



北京大学

# 本科生毕业论文

题目：在 CMS 实验中对希格斯粒子  
的双 W 玻色子散射的探测与  
其末态标记技术的研究

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

中文摘要部分...

关键词：是，楷，体，吗



---

**ABSTRACT**

xxxxxxxxxx

xxxxxx (Physics)

Directed by xxxx

**ABSTRACT**

英文摘要部分...

KEY WORDS: A,B,C,D



## 目录

<b>第一章 引言</b>	<b>1</b>
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 g-2 实验结果与标准模型理论值的偏差 . . . . .	6
1.2.2 超重的 W 玻色子 . . . . .	7
1.3 LHC 上的 CMS 实验 . . . . .	9
1.3.1 大型强子对撞机 (LHC) . . . . .	9
1.3.2 紧凑缪子螺线管实验 (CMS) . . . . .	10
<b>第二章 大动量希格斯粒子的产生和到 WW 散射的物理</b>	<b>13</b>
<b>第三章 CMS 实验上的喷注标记技术发展介绍</b>	<b>15</b>
3.1 重建喷注的 Anti- $k_T$ 算法 . . . . .	15
3.2 基于理论的高级变量的筛选条件算法 . . . . .	17
3.2.1 soft-drop mass 算法 . . . . .	17
3.2.2 N-subjettiness 算法 . . . . .	17
3.2.3 ECF: $N_2$ 算法 . . . . .	18
3.3 基于机器学习的高级变量算法 . . . . .	19
3.3.1 $N_3$ -BDT 算法 . . . . .	19
3.3.2 HOTVR 算法 . . . . .	19
3.3.3 BEST 算法 . . . . .	19
3.4 基于深度学习的初级变量算法 . . . . .	20
3.4.1 IamgeTop 算法 . . . . .	20
3.4.2 DeepAK8 算法 . . . . .	20
<b>第四章 用于喷注标记的 ParticleNet 深度神经网络</b>	<b>21</b>
4.1 喷注表示方式 . . . . .	21
4.2 边卷积 (EdgeConv) . . . . .	22
4.3 ParticleNet 网络架构 . . . . .	23
4.4 ParticleNet 在部分标记任务上达到最佳 (SOTA) . . . . .	24

<b>第五章 开发首个 <math>H \rightarrow WW</math> 质量去相关的多分类标记器（基于 ParticleNet）</b>	<b>25</b>
5.1 质量去相关技术 . . . . .	25
5.2 分类标签 . . . . .	26
5.2.1 信号分类标签 . . . . .	26
5.2.2 本底分类标签 . . . . .	26
5.3 数据集 . . . . .	26
5.3.1 训练集和验证集 . . . . .	26
5.3.2 测试集 . . . . .	27
5.4 标注器设置 . . . . .	27
5.4.1 预挑选条件 . . . . .	27
5.4.2 重加权设置 . . . . .	28
5.4.3 神经网络输入 . . . . .	28
5.5 标记器在测试集上效果 . . . . .	28
5.6 在分析中的初步应用效果 . . . . .	32
<b>第六章 总结和展望</b>	<b>33</b>
<b>致谢</b>	<b>35</b>
<b>北京大学学位论文原创性声明和使用授权说明</b>	<b>37</b>

## 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、朝勇振一郎、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 $\mu$ 子磁矩 $g - 2$ 反常，弱相互作用传播子W玻色子质量超出标准模型预言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 $\mu$ 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散射过程，研究该过程最新最好的喷注标记技术，同时也是第一个对HWW衰变的多衰变道标记器开发，希望以此为W玻色子相关测量提供更多符合标准模型的证据，或者揭开W玻色子超出标准模型的更多迹象！

### 1.1 标准模型

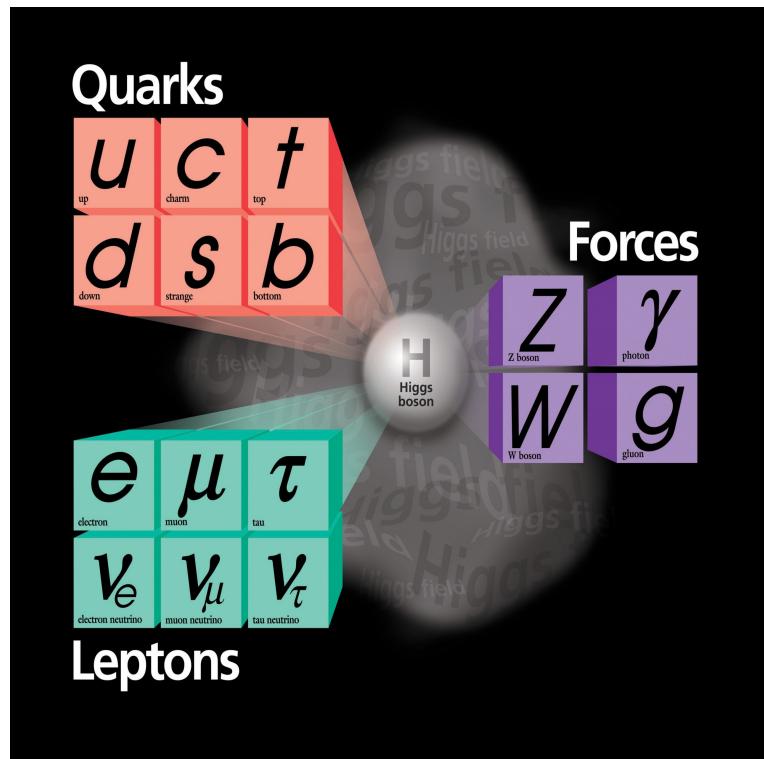
标准模型是当前粒子物理学（也称作高能物理）中得到广泛认可的理论框架，主要包括两方面的内容：第一，给出了构成我们大千世界的基本粒子，包括组成物质的费米子和负责传递各种相互作用的玻色子；第二，它统一了四种基本相互作用中的三种：电磁相互作用、弱相互作用和强相互作用，其中电磁相互作用和弱相互作用在标准模型中通过电弱统一理论进行描述，强相互作用通过量子色动力学进行描述。并且预言了赋予基本粒子质量的希格斯机制和希格斯粒子（也叫“上帝粒子”）。

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，tau子（1975）、顶夸克（1995）、中微子（2000）和希格斯玻色子（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  $\nu_e$ , 缪子中微子  $\nu_\mu$ , 陶子中微子  $\nu_\tau$ ）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  $\lambda^a$  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

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

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 ( $r$ )，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  $R_c \sim 1 \text{ fm} (= 10^{-15} \text{ m})$ 。

### 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  $Q = Y/2 + I_3$ ，而其中 Y 是  $U(1)_Y$  (叫弱超荷) 的生成元， $I_3$  (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

$$\mathcal{L} = i\bar{\Psi}\gamma^\mu\partial_\mu\Psi \quad (1.4)$$

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  $\mathcal{D}_\mu$  以代替  $\partial_\mu$ ：

$$\mathcal{D}_\mu = \partial_\mu - i\frac{g}{2}gT^aW_\mu^a - i\frac{g'}{2}YB_\mu \quad (1.5)$$

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， $T$  是弱同位旋矩阵，形式上等同于泡利矩阵， $a$  是欧式指标可取 1,2,3 (度量矩阵为 3 阶单位阵)， $\mu, \nu$  是四维指标。从而将带电弱相互作用的费米子场拉氏量改写为

$$\mathcal{L} = i\bar{\Psi}\gamma^\mu\mathcal{D}_\mu\Psi - \frac{1}{4}W_a^{\mu\nu}W_{a\mu\nu}^a - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \quad (1.6)$$

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

$$W^\pm = \frac{1}{\sqrt{2}}(W^1 \mp iW^2) \quad (1.7)$$

相应地，Z 和  $\gamma$  光子可表示为

$$\begin{pmatrix} \gamma \\ Z \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B \\ W_3 \end{pmatrix} \quad (1.8)$$

其中  $\theta_W$  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

$$\mathcal{L} = (\partial_\alpha \phi)^* (\partial^\alpha \phi) - m^2 \phi^* \phi - V(\phi^* \phi) \quad (1.9)$$

反过来, 如果一个拉氏量中有  $\phi$  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  $\phi$ , 并将该场的拉氏量写为

$$\mathcal{L} = \partial_\mu \bar{\phi} \partial^\mu \phi + \mu^2 \bar{\phi} \phi - \lambda (\bar{\phi} \phi)^2 \quad (1.10)$$

其中  $\lambda, \mu$  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  $U(1)$  的局域对称变换。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  $\phi$  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  $\phi$  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

$$\begin{aligned} \mathcal{L} = & \frac{1}{2} [(\partial_\alpha + iqA_\alpha)(\varphi_1 + v + i\varphi_2)]^* [(\partial^\alpha + iqA^\alpha)(\varphi_1 + v + i\varphi_2)] - \frac{1}{4} F_{\alpha\beta} F^{\alpha\beta} \\ & + \frac{\mu^2}{2} (\varphi_1 + v + i\varphi_2)^* (\varphi_1 + v + i\varphi_2) - \frac{\lambda}{4} [(\varphi_1 + v + i\varphi_2)^* (\varphi_1 + v + i\varphi_2)]^2 \end{aligned} \quad (1.11)$$

化简后有

$$\mathcal{L} = \frac{1}{2} (\partial_\alpha \varphi_1) (\partial^\alpha \varphi_1) - \mu^2 \varphi_1^2 + \frac{1}{2} (\partial_\alpha \varphi_2) (\partial^\alpha \varphi_2) - \frac{1}{4} F_{\alpha\beta} F^{\alpha\beta} + \frac{1}{2} q^2 v^2 A_\alpha A^\alpha + \mathcal{L}_{int} \quad (1.12)$$

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, 于是我们在局域变换中设定相位  $\theta = -\arctan(\phi_2/\phi_1)$ , 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

$$\mathcal{L} = \frac{1}{2} (\partial_\alpha \varphi_1) (\partial^\alpha \varphi_1) - \mu^2 \varphi_1^2 - \frac{1}{4} F_{\alpha\beta} F^{\alpha\beta} + \frac{1}{2} q^2 v^2 A_\alpha A^\alpha + \mathcal{L}_{int} \quad (1.13)$$

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

现在我们得到,  $\varphi_1$  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

$A_\alpha$  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  $\varphi_1$  变为了复值的

$$\phi(x) = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \quad (1.14)$$

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  $\gamma$ 。

## 1.2 超出标准模型的迹象

### 1.2.1 g-2 实验结果与标准模型理论值的偏差

$\mu$  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,  $\mu$  子的质量大约是电子的 200 倍。 $\mu$  子与同属于轻子的电子和  $\tau$  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,  $\mu$  子的行为就好像它们有一个微小的内部磁铁。在强磁场中,  $\mu$  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  $\mu$  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, 普适的 g 的物理定义是

$$\boldsymbol{\mu} = g \frac{e}{2m} \mathbf{S} \quad (1.15)$$

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,  $\boldsymbol{\mu}$  是磁矩。按照狄拉克方程的计算, 基本费米子的  $g=2$ 。

当宇宙射线撞击地球大气层时会自然产生  $\mu$  子, 费米实验室的粒子加速器可以大量产生它们。

根据量子场论的计算 (由于真空态和量子涨落), 对于  $\mu$  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。

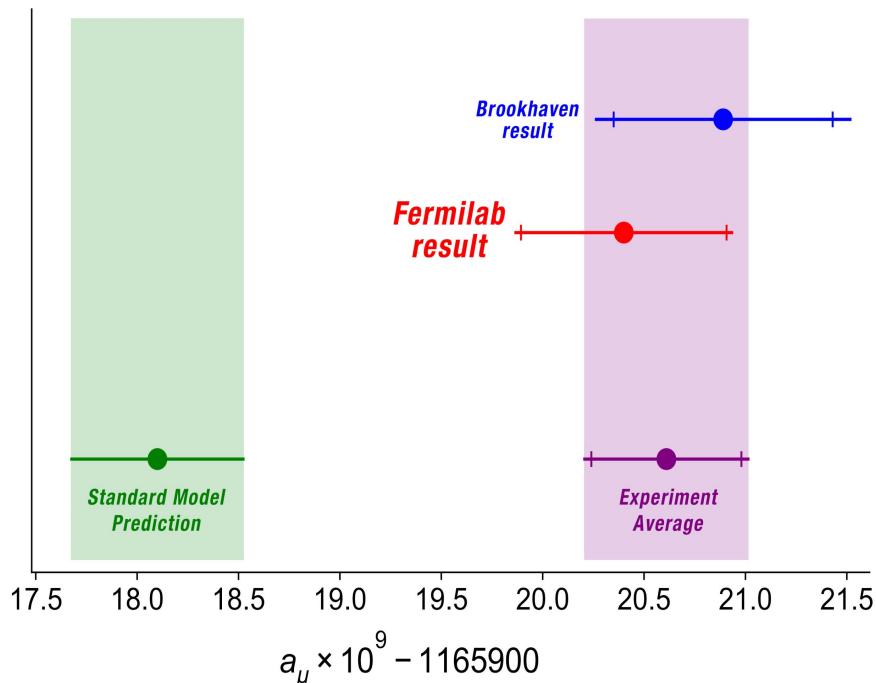
2021 年, 美国费米国家实验室的 Muon g-2 实验备受期待的首个结果表明, 在前所未有的精确度测量下,  $\mu$  子的  $g$  因子与标准模型计算结果存在较大偏差。几十年来,  $g-2$  的实验和理论差异就已经被测量过, 这个里程碑式的结果, 更是进一步确认了这个差异。

当  $\mu$  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 $\mu$ 子的进动加速或减速非常轻微。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 $\mu$ 子的g因子大小。

公认的 $\mu$ 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： $g=2.00233184122(82)$ 。（见下图1.2）

图 1.2 g 因子的标准模型理论值与实验的偏差



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 $4.2\sigma$ （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 $5\sigma$ 阈，但 $4.2\sigma$ 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2018年运行的第一年，费米实验室收集的数据比之前所有 $\mu$ 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 $\mu$ 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， $\mu$ 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。

### 1.2.2 超重的W玻色子

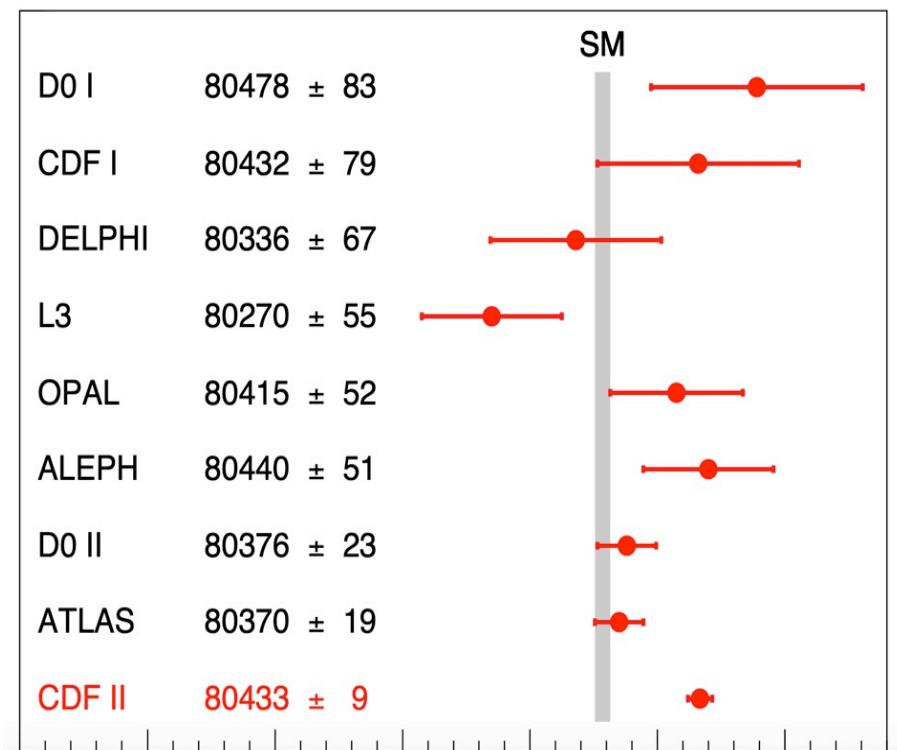
W玻色子是弱相互作用的媒介粒子。它参与太阳发光和粒子衰变的反应过程。标准模型给出的W质量理论计算值为 $M_W = (80357 \pm 6)\text{MeV}$ 。（该值基于复杂的标准模

型计算，该计算通过将 W 玻色子的质量与其他两个粒子的质量的测量错综复杂地联系起来：顶夸克，于 1995 年在费米实验室的 Tevatron 对撞机中发现；希格斯玻色子，在 2012 年欧洲核子研究中心的大型强子对撞机上发现）

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。但是 2022 年 4 月公布的美国费米国家实验室的 W 质量测量达到了更高的精度——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与 CDF 合作的科学家今天宣布，他们已经实现了迄今为止对 W 玻色子质量的最精确测量。

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： $M_W = (80433.5 \pm 9.4)\text{MeV}$ 。（见图1.3）

图 1.3 不同实验对 W 质量测量结果的比较<sup>Wmass</sup>



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  $7\sigma$  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巨大冲突。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。

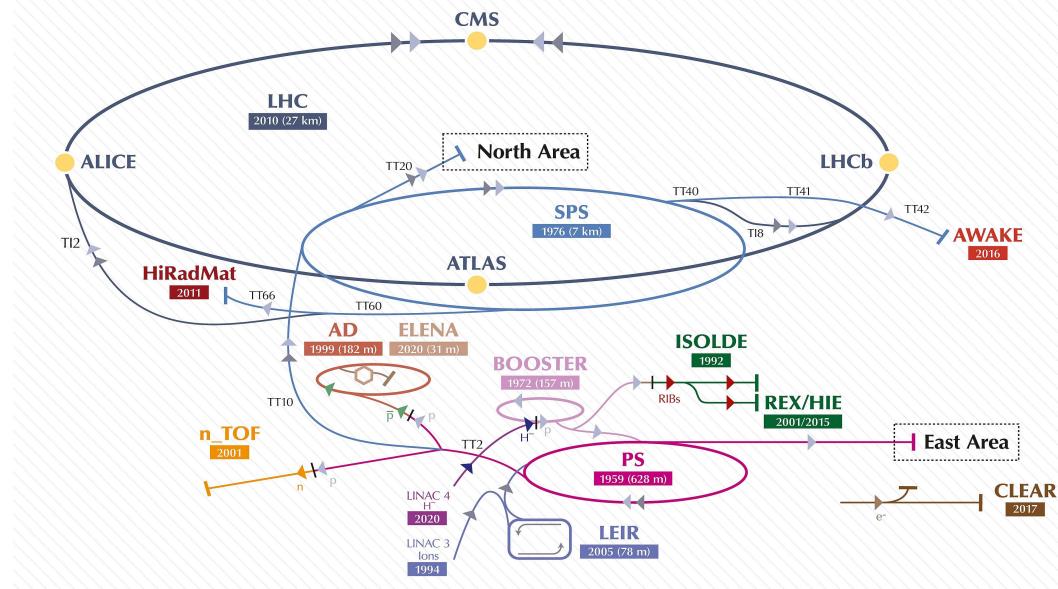
科学界对这个结果也是众说纷纭——费米实验室副主任 Joe Lykken 指出：“虽然这是一个有趣的结果，但测量结果需要通过另一个实验来确认，然后才能完全解释。”德克萨斯农工大学 CDF 联合发言人 David Toback 补充道：“如果实验值和预期值之间的差异是由于某种新的粒子或亚原子相互作用造成的，这是一种可能性，那么它很有可能在未来的实验中被发现。”

### 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。

图 1.4 LHC 的探测器和部门分布



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。

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。

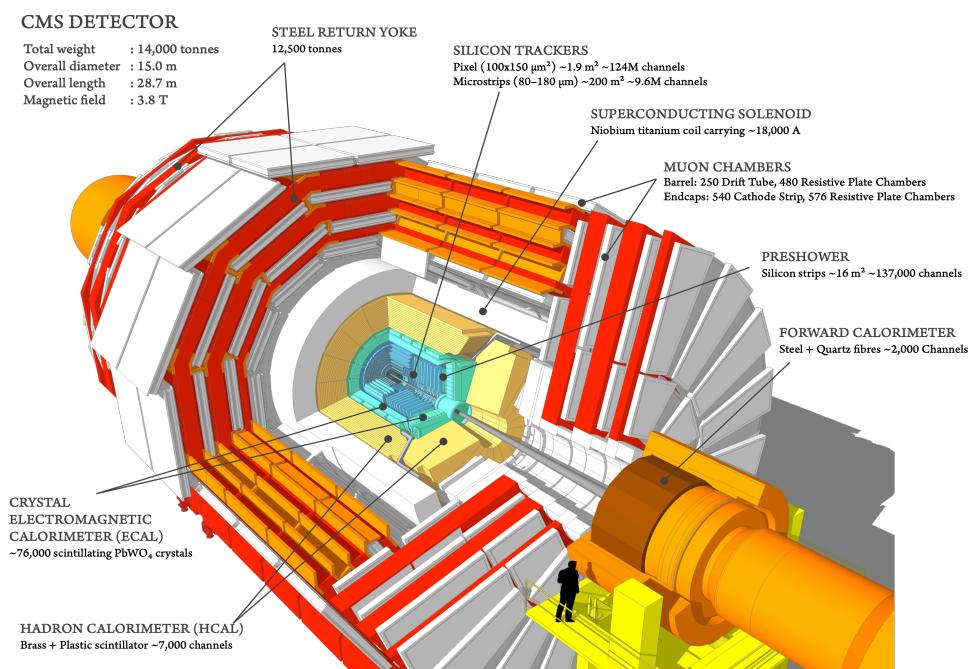
CMS 长 21m，直径 15m，重约 14000 吨。来自 47 个国家/地区的 206 个科研机构的 4000 多人组成的 CMS 合作组建造并运行了探测器。2012 年 7 月，CMS 与 ATLAS 一起初步发现了希格斯玻色子。

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 1.5 CMS 探测器剖面图



CMS 被设计为通用探测器，能够在 LHC 粒子加速器的质心能量 0.9-13 TeV 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 吨重量的探测器主体——钢轭内。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、 $\mu$  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 大动量希格斯粒子的产生和到 WW 散射的物理

第二章部分...



## 第三章 CMS 实验上的喷注标记技术发展介绍

### 3.1 重建喷注的 Anti- $k_T$ 算法

Jet 聚类算法是分析强子碰撞数据的主要工具之一。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：

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \quad (3.1)$$

$$d_{iB} = k_{ti}^{2p} \quad (3.2)$$

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

聚类算法是这么执行的：

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。

通过以上算法，我们就得到了粒子流中的所有喷注。

这里还没有完，对于(3.1)和(3.2)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 anti- $k_t$  算法，也叫 AK 算法。

$R$  通常会取 0.4, 0.8, 1.5 三个值，分别对应的算法是 AK4, AK8, AK15  
anti- $k_t$  算法与其他算法重建喷注效果比较如下图所示

图 3.1 用 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  $|y| < 2$ 。逆反应用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）。

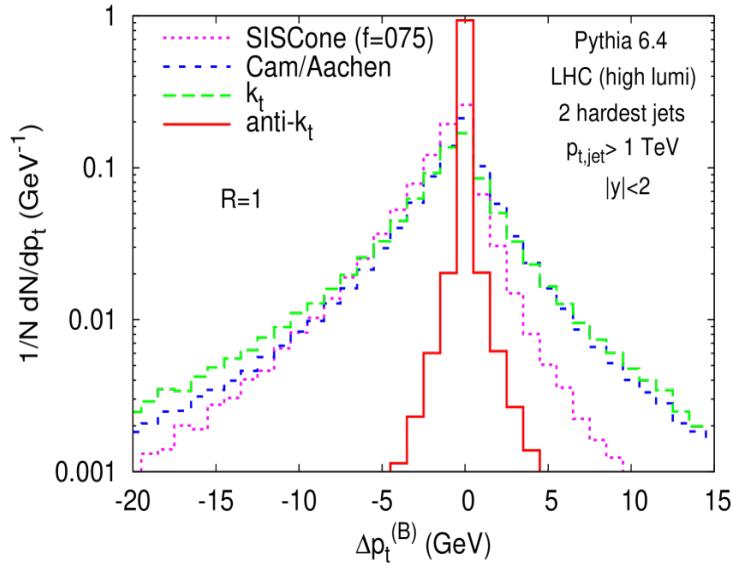
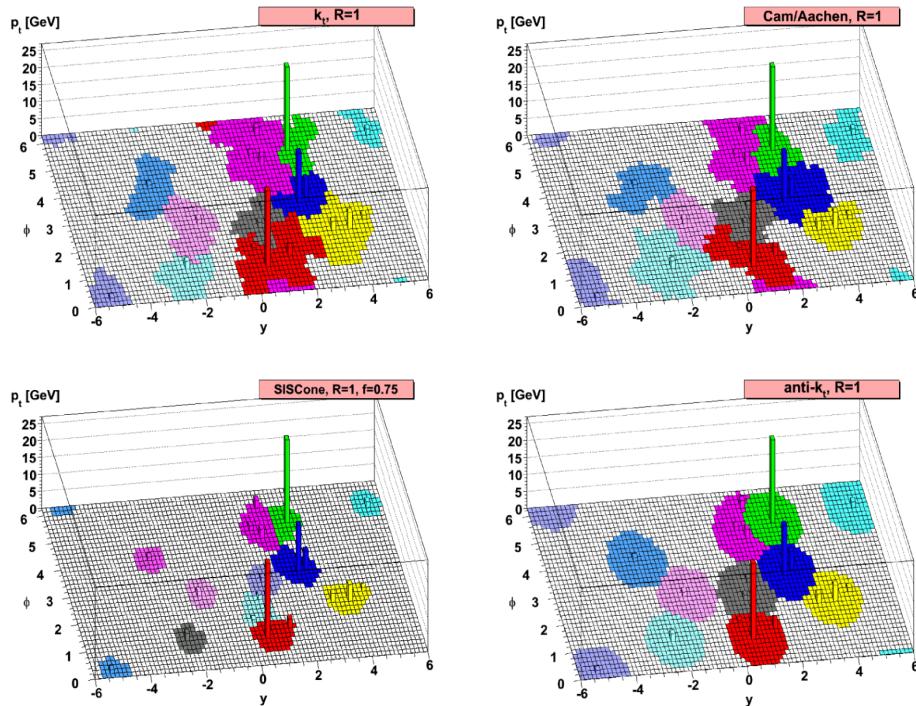


图 3.2 用 anti- $k_t$  算法和  $k_t$ , Cambridge/Aachen, SISCone 算法比较得到的逆反应分布。这是用 Pythia 6.4 模拟的双喷注事件计算的，其中最硬的两个喷注满足  $p_T >$  并且都位于  $|y| < 2$ 。逆反应用两个最硬喷流中每一个的净横向动量变化，这是由于当高亮度 LHC 堆积被加入事例时非堆积粒子的重新分配（每束交叉时约有 25 个 pp 相互作用）。



## 3.2 基于理论的高级变量的筛选条件算法

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

### 3.2.1 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：

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut} \left( \frac{\Delta R_{12}}{R_0} \right)^\beta \quad (3.3)$$

这里

$$R_{12} = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2} \quad (3.4)$$

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(3.3)简化为

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > 0.01 \quad (3.5)$$

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。

### 3.2.2 N-subjettiness 算法

高级变量 N-subjettiness 定义为

$$\tau_N = \frac{1}{d_0} \sum_i p_{T,i} \min [\Delta R_{1,i}, \Delta R_{2,i}, \dots, \Delta R_{N,i}] \quad (3.6)$$

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  $\tau_N$  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  $\tau_N$  之间的比值获得更有鉴别效果的变量，例如

#### (1) 定义

$$\tau_{21} = \frac{\tau_2}{\tau_1} \quad (3.7)$$

可以针对鉴别二分叉喷注（如 W, Z, H）。

## (2) 定义

$$\tau_{32} = \frac{\tau_3}{\tau_2} \quad (3.8)$$

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  $\tau_{21}$  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。

### 3.2.3 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

$${}^q e_N^\beta = \sum_{1 \leq i_1 < i_2 < \dots < i_N \leq N_c} \left[ \prod_{1 \leq k \leq N} \frac{i_k}{J} \right] \prod_{m=1}^q \min_{i_j < i_k \in \{i_1, i_2, \dots, i_N\}}^{(m)} \left\{ \Delta R_{i_j, i_k}^\beta \right\} \quad (3.9)$$

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

$$N_2^1 = \frac{{}^2 e_3^1}{{}^1 e_2^1)^2} \quad (3.10)$$

与 N-subjettiness 比值  $\tau_{21}$  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

$$N_2^{DDT}(\rho, p_T) = N_2(\rho, p_T) - N_2^{(X\%)}(\rho, p_T) \quad (3.11)$$

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 基于机器学习的高级变量算法

#### 3.3.1 $N_3$ -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：

$$\frac{{}_a e_N^\alpha}{({}_b e_M^\beta)^x}, \text{ where } M \leq N \text{ and } x = \frac{a\alpha}{b\beta}. \quad (3.12)$$

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

$$\begin{aligned} & \frac{{}_1 e_2^{(2)}}{\left({}_1 e_2^{(1)}\right)^2}, \frac{{}_1 e_3^{(4)}}{{}_2 e_3^{(2)}}, \frac{{}_3 e_3^{(1)}}{\left({}_1 e_3^{(4)}\right)^{3/4}}, \frac{{}_3 e_3^{(1)}}{\left({}_2 e_3^{(2)}\right)^{3/4}}, \frac{{}_3 e_3^{(2)}}{\left({}_3 e_3^{(4)}\right)^{1/2}}, \frac{{}_1 e_4^{(4)}}{\left({}_1 e_3^{(2)}\right)^2}, \\ & \frac{{}_1 e_4^{(2)}}{\left({}_1 e_3^{(1)}\right)^2}, \frac{{}_2 e_4^{(1/2)}}{\left({}_1 e_3^{(1/2)}\right)^2}, \frac{{}_2 e_4^{(1)}}{\left({}_1 e_3^{(1)}\right)^2}, \frac{{}_2 e_4^{(1)}}{\left({}_2 e_3^{(1/2)}\right)^2}, \frac{{}_2 e_4^{(2)}}{\left({}_1 e_3^{(2)}\right)^2}. \end{aligned} \quad (3.13)$$

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， $\tau_{32}^{SD}$  和  $f_{rec}$  作为输入。

#### 3.3.2 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 ( $t/W/Z/H$ )。

#### 3.3.3 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。

### 3.4 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。

#### 3.4.1 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)  $\mu$  子个数；(4) 径迹条数。还运用了深度喷注 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。

#### 3.4.2 DeepAK8 算法

DeepAK8 是针对 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) 上无处不在但也特别迷人的对象之一，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，要结合以上两点，关键性的问题就是：**怎样尽可能好地把物理上的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（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。

Wang 等人提出了边卷积 (EdgeConv) 操作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  $\Theta$  参数化的函数，且  $\Theta$  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  $\mathbf{h}_\Theta$  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

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

在这里，方程(4.1)中的邻点的特征向量  $\mathbf{x}_{ij}$  被  $\mathbf{x}_{ij}$  与中心点的特征向量  $\mathbf{x}_i$  的差值所取代，并且  $\text{Conv}_{\Theta}$  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  $\square$ ，ParticleNet 采取的是平均值  $\frac{1}{k} \sum$ 。

EdgeConv 操作的一个重要特点是它可以很容易地堆叠，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，随后可以应用另一个 EdgeConv 操作。这使我们能够使用 EdgeConv 操作构建一个深度网络，该操作可以分层学习点云的特征。

EdgeConv 操作的可堆叠性也带来了另一个有趣的可能性。基本上，EdgeConv 学习到的特征向量可以看作是潜在空间中原始点的新坐标，然后是点之间的距离（用于判断最近的邻居）可以在这个潜在空间中重新定义。换句话说，点的接近程度可以动态地通过 EdgeConv 操作学习到。这也是动态图卷积神经网络的结果 Wang (2018)。其中描述点云的图被动态更新以反映边的变化，即每个点的邻居，这也得到证明比保持静态图的性能更好。Wang 等人 (2018)。

### 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 Dynamic Graph CNN 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。

我们在 EdgeConv 操作块中构建 ParticleNet 架构。图 1 说明了 EdgeConv 块的结构。EdgeConv 块从找到每个粒子的最近相邻粒子。使用 EdgeConv 块的坐标输入计算粒子之间的距离。然后，对粒子的输入特征向量应用 EdgeConv 操作。每个块由多个 EdgeConv 操作组成，可能具有不同数量的卷积核。在每个块内，描述粒子云的图是固定的，即一个粒子总是有相同的最近的邻居。每个 EdgeConv 操作后面都有一个批量归一化层 BatchNorm 【Ioffe 和 Szegedy (2015)】然后是 ReLU 非线性 Glorot 等人。(2011)。与 ResNet 类似，与 EdgeConv 操作并行运行的快捷连接也包含在每个块中。

本文使用的 ParticleNet 架构如图 2 所示。它由三个 EdgeConv 块组成。第一个 EdgeConv 块使用伪快速度-方位角空间中粒子的空间坐标来计算距离，而随后的块使用学习的特征向量作为坐标。最近邻居的数量对于三个块，分别为 8、12 和 16。每个块由三个 EdgeConv 层组成。三个块的 EdgeConv 层的输出通道数为 (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云中所有粒子的学习特征。接下来是两个全连接层，分别有 256 和 512 个单元。两者都使用 ReLU 激活函数。辍学层 Srivastava 等人。(2014) 为两个全连接层添加了下降概率为 0.1 和 0.5 以防止过度拟合。具有 2 个单元的全连接层，后跟一个 softmax 函数，用于生成二进制分类任务的输出。

#### 4.4 ParticleNet 在部分标记任务上达到最佳 (SOTA)

## 第五章 开发首个 $H \rightarrow WW$ 质量去相关的多分类标记器（基于 ParticleNet）

### 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。

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。

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

### 5.2.1 信号分类标签

对于我们关心的  $H \rightarrow WW \rightarrow anything$  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
- **$\tau_e\nu_e qq$** :  $H \rightarrow W(\tau\nu)W(2q)$ ,  $\tau$  接着衰变为电子等产物
- **$\tau_\mu\nu qq$** :  $H \rightarrow W(\tau\nu)W(2q)$ ,  $\tau$  接着衰变为  $\mu$  子等产物
- **$\tau_h\nu qq$** :  $H \rightarrow W(\tau\nu)W(2q)$ ,  $\tau$  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 HWW 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：

1. 样本基于 2017 ultra-legacy，总共有两千五百多万事例。

2. 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，同时设置  $W$  质量为  $m_W = 80[\text{GeV}]$ 。
3. 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。
4. 喷注经过 AK8 算重建法并且通过 MC 事实匹配打上标签。

### 5.3.1.2 本底样本

QCD 本底样本有以下几个特点：

1. 样本基于 2017 ultra-legacy，总共有两千八百多万事例。
2. 喷注经过 AK8 算重建法并且通过 MC 事实匹配打上标签。

### 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  $125[\text{GeV}]$  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：

1. 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
2. 喷注的横向动量满足  $200[\text{GeV}] < p_T < 2500[\text{GeV}]$
3. 喷注的 soft-drop 质量满足  $20[\text{GeV}] \leq m_{SD} < 260[\text{GeV}]$
4. 对于训练样本，喷注要通过事实匹配满足 12 个标签分类之一

### 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[GeV] 到 260[GeV] 每隔 10[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 [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：

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

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。

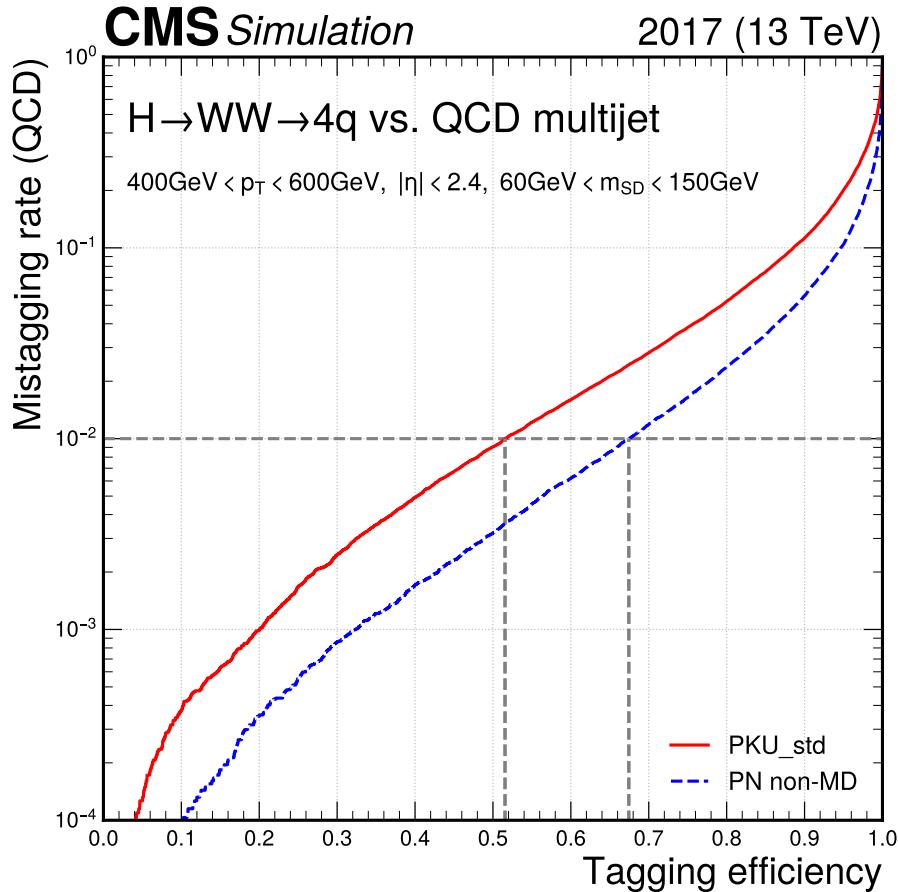
### 5.4.3 神经网络输入

## 5.5 标记器在测试集上效果

ROC 图如下

表 5.1 神经网络输入

对象	变量	描述
粒子候选者	$\eta_{rel}$	相对于 AK8 喷注主轴的赝快度 $\Delta\eta$
	$\phi_{rel}$	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 $\mu$ 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 $\chi^2$
	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 $\eta_{rel}$	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	$\eta_{rel}$	相对于 AK8 喷注主轴的赝快度 $\Delta\eta$
	$\phi_{rel}$	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 $\chi^2$ 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

图 5.1 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 HWW 多分类标记器，PN non-MD 是原生 ParticleNet 的质量相关版本 HWW4q 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5.2所示。

从图5.2可以看到，我们开发的质量去相关标记器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5.3所示。

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。

图 5.2 被 PKU\_std 误标记的 QCD 本底的  $m_{SD}$  分布

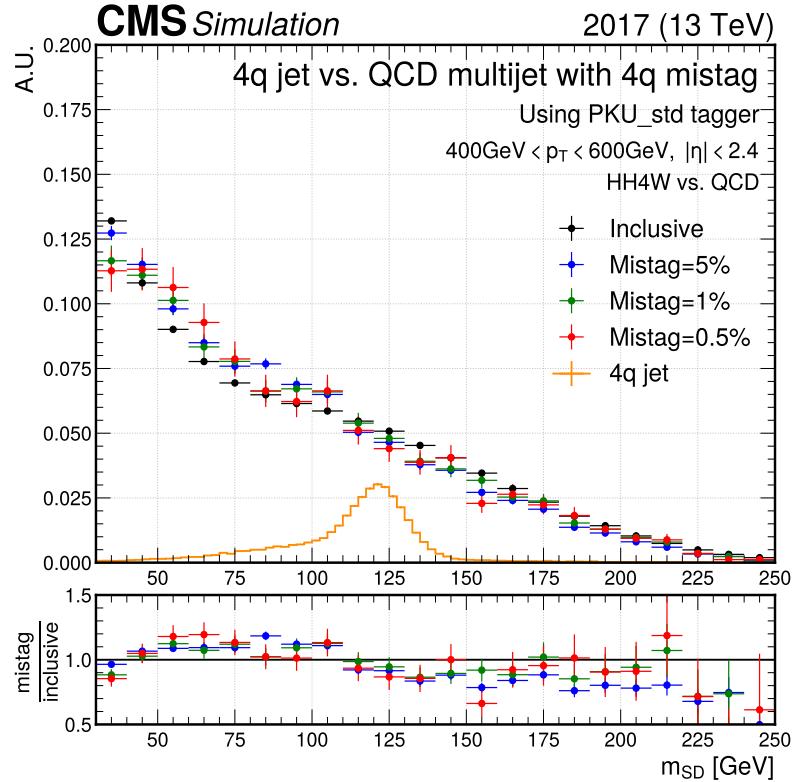
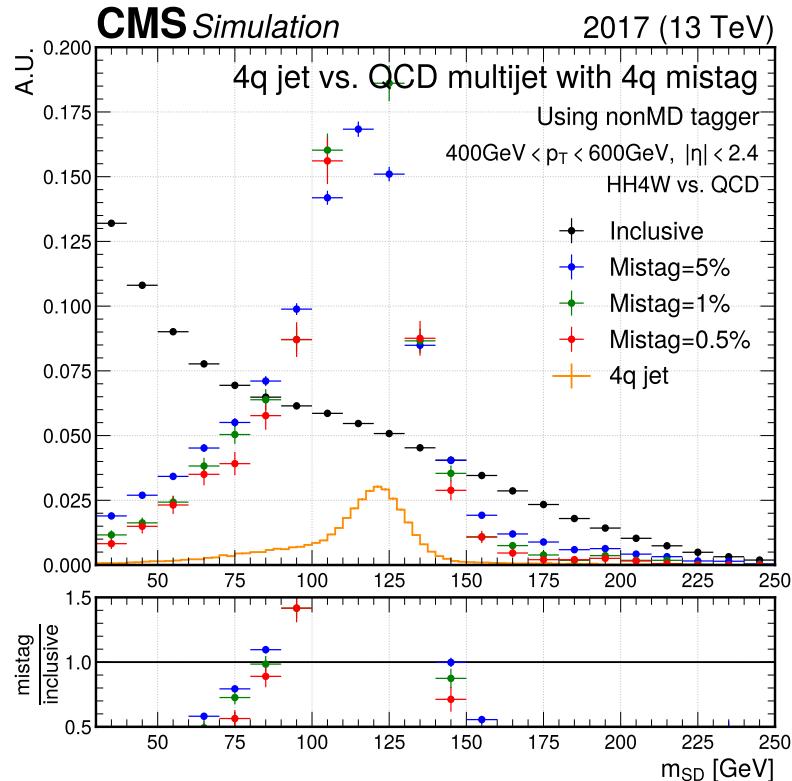
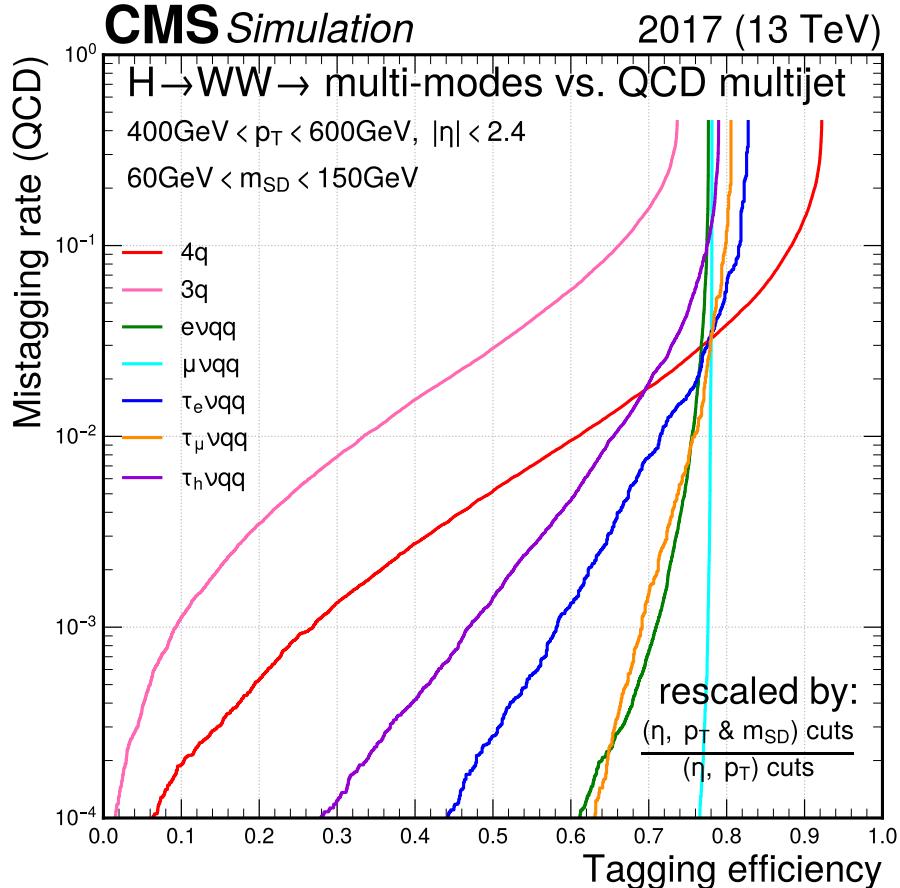


图 5.3 被 PN nonMD 误标记的 QCD 本底的  $m_{SD}$  分布



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

图 5.4 被 PN nonMD 误标记的 QCD 本底的  $m_{SD}$  分布



从这个图里我们可以看到，在 Mistag rate=1% 时，除了 3q 末态（指 HWW 衰变到四个夸克但只有三个出现在被重建的 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

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，目前已经应用的分析为：胶子聚合成产生希格斯，再到 WW 散射的过程。在分析探索中，我们把我们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：

## 第六章 总结和展望

第六章部分...



## 致谢

致谢部分...



# 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