha$

现在我们得到, φ_1 是赋予质量的希格斯粒子场, 其质量为 $\sqrt{2}\mu$, $U(1)$ 规范矢量场 A_α 的质量为 $|q|\mu/\sqrt{\lambda}$, 为对应规范玻色子的质量。

对于 $SU(2) \otimes U(1)$ 群, 希格斯场从 φ_1 变为了复值的

$$\phi(x) = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \quad (1.16)$$

通过对称性自发破缺, 产生了三个带质量的玻色子 W^\pm, Z 和一个无质量玻色子 γ 。

1.2 超出标准模型的迹象

1.2.1 b 夸克衰变中的轻子普适性异常

在标准模型中, 三代带电轻子都具有相同的电弱相互作用耦合强度, 这也是轻子风味普适性 (LFU) 的一个显著特征。但是, 轻子风味普适性 (LFU) 只是标准模型 (SM) 中特有的一种对称性, 这在超出标准模型的理论中可能不成立。

在过去的几年, LHC 上的 LHCb 实验合作组探测了一些以味道改变中性流 (FCNC) 作为传播子的稀有衰变过程^[6], 在这样的过程中来自标准模型的贡献会被压低, 以此探测轻子风味普适性可能出现的异常从而搜寻超出标准模型的物理。2021 年 3 月公布的结果^[6]表明, 来自 LHCb 的结果与之前来自其他 b 夸克工厂的实验结果, 都暗示了可能的的超出标准模型迹象。

在测量实验中，研究人员们测量了 B 介子的衰变到 μ 子与衰变到电子的截面之比，为了简化，定义为 R_K

$$R_K \equiv BR(B^+ \rightarrow K^+\mu^+\mu^-)/BR(B^+ \rightarrow K) \quad (1.17)$$

这里 B^+ 的价夸克是 $u\bar{b}$, K^+ 的价夸克是 $u\bar{s}$, 所以这个过程本质上是 $b \rightarrow s\ell^+\ell^-$ 。在这样的过程中检验轻子风味异常的好处有：充分压低标准模型贡献，同时能精确检验理论预测。在轻子风味普适性的假设下， R_K 理论预言为 $R_K^{th} = 0.997 \pm 0.011$ ，因此 R_K 的测量结果与 1 的任何显著偏离都意味着有超出标准模型的物理。

LHCb 实验的挑战在于，虽然电子和 μ 子以相同的方式参与弱电相互作用，但电子小得多的质量意味着它与探测器材料的相互作用比 μ 子多得多。例如，电子在穿 LHCb 探测器时会辐射出大量的轫致辐射光子，与 μ 子相比，这会降低电子的重建效率和信号分辨率。处理这种效应的关键是通过 $J/\Psi \rightarrow e^+e^-$ 和 $J/\Psi \rightarrow \mu^+\mu^-$ 衰变过程（已知它们具有相同的衰变概率）可用于校准和测试电子重建效率。同时， J/Ψ 的高精度测试与轻子味道普适性兼容，这为实验分析提供了有力的交叉检验。在这次对 R_K 的更高精度与更大统计量测量中，LHCb 合作组发现了更大显著度的轻子风味异常。

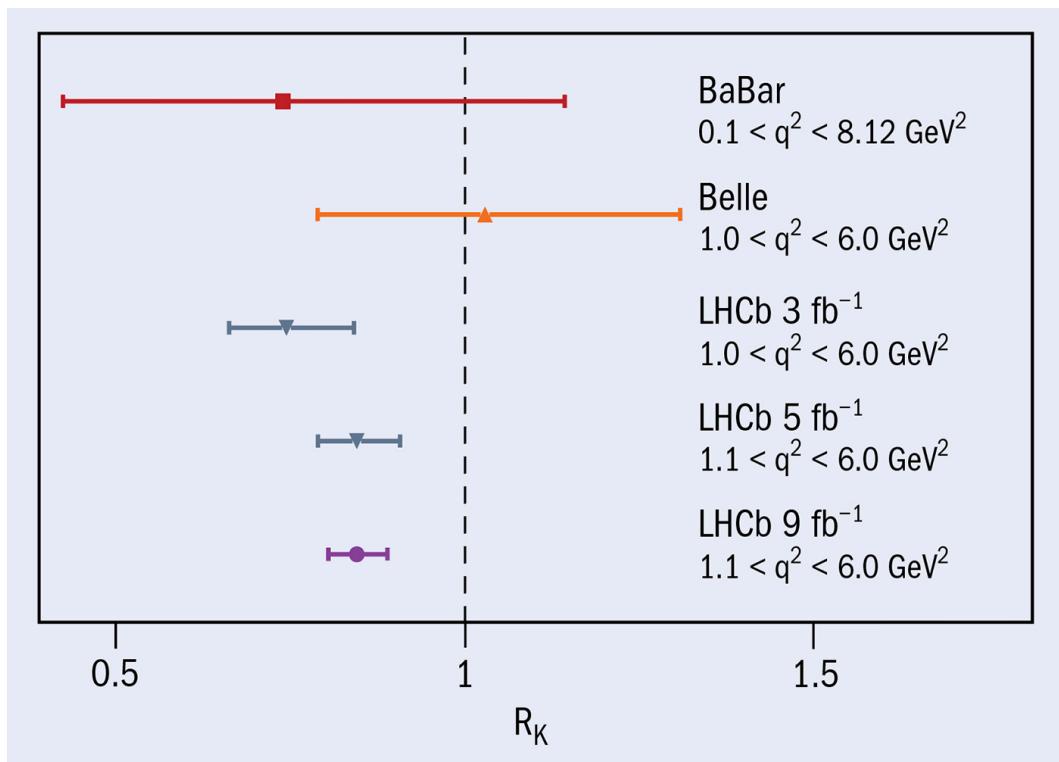


图 1.2 R_K 测量结果的比较^[7]：除了 LHCb 的结果，还展示了 BaBar 和 Belle 合作组的 $B^+ \rightarrow K^+\ell^+\ell^-$ 和 $B^0 \rightarrow K_S^0\ell^+\ell^-$ 测量的联合结果

此前 LHCb 分别在 2019 年^[8]和 2017^[9]年对 R_K 和 R_K^* （基于 $B^0 \rightarrow K^*\ell^+\ell^-$ 衰变过程）的测量已经显示了偏离 1 的迹象。而在这次对 R_K 的最新分析使用了大型强子对撞机

(LHC) 运行周期 RUN 1 和 RUN 2 阶段中收集的完整数据。由于数据集增加了一倍，因此这次与之前的测量相比，精度有了显著提高（如图1.2所示），测量得到的 R_K 比率与标准模型预测的偏离三个标准差（如图1.2所示）。这是第一次在任何单个 B 介子衰变中看到轻子风味普适性的异常高于 3σ ，对应测量值为 $R_K = 0.846^{+0.042}_{-0.039} (\text{stat.})^{+0.013}_{-0.012} (\text{sys.})$ 。

LHCb 实验很好地阐明了这些衰变中可能存在的新物理。现在一系列使用完整 Run 1 和 Run 2 数据集的与 $b \rightarrow s\ell^+\ell^-$ 相关测量正在进行中。在 LHC 第二次长时间关闭期间，对探测器的重大升级将在 Run 3 及以后的阶段提供测量精度上一个阶跃式的变化。

尽管现阶段下结论还为时过早，但轻子风味普适性的这种偏差与过去十年中在 $b \rightarrow s\ell^+\ell^-$ 和类似过程中表现出来的异常模式是一致的。特别是，这个被数据加强的 R_K 异常可以与这个衰变过程中的其他测量结果（包括角不对称性和衰减率）一起用来作为支持新物理的证据。这些都等待未来进一步的分析。

1.2.2 μ 子 g-2 实验结果与标准模型的偏差

μ 子是一种类似于电子的基本粒子，和电子一样带有一个单位负电荷、自旋为 $1/2$ ，但具有更大的质量， μ 子的质量大约是电子的 200 倍。 μ 子与同属于轻子的电子和 τ 子具有相似的性质，人们至今未发现轻子具有任何内部结构。

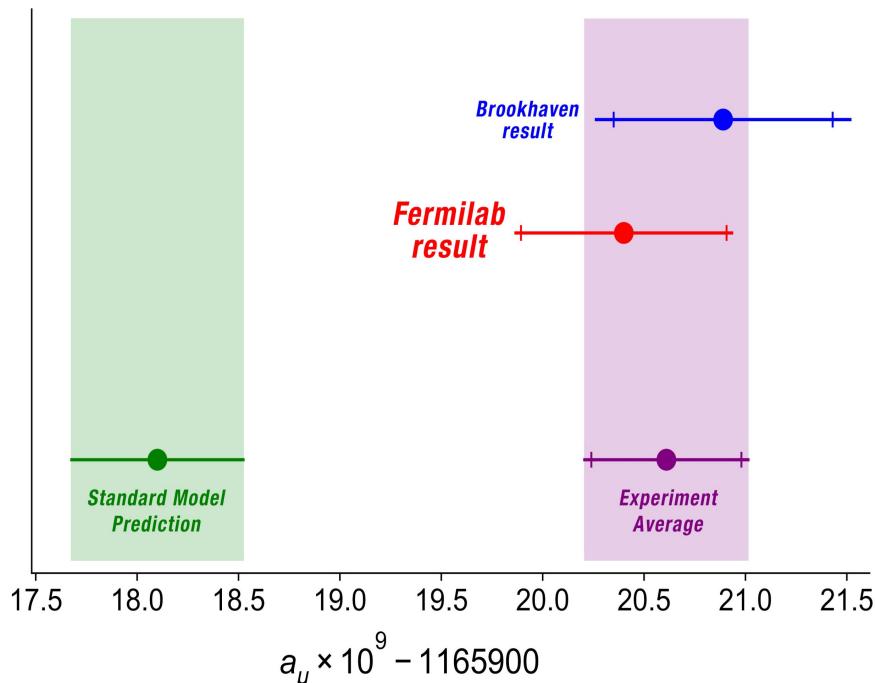
像电子一样， μ 子的行为就好像它们有一个微小的内部磁铁。在强磁场中， μ 子磁铁的方向会进动或摆动，就像陀螺或陀螺仪的轴一样。内部磁铁的强度决定了 μ 子在外部磁场中进动的速率，我们用称为“g 因子”的参数表示这块“磁铁”的强度和旋转速度。这个数字可以用量子场论进行超高精度计算。对于标准模型中的费米子，普通的 g 的物理定义是

$$\mu = g \frac{e}{2m} \mathbf{S} \quad (1.18)$$

其中 e 是单位电荷大小， m 是基本粒子不变质量， \mathbf{S} 是自旋， μ 是磁矩。按照狄拉克方程的计算，基本费米子的 $g=2$ 。

根据量子场论的计算（由于真空态和量子涨落），对于 μ 子， g 的值略大于 2，我们主要关心 g 和 2 的差异，因此得名“g-2”实验。这种与 2 的差异是由量子场论的高阶贡献引起的。在高精度测量 g-2 并将其值与理论预测进行比较时，物理学家将发现实验是否与理论相符。任何超过误差允许的偏差都会暗示着自然界中存在尚未发现的亚原子粒子。

2021 年，美国费米国家实验室的 Muon g-2 实验备受期待的首个结果表明^[1]，在前所未有的精确度测量下， μ 子的 g 因子与标准模型计算结果存在较大偏差。几十年来，g-2 的实验和理论差异就已经被测量过，这个里程碑式的结果，更是进一步确认了这个差异。

图 1.3 g 因子的标准模型理论值与实验的偏差^[10]

当 μ 子在费米实验室的 Muon g-2 磁铁中绕圈旋转时，它们同时会与与量子涨落不断生成或湮灭的亚原子粒子相互作用。与这些短寿命粒子的相互作用会影响 g 因子的值，导致 μ 子非常轻微的进动加速或减速。标准模型极其精确地预测了这种所谓的异常磁矩。但是，如果量子涨落中包含标准模型未考虑的额外相互作用力或粒子，那将进一步影响 μ 子的 g 因子大小。

公认的 μ 子 g 因子理论值为： $g=2.00233183620(86)$ 。而费米实验室的 Muon g-2 实验宣布的新实验结果为： $g=2.00233184122(82)$ 。（见上图1.3）

费米实验室和布鲁克海文的联合结果显示，实验测量与理论的差异显著度为 4.2σ （标准偏差），尽管略低于公认具有说服力的新物理证据所需的 5σ 阈，但 4.2σ 表明这次与标准模型的冲突结果是统计涨落的可能性约为四万分之一。

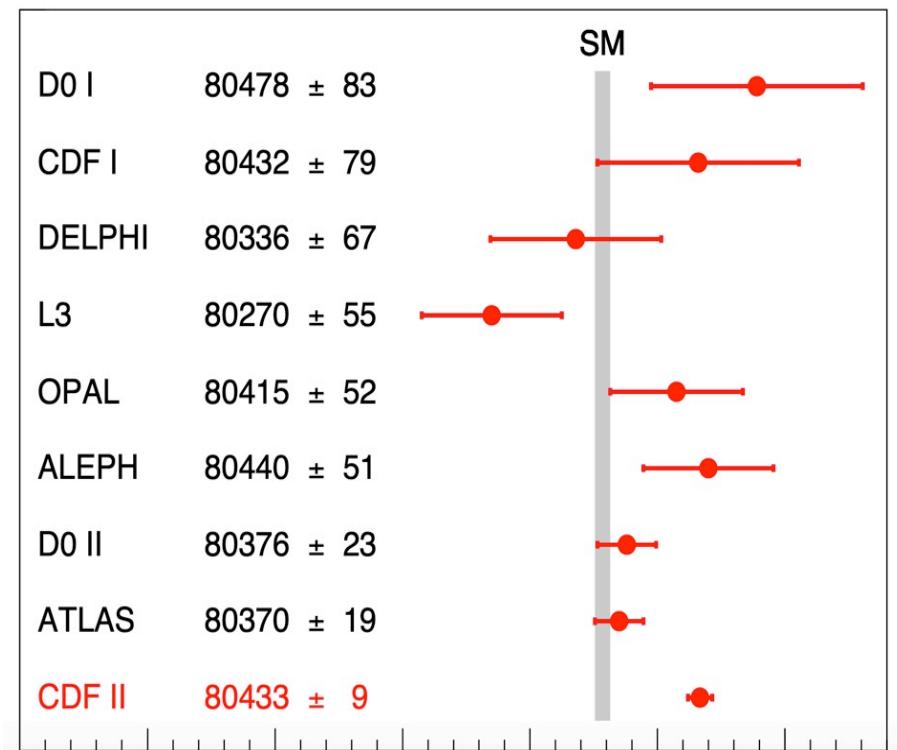
在运行的第一年，即 2018 年，费米实验室收集的数据就比之前所有 μ 子 g 因子实验的总和还要多。目前已完成对第一次运行中超过 80 亿个 μ 子的运动的分析。正在进行第二次和第三次实验的数据分析，第四次正在进行中，第五次正在计划中。到目前为止，只分析了不到 6% 的实验最终将收集的数据。结合所有五次运行的结果， μ 子的 g 因子将得到更加精确地测量，从而更确定地揭示新物理是否隐藏在量子涨落中。

1.2.3 超重的 W 玻色子

W 玻色子是弱相互作用的媒介粒子。它参与太阳发光和粒子衰变的反应过程。标准模型给出的 W 质量理论计算值为 $M_W = (80357 \pm 6)\text{MeV}$ 。该值基于复杂的标准模型计算，通过将 W 玻色子的质量与另外两个粒子的质量的测量联系起来：顶夸克（于 1995 年在费米实验室的 Tevatron 对撞机中发现^[4]），希格斯玻色子（2012 年在欧洲核子研究中心的大型强子对撞机上发现^[5]）。

过去 40 年来，许多对撞机实验都对 W 玻色子质量进行了测量，这些测量都是具有挑战性和十分复杂的。在 2022 年 4 月公布的美国费米国家实验室的 W 质量测量则^[2]达到了更高的精度——经过 10 年的仔细分析和审查，美国能源部费米国家加速器实验室的科学家宣布，他们已经实现了迄今为止对 W 玻色子质量的最精确测量。

图 1.4 不同实验对 W 质量测量结果的比较^[11]



利用费米实验室的 Tevatron 对撞机产生的高能粒子碰撞和对撞机探测器 (CDF) 收集的数据，研究人员收集了 1985 年至 2011 年间包含 W 玻色子的大量数据，它基于对 420 万个 W 玻色子候选者的观察，大约是 2012 年发布的合作分析中使用的数量的四倍。并且花了十年的时间来完成所有的细节和必要的检查，科学家们现在已经以 0.01% 的精度确定了粒子的质量（是之前最佳测量精度的两倍），得到了迄今为止精度最可靠的测量结果： $M_W = (80433.5 \pm 9.4)\text{MeV}$ 。（见上图1.4）

通过合成以上两个数据的独立不确定度，我们可以得到测量值和标准模型预期值

之间的差异存在 7σ 的偏差。这显示出了标准模型框架下，理论和实验的值产生了聚大冲突。如果无潜在错误，此测量表明可能需要改进标准模型计算或扩展模型。

科学界对这个结果也是众说纷纭——费米实验室副主任 Joe Lykken 指出：“虽然这是一个有趣的结果，但测量结果需要通过另一个实验来确认，然后才能完全解释。”德克萨斯农工大学 CDF 联合发言人 David Toback 补充道：“如果实验值和预期值之间的差异是由于某种新的粒子或亚原子相互作用造成的，这是一种可能性，那么它很有可能在未来的实验中被发现。”

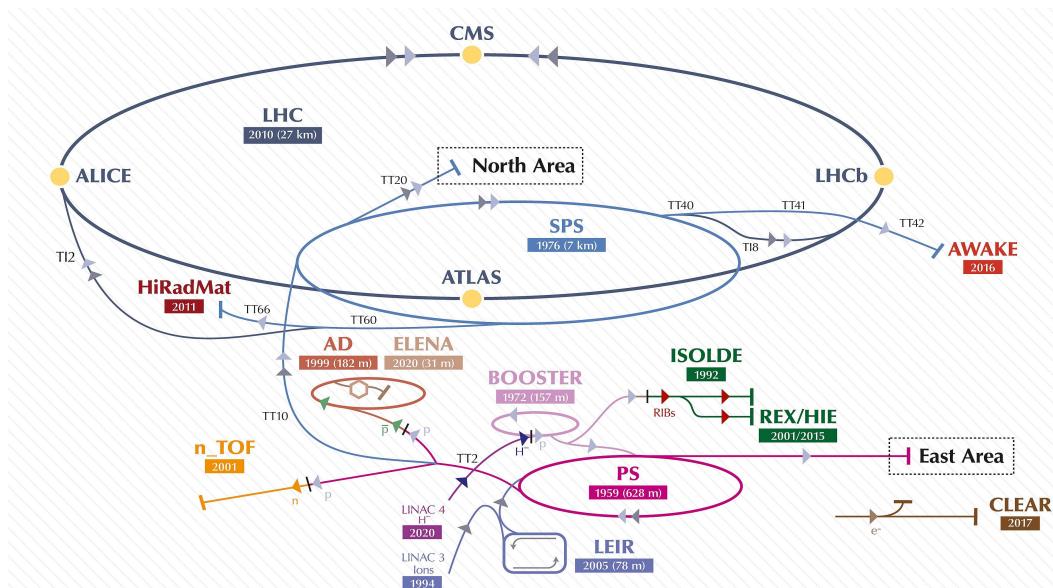
1.3 LHC 上的 CMS 实验

1.3.1 大型强子对撞机 (LHC)

大型强子对撞机 (LHC) 是世界上最大、能量最高的粒子对撞机。它由欧洲核研究组织 (CERN) 于 1998 年至 2008 年间与 10000 多名科学家、数百所大学和实验室以及 100 多个国家合作建造。它位于日内瓦附近法国-瑞士边界下方的一条周长 27 公里 (17 英里)、深达 175 米 (574 英尺) 的隧道中 (如图1.5所示)。

LHC 上的第一轮运行 (RUN I) 开始于 2010 年每个质子束 3.5TeV 的能量对撞，约为之前世界纪录的四倍。升级后，RUN II 达到每质子束 6.5TeV (总碰撞能量 13TeV ，目前世界最高)。2018 年底停产三年，正在进一步升级到 RUN III。

图 1.5 LHC 的探测器和部门分布^[12]



LHC 上有四个交叉点，加速粒子在这些交叉点发生碰撞。LHC 还有七个探测器，每个设用于检测不同的现象，位于交叉点周围。LHC 主要碰撞质子束，但它也可以加

速重离子束：铅-铅碰撞和质子-铅碰撞通常每年进行一个月。

LHC 的目标是让物理学家能够测试不同粒子物理理论的预测，包括测量希格斯玻色子的性质，寻找超对称理论和其他未解决的粒子物理问题。

1.3.2 紧凑缪子螺线管实验 (CMS)

紧凑型介子螺线管 (CMS) 实验是在瑞士和法国欧洲核子研究中心的大型强子对撞机 (LHC) 上建造的两个大型通用粒子物理探测器之一。CMS 实验的目标是研究广泛的物理学，包括寻找希格斯玻色子、额外维度和可能构成暗物质的粒子。

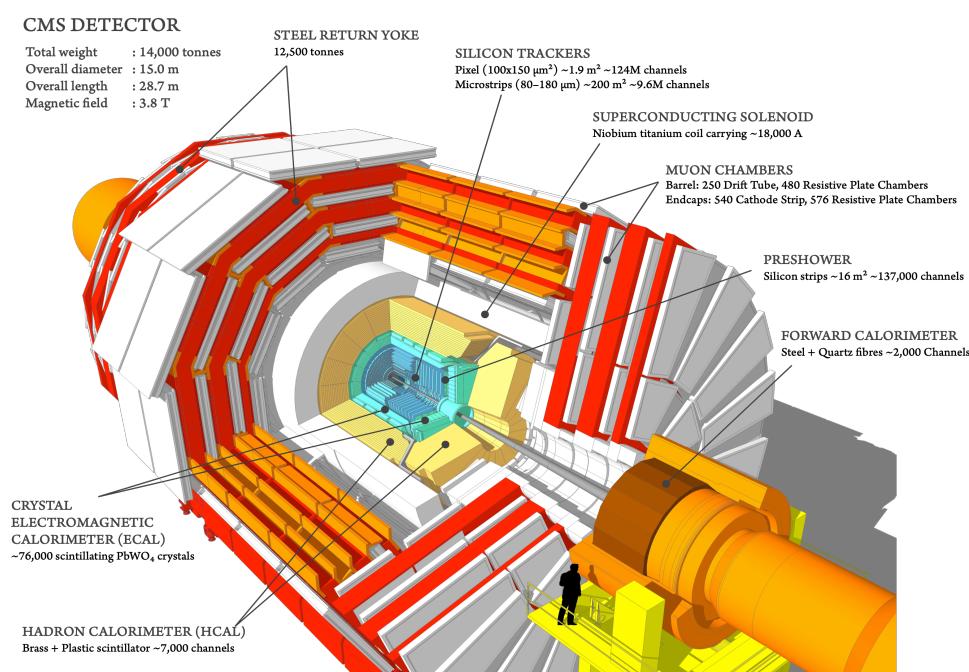
CMS 探测器长 21m，直径 15m，重约 14000 吨。来自 47 个国家/地区的 206 个科研机构的 4000 多人组成的 CMS 合作组建造并运行了探测器。2012 年 7 月，CMS 实验组与 ATLAS 实验组联合声明发现了希格斯玻色子。

CMS 实验的主要目标是：

- 在 TeV 尺度上探索物理学
- 进一步研究 CMS 和 ATLAS 已经发现的希格斯玻色子的性质
- 寻找超出标准模型的物理证据，例如超对称或额外维度
- 研究重离子碰撞的各个方面

位于 LHC 环另一侧的 ATLAS 实验的设计考虑了相似的目标，这两个实验旨在相互补充，以扩大范围并提供对研究结果的证实。CMS 和 ATLAS 使用不同的技术解决方案和探测器磁铁系统设计来实现目标。

图 1.6 CMS 探测器剖面图^[13]



CMS 被设计为通用探测器，能够在 LHC 粒子加速器的质心能量 0.9-13TeV 下研究质子碰撞的许多物理内容。CMS 探测器围绕一个巨大的螺线管磁铁构建。它采用圆柱形超导电缆线圈的形式，产生 4T 的磁场，大约是地球磁场的 100000 倍。磁场被限制产生在 12500 吨重量的的探测器主体——钢轭内（如图1.6所示）。CMS 探测器的一个不同寻常的特点是，它不像大型强子对撞机实验中的其他巨型探测器那样在地下原位建造，而是在地面上建造，然后分 15 个部分被降到地下并重新组装。

CMS 探测器包含用于测量光子、电子、 μ 子和其他碰撞产物的能量和动量的子系统。最内层是硅基径迹探测器，被闪烁晶体电磁量能器所包裹，而闪烁晶体电磁量能器本身又被一个强子采样量能器包围。径迹探测器和量能器足够紧凑，可以安装在 CMS 螺线管内，该螺线管可产生 3.8T 的强大磁场。磁体外部是大型 子探测器，它们位于磁铁的返回轭内。

第二章 大动量希格斯粒子和 $X \rightarrow WW$ 共振态的物理和研究动机

2.1 希格斯粒子的产生和衰变

在标准模型中，夸克、轻子、W、Z玻色子都通过希格斯机制被赋予质量。LHC 以及下一代对撞机实验的重要目标之一就是研究测量希格斯粒子的相关性质，因此，对 LHC 上希格斯粒子的产生和衰变测量本身就是极为重要的一个任务。

本节除了讨论希格斯粒子的产生和衰变，还将讨论 LHC 上大动量希格斯粒子的物理特性和研究动机，并且指出我们感兴趣的 $X \rightarrow WW$ 共振态搜寻的物理背景和动机（包括标准模型的希格斯粒子以及非标准模型的共振态）

2.1.1 希格斯粒子的产生

LHC 上标准模型希格斯粒子的产生道如下表2.1所示，

表 2.1 标准模型中希格斯粒子的产生道和相关参数^[14]

产生道	产生方式	分支比	截面
ggF	胶子通过 b/top 夸克圈聚合产生	~ 87%	48.5 pb
VBF	W/Z 矢量玻色子聚合产生	~ 7%	3.78 pb
WH	W/Z 矢量玻色子与希格斯联合产生	~ 4%	1.37 pb
ttH	正反 top 夸克与希格斯联合产生	~ 1%	0.51 pb
bbH	正反 bottom 夸克与希格斯联合产生	~ 1%	0.49 pb
tH	t,b,q' 与希格斯联合产生	~ 0.1%	0.09 pb

其中每个产生道都有独特的拓扑结构，对于其中的稀有道，尽管很难探测，却是研究超出标准模型物理的重要手段。

2.1.2 希格斯粒子的衰变

标准模型希格斯粒子的衰变道如下表2.2所示，对于其中五个非稀有衰变道，有以下性质：

- $\gamma\gamma$ 和 $ZZ \rightarrow 4\ell$ 道：高分辨率和高信噪比，常用来精确测量希格斯质量和微分散射截面。
- WW 道：高分支比，但衰变末态的中微子导致低信噪比和低分辨率。
- $\tau\tau$ 和 bb 道：高分支比，低信噪比，用于直接探测希格斯粒子与费米子的耦合。

表 2.2 标准模型中希格斯粒子的衰变道和相关参数^[14]

衰变道	主要衰变方式 (领头阶)	分支比	稀有衰变
$H \rightarrow bb$	直接衰变	~ 58.1%	否
$H \rightarrow WW$	直接衰变, 一个在壳一个离壳	~ 21.5%	否
$H \rightarrow \tau\tau$	直接衰变	~ 6.26%	否
$H \rightarrow ZZ$	直接衰变	~ 2.64%	否
$H \rightarrow \gamma\gamma$	通过 $W/t/b/\tau$ 圈间接衰变	~ 0.23%	否
$H \rightarrow \mu\mu$	直接衰变	~ 0.022%	是
$H \rightarrow Z\gamma$	通过 $W/t/b/\tau$ 圈间接衰变	~ 0.154%	是
$H \rightarrow cc$	直接衰变	~ 2.88%	是
$H \rightarrow gg$	通过 top/bottom 圈间接衰变	~ 8.18%	是

2.2 大动量希格斯粒子的物理特性和研究动机

对于大动量希格斯粒子的一个重要特征, 就是原本在常规希格斯粒子衰变中分开的两个喷注, 在大动量希格斯粒子衰变场景下, 会合并成一个喷注, 如图2.1所示, 所以, 对大动量希格斯粒子的标记和分析都存在很多和常规情况截然不同的地方。

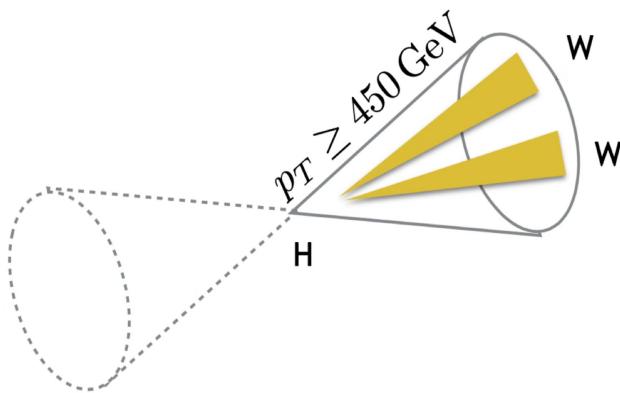


图 2.1 在动量足够大时, 希格斯粒子衰变产物会被重建到同一个喷注中

而对大动量希格斯粒子测量有如下物理动机:

1. 可以提高标准模型希格斯测量的敏感度: 在希格斯粒子 $p_T > 200[\text{GeV}]$ 的区域, $H \rightarrow bb$ 、 $H \rightarrow \tau\tau$ 道还有未分辨出的喷注等待分析; 在更高的 p_T 区域, $H \rightarrow ZZ/WW \rightarrow 4q$ 道也有同样的需求。
2. 检验标准模型在大动量区域的高阶修正的正确性或者发现可能的新物理算符^[15]。
3. 可以用作搜寻超出标准模型新物理的工具, 包括: 辐射子, Randall-Sundrum Bulk Graviton, 复合希格斯粒子, 新的矢量玻色子三重态, 暗物质的探寻 (如可用于解释暗物质和中微子质量的双希格斯二重态模型——2HDM^[16]) 等。

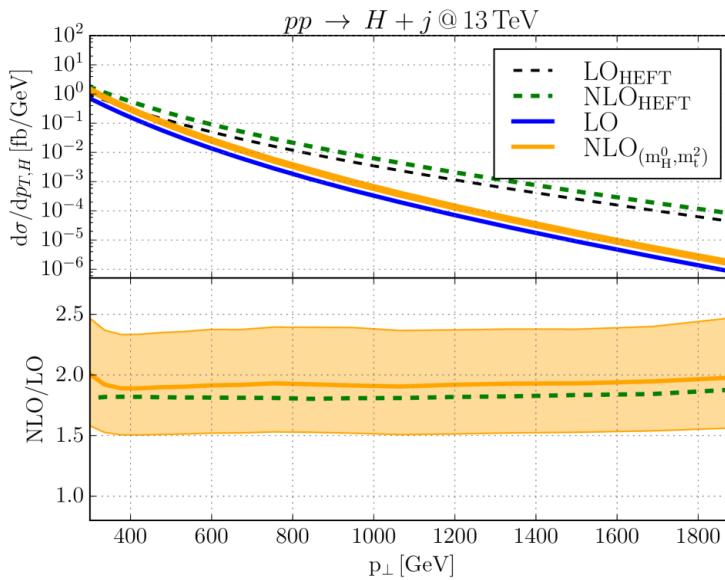


图 2.2 LHC 上质心系能量为 13TeV 时希格斯玻色子的横向动量分布。上图显示标准模型和无穷大 top 质量有效场论 (HEFT) 中领头阶 (LO) 和次领头阶 (NLO) 的预测。下半部分显示了各自的 NLO/LO 校准比值。黄色边带表示由于尺度变化导致的标准模型结果的理论误差。^[17]

对于希格斯粒子的大动量区域，唯象上有以下结论^[17]：标准模型对希格斯粒子动量微分截面的高阶修正在高动量区十分稳定，次领头阶与领头阶的比值固定在一定值范围内（如上图2.1所示）；而部分超标准模型理论^[18]的这一比值则会随动量增大而发生较大的增长（如下图2.2所示）。

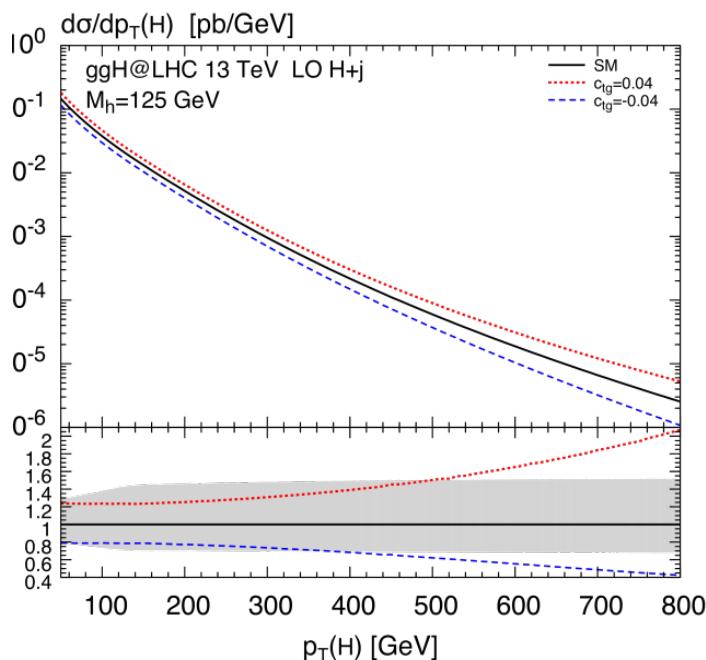


图 2.3 超标准模型的 chromomagnetic 算符对实验允许范围内的希格斯玻色子谱分布。^[18]

所以对大动量区域希格斯粒子的搜寻和研究可以用来检验标准模型的精确性以及提供支持/反对新物理的证据。

除此之外，对大动量希格斯玻色子的搜寻还与其他新物理密切相关，比如测量希格斯粒子的属性与标准模型预言的偏差，有助于探索暗物质的性质和物质-反物质不对称性等谜团，因为高能标尺度的新物理无法在大型强子对撞机上直接观察到，但是，在量子场论中，高能标的新物理会间接产生虚粒子，从而导致产生的希格斯玻色子数量作为 p_T 的函数与标准模型会出现较大偏差。这也是我们研究大动量希格斯粒子的一个重要原因：间接地估计高能标尺度上的新物理敏感性。。

2.3 X→WW 的物理背景及搜寻动机

WW 过程是 LHC 上第一批可观测到的双玻色子末态之一，而研究 WW 末态的最原始动机就是搜寻希格斯玻色子。当下，在 CMS 实验的 RUN III 即将开启运行和未来高亮度 LHC 的展望下，我们研究 WW 共振态的目的就变成了对标准模型希格斯粒子的精确测量和搜寻超标准模型的新 WW 共振态。

2.3.1 标准模型的 H→WW

对于标准模型的 $H \rightarrow WW$ 过程，衰变产物为一个质量在壳的 W 和一个质量离壳的 W^* 。所以也常写作 $H \rightarrow WW^*$ 。

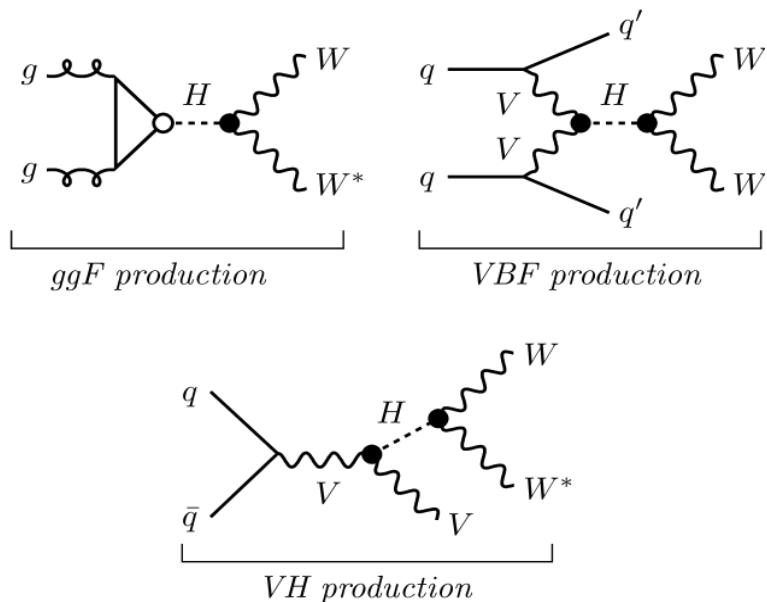


图 2.4 $H \rightarrow WW$ 的主要生产过程的领头阶费曼图：(a) 通过夸克圈的 ggF 生产过程；(b)VBF 生产过程，末态除了两个 W 玻色子之外还有两个夸克喷注；(c)VH 联合产生，末态有三个玻色子

LHC 上最多的 $H \rightarrow WW$ 产生过程就是胶子聚合产生 (ggF)，如图2.4左上所示。初态胶子通过 top 夸克圈耦合成希格斯粒子，然后希格斯粒子继续衰变成一对 W 玻色子。次多产生模式是矢量玻色子聚合产生 (VBF)，如图2.4右上所示，其生产截面大约比 ggF 小一个数量级。两个初态夸克辐射出 W 或 Z 玻色子，然后 W 或 Z 玻色子聚合成希格斯粒子，并进一步衰变成一对 W 玻色子。但 VBF 末态会生成两个喷注加一对 W 玻色子，而 ggF 只会生成一对 W 玻色子。

$H \rightarrow WW^*$ 的主要本底来自于标准模型下的非共振 WW 生产道连续谱。在 LHC 上， WW 生产过程由 $q\bar{q}$ 湮灭过程主导^[19]，如图2.5所示：左边和中间分别是 $qq' \rightarrow WW$ 过程的 t 通道图、s 通道图，其中 s 通道过程对于 WWZ 和 $WW\gamma$ 的三玻色子耦合顶点十分敏感。次领头阶的 WW 本底还有如右边所示的胶子聚合 (ggF) 生成 WW 的 box 图，尽管是次领头阶图，但这个过程被 LHC 上的高亮度胶子产物增强，从而对非共振态 WW 的产生贡献也不可忽视。

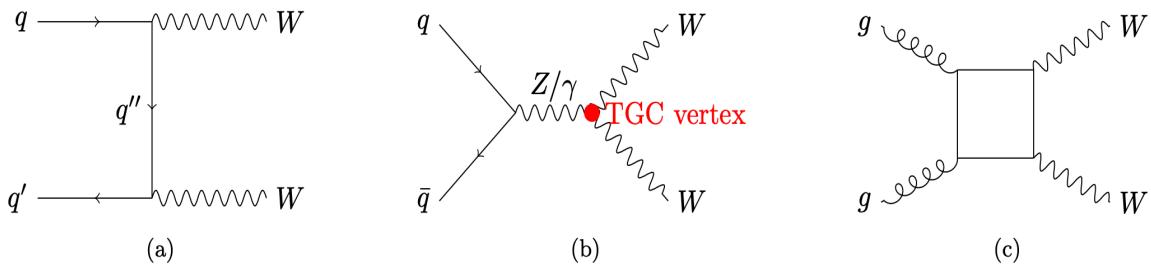


图 2.5 标准模型非共振态 $X \rightarrow WW$ 的主要产生过程^[19]：(a) $qq' \rightarrow WW$ 的 t 通道图；(b) $qq' \rightarrow WW$ 的 s 通道图，有一个三规范玻色子耦合顶点 (TGC)；(c) $gg \rightarrow WW$ 的 box 图，贡献被 LHC 的高亮度胶子抬高。

在标准模型框架下，研究双玻色子产生提供了在 TeV 尺度检验标准模型电弱理论的机会。 WW 过程就是其中重要的一个例子。除了 $H \rightarrow WW$ 产生道之外， WW 产生过程还对三玻色子耦合顶点十分敏感（如图2.5(b) 所示），从而提供了检验标准模型规范对称性（关于三玻色子耦合限制）的重要机会。通过对 WW 产生过程的三玻色子耦合的精确测量，我们有机会探寻到包含规范玻色子的新物理现象。（见2.3.2节）

2.3.2 超标准模型的 $X \rightarrow WW$

对于标准模型的 $H \rightarrow WW$ 共振态和更重质量的 WW 共振态 X ，它们的喷注重建也存在区别，如下图所示，重质量的 $X \rightarrow WW$ 共振态会产生大动量的 W 玻色子从而导致每个 W 衰变的两夸克喷注被重建在同一个 W 喷注里。

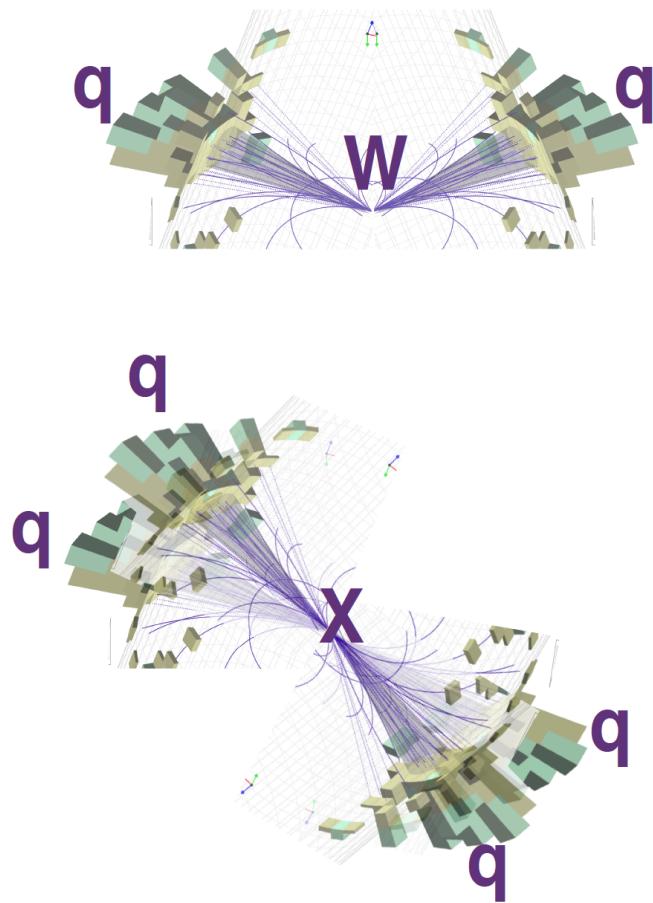


图 2.6 低 p_T 的 W 玻色子衰变出的两个夸克会被重建为两个喷注，重质量的 $X \rightarrow WW$ 共振态会产生两个大动量的 W 玻色子从而导致每个 W 衰变的两夸克被重建合并为一个喷注。^[20]

在这样的物理背景下，对超标准模型的 $X \rightarrow WW$ 共振态搜寻有助于探寻 V_{kk} 等重质量新物理粒子，如下图2.7所示

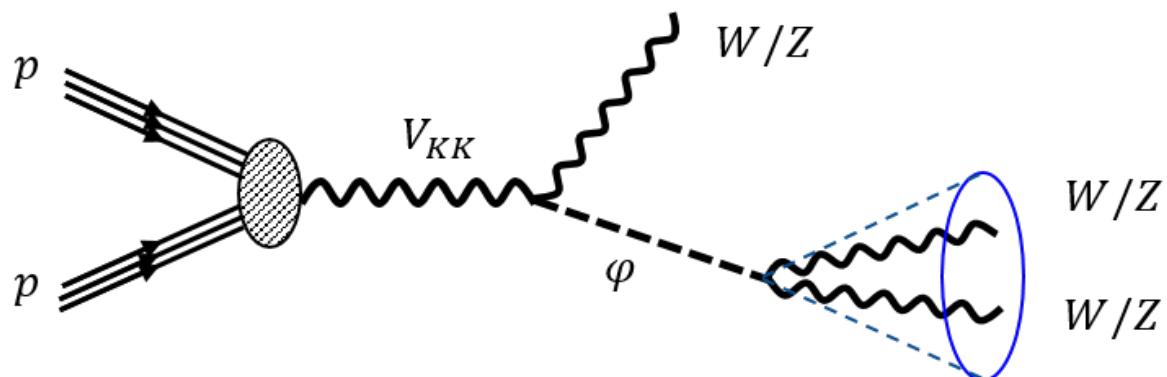


图 2.7 可考虑的超出标准模型物理过程，基于扭曲的额外维 RS 模型的扩展。 V_{kk} 表示 W/Z 的 Kaluza-Klein(KK) 模式，并且对应于较重的母粒子。 ϕ 代表扮演辐射子(radion)角色的标量粒子。两个 W/Z 粒子周围来自 ϕ 衰变的圆锥体表明它们是高度准直的。^[21]

除了对重质量新物理粒子的搜寻, $X \rightarrow WW$ 共振态还可以用来研究非标准模型质量希格斯粒子的衰变分支比关系, 如图2.8展示了希格斯衰变分支比作为希格斯质量的函数。可以看到双 W 衰变道在很大的质量范围内都是最大分支比。特别是当 $M_H > 130[\text{GeV}]$ 时, 就主要是 $H \rightarrow WW$ 衰变了。在 $2m_W < M_H < 2m_Z$ 范围内, $H \rightarrow WW$ 的分支比非常显著。

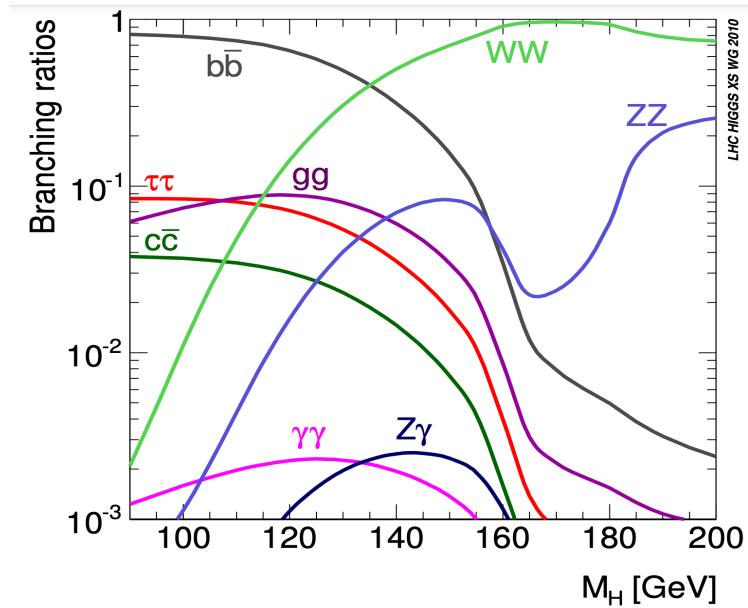


图 2.8 不同质量希格斯粒子的衰变分支比^[19]

因为我们知道 $H \rightarrow WW$ 产生的速率由希格斯到 WW 的分支比决定, 所以如果测量得到的 $H \rightarrow WW$ 数目与标准模型预言有较大偏差, 就暗示着有可能存在着非标准模型质量的伴生/复合希格斯粒子。

如下图2.9所示, 展示了一种可能的复合希格斯粒子的新物理模型^[20]

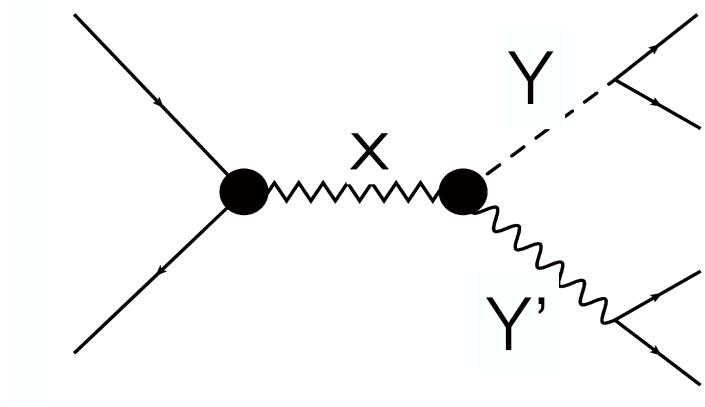


图 2.9 X 代表 TeV 能量尺度的新物理粒子, Y/Y'/Z/H

对 $X \rightarrow WW$ 共振态的搜寻，可以提供类似上图新物理过程的证据，发现可能的额外维度拉格朗日项、复合的希格斯粒子以及扩展的规范对称性（更多种类的 Z/W 玻色子）。

第三章 CMS 实验的重建与标记技术介绍

3.1 重建事例时缓解顶点堆积的 PUPPI 算法

3.1.1 顶点堆积 (pile-up)

事例堆积 (pile-up) 是指：大型强子对撞机（以及所有粒子物理加速器）上的质子流是一束一束的，而不是连续的。每次当一个质子流中的一束遭遇另一个质子流中的一束时，会有多个质子彼此相互作用。通常，只有其中的硬散射才能产生我们感兴趣的粒子，但在两束质子遭遇时一般只有小于等于一个相互作用顶点是“硬散射”。尽管如此，硬散射以外的相互作用仍然会大量发生，并且每个相互作用都会产生一些新粒子。这就是我们所说的堆积 (pile-up)。（另一种说法是“多重相互作用”）

堆积对于我们重建事例会造成困难，一个通俗的理解方式就是烟花，如下图3.1所示



图 3.1 把烟花绽放顶点比作事例顶点，烟花出射轨迹比作事例出射粒子/喷注

我们在按照出射轨迹重建“烟花”顶点时，还会受到周围其他“烟花”顶点出射轨迹的影响，从而对我们重建感兴趣的“烟花”顶点造成很大困难。

3.1.2 PUPPI 算法

现在 CMS 实验中使用 PUPPI 算法来解决堆积困难，其全称是 Pile-Up Per Particle Identification^[22]，以下是对该算法的介绍。

对于一个事例中的每个粒子 i , 我们可以定义一个变量 α_i 满足

$$\alpha_i = \log \sum_{j \in \text{Event}} \frac{p_{Tj}}{R_{ij}} \times \Theta(R_{\min} \leq \Delta R_{ij} \leq R_0) \quad (3.1)$$

这里 $\Theta(R_{\min} \leq \Delta R_{ij} \leq R_0) \equiv \Theta(\Delta R_{ij} - R_{\min}) \times \Theta(R_0 - \Delta R_{ij})$, Θ 是阶跃函数, ΔR_{ij} 是粒子 i 和 j 在 $\eta - \phi$ 空间的距离, p_{Tj} 是粒子 j 的横向动量 (单位为 [GeV]), R_0 定义了每个粒子 i 附近的圆锥所以只有在圆锥内的粒子才会参与 α_i 的计算。除此之外, 离粒子 i 距离小于 R_{\min} 的粒子也会被从求和中舍弃, R_{\min} 作为粒子 i 的共线分裂调整器。通常取 $R_0 = 0.3$, $R_{\min} = 0.02$ 。

在 Particle-Flow 中, 我们可以把所有粒子候选者分为三类: 中性粒子; 来自领头顶点的带电强子; 来自堆积顶点的带电强子。这样我们可以把(3.1)中的求和分解为

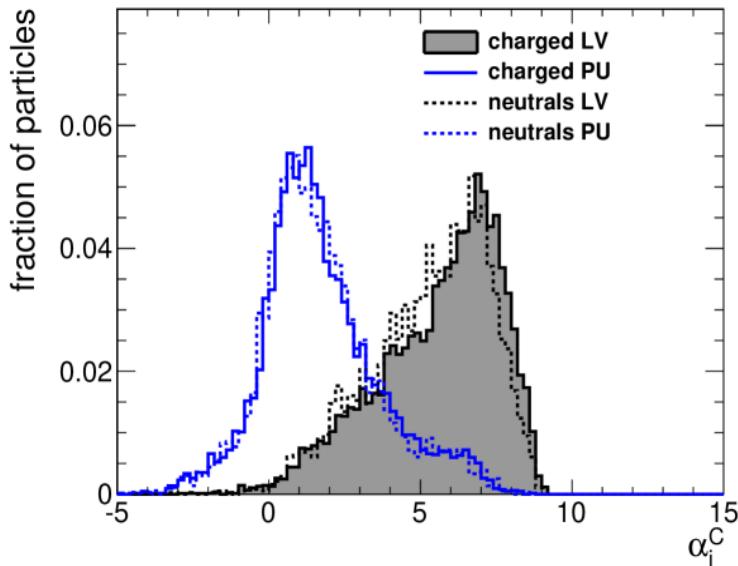
$$\sum_{j \in \text{Event}} = \sum_{j \in \text{Ch,PU}} + \sum_{j \in \text{Ch,LV}} + \sum_{j \in \text{Neutral}} \quad (3.2)$$

因此, 我们可以使用来自领头顶点的带电粒子作为来自领头顶点的所有粒子的代理。

当带电粒子径迹信息可用时, 我们可以计算出粒子 i 与所有来自领头顶点的带电粒子的关系如下

$$\alpha_i^C = \log \sum_{j \in \text{Ch,LV}} \frac{p_{Tj}}{R_{ij}} \times \Theta(R_{\min} \leq \Delta R_{ij} \leq R_0) \quad (3.3)$$

图 3.2 $p_T > 1 \text{ GeV}$ 时, 领头顶点和堆积顶点出射粒子的 α_i^C 变量分布^[22]



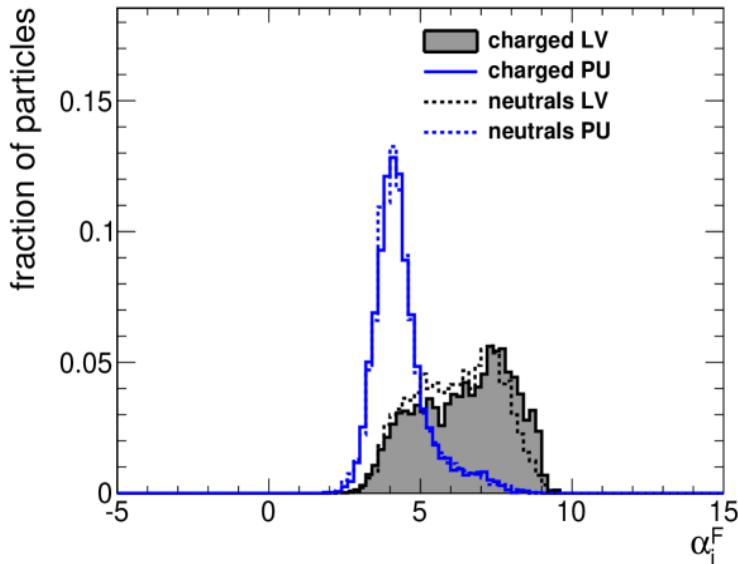
当带电粒子径迹信息不可用时, 我们只能计算出粒子 i 与来自领头顶点的所有粒

子的关系如下

$$\alpha_i^F \equiv \alpha_i = \log \sum_{j \in \text{Event}} \frac{p_{Tj}}{R_{ij}} \times \Theta(R_{\min} \leq \Delta R_{ij} \leq R_0) \quad (3.4)$$

同时，我们假设总和中的所有粒子都来自领头顶点，虽然有来自堆积顶点的噪音贡献，但相对于领头顶点粒子的贡献会被分子上的 p_{Tj} 所抑制。这样的话算法就仍可以为没有径迹信息的区域分配权重。

图 3.3 $p_T > 1 \text{ GeV}$ 时，领头顶点和堆积顶点出射粒子的 α_i^F 变量分布^[22]



利用带电粒子径迹信息可以获得一个好处：对于带电粒子，我们可以利用径迹信息知道其是否来自堆积顶点，从而可以计算 α_i^F 和 α_i^C ；对于中性粒子，我们在重建水平无法知道其是否来自堆积顶点，所以我们可以用带电粒子的 α_i^F 和 α_i^C 估计中性粒子的对应分布（这暗示了带电粒子和中性粒子具有相同分布，这一点可以从图2.2和从图2.3中看出）。

现在，我们用中位数和标准差表征领头顶点和堆积顶点的 α_i^F 和 α_i^C ：

$$\bar{\alpha}_{\text{PU}}^F \text{median}\{\alpha_{i \in \text{Ch,PU}}^F\}, \quad \sigma_{\text{PU}}^F = \text{RMS}\{\alpha_{i \in \text{Ch,PU}}^F\} \quad (3.5)$$

$$\bar{\alpha}_{\text{PU}}^C \text{median}\{\alpha_{i \in \text{Ch,PU}}^C\}, \quad \sigma_{\text{PU}}^C = \text{RMS}\{\alpha_{i \in \text{Ch,PU}}^C\} \quad (3.6)$$

接着为了区分来自领头顶点和来自堆积顶点的粒子，我们引入了一个变量以区分这两类，并使用它来计算每个粒子的权重。该变量如下

$$\chi_i^2 = \Theta(\alpha_i - \bar{\alpha}_{\text{PU}}) \times \frac{(\alpha_i - \bar{\alpha}_{\text{PU}})^2}{\sigma_{\text{PU}}^2} \quad (3.7)$$

这里 Θ 是阶跃函数， χ_i^2 用来度量 α_i 与中位数 $\bar{\alpha}_{\text{PU}}$ 的偏离程度。

然后是对粒子权重的定义：对于领头顶点粒子，理想权重是 1，对于堆积顶点粒子理想权重是 0。在实际中，为了在有限信息下估计一个粒子有多大可能来自堆积顶点，权重可以是介于 0 和 1 之间的任意值。我们把权重记为 w_i ，如下计算：

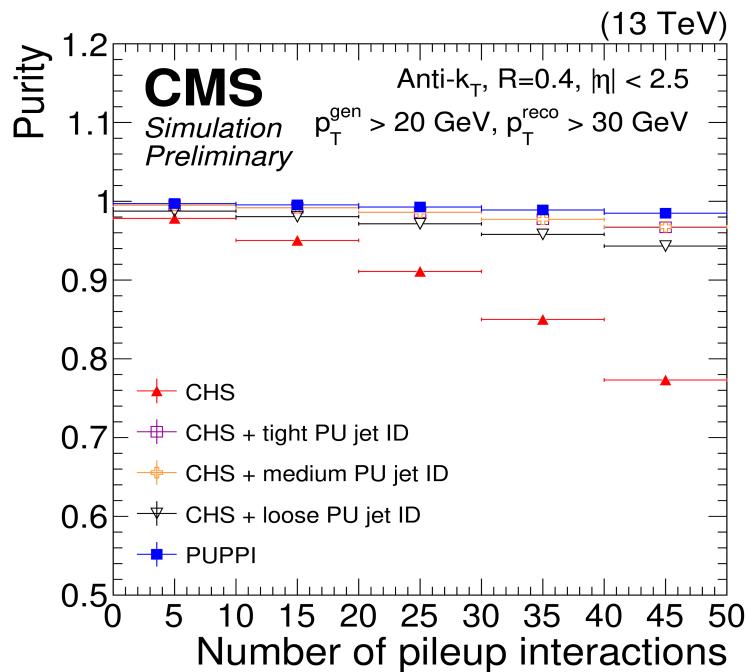
$$w_i \equiv F_{\chi^2, \text{NDF}=1}(\chi_i^2), \quad (3.8)$$

这里 F_{χ^2} 是变量 χ^2 的累积分布函数。

这样我们就知道，对于任何高于中位数的 α_i ， chi_i^2 都会对应于一个介于 0 和 1 之间的权重，并被认为可能来自于领头顶点；反之则对应权重为 0 的情况，并被认为来自于堆积顶点。

puppi 算法在 CMS 实验上的效果如下图2.4所示，可以看到在堆积顶点数目上升时，重建顶点是领头顶点的纯度依然非常稳定接近于 1

图 3.4 PUPPI 算法效果^[23]



至此，我们可以总结 PUPPI 算法流程如下：

1. 计算事例中所有带电粒子的 α_i^F 和 α_i^C 以及对应的中位数和标准差。
2. 所有来自堆积顶点的带电粒子权重都设为 0，所有来自领头顶点的带电粒子权重都设为 1。
3. 然后使用方程(3.7)和方程(3.8)来计算剩余粒子的权重。
4. 所有粒子的四动量都进行重新加权： $p_i^\mu \rightarrow w_i p_i^\mu$ 。

5. 权重或重加权动量小于一定阈值的粒子则被舍弃: $w_i < w_{cut}$ 或 $p_{Ti} < p_{T,cut}$ 。
6. 剩余被重加权粒子的集合则被当作是堆积修正了的事例。

3.2 重建喷注的 anti-kT 算法

Jet 聚类算法是用于分析强子碰撞数据的主要工具之一，下面我们将介绍目前主流的 anti-kT 算法^[24]是如何重建喷注的。

首先我们有一个待处理列表，里面包含所有待处理的物理对象（包括粒子和已经定义的喷注）。接着对于聚类算法，我们要定义两个物理对象 i, j 之间的距离 d_{ij} ，还要额外定义每个物理对象 i 和入射束流 B 之间的距离 d_{iB} 。这两类距离的定义如下：

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \quad (3.9)$$

$$d_{iB} = k_{ti}^{2p} \quad (3.10)$$

这里 $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ ，并且 k_{ti} , y_i , ϕ_i 分别是横向动量, 快度, 方位角。

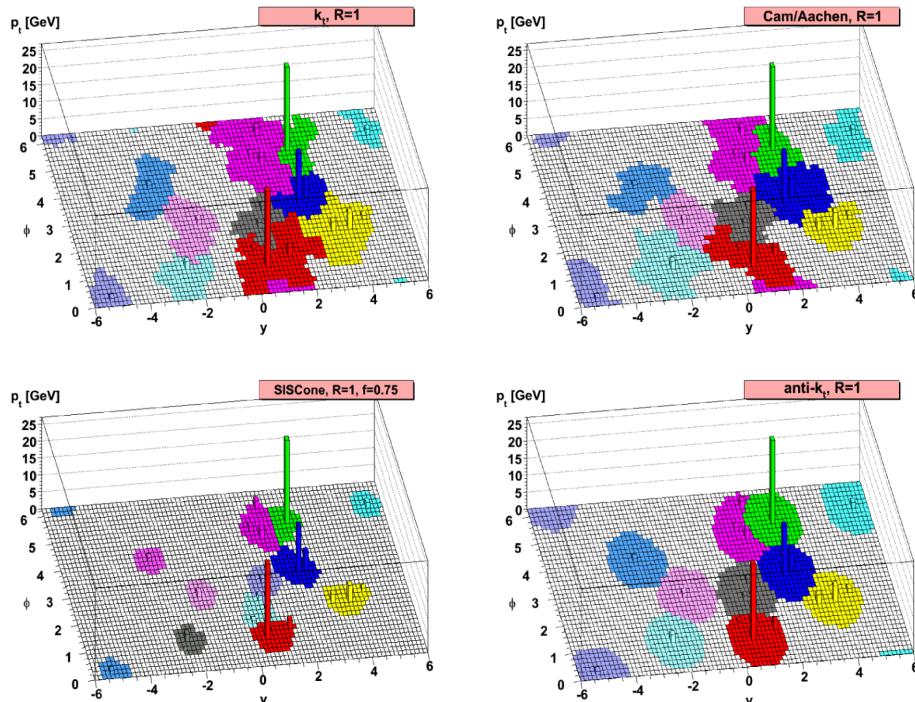


图 3.5 用 anti-kT 算法和 kT, Cambridge/Aachen, SIScone 算法比较得到的逆反应分布。这是用 Pythia6.4 模拟的双喷注事件计算的，其中最硬的两个喷注满足 $p_T > 200[GeV]$ 并且都位于 $y < 2$ 区域。逆反应用对应于两个最硬喷流中每一个的净横向动量变化，这是由于当高亮度 LHC 堆积粒子被加入事例时非堆积粒子的重新加权（每束交叉时约有 25 个 pp 相互作用）。^[24]

对于方程(3.9)和方程(3.10)中的幂指数 $2p$, 如果我们取 $p=1$, 这就是常规的 k_T 算法, 并且对于任意 $p>0$ 的取值都有类似的表现, 只有 $p=0$ 的时候才会变成对应的“Cambridge/Aachen”算法。但是 p 还有一种取值, 就是取 $p<0$, 这里对于所有 $p<0$ 的取值, 软辐射的行为都是类似的, 我们专门取 $p=-1$, 并且称之为 $\text{anti-}k_t$ 算法, 也叫 AK 算法。 R 通常会取 0.4、0.8、1.5 三个值, 分别对应的重建出的喷注名称是 AK4 喷注、AK8 喷注、AK15 喷注。

我们关心的 $\text{anti-}k_T$ 聚类重建算法是这么执行的:

1. 在所有的距离中找到最小的距离, 如果:
 - (1) 最小距离是来自两个物理对象 i, j 的距离 d_{ij} , 那我们就把这两个物理对象 (i,j) 从待处理列表中取出, 合并定义为新喷注, 再放回待处理列表;
 - (2) 最小距离是来自物理对象 i 和入射束流 B 之间的距离 d_{iB} , 那我们就把物理对象 i 定义为一个喷注, 同时把它从待处理列表中移去 (表示已处理完)
2. 重新计算待处理列表中的所有距离 (包含 d_{ij} 和 d_{iB} 两类) 并重复 1. 步骤, 直到待处理列表中没有任何物理对象存在。

通过以上流程, 我们就重建了粒子流中的所有 AK 喷注。

$\text{anti-}k_T$ 算法与其他算法重建喷注效果在蒙特卡洛模拟样本上的比较如下图2.6所示

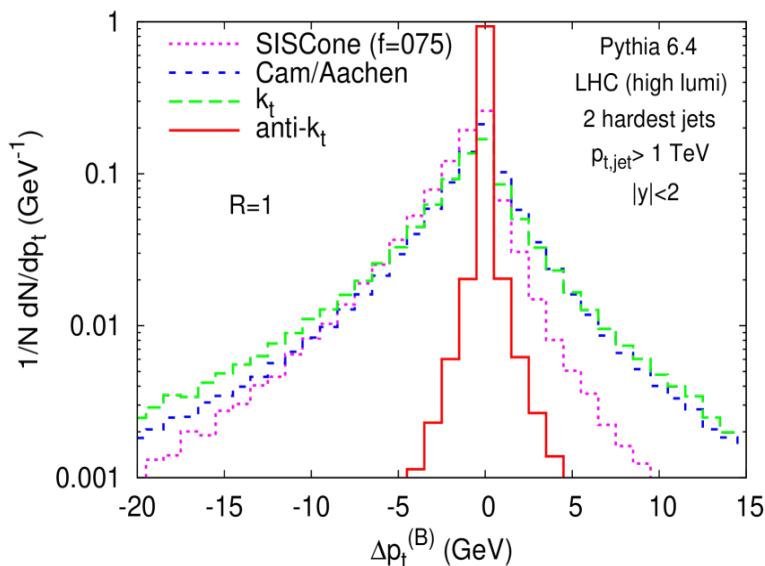


图 3.6 用 $\text{anti-}k_T$ 算法和 k_T , Cambridge/Aachen, SISCone 算法比较得到的逆反应分布。这是用 Pythia 6.4 模拟的双喷注事件计算的, 其中最硬的两个喷注满足 $p_T >$ 并且都位于 $y < 2$ 。逆反应对应于两个最硬喷流中每一个的净横向动量变化, 这是由于当高亮度 LHC 堆积被加入事例时非堆积粒子的重新分配 (每束交叉时约有 25 个 pp 相互作用)。^[24]

可以看到用 anti-kT 算法重建出来的喷注具有更好的分辨率。

3.3 喷注标记算法历史发展^[25]

按照历史发展，对于喷注标记技术，先后出现了基于理论的高级变量算法，基于机器学习的高级变量算法和基于深度学习的初级变量算法^[25]。下面我们将分别介绍其中的代表性算法。

3.3.1 基于理论的高级变量算法

基于筛选条件的标记算法从理论灵感提供的变量出发，在实验和理论两方面都经受了广泛的研究，具有鲁棒性和易于理解和实现的特点，同时也是与新算法比较的基准线。

3.3.1.1 soft-drop mass 算法

我们知道一个喷注由许多的子喷注构成，怎么取舍目标喷注中的子喷注便成了算法的核心问题。

soft-drop 算法，就是丢弃掉喷注当中较软且偏离喷注中心较远的子喷注，然后计算剩余较硬且集中在喷注内部的部分的不变质量，其中，剩余的部分子喷注要满足如下的软滴条件：

- soft-drop 条件：

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut} \left(\frac{\Delta R_{12}}{R_0} \right)^\beta \quad (3.11)$$

这里

$$R_{12} = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2} \quad (3.12)$$

是子喷注 1 和子喷注 2 之间的角距离， R_0 是要求的某个阈值。对于 CMS 实验，通常我们取 $\beta = 0, z_{cut} = 0.1$, soft-drop 条件(3.11)简化为

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > 0.01 \quad (3.13)$$

通过 soft-drop mass 算法，我们可以得到我们想要的子喷注构成的喷注，从而计算出喷注对应的软滴质量 (soft drop mass)，软滴算法可以大大减少喷注质量分布中的“Sudakov”峰结构，使信号分辨率更明显。

3.3.1.2 N-subjettiness 算法

高级变量 N-subjettiness 定义为

$$\tau_N = \frac{1}{d_0} \sum_i p_{T,i} \min [\Delta R_{1,i}, \Delta R_{2,i}, \dots, \Delta R_{N,i}] \quad (3.14)$$

这里 $\Delta R_{j,i}$ 是指第 j 的子喷注到第 i 个子喷注的角距离。通过 τ_N 这个变量，我们可以量化一个喷注拥有 N 个子喷注的兼容性。

进一步地，我们可以通过不同 τ_N 之间的比值获得更有鉴别效果的变量，例如

(1) 定义

$$\tau_{21} = \frac{\tau_2}{\tau_1} \quad (3.15)$$

可以针对鉴别二分叉喷注（如 W, Z, H）。

(2) 定义

$$\tau_{32} = \frac{\tau_3}{\tau_2} \quad (3.16)$$

可以针对鉴别三分叉喷注（如 top），同时对 top 喷注的 b 夸克子喷注还可以运用 τ_{21} 来进一步改善效果。

在重建中，我们一般在应用 soft-drop mass 算法后计算喷注的 ECF 比率，这提高了 ECF 作为喷注质量和 p_T 函数的稳定性。

3.3.1.3 ECF: N_2 算法

这里我们要用到泛化能量关联函数 (ECF)，对于一个包含 N_c 个子喷注的喷注，它的 ECF 如下定义

$$q e_N^\beta = \sum_{1 \leq i_1 < i_2 < \dots < i_N \leq N_c} \left[\prod_{1 \leq k \leq N} \frac{i_k}{J} \right] \prod_{m=1}^q \min_{i_j < i_k \in \{i_1, i_2, \dots, i_N\}} \left\{ \Delta R_{i_j, i_k}^\beta \right\} \quad (3.17)$$

这个变量可以用来测试喷注有 N 个辐射中心的兼容性，与 N-subjettiness 变量有点相似，但是 ECF 是无轴方法，并且对于 N 分叉喷注，如果 $N > M$ ，我们会有 $e_N \gg e_M$ 。

对于二分叉标记喷注 (W/Z/H)，可以定义 ECF 比值为

$$N_2^1 = \frac{{}_2 e_3^1}{{}_1 e_2^1)^2} \quad (3.18)$$

与 N-subjettiness 比值 τ_{21} 相比，其优点是它作为喷注质量和 p_T 的函数更稳定，这种方法也被称为 “ $m_{SD} + N_2$ ” 算法。

本算法还有质量去相关的版本，便于压低峰状本底，具体做法如下，定义“设计

去相关标记器” 变量为

$$N_2^{DDT}(\rho, p_T) = N_2(\rho, p_T) - N_2^{(X\%)}(\rho, p_T) \quad (3.19)$$

此处 $\rho = \ln(m_{SD}^2/p_T^2)$ 是一个无量纲变量, $N_2^{(X\%)}$ 是模拟 QCD 样本中 N_2 分布的 X 百分位的取值。这确保了筛选条件 $N_2^{DDT} < 0$ 会导致在考虑的质量与横向动量范围内 QCD 的标记效率为恒定 X%, 并且没有标记性能上的损失。

3.3.2 基于机器学习的高级变量算法

3.3.2.1 N_3 -BDT 算法

我们想考虑具有尺度不变性的 ECF 比率, 可以通过以下式子定义的变量来构造:

$$\frac{{}_a e_N^\alpha}{({}_b e_M^\beta)^x}, \text{ where } M \leq N \text{ and } x = \frac{\alpha\alpha}{b\beta}. \quad (3.20)$$

对于 top 夸克喷注标记算法, 仅考虑彼此不高度相关的那些变量, 并且丢弃无法很好被模拟定义的比值变量, 这样我们得到如下 11 个比值变量

$$\begin{aligned} & \frac{{}_1 e_2^{(2)}}{{}_1 e_2^{(1)}}^2, \frac{{}_1 e_3^{(4)}}{{}_2 e_3^{(2)}}, \frac{{}_3 e_3^{(1)}}{{}_1 e_3^{(4)}}^{3/4}, \frac{{}_3 e_3^{(1)}}{{}_2 e_3^{(2)}}^{3/4}, \frac{{}_3 e_3^{(2)}}{{}_3 e_3^{(4)}}^{1/2}, \frac{{}_1 e_4^{(4)}}{{}_1 e_3^{(2)}}^2, \\ & \frac{{}_1 e_4^{(2)}}{{}_1 e_3^{(1)}}^2, \frac{{}_2 e_4^{(1/2)}}{{}_1 e_3^{(1/2)}}^2, \frac{{}_2 e_4^{(1)}}{{}_1 e_3^{(1)}}^2, \frac{{}_2 e_4^{(1)}}{{}_2 e_3^{(1/2)}}^2, \frac{{}_2 e_4^{(2)}}{{}_1 e_3^{(2)}}^2. \end{aligned} \quad (3.21)$$

基于 ECF 的 top 夸克标记器, 称为 “ N_3 -BDT (CA15)”, 使用扩展决策树模型, 以这 11 个 ECF 比值变量加上, τ_{32}^{SD} 和 f_{rec} 作为输入。

3.3.2.2 HOTVR 算法

全称为 “带 R 变量的重对象标记器” (Heavy Object Tagger with Variable R), 带 p_T 无关的变量距离参数 R 的喷注簇射和经过 Puppi 算法修正 Pile-up 的 ParticleFlow 候选者, 在这个过程中, 软簇射会被丢弃掉, 从而得到稳定的喷注质量分布, 同时阻止额外的辐射进入喷注。

可以被用于标记不同的重共振态 ($t/W/Z/H$)。

3.3.2.3 BEST 算法

全称是 “扩展事例形状标记器” (Boosted Event Shape Tagger), 是针对 top/W/Z/Higgs 的多分类标记器, 在参考坐标系中计算喷注运动学/形状的变量, , 并且把参考坐标系分别按 top/W/Z/Higgs 喷注假设变换为四个静止坐标系, 如果被变换到了正确的静止坐标系, 那么喷注的子组分就应该是各向同性并且会展示出预期的 N 分叉结构。

我们使用神经网络训练这些运动学变量和子喷注的 b 标记判别式，这个神经网络由三个全连接层构成，每层带有 40 个节点。

3.3.3 基于深度学习的初级变量算法

这里实际上已经开始进入深度学习时代，基于深度学习的新标记算法在最近几年已经被提上预案并且受到了大量关注，基本思想就是使用初级变量加上深度神经网络，对于喷注标记，有两种深度神经网络的路径：

1. 基于图像：

把喷注转化为使用量能器能量沉积得到的图像，利用计算机视觉技术——通常是二维卷积神经网络。但是由于图像的稀疏性和异构探测器，仍然挑战和困难重重。

2. 基于粒子：

把喷注当成它自己组分粒子的集合，这样可以利用循环神经网络，一维卷积神经网络和图神经网络等等技术。同时还可以通过 CMS 的 Particle-Flow 重建流程产生诱导出更多自然的想法，合并所有子探测器的信息并且充分利用粒度。

现在这两条算法路径在 CMS 实验中都在开发，以下将通过两个例子分别介绍这两条路径的情况。

3.3.3.1 ImageTop 算法

这是基于喷注图像的 top 夸克标记算法，喷注图像基于喷注横向动量自适应缩放以增加高 p_T 区域的准直。喷注图像有四个“颜色”通道：(1) 中性 p_T (2) 径迹 p_T (3) μ 子个数 (4) 径迹条数。此算法还利用了深度喷注 b 夸克标记的判别式。

质量去相关的 ImageTop 算法：训练时重新加权 QCD 样本使得本底的质量分布与 top 夸克的质量分布相匹配，从而获得 ImageTop-MD 标记器。

3.3.3.2 DeepAK8 算法

DeepAK8 是基于 AK8 喷注针对 top/W/Z/Higgs 标记任务的多分类标记器，其中还会按照衰变道进行进一步的子分类（例如， $Z \rightarrow bb$, $Z \rightarrow cc$, $Z \rightarrow qq$ 等）。此算法直接用喷注组分（如 ParticleFlow 候选者，二级顶点等）作为输入，采用一维卷积神经网络作为架构

质量去相关的 DeepAK8 算法：使用对抗训练技术，训练时重新加权本底样本和信号样本获得 m_{SD} 和 p_T 的二维平分布以辅助训练。

第四章 用于喷注标记的 ParticleNet 深度神经网络

喷注是 LHC (大型强子对撞机) 上无处不在且蕴含大量粒子信息的物理对象，因此，关于喷注的标记就是许多潜在新物理的探寻与标准模型测量检验的关键，而喷注标记任务主要分为以下三类：

1. 重夸克喷注标记
2. 重共振态标记
3. 夸克/胶子的区分鉴别

而分类标记任务，也是机器学习领域最活跃的方向之一。因此，要结合以上两个领域，关键性的问题就是：**怎样尽可能物理地把喷注表示成机器学习中的对象？**

4.1 喷注表示方式

从深度学习的计算机视觉领域出发，最经典的方式是表示成图像，如1所介绍。还有一种路径是把喷注表示成粒子的集合，如2所介绍。这两种路径的表示方法在 CMS 实验分析中已经有了相关的探索和标记器的开发。但我们还应该保持好奇：是否还有更好的表示方法和与之对应更好的网络架构呢？

这就是我们开发的标记器所采用的喷注表示方法的基础：点云（Particle Clouds）。点云，就是空间中数据点的集合，这类数据结构的收集方式是通过三维扫描测量物体表面周围的大量点而得到。但对于我们关心的物理喷注，应当不仅仅用点云，或者说，应该用点云的一个喷注适应版：粒子云^[26]（Particle Cloud）。

喷注（或者说，粒子云）就是空间中粒子的集合。粒子云的收集方式是用粒子探测器测量到的大量粒子的聚类。

经过喷注适应后的粒子云表示方式和点云表示方式有如下的关联：

- 共同点：点云中的点和粒子云中的粒子都是内禀无序的。
- 不同点：点云中的基本信息是 xyz 空间的 3 维坐标；粒子云中的基本信息是 $\eta - \phi$ 空间的二维坐标，但同时还具有许多其他特征，如：能动量，电荷，粒子鉴别（Particle ID），径迹质量，探测器受击参数等。

粒子云的置换对称性使其成为喷注最自然和有希望的表示。为了实现粒子云表示的最佳性能，必须仔细设计神经网络的架构以充分利用这种表示的潜力。在本节中，我们

将介绍 ParticleNet，这是一种类似于 CNN 的深度神经网络，使用粒子云数据进行喷注标记。

4.2 边卷积 (EdgeConv)

CNN 在计算机视觉的各种机器学习任务中取得了压倒性的成功。CNN 的两个关键特征对其成功做出了重大贡献。首先，卷积操作通过在整个图像中使用共同核函数来利用图像的平移对称性。这不仅大大减少了网络中的参数数量，而且可以更有效地学习参数，因为每个卷积矩阵都会使用图像的所有位置进行学习。其次，CNN 利用分层方法学习图像特征。卷积操作可以有效堆叠形成深度网络。CNN 中的不同层具有不同的感知范围，因此可以学习不同尺度的特征，浅层利用局部定域信息，深层学习更多全局结构。这种分层方法被证明是了一种学习图像的有效方法。

受到 CNN 的启发，ParticleNet 中采用了类似的方法来学习粒子云数据。然而，常规的卷积运算不能应用于粒子云，因为粒子云中的点可以不规则不均匀地分布，而不是像图像中的像素一样被划分为均匀的网格。因此对于粒子云结构，其卷积操作的基础，即卷积核函数如何作用于不均匀不规整的局域数据点，仍然有待定义。此外，一个通常的卷积操作，是这样的形式 $\sum_i K_i x_i$ ，这里 x_i 表示局域某个点的特征， K_i 是核函数对应该点的矩阵元。我们可以看到，这个形式在点的置换操作下是会变的（交换 x_i , x_j 而不交换 K_i , K_j ）。因此，适应于粒子云的“卷积”操作也需要修改以考虑到粒子云内的交换对称性。

边卷积 (EdgeConv) 操作^[27]被提出作为点云结构的类卷积操作。EdgeConv 首先将点云表示为图结构，其顶点 (Vertex) 是点本身，为每个点与其 k 个最近的相邻点的连线被构造为图的边 (Edge)。这样，为每个点定义了点云卷积所需的局部补丁作为与之相连的最近邻点。每个点的 EdgeConv 操作有形式

$$\mathbf{x}'_i = \bigcup_{j=1}^k \mathbf{h}_\Theta(\mathbf{x}_i, \mathbf{x}_{i_j}), \quad (4.1)$$

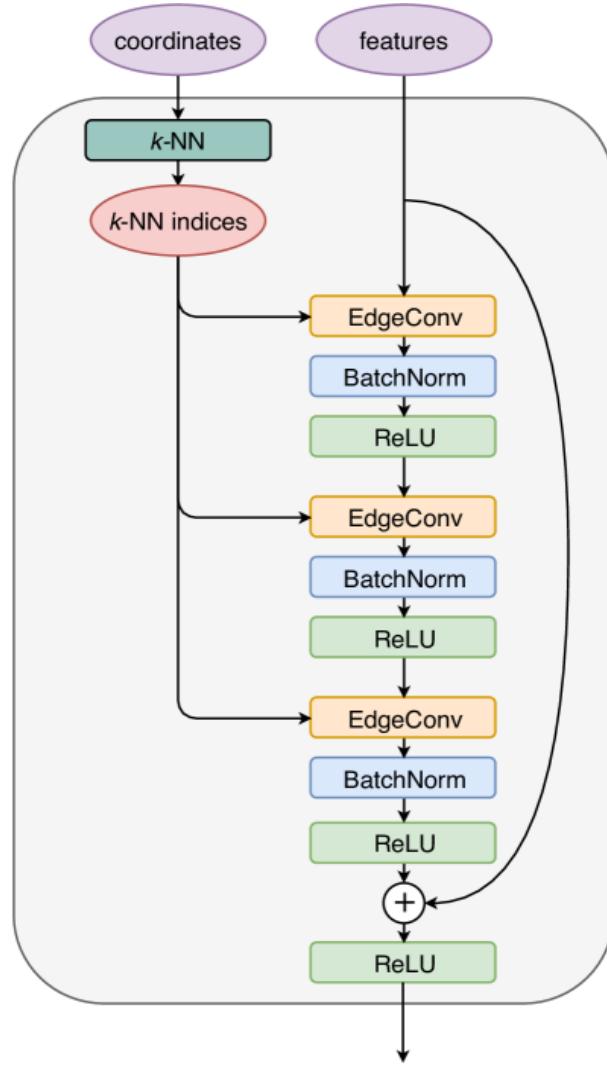
这里 $\mathbf{x}_i \in \mathbb{R}^F$ 表示点 x_i 的特征向量， $\{i_1, \dots, i_k\}$ 是点 x_i 的 k 最近邻点的索引，边函数 $\mathbf{h}_\Theta : \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}^{F'}$ 代表着一系列可被 Θ 参数化的函数，且 Θ 本身属于可学习参数。 \bigcup 是逐通道的对称聚合操作（例如 max, mean, sum 等）。并且边函数 \mathbf{h}_Θ 对于点云中的所有点都是相同的，这样和对称聚合操作 \bigcup 一起使边卷积 (EdgeConv) 成为了点云上的置换对称操作。

在 ParticleNet 模型中，遵循了以上原则，并且使用了特殊化的边函数

$$\mathbf{h}_\Theta(\mathbf{x}_i, \mathbf{x}_{i_j}) = \text{Conv}_\Theta(\mathbf{x}_i, \mathbf{x}_{i_j} - \mathbf{x}_i) = \sum_c \theta_c^a x_{i,c} + \sum_c \theta'_c (x_{i_j,c} - x_{i,c}) \quad (4.2)$$

在这里，方程(4.1)中的邻点的特征向量 \mathbf{x}_{i_j} 被 \mathbf{x}_{i_j} 与中心点的特征向量 \mathbf{x}_i 的差值所取代，并且 $Conv_{\Theta}$ 仅是常规形式下特征向量的加权和。c 是输入特征向量序列的索引，a 是核函数序列的索引。对于方程(4.1)中的对称聚合操作口，ParticleNet 采取的是平均值 $\frac{1}{k} \sum$ 。

图 4.1 边卷积 (EdgeConv) 操作的结构^[26]



EdgeConv 操作的一个重要特点是它可以很容易地堆叠起来，就像常规卷积一样。这是因为 EdgeConv 可以看作是从一个点云到另一个具有相同数目点的点云的映射（只是可能会改变每个点的特征向量的维度）。因此，就可以紧接着一个 EdgeConv 操作应用另一个 EdgeConv 操作，这使我们能够使用 EdgeConv 操作构建一个深度网络，从而分层级地学习点云的特征。

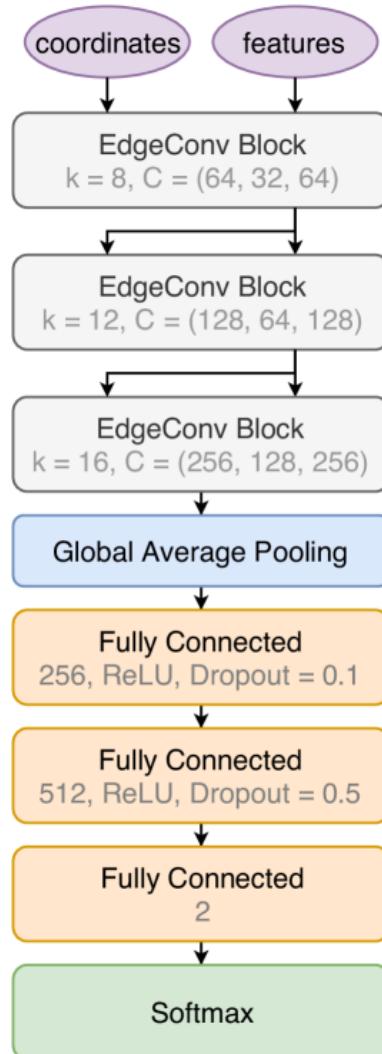
EdgeConv 操作的可堆叠性也带来了另一个有趣的可能性。EdgeConv 学习到的特征向量基本可以看作是潜在空间中原始点的新坐标，这样的话，点云之间的距离（用

于判断最近的 k 个邻点) 就可以在这个潜在空间中重新定义。换句话说, 点的接近程度可以通过 EdgeConv 操作动态地被学习到。这就得到了动态图的卷积神经网络^[28], 其中描述点云的图被动态更新以反映边的变化, 即每个点的邻点, 这也被证明比使用静态图的表现更好^[28]。

4.3 ParticleNet 网络架构

ParticleNet 架构大量使用了 EdgeConv 操作, 也采用了动态图更新方法。然而, 与原始的动态图卷积神经网络相比, ParticleNet 中做出了许多不同的设计选择, 以更好地适应喷注标记任务。

图 4.2 ParticleNet 深度网络架构^[26]



受到 ResNet^[29]网络架构的启发, ParticleNet 基于大量 EdgeConv 操作块构建, EdgeConv 块的结构如图4.1所示。EdgeConv 块首先找到每个粒子的 k 个最近相邻粒子, 然

后通过输入 EdgeConv 块的坐标计算粒子之间的距离。然后，再把 EdgeConv 操作应用在输入的粒子的特征向量上。每个 Edge 块都由多个 EdgeConv 操作组成，并且具有不同数量的卷积核。在每个 Edge 块内，描述粒子云的图是固定的，即一个粒子总是有相同的 k 个最近相邻粒子。每个 EdgeConv 操作后面都跟着一个批量归一化层^[30](batch normalization) 和一个整流激活函数^[31](ReLU)。同时，与 ResNet^[29]类似，每个 EdgeConv 块中还包含着一条与 EdgeConv 操作并行的从输入特征到最终 ReLU 层的快捷连接（如图4.1所示）。

本文使用的 ParticleNet 架构如图4.2所示。它由三个 EdgeConv 块组成。第一个 EdgeConv 块使用 $\eta - \phi$ 空间中粒子的坐标来计算距离，而随后的块使用学习到的特征向量作为坐标。对于从上到下的三个 Edge 块，最近邻的参数 k 分别取为 8、12 和 16。每个块由三个 EdgeConv 层组成。三个块的 EdgeConv 层的输出矩阵维度为 (64, 32, 64), (128, 64, 128) 和 (256, 128, 256)。在 EdgeConv 块之后，应用全局平均池化操作来聚合喷注中所有粒子学习到的特征。接下来是两个全连接层，两者都使用 ReLU 激活函数，并且分别添加概率为 0.1 和 0.5 的 dropout 层^[32]以防止过拟合。然后是仅具有 2 个单元的全连接层，后跟一个 softmax 激活函数，用于生成二分类任务的输出。

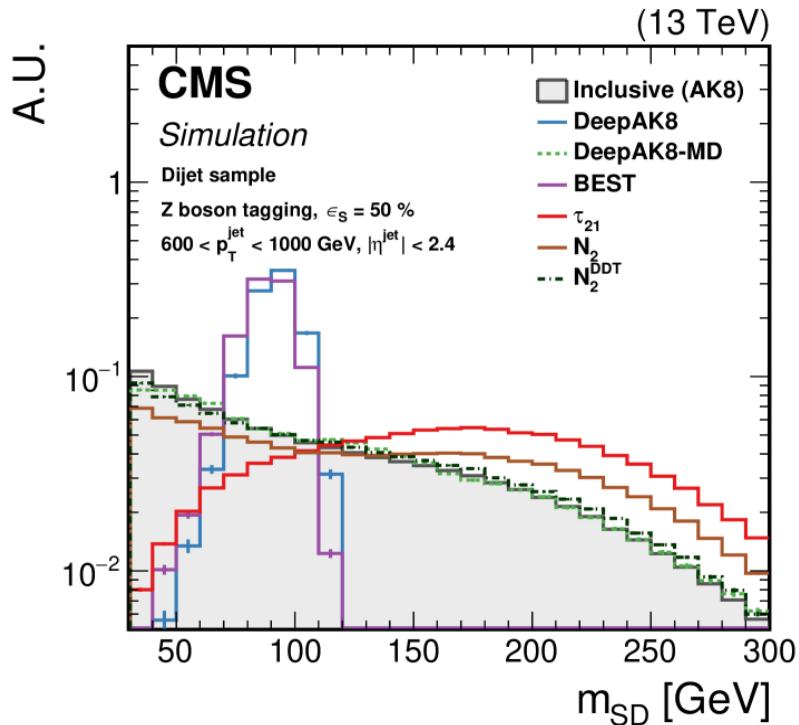
第五章 开发 $H \rightarrow WW$ 质量去相关多分类标记器

我们基于上一章的 ParticleNet 网络架构，通过修改网络输入输出，产生专用训练样本，并结合质量去相关技术，开发了 CMS 实验中首个针对大动量 $H \rightarrow WW$ 的质量去相关版本多分类标记器。

5.1 质量去相关技术

对于质量相关版本的标记器，神经网络会学习到信号样本与本底样本质量分布的差异，并且把它作为区分信号与本底的潜在判别条件，对于这样训练出来的标记器，在把本底事件误鉴别为信号事件时，会更倾向于挑选出质量分布接近信号样本的本底事件，从而对于数据中通过标记器的本底事件，在质量谱上会形成信号峰下的峰状本底，这种情况也叫做质量雕刻 (mass-sculpting)。这显然对于我们从质量谱中提取信号造成了不小的困难，因此质量去相关的标记器就显得尤为重要。

图 5.1 质量关联标记器对 QCD 本底的质量雕刻现象（与质量去相关标记器的对比）^[25]



从以上信息我们知道，当训练中的信号和本底样本有明显不同的质量分布时，训练出来的标记器在推理筛选本底样本时就会出现质量雕刻的现象。所以，如果训练样

本的信号和本底的质量分布一致，标记器则不会学习到二者的质量信息作为区分条件，从而不会在推理筛选本底样本时出现质量雕刻现象^[33]。

要实现这点，最简单直接的办法就是在训练时对信号和本底的分布进行重新加权，但是这往往还有其他隐患，例如：

- (1) 如果信号样本的质量分布峰过于集中，也就是说，在信号质量窗外的事例统计量过低，这样对信号重加权时，会导致信号窗内的高峰被极大压低到和信号窗外事件相同的高度，从而造成大量真实信号没有被选中参与训练，是一种极大的浪费和欠拟合的隐患。
- (2) 同时，对于远离信号窗的事例，也很难被重建出来，进一步加大了训练和推理的难度。

综合以上两点原因，最好的做法就是尽可能使用于训练的信号样本的质量分布尽可能平滑（避免极端事例的低统计量过度压低加权），信号窗尽可能大（避免远离信号窗事例难以重建）。所以我们的做法是：设计带有一维 m_{SD} 平分布的专用 MC 样本以供训练，可以通过产生并合并不同共振态质量的样本实现。（例如，产生变质量 m_X 的 $X \rightarrow bb$, $X \rightarrow cc$, $X \rightarrow qq$ 样本以训练通用二分叉喷注标记）

5.2 分类标签

因为我们采用的 ParticleNet 神经网络属于监督学习范畴，所以对于训练样本，我们在产生的同时还要给它们贴上正确的标签，才能使得网络学到如何区分不同标签的喷注。

对于训练样本的喷注标记，我们使用了一套专门设计的代码^[34]以产生样本和打上训练标签，在训练样本的粒子层级，利用粒子的产生信息，给粒子组成的喷注打上适当的标签，同时要求打上标签的喷注都满足 $\Delta R = 0.8$ 以产生 AK8 喷注。该产生代码的喷注标签部分见附录A.1。

5.2.1 信号分类标签

对于我们关心的 $H \rightarrow WW \rightarrow \text{anything}$ 的信号，我们主要关心两种衰变场景：一种是两个 W 都进行强子化衰变，也称作全强子衰变；另一种是一个 W 进行强子化衰变一个 W 进行轻子化衰变，也被称作半轻衰变）。所以，我们采用了以下的末态分类标签：

- **4q**: $H \rightarrow W(2q)W(2q)$, 并且重建出 4 分叉的 AK8 喷注
- **3q**: $H \rightarrow W(2q)W(2q)$, 但仅重建出 3 分叉的 AK8 喷注

- $e\nu_e qq$: H \rightarrow W($e\nu_e$)W(2q) 喷注
- $\mu\nu_\mu qq$: H \rightarrow W($\mu\nu_\mu$)W(2q) 喷注
- $\tau_e\nu_e qq$: H \rightarrow W($\tau\nu$)W(2q), τ 接着衰变为电子等产物
- $\tau_\mu\nu qq$: H \rightarrow W($\tau\nu$)W(2q), τ 接着衰变为 μ 子等产物
- $\tau_h\nu qq$: H \rightarrow W($\tau\nu$)W(2q), τ 接着进行强子化衰变

5.2.2 本底分类标签

对于我们关注的 QCD 本底喷注，我们可以按喷注中的子喷注结构对其进行分类如下：

- **QCD(bb)**: QCD 喷注中有两个 b 夸克子喷注
- **QCD(cc)**: QCD 喷注中有两个 c 夸克子喷注
- **QCD(b)**: QCD 喷注中有一个 b 夸克子喷注
- **QCD(c)**: QCD 喷注中有一个 c 夸克子喷注
- **QCD(others)**: 具有其他子结构的 QCD 喷注

5.3 数据集

5.3.1 训练集和验证集

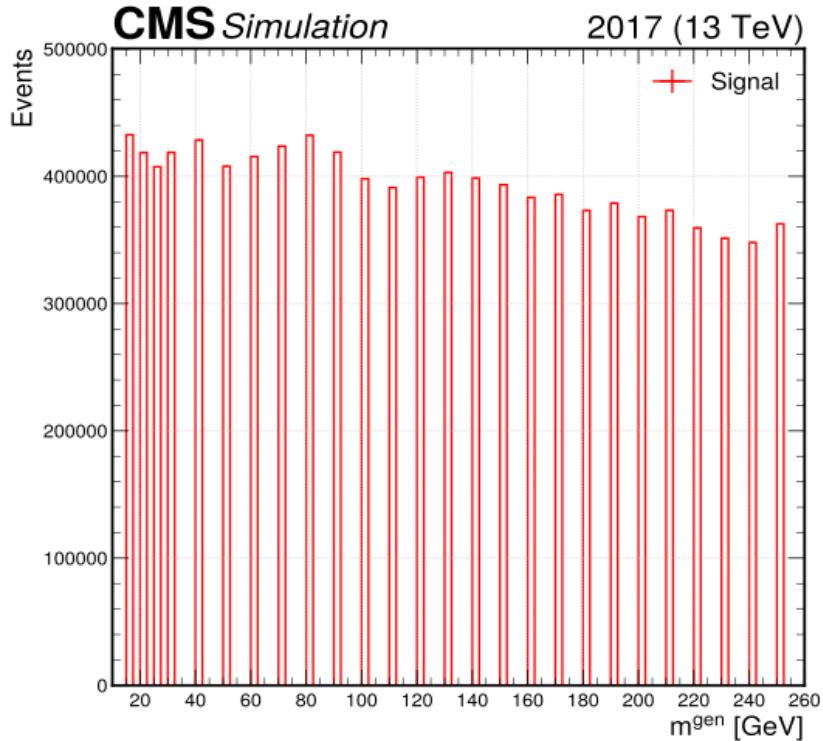
我们产生了专门设计的私人 H \rightarrow WW 信号样本和基于官方设置产生的 QCD 本底样本，同时按照 15:1 的比例把它们分成训练集和验证集供训练使用。

5.3.1.1 信号样本

因为我们的目标是开发一个质量去相关的标记器，所以我们的信号 MC 样本有以下特点：

- 样本基于 2017 ultra-legacy，总共有两千五百多万事例。
- 使用变质量 X 的 X \rightarrow WW 衰变样本，设置的产生级别质量分布区间为 $15[\text{GeV}] \leq m_X \leq 250[\text{GeV}]$ ，并保证合并起来的总信号样本具有平坦的产生级别质量分布（如图5.2所示），同时设置 W 质量为 $m_W = 80[\text{GeV}]$ 。
- 使用 JHUGen 产生子实现 X \rightarrow WW \rightarrow 4q/ $\ell\nu qq$ 衰变以更好模拟衰变产物的自旋关联。
- 喷注通过 $\Delta R = 0.8$ 的事实匹配以满足 AK8 喷注的要求。
- 喷注通过粒子级别的事实匹配打上 4q, 3q, $e\nu qq$, $\mu\nu qq$, $\tau_e\nu qq$, $\tau_\mu\nu qq$, $\tau_h\nu qq$ 七个标签之一。

图 5.2 我们产生的具有平坦质量分布的信号样本用于训练



5.3.1.2 本底样本

QCD 本底样本有以下几个特点：

- 样本基于 2017 ultra-legacy，总共有两千八百多万事例。
- 喷注通过 $\Delta R = 0.8$ 的事实匹配以满足 AK8 喷注的要求。
- 喷注通过粒子级别的事实匹配打上 QCD(bb,cc,b,c,others) 五个标签之一。

5.3.2 测试集

测试集以训练集和验证集相同的方式产生。对于测试集中的信号样本，Higgs 的产生级别质量固定在标准模型 125[GeV] 处（训练集中不包含这个质量点产生的样本），包含约四十万个事例。对于测试集中的 QCD 本底样本，以和训练集相同的方式产生，共有约五百万左右事例。

5.4 标注器设置

5.4.1 预挑选条件

我们要求进入标注器的事例满足以下条件：

- 每个事例仅由一个喷注构成，并且这个喷注通过了 Pile-up 的紧挑选条件

- 喷注的横向动量满足 $200[\text{GeV}] < p_T < 2500[\text{GeV}]$
- 喷注的 soft-drop 质量满足 $20[\text{GeV}] \leq m_{SD} < 260[\text{GeV}]$
- 对于训练样本，喷注要通过粒子级别的事实匹配满足 12 个标签分类之一

以此保证我们设计的标记器是针对大动量 $H \rightarrow WW$ 场景，并且以此通过预筛选去掉大量的无关本底，从而提高标记器的鉴别能力并降低训练速度。

5.4.2 重加权设置

训练用的信号样本和本底样本合起来共有 12 个分类标签，我们要产生质量去相关的标记器，就得对用于训练的信号和本底进行质量和横向动量的二维分区间重加权。

重加权的定义是：对于某个被重加权的标签分类，指定分布上的每个区间都要持有相同数量的事例数，并且来自每个分类的事例数占比要符合我们预定义的分类权重

现在我们要对信号和 QCD 本底样本同时在 $[p_T, m_{SD}]$ 二维分布上做重加权操作，对 soft-drop 质量 m_{SD} 的分 bin 区间为从 $20[\text{GeV}]$ 到 $260[\text{GeV}]$ 每隔 $10[\text{GeV}]$ 分一个 bin，对横向动量 p_T 的分 bin 区间为 $[200, 251, 316, 398, 501, 630, 793, 997, 1255, 1579, 1987, 2500]$ ，单位为 $[\text{GeV}]$ 。（值得注意的是，对 p_T 分 bin 按照对数等 bin 宽的选择，这是因为有对 QCD 的 p_T 分布呈指数衰减的经验分布，所以采用对数等 bin 宽可以尽可能使得分 bin 后直方图高度均匀）

定义各个分类标签定义的重加权权重时，我们把(5.2.2)中的五个 QCD 子分类都合并成 QCD 标签分类参与加权。最后得到分类权重的如下：

$$\begin{aligned} 4q : 3q : evqq : \mu\nu qq : \tau_e vqq : \tau_\mu vqq : \tau_h vqq : QCD \\ = 0.34 : 0.08 : 0.2 : 0.2 : 0.03 : 0.03 : 0.12 : 1 \end{aligned} \tag{5.1}$$

这里各个信号子分类的权重经过我们精心挑选，使得每个权重与该分类在信号中的占比接近，从而提高训练速度。

5.4.3 神经网络输入

我们针对粒子候选者（ParticleFlow candidates）和重建出的次级顶点两类对象，定义如下变量作为标记器神经网络的输入，如下表5.1所示。

表 5.1 神经网络输入^[35]

对象	变量	描述
粒子候选者	η_{rel}	相对于 AK8 喷注主轴的赝快度 $\Delta\eta$
	ϕ_{rel}	相对于 AK8 喷注主轴的方位角 $\Delta\phi$
	$\log p_T$	p_T 的对数
	$\log E$	能量的对数
	$ \eta $	赝快度的绝对值
	charge	电荷
	isEl	是否被鉴别为电子
	isMu	是否被鉴别为 μ 子
	isGamma	是否被鉴别为光子
	isChargedHad	是否被鉴别为带电强子
	isNeutralHad	是否被鉴别为中性强子
	VTX_ass	初级顶点的关联品质
	lostInnerHits	内部硅径迹器的击中数信息
	normchi2	径迹拟合的归一化 χ^2
	quality	径迹品质
	dz	纵向冲击参数：在 z 方向到初级顶点的最近距离
	dzsig	纵向冲击参数显著度
	dxy	横向冲击参数：在横切面到束流的最近距离
	dxysig	横向冲击参数显著度
	BTag η_{rel}	径迹相对 AK8 喷注主轴的赝快度 $\Delta\eta$
	BTag p_T ratio	径迹垂直 AK8 喷注主轴的分动量与合动量之比
	BTag $p_{ }$ Ratio	径迹平行 AK8 喷注主轴的分动量与合动量之比
	BTag Sip3dVal	径迹的三维正负冲击参数
	BTag Sip3dSig	径迹的三维正负冲击参数显著度
	BTag JetDistVal	径迹到 AK8 喷注主轴的最小接近距离
次级顶点	η_{rel}	相对于 AK8 喷注主轴的赝快度 $\Delta\eta$
	ϕ_{rel}	相对于 AK8 喷注主轴的方位角 $\Delta\phi$
	m_{SV}	次级顶点不变质量
	$\log p_T$	p_T 的对数
	$ \eta $	赝快度的绝对值
	N_{track}	径迹条数
	normchi2	顶点拟合的 χ^2 除以自由度
	dxy	横向飞行距离
	dxysig	横向飞行距离显著度
	d3d	三维飞行距离
	d3dsig	三维飞行距离显著度

5.4.4 神经网络输出

我们的标记器（深度神经网络）的输出为给每个喷注打上的以下 12 个不同分类的分数，分数越高代表越有可能属于这个类

$$\text{SCORE} : \left\{ \begin{array}{l} H \rightarrow WW(4q) \\ H \rightarrow WW(3q) \\ H \rightarrow WW(evqq) \\ H \rightarrow WW(\mu\nu qq) \\ H \rightarrow WW(\tau_e\nu qq) \\ H \rightarrow WW(\tau_\mu\nu qq) \\ H \rightarrow WW(\tau_h\nu qq) \\ QCD(bb) \\ QCD(cc) \\ QCD(b) \\ QCD(c) \\ QCD(others) \end{array} \right.$$

然后对于某个信号道 vs.QCD 本底的标记任务，我们可以定义判别分为：

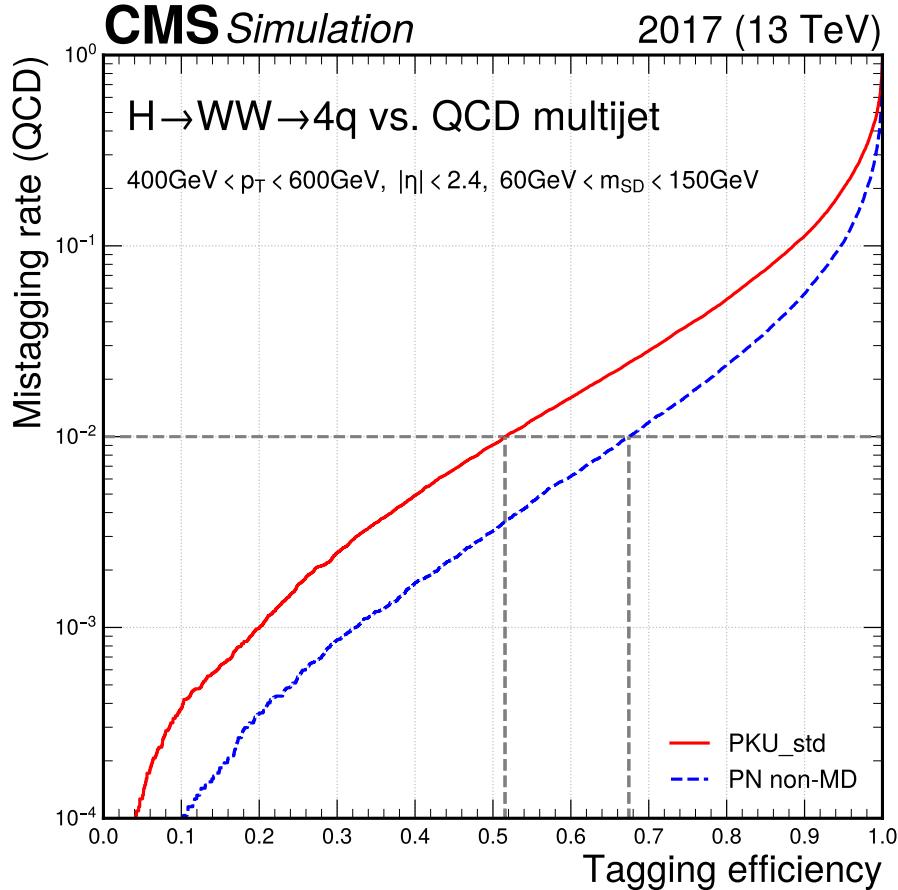
$$\text{discriminant_score} \equiv \frac{\text{score_signal}}{\text{score_signal} + \text{score_QCD}} \quad (5.2)$$

并且通过在判别分数变量上设置阈值，以达到筛选信号和本底的效果，判别分数越接近 1，说明喷注越有可能是信号喷注，反之更可能是本底喷注。

而通过扫描判别分数阈值，我们就可以获得不同的（信号标记效率，本底误标记效率），从而可以画出表示标记器表现效果的 ROC(Receiver Operating Characteristic) Curve 图。一般来说，本底误标记效率一定时，信号标记效率越大，标记器效果越好。

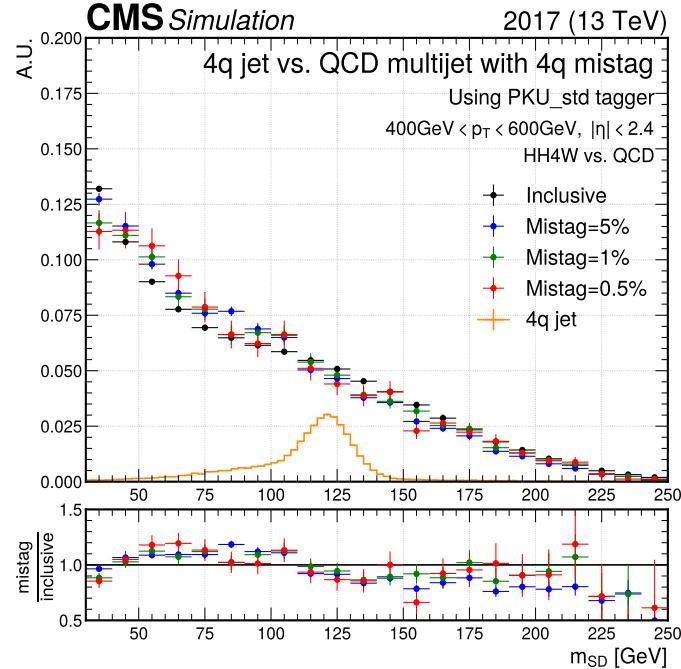
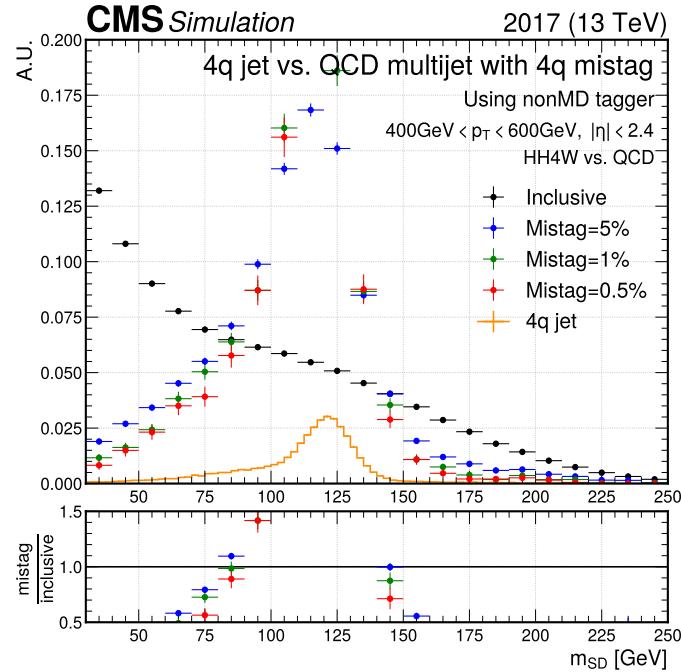
5.5 标记器在测试集上效果

同时对信号样本标记与本底样本标记的 ROC 图如下，其纵轴是 QCD 被误标记信号的效率，横轴是信号被标记的效率

图 5.3 标记器的 $H \rightarrow WW \rightarrow 4q$ 标记效果

这里 PKU_std 是我们开发的质量去相关版本 $H \rightarrow WW$ 多分类标记器，PN non-MD 是原生 ParticleNet 的质量相关版本 $H \rightarrow WW \rightarrow 4q$ 单道标记器。在 $Mistag rate = 1\%$ 时，PKU_std 的 Tagging efficiency $\approx 52\%$ ，PN non-MD 的 Tagging efficiency $\approx 68\%$ 。虽然 PKU_std 比 PN non-MD 的效果要稍微差一些，但这正是质量去相关标记器的代价，换来的是被误鉴别的 QCD 本底没有接近信号峰的质量雕刻。如图5.4所示。

从图5.4可以看到，我们开发的质量去相关标记器 PKU_std 的质量去相关效果非常好，在 4q 喷注的信号峰附近，QCD 本底仍然保持了它原本的分布，没有形成类似 4q 喷注的质量峰分布，从而有利于我们在实验数据中提取 4q 信号。而被 PN non-MD 误标记的 QCD 本底就有非常明显的质量雕刻现象，如图5.5所示。

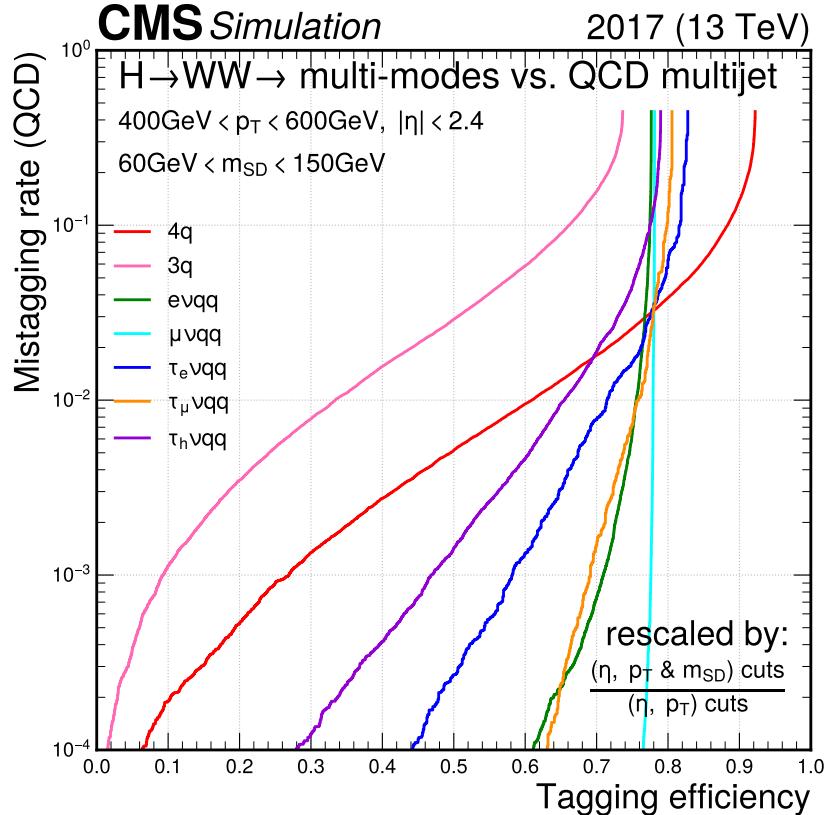
图 5.4 被 PKU_std 误标记的 QCD 本底的 m_{SD} 分布

 图 5.5 被 PN nonMD 误标记的 QCD 本底的 m_{SD} 分布


比较以上两图我们可以看到，虽然 PKU_std 比 PN non-MD 的效果要稍微差一些，但这正是质量去相关标记器的代价，以效率上 20% 的损失换来了从无到有的质量去相关效果，极大地降低了分析难度。

除此之外，原有的 PN non-MD 只是 $H \rightarrow W(2q)W(2q)$ 的单衰变道标记器，但我们

开发的标记器是 $H \rightarrow WW$ 的多衰变道标记器，所以在其他 $H \rightarrow WW$ 衰变道上也有用武之地。

图 5.6 我们开发的 $H \rightarrow WW$ 标记器的对多个衰变道的标记性能

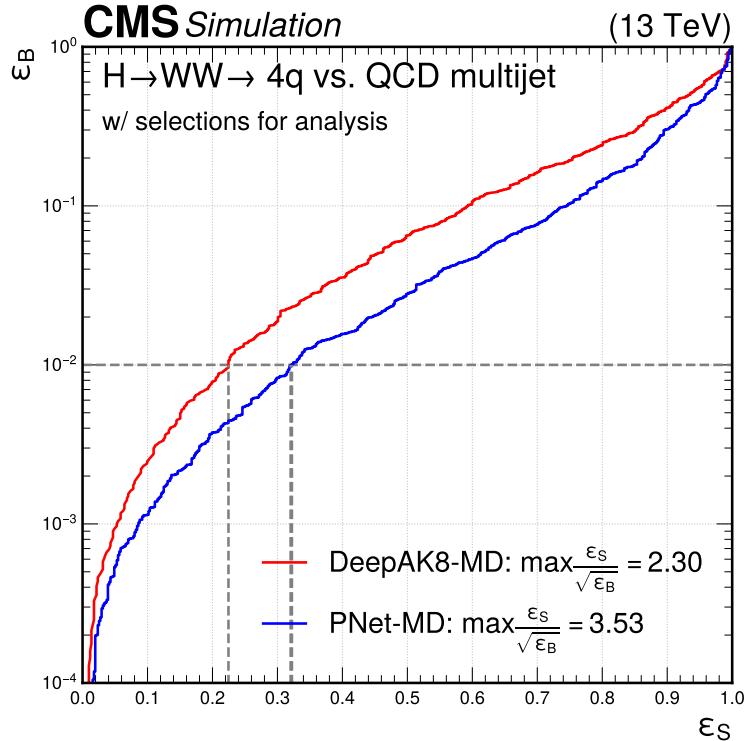


从这个图里我们可以看到，在 Mistag rate=1% 时，除了 3q 末态（指 $H \rightarrow WW$ 衰变到四个夸克但只有三个被重建在 AK8 喷注中）的标记效率只有 33% 左右，其他几个道的标记效率都在 60%~80% 之间，更让我们对开发的标记器充满信心。

5.6 在分析中的初步应用效果和前景

我们目前已经将开发的标记器投入正式的 CMS 实验分析中使用，目前已经初步应用的分析场景为：胶子聚合 (ggF) 产生大动量希格斯粒子，再到 WW 衰变的过程。其费曼图如图2.4左上所示。

在此分析中，我们把我们开发的基于 ParticleNet-MD 的标记器和目前官方分析中针对 AK8 喷注常用的 DeepAK8-MD 标记器进行了比较。比较结果如下图5.7所示

图 5.7 开发的 PNet-MD 标记器与 DeepAK8 标记器在 H \rightarrow WW \rightarrow 4q 道的标记表现比较

可以看到，我们开发的标记器与以往官方使用的标记器 DeepAK8 相比，在 g+g \rightarrow H \rightarrow WW 标记任务上有接近 50% 的性能提升，并且随着日后更多的研究和采用更大规模网络模型，标记器的表现还有更大的提升空间。

除此之外，我们的 H \rightarrow WW 标记器还有更多值得期待的分析应用场景，包括：

- 标记 VBF、WH 等大动量希格斯粒子产生过程：以研究大 p_T 希格斯粒子的物理性质和间接搜寻高能标的超标准模型物理迹象。
- HH \rightarrow WWWW 和 HH \rightarrow bbWW 过程：可以在这样的过程中测量希格斯粒子自耦合强度，同时搜寻双希格斯共振态。
- 对暗物质的搜寻：因为我们开发的标记器的输入不包括中微子信息，所以设计场景中存在消失的横向动量 (MET)，可以借此搜寻 H+MET (Dark Matter) \rightarrow WW 的暗物质过程。

第六章 总结和展望

本文从标准模型出发，先回顾了标准模型的基本内容，包括粒子组成和相互作用。然后结合近年来超出标准模型预测的实验结果（ μ 子 g-2 实验，超重的 W 玻色子），根据大型强子对撞机（LHC）上的紧凑谬子螺线管（CMS）实验优势（对 μ 子探测的优越性），提出了对标准模型检验和新物理搜寻的一条关键路径：对大动量希格斯粒子的 WW 衰变测量以及 X→WW 共振态的搜寻。

在大动量下，希格斯粒子的衰变产物会形成合并喷注，后续衰变出的夸克/轻子喷注也会形成合并子喷注，从而具有独特的相空间结构，是研究标准模型高阶修正项和开发新标记技术的极佳场景。而 X→WW 共振态的搜寻也能为 V_{kk} 玻色子、复合希格斯粒子、额外维算符等新物理模型提供存在证据或给出存在上限。二者可通过质量去相关的 H→WW 标记技术有机结合起来，达到事半功倍的效果。同时还可以借鉴利用 $W_{kk} \rightarrow RW \rightarrow WWW$ 三玻色子过程的分析技术，加速得到具有显著性的结果。

在 CMS 实验 RUN2 到 RUN3 的阶段背景和高亮度 LHC 的前瞻背景下，大动量希格斯粒子作为尚未得到完全开发的研究领域具有十分明朗的研究前景，而 X→WW 共振态的搜寻随着高亮度数据量的大幅上升也能给出更具有显著度的结果。

接着，本文落脚到 CMS 实验的重建和标记技术上，回顾了目前官方重建事例时缓解顶点堆积的 PUPPI 算法和用于重建喷注的 anti-kT 算法。在此基础上，以过往喷注标记技术作为对比（从基于理论的高级变量算法到基于机器学习的高级变量算法再到目前基于深度学习的初级变量算法），指出了新型图神经网络技术 ParticleNet 用于喷注标记的物理优越性和更好的标记表现，并且进一步剖析了其网络架构。

基于 ParticleNet，通过修改网络输入输出，微调权重，设计专门训练样本，我们开发了针对大动量希格斯粒子到 WW 衰变场景的多分类标记器，对 H→WW 的全强子道末态和一轻子道末态进行标记。并且是利用质量去相关技术开发的质量去相关标记器，有两方面的巨大优势：1. 不会在 QCD 本底上雕刻出信号峰，方便后续分析；2. 对质量无关的希格斯标记器可进一步用于 X→WW 共振态的标记上，以搜寻新物理。

最后，通过在分析上的初步应用于对比，我们开发的 ParticleNet-MD 标记器在标记性能优于目前传统的 DeepAK8-MD 标记器近 50%，是一个巨大的提高，并且希冀未来在更大规模的神经网络架构上，开发出针对更多衰变道分类的通用标记器，以尽可能实现 H→WW 乃至大动量希格斯粒子的通用标记任务，将会是深度学习技术在高能物理领域的一次成功而富有意义的应用，从而更好更快地推动高能物理领域前沿发展。

参考文献

- [1] ABI B, ALBAHRI T, AL-KILANI S, et al. Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm[J/OL]. Phys. Rev. Lett., 2021, 126: 141801. <https://link.aps.org/doi/10.1103/PhysRevLett.126.141801>. DOI: 10.1103/PhysRevLett.126.141801.
- [2] AALTONEN T, AMERIO S, AMIDEI D, et al. High-precision measurement of the W boson mass with the CDF II detector[J]. Science, 2022, 376(6589): 170-176. DOI: 10.1126/science.abk1781.
- [3] PERL M L, ABRAMS G S, BOYARSKI A M, et al. Evidence for Anomalous Lepton Production in $e^+ - e^-$ Annihilation[J/OL]. Phys. Rev. Lett., 1975, 35: 1489-1492. <https://link.aps.org/doi/10.1103/PhysRevLett.35.1489>. DOI: 10.1103/PhysRevLett.35.1489.
- [4] ABACHI S, ABBOTT B, ABOLINS M, et al. Observation of the Top Quark[J/OL]. Phys. Rev. Lett., 1995, 74: 2632-2637. <https://link.aps.org/doi/10.1103/PhysRevLett.74.2632>. DOI: 10.1103/PhysRevLett.74.2632.
- [5] CHATRCHYAN S, KHACHATRYAN V, SIRUNYAN A M, et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC[J/OL]. Physics Letters B, 2012, 716(1): 30-61. <https://www.sciencedirect.com/science/article/pii/S0370269312008581>. DOI: <https://doi.org/10.1016/j.physletb.2012.08.021>.
- [6] AAIJ R, BETETA C A, ACKERNLEY T, et al. Test of lepton universality in beauty-quark decays[J]. Nature Physics, 2022, 18(3): 277-282. DOI: 10.1038/s41567-021-01478-8.
- [7] New data strengthens RK flavour anomaly –[EB/OL]. 2022. <https://cerncourier.com/a/new-data-strengthens-rk-flavour-anomaly/>.
- [8] AAIJ R, ABELLÁN BETETA C, ADEVA B, et al. Search for Lepton-Universality Violation in $B^+ \rightarrow K^+ \ell^+ \ell^-$ Decays[J/OL]. Phys. Rev. Lett., 2019, 122: 191801. <https://link.aps.org/doi/10.1103/PhysRevLett.122.191801>. DOI: 10.1103/PhysRevLett.122.191801.
- [9] AAIJ R, ADEVA B, ADINOLFI M, et al. Test of lepton universality with $B \bar{0} \rightarrow K^* \ell^+ \ell^-$ decays[J]. Journal of High Energy Physics, 2017, 2017(8). DOI: 10.1007/jhep08(2017)055.
- [10] M . First results from Fermilab’s Muon g-2 experiment strengthen evidence of new physics[EB/OL]. (2021-06-29). <https://news.fnal.gov/2021/04/first-results-from-fermilabs-muon-g-2-experiment-strengthen-evidence-of-new-physics/>.
- [11] L . CDF collaboration at Fermilab announces most precise ever measurement of W boson mass to be in tension with the Standard Model[EB/OL]. (2022-04-07). <https://news.fnal.gov/2022/04/cdf-collaboration-at-fermilab-announces-most-precise-ever-measurement-of-w-boson-mass/>.
- [12] BROECK V . THE CERN ACCELERATOR COMPLEX[EB/OL]. (2019-09-19). <https://cds.cern.ch/record/2693837/>.
- [13] SAKUMA T . Cutaway diagrams of CMS detector[EB/OL]. (2019-05-03). <https://cds.cern.ch/record/2665537/>.
- [14] FLORIAN D D, FOR O. Handbook of LHC Higgs cross sections : 4. Deciphering the nature of the Higgs sector[M]. [S.I.]: Cern, 2017.

- [15] BATTAGLIA M, GRAZZINI M, SPIRA M, et al. Sensitivity to BSM effects in the Higgs pT spectrum within SMEFT[J]. *Journal of High Energy Physics*, 2021, 2021(11). DOI: 10.1007/jhep11(2021)173.
- [16] CAMARGO D A, CAMPOS M D, de MELO T B, et al. A two Higgs doublet model for dark matter and neutrino masses[J]. *Physics Letters B*, 2019, 795: 319-326. DOI: 10.1016/j.physletb.2019.06.020.
- [17] KUDASHKIN K, LINDERT J M, MELNIKOV K, et al. Higgs bosons with large transverse momentum at the LHC[J]. *Physics Letters B*, 2018, 782: 210-214. DOI: 10.1016/j.physletb.2018.05.009.
- [18] GRAZZINI M, ILNICKA A, SPIRA M. Higgs boson production at large transverse momentum within the SMEFT: analytical results[J]. *The European Physical Journal C*, 2018, 78(10). DOI: 10.1140/epjc/s10052-018-6261-7.
- [19] ALISON J. *The Road to Discovery: Detector Alignment, Electron Identification, Particle Misidentification, WW Physics, and the Discovery of the Higgs Boson* (Springer Theses)[M]. Softcover reprint of the original 1st ed. 2015. [S.l.]: Springer, 2016.
- [20] KILMINSTER B. Boosting the Higgs boson[EB/OL]. 2018. <https://indico.cern.ch/event/732102/contributions/3092580/attachments/1759641/2854473/Higgs-Couplings-boosted-v4.pdf>.
- [21] AGASHE K, COLLINS J H, DU P, et al. Detecting a boosted diboson resonance[J]. *Journal of High Energy Physics*, 2018, 2018(11). DOI: 10.1007/jhep11(2018)027.
- [22] BERTOLINI D, HARRIS P, LOW M, et al. Pileup per particle identification[J]. *Journal of High Energy Physics*, 2014, 2014(10). DOI: 10.1007/jhep10(2014)059.
- [23] How CMS weeds out particles that pile up | CMS Experiment[EB/OL]. 2018. <https://cms.cern/news/how-cms-weeds-out-particles-pile>.
- [24] CACCIARI M, SALAM G P, SOYEZ G. The anti-kt jet clustering algorithm[J]. *Journal of High Energy Physics*, 2008, 2008(04): 063-063. DOI: 10.1088/1126-6708/2008/04/063.
- [25] SIRUNYAN A, TUMASYAN A, ADAM W, et al. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques[J/OL]. *Journal of Instrumentation*, 2020, 15(06): P06005-P06005. <https://doi.org/10.1088/1748-0221/15/06/p06005>. DOI: 10.1088/1748-0221/15/06/p06005.
- [26] QU H, GOUSKOS L. Jet tagging via particle clouds[J]. *Physical Review D*, 2020, 101(5). DOI: 10.1103/physrevd.101.056019.
- [27] WANG Y, SUN Y, LIU Z, et al. Dynamic Graph CNN for Learning on Point Clouds[J/OL]. *ACM Trans. Graph.*, 2019, 38(5). <https://doi.org/10.1145/3326362>. DOI: 10.1145/3326362.
- [28] WANG Y, SUN Y, LIU Z, et al. Dynamic Graph CNN for Learning on Point Clouds[J]. *ACM Transactions on Graphics*, 2019, 38(5): 1-12. DOI: 10.1145/3326362.
- [29] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[EB/OL]. arXiv. 2015. <https://arxiv.org/abs/1512.03385>.
- [30] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[EB/OL]. arXiv. 2015. <https://arxiv.org/abs/1502.03167>.
- [31] GLOROT X, BORDES A, BENGIO Y. Deep Sparse Rectifier Neural Networks[C/OL]//GORDON G, DUNSON D, DUDÍK M. *Proceedings of Machine Learning Research: Proceedings of the Four-*

- teenth International Conference on Artificial Intelligence and Statistics: vol. 15. Fort Lauderdale, FL, USA: PMLR, 2011: 315-323. <https://proceedings.mlr.press/v15/glorot11a.html>.
- [32] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J/OL]. Journal of Machine Learning Research, 2014, 15(56): 1929-1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
- [33] Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques[J/OL], 2020. <http://cds.cern.ch/record/2707946>.
- [34] LI C, QU H. GitHub - colizz/DNNTuples at dev-UL-hww[EB/OL]. 2022. <https://github.com/colizz/DNNTuples/tree/dev-UL-hww>.
- [35] CERN Open Data Portal[EB/OL]. 2019. <http://opendata.cern.ch/record/12102>.

附录 A 关键代码

A.1 训练样本的喷注标签部分代码

来自 DNNTuples^[34]/FatJetHelpers/src/FatJetMatching.cc 文件

```

1 #include "DeepNTuples/FatJetHelpers/interface/FatJetMatching.h"
2
3 #include <unordered_set>
4 #include "TString.h"
5 #include "DataFormats/HepMCandidate/interface/GenParticle.h"
6
7 using namespace deepntuples;
8
9
10 std::pair<FatJetMatching::FatJetLabel, const reco::GenParticle*>
11   ↵ FatJetMatching::higgs_label(const pat::Jet* jet, const reco::GenParticle *parton, double
12   ↵ distR)
13 {
14
15     auto higgs = getFinal(parton);
16
17     if (debug_){
18         using namespace std;
19         cout << "jet: " << jet->polarP4() << endl;
20         cout << "H: " ; printGenParticleInfo(higgs, -1);
21     }
22
23     bool is_hVV = false;
24     if (higgs->numberOfDaughters() >= 3) {
25         // e.g., h->Vqq or h->qqqq
26         is_hVV = true;
27     }else {
28         // e.g., h->VV*
29         for (const auto &p : higgs->daughterRefVector()){
30             auto pdgid = std::abs(p->pdgId());
31             if (pdgid == ParticleID::p_Wplus || pdgid == ParticleID::p_Z0){
32                 is_hVV = true;
33                 break;
34             }
35         }
36     }
37
38     if (is_hVV){
39         // h->WW or h->ZZ
40         std::vector<const reco::GenParticle*> hVV_daus;
41         int n_el = 0, n_mu = 0, n_tau = 0, n_quarks = 0;
42         bool is_lhe_hvv = false;

```

```

41     for (unsigned idau=0; idau<higgs->numberOfDaughters(); ++idau){
42         const auto *dau = dynamic_cast<const reco::GenParticle*>(higgs->daughter(idau));
43         auto pdgid = std::abs(higgs->daughter(idau)->pdgId());
44         if (pdgid >= ParticleID::p_d && pdgid <= ParticleID::p_b) {
45             ++n_quarks;
46             hVV_daus.push_back(dau);
47         }else if (pdgid >= ParticleID::p_eminus && pdgid <= ParticleID::p_nu_tau){
48             if (pdgid == ParticleID::p_eminus) ++n_el;
49             else if (pdgid == ParticleID::p_muminus) ++n_mu;
50             else if (pdgid == ParticleID::p_tauminus) ++n_tau;
51             hVV_daus.push_back(dau);
52         }else{
53             // const auto d = getDaughterQuarks(getFinal(dau));
54             // hVV_daus.insert(hVV_daus.end(), d.begin(), d.end());
55             is_lhe_hvv = true;
56             auto daufinal = getFinal(dau);
57             for (unsigned j=0; j<daufinal->numberOfDaughters(); ++j){
58                 const auto *ddau = dynamic_cast<const reco::GenParticle*>(daufinal->daughter(j));
59                 auto dpdgid = std::abs(ddau->pdgId());
60                 if (dpdgid >= ParticleID::p_d && dpdgid <= ParticleID::p_b){
61                     ++n_quarks;
62                     hVV_daus.push_back(ddau);
63                 }else if (dpdgid >= ParticleID::p_eminus && dpdgid <= ParticleID::p_nu_tau){
64                     if (dpdgid == ParticleID::p_eminus) ++n_el;
65                     else if (dpdgid == ParticleID::p_muminus) ++n_mu;
66                     else if (dpdgid == ParticleID::p_tauminus) ++n_tau;
67                     hVV_daus.push_back(ddau);
68                 }
69             }
70         }
71     }
72     auto is_neutrino = [&](int pdgid){
73         pdgid = std::abs(pdgid);
74         return pdgid == ParticleID::p_nu_e || pdgid == ParticleID::p_nu_mu || pdgid ==
75             ParticleID::p_nu_tau;
76     };
76     // auto printDaus = [&](std::vector<const reco::GenParticle*>& parts, const pat::Jet*
77     // → jet){
78     //     using namespace std;
79     //     cout << "Found " << parts.size() << " quarks from Higgs decay" << endl;
80     //     for (const auto * gp : parts){
81     //         using namespace std;
82     //         printGenParticleInfo(gp, -1);
83     //         cout << "... dR(q, jet) = " << reco::deltaR(*gp, *jet) << endl;
84     //     }
85     // };
85     // if four daughters are all quarks
86     if (n_quarks == 4){
87         int n_daus_in_jet = 0, idx = 0;
88         std::vector<int> idx_daus_in_jet;
89         for (const auto *gp : hVV_daus){

```

```

90     auto dr = reco::deltaR(*gp, *jet);
91     if (dr < distR){
92         ++n_daus_in_jet;
93         idx_daus_in_jet.push_back(idx);
94     }
95     ++idx;
96 }
97 if (n_daus_in_jet >= 4){
98     // std::cout << "4q!" << std::endl;
99     return std::make_pair(FatJetLabel::H_ww4q, higgs);
100 }
101 else if (n_daus_in_jet >= 3){
102     // std::cout << "3q!" << std::endl;
103     return std::make_pair(FatJetLabel::H_ww3q, higgs);
104 }
105 else if (n_daus_in_jet >= 2 && is_lhe_hvv){
106     if (idx_daus_in_jet == std::vector<int>({0,1}) || idx_daus_in_jet ==
107         std::vector<int>({2,3})){
108         // std::cout << "2q from same W!" << std::endl;
109         return std::make_pair(FatJetLabel::H_ww2qsame, higgs);
110     }else {
111         // std::cout << "2q from separate Ws!" << std::endl;
112         return std::make_pair(FatJetLabel::H_ww2qsep, higgs);
113     }
114 }else if (n_quarks == 2 && n_tau == 0) { // 1vqq
115     int n_daus_in_jet = 0;
116     for (const auto *gp : hVV_daus){
117         if (is_neutrino(gp->pdgId())){
118             continue;
119         }
120         auto dr = reco::deltaR(*gp, *jet);
121         if (dr < distR){
122             ++n_daus_in_jet;
123         }
124     }
125     if (n_daus_in_jet >= 3){
126         // std::cout << "1vqq!" << std::endl;
127         if (n_el > 0){
128             return std::make_pair(FatJetLabel::H_wwevqq, higgs);
129         }else if (n_mu > 0){
130             return std::make_pair(FatJetLabel::H_wwmqq, higgs);
131         }
132     }
133 }else if (n_quarks == 2 && n_tau == 1) { // tauvqq
134     int tau_pos = 0;
135     int n_daus_in_jet = 0;
136     for(std::size_t igp = 0; igp < hVV_daus.size(); ++igp) {
137         auto gp = hVV_daus[igp];
138         if (is_neutrino(gp->pdgId())){
139             continue;

```

```

140 }
141 if (std::abs(gp->pdgId()) == ParticleID::p_tauminus){
142     tau_pos = 1;
143     continue;
144 }
145 auto dr = reco::deltaR(*gp, *jet);
146 if (dr < distR){
147     ++n_daus_in_jet;
148 }
149 }
150 auto tau = getFinal(hVV_daus[tau_pos]);
151 if (n_daus_in_jet >= 2){ // both two quarks are in
152     bool is_leptau = false;
153     for (unsigned j=0; j<tau->numberOfDaughters(); ++j){
154         const auto *tdau = dynamic_cast<const reco::GenParticle*>(tau->daughter(j));
155         auto tdaupdgid = std::abs(tdaupdgid);
156         if (tdaupdgid == ParticleID::p_eminus || tdaupdgid == ParticleID::p_muminus){
157             is_leptau = true;
158             auto dr = reco::deltaR(*tdau, *jet);
159             if (dr < distR){
160                 // std::cout << "leptau!" << std::endl;
161                 if (tdaupdgid == ParticleID::p_eminus){
162                     return std::make_pair(FatJetLabel::H_wwleptauevqq, higgs);
163                 }else if (tdaupdgid == ParticleID::p_muminus){
164                     return std::make_pair(FatJetLabel::H_wwleptaumvqq, higgs);
165                 }
166             }
167         }
168     }
169     if (!is_leptau){ // hadronic taus
170         auto dr = reco::deltaR(*tau, *jet);
171         if (dr < distR){
172             // std::cout << "hadtau!" << std::endl;
173             return std::make_pair(FatJetLabel::H_wwhadtauvqq, higgs);
174         }
175     }
176 }
177 }else {
178     // std::cout << "***undefined!" << std::endl;
179     return std::make_pair(FatJetLabel::Invalid, higgs);
180     // throw std::logic_error("[FatJetMatching::higgs_label] Illegal H->WW mode");
181 }
182 // std::cout << "***unmatch!" << std::endl;
183 return std::make_pair(FatJetLabel::Invalid, higgs);

184
185 // if (debug_){
186 //     using namespace std;
187 //     cout << "Found " << hVV_daus.size() << " quarks from Higgs decay" << endl;
188 //     for (const auto * gp : hVV_daus){
189 //         using namespace std;
190 //         printGenParticleInfo(gp, -1);

```

```

191     //     cout << " ... dR(q, jet) = " << reco::deltaR(*gp, *jet) << endl;
192     //
193     //
194
195     // unsigned n_quarks_in_jet = 0;
196     // for (const auto *gp : hVV_daus){
197     //     auto dr = reco::deltaR(*gp, *jet);
198     //     if (dr < distR){
199     //         ++n_quarks_in_jet;
200     //     }
201     // }
202     // if (n_quarks_in_jet >= 4){
203     //     return std::make_pair(FatJetLabel::H_qqqq, higgs);
204     // }
205
206 }else if (isHadronic(higgs)) {
207     // direct h->qq
208
209     auto hdaus = getDaughterQuarks(higgs);
210     if (hdaus.size() < 2) throw std::logic_error("[FatJetMatching::higgs_label] Higgs decay
211     ↪ has less than 2 quarks!");
212     // if (zdaus.size() >= 2)
213     {
214         double dr_q1    = reco::deltaR(jet->p4(), hdaus.at(0)->p4());
215         double dr_q2    = reco::deltaR(jet->p4(), hdaus.at(1)->p4());
216         if (dr_q1 > dr_q2){
217             // swap q1 and q2 so that dr_q1<=dr_q2
218             std::swap(dr_q1, dr_q2);
219             std::swap(hdaus.at(0), hdaus.at(1));
220         }
221         auto pdgid_q1 = std::abs(hdaus.at(0)->pdgId());
222         auto pdgid_q2 = std::abs(hdaus.at(1)->pdgId());
223
224         if (debug_){
225             using namespace std;
226             cout << "deltaR(jet, q1)    : " << dr_q1 << endl;
227             cout << "deltaR(jet, q2)    : " << dr_q2 << endl;
228             cout << "pdgid(q1)        : " << pdgid_q1 << endl;
229             cout << "pdgid(q2)        : " << pdgid_q2 << endl;
230         }
231
232         if (dr_q1<distR && dr_q2<distR){
233             if (pdgid_q1 == ParticleID::p_b && pdgid_q2 == ParticleID::p_b) {
234                 return std::make_pair(FatJetLabel::H_bb, higgs);
235             }else if (pdgid_q1 == ParticleID::p_c && pdgid_q2 == ParticleID::p_c) {
236                 return std::make_pair(FatJetLabel::H_cc, higgs);
237             }else {
238                 return std::make_pair(FatJetLabel::H_qq, higgs);
239             }
240         }
241     }
242 }
```

```

241 }else {
242     // test h->tautau
243     std::vector<const reco::GenParticle*> taus;
244     for (unsigned i=0; i<higgs->numberOfDaughters(); ++i){
245         const auto *dau = dynamic_cast<const reco::GenParticle*>(higgs->daughter(i));
246         if (std::abs(dau->pdgId()) == ParticleID::p_tauminus){
247             taus.push_back(dau);
248         }
249     }
250     if (taus.size() == 2){
251         // higgs -> tautau
252         // use first version or last version of the tau in dr?
253         double dr_tau1    = reco::deltaR(jet->p4(), taus.at(0)->p4());
254         double dr_tau2    = reco::deltaR(jet->p4(), taus.at(1)->p4());
255
256         if (debug_){
257             using namespace std;
258             cout << "deltaR(jet, tau1) : " << dr_tau1 << endl;
259             cout << "deltaR(jet, tau2) : " << dr_tau2 << endl;
260         }
261
262         auto isHadronicTau = [] (const reco::GenParticle* tau){
263             for (const auto &dau : tau->daughterRefVector()){
264                 auto pdgid = std::abs(dau->pdgId());
265                 if (pdgid==ParticleID::p_eminus || pdgid==ParticleID::p_muminus){
266                     return false;
267                 }
268             }
269             return true;
270         };
271
272         auto tau1 = getFinal(taus.at(0));
273         auto tau2 = getFinal(taus.at(1));
274         if (dr_tau1<distR && dr_tau2<distR){
275             if (isHadronicTau(tau1) && isHadronicTau(tau2)) {
276                 return std::make_pair(FatJetLabel::H_tautau, higgs);
277             }
278         }
279     }
280 }
281
282 return std::make_pair(FatJetLabel::Invalid, nullptr);
283
284 }
285
286
287 std::pair<FatJetMatching::FatJetLabel, const reco::GenParticle*>
288     ↵ FatJetMatching::qcd_label(const pat::Jet* jet, const reco::GenParticleCollection&
289     ↵ genParticles, double distR)
290 {

```

```

290 const reco::GenParticle *parton = nullptr;
291 double minDR = 999;
292 for (const auto &gp : genParticles){
293     if (gp.status() != 23) continue;
294     auto pdgid = std::abs(gp.pdgId());
295     if (!(pdgid<ParticleID::p_t || pdgid==ParticleID::p_g)) continue;
296     auto dr = reco::deltaR(gp, *jet);
297     if (dr<distR && dr<minDR){
298         minDR = dr;
299         parton = &gp;
300     }
301 }
302 if (debug_){
303     using namespace std;
304     if (parton){
305         cout << "parton"; printGenParticleInfo(parton, -1);
306         cout << "dr(jet, parton): " << minDR << endl;
307     }
308 }
309
310 auto n_bHadrons = jet->jetFlavourInfo().getbHadrons().size();
311 auto n_cHadrons = jet->jetFlavourInfo().getcHadrons().size();
312
313 if (n_bHadrons>=2) {
314     return std::make_pair(FatJetLabel::QCD_bb, parton);
315 }else if (n_bHadrons==1){
316     return std::make_pair(FatJetLabel::QCD_b, parton);
317 }else if (n_cHadrons>=2){
318     return std::make_pair(FatJetLabel::QCD_cc, parton);
319 }else if (n_cHadrons==1){
320     return std::make_pair(FatJetLabel::QCD_c, parton);
321 }
322
323 return std::make_pair(FatJetLabel::QCD_others, parton);
324 }

```


致谢

以这篇致谢作为对我本科四年的总结。我自己是通过竞赛降分进入的北大，高考也没有同过北大的裸分线，所以一开始进北大也不在物理学院，而是被调剂到了工学院。因为竞赛的失利，所以高三末期到刚进北大，也没有特别大的愿望要来到物理学院。

大一在工学院的时候，一开始也没有特别宏大的目标，只是正常地学习生活玩耍，同时在因为工学院的部分专业要求，我在大一还修了数学分析和高等代数，在大一并不算忙碌的时间里，这两门纯粹的数学课让我遨游在思考数学的海洋里，而不用关心现实中繁杂的事务，至今回想起来仍然觉得十分美好，并且特别感谢教我数学分析的史一鹏老师，他是一位十分善良和蔼关心学生的老师，为我在北大碰到的教师群体定下了主色调。同时，由于工学院还要修适当的物理课，我在旁听了一门工学院开的物理课后，觉得还是理科学院系开的物理课更对我的胃口，便决定选了物理学院的力学课，这门课的老师是孟策老师，他是一位特别爱笑、心理年龄十分年轻的中年教师，至今我仍然常常回忆起他坐在转椅上拿着杠铃，一边自我旋转一边向我们讲授角动量的概念，让人忍俊不禁。

到了大一的下学期，我依然没有强烈的要去哪个方向学习/转专业的想法，只是忽然看见转专业通知，便又想起了我的老本行：物理。“要不要到物理学院去看看呢？可以先报个名，之后再慢慢做决定吧”，这就是我当时的想法，于是我顺利通过了转专业的笔试和面试，记得面试我的是王稼军奶奶，刘树新老师（留着哈利波特中斯内普教授一样的发型）以及实验中心的张朝晖老师。他们问我以后到物理学院想来学什么方向，做实验还是理论，我说我觉得实验理论都可以吧（但其实那时候我知道我实验动手能力很差），方向还没有想好，但是想做和量子有关的东西（因为那时候看新闻上的量子计算量子通信特别火热）。刘树新老师笑了笑：“物理什么东西和量子无关呀？”

于是，对于我这种选择困难症来说，最后也就在大二开学时顺其自然地来到了物理学院，和我一起从工学院转来的有谭奕和许安冬，二者后来都和我成为了很好的朋友。这里补充一个有趣的事，大一下时我知道自己报名了转专业可能转到物院，于是就想着要开始选量子力学（四大我第一门学的课），因为知道自己是平转所以比同级的同学少了一年在物院的学习，要抓紧赶上才能尽快投入本研。可惜运气不好的是正好碰上了刘玉鑫老师开的量子力学，刘老师刀子嘴豆腐心，对学生十分关心尊重，但他上课起来却毫不留情，巨大的作业量和凶残的小测让我很快就感觉自己没有跟上进度，只得第一次动用期中退课的选项。

到了大二上的时候，我再次选了量子力学，这次碰上了年纪稍大但是为人幽默风趣，操着一口京腔官话（并不是现在的北京话）的田光善老师，他制作了详细精美的量子力学笔记，同时辅以适当的作业，让我感觉课程难度放缓了许多，同时也激发了我认真学习的兴趣，这门考试我最后拿了 97 分，在我四年的成绩单中也算得上是一个漂亮的数字，大大地鼓舞了我学物理的兴趣。在课程结束后，寒假期间，我去找了这位幽默可爱的田老师聊天（聊科研聊人生），并且尝试问他是否还招本科生做本研，因为我个人很喜欢他的幽默特质，具有很强的人格魅力。可惜田老师说，他 2018 年就已经退休了，自此不再带学生了。田老师自己是做凝聚态理论的，我至今还在常常想：如果田老师当时年轻几岁，还带学生，也许我就会一直跟着他做本研甚至读博士了，那我就应该从此钻研与凝聚态理论，而与现在的人生轨迹截然不同了。当时，我在叹息之余还是和他聊了许多和物理有关无关的事，那次谈话至今想起来都是本科生涯最美好的回忆之一。

和田老师谈完之后，我回到湖北家中，这时候正好爆发了新冠疫情，我回家的那两天正好就是官方宣布不明原因肺炎为新冠肺炎，并且开始动用全国力量支援武汉疫情。回到家中一周后，武汉开始封城，同时过了几天我所在的荆州市也开始封城，于是我就在家中开始了我的大二下学期，我知道我到了该进入本研的时间节点，于是在家中也开始疯狂联系各个方向老师，包括做凝聚态实验的老师，以及做高能实验的高原宁院长、冒亚军教授等，最后因为要远程开展科研的原因，我便进入了做高能实验分析的冒亚军老师组，做了大约一年的 BESIII 实验相关课题，在这个过程中感谢全程带我入门的宋昀轩师兄，为人无私热情，技术高超。

同时，在大三回到校园后，我又在办公室认识了本科期间重要的一位朋友——李聪乔师兄，当时他是隔壁李强老师组做 CMS 实验的，和我一见如故，常常约饭聊天。在 BSEIII 上的课题做了很久，有了一定的进展，但我卡在课题上的一个地方卡了很久，由于只有我一个人在推动这个课题，所以对我来说是个很巨大的挑战。之后由于一些原因，我个人就逐渐丧失了对 BES 实验的热情，正好这个时候到了夏令营保研的阶段，我想试试换个方向，于是李聪乔师兄就把我推荐到了李强老师那边。而正好我的导师冒亚军老师和李强老师有合作，我便在冒老师名下跟着李强老师从大三暑假开始了我在 CMS 实验的探索。

在这个探索过程中，我也就做出了本篇毕业论文的工作，这个工作成功地和 UCSD 的 Javier 组建立了合作关系，并且在 CMS 官方会议上给予了报告，作为博士生涯的铺垫来说是一个不错的开端，也为以后更多的合作奠定了基础

于是，在这篇论文的最后，我要感谢全程指导我科研工作的冒亚军老师、李强老师，为我科研工作牵线搭桥不吝赐教的李聪乔师兄，以及两年多来在办公室与我一起奋战在科研一线并且帮助过我的宋昀轩师兄、郭启隆师兄、赵宇哲师兄、邓森师兄、钱

致谢

思天师兄、李彦谷师兄、谢昕海师兄、相腾师兄和王轩师姐（以上排名不分先后），以及给了我不少人生建议的安莹师姐。

回首四年，我已与刚踏进北大时的自己发生了很多天翻地覆的改变，比起刚进来时的随波逐流，我现在更感到自己有做出成就改变世界的可能性，并且愿意为此奉献汗水和努力。希望我自己在另一个四年后，能够交出一份让现在的我满意的人生答卷，与世界共勉。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

(必须装订在提交学校图书馆的印刷本)

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因须要延迟发布学位论文电子版，授权学校在 一年 / 两年 / 三年以后在校园网上全文发布。

(保密论文在解密后遵守此规定)

论文作者签名： 导师签名： 日期： 年 月 日