



北京大学

博士研究生学位论文

题目：大半径喷注标记的现代深度学习技术——算法、校准方法及CMS 实验中的应用

姓 名：李聪乔

学 号：1901112050

院 系：物理学院

专 业：粒子物理与原子核物理

研究方向：高能实验物理

导 师：李强 教授

☒ 学术学位 ☐ 专业学位

二〇二四 年 六 月



北京大學

Doctoral Dissertation

Title: Modern deep learning for large- R jet
tagging — algorithms, calibration
methods, and applications in the CMS
experiment

Name: Congqiao Li

Student ID: 1901112050

Affiliation: School of Physics

Major: Particle Physics and Nuclear Physics

Research Division: High Energy Experimental Physics

Directed by: Prof. Qiang Li

☒ Academic Degree

☐ Professional Degree

June, 2024

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，引起有碍作者著作权之问题，将可能承担法律责任。



摘要

探索构成世界的最基本的物质单元是人类永恒的追求。尽管人类的思考与探索持续了上千年，但是直到 19 世纪末电子的发现，人类才进入了亚原子时代，从此对开启了粒子物理科学探索的征程。20 世纪是粒子物理辉煌发展时期。从量子力学的奠定，到描述电磁作用的量子电动力学的完善和一系列对量子场论理论的扩充，伴随着由加速器技术为实验粒子物理带来的井喷式发展，人们对微观世界的认识逐渐完善，最终推向粒子物理标准模型（SM）的建立。SM 是人类迄今为止描述微观世界的最丰富和权威的模型。它涵盖了目前已知的所有构成世界的基本粒子，分为费米子（夸克、轻子）和玻色子（胶子、光子、 W 、 Z 玻色子和 Higgs 玻色子）；统一地表述了强、弱和电磁三种相互作用；并基于基本规范对称性建立，以 Higgs 机制为其理论核心。半个世纪来，SM 的预言的物理现象在对撞机实验中被纷纷证实，其中，在 2012 年大型强子对撞机（LHC）发现了 Higgs 玻色子，标志了验证 SM 的最后一块拼图得以完成，是粒子物理发展的重要里程碑。

然而，尽管 SM 已经取得了空前的成功，对微观世界的探索仍在继续，因为 SM 仍有许多已知的缺陷。例如，SM 无法为暗物质提供解释，无法统一引力和其它三种作用力，也无法预言中微子的质量等等。理论学家就此提出了很多扩展 SM 的新理论，来涵盖这些可能的新粒子、解释这些新现象。因此，当前在 LHC 实验中寻找这些超出标准模型的新物理的存在证据是至关重要的。此外，对 SM 参数的精确测量，尤其是对 Higgs 玻色子性质方面的精确测量，可以进一步验证 SM 理论的正确性；利用 SM 理论预言与实验观测的潜在差异，可以寻找新物理可能存在的方向。这些方面是 LHC 当前进行实验的重要物理动机。

ATLAS 和 CMS 是 LHC 上的两个通用实验。它们利用各自的探测器连续多年记录 LHC 中的质子束在 TeV 能标下对撞产生的数据，用于进行成百上千项物理分析，从各个方面推进着标准模型的精确测量，并全面地寻找可能存在的新物理。利用采集到的数据，不断地提升数据分析手段，以获得更精确的测量结果、实现更灵敏的新物理寻找，是实验粒子物理学家的长期追求。自 2012 年以来，世界见证了现代深度学习技术的飞跃式发展。深度学习为数据统计、分析领域提供了全新的手段。它的经验表明，先进的深度神经网络（DNN）在处理具有复杂形式的数据时，可以表现出尤其巨大的潜力；而且数据本身越复杂，DNN 所能达到的能力相较传统方法的能力则越突出。近期

① This section presents an extended abstract in Chinese, formatted in compliance with the dissertation requirements. The corresponding English version is provided in the subsequent section.

CMS 实验在类似的场景下运用前沿的深度学习技术取得了一系列物理结果，实现了相应的物理测量灵敏度的大幅提升，展示着深度学习对 LHC 实验数据分析的巨大潜力。而在深度学习技术仍在飞速发展的今天，它为 LHC 实验带来的潜力还在不断地被挖掘。探索前沿的深度学习技术如何应用于 LHC 的物理分析，甚至改变其未来分析的范式，以释放 LHC 在物理测量方面的潜力，具有重要的现实意义。

本博士论文聚焦在大半径喷注标记的现代深度学习技术。在 CMS 实验采集的事例信息中，初始过程中产生的夸克或胶子，经过强子化过程后会产生一束聚拢的粒子簇，称为喷注。喷注是十分常见，却也是十分复杂且难以处理的物理对象。一个喷注中可能由几十个重建粒子所组成。相比于实验中常用的小半径喷注，大半径喷注是由更大聚类的半径 R 重建出来的，它所包含的粒子数量可多达 100 个，其子结构也更加复杂。在实验中，一个大半径喷注可以重建出高 Lorentz-boosted 的共振态——具体来说，在 LHC 中若产生几百 GeV 到几 TeV 能量的高能共振态粒子，它经过一系列衰变，并经强子化过程后所产生所有末态粒子，可以被重建到单个大半径喷注中。研究表明，使用大半径喷注在寻找某些事例相空间中具备独特优势。在相应物理实验中，为了区分信号和本底，需要对大半径喷注进行标记，即从大量本底过程的喷注中找到更具备信号特征的喷注。如何构建能够反映喷注性质的物理量，由此进行喷注标记，是 LHC 上的一个历时已久的复杂问题。很明显，这恰好为现代深度学习技术提供了土壤。近五年来，人们发现直接用前沿的 DNN 处理大半径喷注所包含的全部粒子的底层特征，进行喷注标记，可以表现出很大的性能提升。这蕴含着在大半径喷注标记这样的复杂任务中现代深度学习方法的巨大潜力。

围绕大半径喷注标记的现代深度学习技术，本论文旨在系统性地总结它所涉及的前沿 DNN 算法、校准方法、在 CMS 实验中取得的突破性进展，以及更进一步地，它为未来 CMS 实验带来的全新潜力。论文将以四个部分分别就四方面进行展开。前三个方面本质上涉及了前沿的 DNN 大半径喷注标记器参与到一个真实的物理分析时，涌现出的需要研究的各方面问题——首先，DNN 算法的设计是喷注标记器具备高性能的保障；其次，将标记器用于筛选实验数据时，对筛选效率在模拟数据与真实数据的差异进行校准是每个物理分析必备的步骤；最后，当它被用于物理分析时，要考虑整个实验的综合设计以确保标记器及其它工具的正确使用，得到可靠的物理结果。在第四个方面，文中将结合近年来深度学习和人工智能的新进展，开创性地阐述未来的喷注标记器的可能模式及其在实验中应用，即基于“大模型、多分类”的模式进行一个喷注大模型的预训练，并利用微调的技术用于具体的物理分析之中。后文中将对四个方面进行独立地论述。

首先，论文的第一部分陈列了文中所用到的实验粒子物理的基础和深度学习的基

础。在粒子物理方面，文中回顾了标准模型、LHC、CMS 探测器的基本知识，介绍了 CMS 的蒙特卡洛事例模拟、事例重建的情况。在深度学习方面，文中回顾了从上世纪来深度学习跌宕起伏的发展经过，强调了自 2012 年以来现代深度学习技术的飞跃式发展历程。

在第二部分，本论文旨在讲解近三年内推动大半径喷注标记算法领域进步的新进展。首先，作为历史回顾，介绍了早期基于规则的算法（如基于喷注子结构的 N -subjettiness 技术）、由“较浅”的机器学习方法实现的算法（如提升决策树方法）、和深度学习技术运用初期发展的算法，后者包括使用卷积神经网络（CNN）或循环神经网络（RNN）以图像、序列等形式表征喷注输入的算法；它们取得了相较之前方法的很大进步，但被后续算法大幅超越。这蕴含了重要理念的诞生，即设计更佳的适于喷注数据的 DNN 重点在于找到更好的底层特征表示，并将适合处理这种表示的前沿架构应用到喷注标记任务中。自 Particle Flow Network、ParticleNet 等网络诞生以来，人们聚焦使用点云或图的形式表示喷注，取得了算法设计的突破。本部分将重点聚焦在这一阶段之后发展起来的有助于算法提升的两个重要方面，即算法对于喷注数据内禀的 Lorentz 对称性的尊重，以及基于 Transformer 架构的算法设计。对于第一方面，在算法设计中遵守数据内禀的对称性有助于性能提升是深度学习领域的共识。在喷注网络的发展中，部分工作考虑了算法应对喷注进行特殊的旋转或 boost 不变的问题，然而对这些变换所从属的更普适的 Lorentz 变换，并未引起普遍重视。文中首先介绍了遵守 Lorentz 对称性的性能优异的典型网络 LorentzNet；又通过系统性地研究，通过设计方案不断使网络获得遵守新的 Lorentz 子对称性的能力，阐述了遵守 Lorentz 对称性如何作为一项归纳偏置具有广泛地提升网络性能、增强网络稳健性的效果。这一结论可以为后续算法的设计提供普适的指导。对于第二方面，本文介绍了 Transformer 算法在当下深度学习领域的特殊地位，并指出它在喷注数据中的适用性；介绍 Particle Transformer (ParT) 这一典型的基于 Transformer 架构的网络，讨论其设计形式及特殊性质，强调其在表现和计算性能方面取得的双重优势。尤其是，Transformer 架构在设计大模型方面具备独特优势，这与论文第五部分联系密切。基于这些进展，文中最后简要分析了未来 DNN 算法设计的发展方向。

在第三部分中，本文介绍了在 CMS 实验中校准基于先进 DNN 的大半径喷注标记器的方法上取得的新进展。在将高级的喷注标记算法用于物理分析中，测量其在真实数据和模拟数据上的选择效率的差异，计算用来修正这一差异的缩放因子 (SF)，是不可或缺的步骤。本部分聚焦在校准一类特殊的喷注标记算法，即 boosted 喷注的味道标记算法，它用来标记大半径喷注是否来源于一个共振态 X 衰变到正反底夸克对 ($b\bar{b}$) 或粲夸克对 ($c\bar{c}$) 的过程。这一类标记算法在 Higgs 物理和新物理的分析中具有广泛

应用。对此类标记器的校准本身难度较大，因为难以在数据上找到合适的 $X \rightarrow b\bar{b}/c\bar{c}$ 的代理。利用量子色动力学 (QCD) 导致的多喷注过程中非常常见的，由胶子劈裂，即 $g \rightarrow b\bar{b}$ 诱发的喷注，经过合适的选择条件后作为 $X \rightarrow b\bar{b}$ 的代理，是此前尝试的方案，但是这一校准方法对应用在目前高性能的（如基于 ParticleNet 算法的）标记器上存在挑战。为解决这一难题，本文介绍了全新的 sfBDT 校准方法。它仍基于由 $g \rightarrow b\bar{b}$ 或 $c\bar{c}$ 喷注作为信号 $X \rightarrow b\bar{b}$ 或 $c\bar{c}$ 的代理的思路，但创新性地使用一个基于提升决策树 (BDT) 的多变量分析方法，从 $g \rightarrow b\bar{b}$ 或 $c\bar{c}$ 喷注中选取了更接近 $X \rightarrow b\bar{b}$ 或 $c\bar{c}$ 喷注的独特相空间并以此作为代理。在 BDT 的设计方面，本工作采用产生子水平的变量，将 QCD 多喷注过程产生的喷注分为两类，一类的相空间更接近于信号特征，一类则相反；而后训练 BDT，利用关联的次级顶点、径迹上具备的特征作为训练输入，以区分这两个类别，获得的 BDT 变量即可用于选择出合适的 $g \rightarrow b\bar{b}$ 或 $c\bar{c}$ 相空间作为代理。本工作研发了两代关于 BDT 中用于区分两类的产生子水平变量的设计方案。基于训练好的 BDT，也设计了两种基于 BDT 做事例选择的方式。本校准工作中，对 SF 的提取仍如先前方法一样通过基于模版的最大似然拟合的方法实现，但引入诸多形式的不确定度，例如充分考虑 BDT 选择阈值对 SF 的影响、模拟数据和真实数据在模版上分布的差异对 SF 的影响等，对提取的 SF 施加额外的误差项，使得结果更加可靠。最终，利用 sfBDT 方法得到的 SF 相较 CMS 实验组中另外两个方法是最准确、稳健的，在 CMS 的诸多分析中得到应用。本论文最后综合地展示了利用 sfBDT 方法，对多种 CMS 实验在用的 boosted 喷注味道标记器，在不同选择工作点、不同的喷注横动量区间、不同的取数年份条件下的校准结果。

在第四部分中，本文介绍了在 CMS 实验中利用前沿的 DNN（基于 ParticleNet 网络）训练的大半径喷注标记算法在测量 Higgs 和粲夸克的 Yukawa 耦合强度方面取得的重要物理成果。该工作利用了高性能 $H \rightarrow c\bar{c}$ 喷注标记器 ParticleNet-MD 算法及第三部分讲述的 sfBDT 校准方法。本文首先讲解了该 Higgs 性质测量工作的物理动机。Higgs 机制是标准模型的理论核心，在 高能实验中测量 Higgs 玻色子与标准模型费米子的耦合强度，验证它是否与费米子的质量成正比，是一个用于验证 Higgs 理论、检验标准模型正确性的长久且重要的课题。目前，Higgs 玻色子与第三代费米子的耦合均已被实验证实，因此测量其与第二代费米子（ μ 子、粲夸克和奇夸克）的耦合是下一个里程碑。本工作介绍的是 CMS 实验使用全部 Run-2 数据，利用 Higgs 玻色子衰变到正反粲夸克对（ $H \rightarrow c\bar{c}$ ）并伴随轻子道衰变的 W 或 Z 玻色子产生的物理道测量（称为 VHcc 分析道）得到的对 Higgs 玻色子与粲夸克耦合强度进行限制的结果。这一结果的精度大幅超过了先前在各实验组已经取得的结果，获得较多关注。本文着重展示的分析道是产生的 Higgs 玻色子具有高 Lorentz-boost 的分析道。在此情形下，实验利用一个大

半径喷注来重建具有高横动量 Higgs 玻色子衰变成 $c\bar{c}$ 并强子化后产生的全部末态。这里半径 R 选取为更大的值，即 1.5。在本道中，实验采用了 ParticleNet 架构设计 DNN 进行 $H \rightarrow c\bar{c}$ 大半径喷注进行鉴别，并用回归的方式对 $H \rightarrow c\bar{c}$ 喷注进行质量重建，大幅提升了物理分析敏感度：在 $H \rightarrow c\bar{c}$ 鉴别效率方面，实验实现了约 5 倍的本底过程的压低；在质量重建方面，获得了更高分辨率的质量峰，对最终信号强度的限制结果带来 20–25% 的提升。本分析确立了基于三个重要变量进行事例选择、本底控制和信号抽取。首先，本分析设计了基于事例的运动学变量作为输入（即不涵盖喷注味道、质量信息）训练了 BDT，粗略区分 VHcc 信号过程和主要本底。其次，实验使用了基于 ParticleNet 的标记器极高效地选择出接近 $H \rightarrow c\bar{c}$ 特征的喷注。最后，对信号的提取是通过在由 ParticleNet 算法重建的质量谱上进行拟合得到的。关于本底控制方面，实验通过设计相应的控制区间（CR）以突出不同本底的贡献，并利用信号区间（SR）和控制区间的同时拟合来控制各本底的贡献，其中 CR 主要由翻转 BDT 选择条件来获得。在测量 $VH (H \rightarrow c\bar{c})$ 过程之前，实验还采用了 $VZ (Z \rightarrow c\bar{c})$ 过程对整个分析的策略进行验证。最终的测量结果由本分析道和对 Higgs 玻色子以两个独立的小半径喷注重建的分析道进行同时拟合而获得的，两个分析道选取了相互没有重叠的事例相空间。结果表明，对 $H \rightarrow c\bar{c}$ 过程，实验测得 $VH (H \rightarrow c\bar{c})$ 信号强度在 95% 置信区间的上限值的观测（预期）结果是 14 (7.6)，确立了 Higgs 玻色子与粲夸克耦合的修正比例因子 κ_c 在 95% 的置信区间内的观测（预期）结果是 $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$)，系目前世界上最佳的直接测量 Higgs 与 $c\bar{c}$ 耦合的结果。作为物理验证，在 $Z \rightarrow c\bar{c}$ 伴随矢量玻色子产生的物理道上，实现了强子对撞机上对 $Z \rightarrow c\bar{c}$ 的物理过程的首次观测，其统计显著度达到 5.9σ 。这一结果彰显了前沿 DNN 算法在处理大半径喷注方面，尤其是对喷注进行高性能标记的方面具备的突出优势。

在第五部分中，本文在上面讨论的 ParticleNet 算法在 VHcc 测量的实验应用的基础上更进一步，提出了一项大半径喷注标记相关的全新技术，预期对未来的 CMS 实验具有重要潜力。本章将讨论如何开发一个预训练的、全面的喷注大模型并利用“预训练—微调”范式将其应用到广泛的物理分析中。在本章开始，首先从统计原理出发重新审视了喷注标记任务，对它可能达到的理论上限进行理解，阐述 DNN 算法为何是目前唯一可行的方法来逼近这一上限；随后讨论了未来 DNN 喷注模型的发展方向的诸多观点，并揭示了微调范式如何蕴含在已有的分析设计思路中。本文随后重点讨论了基于 CMS 实验，运用“大模型、多分类”的思路，利用大半径喷注末态进行精细分类后的上百类的喷注样本，使用分类和回归任务混合的方式训练一个通识的、具备预训练特征的喷注大模型 Global Particle Transformer (GloParT)。文中详细介绍了样本产生、喷注分类、网络输入、模型训练等方面的细节，并对模型本身具备的分类和回归的性

能进行展示，显示它几乎所有任务上都获得最佳性能。随后，文中运用两个实例介绍如何用微调的方式开发 GloParT 在具体的 CMS 实验应用当中的潜能。本文工作聚焦在采用迁移学习的方式，即以 GloParT 的潜空间的特征向量作为后续网络的输入的方式进行微调的案例。第一个实例探究了将模型微调到在训练样本中不包含的全新分类任务，即 leptoquark 共振态衰变到 $b\tau$ 末态的喷注标记任务中的能力。第二个实例研究将模型微调以得到其它 CMS 实验中常用的喷注模型的能力，以彰显该模型作为应用层面的“基础喷注模型”的能力。具体地，实验研究了利用 GloParT 潜空间特征并结合额外的喷注特征输入，迁移得到用于顶夸克喷注标记且不与质量去相关的模型。这些研究表明部署一个预训练喷注模型，并借助“预训练—微调”范式预期在未来 CMS 实验中具有广泛的应用价值。本部分最后一章介绍了 GloParT 的训练与应用在基于 DELPHES 快速模拟的数据集上的复现。本文将介绍为此工作研发的 JETCLASS-II 数据集：一个涵盖极宽范围的喷注横动量和质量参数、且覆盖全面的喷注相空间的大半径喷注数据集，及基于此开发的预训练 Sophon 模型。本文展示了利用 Sophon 模型的多输出值与利用其潜空间表示进行模型迁移，为 LHC 上广泛的模型相关与模型无关的新共振态寻找带来的巨大潜力。最后，本文展望了 GloParT 为 CMS 实验的 Run 3 物理分析带来的丰富可能性，以及 JETCLASS-II 数据集与 Sophon 模型可能为未来高能物理基础模型的训练带来的价值。

综上所述，本文从前沿深度学习技术用于大半径喷注标记这一角度切入，综合地总结了在算法开发、CMS 实验上的校准技术与物理应用、以及未来应用潜力等方面，在近年来取得的进展。这些内容系作者在博士研究期间深度参与或主要推动的课题。作为结语，在深度学习技术和人工智能对人类社会愈发产生深刻影响的当下，思考它们会对人类的大科学装置和实验进行怎样的塑造、会使人类对基础科学的探索带来怎样的变化，将具有特殊的时代意义。作者将在后续的研究中继续为推进这一变革做出努力。

关键词：大半径喷注，深度学习，Transformer，校准技术，Higgs 物理，预训练和微调

Extended Abstract (English Edition)

Exploring the most fundamental building block of our universe is humanity's eternal pursuit. While philosophical and scientific exploration spanned centuries, the discovery of the electron in the late 19th century marked the beginning of the subatomic era, paving the way for modern particle physics. The 20th century witnessed extraordinary progress in this field. Key developments included the emergence of quantum mechanics, the formulation of quantum electrodynamics, and the refinement of quantum field theory. These theoretical advances went hand in hand with experimental breakthroughs, many driven by advances in particle accelerator technology. Together, they led to the creation of the standard model (SM) of particle physics—a comprehensive framework that describes the microscopic world. The SM classifies all known fundamental particles into fermions (quarks and leptons) and bosons (gluons, photons, W , Z , and Higgs bosons) and unifies the strong, weak, and electromagnetic interactions. Rooted in fundamental gauge symmetries and the Higgs mechanism, the SM has become a cornerstone of modern physics. Over the last fifty years, its predictions have been consistently confirmed in collider experiments, with the discovery of the Higgs boson at the Large Hadron Collider (LHC) in 2012 serving as a landmark achievement in particle physics.

Despite its remarkable success, the SM is not the ultimate framework for understanding nature. It has significant limitations, including its inability to explain dark matter, unify gravity with the other fundamental forces, or account for neutrino masses, among other issues. To address these gaps, theorists have proposed numerous extensions to the SM that predict new particles and phenomena. Searching for evidence of physics beyond the SM at the LHC has therefore become a top priority. Additionally, precise measurements of SM parameters, especially those involving the Higgs boson, serve as rigorous tests of the theory. Any deviations from theoretical predictions in experimental results could provide critical clues to new physics. These goals define the current experimental programmes at the LHC.

The LHC is home to two general-purpose experiments, ATLAS and CMS, which have been recording proton-proton collision data at TeV energy scales for over a decade. Thousands of physics analyses have been conducted to improve the precision of SM measurements and search for potential signs of new physics. A key challenge for experimental physicists is to continually enhance data analysis methods to achieve greater precision and sensitivity. Since

2012, the rise of modern deep learning has revolutionized statistical and analytical techniques, enabling unprecedented capabilities in handling complex data. Deep neural networks (DNNs) have proven particularly effective in analyzing intricate datasets, often surpassing traditional methods as data complexity increases. In recent years, cutting-edge deep learning applications in the CMS experiment have significantly improved measurement precision and boosted sensitivity to potential new physics signals. As these methods continue to advance, their potential to reshape LHC analyses remains vast. Exploring the full capabilities of deep learning in particle physics holds great promise for transforming experimental approaches and unlocking new discoveries at the LHC.

This dissertation explores state-of-the-art deep learning techniques for large-radius (large- R) jet tagging. In the CMS experiment, quarks or gluons produced in high-energy collisions undergo hadronization, forming collimated sprays of particles known as jets. Jets are the most ubiquitous objects in collision events but pose significant challenges for analysis due to their complex structure. While small-radius jets typically consist of a few dozen reconstructed particles, large- R jets—clustered with a larger radius R —can contain over 100 particles and exhibit intricate internal substructures. These jets are particularly crucial for reconstructing highly Lorentz-boosted resonances. Specifically, high-energy resonance particles with hundreds of GeV to TeV-scale energies produced at the LHC often decay into their full set of final-state particles, which can be entirely contained within a single large- R jet. Such jets offer unique advantages for specific analyses; however, distinguishing signal-like jets from substantial background processes requires sophisticated tagging algorithms. Designing observables to characterize jet properties has long been a central challenge in LHC physics, making this domain highly suited to the application of modern deep learning techniques. Over the past five years, the direct application of advanced DNNs to low-level particle features within large- R jets has delivered significant performance gains, highlighting the transformative potential of deep learning in this area.

This dissertation provides a comprehensive study of advanced deep learning methods for large- R jet tagging, integrating recent developments in DNN algorithms, a novel calibration technique, an experimental breakthrough achieved within the CMS collaboration, and a novel paradigm for future applications. The content is organized into four parts, each addressing a key aspect of the subject. The first three parts focus on challenges encountered when deploying DNN-based taggers in practical physics analyses: (1) algorithm design, which is fundamental to developing high-performance taggers; (2) calibration, a critical step for correcting

simulation-to-data discrepancies when applying taggers to real LHC data; and (3) experimental design, which ensures reliable physics results. The fourth part explores future directions, proposing a paradigm shift in jet tagging through the pre-training of a universal jet model. This approach utilizes a “large models for large-scale classification” methodology, with subsequent fine-tuning tailored to specific analysis needs.

Part I of this dissertation introduces the foundational knowledge of experimental particle physics and deep learning. On the particle physics side, it provides an overview of the SM, the LHC, and the CMS detector, along with Monte Carlo event simulation and reconstruction processes in the CMS experiment. On the deep learning side, it traces the rise of modern deep learning since 2012, focusing on the vital technological advancements that have driven its rapid evolution.

Part II highlights recent advancements in large- R jet tagging algorithms over the past three years. It begins with a historical overview of early methodologies, including rule-based algorithms (e.g. N -subjettiness for jet substructure analysis), approaches utilizing “shallow” machine learning techniques such as boosted decision trees (BDTs), and the initial adoption of deep learning. Early deep learning efforts explore convolutional neural networks and recurrent neural networks to represent jets in formats such as images or sequences, achieving notable performance improvements. However, these methods were eventually surpassed by newer techniques, leading to the emergence of a critical insight: the design of superior DNNs for jet data depends critically on finding better low-level jet feature representations and applying advanced architectures suited to the task. The introduction of the Particle Flow Network and ParticleNet introduced point-cloud and graph-based representations for jets, marking a leap in algorithm design. This part focuses on two critical advancements that have further driven progress in this field: (1) the incorporation of Lorentz symmetry into jet tagging algorithms and (2) the adoption of Transformer-based architectures. For the first advancement, it is widely recognized in deep learning that respecting the intrinsic symmetries of data can enhance both performance and robustness. In the context of jet tagging, earlier efforts focused on specific invariances, such as rotational or boost invariance, but often overlooked the broader and more fundamental invariances under the full set of Lorentz transformations. This part highlights LorentzNet, a network explicitly designed to respect Lorentz symmetry, and demonstrates its superior performance. A systematic study that follows demonstrates how incorporating Lorentz symmetry as an inductive bias can substantially improve network accuracy and robustness. This approach offers general guidelines and insights for integrating symmetry principles into future algorithm

designs. For the second advancement, this dissertation explores the growing prominence of Transformer architectures in modern deep learning and their application to jet tagging. The Particle Transformer (ParT), a representative network built on Transformer principles, is introduced and analyzed. The discussion emphasizes ParT's unique design features, particularly its ability to preserve underlying symmetries through the attention mechanism, and highlights its dual advantages of enhanced expressiveness and computational efficiency. Transformers are particularly well-suited for scaling to larger models, a capability that is explored further in Part V of this dissertation. Based on these developments, the part concludes with a discussion of future directions in DNN algorithm design.

Part III introduces new advancements in calibrating advanced DNN-based large- R jet taggers in the CMS experiment. Measuring the difference in selection efficiency between data and simulation, in terms of the scale factor (SF), for the employed jet tagging algorithms is a critical step in physics analyses. This part focuses on calibrating a specific subset of jet tagging algorithms, namely, the boosted-jet flavour tagging algorithms, which aim to identify whether a large- R jet originates from a resonance X decaying into a bottom or charm quark-antiquark pair ($X \rightarrow b\bar{b}/c\bar{c}$). Such tagging algorithms are widely applied in Higgs physics studies and searches for new physics. Calibrating these taggers is challenging, primarily due to the difficulty in finding suitable proxies for $X \rightarrow b\bar{b}/c\bar{c}$ in data. A previous approach employed jets originating from gluon splitting, specifically $g \rightarrow b\bar{b}$, which are abundant in quantum chromodynamics (QCD) multijet events, as proxies for $X \rightarrow b\bar{b}$ after applying appropriate selection criteria. However, this method faces limitations when applied to high-performance taggers (e.g. those based on the ParticleNet algorithm). To address this challenge, this part introduces a novel calibration method, named the sfBDT method. While retaining the use of $g \rightarrow b\bar{b}/c\bar{c}$ jets as proxies for $X \rightarrow b\bar{b}/c\bar{c}$ signal jets, this method leverages a multivariate analysis approach based on a boosted decision tree (BDT) to identify a unique phase space closely resembling $X \rightarrow b\bar{b}/c\bar{c}$ jets. The BDT is trained to distinguish between two categories of jets from QCD multijet events, defined by a generator-level variable: one category exhibits signal-like characteristics, while the other does not. The BDT training exploits features derived from associated secondary vertex and track information. The resulting BDT discriminant is then used to select an optimized $g \rightarrow b\bar{b}/c\bar{c}$ phase space to act as a proxy. This work introduces two generations of generator-level variable designs for the sfBDT method, along with two approaches for event selection using the BDT discriminant. In the calibration procedure, SFs are extracted via a template maximum likelihood fit, augmented with additional sources of uncertainties, including

the one arising from variations in the BDT selection thresholds. These additional uncertainty components contribute to the robustness and reliability of the extracted SFs. The SFs resulting from the sfBDT method reach good accuracy and robustness and find applications in numerous CMS analyses. The part concludes by presenting the comprehensive calibration results using the sfBDT method for seven $X \rightarrow b\bar{b}/c\bar{c}$ tagging discriminants utilized in CMS analyses, across different working points, jet transverse momentum ranges, and data-taking conditions.

Part IV of this dissertation presents recent progress in measuring the Yukawa coupling strengths of the Higgs and charm quarks within the CMS experiment. This work employed the high-performance $H \rightarrow c\bar{c}$ large- R jet tagger based on the ParticleNet algorithm and the sfBDT calibration method described in Part III. It begins by outlining the physics motivation behind this measurement of the Higgs-charm coupling. As the Higgs mechanism is at the core of the SM, measuring the Higgs boson coupling strengths with SM fermions and verifying their proportionality to the fermions' masses represents a crucial test of the SM's accuracy. To date, the couplings of the Higgs boson with third-generation fermions have been experimentally confirmed; thus, measuring its coupling with second-generation fermions (including muons, charm and strange quarks) will be the next milestone. This work presents the analysis using the full CMS data collected during Run 2 to measure the Higgs boson coupling strength to charm quarks through the $H \rightarrow c\bar{c}$ decay accompanied by the leptonically decayed W or Z boson (referred to as the VHcc analysis). The VHcc analysis achieves a substantially higher-precision measurement compared with previously established measurements. The work presented in this dissertation focuses on scenarios where the produced Higgs boson has a high Lorentz boost. In these cases, the analysis employs a large- R jet to reconstruct all final states produced by the hadronization of the high-momentum Higgs boson decaying into $c\bar{c}$, with a larger radius R of 1.5. The analysis exploits a DNN with the ParticleNet architecture for $H \rightarrow c\bar{c}$ large- R jet tagging and mass reconstruction via regression, significantly enhancing the physics sensitivity: in terms of $H \rightarrow c\bar{c}$ identification, the experiment achieved about a fivefold suppression of background events compared to the previous analysis; for mass reconstruction, it obtained a peak with higher resolution, enhancing the signal strength upper limit extraction by 20–25%. This analysis establishes the event selection, background estimation, and signal extraction based on three key variables. First, a BDT is trained with kinematic variables (excluding jet flavour and mass information) to provide a coarse separation between the VH ($H \rightarrow c\bar{c}$) signal and the main background processes. Second, the ParticleNet-based tagger is exploited to efficiently select jets closely resembling $H \rightarrow c\bar{c}$ characteristics. Finally, signal extraction is achieved

by fitting the reconstructed mass spectrum regressed via the ParticleNet algorithm. For background estimation, dedicated specific control regions are designed to highlight contributions from different backgrounds. A simultaneous fit of all signal and control regions is performed to control the background for each process. The control regions are constructed by reversing BDT selections and other kinematic selections. Before measuring the VH ($H \rightarrow c\bar{c}$) process, the entire analysis strategy is validated using the VZ ($Z \rightarrow c\bar{c}$) process. The $VH \rightarrow c\bar{c}$ result is obtained by performing a simultaneous fit with this analysis channel, dubbed the merged-jet analysis channel, and the other resolved-jet analysis channel that reconstructs the Higgs boson with two separate small- R jets, using orthogonal event phase spaces between the two channels. The results show that for the $H \rightarrow c\bar{c}$ process, the observed (expected) upper limit on the $VH \rightarrow c\bar{c}$ signal strength at a 95% C.L. was 14 (7.6), establishing the observed (expected) range for the Higgs-charm coupling modifier κ_c within a 95% C.L. as $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$), representing the best direct measurement of the Higgs boson to $c\bar{c}$ coupling to date. As a validation of the analysis strategy, the process of $Z \rightarrow c\bar{c}$ production accompanied by a vector boson achieved the first observation of the $Z \rightarrow c\bar{c}$ process at a hadron collider, with a statistical significance of 5.9σ . This result highlights the exceptional advantage of advanced DNN algorithms in processing large- R jets, especially in high-performance jet tagging.

Part V introduces a novel approach to large- R jet tagging, building upon the experimental application of the ParticleNet algorithm in the VH_{cc} measurement and integrating advancements in deep learning. The work explores the development of a pre-trained, comprehensive and large jet model for large- R jets and employs the “pre-training and fine-tuning” paradigm to extend its applicability across a wide range of physics analyses. The discussion begins by revisiting the jet tagging problem from statistical principles, understanding the theoretical limits of the problem and explaining why DNN algorithms are helpful in approaching this limit. It further examines various potential directions for the future development of DNN-based jet models, with an emphasis on the emergence of the fine-tuning paradigm, inspired by established analysis design principles. The focus then shifts to developments within the CMS experiment, leveraging the “large model for large-scale classification” strategy. This involves a combination of classification and regression tasks to train a versatile, pre-trained large jet model named Global Particle Transformer (GloParT), developed using hundreds of finely categorized large- R jet classes. Detailed information is provided on the sample production process, jet categorization, model input design, and training methodology. The model’s performance across a wide range of classification and regression tasks is showcased, demon-

strating its ability to achieve optimal performance in nearly all evaluated scenarios. On top of this, the following chapter uses two examples to demonstrate how fine-tuning can be employed to unlock the potential of GloParT for specific applications within the CMS experiment. These examples focus on cases of transfer learning, where GloParT’s latent space features serve as input for subsequent networks, a suitable approach for integration into the current experimental workflow. The first example explores fine-tuning the model to identify jets from a new category that is not included in the training, i.e. a leptoquark resonance decaying into a $b\tau$ final state. The second example investigates the model’s ability to derive other jet models commonly used in CMS analyses, showcasing its capabilities as a “base jet model” for analysis-level applications. Specifically, the model is transferred to a top-jet tagging model that does not require mass decorrelation, utilizing GloParT’s latent features combined with additional jet mass and momentum inputs. These studies demonstrate that deploying a pre-trained jet model and facilitating the “pre-training and fine-tuning” paradigm is expected to have a broad application value in future CMS analyses. The final chapter of this part presents the reproduction of the GloParT training and application scenarios on datasets generated via DELPHES’s fast simulation. This work introduces the development of the JETCLASS-II dataset—a large- R jet dataset spanning an extensive range of transverse momentum and mass values and covering comprehensive jet phase space—and the Sophon model, which is pre-trained on this dataset. This work highlights the Sophon model’s capability of producing multi-output probabilities and utilizing its latent space representations for model transfer, revealing its tremendous potential for both model-specific and model-agnostic searches for new resonances at the LHC. The part concludes with a discussion on the large potential of GloParT to contribute to Run 3 physics analyses in the CMS experiment, along with the significant value of the JETCLASS-II dataset and the Sophon model for training foundation models in future high-energy physics research.

In summary, this dissertation approaches the topic from the perspective of cutting-edge deep learning technologies applied to large- R jet tagging, providing a comprehensive overview of recent advancements in algorithm development, calibration techniques within the CMS experiment, the physics application, and its potential for future applications. These topics reflect research projects in which the author was deeply involved or served as a principal contributor during the doctoral studies. As a final remark, it is a compelling era to witness the transformative impact of deep learning technologies on scientific research and beyond. Their capacity to drive innovation in large-scale scientific programmes and to enhance our understanding of fundamental science demonstrates their profound significance. The author aims to continue

contributing to this transformative journey through future research endeavours.

KEY WORDS: large- R jets, deep learning, Transformer, calibration, Higgs physics, pre-training and fine-tuning

Abstract

Modern deep learning technologies have demonstrated remarkable potential in analyzing complex LHC data, accelerating foundational scientific discoveries at the LHC. This doctoral dissertation focuses on modern deep learning techniques for tagging large- R jets, one of the most challenging physics objects to analyze at the LHC. It discusses novel methods, applications, and potentials in jet research. Surrounding this technology, the dissertation systematically summarizes the forefront progress in various aspects, involving advanced deep neural network (DNN) algorithms, calibration methods developed in the CMS experiment, breakthroughs in measuring Higgs boson properties, and discusses the new application paradigms of this technology in future experiments.

The dissertation is organized into five parts. The first part introduces the fundamentals of experimental particle physics and deep learning. The second part discusses new developments in the field of large- R jet tagging algorithms in the past three years, with a special focus on the algorithm designs that respect Lorentz symmetry and DNN algorithms based on the Transformer architecture. The third part introduces a new method for calibrating advanced DNN-based $X \rightarrow b\bar{b}$ or $c\bar{c}$ jet tagging algorithms in the CMS experiment, namely the sfBDT calibration method: this method uses multivariate analysis to select a phase space closer to $X \rightarrow b\bar{b}$ or $c\bar{c}$ jets in gluon-splitting $b\bar{b}$ or $c\bar{c}$ jets as a proxy for calibration, tackling the problem of previous methods failing to calibrate advanced DNN-based taggers. The fourth part presents significant physics results obtained using cutting-edge DNN algorithms in the CMS experiment in measuring the Higgs-charm Yukawa coupling strengths: this work aims to explore the phase space of highly Lorentz-boosted Higgs boson decays into a charm quark-antiquark pair ($c\bar{c}$) reconstructed by a large- R jet, using the advanced ParticleNet algorithm for $H \rightarrow c\bar{c}$ jet tagging and mass regression, resulting in the best direct measurement of κ_c and marking the first observation of the $Z \rightarrow c\bar{c}$ process at a hadron collider. Finally, the fifth part introduces a novel application paradigm for large- R jet tagging by leveraging advancements in deep learning. It emphasizes the development of pre-trained, comprehensive jet models and the adoption of the “pre-training and fine-tuning” paradigm to extend their applicability to diverse downstream tasks. Key contributions include the Global Particle Transformer (GloParT) model, developed within the CMS collaboration, and the Sophon model, trained on the newly

developed JETCLASS-II dataset. Extensive experiments demonstrate their significant potential for CMS Run 3 analyses and future research.

Through a comprehensive discussion of the forefront progress of modern deep learning technology applied to large- R jet tagging, the dissertation reveals their transformative impact on particle physics, particularly in the CMS experiment. It highlights the growing role of deep learning in contemporary scientific research patterns.

KEY WORDS: large- R jets, deep learning, Transformer, calibration, Higgs physics, pre-training and fine-tuning

Contents

Part I	Foundations	1
Chapter 1	Experimental Particle Physics	3
1.1	The standard model	3
1.2	The Large Hadron Collider	9
1.3	The CMS detector	10
1.4	Monte Carlo simulation in CMS	15
1.5	Event reconstruction in CMS	17
Chapter 2	Deep Learning	25
2.1	The path to modern deep learning	25
2.2	Advances in deep learning	26
2.3	Deep learning dataset	31
Part II	Algorithms for Jet Tagging	35
Chapter 3	An Overview on Algorithms	37
3.1	Rule-based algorithms	37
3.2	Shallow machine-learning algorithms	41
3.3	Deep learning algorithms	42
Chapter 4	New Advances in Deep Learning Algorithms	47
4.1	Algorithms preserving Lorentz symmetry	47
4.2	The Transformer algorithm	61
Part III	Calibration of Jet Taggers	69
Chapter 5	Introduction to Boosted-jet Flavour Tagger Calibration	71
5.1	Overview of boosted-jet flavour taggers	71
5.2	Overview of tagger calibration	76

Chapter 6	A Novel Calibration Method: sfBDT	79
6.1	Data and simulated samples	79
6.2	Calibrated taggers in CMS	82
6.3	Event and jet selection	85
6.4	Scale factor derivation	102
6.5	Calibration results and discussions	108

Part IV Application in Higgs-Charm Coupling Measurement 115

Chapter 7	Introduction to Higgs-Charm Coupling Measurement	117
7.1	Physics motivation	117
7.2	Overview of relevant measurements	119
7.3	Implications of deep learning	120
Chapter 8	Analysis Method	123
8.1	Data and simulated samples	124
8.2	Jet tagging and mass regression techniques	124
8.3	Reconstruction of objects and events	134
8.4	Analysis selection	136
8.5	Signal extraction	148
8.6	Systematic uncertainties	151
Chapter 9	Analysis Results	155
9.1	$VZ \rightarrow c\bar{c}$ validation results	155
9.2	$VH \rightarrow c\bar{c}$ results	159
9.3	Combined $VH \rightarrow c\bar{c}$ results with resolved-jet analysis	165
9.4	Discussion and outlook	166

Part V Application Paradigm for Pre-training and Fine-tuning 169

Chapter 10	Perspectives on Future Jet Models	171
10.1	Jet tagging problem revisited	171
10.2	Future models	173
10.3	Fine-tuning techniques	176

Chapter 11	GloParT: Model Pre-training.....	179
11.1	Generated samples and jet labels.....	179
11.2	Development details	187
11.3	Overall performance	197
11.4	Ablation experiments	204
Chapter 12	GloParT: Model Fine-tuning.....	211
12.1	Application on novel tagging.....	212
12.2	Application on deriving new models.....	213
Chapter 13	The Sophon model	221
13.1	The JetClass-II dataset.....	222
13.2	Pre-training of the Sophon model	225
13.3	Performance benchmark	226
13.4	Implications to LHC resonance search.....	230
13.5	Discussion and outlook	239
Chapter 14	Outlook	243
Appendix A	Additional sfBDT Calibration Results.....	247
Appendix B	Statistical Framework of Classification and Regression	251
B.1	Multiclass classification	251
B.2	Regression	252
References		255
作者贡献	Author Contributions.....	269
致谢	Acknowledgement	275
北京大学学位论文原创性声明和使用授权说明		281

List of Tables

Table 2.1	Summary of the particle input features used for the Top, QG (QG_{exp} and QG_{full}), and JETCLASS dataset, adapted from Ref. ^[65] . Customizations are done including the expansion of kinematic features and the use of logarithms and hyperbolic tangents for input standardization.	33
Table 4.1	Invariance property of the pairwise variables between particles when the jet undergoes a certain type of Lorentz transformation.	55
Table 4.2	Comparison of network performance of the ParticleNet and LorentzNet _{base} variants which incorporate different pairwise features. For each metric, the best-performing network from each baseline type is highlighted in bold text.	57
Table 4.3	Comparison of network performance of the PFN, ParticleNet and LorentzNet _{base} variants incorporating different nodewise features. For each metric, the best-performing network from each baseline type is highlighted in bold text.	61
Table 4.4	Performance of ParticleNet and ParT trained with different data sizes of JETCLASS. $\text{Rej}_{X\%}$ refers to $1/\epsilon_{q/g}$ when $\epsilon_S = X\%$	62
Table 4.5	Comparison of jet tagging performance on the top tagging dataset with multiple algorithms. The best-performing model is highlighted in bold text.	66
Table 4.6	Comparison of jet tagging performance on the quark–gluon tagging dataset with multiple algorithms. The subscript “exp” and “full” distinguish models using partial or full particle identification information. The best-performing model is highlighted in bold text.	67
Table 4.7	Comparison of jet tagging performance on the JETCLASS dataset with multiple algorithms, including LorentzNet. The best-performing model is highlighted in bold text. $\text{Rej}_{X\%}$ refers to $1/\epsilon_{q/g}$ when $\epsilon_S = X\%$	67
Table 4.8	Number of trainable parameters and floating-point operations per second (FLOPs).	67
Table 6.1	A description of the high-level triggers used in this sfBDT study in different data-taking conditions.	80

Table 6.2	The list of simulated MC signal samples used for deriving the tagger working points, for the data-taking year 2016–2018. The matrix-element (ME) and parton-shower (PS) generators are also shown.	80
Table 6.3	The list of simulated MC samples for the sfBDT calibration study in data-taking year 2016–2018. The matrix-element (ME) and parton-shower (PS) generators, prediction order for the total cross section, and the cross section values are shown.	81
Table 6.4	Definition of the three working points (WPs) for each calibrated $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ tagging discriminants to derive the scale factors, shown in four data-taking year conditions. The three WPs correspond to selecting with the efficiency of 40%, 60% and 80% of the $H \rightarrow b\bar{b}$ signal jets, and the efficiency of 15%, 30% and 50% of the $H \rightarrow c\bar{c}$ signal jets.	83
Table 6.5	Details of the 3D binning on (event H_T , jet p_T , jet index $\in \{1, 2\}$) used for the MC-to-data reweighting. When the jet index and p_T range are fixed, the H_T adopts an equal-width binning with a varying upper and lower bound and a fixed width of 50 GeV with flows included. The H_T boundaries in the 3D grid are shown.	88
Table 6.6	The truth-matching criteria to assign a large- R jet from the QCD multijet events to one of the five exclusive categories: bb, b, cc, c, and light.	89
Table 6.7	The list of variables used in the sfBDT training to discriminate the signal-like jet versus gluon-contaminated jet.	92
Table 6.8	Summary of the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately. These SFs are part of the published results in Ref. ^[124]	111
Table 6.9	Summary of the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately. These SFs are part of the published results in Ref. ^[124]	112
Table 8.1	The input list of variables for the ParticleNet-MD AK15 tagger training, including features of each jet constituent PF candidate and SVs. The track-based variables are only defined for charged PF candidates, and a value of 0 is assigned to neutral candidates.....	127
Table 8.2	Definition of the three $X \rightarrow c\bar{c}$ tagging categories for the $VH \rightarrow c\bar{c}$ boosted-jet analysis.	128
Table 8.3	The calibration results of the $X \rightarrow c\bar{c}$ tagging discriminant on large- R ($R = 1.5$) $H \rightarrow c\bar{c}$ signal jets in terms of scale factors. The calibration is dependent on jet p_T and data-taking conditions and is performed using the sfBDT method.	130

Table 8.4	The regressed jet mass JMS and JMR (relative to the nominal 5% and 10%) obtained from the template fit of the $W \rightarrow cx$ component.	134
Table 8.5	A list of variables used in the training of the kinematic BDT for the 0L, 1L and 2L channels.	147
Table 8.6	A summary of event selection criteria for the signal region. Note that the values listed for kinematic variables are in units of GeV, and for angles in units of radians; the selections are listed in the format of (muon, electron) if they are varied between lepton flavours.	147
Table 9.1	Expected and observed significance and best-fit signal strength for $VZ \rightarrow c\bar{c}$.	156
Table 9.2	Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2016 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.	162
Table 9.3	Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2017 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.	162
Table 9.4	Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2018 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.	163
Table 9.5	Observed and expected upper limits at 95% CL on the $VH \rightarrow c\bar{c}$ signal strength $\mu_{VH \rightarrow c\bar{c}}$ and the best-fit value.	163
Table 11.1	The model parameters and training configurations for the ParT model in three sizes: ParT-lite, ParT-B and ParT-L (GloParT).	191
Table 11.2	The list of input variables used in the GloParT model development, grouped into three collections: the charged PF candidates + lost tracks, the neutral PF candidates, and the SVs. The input collections or features are marked as (★) to highlight the difference with the ParticleNet training inputs listed in Table 8.1.	193
Table 13.1	Summary of the 188 jet labels in the JETCLASS-II dataset.	224
Table A.1	Summary of the DeepDoubleX $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.	247

Table A.2	Summary of the DeepAK8-MD $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.	248
Table A.3	Summary of the double-b $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.	248
Table A.4	Summary of the DeepDoubleX $X \rightarrow c\bar{c}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.	249
Table A.5	Summary of the DeepAK8-MD $X \rightarrow c\bar{c}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.	249

List of Figures

Figure 1.1	A visualization of all elementary particles in the SM and their properties summarized in Ref. ^[17]	4
Figure 1.2	The interaction of SM particles representative by the five terms in the SM Lagrangian in the format of Feynman diagrams ^[18]	5
Figure 1.3	The Feynman diagrams of the main Higgs production modes (left) and the Higgs decay modes (right). Image adapted from Ref. ^[19]	7
Figure 1.4	The Higgs boson production cross sections in various production modes as a function of the centre-of-mass energy ^[20] (left), and the branching functions as a function of the Higgs boson mass ^[20] (right).	8
Figure 1.5	A schematic view of the LHC ^[21]	9
Figure 1.6	Timeline for the LHC and the High-Luminosity LHC. Source: <i>CERN website</i> . 10	
Figure 1.7	A cutaway diagram of the CMS detector ^[22]	11
Figure 1.8	Illustration of CMS coordinate system.	12
Figure 1.9	An R - z cross section of a half of the CMS detector, showing the layout of the inner tracker system ^[23]	13
Figure 1.10	An R - z cross section of a quadrant of the CMS detector, showing the layout of the muon detector ^[24]	15
Figure 1.11	A pie chart illustrating the distribution of usage across various event generators and their combinations, represented by the number of events generated.	17
Figure 1.12	An event display of a toy jet initiated by two charged particles π^+ and π^- , one neutral hadron K_L^0 , and two photons from the π^0 decay, shown in the (x, y) coordinate ^[31] . The symbols $T_{1,2}$, $E_{1,2,3,4}$ and $H_{1,2}$ represent tracks and energy deposit in ECAL and HCAL measured by the CMS subdetectors corresponding to these particles.	19
Figure 2.1	A view of a typical neural network. Source: <i>Mathworks</i>	25
Figure 2.2	The prototype of convolutional neural networks and recurrent neural networks.	28
Figure 2.3	The prototype of a message-passing layer in the graph neural networks ^[55]	29
Figure 2.4	Illustration of the Transformer self-attention mechanism.	30

Figure 3.1	A schematic image of the four jet grooming algorithms: jet trimming, jet pruning, jet filtering, and the soft drop algorithm. Image source: <i>CMS Open Data workshops</i>	38
Figure 3.2	The shape comparison for multiple resonant types of jets and QCD background jets on the N -subjettiness ratio τ_{21} and τ_{32} , and generalized energy correlation function ratio $N_2^{[70]}$	39
Figure 3.3	Illustration of a heavy-flavour jet initiated from the decay of a b or c hadron resulting in tracks that are displaced with respect to the primary vertex (PV), and possibly reconstructed SV within the jet ^[75]	41
Figure 3.4	The image representation of top jets shown as an ensemble of overlaid images (left) and a schematic diagram of the ImageTop algorithm for boosted top jet tagging (right) ^[70]	43
Figure 3.5	A schematic diagram of the DeepDoubleX algorithm (V2) ^[86] for boosted heavy-resonant $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging.	43
Figure 3.6	A detailed explanation of the ParticleNet architectural design.	45
Figure 4.1	Illustration of a toy jet on the η - ϕ plane (a) and its behaviour when it undergoes the four types of Lorentz transformation with its centre fixed at the origin of the η - ϕ plane, which is y - z rotation (b), x - t boost (c), z tilt (d), and y tilt (e).	49
Figure 4.2	A schematic diagram of the Lorentz group equivariant block (LGEB) within LorentzNet, highlighted by three steps. Image adapted from Ref. ^[94]	51
Figure 4.3	Comparison of jet tagging performance in terms of tagging accuracy with different algorithms when input jets undergo a x - t boost with parameter $\beta = v/c$ ^[94]	53
Figure 4.4	A schematic diagram of the modified ParticleNet architecture to incorporate the pairwise features. The modification is drawn in a red background to be distinguished from the original ParticleNet structure.	56
Figure 4.5	Top: Comparison of network performance on Top dataset in terms of AUC among different training data size, selected among 6 k, 20 k and 60 k, for the baseline network as ParticleNet (left) and LorentzNet _{base} (right). Bottom: Extended comparison delivered on JETCLASS dataset with larger data size, selected among 60 k, 200 k, 500 k, 2 M, 20 M and 100 M.	58
Figure 4.6	Comparison of network performance in terms of AUC evaluated under various types of Lorentz transformation applied to the test dataset. They correspond to the input jet undergoing a y - z rotation with angle α_x (left), an x - t boost with rapidity w_x (middle), and a z -tilt transformation, i.e., a z - t boost with rapidity w_z followed by a x - z rotation to restore the jet direction (right). The baseline is chosen as ParticleNet (top) or LorentzNet _{base} (bottom).	59

Figure 4.7	A schematic diagram of the universal network-modifying recipe for all baseline networks to incorporate additional nodewise features. The modification is drawn in a red background to be distinguished from the original network structure.	60
Figure 4.8	A schematic diagram of the Particle Transformer architecture (a), the Particle Attention Block (b) and Class Attention Block (c).	63
Figure 5.1	Illustration on large- R jet topologies with different origins. The first two cases fall under the category of boosted-jet flavour tagging.	72
Figure 5.2	The normalized distributions of the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants on the simulated signal $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets, the bb and cc components of QCD multijet background jets, and inclusive QCD jets. From top to bottom: the DeepAK8-MD, DeepDoubleX, and ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants.	77
Figure 6.1	Performance comparison of the $X \rightarrow b\bar{b}$ (top) and $X \rightarrow c\bar{c}$ (bottom) identification algorithms in terms of ROC curves for $H \rightarrow b\bar{b}$ or $H \rightarrow c\bar{c}$ signal jets versus the inclusive QCD jets as background. Performance is shown in the $450 < p_T < 600$ GeV (left) and $p_T > 600$ GeV (right) regions.	84
Figure 6.2	The mass decorrelation performance on the QCD multijet background for various $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants. From top to bottom: The ParticleNet-MD, DeepDoubleX, DeepAK8-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants.	86
Figure 6.3	The number of data events passing the jet H_T triggers with different thresholds and passing the logical OR of all triggers, for the 2016 post-VFP (left) and 2018 (right) detector conditions as an example. The prescaling effect is demonstrated by low-threshold HLTs, and the number in parentheses indicates the prescaling factor for the corresponding trigger.	87
Figure 6.4	Summary of the MC and data yields (upper panel for each plot) and the MC-to-data reweighting factors (lower panel) for the 2018 data on the 3D variables (event H_T , jet p_T , jet index). The H_T ranges are shown in the x axis; jet p_T ranges are separated by plots where three example ranges of (200, 250), (400, 450), and (600, 700) GeV are shown from left to right; the jet index takes the value of 1 and 2.	87
Figure 6.5	The κ_g distributions of large- R jets for the signal $H \rightarrow c\bar{c}$ jets and the $g \rightarrow c\bar{c}$ jets selected from QCD multijet events.	91

Figure 6.6	A 2D histogram on the tagging discriminant (ParticleNet-MD $X \rightarrow c\bar{c}$ developed for AK15 jets) versus the sfBDT score, for the $H \rightarrow c\bar{c}$ (left) and the $g \rightarrow c\bar{c}$ (right) jets ^[125] . The $g \rightarrow c\bar{c}$ histogram has two enriched regions, one in the top-right region that resembles $H \rightarrow c\bar{c}$ signal jets, and one in the bottom-left corner representing jets with more gluon contamination. A selection involving the sfBDT helps to select a dedicated phase space that improves the proxy and signal jet similarity. Two types of selection are considered depending on the specific signal and proxy condition, as illustrated in the right plot.	93
Figure 6.7	The distributions of the tagging discriminant (ParticleNet-MD $X \rightarrow c\bar{c}$ developed for AK15 jets) ^[125] . Both equal-width binning (left) and analysis-defined working points ^[112] (right) are shown. The histograms demonstrate that the sfBDT is capable of tuning the tagger shape and improving the similarity between the signal and proxy shape.....	93
Figure 6.8	The cartoon of a $g \rightarrow b\bar{b}$ ($c\bar{c}$) jet with an additional parton emission in the format of Feynman diagrams.	94
Figure 6.9	The example plots for the tagger transformation map, the original tagger shape and the transformed tagger shape for the ParticleNet-MD (top) and DeepDoubleX (bottom) $X \rightarrow b\bar{b}$ discriminants.....	96
Figure 6.10	The 2D histogram of the tagger–sfBDT variable (the coloured pixels), and Illustration of the “sfBDT coastlines” (the curves). The 2D histogram showcases the 2D PDF $f(x, y)$, and the coastlines are mathematically defined as the contour of the accumulative PDF $F_y(x, y) = \int_y^1 f(x, y') dy'$	97
Figure 6.11	The sfBDT coastlines derived in the context of calibrating the ParticleNet-MD $X \rightarrow b\bar{b}$ discriminants in three exclusive p_T bins.	97
Figure 6.12	The sfBDT coastlines derived in the context of calibrating the ParticleNet-MD $X \rightarrow c\bar{c}$ discriminants in three exclusive p_T bins.	98
Figure 6.13	The distribution of the tagging discriminant of proxy jets with (left) and without (right) normalized to unity, in the case of calibrating ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant at HP and after applying the sfBDT coastline selection in nine options. In the normalized plot, the signal $H \rightarrow b\bar{b}$ jet distribution is also included for comparison.	98
Figure 6.14	Comparison of the signal and proxy jets on a collection of jet observables. The observables are related to high-level properties of the jet, subjets, tracks, and SVs. See the detailed observable list in the text.	99
Figure 6.15	Illustration on the choices of “sfBDT coastline” selections applied for an individual fit.	101

Figure 6.16	The separation power of signal and proxy jets illustrated in terms of the ROC curve, where the proxy jet collection is defined by a selection on the sfBDT coastline over all 81 possible options. The plot corresponds to the calibration of the ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant at HP.	101
Figure 6.17	Data and MC comparison on various jet kinematics and sfBDT in 2018 data-taking condition, before the sfBDT selection is applied. The distributions for the jet variable m_{SD} , η , and sfBDT discriminant is shown, under two example p_T ranges.	103
Figure 6.18	Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow b\bar{b}$ (left) and $X \rightarrow c\bar{c}$ (right) tagging discriminants in 2018 data-taking condition, before the sfBDT selection is applied.	104
Figure 6.19	Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminants in 2018 data-taking condition, after the sfBDT coastline selection applied on the leftmost (left), the middle (middle), and the rightmost (right) positions.	104
Figure 6.20	Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminants in 2018 data-taking condition, after the sfBDT coastline selection applied on the leftmost (left), the middle (middle), and the rightmost (right) positions.	105
Figure 6.21	Data and MC comparison on the fit variable for the inclusive “pass”+“fail” region, passing the middle sfBDT coastline selection.	105
Figure 6.22	Pre-fit and post-fit distributions on the fit variable for the “pass” (left) and “fail” (right) region in the context of calibrating the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant at HP, under the 2018 data-taking conditions. The fit corresponds to applying the middle sfBDT coastline selection and is performed under three jet p_T ranges (from top to bottom).	109
Figure 6.23	Pre-fit and post-fit distributions on the fit variable for the “pass” (left) and “fail” (right) region in the context of calibrating the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant at HP, under the 2018 data-taking conditions. The fit corresponds to applying the middle sfBDT coastline selection and is performed under three jet p_T ranges (from top to bottom).	110
Figure 6.24	Summary of the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant SFs in the plot format, separately shown for four data-taking conditions. The displayed results are excerpted from those published in Ref. ^[124] .	113
Figure 6.25	Summary of the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant SFs in the plot format, separately shown for four data-taking conditions. The displayed results are excerpted from those published in Ref. ^[124] .	114

Figure 7.1	The measurement for the reduced Higgs-coupling modifiers extracted for fermions and weak bosons ^[156]	118
Figure 7.2	The Feynman diagrams of the $VH \rightarrow c\bar{c}$ process at the leading order.	119
Figure 7.3	The simultaneous fit result of κ_b and κ_c , which are indirectly constraint by the shape and normalization of the differential Higgs boson cross section on p_T ^[167]	121
Figure 8.1	Performance of ParticleNet-MD AK15 $X \rightarrow c\bar{c}$ tagging discriminant for identifying $H \rightarrow c\bar{c}$ jets versus the V +jets and the $H \rightarrow b\bar{b}$ backgrounds in terms of the ROC curve.	128
Figure 8.2	The mass resolution obtained with the soft-drop (left) and mass regression (right) algorithms on large- R ($R = 1.5$) $H \rightarrow c\bar{c}$ jets for various p_T ranges.	129
Figure 8.3	The mass resolution obtained with the soft-drop (left) and mass regression (right) algorithms on large- R ($R = 1.5$) QCD jets for various p_T ranges.	129
Figure 8.4	Comparison of the log-cosh loss function shape with the L1 and L2 loss function shapes.	131
Figure 8.5	The mass resolution as a function of jet p_T obtained with the soft drop and the mass regression algorithms (left); the efficiency of an event selection $ m - m_H < 2\sigma_{\text{eff}}$ as a function of jet p_T (right).....	132
Figure 8.6	The expected gain in significance by using the regressed mass compared with using the soft-drop mass as a function of p_T	133
Figure 8.7	The post-fit distributions and data–simulation agreement in the single-muon $t\bar{t}$ control sample as a function of leading large- R jet regressed mass after passing (top) and failing (bottom) the ParticleNet $X \rightarrow c\bar{c}$ tagging discriminant at LP.....	134
Figure 8.8	The distributions of the input variables used for training the kinematic BDT in the 2L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds. The definition of the variables can be found in Table 8.5.	138
Figure 8.9	The kinematic BDT output distributions after four sets of selections in the 2L channel, i.e. the baseline selection (top left, from Ref. ^[113]), the $X \rightarrow c\bar{c}$ tagging category selection at HP (top right), the selection at MP (bottom left), and the selection at LP (bottom right). Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds.....	139
Figure 8.10	Comparison of the kinematic BDT output distribution in the 2L channel before and after the requirement on the $X \rightarrow c\bar{c}$ tagging discriminant.	140

Figure 8.11	Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high kinematic BDT output scores in the 2L channel ^[113] .	140
Figure 8.12	The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 1L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.	141
Figure 8.13	The kinematic BDT output distributions after the baseline selection in the 1L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds ^[113] .	143
Figure 8.14	Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high (left, from Ref. ^[113]) and medium (right) kinematic BDT output scores in the 1L channel.	144
Figure 8.15	The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 1L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.	144
Figure 8.16	The kinematic BDT output distributions after the baseline selection in the 0L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds ^[113] .	145
Figure 8.17	Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high kinematic BDT output scores in the 0L channel ^[113] .	146
Figure 8.18	The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 0L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.	146
Figure 8.19	Illustration of the setup of the signal region and control regions.	149

Figure 8.20 Comparison of the H_{cand} mass shapes between events with $N_j < 2$ and $N_j \geq 2$ for the top-matched (left), W -matched (middle) and other (right) components of the $t\bar{t}$ process in the 0L channel.	150
Figure 9.1 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the low-BDT control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$	157
Figure 9.2 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the high N_j control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$	157
Figure 9.3 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the high-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$	158
Figure 9.4 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the medium-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$	158
Figure 9.5 Post-fit plots for the low-BDT control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$. ..	159
Figure 9.6 Post-fit plots for the high N_j control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$. ..	160
Figure 9.7 Post-fit plots for the high-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$. ..	160
Figure 9.8 Post-fit plots for the medium-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(ev)$. ..	161
Figure 9.9 Post-fit values and the impacts of the selected nuisance parameters on the $VH \rightarrow c\bar{c}$ signal strength, focusing on those with the largest impacts.	164
Figure 9.10 The profile likelihood scan on the signal strengths $\mu_{VH \rightarrow b\bar{b}}$ and $\mu_{VH \rightarrow c\bar{c}}$ projected to HL-LHC corresponding to the data of 3000 fb^{-1} . The projection is based on the merged-jet analysis workflow with proper estimation of the statistical uncertainties ^[113, 186]	167
Figure 10.1 The evolution of the large language model size (by 2022). Source: <i>Hugging face</i>	175
Figure 10.2 Illustration on the connection from the conventional analysis-level neural network training strategy to the concept of transfer learning.....	177

Figure 11.1	The 314 jet categories for multiclass classification used in the GloParT model development.	180
Figure 11.2	The GEN-level m_H and m_W configurations for all produced $H \rightarrow W^{(*)}W^{(*)}$ samples used for training and evaluating the GloParT model.....	183
Figure 11.3	The 2D distribution of $(m_H - m_{W_1} - m_{W_2})/m_H$ and $(m_{W_1} - m_{W_2})/(m_{W_1} + m_{W_2})$ for $H \rightarrow W^{(*)}W^{(*)}$ jets in the spanned- m_W/m_H sample.	184
Figure 11.4	A illustration of the matching criteria from the GEN-level quarks or leptons to the jet axis to define the category. The subplots show four possible final states resulting from $H \rightarrow W^{(*)}W^{(*)}$ jets.....	186
Figure 11.5	Illustration on the fact that each class of jets corresponds to a PDF in an abstract jet phase space Ω , and that a binary or multiclass classifier network trained with the cross-entropy loss is learned to estimate the a normalized PDF ratio.....	189
Figure 11.6	Illustration of the parameter details of the ParT-L (GloParT) model. The outline of the ParT architecture is adapted from Ref. ^[65]	191
Figure 11.7	The normalized distributions of the number of charged PF candidates + lost tracks, the neutral PF candidates, and the SVs for several typical jet categories.....	194
Figure 11.8	The point-cloud visualization of selected jets in different categories. In each visualization, the position of all PF candidates and SVs is marked on the η - ϕ plane, displayed with their energies or other properties (see text for details). The GEN-level properties of the jet are also annotated in the figure.	195
Figure 11.9	A summary of the reweighting classes and class weights used in the GloParT model development.....	196
Figure 11.10	The metrics monitored during the training of the ParT-lite (orange), ParT-B (blue), and ParT-L (red) model, including the classification accuracy (top), the classification loss (middle), and the regression loss (bottom), for each epoch of the training (left) and evaluation (right) procedure.	198
Figure 11.11	Illustration on the fact that the binary score $A/(A + B)$ constructed from the model should represent an estimate of $\rho_A(x)/(\rho_A(x) + \rho_B(x))$ regardless of additional categories included in the training.	199
Figure 11.12	The GloParT's tagging performance for distinguishing $H \rightarrow b\bar{b}$ jets from QCD jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.....	200
Figure 11.13	The GloParT's tagging performance for distinguishing $H \rightarrow c\bar{c}$ jets from QCD jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.....	200

Figure 11.14	The GloParT's tagging performance for distinguishing $H \rightarrow c\bar{c}$ jets from $H \rightarrow b\bar{b}$ jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.	201
Figure 11.15	The mass decorrelation performance of the $X \rightarrow b\bar{b}$ tagging discriminant for the CMS ParticleNet-MD tagger (left) and the GloParT tagger (right). normalized distributions of the QCD jet m_{SD} on the test dataset, passing multiple thresholds of the tagging discriminant, are shown.	201
Figure 11.16	The mass decorrelation performance of the $X \rightarrow c\bar{c}$ tagging discriminant for the CMS ParticleNet-MD tagger (left) and the GloParT tagger (right). normalized distributions of the QCD jet m_{SD} on the test dataset, passing multiple thresholds of the tagging discriminant, are shown.	202
Figure 11.17	A comparison of reconstructed mass resolution in $H \rightarrow b\bar{b}$ jets among three algorithms: the soft-drop, the CMS ParticleNet regression algorithm, and the GloParT model with the regression target being the GEN-level resonance (left) and the visible-particle mass (right).	203
Figure 11.18	A comparison of reconstructed mass resolution in $H \rightarrow c\bar{c}$ jets among three algorithms: the soft-drop, the CMS ParticleNet regression algorithm, and the GloParT model with the regression target being the GEN-level resonance (left) and the visible-particle mass (right).	203
Figure 11.19	A comparison of tagging performance when additional irrelevant categories are added to the classifier model. The performance is compared for the hadronically decayed $H \rightarrow WW^* \rightarrow 4q$ (upper left), hadronically decayed $H \rightarrow WW^*$ with a quark outside of the jet cone (denoted as $3q$) (upper right), and $H \rightarrow WW^* \rightarrow \tau_h \nu qq$ against the QCD jets.....	205
Figure 11.20	A comparison of $H \rightarrow WW^* \rightarrow 4q$ versus QCD jets tagging performance for using the three discriminants defined on different truth $H \rightarrow W^{(*)}W^{(*)}$ labels, i.e. the SM $H \rightarrow WW^*$ class and two BSM classes $H \rightarrow WW$ and $H \rightarrow WW^*$. The generator-level m_H is configured as 50 (upper left), 100 (upper right), and 200 (lower) GeV to generate different kinematics of $H \rightarrow W^{(*)}W^{(*)}$ decays.....	207
Figure 11.21	A comparison of tagging performances on multiple tasks when the model scales from ParT-lite, to ParT-B, then to ParT-L (GloParT).	208
Figure 11.22	A comparison of reconstructed mass resolution on $H \rightarrow b\bar{b}$ (left) and $H \rightarrow c\bar{c}$ (right) jets when the model scales from ParT-lite, to ParT-B, then to ParT-L (GloParT). The regression target is the GEN-level visible-particle mass.....	209

Figure 12.1	Illustration on the fine-tuning experiment of GloParT to perform jet tagging on novel jet categories.	213
Figure 12.2	Metrics and the per-batch learning rate (bottom) monitored during the fine-tuning experiment of GloParT to perform tagging on novel jet categories. The metrics include the per-epoch classification accuracy (top) and the per-epoch loss (middle) for the training (left) and evaluation (right) procedure.....	214
Figure 12.3	The performance on tagging a new class of jet, $LQ \rightarrow b\tau_h$, versus the QCD, $H \rightarrow b\bar{b}$, $H \rightarrow \tau_h\tau_h$, or $t \rightarrow bW \rightarrow \tau_h\nu$ jets as backgrounds. The performance is compared with a training-from-scratch GloParT model with the new classes initially added.	215
Figure 12.4	Illustration on the fine-tuning experiment of GloParT to derive the mass-correlated top-jet tagging model.....	217
Figure 12.5	A visualization of $\exp(\alpha_i + \Delta)$, where α_i is the output value of the auxiliary network for class i , and it is shifted by a constant Δ to ensure the minimum value is zero. An annotation about how to explain the performance of different top tagging classes is shown in the plot.	219
Figure 12.6	A comparison of performance on top versus QCD jet tagging among the CMS ParticleNet top-jet tagging model, the GloParT model without fine-tuning such that the tagging is decorrelated with jet mass, and the GloParT model with fine-tuning (GloParT f.t.). The ROC curve is displayed in both low (left) and high (right) p_T region.....	220
Figure 13.1	(a) Pre-training of the Sophon model as a multiclass classifier. (b) Application of the Sophon model via the construction of discriminants from selected outputs and transfer learning performed using the hidden representations.	227
Figure 13.2	Top: Comparison of $X \rightarrow b\bar{b}$ signal discovery significance as a function of the SM background efficiency across different algorithms. Bottom: Stacked distributions of various SM background processes and the $X \rightarrow b\bar{b}$ signal process, selected using the Sophon $X \rightarrow b\bar{b}$ vs. QCD discriminant at $\epsilon_B = 10^{-4}$	229
Figure 13.3	Top: Comparison of $X \rightarrow bs$ signal discovery significance as a function of the SM background efficiency across different algorithms. Bottom: Stacked distributions of various SM background processes and the $X \rightarrow bs$ signal process, selected using the fine-tuned Sophon $X \rightarrow bs$ vs. QCD discriminant at $\epsilon_B = 10^{-4}$	231
Figure 13.4	Illustration of the application of various Sophon tagging discriminants at different selection efficiencies ($10^{-4} < \epsilon < 10^0$) to unveil the hadronic peak structures of the Z , W bosons, and the t quark.....	232

Figure 13.5	Top: Discovery significance for the $W' \rightarrow W\phi \rightarrow WWW$ signal as a function of the SM background efficiency, utilizing an optimized Sophon tagging discriminant. Middle and bottom: Stacked distributions of various SM background processes and the signal process, selected using the discriminant at $\epsilon_B = 10^{-4}$ and 10^{-5} , respectively.....	235
Figure 13.6	Illustration of the anomaly detection method employed in the weakly supervised approach ^[201]	236
Figure 13.7	Results from Ref. ^[200] showcasing the capability to search for $W' \rightarrow W\phi \rightarrow WWW$ within the m_{jj} spectrum using the weakly supervised anomaly detection technique.	236
Figure 13.8	Comparison of the Sophon model's effectiveness in model-specific searches and its model-agnostic capabilities, alongside the conventional method, for probing $W' \rightarrow W\phi \rightarrow WWW$ within the m_{jj} spectrum.....	239



Part I

Foundations

Chapter 1

Experimental Particle Physics

1.1 The standard model

1.1.1 The path to the SM

The journey of particle physics to the standard model (SM) can be traced back to the 19th century with the discovery of the electron by J. J. Thomson^[1], marking the divisibility of the atom. This was followed by the discovery of the nucleus in 1911^[2] and the discovery of the neutron in 1932^[3]. On the theory's side, quantum mechanics emerged in the 1920s and was found suitable for describing the atomic and subatomic processes. It soon evolved into quantum field theory (QFT), a specific form of which is the quantum electrodynamics (QED) established in the 1940s^[4-7] that describes electromagnetic interactions within a quantum framework and is integrated with special relativity. In the 1950s, the Yang-Mills theory was introduced^[8], incorporating QFT with the non-Abelian gauge group, which was crucial to the later formation and unification of electroweak interactions^[9, 10]. At the same time, advances in accelerator technology boosted the discovery of a large number of subatomic particles, which led to the quark model^[11] and the formulation of the quantum chromodynamics (QCD)^[12, 13]. Following this, the formulation of the Higgs mechanisms in the 1960s^[14-16] explains how particles acquire mass through spontaneous symmetry breaking in gauge theories. These theoretical and experimental milestones paved the way for the SM, a comprehensive theory that unifies all known elementary particles and forces except for gravity.

A general picture of the SM is provided in Fig. 1.1 to showcase the particle constituents and their interactions. At its heart is the Higgs boson, which gives masses to the other particles. W and Z bosons are the carriers of the weak interaction, and they acquire mass via spontaneous symmetry breaking in the Higgs field. The photon, which mediates the electromagnetic force, remains massless, preserving the $U(1)$ gauge symmetry associated with electromagnetism. The gluon is also a massless boson that mediates the strong interaction. It holds quarks together within protons, neutrons, and other hadrons. In addition, fermions are the building blocks of matter and are divided into quarks and leptons. They all constitute three generations. The first generation includes up and down quarks, which make up protons and neutrons, and the electron

along with its neutrino. The second and third generations are heavier counterparts, including the charm and strange quarks, the muon and its neutrino, for the second generation; and the top and bottom quarks, and tau and its neutrino, for the third generation.

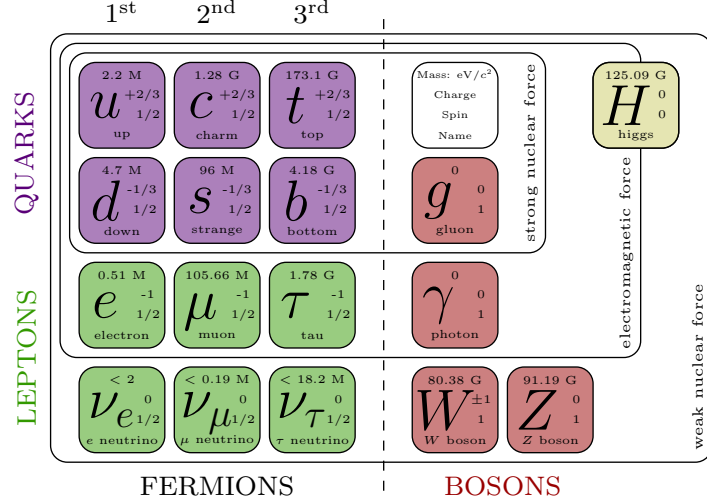


Figure 1.1 A visualization of all elementary particles in the SM and their properties summarized in Ref.^[17].

1.1.2 The SM Lagrangian

This principle of symmetry is the core of the SM framework. The Lagrangian of the SM is built to respect the $SU(3) \times SU(2) \times U(1)$ gauge symmetry. It describes the fermions constituting the world and their interactions mediated by bosons. The $SU(3)$ describes the nature of quantum chromodynamics, and $SU(2) \times U(1)$ reveals the unified electroweak interactions. Given that discussions of the SM framework have been widely covered in various literature over several decades, a concise summary is provided as follows.

The compact form of the SM Lagrangian is written as

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F_a^{\mu\nu} + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi + (y_{ij}\bar{\psi}_i\phi\psi_j + h.c.) + |D_\mu\phi|^2 - V(\phi). \quad (1.1)$$

Here, the first term corresponds to the kinetic term for gauge field, where $F_{\mu\nu}$ is the field strength tensor that covers the electroweak force and the strong force. The second term is the coupling of fermions, such as leptons and quarks, to the gauge field. The third term characterizes the Yukawa coupling, and it describes the interaction of the fermion fields with the Higgs field. The fourth term is the Higgs interaction with the boson field. The fifth term is the Higgs field potential. Each term maintains its form under the gauge transformations. These interactions are depicted in Fig. 1.2, in the format of Feynman diagrams, a tool to graphically present

the quantum fields and their interaction within a Lagrangian.

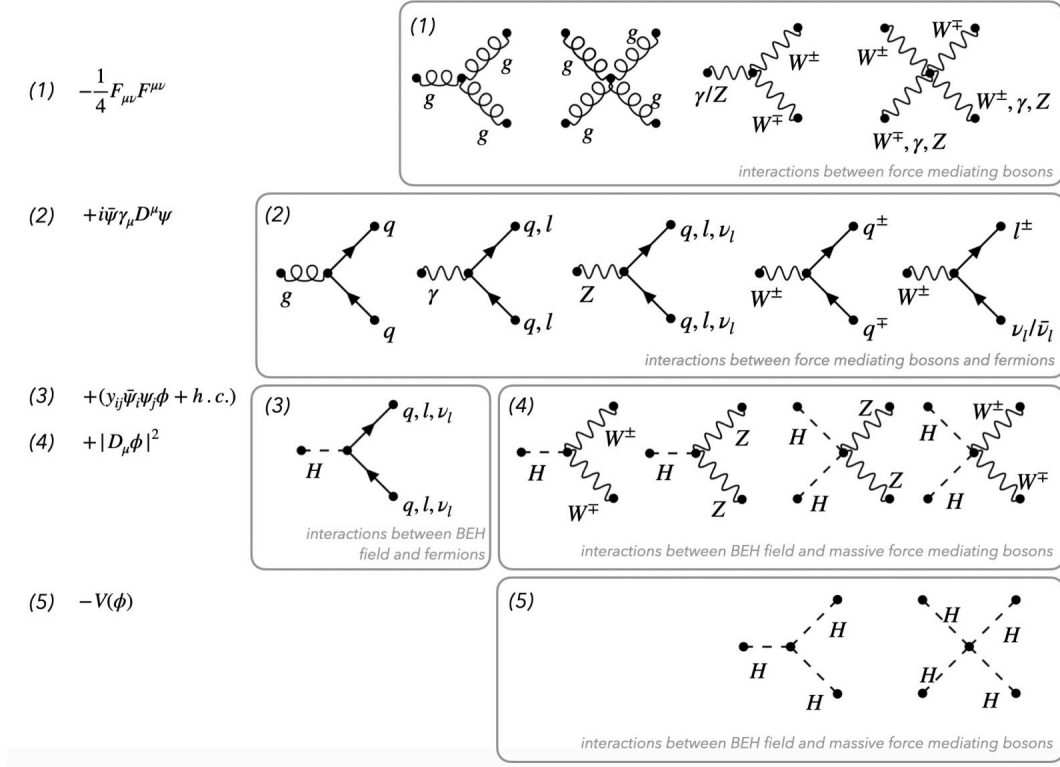


Figure 1.2 The interaction of SM particles representative by the five terms in the SM Lagrangian in the format of Feynman diagrams^[18].

1.1.3 Higgs mechanism and the origin of mass

Allowing particles to acquire mass without breaking the gauge symmetry was a challenge in history, and the Higgs mechanism provides a viable solution. The Higgs mechanism introduces a self-interacting complex scalar field ϕ , which spontaneously breaks the $SU(2) \times U(1)$ symmetry. A concise summary is presented below.

The Higgs field is a complex scalar doublet field,

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \quad (1.2)$$

With the Higgs field, the Higgs potential term in SM Lagrangian is designated in the form of

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \quad (1.3)$$

The minimum potential is not at the zero point when $\mu^2 < 0$ and $\lambda > 0$. Without loss of generality, the physics vacuum state where the potential reaches the minimum can be written

as

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad \text{where } v = \sqrt{-\frac{\mu^2}{\lambda}}. \quad (1.4)$$

The choice of a specific vacuum state indices the spontaneous symmetry breaking. The scalar field can be expanded around the vacuum expectation value as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ v + h + ia \end{pmatrix}, \quad (1.5)$$

where h is the physical Higgs bosons. These Goldstone bosons associated with the fields ϕ_1 , ϕ_2 , and a^0 are absorbed by the electroweak gauge fields, providing the necessary longitudinal components for the $W^\pm$ and Z bosons. This is the procedure through which they acquire masses. The mass equations for the $W^\pm$ and Z bosons are directly proportional to the VEV, with

$$m_W^2 = \frac{g^2 v^2}{4}, \quad m_Z^2 = \frac{(g^2 + g'^2) v^2}{4}, \quad (1.6)$$

where g and g' are the gauge couplings for the SU(2) and U(1) gauge fields. Here, h denotes the single scalar field degree of freedom, and it emerges as the Higgs boson with a mass determined by the Higgs potential parameters, given by $m_H = v\sqrt{2\lambda}$. After the electroweak symmetry breaking, fermions gain mass through Yukawa interactions with the Higgs field. This fermion acquire masses (m_f) as

$$m_f = \frac{1}{\sqrt{2}} v y_f, \quad (1.7)$$

which is proportional to y_f , the corresponding Yukawa coupling strength. Part IV in this dissertation focuses on the measurement of the Yukawa coupling between the Higgs boson and the charm quark, y_c . Eqs. (1.6) and (1.7) highlight the structure of the SM related to the Higgs mechanism and its foundation role in the generation of particle masses for the SM weak bosons and fermions.

1.1.4 Higgs boson phenomenology

In the Large Hadron Collider (LHC) experiments, a thorough understanding of the Higgs boson's phenomenology is essential for finding the most promising final states and implementing optimal event reconstruction methods, which enable precise measurements of its properties. This section highlights the key aspects of the phenomenology of the Higgs physics at the LHC, with a focus on Higgs boson production and decay modes. In a proton-proton collision at the LHC, Higgs bosons are primarily produced through four mechanisms:

Gluon fusion (ggF) The most dominant Higgs production mode at the LHC involves the fusion of two gluons mediated by a virtual quark loop, depicted in Fig. 1.3 (a). This method is known as the gluon fusion process, often denoted by ggF.

Vector boson fusion (VBF) The VBF process produces a Higgs boson through the interaction of two vector bosons, which are emitted by quarks, as shown in Fig. 1.3 (b). The process is characterized by the jets induced from the emitted quarks being widely separated in pseudo-rapidity η and having a large invariant mass.

Higgs boson production associated with a vector boson (VH) The Higgs boson can be produced in association with a vector boson W or Z , denoted by VH . The Feynman diagram of this process is shown in Fig. 1.3 (c). It is also referred to as the Higgsstrahlung process. The emitted vector boson can provide a triggering of the Higgs boson to facilitate the experimental search. This production channel is the main research target in the physics measurement discussed in Part IV.

Higgs boson production associated with a top quark-antiquark pair ($t\bar{t}H$) The $t\bar{t}H$ process involves the Higgs boson production in association with a top quark-antiquark pair, illustrated partially in Fig. 1.3 (d). It is notable for its lower cross section owing to the substantial mass of the resulting particles.

The above processes are the dominant Higgs production mode at the LHC. The cross sections of these processes are compared as a function of centre-of-mass energy in Fig. 1.4 on the left plot. Additional production modes include the H production associated with $b\bar{b}$, which is also shown in Fig. 1.3 (d), or associated with a single top quark as shown in the Fig. 1.3 (e, f).

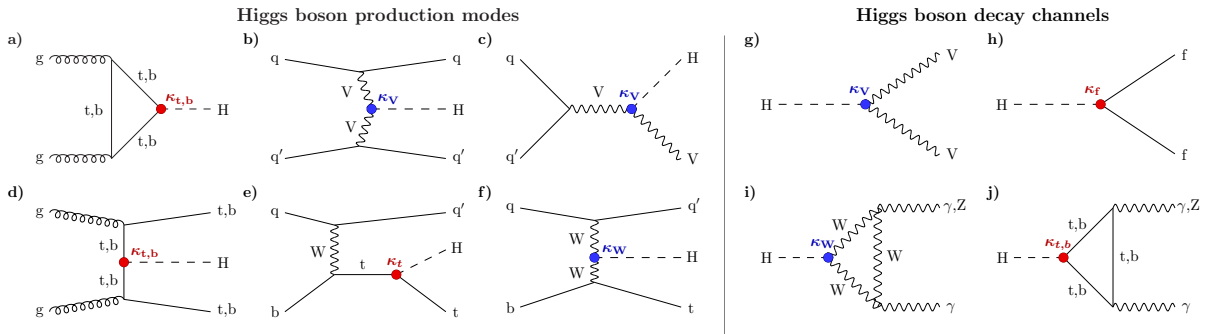


Figure 1.3 The Feynman diagrams of the main Higgs production modes (left) and the Higgs decay modes (right). Image adapted from Ref.^[19].

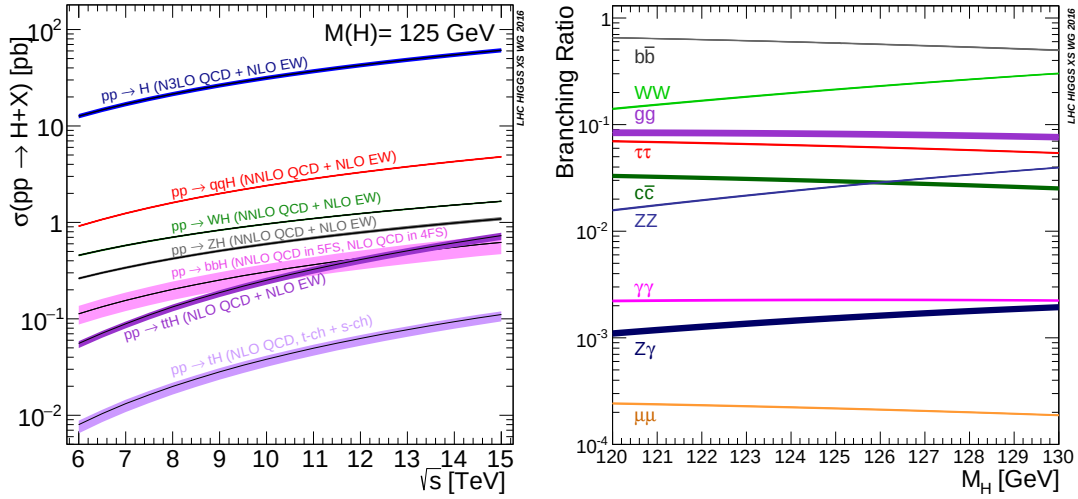


Figure 1.4 The Higgs boson production cross sections in various production modes as a function of the centre-of-mass energy^[20] (left), and the branching functions as a function of the Higgs boson mass^[20] (right).

Regarding the decay of the Higgs boson, it can decay through multiple channels due to its interactions with leptons, quarks, and vector bosons, each with distinct branching fractions shown in Fig. 1.4 on the right plot. In general, the decay of the Higgs boson prefers heavier final-state particles but is subject to phase-space limitations. The decay modes are briefly outlined below.

Decays to vector bosons The Feynman diagram for Higgs boson decay into a pair of vector bosons is shown in Fig. 1.3 (g). The mass constraints of the W and Z bosons forbid Higgs boson's onshell decay into vector bosons. However, the decay branching ratios to the onshell-offshell WW^* and ZZ^* pairs are still considerable. In addition, the decay to a pair of Z bosons plays an important role at the LHC in the Higgs boson due to a very clean experimental signature in the four-lepton final state.

Decays to fermions The Higgs boson decays to fermions through a direct Yukawa coupling, depicted in Fig. 1.3 (h). Given that the Yukawa coupling strength is proportional to the fermion mass, the Higgs boson's decay to first-generation fermions, i.e. $b\bar{b}$ and $\tau^+\tau^-$ pairs, are significant. These decays are predominant but challenging to isolate in hadron collider environments due to the difficulty in constructing their final state signatures. Searching for their decay into second-generation fermions can be more challenging as the branching ratio is small. Nevertheless, they can be the next important milestone in uncovering the Higgs boson property. Part IV of this dissertation discusses the recent measurement of the Higgs boson decaying to $c\bar{c}$ within the CMS experiment.

Loop-induced decay processes The Higgs boson decay into massless particles, including gluons and photons, is possible through a loop-induced process. The Higgs boson to gg decay takes up a considerable proportion but its direct measurement is difficult. Meanwhile, the Higgs boson decaying to diphoton through a vector-boson loop or a fermion loop, as shown in Fig. 1.3 (i, j), provides clear modelling of a peak structure on the diphoton invariant mass and is the golden channel to probe the Higgs boson at the LHC.

1.2 The Large Hadron Collider

The LHC is the world's largest and highest-energy particle accelerator. Located on the border of Switzerland and France, it spans a total length of 27 km and is designed for accelerating heavy ions for collision, primarily focusing on proton-proton collisions. The LHC is designed to reach collision energies of 14 TeV, equivalent to accelerating protons to 99.9999991% of the speed of light.

The acceleration of protons in the LHC is achieved through multiple stages, including initial acceleration in Linac 2, boosting in the Proton Synchrotron Booster (PSB), further acceleration in the Proton Synchrotron (PS) and Super Proton Synchrotron (SPS), and finally injected into LHC's vacuum tubes. A schematic view of the LHC with various accelerators at CERN is shown in Fig. 1.5.

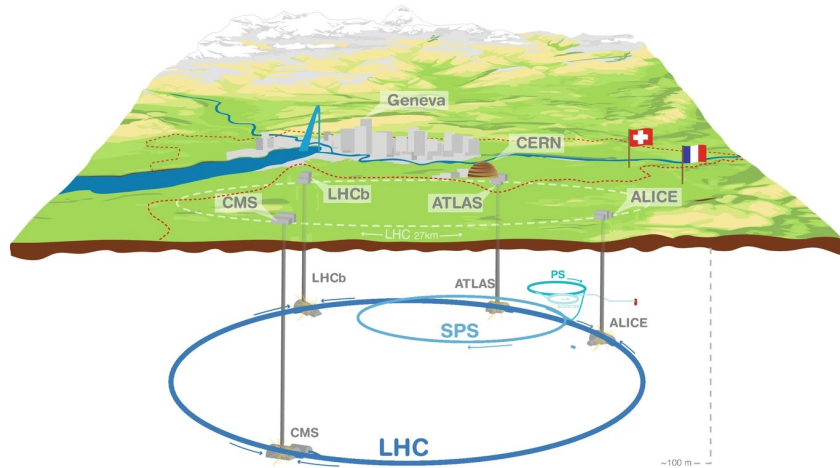


Figure 1.5 A schematic view of the LHC^[21].

To increase the probability of proton collisions, the protons in the accelerator are squeezed into a very small space before colliding. Therefore, the collisions at the collision points involve proton bunches. The collision of proton bunches (bunch crossing) is every 25 nanoseconds or a collision frequency of 40 MHz. Proton collisions occur at four particle detectors on the LHC:

ATLAS (A Toroidal LHC Apparatus), CMS (Compact Muon Solenoid), ALICE (A Large Ion Collider Experiment), and LHCb (Large Hadron Collider beauty). ALICE is a heavy-ion detector primarily focused on studying quark-gluon plasma at extreme energy densities. LHCb explores the differences between matter and antimatter through b -quark physics. ATLAS and CMS are two general-purpose detectors with the same research scope and physical objectives.

The LHC programme's physics goals range widely, including the detailed study of SM, properties of the Higgs boson, and the extensive search of the beyond-the-SM (BSM) signatures. The latter includes searching for dark matter, exploration of supersymmetry, searching for extra dimensions, etc. It allows physicists to test the predictions of the particle physics theories. The independent discovery of the Higgs boson in 2012 by the ATLAS and CMS experiments marks a major milestone in the success of the LHC programmes.

The LHC programmes have evolved through various operational phases. The Run 2 data-taking period (2015–2018) was operated at the centre-of-mass energy of 13 TeV and collected data corresponding to an integrated luminosity of around 140 fb^{-1} . In the year 2024, the Run 3 period (2022–2025) is ongoing, with the LHC's increased luminosity and energy. The peak luminosity can reach around $2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, and the energy at a centre-of-mass energy of 13.6 TeV. The High-Luminosity LHC (HL-LHC) project is a substantial upgrade planned to start in 2029 following the long shutdown 3 (LS3). The HL-LHC aims to significantly increase the luminosity, potentially by a factor of 10, extending the LHC's operational life and discovery potential out to 2041. Figure 1.6 shows the timeline of the LHC and HL-LHC programmes.

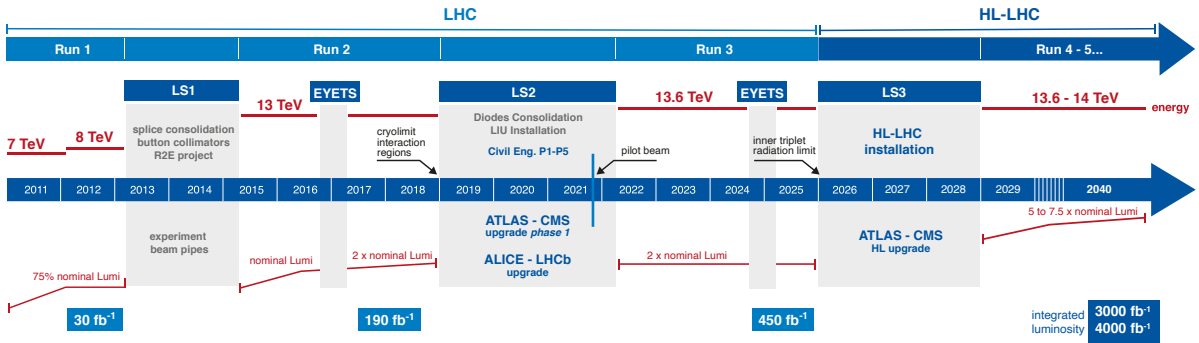


Figure 1.6 Timeline for the LHC and the High-Luminosity LHC. Source: *CERN website*.

1.3 The CMS detector

Located in Cessy, France, the CMS detector is one of two general-purpose colliders whose physics targets include precision measurement of the SM and the search for new physics.

The CMS detector is 21 m long, 15 m in diameter and weighs about 14000 tons. It constitutes the following subdetectors: the inner tracking system including the silicon pixel detector and silicon stripe detector, an electromagnetic calorimeter, a hadron calorimeter, and a muon system. A cutoff view of the CMS detector with its subcomponents is shown in Fig. 1.7. Each subdetector consists of a small and individual hardware unit covering most of the area around the collision point. All subdetectors are inside a superconducting solenoid at 3.8 T, except for the muon system and a minor part of the hadron calorimeter. When charged particles pass through the trackers, the large magnetic field induces a curved trajectory, which enables a precise measurement of the particle momenta.

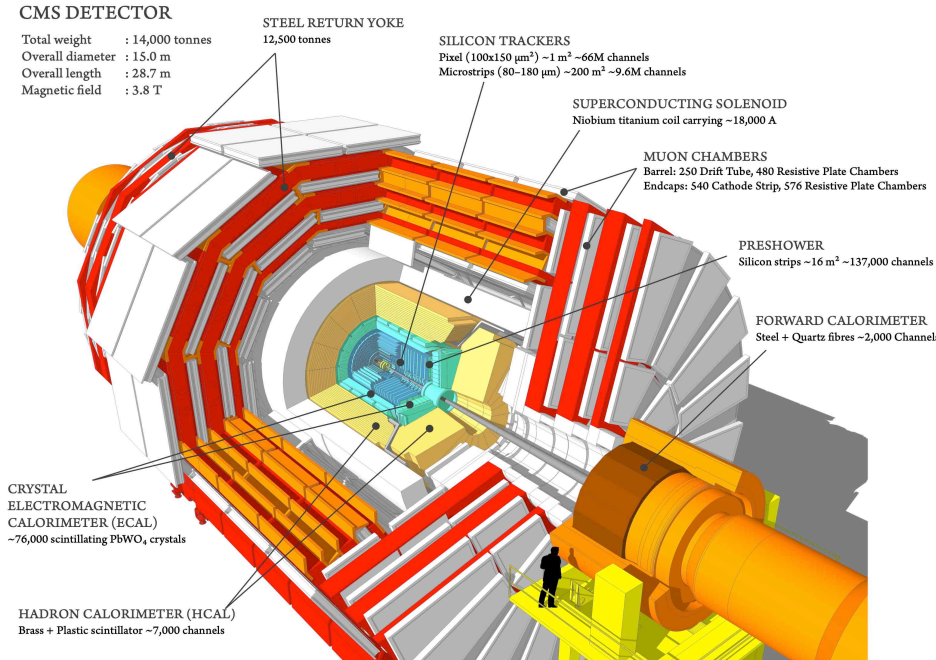


Figure 1.7 A cutaway diagram of the CMS detector^[22].

CMS coordinate system The CMS experiment uses a right-handed coordinate system. In this system, the origin lies in the nominal interaction point. The z axis runs counterclockwise along the beam direction and the x - y plane is the transverse plane. The x axis points to the centre of the LHC ring, and the y axis points upward. The ϕ and r characterize the azimuthal angle and the radial coordinate on the x - y plane. The coordinate system is depicted in Fig. 1.8. Besides, in the three-dimensional coordinate, the polar angle is denoted θ , and the pseudorapidity η is defined as $\eta = -\log(\tan(\theta/2))$.

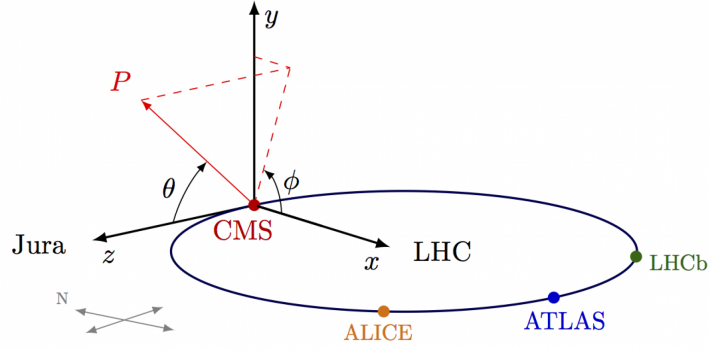


Figure 1.8 Illustration of CMS coordinate system.

1.3.1 Inner tracker

The inner tracking system measures the trajectories of charged particles. The tracker is divided into several subregions, as illustrated in Fig. 1.9. Closest to the collision point lies the silicon pixel detectors, each with a pixel size of $100 \times 150 \mu\text{m}^2$. The pixel detector has a total surface area of approximately 1 m^2 with around 6.6×10^7 pixels. The silicon strip tracker is placed outside of the pixel detector, composed of 80-to-180 μm -wide strips. It is subdivided into the inner barrel part (TIB), the inner disks (TID), the outer barrel (TOB) and outer end-caps (TEC), with the configuration outlined in Fig. 1.9. The silicon strip tracker spans 200m^2 with about 9.6×10^6 strips.

The inner tracker covers a spatial extent of $|\eta| < 2.4$. This setup enables the tracking of particles through electrical signals generated by their interaction with the detector, with material thickness varying from about $0.4X_0$ at $\eta = 0$, to around $1X_0$ at $|\eta| = 1.6$, and back to $0.6X_0$ at $|\eta| = 2.5$, where X_0 denotes the radiation length.

The pixel detector encountered an upgrade in the LHC's year-end technical stop in 2016–2017, during the Run 2 data-taking period. The new detector is named the Phase I pixel detector, and it adds a new pixel layer in the barrel and endcap section, increasing the coverage of pseudorapidity from $|\eta| = 2.5$ to $|\eta| = 3$.

1.3.2 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is built with lead tungstate (PbWO_4) crystals, which accurately measure the energy of electrons and photons. It utilizes a dense yet optically transparent material to stop high-energy electrons and photons. The material is primarily metallic, heavier than stainless steel but highly transparent. It emits scintillation light proportional to the energy deposited by electrons and photons, making it an accurate, fast, and

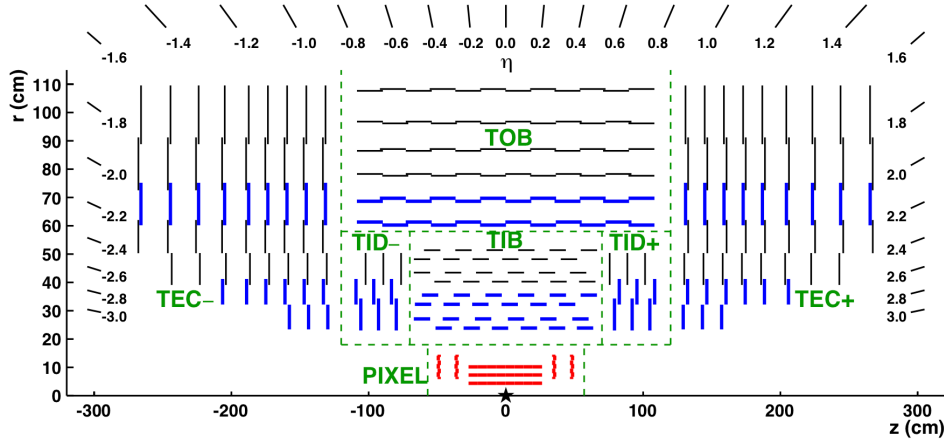


Figure 1.9 An R - z cross section of a half of the CMS detector, showing the layout of the inner tracker system^[23].

compact detector. This energy is then collected via photodiodes, which estimate the energy of an incident particle.

The ECAL comprises a barrel and two endcaps positioned between the tracker and the hadron calorimeter. The barrel covers $|\eta| < 1.479$ and the two endcap disks $1.479 < |\eta| < 3.0$. In ECAL barrel section contains 61200 crystals, and the traverse size of the crystal measures $22 \text{ mm} \times 22 \text{ mm}$ at the front face, corresponding to 0.0174×0.0174 on the (η, ϕ) space, with a length of 230 mm. The ECAL endcaps are made up of nearly 15000 crystals. The relative energy resolution σ/E shows

$$\frac{\sigma}{E} = \frac{2.8\%}{\sqrt{E/\text{GeV}}} \oplus \frac{12\%}{E/\text{GeV}} \oplus 0.3\%. \quad (1.8)$$

To achieve higher spatial resolution, ECAL includes a preshower detector with superior granularity located in front of the endcaps, made of lead block and silicon sensor strips. It allows for distinguishing between single high-energy photons and two closely spaced low-energy photons and improves the separation power between photons and neutral pions.

1.3.3 Hadron calorimeter

The hadron calorimeter (HCAL) is a hermetic sampling calorimeter built around the ECAL and inside the solenoid, used to measure the energy of outgoing hadrons. It is built by alternating brass absorber and plastic scintillator layers, the later emit light when a particle passes through, collected by optic fibers into photodetectors for amplification. The light across is summed over multiple layers and used to estimate the particle's energy. The HCAL has a bar-

rel covering $|\eta| < 1.3$ and two endcaps covering $1.3 < |\eta| < 3.0$. In addition, the HCAL Outer (HO) is installed outside the solenoid and serves as an additional absorber covering $|\eta| < 1.3$. The HCAL segmentation in the (η, ϕ) space is 0.087×0.087 for $|\eta| < 1.6$ and 0.17×0.17 for higher pseudorapidities, which is much more coarse compared to the ECAL. The energy resolution σ/E for the combined ECAL and HCAL measured in a pion test beam is

$$\frac{\sigma}{E} = \frac{110\%}{E/\text{GeV}} \oplus 9\%. \quad (1.9)$$

1.3.4 Muon chambers

The CMS experiment has the best muon identification system at the LHC. Muons can penetrate the calorimeters, so they are detected at the very edges of the CMS using a dedicated muon system. The system consists of a barrel and two endcaps. Its detection planes are interleaved with the steel yoke that returns the solenoid's magnetic flux. Three types of gaseous particle detectors are used in the muon system, outlined in Fig. 1.10. The drift tube (DT) chambers are used for precise trajectory measurements in the central barrel region ($|\eta| < 1.2$) of the detector. In contrast, the cathode strip chambers (CSC) are used in the endcaps ($0.9 < |\eta| < 2.4$), areas with uneven magnetic fields and high radiation levels. The DT and CSC subsystems can not only provide measurements on muons. Still, they can also be used to trigger the p_T of muons with excellent efficiency and high background rejection, independent of the rest of the detector. The resistive plate chambers (RPC) are installed in both the barrel and endcaps covering $|\eta| < 1.6$, providing rapid response for the CMS trigger system.

1.3.5 Trigger and data acquisition

The rate of collisions (40 MHz) at the LHC will result in a data output of 15 PB per year, which is far beyond the capability of mass storage systems. The CMS trigger and data-acquisition system scrutinizes each event to determine which ones to store for analysis.

The CMS trigger system is split into two levels: the Level-1 trigger (L1T) and the high-level trigger (HLT). The L1T is built with custom hardware processors to collaborate with the calorimeters and muon detectors to filter event rates down to less than 100 kHz within a $4 \mu\text{s}$ latency. It selects events based on predefined energy thresholds or momentum for objects like electrons, photons, muons, and jets, using logic encoded in application-specific integrated circuits (ASICs) and programmable memory lookup tables (LUT) or field-programmable gate arrays (FPGA) for implementations. The HLT further reduces the L1-accepted rate of 100 kHz to 100 Hz for the final output, processing full events using large computing farms for tasks

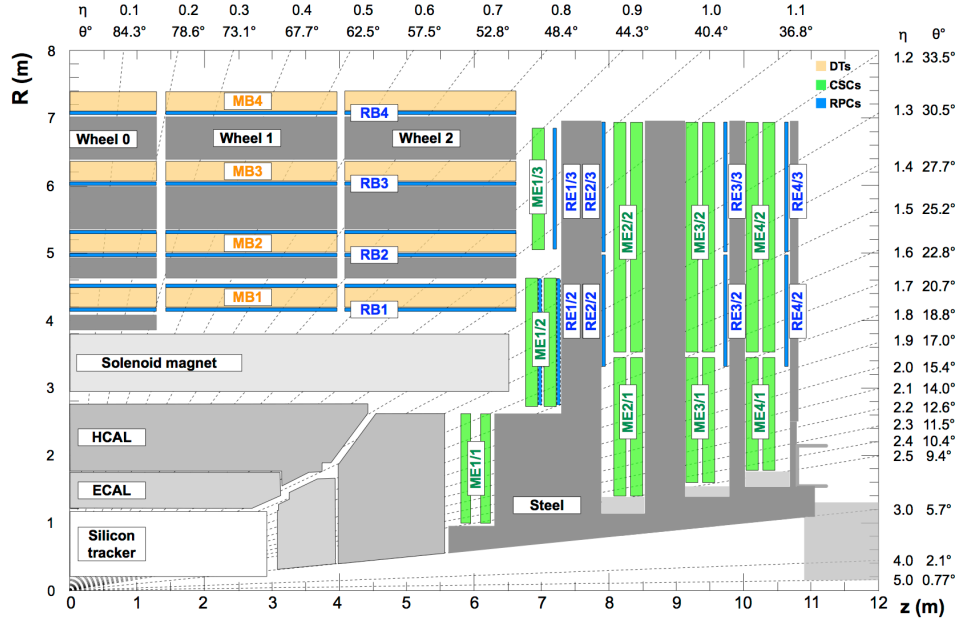


Figure 1.10 An R - z cross section of a quadrant of the CMS detector, showing the layout of the muon detector^[24].

exploiting full detector information and making the software-level decisions on event storage. It uses multiple dedicated streams or trigger paths for event selection at the analysis level.

1.4 Monte Carlo simulation in CMS

It is essential to compare collected data with SM predictions in collider experiments. The CMS experiment employs extensive Monte Carlo (MC) simulations to model phenomena expected in the detector. MC simulations are particularly valuable when analytical solutions are infeasible due to the system's complexity. They numerically estimate observables by sampling pseudorandom numbers according to the phase space and using the acceptance-rejection method to approximate integrals.

In the CMS experiment's typical workflow, MC simulations include multiple stages. First, the simulation of a collision event produced by a proton-proton collision is delivered by event generators, as detailed in Section 1.4.1. The simulation of the detector effect and the mixing of pileup is introduced in Section 1.4.2.

1.4.1 Event generators in CMS

Event generators are MC simulation tools used for exclusive collision event generation. This process involves both the hard-scattering process and the parton showering.

The inclusive cross section of a hard-scattering process can be analytically calculable by QFT at the fixed order in α_s . However, for exclusive event simulations in collider experiments, MC simulations are employed, with physics models and the rules of QFT at their core. The frequently used event generators at this stage within the CMS experiment mainly include MADGRAPH5_aMC@NLO^[25] (MG) and POWHEG^[26, 27]. MG is a versatile event generator capable of simulating hard-scattering processes at the leading order (LO) and next-to-leading order (NLO) accuracy. It has flexible support for various physics models. POWHEG, on the other hand, focuses on generating events with NLO accuracy in QCD while preserving the correct behaviour of parton showers.

The parton showering mainly includes fragmentation, a process to add new emissions of quarks and gluons when the energy scale goes lower, and the hadronization process that constructs colourless hadrons from the partons. They handle the non-perturbative QCD effect at the low energy scale and are archived by dedicated phenomenological models. The PYTHIA generator^[28] is most frequently used within the CMS experiment, and HERWIG^[29] is often used alternatively as a validation. The event generators for parton showering can also simulate the complex case for initial- or final-state radiation of the hard-scattered quarks or gluons, and the particles that come from the break-up of the colliding protons. These processes are often accompanied by the hard-scattering process and are called underlying events. In addition, the PYTHIA and HERWIG generators include built-in matrix element methods that can compute simpler processes at LO accuracy and subsequently evolve them through parton showering. In the CMS experiment, the QCD multijet process, which has the highest statistical demand in simulations, is primarily generated by PYTHIA, with HERWIG serving as an alternative option.

In the CMS experiment, the physics events generated using the aforementioned methods, represented by their respective event counts, are summarized in Fig. 1.11. From a technical perspective, the wide variety and combinations of event generators are integrated into the CMS software (CMSSW) framework through various modules. These modules leverage various forms of multithread computing utilities, which represent a key computational advantage at the core of CMSSW^[30]. The implementation of multithreaded event generation support was first introduced during the Run-2 ultra-legacy data reprocessing campaign, where the author contributed to its development.

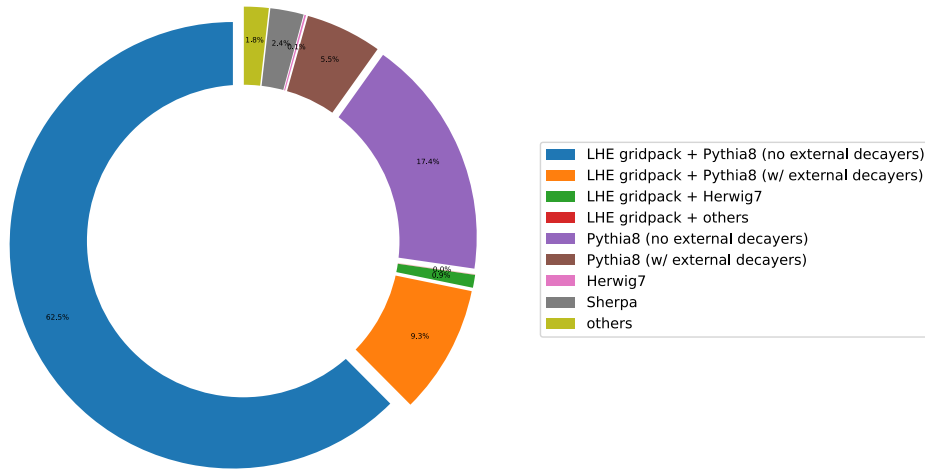


Figure 1.11 A pie chart illustrating the distribution of usage across various event generators and their combinations, represented by the number of events generated.

1.4.2 Detector simulation

The hadronized particles induced by the hard-scattering and accompanying underlying events are propagated to the location of multiple CMS subdetectors. This detector effect is delivered by an MC simulation toolkit, `GEANT4`, which carries the detector layout of all CMS subdetectors and simulates the interaction of these particles with the detector material.

The detector simulation in CMS integrates the effect of pileup. Pileup describes the additional proton-proton collisions when two beams of protons are made to collide. It can be induced from the same bunch crossing (in-time pileup) or in the adjacent bunch crossing (out-of-time pileup) that can overlap with the current readout chain. To simulate pileup, the inelastic events are generated first with `PYTHIA`, proceeded to `GEANT4` for detector simulation, and then superimposed on the simulated hard-scattering event.

1.5 Event reconstruction in CMS

Event reconstruction in the CMS experiment is essential for transforming raw detector data into meaningful physics objects, such as electrons, muons, photons, and jets, enabling effective analysis of particle collisions. At the heart of CMS event reconstruction lies the particle-flow (PF) algorithm^[31], which combines information from all subdetectors to identify and reconstruct individual particles within each event. This leads to the creation of PF candidates, providing a global description of the event (detailed in Section 1.5.1). The reconstruction processes for electrons, muons, secondary vertices (SVs), jets, and missing transverse

momentum are briefly described in Sections 1.5.2–1.5.6.

To support the development of jet tagging tools in both Part IV and Part V, generator-level jets (GEN-jets) are produced directly from stable hadronized particles. These GEN-jets serve as truth-level references for jets and are detailed in Section 1.5.7.

1.5.1 Particle-flow candidate

PF candidates are unified objects representing a reconstructed particle. They play a central role throughout this dissertation, serving as the fundamental building blocks for jet reconstruction in the CMS experiment and providing the primary input features for modern deep learning-based jet processing.

The PF candidates are constructed using the PF algorithm, which integrates information from all subdetectors. For illustrative purposes, Fig. 1.12 (adapted from Ref.^[31]) depicts a toy jet composed of charged and neutral hadrons, as well as photons. The charged hadrons (π^+ and π^- in the figure) are identified through geometrical associations between the tracks and calorimeter clusters (as well as the absence of a signal in the muon detectors). Photons and neutral hadrons are linked to ECAL and HCAL clusters without associated tracks. In addition, electrons are identified by the presence of a track and an ECAL cluster, while muons are associated with tracks in both the inner tracker and the muon detectors (not depicted in the figure).

The process begins with the reconstruction of PF elements, including track reconstruction in the inner tracker (detailed in Section 1.5.1.1), with specialized treatments for electron and muon tracks, and calorimeter clusters in the electromagnetic and hadronic calorimeters (Section 1.5.1.2). The PF elements from different subdetectors are linked together in a predefined order to produce the PF candidates, as described in Section 1.5.1.3.

1.5.1.1 Track reconstruction

Charged-particle track reconstruction is crucial for understanding particle trajectories, their origins, and properties like momentum and direction. The method utilizes a Kalman Filtering (KF) based combinatorial track finder that operates in three main stages: initial seed generation from a few compatible hits, trajectory building to collect hits across the tracker layers, and final fitting to determine the particle properties. Reconstructed tracks must pass several criteria, including a minimum of eight hits for a track to be considered, originating from a specific area around the beam axis, and having a certain minimum p_T . Specifically, the combinatorial track finder uses an iterative procedure. It was developed to enhance track-

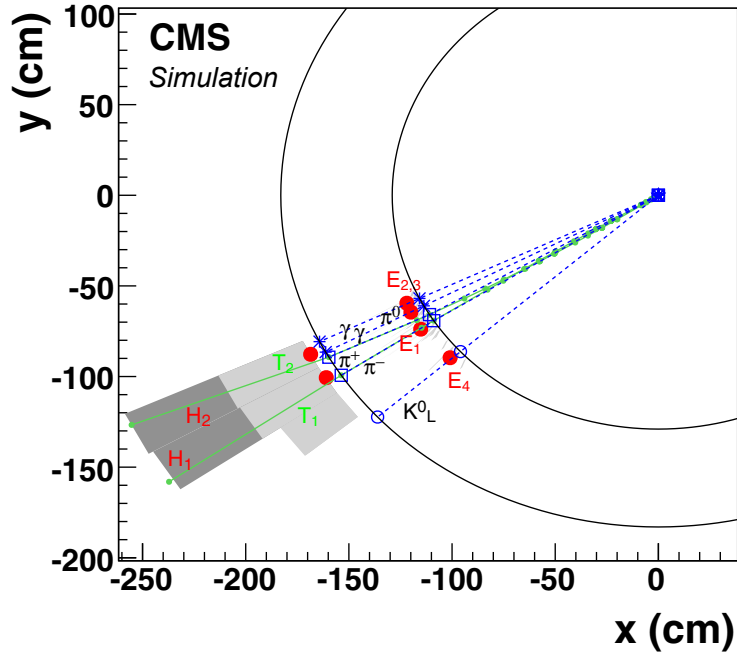


Figure 1.12 An event display of a toy jet initiated by two charged particles π^+ and π^- , one neutral hadron K_L^0 , and two photons from the π^0 decay, shown in the (x, y) coordinate^[31]. The symbols $T_{1,2}$, $E_{1,2,3,4}$ and $H_{1,2}$ represent tracks and energy deposit in ECAL and HCAL measured by the CMS subdetectors corresponding to these particles.

ing efficiency without significantly increasing the misreconstruction rate. It employs multiple rounds of tracking with gradually relaxed criteria. The later iterations aim to resolve challenges in finding very displaced tracks and in distinguishing tracks within the dense high- p_T jets.

Electron tracking benefits from both the ECAL-based approach, relying on energetic ECAL clusters, and a tracker-based method, which uses iterative tracking for higher efficiency. The tracker-based approach improves electron detection, particularly in nonradiated electrons and radiated ones producing lower p_T tracks. In this approach, the probability of radiation within the tracker material is used to distinguish electrons from hadrons. Selected tracks are refitted with a Gaussian-sum filter (GSF) and required to pass a BDT selection.

Muon tracking uses the muon spectrometer and the inner tracker for high-efficiency identification over a wide momentum range. Different muon types, i.e. standalone, global, and tracker muons, are identified based on different matchings between the inner track and muon spectrometer track segment. The high efficiency of muon tracking is attributed to the precise measurement capabilities of the inner tracker and the specific use of the muon spectrometer to improve momentum resolution.

1.5.1.2 Calorimeter clusters reconstruction

Energy deposits in calorimeters form the clusters, which is crucial to analyses as they measure the energy and direction of stable neutral particles such as photons and neutral hadrons. The clustering algorithm distinguishes these neutral particles from charged particle energy deposits. It helps better reconstruct electrons and any bremsstrahlung photons that accompany them and assists in measuring charged hadrons' energy.

In PF element reconstruction, a dedicated clustering algorithm is developed and performed separately in each subdetector, including the ECAL barrel and endcaps, the HCAL barrel and endcaps, and the two preshower layers. The clustering algorithm identifies cluster seeds in cells with a minimum energy threshold and other requirements. Topological clusters then grow from these seeds by including neighbouring cells. The clusters are reconstructed from these topological clusters with an expectation-maximization algorithm based on a Gaussian-mixture model. This model assumes energy deposits in cells follow Gaussian energy distribution centred around the seeds. In an iterative procedure, it adjusts the Gaussian distribution parameters to fit the observed energy deposits.

Calibrating the reconstructed calorimeter clusters is crucial to acquire accurate reconstruction of photons and neutral hadrons. ECAL calibration includes the absolute calibration of ECAL response to electrons and photons, cell-to-cell calibration, and refinement with the collision data. Energy deposit in HCAL is adjusted to account for the nonlinear response to hadrons, separating the case where hadrons deposit energy only in the HCAL and in both ECAL and HCAL.

1.5.1.3 Particle-flow element linkage

PF elements are linked through a set of specific conditions. Tracks from the central tracker are extrapolated to the calorimeters and linked to clusters based on proximity to the cluster area, which is adjusted by considering uncertainties due to gaps between cells, or effects on multiple scattering. For electron bremsstrahlung photons, tangents to the GSF tracks are linked to ECAL clusters if they lie within the cluster boundaries. Conversion finder algorithms link tracks resulting from photon conversions. Calorimeter cluster-to-cluster links are established between HCAL and ECAL clusters and between ECAL and preshower clusters, preferring the links with the smallest distances. Charged tracks can be linked through common secondary vertices, indicating nuclear interactions. The tracks are also linked to the muon detector information for reconstructing muons.

PF blocks result from these linkages, including PF elements related directly or indirectly. The reconstruction sequence begins with muon identification, followed by electron reconstruction, which includes recovering bremsstrahlung photons. Isolated energetic photons are identified next. Then, tracks with high p_T uncertainty are masked, and the remaining elements undergo cross-identification for charged and neutral hadrons and photons. The reconstructed PF blocks form the PF candidates as a global event description. As a final step, these candidates are postprocessed to alleviate misreconstruction by correcting the large artificial missing transverse momentum.

1.5.2 Muons

The muon candidates are selected on top of the global and tracker muons with additional criteria. To determine the isolated global muons, muons with a large energy deposit in their vicinity are vetoed to reject more hadrons misidentified as muons. The selection of muons inside a jet, which is crucial for identifying heavy flavours of jets, also requires more stringent criteria. For muon momentum, the inner track's p_T is used if it is below 200 GeV; above this threshold, the choice is based on the lowest χ^2 probability among various track fits.

1.5.3 Electrons

An electron candidate originates from a GSF track if its associated ECAL cluster is not connected to three or more tracks. ECAL clusters in the PF block, linked to the supercluster or a GSF track tangent, are assigned to the candidate. Tracks linked to these ECAL clusters are considered part of the candidate if their momentum and the ECAL cluster's energy support the electron hypothesis. ECAL clusters and tracks from identified photon conversions connected to the GSF track tangents are also included. The summed energy of these ECAL clusters is adjusted for any missed energy during association. The electron candidate further undergoes requirements on isolation and its energy deposit. The electron's final energy is obtained by combining the corrected ECAL energy with the momentum of the GSF track, and the electron's direction is chosen to be that of the GSF track. The electron is further selected with a BDT using features associated with the GSF track and ECAL cluster.

1.5.4 Secondary vertices

The SVs can indicate the commonality of displaced tracks. The presence of SVs provides a good handle to identify a jet initiated from heavy-flavour hadrons. SVs are reconstructed by the inclusive vertex finding (IVF) algorithm. The algorithm uses all reconstructed tracks

in an event with loose criteria as input. Starting with identifying a collection of seed tracks, it proceeds with the track clustering step, SV fitting and cleaning with the adaptive vertex fitter^[32], track arbitration, and the SV refitting and cleaning procedure.

1.5.5 Jets

Jets are clustered from PF candidates using the anti- k_T algorithm^[33] with a distance parameter of R . This algorithm runs in an iterative clustering procedure to either merge two input constituents into one or merge one constituent with the beam, which means that the constituent is declared as a jet. In each step, the anti- k_T distance is computed for each pair of particles (i, j),

$$d_{ij} = \min(p_{T,i}^{-2}, p_{T,j}^{-2}) \frac{\Delta R_{ij}^2}{R^2}, \quad (1.10)$$

and between a particle and the beam,

$$d_{iB} = p_{T,i}^{-2}, \quad (1.11)$$

and proceed to the merging step. The procedure stops until there is no particle left. When applying the anti- k_T algorithm, the small- R jet collection is obtained from parameter $R = 0.4$ and is referred to as AK4 jets. In the CMS experiment, the large- R jet collection typically results from a clustering parameter $R = 0.8$ (AK8 jets) but can be customized by using an even larger radius $R = 1.5$ (AK15 jets) for certain analyses, e.g. in the Higgs-charm coupling measurement discussed in Part IV.

The particles originating from pileup are mitigated through the charged hadron subtraction^[34] algorithm for AK4 jets and the pileup per-particle identification (PUPPI)^[35, 36] algorithm for AK8 or AK15 jets. The CHS algorithm first identifies charged PF candidates originating from pileup and removes them before the jet clustering. After AK4 jets are clustered, an event-by-event correction based on the jet area is applied to the jet four-momentum to remove the remaining energy from pileup, mostly from the neutral particles. In the PUPPI algorithm, charged particles identified to be originating from pileup are discarded, and a weight between zero and one is assigned to each neutral particle as the probability of the particle originating from the primary vertex. The four-momenta is rescaled per particle to form the new PF candidate collection, which is used as input to cluster AK8 or AK15 jets.

The soft drop algorithm^[37-39] is applied to each large- R jet to acquire the soft-drop mass in a procedure to remove the soft radiation patterns within the jet progressively. This algorithm first decluster a jet into its constituents and recluster them using the Cambridge/Aachen clus-

tering algorithm. This algorithm is similar to the anti- k_T one, except that the power index in Eqs. (1.10, 1.11) on p_T is changed from -2 to 0. This results in a clustering tree that represents the merging history of the constituents. An iterative procedure runs on the reversed merging history to split one constituent into two; for each splitting, the criteria

$$\frac{\min(p_{T_1}, p_{T_2})}{p_{T_1} + p_{T_2}} > z_{\text{cut}} \left(\frac{\Delta R}{R_0} \right)^\beta \quad (1.12)$$

is examined, with $\beta = 0$ and $z_{\text{cut}} = 0.1$, and $R = 0.8$ or 1.5 for AK8 and AK15 jets. If the criteria are failed, it indicates that one branch is comparable soft and is discarded, and the procedure continues with the remaining branch. The procedure stops until the requirement is satisfied; the remaining two branches are considered two subjects of the given net, and the soft-drop mass is designated as the invariant mass of two subjects.

1.5.6 Missing transverse momentum

The missing transverse momentum ($\vec{p}_T^{\text{miss}}$) of an event is reconstructed as the negative vector sum of all PF candidates

$$\vec{p}_T^{\text{miss}} (\text{raw}) = - \sum_i^{N_{\text{PF cands.}}} \vec{p}_{T,i}, \quad (1.13)$$

to describe the imbalance of transverse momentum due to the presence of particles that does not interact with detectors, such as neutrinos. For analysis use, the energy corrections applied to jets can be propagated to $\vec{p}_T^{\text{miss}}$:

$$\vec{p}_T^{\text{miss}} = - \sum_i^{N_{\text{PF cands.}}} \vec{p}_{T,i} - \sum_j^{N_{\text{jets}}} (\vec{p}_{T,j}^{\text{corr}} - \vec{p}_{T,j}). \quad (1.14)$$

1.5.7 Generator-level objects

The GEN-jets are clustered from all stable final particles generated by PYTHIA, with the anti- k_T algorithm and $R = 0.8$ or 1.5 . Two cases are considered, one in which the neutrinos are excluded from the particle list (denoted as “no- ν ”), and another without the exclusion (denoted as “with- ν ”). The former is the default configuration, and the clustered GEN-jets provide a reference for the reconstruction accuracy. For both cases, a soft drop algorithm with the same settings as described above is adopted to mimic our handling of the reconstructed-level jets.

The “no- ν ” and “with- ν ” GEN-jet collections after soft-drop are matched to the large- R jet collection using the standard CMS convention. To perform the matching, each jet is matched with the GEN-jet with the smallest ΔR with the jet, also requiring $\Delta R < R_0$. This

matching process is carried out for each jet, following a sequence sorted by decreasing jet p_T . Some jets may have no matched GEN-jet after this procedure. The “no- ν ” and “with- ν ” GEN-jet soft-drop mass for each jet is determined as the mass of the matched GEN-jet for the two collections.

Chapter 2

Deep Learning

2.1 The path to modern deep learning

Deep learning is a specific type of machine learning that utilizes the ability of neural networks (see Fig. 2.1), usually with multiple layers, to extract progressively higher-level features from data. The journey of building artificial neural networks was initiated in the 1940s when the first effort of making brain-like neurons was made^[40]. It provides a mathematical representation of brain cells and initiates interest in creating networks that could simulate brain functions. The first perceptron was devised later in the 1950s, which marks an early learning algorithm that adjusts its learned parameters with data^[41]. The first hierarchical, multilayer perceptrons (MLP) were devised in 1965 and are generally considered the work pioneering the first deep learning architectures. Despite this early success, the community soon encountered criticism of the effectiveness of neural networks and witnessed the first “artificial intelligence (AI) winter”. There were no further leaps until the 1980s, when the backpropagation algorithm was developed^[42], allowing for more complex networks to learn. Another breakthrough during this period was the introduction of the renowned convolutional neural networks designed for image recognition tasks utilizing the backpropagation technique^[43]. The long short-term memory (LSTM) network, designed in the late 1990s, is an efficient structure to handle sequence data^[44]. In 2006, the introduction of deep belief networks^[45] demonstrated the potential of deep architectures, advancing the arrival of the modern era of deep learning.

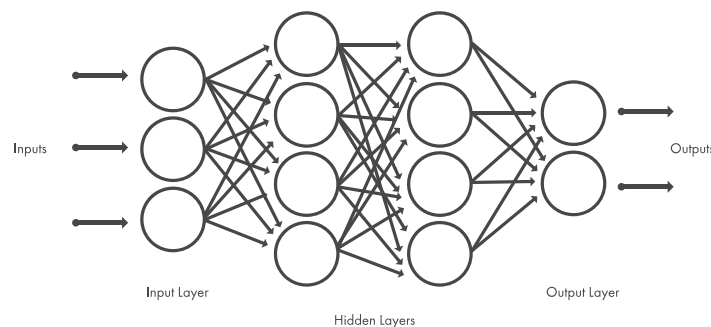


Figure 2.1 A view of a typical neural network. Source: *Mathworks*.

In 2012, the introduction of AlexNet^[46] marked a significant milestone in deep learning, notably outperforming existing algorithms on the ImageNet challenge^[47]. This breakthrough led to a surge in deep learning research, where DNNs began to surpass traditional methods consistently. Advances in hardware technology, such as graphic processing units (GPUs), have also enhanced networks, which have grown larger and deeper. This allows substantial methodological advancements across various domains, with a brief overview presented in the next section. Today, we are on the verge of an era dramatically transformed by modern deep learning, driven by its unprecedented capabilities to analyze, understand, and interact with data in ways that mimic human intelligence.

2.2 Advances in deep learning

The rapid development of deep learning over the past decade has reshaped the academic landscape and its applications to broader fields. This section provides a selective and concise summary directly relevant to the topic of this dissertation. It highlights advancements in training techniques and methodologies, algorithm developments, computational progress, and the rapid expansion of application domains.

The advancements in training methodologies have been the early driving force behind the surge in deep learning applications. Historically, there have been many practical challenges to train DNNs efficiently and to achieve high performance with DNNs. The challenges involve the overfitting issue, vanishing or exploding gradient problems, the imbalance in datasets, etc. Several proposed solutions have become standard within the community for present DNN developments. The innovative methodology in DNN training also includes transfer learning, which leverages pre-trained models to achieve high performance in data-constrained environments (the one we will discuss in Part V), the model compression techniques that aim to reduce model size without sacrificing accuracy, etc. They lead the advancement in the recent deep learning development in large datasets and large model eras.

Training algorithms involve engineering methods that develop network architectures and training strategies that enhance learning efficiency and capabilities. The field is witnessing a progressive development in the algorithm of DNNs. Several key milestones have been established within the deep learning community. Early landmarks in the DNN design include the development of convolutional neural networks (CNNs) for image recognition^[43, 48] and recurrent neural networks (RNNs), such as LSTM networks^[44], for natural language processing (NLP). Recent achievements involve the generative adversarial neural networks (GANs)^[49],

which start a new era of building generative models, and Transformers, which revolutionized NLP by enabling models like BERT^[50] and GPT^[51] to understand and generate human-like text. Section 2.2.1 will introduce typical DNN algorithms in the modern era that emerged from the deep learning community. They also have a significant impact in their applications to experimental particle physics.

The computational advances have enabled the model to train faster and deploy more efficiently, allowing the model to become larger, more complex, and thus more performant. The evolution of hardware such as GPUs and field-programmable gate arrays (FPGAs) has played a vital role in this transformation. GPUs, with their parallel processing abilities, have proven to be very effective for parallelizable tasks, especially for the training of DNNs.

The deep learning applications were centred in the domain of computer vision (CV) and NLP in their early development, but it has now found applications across a wide range of domains involving all aspects of human lives and work. Deep learning applications for science are especially intriguing as they can potentially transform how we do scientific research or deliver scientific discovery. One example is the invention of AlphaFold^[52, 53], which sets new standards in protein structure prediction and accelerates biological research. In Part IV (Sections 7–9), a detailed discussion of its role in improving the understanding of the Higgs boson properties, which is a central focus of this dissertation, will be presented.

The following subsections will discuss detailed concepts in algorithm developments, which are crucial to the rest of the dissertation.

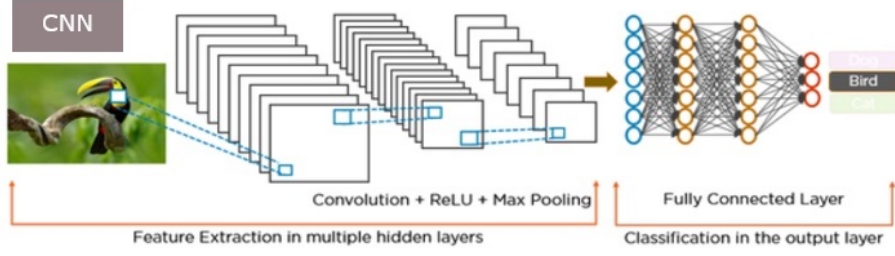
2.2.1 Modern deep learning algorithms

The MLP is the prototype of the neural network design, which is good at handling fixed-dimensional data. The real-world data often manifest structural patterns and inherent hierarchies. Various algorithms are designed to accommodate different types of data.

CNNs are designed to process data with grid-like structures, such as images. They employ layers of convolutional filters to capture spatial hierarchies in the data. RNNs are designed to process sequential data such as text or speech, and they have the unique ability to maintain a memory of previous inputs through loops in their architecture. Figure 2.2 shows the prototype of the two network classes. They drive the early advancements in CV and NLP, respectively.

Following the widespread adoption of CNNs and RNNs, more expressive network architectures tailored for specific tasks were introduced, notably graph neural networks (GNNs) and Transformers. These architectures have also substantially advanced the techniques in particle

Convolutional Neural Network



Recurrent Neural Network

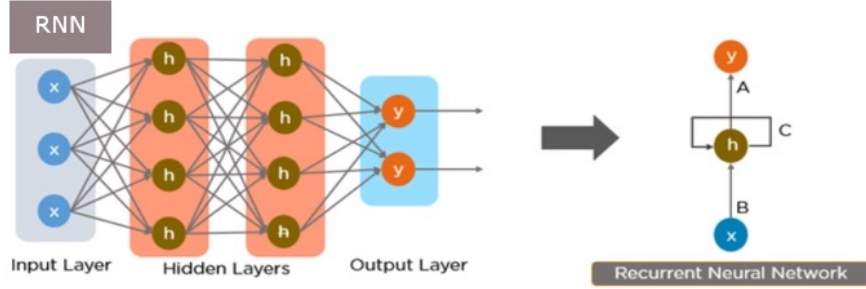


Figure 2.2 The prototype of convolutional neural networks and recurrent neural networks.

physics. A detailed review of the structures and principles in GNNs and Transformers will be provided below.

Graph neural networks GNNs (initiated in, e.g. Ref.^[54]) extend deep learning frameworks to graph-structured data. The algorithm efficiently proceeds with the local node features, the edge features between pairwise nodes, and the global graph features by dedicated network designs, mainly through aggregating information from a node's neighbours (known as message passing). This allows GNNs to be especially suitable for node classification, link prediction, and graph classification. To be specific, consider the graph $G = (V, E)$ built from a list of vertices and edges. In its l -th layer, denote the node feature as v_i^l for the node i , and the edge feature e_{ij}^l connecting the nodes i and j , the message passing from all neighbouring nodes to a given node can be generally described as an edge-to-node and a node-to-edge message passing, written as

$$\begin{aligned} \tilde{e}_{ij}^l &= \rho_{V \rightarrow E}^l(v_i^l, v_j^l, e_{ij}^l), \\ \tilde{v}_i^l &= \sum_{j \in \mathcal{N}(i)} \rho_{E \rightarrow V}^l(v_i^l, v_j^l, \tilde{e}_{ij}^l), \end{aligned} \quad (2.1)$$

then followed by the encoding layer

$$\begin{aligned} e_{ij}^{l+1} &= f_E(\tilde{e}_{ij}^l), \\ v_i^{l+1} &= f_V(\tilde{v}_i^l), \end{aligned} \quad (2.2)$$

where $\mathcal{N}(i)$ is the set of neighbours for node i , and $\rho_{V \rightarrow E}^l(\cdot)$, $\rho_{E \rightarrow V}^l(\cdot)$, $f_E(\cdot)$ and $f_V(\cdot)$ are the neural network functions. The prototype of a layer in GNNs can be found in Fig. 2.3, which depicts the two message-passing steps.

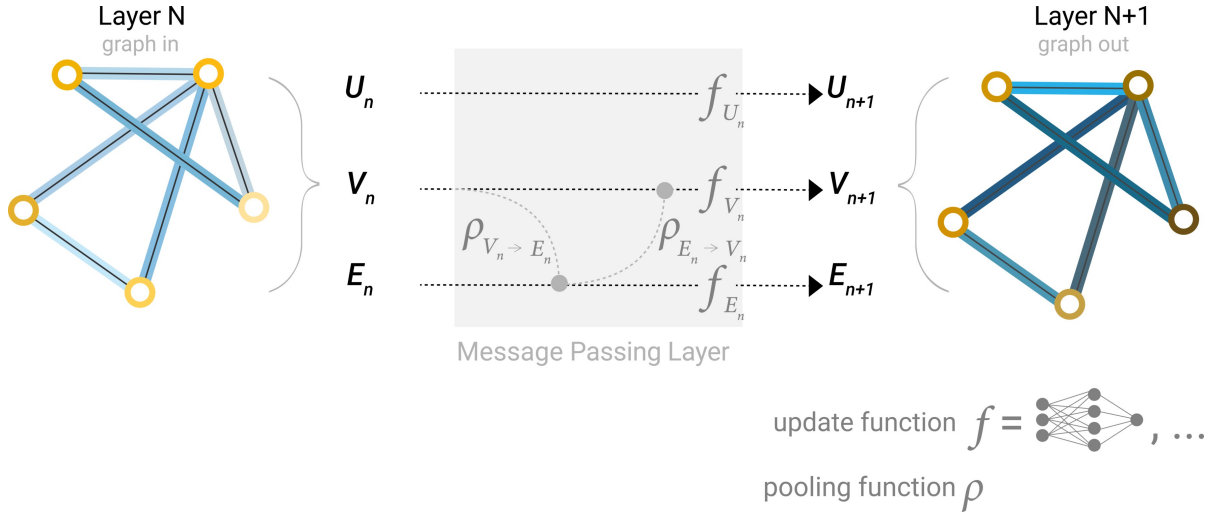


Figure 2.3 The prototype of a message-passing layer in the graph neural networks^[55].

Transformers The Transformer network^[56] is a milestone in DNN algorithm design and it started by setting a new standard in NLP by purely exploiting self-attention mechanisms. This algorithm enables parallel processing and data and reduces training times. It brings the birth of the large language models such as BERT^[50] and GPT^[51] that begin to transform human lives in many aspects. Transformers have also rapidly manifested themselves as a universal algorithm, suitable to be applied to many other domains such as CV (ViT^[57] with its variants, e.g. Refs.^[58, 59]), graph learning (e.g. Ref.^[60]), point cloud recognition (e.g. Ref.^[61]), etc. Transformer models have also shown their power in solving science problems. Examples include AlphaFold 2^[53] for accelerating protein structure prediction, and Pangu-Weather^[62] for improving global weather forecasting, etc.

For generality, the multi-head attention operates on tokens represented by three vectors as

input: the query ($\mathbf{Q}$), key ($\mathbf{K}$), value ($\mathbf{V}$) (which can be equal in practice),

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O, \quad (2.3)$$

where $\text{head}_i = \text{Attention}(\mathbf{Q} \mathbf{W}_i^Q, \mathbf{K} \mathbf{W}_i^K, \mathbf{V} \mathbf{W}_i^V),$

Here, $\mathbf{W}^O, \mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ are parameter matrix of the linear layer, and the attention mechanism is written as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \mathbf{K}^T}{\sqrt{d}}\right) \mathbf{V}. \quad (2.4)$$

Figure 2.4 visualizes this procedure in a schematic diagram. The feed-forward network (FFN) is a two-layer MLP:

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x} \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2. \quad (2.5)$$

It is connected after the self-attention block and forms a Transformer block. The stacking of the Transformer block with residual paths is the prototype of the Transformer. It is able to handle a wide array of data formats, ranging from visual inputs, where tokens represent individual image patches, to textual content, with tokens corresponding to discrete pieces of text, or for the particle-format data (as shown in Fig. 2.4) with tokens represent particle features.

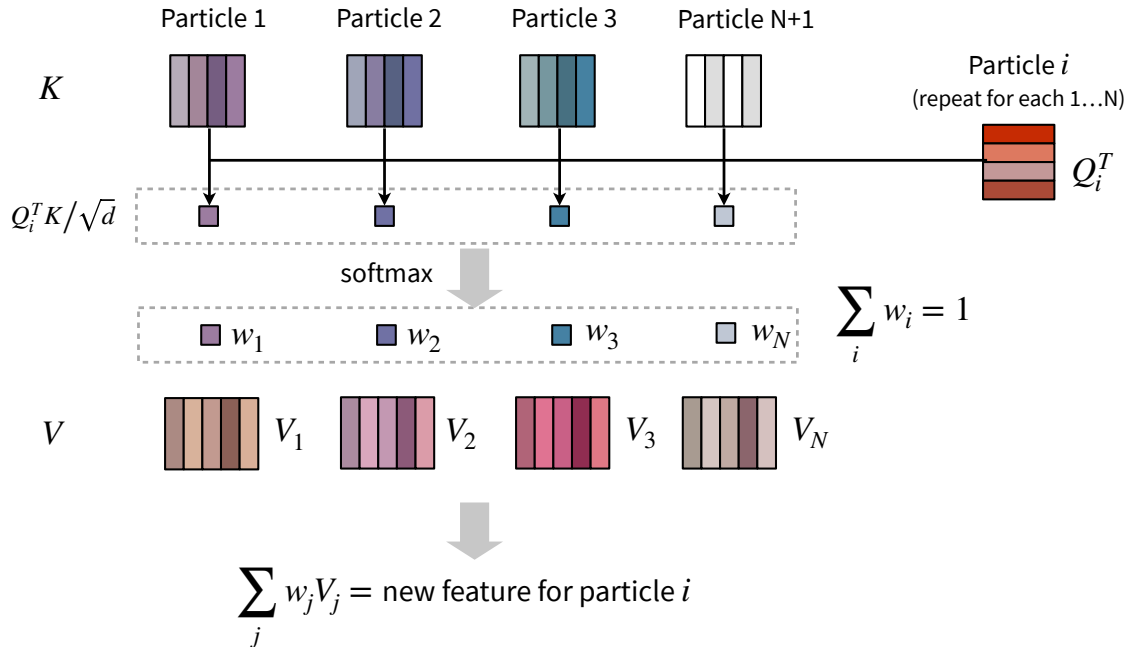


Figure 2.4 Illustration of the Transformer self-attention mechanism.

2.3 Deep learning dataset

Data serves as the essential fuel for training performant models in the deep learning era. The ImageNet dataset^[47] is widely acknowledged as a key driver behind the rapid advancement of modern deep learning algorithms. The quest for large, labelled datasets containing million-level samples or beyond is crucial in almost every domain to advance deep learning progress. Datasets produced outside of the experimental groups are widely used to develop jet-tagging DNNs. This section introduces the datasets specifically developed for the jet tagging domain, which will be revisited in both Part II and V.

Unlike the CV and NLP domains, where unlabelled datasets are most common and abundant, public datasets in the high-energy physics domain mostly come from simulated events that are naturally labelled. These datasets are simulated with MC event generators for hard-scattering process and DELPHES^[63] for fast simulation of detector effects and reconstruction of common analysis-level objects. DELPHES is incredibly useful for delivering phenomenological studies in LHC physics and deep learning. In DELPHES, stable particles produced by event generators are propagated through the detectors configured with the DELPHES card and interacted with the tracker, ECAL and HCAL modules sequentially to build charged-particle tracks and calorimeter towers with simplified algorithms, based on predefined detector geometry, granularity and resolution, etc. The processed tracks and towers are merged into the E-flow object collection, which is considered the reconstructed particles using all the detector information, in analogy with the PF candidates in the CMS experiment. Jets are clustered from E-flow objects with the jet clustering algorithm, and they constitute the samples in the jet tagging dataset.

The following paragraphs overview three jet tagging datasets.

Top tagging dataset The top tagging dataset (abbreviated as the Top dataset) comprises 1.2/0.4/0.4 M jets for training, validation, and testing. It includes two sets of large- R jets, one initiated from the top quark and the other from the QCD multijet events initiated by quarks and gluons. Events in this dataset are generated by PYTHIA8^[64] and passed to DELPHES^[63] for fast simulation of the detector effect. Jets are clustered from the E-flow objects using the anti- k_T algorithm with $R = 0.8$. The dataset includes kinematics information of jet constituent, with the 4-momenta saved in Cartesian coordinates (E, p_x, p_y, p_z) for up to 200 highest- p_T constituents.

Quark–gluon tagging dataset The quark–gluon tagging dataset (abbreviated as the QG dataset) includes 1.6/0.2/0.2 M jets for training, validation, and test. It includes two sets of small- R jets, the quark jets and gluon jets, initiated from $Z(\rightarrow \nu\nu) + u/d/s$ and $Z(\rightarrow \nu\nu) + g$ generated by PYTHIA v8^[64], respectively. Jets are clustered from the E-flow objects with anti- k_T and $R = 0.4$.

The dataset includes kinematics information of jet constituent and particle identification (PID) information according to the reconstructed E-flow object. The PID varies from 8 types, including electrons, muons, photons, charged hadrons (including charged pions, kions and protons), and neutral hadrons (including neutral pions and neutrons). The PID is presented by an integer following the Particle Data Group^[17] numbering scheme. Particles are assigned a positive PID, antiparticles a negative one. For reference, the absolute value of PID for $e, \mu, \gamma, \pi^\pm, \pi^0, K^\pm, p, n$ are 11, 13, 22, 211, 130, 321, 2212, 2112.

Based on the earlier usage of the PID information for the QG dataset, two scenarios are defined according to the convention in Ref.^[65]. The first scenario is abbreviated as QG_{exp} , where the information is restricted into only five classes, including electrons, muons, photons, charged hadrons, and neutral hadrons, according to the detector limitation in real LHC experiments further to distinguish the hadron types with charged and neutral ones. The second scenario is named QG_{full} , where full identification of hadron types presented in this dataset is used. Table 2.1 indicates the difference between the two scenarios.

JetClass dataset The JETCLASS dataset is large, comprising 100/20/20 M for training, validation, and testing. It is composed of ten classes of large- R jets, including five decay modes of the Higgs boson, two decay modes of the top quark, two initiated by Z and W bosons, and one from the QCD event initiated by quarks and gluons. Events in this dataset are first generated by MG^[25] for resonance production and their decay, then they proceed to PYTHIA v8^[64] for parton showering and DELPHES^[63] for detector effect simulation. Jets are clustered similarly from the E-flow objects with $R = 0.8$. The constituent-level features are used as the input to the jet network, including the kinematics information, particle identification flags, and trajectory displacement features, which are close to the feature list for the tagging development within the CMS experiment.

Due to the nature of multiclass classification, the recommended metrics are in the form of binary background rejection, which evaluates the ability to separate two certain classes, S (signal) and B (background), where B is typically the QCD class. Evaluating this metric

involves first defining binary S vs. B discriminant as

$$\text{disc.}(A \text{ vs. } B) = \frac{\text{score}(A)}{\text{score}(A) + \text{score}(B)}, \quad (2.6)$$

then calculating the background rejection rate ($1/\epsilon_B$) under given signal efficiency ϵ_S at $X\%$, similar to metrics defined in the Top and QG datasets.

As a summary of the included features, Table 2.1 showcases the particle features in the Top, QG (QG_{exp} and QG_{full}), and JETCLASS datasets. It includes several customizations to standardise the inputs for the ease of training, e.g. the use of logarithm on momentum and energy, and the hyperbolic tangent on impact parameters. This will be the main configuration adopted in the studies in Part II unless specifically mentioned.

Table 2.1 Summary of the particle input features used for the Top, QG (QG_{exp} and QG_{full}), and JETCLASS dataset, adapted from Ref.^[65]. Customizations are done including the expansion of kinematic features and the use of logarithms and hyperbolic tangents for input standardization.

Category	Variable	Definition	Top	QG _{exp}	QG _{full}	JETCLASS
Kinematics	$\Delta\eta$	difference in pseudorapidity η between the particle and the jet axis	✓	✓	✓	✓
	$\Delta\phi$	difference in azimuthal angle ϕ between the particle and the jet axis	✓	✓	✓	✓
	$\log p_T$	logarithm of the particle's p_T	✓	✓	✓	✓
	$\log E$	logarithm of the particle's energy	✓	✓	✓	✓
	$\log \frac{p_T}{p_T(\text{jet})}$	logarithm of the particle's p_T relative to the jet p_T	✓	✓	✓	✓
	$\log \frac{E}{E(\text{jet})}$	logarithm of the particle's energy relative to the jet energy	✓	✓	✓	✓
	ΔR	angular separation between the particle and the jet axis ($\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$)	✓	✓	✓	✓
Particle identification	charge	electric charge of the particle	—	✓	✓	✓
	Electron	if the particle is an electron ($ PID ==11$)	—	✓	✓	✓
	Muon	if the particle is a muon ($ PID ==13$)	—	✓	✓	✓
	Photon	if the particle is a photon ($PID==22$)	—	✓	✓	✓
	CH	if the particle is a charged hadron ($ PID ==211$ or 321 or 2212)	—	✓	✓ ^①	✓
	NH	if the particle is a neutral hadron ($ PID ==130$ or 2112 or 0)	—	✓	✓ ^②	✓
Trajectory displacement	$\tanh d_0$	hyperbolic tangent of the transverse impact parameter	—	—	—	✓
	$\tanh d_z$	hyperbolic tangent of the longitudinal impact parameter	—	—	—	✓
	σ_{d_0}	error of the measured transverse impact parameter	—	—	—	✓
	σ_{d_z}	error of the measured longitudinal impact parameter	—	—	—	✓

① $(|PID|==211) + (|PID|==321)*0.5 + (|PID|==2212)*0.2$

② $(|PID|==130) + (|PID|==2112)*0.2$.

Part II

Algorithms for Jet Tagging

Chapter 3

An Overview on Algorithms

Jet tagging aims to identify a jet's origin using the available properties associated with a jet. Developing algorithms for large- R jet tagging is more challenging due to the larger complexity of the jet constituent particles and the diversity in the structure. This chapter will focus on algorithms developed for the large- R jet tagging. First, a review of its development history will be presented, spanning from the earlier rule-based criteria to the flourishing of algorithms when machine learning approaches are widely considered.

Especially, in the modern deep learning era, new opportunities arise when one can dive deep into the low-level particle features and use an advanced neural network to explore the comprehensive relations among the input features and different particles. This also brings new challenges in making the algorithms more performative, where the knowledge is rooted in the intersection of deep learning advances and particle physics. This chapter will review the development path of deep learning algorithms. In Chapter 4, the recent advances on this forefront will be highlighted.

3.1 Rule-based algorithms

Rule-based algorithms have been developed to reflect jet properties from various aspects. This section focuses on three aspects of them.

3.1.1 Jet grooming

Jet grooming techniques^[37-39, 66-68], initiated a decade ago, particularly aid in mass reconstruction when the large- R jet is initiated from a resonance. These techniques aim to remove soft emissions within a jet that come from the pileup or underlying events. Including the soft particles in the mass calculation may result in a smeared mass distribution.

Four methods are highlighted and have been explored in the CMS and ATLAS experiment, namely, jet trimming, jet pruning, jet filtering, and the soft drop algorithm. To probe the soft patterns within a jet, these algorithms utilize the technique to recluster all constituents within

a jet to form subjects to identify the soft ones, or decluster the jet into constituent particles and find the steps with soft radiation pattern. The procedure is briefly outlined in Fig. 3.1.

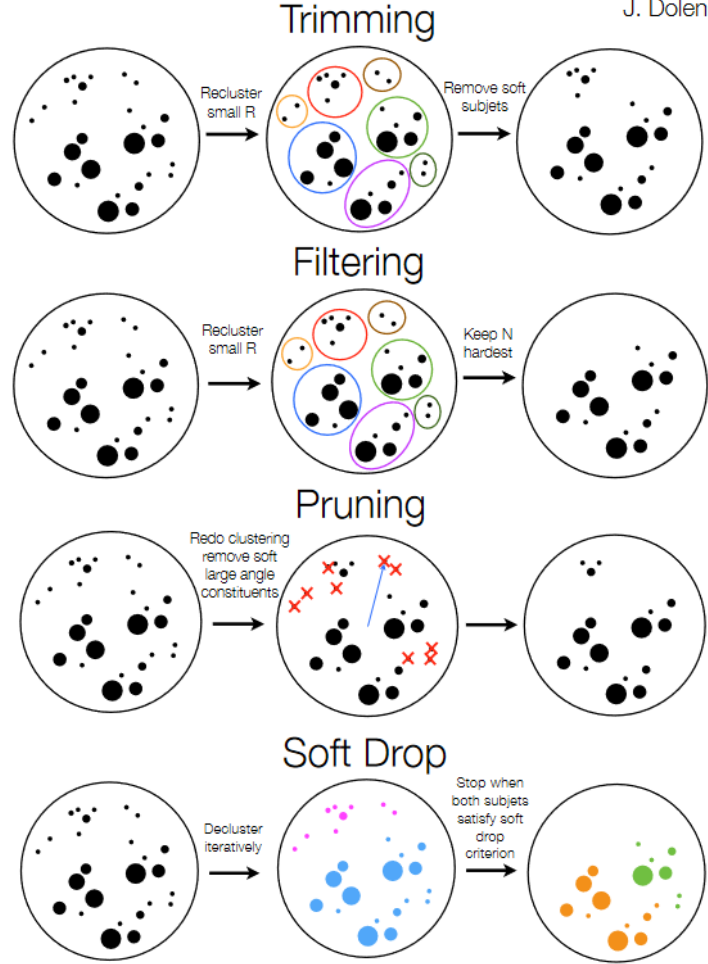


Figure 3.1 A schematic image of the four jet grooming algorithms: jet trimming, jet pruning, jet filtering, and the soft drop algorithm. Image source: *CMS Open Data workshops*

The soft drop algorithm among the four is most frequently used in the CMS experiment, and its details can be found in Section 1.5.5. It is also called the modified mass drop tagger (mMDT)^[37]. The variable can be incorporated with a multivariate analysis or used for likelihood fit to extract the signal jets directly.

3.1.2 Jet substructure

Probing the jet substructure is a good handle to separate signal jets from backgrounds, especially for the large- R jets where the substructure can reveal more complexity. One widely used approach is to define N -subjettiness^[69] to characterize the jet substructure. This variable

provides a measure of the N subjets that can be found inside the jet, defined as

$$\tau_N = \frac{1}{d_0} \sum_{i=1}^M p_{T,i} \min\{\Delta R_{i,sj_1}, \Delta R_{i,sj_2}, \dots, \Delta R_{i,sj_N}\}, \quad (3.1)$$

where M is the number of constituents in the jet, $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ represents the angular separation between each constituent and the subjets. The subjets within the jet are determined from an exclusive k_T algorithm on the original jet constituent before jet grooming is applied. If the jet resembles an N -subjett structure, the constituents will be centred to the N axis, resulting in a smaller τ_N value. On top of this, the N -subjettiness ratio $\tau_{NM} = \tau_N/\tau_M$ is usually defined to optimize for separating two classes of jets.

The N -subjettiness has been widely used in the CMS experiment to identify the 2-prong or multipronged signatures originating from a resonance. For example, τ_{21} is designed to tag 2-prong hadronically decayed $W/Z/H$ jets against the QCD multijet background, and τ_{32} can be designed to identify 3-prong hadronically decayed top jets. Figure 3.2 shows the τ_{21} and τ_{32} normalized distributions for the $W/Z/H$ jets and top jets, respectively, comparing them with the distribution of the QCD background jets to demonstrate the discriminating power.

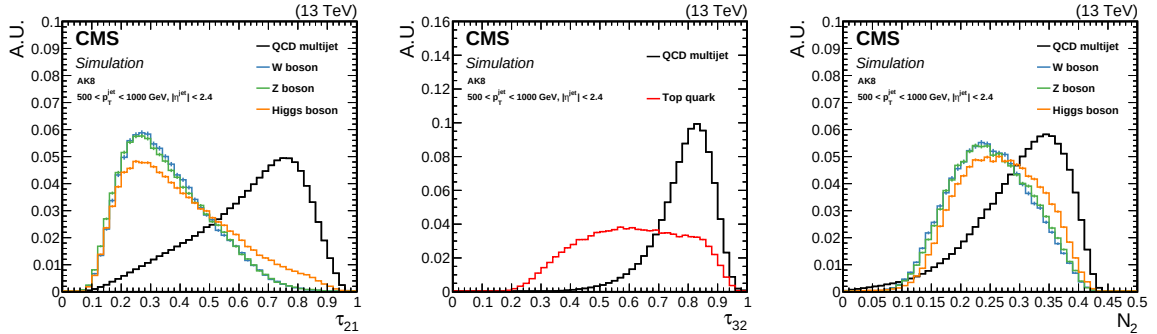


Figure 3.2 The shape comparison for multiple resonant types of jets and QCD background jets on the N -subjettiness ratio τ_{21} and τ_{32} , and generalized energy correlation function ratio N_2 [70].

Another set of techniques reveals the energy distribution within a jet without the need to reconstruct subjet axes. The construction is initiated from the C -correlators [71],

$$C_N^{f_N} = \sum_{i_1=1}^M \dots \sum_{i_N=1}^M E_{i_1} \dots E_{i_N} f_N(\hat{p}_{i_1}^\mu, \dots, \hat{p}_{i_N}^\mu), \quad (3.2)$$

where f_N is a function on the pure direction of the particles, $\hat{p}_i^\mu = p_i^\mu/E_i$. This function correlates N particles among M ones using a customized function f_N to characterize their

spatial distributions, and it has good properties such as infrared and collinear safety. This induces the energy correlation functions (ECFs)^[72, 73] by choosing

$$f_N(\Delta R_{ij}) = \prod_{1 \leq i < j \leq M} \Delta R_{ij}^\beta, \quad (3.3)$$

where ΔR_{ij} is again used to describe the distance on the 2D plane spanned by the pseudo-rapidity and azimuthal angle (η - ϕ plane) to adapt the hadron collider's configuration. ECF variables facilitate tagging the N -prong structure within the jet. ECFs can be generalized to choosing only the smallest m values of ΔR_{ij} to make the product in Eq. (3.3), instead of using all $M(M-1)/2$ values. For example, a widely used 2-prong jet tagger in CMS N_2 , i.e. a ratio constructed from generalized ECFs using $m = 2$ and 1, is written as

$$N_2 = \frac{\sum_{1 \leq a < b < c \leq M} \frac{p_{T,a} p_{T,b} p_{T,c}}{p_{T,J}^3} \min\{\Delta R_{ab} \Delta R_{ac}, \Delta R_{ab} \Delta R_{bc}, \Delta R_{ac} \Delta R_{bc}\}}{\left(\sum_{1 \leq a < b \leq M} \frac{p_{T,a} p_{T,b}}{p_{T,J}^2} \Delta R_{ab} \right)^2}. \quad (3.4)$$

The distributions of N_2 on $W/Z/H$ jets and QCD jets are shown in Fig. 3.2 on the right plot. The series of generalized ECFs and their ratios constructed like N_2 can be exploited in a multivariate approach to yield better discrimination. Besides, another extension of ECFs is to define the energy flow polynomials (EFPs)^[74], where the used ΔR_{ij} for constructing the product in Eq. (3.3) is selected from the edge of a given graph G . The ECFs then correspond to the special case defined on the fully connected graph. Advances in defining rule-based algorithms to probe the jet energy distribution are still rapidly growing.

3.1.3 Heavy-flavour jet properties

Identifying heavy flavour jets, i.e. those initiated from a b or c quark, is a common practice for small- R jet tagging. A b or c quark produced in the hard-scattering process will form a b - or c -hadron, and they will subsequently decay via the weak interaction to the light hadrons (pions or kaons), potentially along with the production of a nonprompt lepton with low p_T (known as the soft leptons). This results in a relatively longer lifetime for these hadrons, so their decay produces tracks which are displaced from the PV. Also, an SV can result from those displaced tracks. Figure 3.3 illustrates the key features of typical heavy flavour jets.

Efforts in CMS heavy-flavour jet tagging rely heavily on the track displacement, the existence of soft leptons, and the SV features. Below is a brief overview of these approaches. First, algorithms such as jet probability (JP) and jet b probability (JBP) for AK4 tagging have used the impact parameter significance of a track to reconstruct a track probability^[75]. This

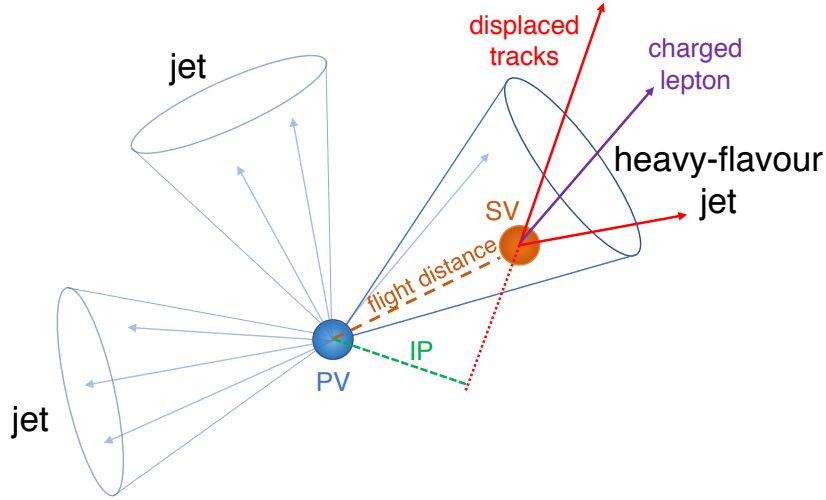


Figure 3.3 Illustration of a heavy-flavour jet initiated from the decay of a b or c hadron resulting in tracks that are displaced with respect to the primary vertex (PV), and possibly reconstructed SV within the jet^[75].

track property will differentiate tracks decayed from heavy-flavour hadrons with light-flavour ones, and it is used to construct a jet probability for b or c jet tagging. Besides, the existence of soft leptons has been used as a requirement to select jets from QCD multijet events with less light jet contamination, aiding the calibration of AK4 b or c taggers in CMS^[75-77]. In the CMS experiment, due to the high efficiency and accuracy in identifying final state muons, dedicated HLT paths are designed to select muon-enriched AK4 or AK8 jets^[78]. The features of SV are also good discriminants to select b jets. These include the corrected SV mass, the number of tracks included in SV, and the SV's azimuthal relation to the jet. In the CMS jet tagging practice, these criteria are often combined in a multivariate approach thanks to machine learning techniques. These techniques to identify jets initiated from a single b or c hadron are also extendable to tag large- R jets, such as jets initiated from the $H \rightarrow b\bar{b}$ or $H \rightarrow c\bar{c}$ signature.

3.2 Shallow machine-learning algorithms

As many hand-crafted variables can be used to reflect some aspects of jets, multivariate analyses can be performed to combine the benefits from all variables. This approach is referred to as “shallow machine-learning algorithms” in this dissertation and the method has been widely adopted in the CMS experiment to determine the strongest tagger. For example, early top tagger exploits all m_{SD} , τ_{32} and b -tagging properties to enhance the distinguishing ability for top jets versus other jets^[70]. The series of N -subjettiness ratios and generalized EFP

ratios have also been exploited together to construct a boosted design tree (BDT), achieving superior results compared to every single one.

In addition to these developments, the variables can be constructed in a neural network approach using a few-layer MLP. Using a neural network also allows for multiclass classification, paving the way towards a unified tagger capable of identifying several types of jets. Such algorithms include the boosted event shape tagger (BEST)^[70] that is able to classify all $W/Z/H/t$ jets and QCD jets. The input to the MLP algorithm includes the jet kinematic variables and m_{SD} , the b -tagging properties, the invariant masses for the different pairing of its subjets, and other high-level jet features such as sphericity and trust. The variables are calculated several times by Lorentz-transform the jets into the rest frame in assumption that the jet mass is configured to different resonant mass.

3.3 Deep learning algorithms

The methodologies for jet tagging have been revolutionized in the age of deep learning. Because a DNN can proceed directly with the input as a low-level representation of the jet, there is less need to construct hand-crafted variables. Instead, one can directly input the underlying low-level variables and design an efficient network to extract useful features. It is worth noting that these methods are beyond the early MLP-based deep learning approach, which works like a multivariate analysis. Instead, they exploit the larger, deeper network with an architectural upgrade to proceed directly with the hierarchical particle-format input within a jet.

3.3.1 Jet representation of images and sequences

The initial work of applying modern DNN architectures in jet tagging utilizes networks like CNNs and RNNs. The initiative to use CNN and proceed with the jet as an image is rooted in the fact that one fundamental form of the raw jet input is the calorimeter energy deposit, which has a grid-like structure. For the use of RNN, there is an analogue that jets can be viewed as sentences where the QCD is the “grammar” of the language. Early work adopts RNN to deliver jet tagging to attempt to understand the jet’s grammatical structure.

For the CNN architecture, in addition to representing a jet on the 2D η - ϕ grid, it can also be presented on the 1D sequence of the particles, usually ordered inversely by p_T , such that the convolutional filter acts on a particle its neighbouring particle (with similar p_T) that follows the predefined order.

A variety of studies has been presented using CNNs^[79-82] or RNNs^[83, 84] for jet tagging.

Experimentally, applications of the CNN via the image-like representation include the ImageTop tagger^[70, 82] developed in CMS. As briefly demonstrated in Fig. 3.4, this approach defines customized pixels to “rasterize” the particles to extract the top-jet properties. At the same time, it also includes a second neural network path to embed the features related to the b -tagging of subjets. The typical CNN developed under the particle’s 1D representation in CMS is the DeepAK8 algorithm^[70]. The DeepDoubleX algorithm^[85, 86] developed in CMS combines CNN and the gated recurrent unit (GRU, a type of RNN layer) and achieves an excellent performance among methods under these representations. A schematic architectural outline can be found in Fig. 3.5.

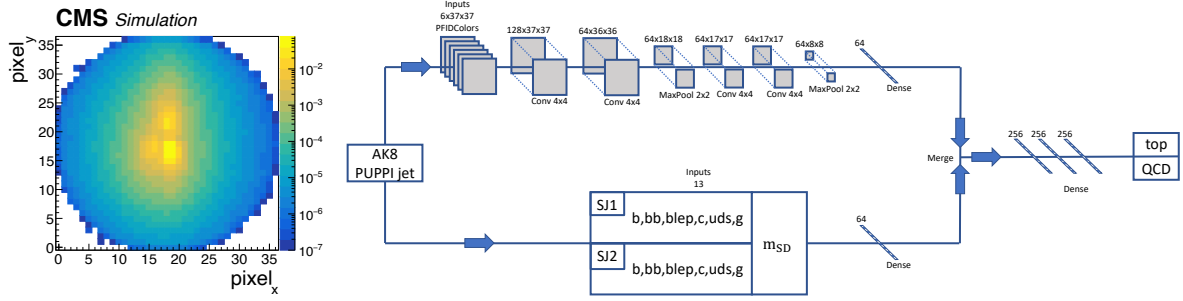


Figure 3.4 The image representation of top jets shown as an ensemble of overlaid images (left) and a schematic diagram of the ImageTop algorithm for boosted top jet tagging (right)^[70].

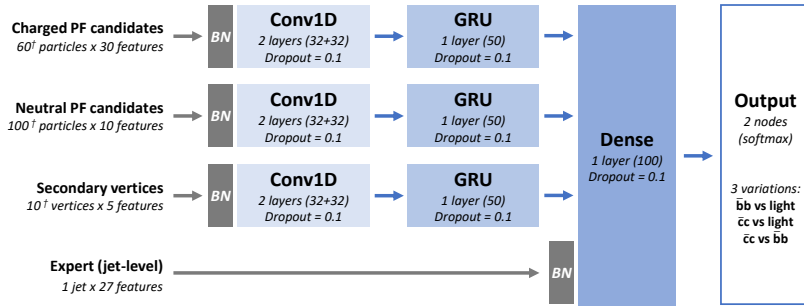


Figure 3.5 A schematic diagram of the DeepDoubleX algorithm (V2)^[86] for boosted heavy-resonant $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging.

3.3.2 Jet representation of sets and graphs

It has been recognized that using images or sequences to represent jets had inherent deficiencies regarding how the data is presented. For instance, jet images are sparse, i.e. many grid points remain unfilled by jet particles, making image representation less efficient. In addition,

when using RNNs or CNNs based on 1D sequential representations to process jets, a major issue is that a predefined sequence must order them. This suggests that unlike text structured by specific grammar, the particles representing a jet are inherently permutation symmetric and do not require a dedicated sequence to confer meaning.

This insight led to the representation of jets through sets (equivalently referred to as point clouds) or graphs. This approach eliminates redundant data and adheres to the permutation invariance. Early notable works include the Particle Flow Network (PFN)^[87]. It first performs elementwise MLP embedding to all particles, and then pooling on the particle's dimension, followed by a final further MLP embedding. This is an early attempt to respect permutation symmetry, yet it succeeds on the quark–gluon tagging benchmark. Meanwhile, as GNNs (typically full-connected GNNs) have become mature in deep learning and naturally provide a permutation invariant way to process jets, their applications to jet tagging have seen a boost (e.g. Refs.^[88, 89]). ParticleNet^[90] is among those that bring jet tagging to the next performance level, and it also witnesses a large success when applied in the CMS experiment for the $W/Z/H/t$ tagging and for heavy-resonant $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging (technical details will be revisited in Section 5.1). Based on the successful Dynamic Graph CNN (DGCNN) developed for processing point clouds^[91], ParticleNet dynamically identifies and connects each particle to its k nearest neighbours in η – ϕ space, extracting local features efficiently. Figure 3.6 provides a detailed explanation of the ParticleNet architectural design.

Studies at this stage highlight the importance of choosing representations close to the data for designing efficient and performative networks. Following ParticleNet, graph-based networks like LundNet^[92], ABCNet^[93], LorentzNet^[94], HMPNet^[95] etc. further iterated on this approach, yielding improvements or similar performance on the jet tagging benchmarks. They prompt reflection on the key to these effective strategies for jet tagging with the modern DNN approach.

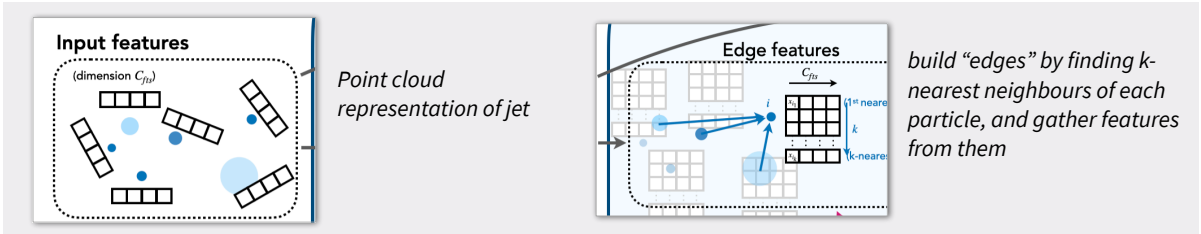


Figure 3.6 A detailed explanation of the ParticleNet architectural design.

Chapter 4

New Advances in Deep Learning Algorithms

Following the discussion above, the community for jet tagging using DNNs has widely recognized that sets (point clouds) or graphs are more effective for representing particle-format data. This chapter will highlight new insights in deep learning algorithms developed for jet tagging, after this phase of progress. The topic will be brought up from two aspects, the consideration of symmetry preservation within the network design, and the advances driven by the new architecture, Transformers.

4.1 Algorithms preserving Lorentz symmetry

4.1.1 Preserving symmetries

It is commonly understood that designing networks to adhere to the intrinsic symmetries within data acts as an inductive bias, and the principle works as follows. For data in any domain, it may have an inherent property that its attribute can remain unchanged if it undergoes a certain type of transformation. This is regarded as an inherent symmetry that the data possesses. If the network is designed so that its neuron activations, or even its overall output, remain unchanged when the input data are transformed in this manner, it becomes more robust to these transformations and thus aligns with the properties of the data. This provides specific knowledge to the network algorithm, which can often improve the network's performance. An example from early deep learning experiences is the notable success of CNN for image recognition. The success can be attributed to the symmetry preservation property, i.e. the CNNs are robust to variations in images, such as shifts and distortions of some elements within the image, due to their unique property of convolutional filter design.

From this viewpoint, the permutational symmetry of particles within jet data can be recognized as the fundamental one in particle-format data. This leads to a aforementioned change in how to represent a jet, i.e. from an ordered list towards an unordered set.

It has also been recognized that a jet's property should remain invariance when the entire jet undergoes a boost along the beam direction, i.e. the z axis (denoted as z - t boost), or a rotation around that axis (denoted as x - y rotation). There is an inherent aspect of hadron colliders

that the physics properties of an event should remain unchanged when all postcollision particles undergo z -axis boost and x - y rotation. Therefore, many early algorithms (e.g. Refs.^[79, 80]) have already brought in a data preprocessing step to undergo these two transformations into a reference direction. Viewing a jet on the η - ϕ plain, this procedure means to shift the input jet towards the centre of the plane.

The works based on jet image representation consider another preprocessing that includes an additional rotation on the jet image^[79, 80], or equivalently, on the η - ϕ plane. Studies based on set (point cloud) or graph representation that emerged in recent years also consider retaining invariance under dedicated jet transformations via special network designs. These includes preserving symmetries under boosts on z axis or rotation on the x - y plane^[96], rotation on the η - ϕ plain^[97] (or similarly, around the jet axis^[98]), boost along the jet axis^[98], or even under full Lorentz transformations^[94, 99]. The experiment in each work reveals some gain when such symmetry is incorporated.

It can be easily recognized that all these transformations belong to the general Lorentz transformations; hence, the symmetry group that this network is designed to adhere to is a subgroup of the Lorentz group. Considering a general way to describe the symmetries, Fig. 4.1 provides a basis for how additional symmetries on top of the invariant under z - t boost and x - y rotation can be classified into four types. Note that, as the 2 degrees of freedom are frozen out of 6, there are 4 additional degrees of freedom to Lorentz transform the jet, and they are the y - z rotation, x - t boost, z -tilt, and y -tilt. The images in Fig. 4.1 represent all constituents within the jet on the η - ϕ space, with the marker area proportional to the constituent's p_T . It is also worth noting that the latter two are a mixture of z - t boost with x - z rotation and a mixture of y - t boost with y - z rotation—the z - t boost and x - z rotation are not commutable, similar to y - t boost with y - z rotation, hence a convention is made first to deliver the boost, followed by the rotation in the following context.

There are several interesting relations one can find with the basis presented in Fig. 4.1 and the symmetries studied in the previous works^[94, 96-99]. First, in the relativistic limit,

$$\begin{aligned}\eta &= \operatorname{arctanh} \frac{p_z}{E}, \\ \phi &= \operatorname{arctan} \frac{p_y}{p_x}.\end{aligned}\tag{4.1}$$

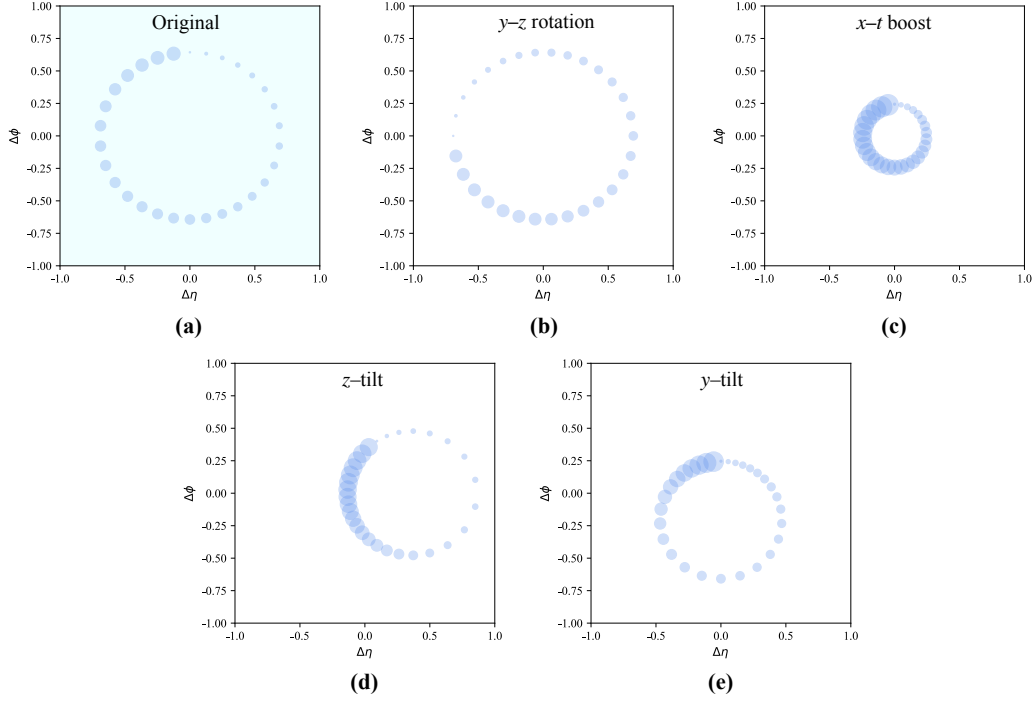


Figure 4.1 Illustration of a toy jet on the η - ϕ plane (a) and its behaviour when it undergoes the four types of Lorentz transformation with its centre fixed at the origin of the η - ϕ plane, which is y - z rotation (b), x - t boost (c), z tilt (d), and y tilt (e).

In the limit of $p_y \sim o(E)$ and $p_z \sim o(E)$ (i.e., $p_{y,z} \ll E$). Define rapidities

$$\begin{aligned} y_y &= \operatorname{arctanh} \frac{p_y}{E}, \\ y_z &= \operatorname{arctanh} \frac{p_z}{E}, \end{aligned} \quad (4.2)$$

they should satisfy $y_{y,z} \sim o(1)$. It is easy to see that

$$\begin{aligned} \eta &= \frac{p_z}{E} + o(y_z), \\ \phi &= \frac{p_y}{E} + o(y_y, y_z), \end{aligned} \quad (4.3)$$

and this proves that η - ϕ rotation is essentially an approximate y - z rotation, which is illustrated also in Fig. 4.1 (b). The approximation holds better when the jet constituents are concentrated within a small cone. In this regard, given that the jet axis should always point to the x axis (or its centre should point to $(0, 0)$ on the η - ϕ space), a Lorentz boost on the jet axis essentially corresponds to a x - t boost, illustrated in Fig. 4.1 (c). On top of these two transformations, there are an additional two indicated in (d, e).

This discussion provides a general framework to quantify which subsymmetries previ-

ous networks have adhered to. This implies that adhering to the full Lorentz symmetry in the network has the most potential for achieving high performance. Therefore, it inspires the community to design a systematic study to understand the effect of preserving these symmetries in the network. Section 4.1.2 first showcases the design of LorentzNet, a specific high-performing network design that incorporates full Lorentz symmetry. Section 4.1.3 will then introduce a systematic study on Lorentz symmetry preservation, where one can see how the performance evolves when the network progressively incorporates more types of the four subsymmetries.

4.1.2 LorentzNet

LorentzNet is a recent GNN-based network with built-in Lorentz symmetry preservation. It closely follows the prototype of GNNs, i.e. each layer has its node features and edge features with message-passing performed between them. In particular, the graph for LorentzNet is fully connected. The difference is that, to preserve full Lorentz symmetries, the node features belong to either Lorentz scalars or Lorentz vectors, while the edge features are pure Lorentz scalars. The message-passing neural network functions in Eq. (2.1) are designed to respect the invariant or equivariant properties of a scalar and vector, respectively.

The Lorentz group equivariant block

The unit block of LorentzNet is the Lorentz group equivariant block (LEGB). It proceeds with three steps, namely, the feature updates of the (1) edge features, (2) the vector node features, and (3) the scalar node features. The procedures are outlined in Fig. 4.2 and highlighted in three individual parts.

Specifically, for the l -th layer, denoting the edge features connecting nodes i and j as m_{ij}^l , the vector features as x_i^l , and the node features as h_i^l , the update procedures are as follow.

1. The first step in Fig. 4.2 aims to update m_{ij} as

$$m_{ij}^l = \phi_e(h_i^l, h_j^l, \psi((x_i^l - x_j^l)^2), \psi(x_i^l \cdot x_j^l)), \quad (4.4)$$

where $\phi_e(\cdot)$ represents a neural network function and $\psi(\cdot) = \text{sgn}(\cdot) \log(|\cdot| + 1)$ serves for feature standardization to ease the training. To maintain that m_{ij} is a scalar, the last two input arguments of $\phi_e(\cdot)$ should also be scalars, and they are constructed from Minkowski dot product $x_i^l \cdot x_j^l$ and $(x_i^l - x_j^l)^2$ using the vectors.

2. The second step to update the vector x_i writes

$$x_i^{l+1} = x_i^l + c \sum_{j \in \mathcal{N}(i)} \phi_x(m_{ij}^l) \cdot x_j^l, \quad (4.5)$$

where $\phi_x(\cdot)$ is a neural network function, $\mathcal{N}(i)$ is the set of neighbours of the node i (i.e. the full nodes in the case of a fully connected graph). This step is a pooling step that gathers the features of neighbouring nodes or the edge connecting them to the node i . To maintain the Lorentz vector property, the output vector can be constructed only with the weighted sum of all input vectors, where weights can be learned from the scalar features m_{ij} .

3. The third step to update the scalar h_i writes

$$h_i^{l+1} = h_i^l + \phi_h(h_i^l, \sum_{j \in \mathcal{N}(i)} w_{ij} m_{ij}^l), \quad (4.6)$$

where $\phi_h(\cdot)$ again models a neural network function, and the weight w_{ij} is modelled by $w_{ij} = \phi_m(m_{ij}) \in [0, 1]$. This is again a pooling step, but the output is guaranteed to be a scalar.

One can see that the edge features are constructed immediately to facilitate updating the node features, and they are not passed through different LGEBs.

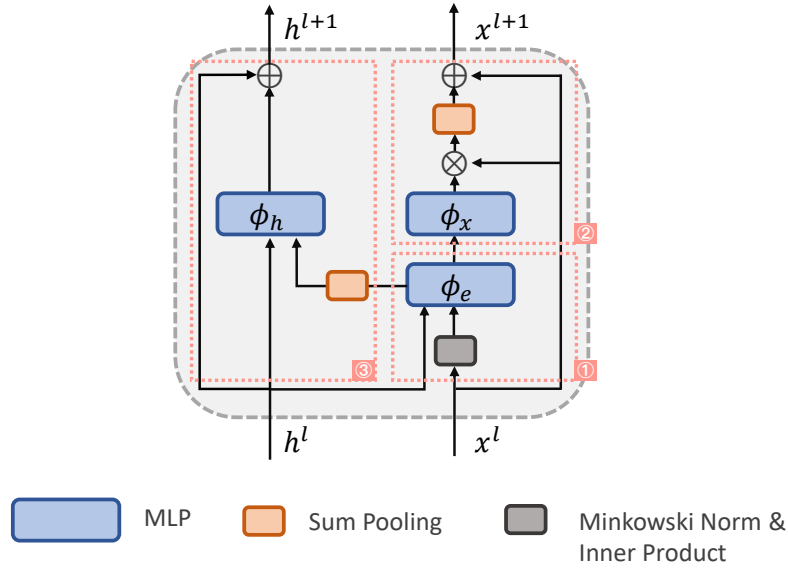


Figure 4.2 A schematic diagram of the Lorentz group equivariant block (LGEB) within LorentzNet, highlighted by three steps. Image adapted from Ref.^[94].

LorentzNet implementation

The LorentzNet model is constructed by stacking $L = 6$ LGEBs to convert the initial node features (h^0, x^0) to (h^L, x^L) . Before proceeding to the first LEGB, the scalar is embedded

through a single-layer MLP, denoted as $\text{Linear}(\text{scalar_dim}, 72)$, where scalar_dim is the input scalar dimension. Within each LEGB, the ϕ_h, ϕ_e, ϕ_h is implemented as

$$\text{Linear}(72, 72) \rightarrow \text{ReLU} \rightarrow \text{BatchNorm1D}(72) \rightarrow \text{Linear}(72, 72), \quad (4.7)$$

and ϕ_m is implemented as $\text{Linear}(72, 72) \rightarrow \text{SigMoid}$, using the pseudocode covention of PYTORCH. After passing L LEGBs, the fully-connected MLP is implemented as

$$\text{Linear}(72, 72) \rightarrow \text{ReLU} \rightarrow \text{Linear}(72, \text{output_dim}), \quad (4.8)$$

where a dropout rate of 0.2 is adopted, and the final output logits are in the dimension of output_dim .

Experiments

LorentzNet on Top and QG_{full} dataset The performance of LorentzNet is benchmarked on the Top and QG datasets (see the definition of datasets in Section 2.3), with details in Ref.^[94]. Initial scalar and vector node features are constructed as follows. The four-momenta of particles in both datasets are the initial vector features. For the scalar node features, the PID information for the QG dataset is used, implemented through the one-hot encoding, hence falling to the QG_{full} scenario when using this information. In the Top dataset, the zero vector is used due to the lack of scalar variables presented.

The LorentzNet training on the Top and QG_{full} datasets adopt the ADAMW optimizer, with a weight decay of 0.01 and a target to minimize the binary cross-entropy loss. It is trained with 35 epochs with an initial learning rate (LR) 1×10^{-3} , and the COSINEANNEALINGWARM-RESTARTS^[100] scheduler equipped with a linear warm-up at the beginning and exponential decay at the end (detailed can be found in Ref.^[94]).

The jet tagging performance on the Top and QG dataset in terms of the classification accuracy, the area under receiver operating characteristic (ROC) curve (abbreviated as AUC), and background rejection rate at certain signal efficiency working points are shown in Tables 4.5 and 4.6, also included with performance from other models. It can be found that LorentzNet achieves superior performance, surpassing ParticleNet by a large margin. In addition, as LorentzNet is built with full Lorentz symmetry preservation, Fig. 4.3 studies the impact of tagging performance when the input jets undergoing a Lorentz boost on x axis (showcased in Fig. 4.1 (c)) with the parameter $\beta = v/c$. It shows the performance remains constant as expected, while the algorithms for comparison, including ParticleNet, manifest a performance drop.

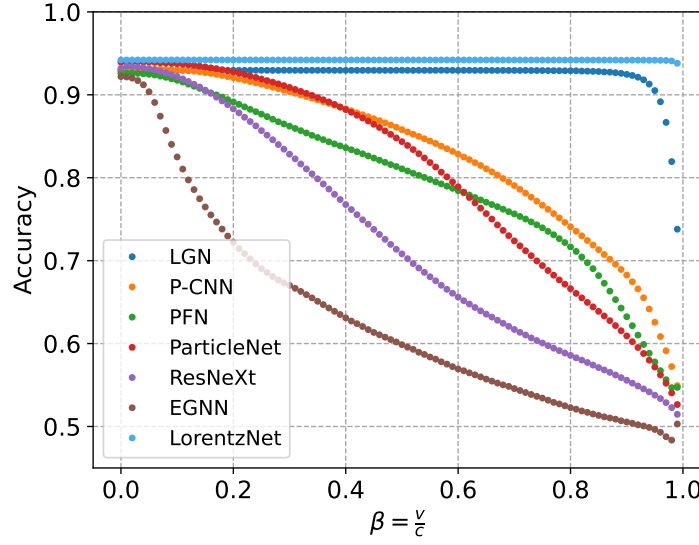


Figure 4.3 Comparison of jet tagging performance in terms of tagging accuracy with different algorithms when input jets undergo a x - t boost with parameter $\beta = v/c$ ^[94].

LorentzNet on JetClass dataset With the recent development of the JETCLASS dataset, LorentzNet has also been benchmarked on this dataset to compare with other algorithms, especially ParticleNet and Particle Transformer (detailed in Section 4.2). The input variables differ from the implementation version in Top and QG datasets as introduced in Ref.^[94]. It is observed that supplementing the scalar features with the full list of inputs shown in Table 2.1 does not impact the LorentzNet performance, although this breaks the perfect Lorentz equivariance property. This is understandable because, in practice, it is unnecessary to strictly require the network to adhere to the Lorentz symmetry to achieve high performance. In contrast, it works fine to hint at the Lorentz-symmetric properties by providing an inductive bias to the network. This supplementation of inputs allows one to synchronize the features used to compare the ParticleNet, LorentzNet, and ParT network performance on JETCLASS and adapt LorentzNet to the JETCLASS dataset where the nonscalar track trajectory features are present.

The LorentzNet training on the JETCLASS dataset is synchronized with the training configuration of ParticleNet^[90] and ParT^[65]. Specifically, The training is performed on the 100 M dataset for 50 epochs, each defined as iterating over 10 M samples. The training adopts the LOOKAHEAD optimizer^[101] with $k = 6$ and $\alpha = 0.5$ and the inner optimizer is RADAM^[102] with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-5}$. The batch size of 256 and initial LR of 2×10^{-3} is adopted. The LR starts to decay exponentially when finishing 70% of the epochs, changing at the beginning of every epochs and dropping to 1% of its initial rate in the last epoch.

The performance on LorentzNet can be found in Table 4.7. Again, a substantial perfor-

mance gain can be found compared with ParticleNet. Further discussions will be provided in Section 4.2.

4.1.3 A systematic study in preserving Lorentz symmetry

The success of LorentzNet implies the importance of preserving the Lorentz symmetry within a high-performant network to achieve even higher performance. Meanwhile, an upgraded version of ParticleNet, named ParticleNeXt, indicates that incorporating the pairwise particle invariant masses can bring a visible improvement. Given that these pairwise masses span a basis to construct all Lorentz scalars available for each jet, it leads one to suspect that Lorentz symmetry preservation is the underlying fact that helps boost the network performance. This motivates a systematic study on Lorentz symmetry preservation.

The design of the study adopts the paradigm as follows. First, a baseline network is adopted without emphasizing the Lorentz symmetries or the subsymmetries corresponding to the four transformations in Fig. 4.1. Building upon this baseline, the study focuses on only a specific part of the network, a “subnetwork”, which is invariant under specific types of Lorentz subsymmetries or full symmetry. This paradigm should cover both means to preserve symmetries as introduced above, either using symmetry-invariant input or realizing the invariance by introducing decimated network modules. The outcome is consistent: a part of the network, along with all its neurons, remains invariant under some types of transformations^①.

Three baseline networks are considered in this study: the PFN, ParticleNet, and a “weakened” LorentzNet model named LorentzNet_{base}. LorentzNet_{base} supplements the scalar features by full feature list as mentioned above while modifying the Lorentz-invariant edge features (in Eq. (4.4)) as

$$\tilde{m}_{ij}^l = \phi_e(h_i^l, h_j^l, P), \quad (4.9)$$

where P is the placeholder to insert either $\psi(x_i^l \cdot x_j^l)$ to preserve full Lorentz symmetry, or other variables preserving certain subsymmetries.

These baselines are incorporated with two network-modifying recipes to incorporate different types of symmetries, which are detailed below.

① Note that, for the latter case, it is possible to incorporate symmetric design using an equivariant module. The studies that embed Lorentz equivariance include, e.g. Refs.^[99, 103], but they are often specialized in designs and hard to generalize to a broad paradigm. This is not discussed in this section.

Table 4.1 Invariance property of the pairwise variables between particles when the jet undergoes a certain type of Lorentz transformation.

Pairwise variable	$z-t$ boost	$x-y$ rotation	$y-z$ rotation ($y_{y,z} \sim o(1)$)	$x-t$ boost ($y_{y,z} \sim o(1)$)	z tilt ($y_{y,z} \sim o(1)$)	y tilt ($y_{y,z} \sim o(1)$)
m_{ij}^2	✓	✓	✓	✓	✓	✓
ΔR_{ij}	✓	✓	✓			
$\Delta R_{ij}(p_{T,i} + p_{T,j})$	✓	✓	✓	✓		
E_{ij} (ablation study)		✓	✓			

Incorporating pairwise features

The first study incorporates different pairwise features into ParticleNet or LorentzNet_{base}. For pairwise features, four of them are considered that preserve certain symmetry or subsymmetries, as detailed in Table 4.1. They are elaborated below.

1. **Pairwise mass.** $m_{ij} = \sqrt{p_i \cdot p_j}$ is known to be invariant under full Lorentz transformations.
2. **Pairwise ΔR .** $\Delta R_{ij} = \sqrt{(\Delta\eta_{ij})^2 + (\Delta\phi_{ij})^2}$ is the angular separation of particles in $\eta-\phi$ space. Considering the Lorentz transformations defined in the basis in Fig. 4.1, one sees that it remains invariant under $\eta-\phi$ rotation, which means that ΔR_{ij} is approximately invariant under $y-z$ rotation when the jet direction points to the x axis.
3. **Pairwise p_T -weighted ΔR .** $\Delta R_{ij}(p_{T,i} + p_{T,j})$ is constructed on the basis of ΔR_{ij} , which preserves one more type of symmetry approximately, under the transformation of $x-t$ boost. From Figs. 4.1 (a) and (c), one sees that an $x-t$ boost in the positive direction reduces the distance between particles while raising the p_T . It can be proven that they are in inverse proportion in the limit of $y_{y,z} \sim o(1)$ for all particles.
4. **Pairwise energy.** To design the ablation experiment, the pairwise energy variables $E_{ij} = E_i + E_j$ are chosen that, in general, violate the Lorentz symmetry. They do not obey the two basic symmetries, i.e., the $z-t$ boost and the $x-y$ rotation.

These pairwise variables are inserted into the ParticleNet and LorentzNet_{base} baselines using the following network-modifying recipes.

1. **Modifying ParticleNet.** The EdgeConv^[91] layers at the heart of ParticleNet include an intrinsic message-passing mechanism for each node to communicate with their k nearest neighbouring particles. Specifically, for each node i carrying features $\mathbf{x}_i$, consider

its neighboring nodes i_j ($j = 1, \dots, k$), the message $(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k})$ is passed to target node i . The designed pairwise features between node i and i_j can, therefore, be incorporated along with existing message-passing. First, the $N(N-1)/2$ pairwise features are calculated, and each of them is embedded with

$$\begin{aligned} & \text{Linear}(1, 64) \rightarrow \text{ReLU} \rightarrow \text{BatchNorm1D}(64) \\ & \left[\rightarrow \text{Linear}(64, 64) \rightarrow \text{ReLU} \rightarrow \text{BatchNorm1D}(64) \right]_{\times 2} \\ & \rightarrow \text{Linear}(64, 64), \end{aligned} \quad (4.10)$$

resulting to $\mathbf{U}_{ij}$. The new message conveyed from neighbouring nodes can be constructed by $(\mathbf{U}_{i,i_1}, \dots, \mathbf{U}_{i,i_k})$. The feature vectors of $\mathbf{U}_{ij}$ are directly added to the original message, after embedded by an additional layer $\text{Linear}(64, 64)$, written as

$$\mathbf{x}'_{i_j} = \mathbf{x}_{i_j} + f_{\text{Linear}(64, 64)}(\mathbf{U}_{i,i_j}) \quad (j = 1, \dots, k). \quad (4.11)$$

Figure 4.4 provides a schematic explanation of this recipe.

2. **Modifying LorentzNet_{base}**. The pairwise features are inserted simply into P in Eq. (4.9). It means that the intrinsic mechanism of LorentzNet is used to incorporate customized pairwise features.

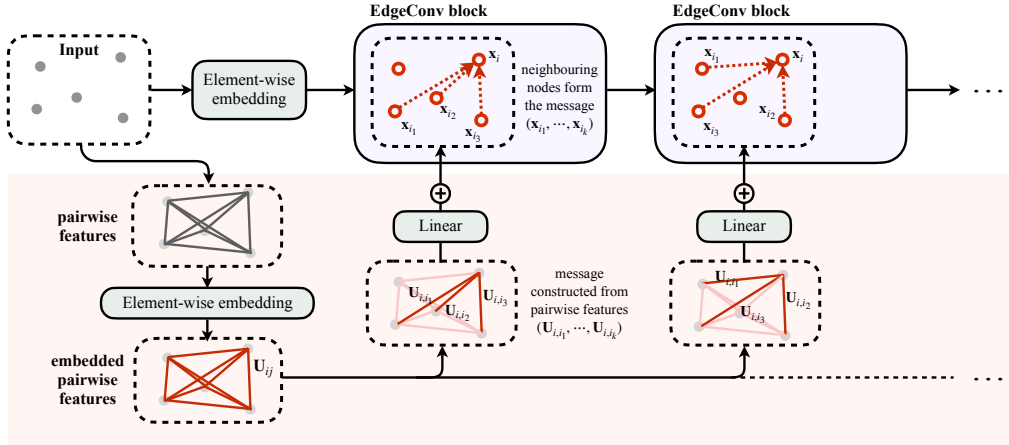


Figure 4.4 A schematic diagram of the modified ParticleNet architecture to incorporate the pairwise features. The modification is drawn in a red background to be distinguished from the original ParticleNet structure.

Experiments

Experiments are delivered for both ParticleNet and LorentzNet_{base} baselines, with different pairwise features exploited. They are performed on both Top and JETCLASS datasets, with the training data size varying across a broad range, from 6 k to 100 M.

The default experiment is performed with the 60 k training sample on the Top dataset. By evaluating the performance in terms of accuracy, AUC, and background rejection rates, shown in Table 4.2, it can be seen that incorporating variables invariant under more types of symmetries boosts network performance. This is the trend as observed from using ΔR_{ij} to $\Delta R_{ij}(p_{T,i} + p_{T,j})$, and eventually to m_{ij} , when removing the effect from the fluctuation. The trend is clearer when the training data size shrinks, as shown in Fig. 4.5 on the top plots. This proves that incorporating the Lorentz symmetries should serve as an inductive bias to the task because it especially helps to improve performance in data-constraint conditions, effectively augmenting input data under the case. Figure 4.5 also shows a trend of performance saturation when data size reaches the full size of 1.2 M. This is found to be an issue of the Top dataset as the high-performing network all reaches near-saturated performance. Hence, studies are also performed on the JETCLASS dataset with a more complex 10-class classification task and larger data size, as shown in Fig. 4.5 on the bottom plot. The conclusion still holds with less saturation found. A detailed analysis of the results can be found in Ref.^[104].

Table 4.2 Comparison of network performance of the ParticleNet and LorentzNet_{base} variants which incorporate different pairwise features. For each metric, the best-performing network from each baseline type is highlighted in bold text.

Base model	Variation	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 50\%$)	$1/\epsilon_B$ ($\epsilon_S = 30\%$)
ParticleNet	—	0.9310(3)	0.9810(2)	198 ± 7	640 ± 29
	+pairwise: m_{ij}	0.9334(8)	0.9820(4)	222 ± 13	722 ± 52
	+pairwise: ΔR_{ij}	0.9334(6)	0.9823(3)	231 ± 10	752 ± 43
	+pairwise: $\Delta R_{ij}(p_{T,i} + p_{T,j})$	0.9337(3)	0.9821(1)	223 ± 6	741 ± 36
	+pairwise: E_{ij}	0.9303(5)	0.9807(2)	200 ± 6	651 ± 23
LorentzNet _{base}	—	0.9276(12)	0.9789(7)	172 ± 13	581 ± 53
	+pairwise: m_{ij}	0.9347(4)	0.9829(2)	260 ± 6	931 ± 50
	+pairwise: ΔR_{ij}	0.9328(4)	0.9819(3)	232 ± 10	807 ± 35
	+pairwise: $\Delta R_{ij}(p_{T,i} + p_{T,j})$	0.9342(4)	0.9826(2)	251 ± 6	919 ± 34
	+pairwise: E_{ij}	0.9243(37)	0.9767(23)	144 ± 29	485 ± 108

Another way to demonstrate the role Lorentz symmetry preservation plays is to study the

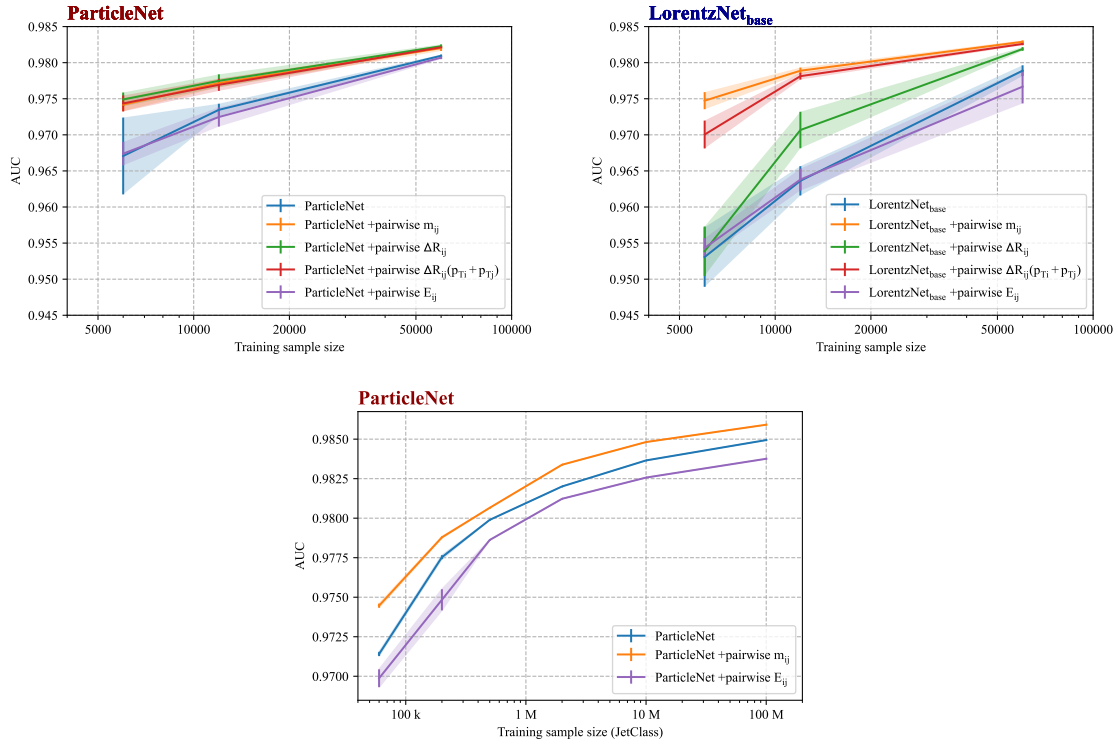


Figure 4.5 Top: Comparison of network performance on Top dataset in terms of AUC among different training data size, selected among 6 k, 20 k and 60 k, for the baseline network as ParticleNet (left) and LorentzNet_{base} (right). Bottom: Extended comparison delivered on JETCLASS dataset with larger data size, selected among 60 k, 200 k, 500 k, 2 M, 20 M and 100 M.

robustness of the network with respect to the certain transformation of inputs, similar to the study delivered in LorentzNet in Fig. 4.3. Figure 4.6 illustrates the robustness when input jets undergo y - z rotation, x - t boost, and the z tilt, namely the procedures depicted in Fig. 4.1 (b), (c), and (d). The expected conclusion can be drawn, i.e. when incorporating features that are invariant under certain symmetries, the entire network manifests robustness under the corresponding transformations. Note that the modification of the network does not guarantee that it fully adheres to that symmetry. Hence, upon its learning, the invariant properties manifest on a “subnetwork” are somehow reflected in the entire network, and this illustrates the important role that symmetry-preserving variables play in shaping the network’s behaviour.

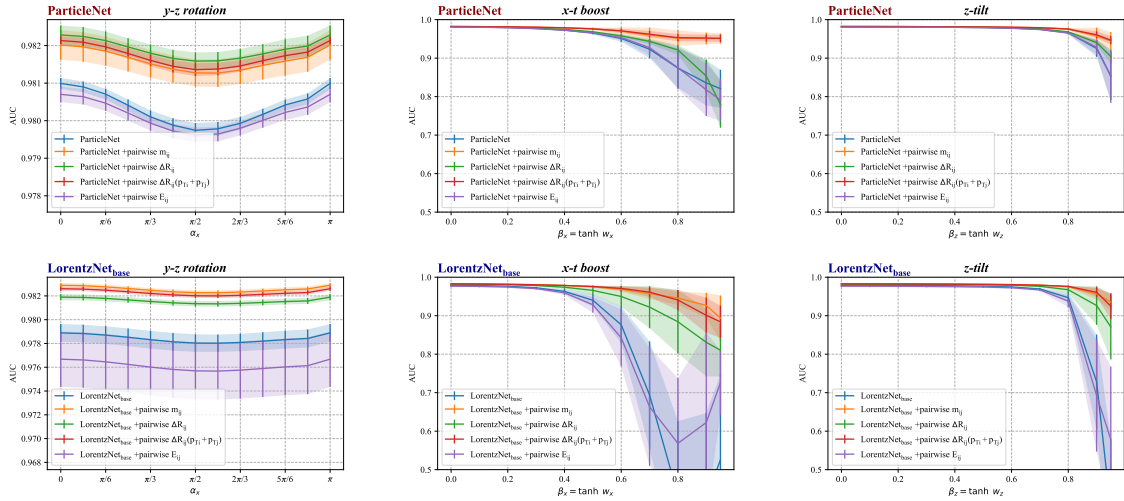


Figure 4.6 Comparison of network performance in terms of AUC evaluated under various types of Lorentz transformation applied to the test dataset. They correspond to the input jet undergoing a y - z rotation with angle α_x (left), an x - t boost with rapidity w_x (middle), and a z -tilt transformation, i.e., a z - t boost with rapidity w_z followed by a x - z rotation to restore the jet direction (right). The baseline is chosen as ParticleNet (top) or LorentzNet_{base} (bottom).

Incorporating nodewise features

To ensure that the modification fits a more generic scenario, a second network-modifying recipe is designed to incorporate nodewise features that every set-based or graph-based jet network should possess. The recipe is outlined in Fig. 4.7. The recipe reads that the nodewise features, which are specially designed to incorporate symmetries, are calculated at the beginning, embedded, and then integrated by “summation” to the latent space features layer by layer. This works for both PFN, ParticleNet, and LorentzNet_{base} as the baseline: for PFN, the unit block is $\Phi(x)^{[87]}$, the elementwise MLP for each particle; For ParticleNet, it is the EdgeConv

operation^[90]; for LorentzNet, it is the LGEB^[94].

On the nodewise feature design, consider the following mass-related variable as a Lorentz invariance. The invariant mass of a particle i with its selected neighbours, G_i , also chosen from a Lorentz-invariant way, can be calculated as

$$m_{G_i} = \sqrt{\left(\sum_{j \in G_i} p_j\right)^2}, \quad (4.12)$$

where $G_i = \{j | p_i \cdot p_j \text{ is among the } k \text{ largest values for all } j\}$.

As an ablation study, the sum of energies of the selected particle is calculated, resulting in E_{G_i} . Table 4.3 summarizes the performance for each baseline network and its variant. It again demonstrates that incorporating Lorentz-symmetry-preserving features improves the network performance. Additional checks similar to the pairwise features studies are delivered and can be found in Ref.^[104].

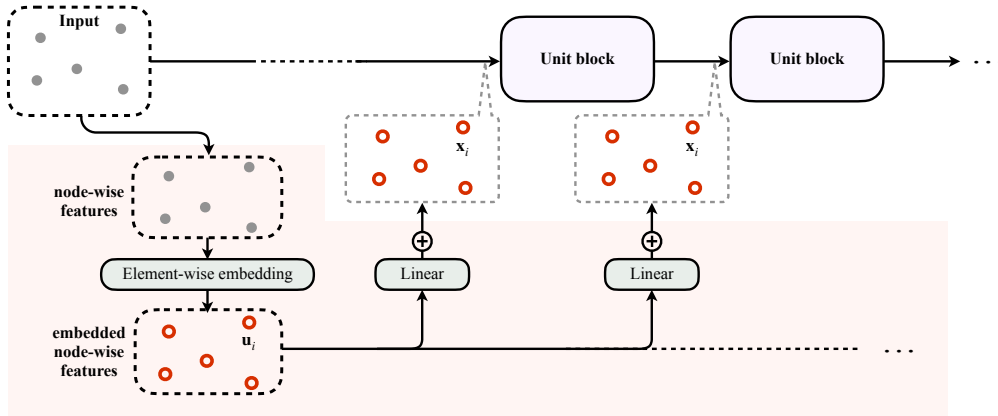


Figure 4.7 A schematic diagram of the universal network-modifying recipe for all baseline networks to incorporate additional nodewise features. The modification is drawn in a red background to be distinguished from the original network structure.

This section will end with a brief outlook. If one reviews the development of the jet tagging network in a broad picture, one sees that the community progressively move from the image or sequence representation to the recent set or graph representation. This indicates a common sense to adhere to the jet's permutational invariance within the network design. Similarly, recent years have witnessed the birth of high-performing networks to incorporate jet's Lorentz symmetry, such as LorentzNet discussed in this dissertation or PELICAN proposed

Table 4.3 Comparison of network performance of the PFN, ParticleNet and LorentzNet_{base} variants incorporating different nodewise features. For each metric, the best-performing network from each baseline type is highlighted in bold text.

Base model	Variation	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 50\%$)	$1/\epsilon_B$ ($\epsilon_S = 30\%$)
PFN	—	0.9104(12)	0.9664(13)	67 ± 5	198 ± 21
	+nodewise: m_{G_i}	0.9281(4)	0.9791(2)	184 ± 5	714 ± 50
	+nodewise: E_{G_i}	0.9207(4)	0.9750(3)	125 ± 3	378 ± 19
ParticleNet	—	0.9310(3)	0.9810(2)	198 ± 7	640 ± 29
	+nodewise: m_{G_i}	0.9313(3)	0.9812(1)	222 ± 5	800 ± 40
	+nodewise: E_{G_i}	0.9300(12)	0.9802(6)	183 ± 12	572 ± 47
LorentzNet _{base}	—	0.9276(12)	0.9789(7)	172 ± 13	581 ± 53
	+nodewise: m_{G_i}	0.9306(3)	0.9809(2)	219 ± 3	887 ± 36
	+nodewise: E_{G_i}	0.9272(3)	0.9788(1)	171 ± 2	562 ± 16

recently^[95]; meanwhile, several works have emphasised certain types of Lorentz subsymmetries. Hence, it may again indicate an awareness of design networks that adhere to *full* Lorentz symmetry for future DNN developments. Studies introduced in this section hope to bring this awareness to a broader community.

4.2 The Transformer algorithm

As introduced in Section 2.2.1, the Transformer algorithm is regarded as a milestone in the DNN algorithm development. It has been proven to be a universal architecture suitable to many data domains. In a broad view, Transformer efficiently captures relations among its tokens, especially long-range dependencies, through the self-attention mechanism depicted in Fig. 2.4. It can be seen that the mechanism connects each token with every other token and treats them equivalently for feature extraction. For its application in CV and NLP, this is unlike the CNNs, where the convolutional filter only acts on a local patch of images, or the RNNs where tokens are processed subsequently so the features get diluted over distance.

This property of self-attention is suitable in jet physics by concept. First, the advanced GNN algorithms developed for jet tagging already target to connect each particle with others to deliver message passing; the self-attention naturally allows its connection—it can be considered to process particle-format data within a fully connected graph, but the “message

Table 4.4 Performance of ParticleNet and ParT trained with different data sizes of JETCLASS. $\text{Rej}_{X\%}$ refers to $1/\epsilon_{q/g}$ when $\epsilon_S = X\%$.

	All classes		$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow gg$	$H \rightarrow 4q$	$H \rightarrow \ell\nu qq'$	$t \rightarrow bqq'$	$t \rightarrow b\ell\nu$	$W \rightarrow qq'$	$Z \rightarrow q\bar{q}$
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
ParticleNet (2 M)	0.828	0.9820	5540	1681	90	662	1654	4049	4673	260	215
ParticleNet (10 M)	0.837	0.9837	5848	2070	96	770	2350	5495	6803	307	253
ParticleNet (100 M)	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT (2 M)	0.836	0.9834	5587	1982	93	761	1609	6061	4474	307	236
ParT (10 M)	0.850	0.9860	8734	3040	110	1274	3257	12579	8969	431	324
ParT (100 M)	0.861	0.9877	10638	4149	123	1864	5479	32787	15873	543	402

passing” is efficiently done via the dot-product operation. Second, it is known that image or textual data have predefined orders in their tokens, so their initial token features should be pre-processed to encode their position. This can be simply omitted in processing particle-format tokens, because they are inherently permutational invariant.

Deep learning experiences suggest that Transformer often stands out when the training data size is massive. The attention module contains more parameters. Therefore, when comparing the Transformer with other architectures, the Transformer is usually more compact in similar computation cost. This requires a larger data size to fit the parameter size. Early Transformer studies, including the Point Cloud Transformer^[105] experiment with the Top and QG benchmark dataset (with $\sim 1\text{--}2$ M training samples) and are found not to achieve superior performance. On the contrary, in the JETCLASS dataset with 100 M training sample, it occurs that Transformer models (ParT) yield substantial improvement and are found to scale better with the training size. For introductory purposes, Table 4.4 compares the performance of ParT and ParticleNet performance when training on the 2 M, 10 M and 100 M samples of the JETCLASS dataset. It is clear that ParT showcases a larger potential for improvements as the data size rises, while ParticleNet is closer to saturation. As the experimental development of jet taggers is generally trained on larger datasets, reaching a hundred million, Transformer models are believed to be very suitable for future experimental applications.

4.2.1 Particle Transformer

This section will highlight the ParT model, a typical application of the Transformer architecture into jet physics, but incorporating pairwise features that bring inductive bias relevant to Lorentz symmetry and its subsymmetries.

Particle Transformer outlines A schematic diagram of the ParT architecture is shown in Fig. 4.8 (a). The basic structure of ParT is to stack $L = 8$ layers of Particle Attention Block and 2 layers of Class Attention Block. The commonality to a conventional Transformer outline is described below, followed by the specificity of the Particle Attention Block design, which incorporates pairwise features.

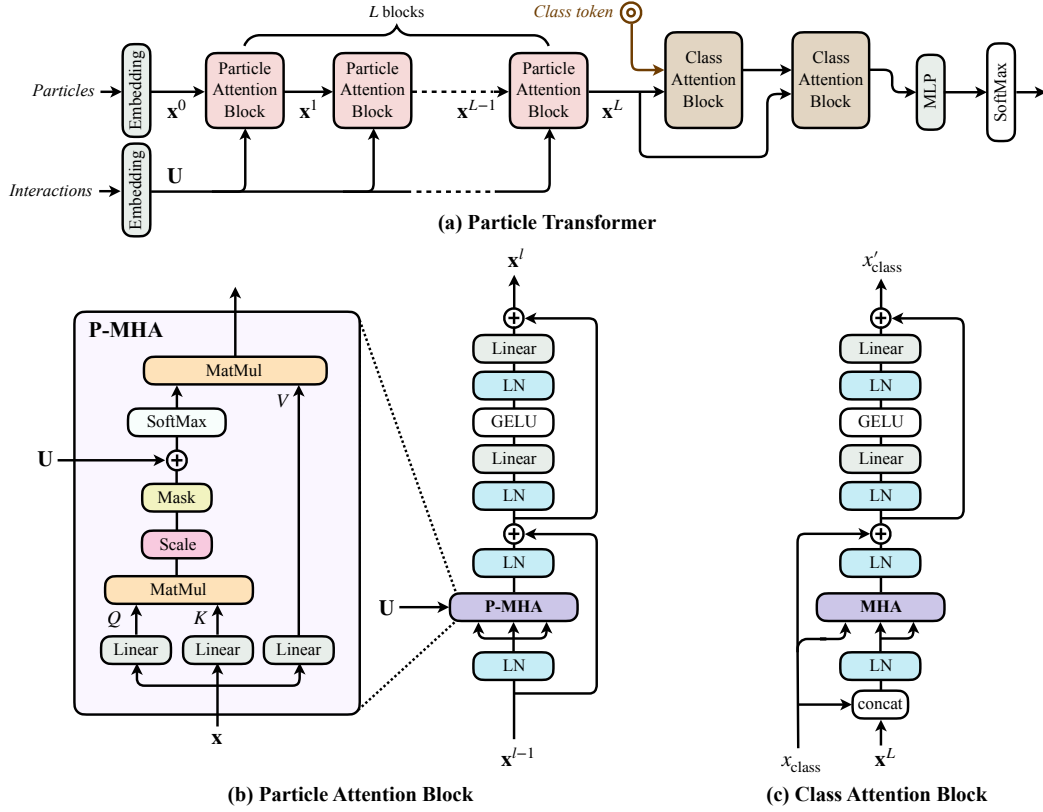


Figure 4.8 A schematic diagram of the Particle Transformer architecture (a), the Particle Attention Block (b) and Class Attention Block (c).

Prior to the attention blocks, the input feature is embedded from an initial dimension of input_dim to the attention embedding dimension $d = 128$, by

$$\begin{aligned}
 & \text{LayerNorm}(\text{input_dim}) \rightarrow \text{Linear}(\text{input_dim}, 128) \rightarrow \text{GELU} \\
 & \rightarrow \text{LayerNorm}(128) \rightarrow \text{Linear}(128, 512) \rightarrow \text{GELU} \\
 & \rightarrow \text{LayerNorm}(512) \rightarrow \text{Linear}(512, 128) \rightarrow \text{GELU}.
 \end{aligned} \tag{4.13}$$

When processing through Particle Attention blocks, the attention is performed on tokens with a particle basis, as illustrated in Fig. 2.4, with an embedding dimension $d = 128$ and the number

of attention heads $n_{\text{head}} = 8$. Within each block, the post-attention FFN is the form of

$$\text{Linear}(128, 512) \rightarrow \text{GELU} \rightarrow \text{LayerNorm}(512) \rightarrow \text{Linear}(512, 128). \quad (4.14)$$

After passing through L Particle Attention Blocks, with the particle tokens sufficiently learned, two Class Attention Blocks are followed to compute the attention between a specifically introduced class token and all particle tokens to update the class tokens. The processed class token is then converted to output logits with a single-layer MLP. The principle here is close to the Transformer’s convention that feature aggregation is not done using global pooling but by introducing the class token. The mechanism to bring the class token in the last few layers is inspired by the CaiT approach^[106].

Particle Attention Block The Particle Attention Block distinguishes ParT from the conventional Transformer by incorporating pairwise features within attention. As introduced in Section 4.1.3, and especially indicated in Table 4.1, pairwise masses provide a basis of all Lorentz scalars within a jet. Incorporating the features within network design may bring special bias on Lorentz-symmetry-preservation to the network and improve its performance. This is also found to be the case for this study. The pairwise features to be incorporated are in the dimension of (N, N, d_{pair}) , where N is the particle length and d_{pair} is the feature dimension, controllable by its embedding network. Notice that, the attention map obtained in the middle of the attention, with every particle has calculated a dot-product attention per head with other particles, is in the dimension of $(N, N, n_{\text{head}} = 8)$. Therefore, the embedded pairwise features, if configured as $d_{\text{pair}} = 8$, can be added straightforwardly to the attention map. This provides an “attentive bias” in the attention mechanism.

Concretely, the attention mechanism for each head, in analogy to Eq. (2.4), can be written as

$$\text{P-MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{U}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{U}\right)\mathbf{V}, \quad (4.15)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ matrices are in dimension (N, d) , and the additional $\mathbf{U}$ matrix is in dimension (N, N) , i.e. a slice of the embedded pairwise feature matrix for a single head.

The pairwise features have the initial dimension 4: in addition to the pairwise mass features, it also includes three other features that are also invariant under certain subsymmetries.

For the particle pair (i, j) , they are defined as

$$\begin{aligned}
 \Delta R_{ij} &= \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}, \\
 k_{T,ij} &= \min(p_{T,i}, p_{T,j}) \Delta R_{ij}, \\
 z_{ij} &= \min(p_{T,i}, p_{T,j}) / (p_{T,i} + p_{T,j}), \\
 m_{ij}^2 &= (p_i + p_j)^2 \quad (p_{i,j} \text{ are the four-momentum}).
 \end{aligned} \tag{4.16}$$

They are embedded to dimension $n_{\text{head}} = 8$ as

$$\begin{aligned}
 &\text{Linear}(4, 64) \rightarrow \text{BatchNorm1D}(64) \rightarrow \text{GeLU} \\
 &\left[\rightarrow \text{Linear}(64, 64) \rightarrow \text{BatchNorm1D}(64) \rightarrow \text{GeLU} \right]_{\times 2}.
 \end{aligned} \tag{4.17}$$

This explains the full structure of the Particle Attention Block.

Experiments The ParT is experimented in the Top, QG, and JETCLASS datasets. For the QG dataset, it adopts the QG_{exp} scenario. Input features for each experiment are summarized in Table 2.1.

The training of ParT adopts the LOOKAHEAD optimizer^[101] with $k = 6$ and $\alpha = 0.5$ and the inner optimizer is RADAM^[102] with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-5}$. For training on the Top and QG dataset, the total epoch is 20, each defined as iterating over the full training sample; for training on the JETCLASS 100 M dataset, the total epoch is 50, each defined as iterating over 10 M samples. A weight decay of 0.01 is configured for training on the Top and QG datasets, but not on the JETCLASS dataset. The batch size of 512 and initial LR of 1×10^{-3} is adopted. The LR starts to decay exponentially when finishing 70% of the epochs, changing at the beginning of every epochs and drop to 1% of its initial rate in the last epoch.

To study the effect of adding pairwise features as attentive bias, the plain Transformer is also trained with the same configuration without adding the $\mathbf{U}$ matrix in the Eq. (4.15). The model is noted as ParT (plain). The jet tagging performance on Top and QG datasets are shown in Tables 4.5 and 4.6, in terms of accuracy, AUC, and background rejections at fixed signal efficiencies. The performance on JETCLASS is also shown in Table 4.7. The ParticleNet and LorentzNet performances are shown in parallel.

Table 4.8 presents a detailed comparison of the model sizes and computational complexities of various algorithms, measured in terms of their number of parameters and floating-point operations per second (FLOPs).

ParT achieves the highest performance while maintaining a relatively modest computational cost. Specifically, while ParT and ParticleNet have comparable FLOPs, ParT requires

Table 4.5 Comparison of jet tagging performance on the top tagging dataset with multiple algorithms. The best-performing model is highlighted in bold text.

	Accuracy	AUC	Rej _{50%}	Rej _{30%}
P-CNN ^[90]	0.930	0.9803	201 $\pm$ 4	759 $\pm$ 24
PFN ^[87]	—	0.9819	247 $\pm$ 3	888 $\pm$ 17
ParticleNet ^[87]	0.940	0.9858	397 $\pm$ 7	1615 $\pm$ 93
JEDI-net (w/ $\sum O$) ^[89]	0.930	0.9807	—	774.6
PCT ^[105]	0.940	0.9855	392 $\pm$ 7	1533 $\pm$ 101
LGN ^[99]	0.929	0.964	—	435 $\pm$ 95
rPCN ^[97]	—	0.9845	364 $\pm$ 9	1642 $\pm$ 93
LorentzNet ^[94]	0.942	0.9868	498 $\pm$ 18	2195 $\pm$ 173
ParT ^[65]	0.940	0.9858	413 $\pm$ 16	1602 $\pm$ 81

six times more parameters than ParticleNet. This highlights the efficiency of Transformer-based architectures compared to GNN-based models. A comparison between LorentzNet and ParticleNet further shows notable differences. Despite they have similar parameter counts, LorentzNet has much higher computational costs. This difference arises from LorentzNet’s reliance on a fully connected GNN architecture, which imposes greater computational overhead during message-passing operations compared to the k -nearest-neighbour-based message-passing mechanism employed in ParticleNet.

Table 4.6 Comparison of jet tagging performance on the quark–gluon tagging dataset with multiple algorithms. The subscript “exp” and “full” distinguish models using partial or full particle identification information. The best-performing model is highlighted in bold text.

	Accuracy	AUC	Rej _{50%}	Rej _{30%}
P-CNN _{exp} ^[90]	0.827	0.9002	34.7	91.0
PFN _{exp} ^[87]	—	0.9005	34.7 ± 0.4	—
ParticleNet _{exp} ^[90]	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3
rPCN _{exp} ^[97]	—	0.9081	38.6 ± 0.5	—
ParT _{exp} ^[65]	0.840	0.9121	41.3 ± 0.3	101.2 ± 1.1
PFN _{full} ^[87]	—	0.9052	37.4 ± 0.7	—
ABCNet _{full} ^[93]	0.840	0.9126	42.6 ± 0.4	118.4 ± 1.5
PCT _{full} ^[105]	0.841	0.9140	43.2 ± 0.7	118.0 ± 2.2
LorentzNet _{full} ^[94]	0.844	0.9156	42.4 ± 0.4	110.2 ± 1.3
ParT _{full} ^[65]	0.849	0.9203	47.9 ± 0.5	129.5 ± 0.9

Table 4.7 Comparison of jet tagging performance on the JETCLASS dataset with multiple algorithms, including LorentzNet. The best-performing model is highlighted in bold text. Rej_{X%} refers to $1/\epsilon_{q/g}$ when $\epsilon_S = X\%$.

	All classes		$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow gg$	$H \rightarrow 4q$	$H \rightarrow \ell\nu qq'$	$t \rightarrow bqq'$	$t \rightarrow b\ell\nu$	$W \rightarrow qq'$	$Z \rightarrow q\bar{q}$
	Accuracy	AUC	Rej _{50%}	Rej _{50%}	Rej _{50%}	Rej _{50%}	Rej _{99%}	Rej _{50%}	Rej _{99.5%}	Rej _{50%}	Rej _{50%}
PFN ^[87]	0.772	0.9714	2924	841	75	198	265	797	721	189	159
P-CNN ^[70]	0.809	0.9789	4890	1276	88	474	947	2907	2304	241	204
ParticleNet ^[90]	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
LorentzNet	0.855	0.9869	9217	3425	117	1550	4425	19807	12500	480	353
ParT (plain) ^[65]	0.849	0.9859	9569	2911	112	1185	3868	17699	12987	384	311
ParT ^[65]	0.861	0.9877	10638	4149	123	1864	5479	32787	15873	543	402

Table 4.8 Number of trainable parameters and floating-point operations per second (FLOPs).

	Accuracy	# params	FLOPs
PFN	0.772	86.1 k	4.62 M
P-CNN	0.809	354 k	15.5 M
ParticleNet	0.844	370 k	540 M
LorentzNet	0.855	233 k	2.01 G
ParT (plain)	0.849	2.13 M	260 M
ParT	0.861	2.14 M	340 M

4.2.2 Discussion on Transformer models

In deep learning, the Transformer architecture exhibits scaling laws, consistently improving performance as data size and model capacity increase. As shown in Table 4.4, JETCLASS, the first dataset with 100 M samples, enables Transformers to outperform GNN-based models such as ParticleNet significantly. This result provides partial evidence supporting scaling laws.

Although Transformers are highly efficient models, their computational complexity remains quadratic, which motivates studies to deliver further optimization. For potential deployment in online computing environments for high-energy physics, improving the efficiency of Transformer models is an important focus.

Based on these considerations, we outline three potential directions for the future development of ParT or Transformer-based models in the particle physics domain.

Investigating the effects of scaling up ParT on a 100 M dataset is crucial to understanding its performance limits at this data size. In addition, whether increased data continues to improve performance is yet another question to understand. This serves as one of the motivations for the study presented in Part V.

Given the substantial capacity of Transformer models to retain large amounts of knowledge, ParT offers the potential to handle increasingly complex tasks. In particle physics, this could enable the development of models to handle more complex tasks or even multimodal models, providing another important motivation explored in Part V.

Improving the efficiency of ParT is an essential objective. On one front, ParT could evolve into a more efficient Transformer by adopting optimization strategies commonly employed in the deep learning community, such as various forms of efficient Transformers. On another front, it is crucial to address that the pairwise feature computations still inherently have quadratic complexity in the number of tokens. These pairwise features and subsequent embeddings are key to ParT’s superior performance, as they enhance the model’s inductive bias for Lorentz symmetry. However, pairwise embedding computations constitute around 20% of ParT’s test-time computing cost, which is a large amount. Therefore, identifying alternative methods to integrate Lorentz symmetry while circumventing the quadratic complexity of pairwise computations will be a promising research direction. This enables the building of more efficient Transformer models dedicated to the domain of particle physics.

Part III

Calibration of Jet Taggers

Chapter 5

Introduction to Boosted-jet Flavour Tagger Calibration

When a jet tagging algorithm is developed and prepared for use in an LHC search or measurement, calibration is a crucial step. The calibration aims to correct the selection efficiency in simulated events to that measured on data. The efficiencies between simulation and data are not necessarily the same because the jet algorithms may utilize higher-level patterns of a jet, which may not be well described by the phenomenological model within the generator used to simulate the jet. ML-based algorithms are trained on simulated jets, so they are more likely to capture special details that only appear in simulated jets but not in real jets. This can enlarge the discrepancy between simulation and data performance. The situation becomes more severe in deep learning algorithms when the network becomes deeper and more convoluted, posing a challenge to using these algorithms in analyses.

This part will focus on the calibration specifics of the algorithms. Specifically, our focus will be the boosted-jet flavour tagging that aims to identify $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ signature (X refers to a heavy resonant particle), which one has witnessed plentiful applications in analyses related to Higgs and BSM physics. In this chapter, the boosted-jet flavour tagging algorithms and their past calibration methods in the CMS and ATLAS experiments are reviewed, emphasizing the progress made and potential limitations. Then, in Chapter 6, a novel calibration method developed within the CMS experiment is introduced. The new method is applicable to calibrate highly performant $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging algorithms and has results in many applications.

5.1 Overview of boosted-jet flavour taggers

The boosted-jet flavour tagging algorithms aim to identify a type of large- R jets initiated from a highly Lorentz-boosted heavy resonance decaying to heavy-flavour quarks such as $b\bar{b}$ or $c\bar{c}$. They are crucial to several physics measurements, including searching for the SM Higgs bosons in their $b\bar{b}$ and $c\bar{c}$ decay modes^[107-114] or a BSM resonance decaying to such final

states^[115].

The essential characteristic of a large- R jet initiated from $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ is that it is likely to have a two-prong structure, and each prong manifests a character of b or c jet, with displacement in tracks or the existence of SV. This is illustrated in the cartoon in Fig. 5.1, along with many other topologies in the broad realms of large- R jet tagging.

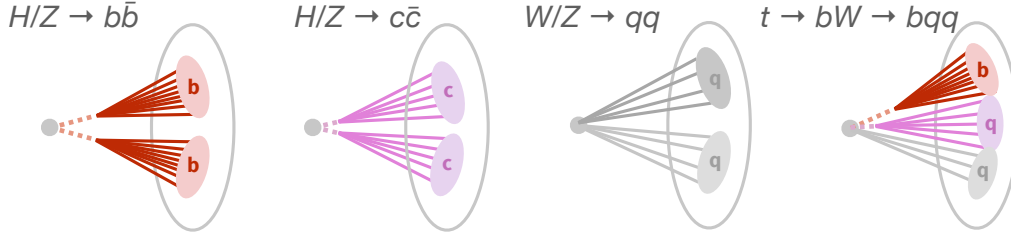


Figure 5.1 Illustration on large- R jet topologies with different origins. The first two cases fall under the category of boosted-jet flavour tagging.

In this regard, an initial attempt to tag $X \rightarrow b\bar{b}$ jet requires two or more representative subobjects within the large- R jet with special characteristics caused by a b quark. The CMS experiment explores the double- b tagger^[75], which exploits two SVs within the jet and the features from the leading tracks in a multivariate approach, and ATLAS developed the double b -tagging algorithm^[116] that matches the jet with variable-radius track-jets and imposes b -tagged requirements on each track-jet.

It is possible that the final state particles from the two prongs may overlap, resulting in a less satisfied separation of these subobjects as well as the b -tagging performance for each subobject. In this regard, the advanced deep learning method consumes the entire constituent list within the large- R jet and lets the neural network learn the inherited correlation between constituents. This was found helpful in the initial CMS efforts, which exploits the advanced CNN in the work of DeepAK8-MD tagger^[70] to process the list of constituents. Meanwhile, a list of SVs is also input to the network as they are crucial for heavy flavour tagging. The architecture has undergone several upgrades in the subsequent efforts to improve the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging performance. These include the DeepDoubleX^[85, 86] (with CNN and RNN backbone) and the ParticleNet^[117] (with graph convolutional network backbone) algorithm, with a detailed description in the following. The ATLAS experiment also explores advanced neural network approaches to deal with whole constituent inputs, including GN2X^[118] (with Transformer backbone).

An important property for all present $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ taggers is its decorrelation

with the jet mass. In their application to Higgs or BSM analysis, the signal process should characterize a resonance (the Higgs boson or a BSM particle) with a specific mass, and the signal extraction is often performed on a distribution of the reconstructed mass. To enable that a selection on the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagger discriminant does not significantly change the background distribution or even fabricate a fake peak on the spectrum, the discriminant should be decorrelated with the mass variable. This property also helps to extend the tagger's usage to BSM physics such that the tagger is sensitive to identify $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ signature with a broad range of m_X . Specific techniques are considered to achieve good mass-decorrelation performance, whether it be choosing variables that do not have a strong dependence on jet mass, using a dedicated training strategy to force this decorrelation, or using input jets that span over a wide mass range and applying a reweighting to enable an identical distribution across the categories.

A short technical summary of the recent $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ taggers developed in CMS is presented below. These taggers also serve for the calibration targets in Chapter 6.

The double-b tagger

The double-b tagger^[75] is developed at CMS in the early Run 2 era. The tagger is trained with a BDT to distinguish $H \rightarrow b\bar{b}$ jets as the signal and QCD jets as the background. The input to the tagger includes constructed high-level variables of leading tracks and secondary vertices associated with the jet, which are chosen as not having a strong dependence on jet p_T and mass. The output BDT score is referred to as the double-b tagging discriminant. The tagger is applied to the early search of the boosted $H \rightarrow b\bar{b}$ signature^[107] within CMS.

DeepAK8-MD

The DeepAK8-MD algorithm^[70] is a DNN-based AK8 jet tagging algorithm developed at CMS in the early Run 2 era, for identifying the resonance decay to $b\bar{b}$ or $c\bar{c}$ flavour. It adopts the architecture constructed by one-dimensional convolutional layers based on the ResNet model^[119]. It is trained with input features of all constituent PF candidates within the jet and the SVs that are matched with the jet axis by $\Delta R < 0.8$.

The DeepAK8-MD algorithm is trained as a multiclass classifier. The output classes comprise five main categories (t , W , Z , H , and others), each of which is subdivided into minor categories, corresponding to the decay mode (e.g. $b\bar{b}$ or $c\bar{c}$ decay). The mass decorrelation is achieved via an adversarial training approach^[70]. Specifically, a mass prediction network

is added to the classification network, and the training target is also modified to include the accuracy of the mass prediction as a penalty. After training, The algorithm outputs probability scores largely independent of jet mass.

The discriminants to identify a resonance decay to $b\bar{b}$ or $c\bar{c}$ are defined by

$$\begin{aligned} \text{DeepAK8-MD bbvsQCD disc.} &= \frac{p(H \rightarrow b\bar{b}) + p(Z \rightarrow b\bar{b})}{p(H \rightarrow b\bar{b}) + p(Z \rightarrow b\bar{b}) + p(\text{QCD})}, \\ \text{DeepAK8-MD ccvsQCD disc.} &= \frac{p(H \rightarrow c\bar{c}) + p(Z \rightarrow c\bar{c})}{p(H \rightarrow c\bar{c}) + p(Z \rightarrow c\bar{c}) + p(\text{QCD})}. \end{aligned} \quad (5.1)$$

The $X \rightarrow b\bar{b}$ tagging discriminants are adopted by CMS searches, including Refs.^[112, 120].

DeepDoubleX

The DeepDoubleX tagging algorithm^[86] is a DNN-based approach for $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ AK8 jet identification in the boosted topology, trained with a combined architecture composed of one-dimensional convolutional layers and the gated recurrent units (GRU). The input to the algorithms includes three groups of low-level input features, i.e. the charged PFs, neutral PFs, and SVs, as well as a group of jet-level observables. The algorithm is updated from a previous version (denoted as V1)^[85], the latter adopted in the search of boosted Higgs boson decay to $b\bar{b}$ signature^[108] in CMS.

The mass decorrelation of the algorithm is achieved by using signal jets that originate from the Higgs boson with a varying mass in the training. This results in a similar flat distribution on m_{SD} for the signal jets and background QCD jets. A reweighting on m_{SD} further guarantees the perfection of mass decorrelation.

Different from ParticleNet-MD, the algorithm is trained in three individual binary jet classification tasks, i.e. for distinguishing $X \rightarrow b\bar{b}$ from QCD jets, $X \rightarrow c\bar{c}$ from QCD jets, and $X \rightarrow c\bar{c}$ from $X \rightarrow b\bar{b}$ jets. This results in three individual trained models, named DeepDoubleBvL (DDBvL), DeepDoubleCvL (DDCvL), and DeepDoubleCvB (DDCvB). The three DeepDoubleX discriminants are DDBvL, DDCvL, and DDCvB, produced from each model. The DDBvL and DDCvL discriminants are the targets for this calibration study in identifying $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ jets.

ParticleNet-MD

The ParticleNet-MD jet tagging algorithm^[117] provides two recommended discriminants, the “ParticleNet-MD bbvsQCD” and “ParticleNet-MD ccvsQCD”, used for $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ identification, respectively. The ParticleNet-MD algorithm is a DNN-based algorithm to

identify two-prong hadronic decays ($b\bar{b}$, $c\bar{c}$, and $q\bar{q}$) of a highly Lorentz-boosted particle in a wide range of resonance mass. The algorithm is trained with the “ParticleNet” neural network architecture^[90], as introduced in Section 3.3.2. The training uses a list of PF candidates and SVs associated with the jet as input, which is similar to the DeepAK8-MD training.

Note that, in the CMS experiment, taggers with the ParticleNet architecture adopted include both the mass-decorrelated (MD) version (i.e. the “ParticleNet-MD” algorithm) and the mass-correlated version (the “ParticleNet” algorithm)^[117]. The former targets tagging $H \rightarrow b\bar{b}$ ($c\bar{c}$) signatures in the mass decorrelated manner, and calibration study is aimed at the former one; the latter aims at identifying hadronic decays of Lorentz-boosted SM particles (t , W , Z , and H) and classifying different decay modes, which is not in the scope of this chapter.

The ParticleNet-MD algorithm is trained on a set of signal jets including $X \rightarrow b\bar{b}$, $X \rightarrow c\bar{c}$, $X \rightarrow q\bar{q}$, and background QCD jets, where X is a variable-mass spin-0 particle, with m_X ranging from 15 GeV to 250 GeV. Jets from both signal and background samples are reweighted to yield flat distributions in both p_T and m_{SD} so as to decorrelate the trained tagger variable with the jet mass. The output of the algorithm provides four probability-like scores: $p(X \rightarrow b\bar{b})$, $p(X \rightarrow c\bar{c})$, $p(X \rightarrow q\bar{q})$, and $p(\text{QCD})$, where the QCD class is divided into five finer categories in the classification, namely, QCD-bb, QCD-cc, QCD-b, QCD-c, and QCD-others. In applying this tagger to physics searches, the QCD jets are usually the major background to be vetoed against. Hence the discriminants to tag $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ jets are defined as binary classification scores:

$$\begin{aligned} \text{ParticleNet-MD bbvsQCD disc.} &= \frac{p(X \rightarrow b\bar{b})}{p(X \rightarrow b\bar{b}) + p(\text{QCD})}, \\ \text{ParticleNet-MD ccvsQCD disc.} &= \frac{p(X \rightarrow c\bar{c})}{p(X \rightarrow c\bar{c}) + p(\text{QCD})}. \end{aligned} \quad (5.2)$$

The $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants are adopted in the physics measurements at CMS, including Refs.^[109, 113, 115], and are the target discriminants for the performance study.

As a summary of the taggers introduced above, Fig. 5.2 shows the distribution of several $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants on the signal $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets, the QCD multijet background jets, and bb and cc components from the QCD jets. Here, QCD jets matched with at least two ghost b hadrons are designated as the bb component; jets with no ghost-matched b hadrons but include at least two ghost-matched c hadrons are designated as the cc component.

It can be seen that $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants show higher values in the corresponding $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ signals, demonstrating the ability to distinguish from the QCD jets and their subcomponents. However, the shape of each tagging discriminant can be different, depending on the training conditions.

5.2 Overview of tagger calibration

Calibration of the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ large- R jet tagging algorithm is a crucial task for their application in physics searches and measurements. Especially, when the algorithms involve deep and complex neural network structures, it is possible that the subtle difference between data and simulation may cause a substantial difference in the neural network output. The difference may come in the splitting of partons or jet hadronization, partly due to the perturbation theory and phenomenological model used in the simulation software, such as PYTHIA or HERWIG. Furthermore, a tight selection is likely to be imposed in physics analyses to achieve greater background rejection. Thus, the data-simulation discrepancy might be further amplified. These make the calibration of vital importance in the physics analyses.

Practically, the calibration is done by measuring the ratio of efficiencies to select the signal events from data and simulation. Here, the ratio is referred to as the scale factor (SF). The SF measurement often depends on the working point of the jet tagger selection.

To measure the SF, the general calibration strategy adopted in all present calibration methods is to select a dedicated phase space from the data events that have the same or similar characteristics to the signal jets. The case is simpler if the signal jet targeted in calibration can find a corresponding phase space on data. For example, for signal jets being the boosted hadronic top/ W jets, one can use the data region enriched in the top quark-antiquark pair ($t\bar{t}$) process to select jets induced from a top quark or W boson^[121, 122]. However, the challenge lies in the $H \rightarrow b\bar{b}$ ($c\bar{c}$) jet calibration because it is almost impossible to obtain a pure sample of $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets in data. In this regard, a different sample of jets with characteristics similar to those of the signal, referred to as “proxy jets”, is used to derive the SFs. The SFs derived from the proxy jets are used to calibrate simulated signal jets. Selecting proxy jets that closely resemble the characteristics of signal jets is of vital importance. Since the calibration tagging discriminant is designed to be decorrelated from the resonance mass, it is possible to select proxy jets with a $b\bar{b}$ ($c\bar{c}$) signature but induced from other processes.

One viable approach is to utilize the gluon-splitting $b\bar{b}$ ($c\bar{c}$) jets, which are abundant in the QCD multijet process. For example, Refs.^[75, 123] explore the gluon-splitting $b\bar{b}$ jets as the

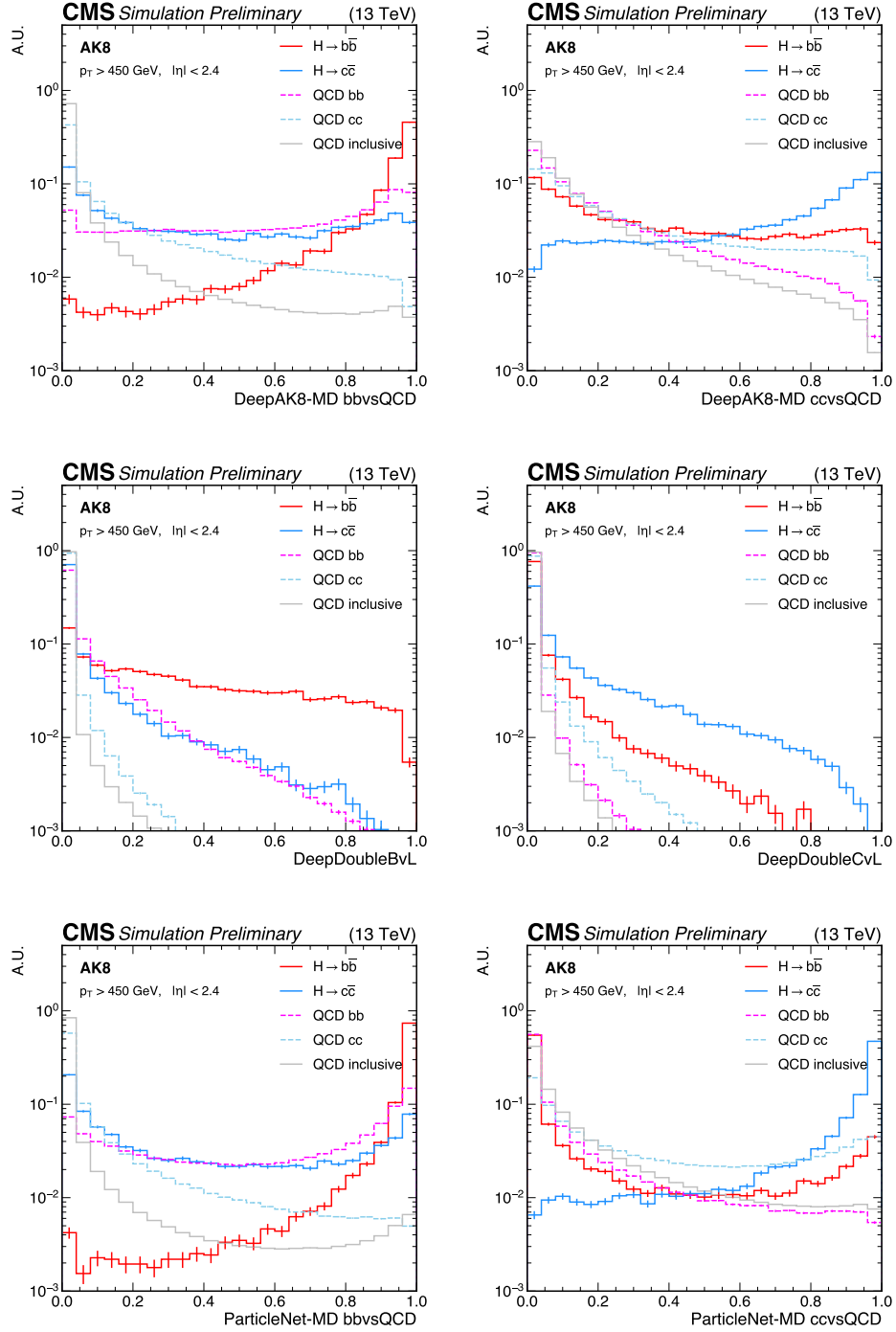


Figure 5.2 The normalized distributions of the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants on the simulated signal $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets, the bb and cc components of QCD multijet background jets, and inclusive QCD jets. From top to bottom: the DeepAK8-MD, DeepDoubleX, and ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants.

proxy to $H \rightarrow b\bar{b}$ jets. In the method developed within CMS^[75] to calibrate the double-b tagging discriminant, a dedicated reweighting procedure is performed on the SV-related variables to draw a closer phase space between $g \rightarrow b\bar{b}$ and $H \rightarrow b\bar{b}$. Similarly, in the early CMS efforts to calibrate the efficiency of DeepAK15-MD^① $X \rightarrow c\bar{c}$ tagging discriminant, The presence of two SVs and a phase-space selection is applied to ensure the similarity between selected $X \rightarrow c\bar{c}$ and $H \rightarrow c\bar{c}$ jets^[112]. However, it has been observed that a more performant $X \rightarrow b\bar{b}$ ($c\bar{c}$) tagging algorithm has a better discrimination power between the $b\bar{b}$ ($c\bar{c}$) jets from a resonance or by gluon-splitting; hence, it becomes more challenging to select $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets as a proxy. To calibrate the ParticleNet-MD $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminant, an upgrade on the earlier methods based on $g \rightarrow b\bar{b}$ ($c\bar{c}$) proxies is demanded.

The other approach is to rely on the $Z \rightarrow b\bar{b}$ jets to calibrate the $H \rightarrow b\bar{b}$ jets. This approach was previously explored in the ATLAS experiment^[123] and recently studied in CMS^[124]. Since Z is a massive particle, the jets directly correspond to the signal jets on which the tagger was trained. Therefore, no special selection needs to be applied to select proxy jets as is done in the method with jets originating from gluon splitting. However, measurement of Z jets comes with fewer statistics than gluon splitting jets, and there is a substantial QCD multijet background. The method, therefore, requires measuring the Z peak on the non-resonant hadronic background and often results in larger SF uncertainties.

The method introduced in Chapter 6 serves to upgrade the method based on $g \rightarrow b\bar{b}$ ($c\bar{c}$) proxies. It characterizes a multivariate approach to select $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets so as to reach a closer similarity with $H \rightarrow b\bar{b}$ ($c\bar{c}$) signals, and extensive studies on the different proxy selections that potentially impact the SF. In addition, this method is designed as a generalized approach to accommodate the calibration of various supported $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ taggers in CMS.

① A tagger developed in a similar approach with DeepAK8-MD, but on AK15 jets.

Chapter 6

A Novel Calibration Method: sfBDT

This chapter introduces a novel calibration method developed in the CMS experiment, the sfBDT method. The method is first recorded in Ref.^[125] and further developed in Ref.^[124]. Its aim is to calibrate the $X \rightarrow b\bar{b}$ ($c\bar{c}$) signal using gluon-splitting jets that pass an additional BDT selection as a proxy. The BDT used to select $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets is referred to as “sfBDT”, meaning the BDT for SF derivation.

In this chapter, Section 6.1 first outlines the data and simulated samples used for performing the calibration study, Section 6.2 describes all $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminants and their working points in which the calibration is performed, Sections 6.3 and 6.4 introduce the sfBDT method, and finally, the results are highlighted in Section 6.5.

6.1 Data and simulated samples

The calibration studies use proton-proton collision data collected during the 2016–2018 data-taking period of LHC Run 2 at 13 TeV. The selected data correspond to an integrated luminosity of 138 fb^{-1} . For data and simulated events, the study adopts the Run-2 UL reprocessing campaign. The events are simulated separately for four data-taking eras during Run 2 with their corresponding conditions, denoted as the 2016 pre-VFP, 2016 post-VFP, 2017, and 2018 eras. The 2016 pre-VFP and post-VFP eras are treated separately because of the substantial change in the strip tracker conditions between them, where VFP stands for feedback preamplifier bias voltage^[126].

Events are selected online using a logical OR of HLTs with different H_T (scalar p_T sum of all AK4 jets) thresholds, starting from 125 GeV (180 GeV) for the 2016 (2017 and 2018) era. The corresponding HLT paths in different data-taking conditions are summarized in Table 6.1.

The MC simulated samples used in this study are the $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ signal events and a list of background processes, including the QCD multijet process, V +jets, the top quark-antiquark pair ($t\bar{t}$), and the single top production process. Their production details, including the specifics on their ME and PS generators, are summarized in Tables 6.2 and 6.3, respectively.

The ggF $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ signal processes are not explicitly used in the calibration

Table 6.1 A description of the high-level triggers used in this sfBDT study in different data-taking conditions.

Conditions	HLT paths	Definition
2016 data-taking condition	PFHT125	$H_T > x$ GeV, (where H_T is the scalar p_T sum of all HLT-level AK4 jets, x refers to the number in the path)
	PFHT200	
	PFHT250	
	PFHT300	
	PFHT350	
	PFHT400	
2017 and 2018 data-taking conditions	PFHT475	same as above
	PFHT600	
	PFHT650	
	PFHT800	
	PFHT900	
	PFHT180	same as above
	PFHT250	
	PFHT350	
	PFHT370	
	PFHT430	
	PFHT510	
	PFHT590	same as above
	PFHT680	
	PFHT780	
	PFHT890	same as above
	PFHT1050	

Table 6.2 The list of simulated MC signal samples used for deriving the tagger working points, for the data-taking year 2016–2018. The matrix-element (ME) and parton-shower (PS) generators are also shown.

Process	ME generator	PS generator
ggF $H \rightarrow b\bar{b}$	HJ-MiNLO ($p_T(H) > 200$ GeV)	PYTHIA
ggF $H \rightarrow c\bar{c}$	HJ-MiNLO ($p_T(H) > 200$ GeV)	PYTHIA

Table 6.3 The list of simulated MC samples for the sfBDT calibration study in data-taking year 2016–2018. The matrix-element (ME) and parton-shower (PS) generators, prediction order for the total cross section, and the cross section values are shown.

Process	ME generator	PS generator	pred. order for total σ	σ [pb]
QCD multijet	MG ($300 < H_T < 500$ GeV)	PYTHIA	LO	322600
	MG ($500 < H_T < 700$ GeV)	PYTHIA	LO	29980
	MG ($700 < H_T < 1000$ GeV)	PYTHIA	LO	6334
	MG ($1000 < H_T < 1500$ GeV)	PYTHIA	LO	1088
	MG ($1500 < H_T < 2000$ GeV)	PYTHIA	LO	99.11
	MG ($H_T > 2000$ GeV)	PYTHIA	LO	20.23
$(V \rightarrow q\bar{q}) + \text{jets}$	MG ($W \rightarrow q\bar{q}$, $400 < H_T < 600$ GeV)	PYTHIA	LO	314.8
	MG ($W \rightarrow q\bar{q}$, $600 < H_T < 800$ GeV)	PYTHIA	LO	68.64
	MG ($W \rightarrow q\bar{q}$, $H_T > 800$ GeV)	PYTHIA	LO	34.7
	MG ($Z \rightarrow q\bar{q}$, $400 < H_T < 600$ GeV)	PYTHIA	LO	145
	MG ($Z \rightarrow q\bar{q}$, $600 < H_T < 800$ GeV)	PYTHIA	LO	34.12
	MG ($Z \rightarrow q\bar{q}$, $H_T > 800$ GeV)	PYTHIA	LO	18.57
$t\bar{t}$	POWHEG v2	PYTHIA	NNLO+NNLL	377.96
single top	POWHEG (tW process)	PYTHIA	NLO	19.6
	POWHEG ($\bar{t}W$ process)	PYTHIA	NLO	19.6
	POWHEG (t-channel t production)	PYTHIA	NLO	136.02
	POWHEG (t-channel $\bar{t}$ production)	PYTHIA	NLO	80.95

study but in the derivation of the standard working point of the jet taggers. The procedure is introduced in Section 6.2.1. As can be seen in the tables, These events are simulated using the HJ-MiNLO^[26, 127, 128] event generator with the Higgs boson mass $m_H = 125$ GeV, interfaced with PYTHIA v8.230 for Higgs boson decaying to $b\bar{b}$ or $c\bar{c}$ and event hadronization.

For the main MC simulated processes used for calibration, the main QCD multijet and $V + \text{jets}$ ($V = Z, W$) processes are modelled at QCD LO accuracy using the MG v2.6.5 generator^[25]. They are simulated with up to 4 (3) extra partons for the QCD ($V + \text{jets}$) process at the matrix element (ME) level. The Z (W) boson is configured to decay to a quark-antiquark pair at the ME level of the $Z + \text{jets}$ ($W + \text{jets}$) event.

For $t\bar{t}$ simulation, the ME generation is performed with POWHEG v2^[26, 27, 129, 130] at NLO accuracy in QCD, and its cross section is scaled to a theoretical prediction at next-to-NLO (NNLO) in QCD, including resummation of next-to-next-to-leading logarithmic soft-gluon terms^[131]. The single top quark production in the t -channel (tW channel) was simulated using POWHEG also in the 4-flavour (5-flavour) scheme^[132-135], with its cross section normalized to the NLO calculations^[136].

The parton showering of all processes is simulated with PYTHIA v8.230^[28], using the CP5

tune^[137] with the NNPDF3.1 NNLO parton distribution function (PDF) set^[138]. For QCD multijet and V +jets events, the matching of jets from ME calculations and those from PS is done with the MLM^[139] technique. Here, it is important to note that a consistent parton shower generator should be used in a Higgs analysis when generating $H \rightarrow b\bar{b}$ ($c\bar{c}$) processes and in our calibration study where $g \rightarrow b\bar{b}$ ($c\bar{c}$) proxy jets are selected. This is because the performance of a DNN jet algorithm on simulated events is found to be very sensitive to the parton shower patterns. In the recent CMS searches for Higgs-related processes^[107-109, 112-115], the same PYTHIA generator is used for parton showering.

6.2 Calibrated taggers in CMS

6.2.1 List of taggers and working points

The $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging discriminants used for performing the calibration study are introduced in Section 5.1. They are four $X \rightarrow b\bar{b}$ discriminants: the double-b, DeepAK8-MD, DeepDoubleX, and ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant, and three $X \rightarrow c\bar{c}$ discriminants: the DeepAK8-MD, DeepDoubleX, and ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant.

For each discriminant, three working points are determined at the given efficiency points of the $H \rightarrow b\bar{b}$ or $H \rightarrow c\bar{c}$ signal jets, namely the high purity (HP), medium purity (MP), and low purity (LP) working points. The signal jets are selected from the MC simulated ggF $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ signal events, requiring the Higgs boson, as well as its two daughters, to be within the jet cone ($\Delta R < 0.8$). For $X \rightarrow b\bar{b}$ ($c\bar{c}$) tagging discriminants, the working points of HP, MP, and LP are defined as $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets at efficiencies of 40% (15%), 60% (30%), and 80% (50%), respectively, according to the common working point choices adopted in various CMS physics searches^[107-109, 112-115, 120]. The working points are determined in four data-taking eras separately.

Table 6.4 summarizes the working points obtained from each era for the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging discriminants, respectively.

6.2.2 Performance

The performances of the $X \rightarrow b\bar{b}$ ($c\bar{c}$) tagging discriminants are compared in the 2018 data-taking condition. The comparison is performed in terms of receiver operating characteristic (ROC) curves, shown in Fig. 6.1, which describes the selection efficiency of $X \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets as a function of background selection rate.

Table 6.4 Definition of the three working points (WPs) for each calibrated $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ tagging discriminants to derive the scale factors, shown in four data-taking year conditions. The three WPs correspond to selecting with the efficiency of 40%, 60% and 80% of the $H \rightarrow b\bar{b}$ signal jets, and the efficiency of 15%, 30% and 50% of the $H \rightarrow c\bar{c}$ signal jets.

$X \rightarrow b\bar{b}$ or $c\bar{c}$ discriminant	High purity (HP)	Medium purity (MP)	Low purity (LP)
UL 2016 pre-VFP			
ParticleNet-MD bbvsQCD	0.9883	0.9737	0.9088
DeepDoubleBvL	0.2739	0.1180	0.0256
DeepAK8-MD bbvsQCD	0.9521	0.8940	0.7021
double-b	0.8926	0.7559	0.3104
ParticleNet-MD ccvsQCD	0.9909	0.9751	0.9252
DeepDoubleCvL	0.2568	0.1043	0.0327
DeepAK8-MD ccvsQCD	0.9326	0.8496	0.6733
UL 2016 post-VFP			
ParticleNet-MD bbvsQCD	0.9883	0.9735	0.9137
DeepDoubleBvL	0.2786	0.1213	0.0270
DeepAK8-MD bbvsQCD	0.9507	0.8877	0.6890
double-b	0.8955	0.7646	0.3279
ParticleNet-MD ccvsQCD	0.9905	0.9743	0.9252
DeepDoubleCvL	0.2505	0.1021	0.0321
DeepAK8-MD ccvsQCD	0.9277	0.8369	0.6458
UL 2017			
ParticleNet-MD bbvsQCD	0.9870	0.9714	0.9105
DeepDoubleBvL	0.3154	0.1566	0.0404
DeepAK8-MD bbvsQCD	0.9546	0.9033	0.7422
double-b	0.9185	0.8237	0.5068
ParticleNet-MD ccvsQCD	0.9909	0.9765	0.9347
DeepDoubleCvL	0.3013	0.1361	0.0457
DeepAK8-MD ccvsQCD	0.9390	0.8677	0.7100
UL 2018			
ParticleNet-MD bbvsQCD	0.9880	0.9734	0.9172
DeepDoubleBvL	0.3140	0.1566	0.0399
DeepAK8-MD bbvsQCD	0.9551	0.9058	0.7432
double-b	0.9175	0.8193	0.4939
ParticleNet-MD ccvsQCD	0.9917	0.9777	0.9368
DeepDoubleCvL	0.3013	0.1322	0.0425
DeepAK8-MD ccvsQCD	0.9390	0.8662	0.7017

The signal jets are the same truth $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets used for defining the working points. The background jets are selected from the leading jets in the QCD multijet events, without passing specific selection. Jets are required to have $p_T > 450$ GeV, $|\eta| < 2.4$, in making the ROC curve. Two sets of ROC curves are shown. Additional requirements involve the mass window selection $90 < m_{SD} < 140$ GeV, which is in line with the convention in Refs.^[75, 117]. Both performances of $X \rightarrow b\bar{b}$ ($c\bar{c}$) versus the entire QCD jet collection and versus the bb (cc) background component are shown using the simulated signal and background events under the 2018 data-taking condition. It can be seen that ParticleNet-MD demonstrates largely enhanced separating power between the signal and QCD background.

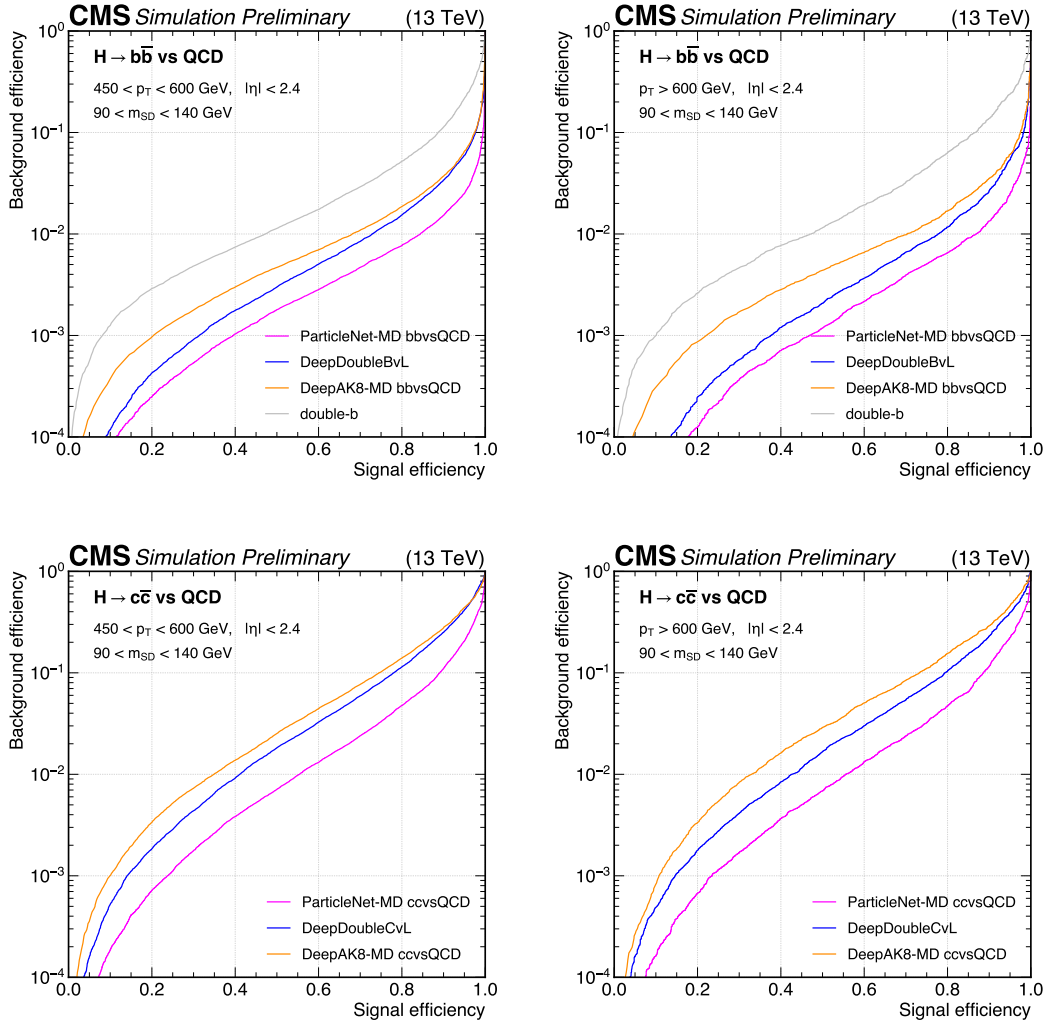


Figure 6.1 Performance comparison of the $X \rightarrow b\bar{b}$ (top) and $X \rightarrow c\bar{c}$ (bottom) identification algorithms in terms of ROC curves for $H \rightarrow b\bar{b}$ or $H \rightarrow c\bar{c}$ signal jets versus the inclusive QCD jets as background. Performance is shown in the $450 < p_T < 600$ GeV (left) and $p_T > 600$ GeV (right) regions.

Figure 6.2 further shows the mass decorrelation performance on the QCD multijet background is checked on the $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants using the 2018 simulation. Here, we see that the DeepAK8-MD taggers result in sculpting on the mass spectrum around the Higgs boson mass, while the effect is less pronounced in the case of ParticleNet-MD and DeepDoubleX. This can be explained as a defect of the adversarial training approach for mass decorrelation compared to the mass spectrum reweighting approach.

6.3 Event and jet selection

This section introduces the event and jet selections for the sfBDT method. Specifically, the design of the sfBDT variable and the choice of selections will be highlighted.

6.3.1 Event selection

In the method using $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets as a proxy to the signal jets, events are triggered by H_T requirements. The MC samples are formed as a combination of QCD (PYTHIA), $V(\rightarrow q\bar{q})$ +jets, $t\bar{t}$, and single top processes. The contribution from QCD is dominant. Other processes contribute overall less than 8% after applying all selections, and less than 20% in the highest p_T bin and tightest tagger WP when QCD events are highly suppressed.

As introduced in Section 6.1, to enlarge the data statistics, we take the logical OR of all the H_T HLTs. Compared to the earlier calibration strategy for DeepAK15 $X \rightarrow c\bar{c}$ discriminants^[112], where only the unscaled H_T trigger with a high threshold is used, the new strategy helps to recover the statistics in the low large- R jet p_T region to a large extent. To further increase jet statistics, both the leading and subleading jets (when $N(\text{jets}) \geq 1$) are considered jet candidates. The jet-based selections are then applied to all jets from the collection of candidates.

However, we note that most H_T HLT paths are prescaled, except a few highest H_T ones. It means that for HLTs with a lower H_T threshold, only a proportion of data events passing the threshold are maintained. Figure 6.3 shows collected events passing jet H_T triggers with different thresholds and the corresponding prescaling factors.

On the other hand, the simulated events are not affected by the trigger prescaling. This causes a much larger event yield in MC simulation. Therefore, to well-estimate the data phase space by MC simulation, we remove all H_T requirements in MC and apply an overall MC-to-data reweighting. Here, the reweighting is done on a jet basis and performed on the three-dimensional grid: (event H_T , jet p_T , jet index $\in \{1, 2\}$). The detailed binning is shown in

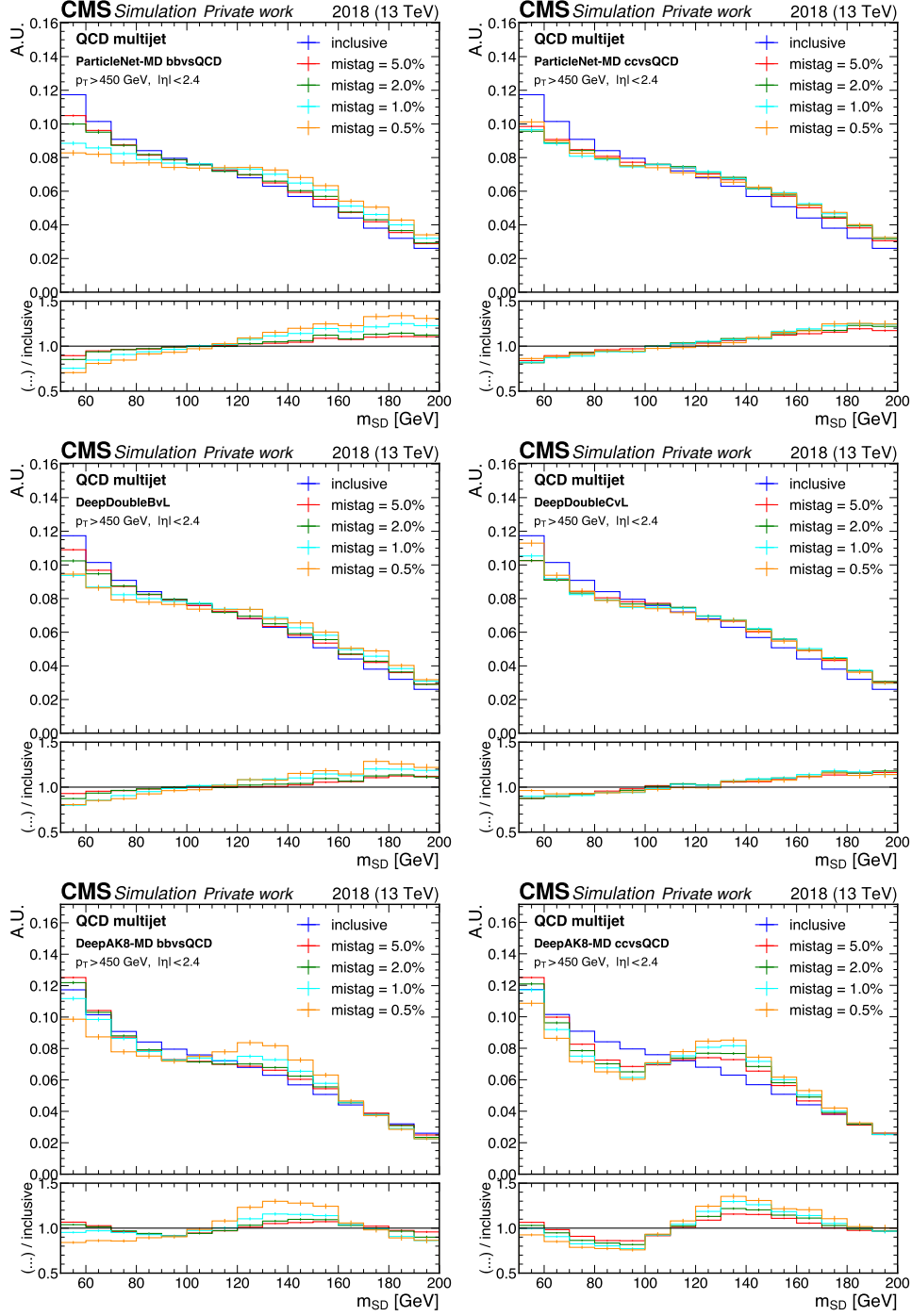


Figure 6.2 The mass decorrelation performance on the QCD multijet background for various $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants. From top to bottom: The ParticleNet-MD, DeepDoubleX, DeepAK8-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants.

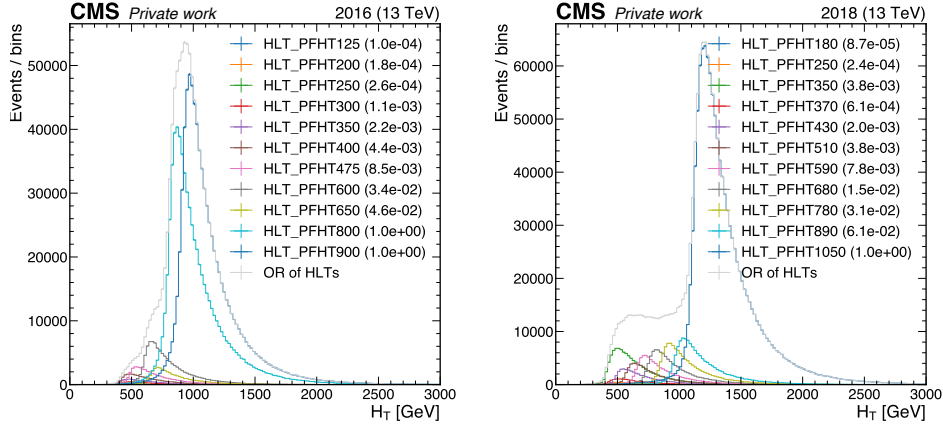


Figure 6.3 The number of data events passing the jet H_T triggers with different thresholds and passing the logical OR of all triggers, for the 2016 post-VFP (left) and 2018 (right) detector conditions as an example. The prescaling effect is demonstrated by low-threshold HLTs, and the number in parentheses indicates the prescaling factor for the corresponding trigger.

Table 6.5. For each bin i , a reweighting factor r_i , which is the ratio of MC and data yields in this bin, is extracted. The procedure is repeated separately for four data-taking conditions and applied to all MC simulated jets (regardless of the MC process). As an example, Fig. 6.4 shows the MC and data yields, where the reweight factors are displayed in the lower panels. It is obvious that the factors in the low event H_T or low jet p_T region have 1 $\sim$ 2 orders of magnitude smaller than 1 due to the use of prescaled triggers for the data selection.

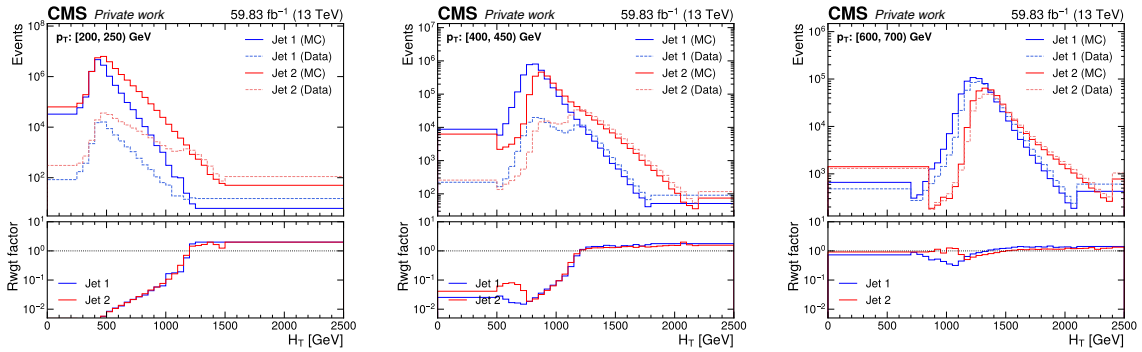


Figure 6.4 Summary of the MC and data yields (upper panel for each plot) and the MC-to-data reweighting factors (lower panel) for the 2018 data on the 3D variables (event H_T , jet p_T , jet index). The H_T ranges are shown in the x axis; jet p_T ranges are separated by plots, where three example ranges of (200, 250), (400, 450), and (600, 700) GeV are shown from left to right; the jet index takes the value of 1 and 2.

Utilizing the reweighting factor is a crucial means in this method to enlarge the data statistics while preserving a reasonable estimation of data events using simulation. Alternative ap-

Table 6.5 Details of the 3D binning on (event H_T , jet p_T , jet index $\in \{1, 2\}$) used for the MC-to-data reweighting. When the jet index and p_T range are fixed, the H_T adopts an equal-width binning with a varying upper and lower bound and a fixed width of 50 GeV with flows included. The H_T boundaries in the 3D grid are shown.

Jet p_T ranges [GeV]	H_T {lower, upper} bounds [GeV]	
	jet index = 1	jet index = 2
(200, 250)	{250, 1250}	{250, 1500}
(250, 300)	{350, 1400}	{350, 1600}
(300, 350)	{400, 1600}	{400, 1800}
(350, 400)	{450, 1700}	{450, 2000}
(400, 450)	{500, 1800}	{500, 2200}
(450, 500)	{550, 1900}	{550, 2400}
(500, 550)	{600, 1900}	{650, 2400}
(550, 600)	{650, 2000}	{750, 2400}
(600, 700)	{700, 2100}	{850, 2400}
(700, 800)	{800, 2200}	{1000, 2400}
(600, $+\infty$)	{1000, 2400}	{1200, 2400}

proaches have been studied, including using only the unrescaled HLT with the highest H_T thresholds with no reweighting needed or changing to simpler reweighting on one-dimensional binning. An agreement can be found in the calibration results between these approaches, while this method results in smaller statistical uncertainty and reveals stronger stability.

6.3.2 Jet selection

Jet selections are applied on top of the event selection. First, all selected jets must pass the kinematics preselection of $p_T > 200$ GeV, $|\eta| < 2.4$, and $50 < m_{SD} < 200$ GeV, where m_{SD} is the groomed jet mass using the soft-drop algorithm.

Additional requirements involve the existence of SV within the jet. We use the same requirement as the previous method for calibration of the DeepAK15 $H \rightarrow c\bar{c}$ tagging discriminant^[112], asking for at least one SV matched to each of the two subjets. The matching criterion is

$$\Delta R(\text{SV}, \text{sj}_i) < \min\{0.4, 0.5 \cdot \Delta R(\text{sj}_{1,2})\}, \quad (6.1)$$

where SV are the secondary vertices, sj_i is the i -th out of two subjets resulting from the soft-drop algorithm, and $\Delta R(\text{sj}_{1,2})$ is the distance between the two subjets. The SVs are required to pass d_{xy} significance > 2 .

On top of the SV requirement, a multivariate approach is adopted to further select those

QCD jets that resemble the two-prong $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets. Specifically, a BDT (named sfBDT) is trained among two subclasses of QCD jets, and it is then applied to select only a portion of jets. The details of the sfBDT design concept, training, and selection criteria will be our focus in the next subsection.

In simulated events, it is necessary to introduce the flavour categorization based on the number and types of heavy flavour hadrons it contains. This is crucial in the definition of “proxy jets” from MC events, which are used to represent the signal $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets. To be specific, each large- R jet is subdivided into five exclusive categories: bb-, b-, cc-, c-, and light-type jets, by matching the jets with the truth hadrons using the ghost-matching technique. The definition is summarized in Table 6.6.

Table 6.6 The truth-matching criteria to assign a large- R jet from the QCD multijet events to one of the five exclusive categories: bb, b, cc, c, and light.

Category	Criteria
“bb”	If each of the two subjects in the large- R jet are ghost-matched to at least one b-hadron.
“b”	Else if the large- R jet is ghost-matched to at least one b-hadron.
“cc”	Else if each of the two subjects in the large- R jet are ghost-matched to at least one c-hadron.
“c”	Else if the large- R jet is ghost-matched to at least one c-hadron.
“light”	Else.

Based on this categorization, the “b” and “bb” categories are merged into “flvB”; the “cc” and “c” are merged into “flvC”; and the “light” forms the “flvL”. Thus, the proxy jets in the case of calibrating the bb (cc) flavour taggers are designated by the QCD jets in flvB (flvC), after passing the SV selection and the sfBDT selection. Thanks to the requirements for the presence of two SVs, the proxy jets are dominated by $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets, with only a very small fraction of jets originating from a single b (c) quark. The sfBDT selection further selects a subset of $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets to make it resemble $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets.

For clarity, we provide a summary of the definitions of “proxy jet” and “signal jet”. The signal jets are selected as truth-matched $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets from the SM ggF $H \rightarrow b\bar{b}$ ($c\bar{c}$) events. The proxy jets, which we select from collected events to mimic the characteristics of signal jets, are determined following the list below. We note that all steps are composed of procedures introduced earlier, but the order is provided here for clarity. The sfBDT selection realizes the mission of drawing the proxy as close as possible to the signal jets.

1. **Datasets:** use the data passing JetHT triggers and MC samples composed of QCD, $V(\rightarrow q\bar{q})$ +jets, $t\bar{t}$, and single top processes.
2. **The event selection:** use the logical OR of all H_T triggers for data, and remove all H_T

triggers for MC.

3. **The jet selection on SV:** among all the leading or subleading large- R jets, select the ones passing the kinematics preselection, the SV requirement.
4. **The MC-to-data reweighting:** apply a reweighting from selected MC to data jets on a three-dimensional grid.
5. **The sfBDT selection:** the additional selection on a BDT variable, discussed in the next subsection.
6. **Specific flavour component:** all events are categorized into flvB, flvC, and flvL. The proxy jets are designated as the flvB (flvC) component in the case of calibrating the $X \rightarrow b\bar{b}$ ($c\bar{c}$) tagging discriminants.

6.3.3 Design of the sfBDT

The design spirit of sfBDT is that we select two subsets of jets in QCD $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets, using criteria from the generator-level information within the jets. One subset is chosen to be more alike to the $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets, and another is less alike. We then train a BDT to distinguish jets between the two subsets. A selection on the trained BDT will, therefore, select the region with more signal-like jets.

The core of the BDT design is to understand, from the truth level, the underlying difference from $b\bar{b}$ ($c\bar{c}$) jets initiated from gluons or the heavy resonance, and then propose generator-level criteria to define the two subsets. In this regard, we introduce two designs of BDT. An initial design is based on the generator-level partons (quarks and gluon) within a jet. This version of sfBDT is documented earlier in Ref.^[125]. A latter version exploits the generator-level hadrons within the jet, documented in Ref.^[124]

6.3.3.1 The sfBDT based on generator-level partons

A generator-level study is first performed to uncover the difference between the $g \rightarrow b\bar{b}$ ($c\bar{c}$) and $H \rightarrow b\bar{b}$ ($c\bar{c}$) large- R jets. The following study is performed on the $c\bar{c}$ final state.

Our first interesting observation is that large- R jets originated by $g \rightarrow c\bar{c}$ are more likely to contain extra fragments originating from additional gluons. These gluons potentially come from the initial- or final-state radiation (ISR or FSR) of the QCD process. We refer to this phenomenon as “gluon contamination” that appears in proxy jets. To visualize this effect, a generator-level variable κ_g , representing the gluon contamination rate, is defined as

$$\kappa_g = \frac{\sum_{i \in \{g\}} p_{T,i}}{\sum_{i \in \{g, q\}} p_{T,i}}. \quad (6.2)$$

Here, the set $\{g\}$ and $\{q\}$ in the equation are formed by taking all gluons and quarks in the PYTHIA final state, that appears as the last copy in the list before their hadronization, satisfy $p_T > 15$ GeV, and have ΔR smaller than the jet radius with respect to the large- R jet axis. Figure 6.5 compares the κ_g distributions between the $H \rightarrow c\bar{c}$ signal and the $g \rightarrow c\bar{c}$ background from the QCD multijet events. It is evident that $g \rightarrow c\bar{c}$ jets generally have a larger gluon contamination effect.

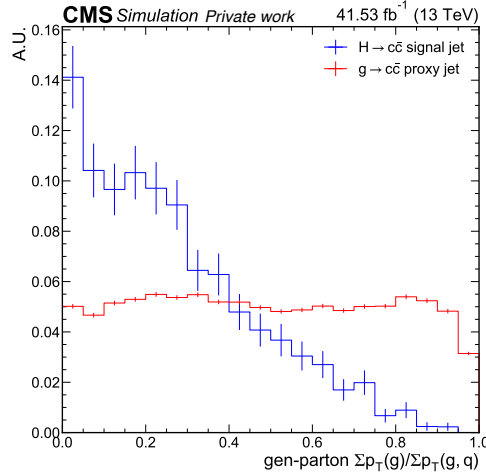


Figure 6.5 The κ_g distributions of large- R jets for the signal $H \rightarrow c\bar{c}$ jets and the $g \rightarrow c\bar{c}$ jets selected from QCD multijet events.

Two subsets from QCD jets are determined, which are used for training a BDT. For this QCD jet selection, the simulated QCD events in the 2017–2018 data-taking conditions are adopted, without passing the reweighting procedure but requiring the jets to pass the SV selection, as mentioned. Only the bb -type or cc -type QCD jets are chosen to form a pure $g \rightarrow b\bar{b}$ and $g \rightarrow c\bar{c}$ jet collection, using the categorization in Table 6.6. Jets that survived the selections are then labelled as the “signal-like jet” if $\kappa_g < 0.15$, or the “gluon-contaminated jet” if $\kappa_g > 0.85$. Here, smaller κ_g indicates smaller gluon contamination, which is more like the signal.

A BDT is trained to discriminate the signal-like jet against the gluon-contaminated jet. The input features of the BDT include several track and SV properties of the large- R jet. After optimization, 6 input observables are chosen, as shown in Table 6.7.

The BDT training and the score prediction are performed with XGBoost 0.72.1^[140]. To avoid any potential bias from overfitting to the training dataset, a 10-fold cross-validation^[141] is used in the development of the BDT, i.e., a total of 10 BDT models are trained with exactly the same input and configurations, but each using only 90% of the events for training and

Table 6.7 The list of variables used in the sfBDT training to discriminate the signal-like jet versus gluon-contaminated jet.

Variable	Description
τ_{21}	The N -subjettiness variable ^[69] defined by τ_2/τ_1
M_{sj1}	The mass of the leading subjet in p_T obtained from the softdrop algorithm
M_{sj2}	The mass of the subleading subjet in p_T obtained from the softdrop algorithm
$N_{\text{tracks from } SV_{1,2}}$	The number of tracks associated with the two SVs which are matched to the two subjets
p_{T, SV_1}	The p_T of the SV matched to the leading subjet in p_T
p_{T, SV_2}	The p_T of the SV matched to the subleading subjet in p_T

predict on the rest 10% of the event. In this way, no events are used in both the training and the application of a BDT model. The sfBDT model is applied to all 2016–2018 simulated and collected data for prediction.

The predicted sfBDT score is capable of identifying genuine bb - (cc -)dominated proxy jets with less gluon contamination. Therefore, by tightening the sfBDT selection threshold, the $g \rightarrow b\bar{b}$ ($c\bar{c}$) proxy jet will be driven more like the signal jet.

Figure 6.6 illustrates the effect of sfBDT in the example of $g \rightarrow c\bar{c}$ and $H \rightarrow c\bar{c}$ jets. The 2D histogram on the ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant versus the sfBDT score is shown. The proxy large- R jets contain both a genuine cc component, with a higher tagging discriminant value and sfBDT score, and a group of gluon-dominated large- R jets with a lower tagging discriminant value and sfBDT score. By using the additional BDT selection, it is clear that the selected $g \rightarrow c\bar{c}$ large- R jet phase space is driven more like the signal one.

As a second illustration, we show in Fig. 6.7 the tagging discriminant shape comparison between the signal and proxy jets, and visualize how the shape changes when an sfBDT selection on multiple choices of thresholds is applied. We see that the threshold on the sfBDT can be tuned so that the dynamic range of the normalized ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant shape for proxy jets can fully cover that for the signal. In general, with a tighter cut on the sfBDT score, the proxy tagger shape is driven to be more similar to the signal.

6.3.3.2 The sfBDT based on generator-level hadrons

Although the exceeds in gluons among all partons in the QCD $g \rightarrow b\bar{b}$ ($c\bar{c}$) jets can be used as a handle to select jets that are more signal-like, there are two reasons one hopes to use hadron-level information than the parton-level ones. First, the partons are considered internal to the generator, and their pattern is generator-dependent, carrying less physics meaning. In contrast, hadrons, which undergo full propagation through strong interactions, carry a greater

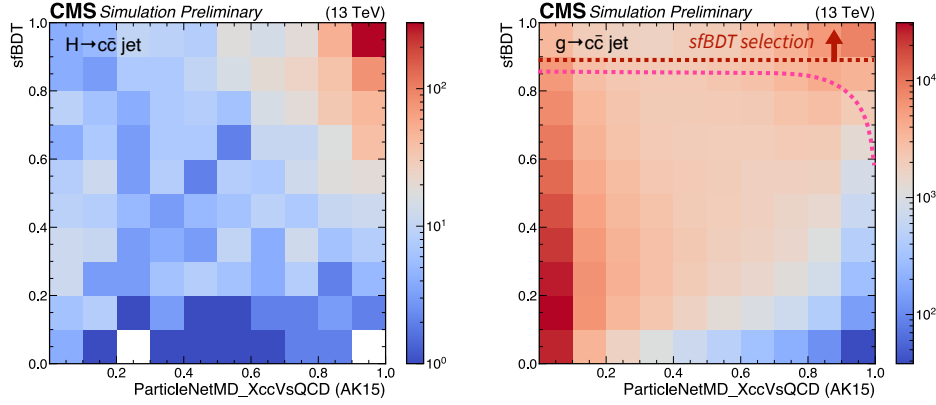


Figure 6.6 A 2D histogram on the tagging discriminant (ParticleNet-MD $X \rightarrow c\bar{c}$ developed for AK15 jets) versus the sfBDT score, for the $H \rightarrow c\bar{c}$ (left) and the $g \rightarrow c\bar{c}$ (right) jets^[125]. The $g \rightarrow c\bar{c}$ histogram has two enriched regions, one in the top-right region that resembles $H \rightarrow c\bar{c}$ signal jets, and one in the bottom-left corner representing jets with more gluon contamination. A selection involving the sfBDT helps to select a dedicated phase space that improves the proxy and signal jet similarity. Two types of selection are considered depending on the specific signal and proxy condition, as illustrated in the right plot.

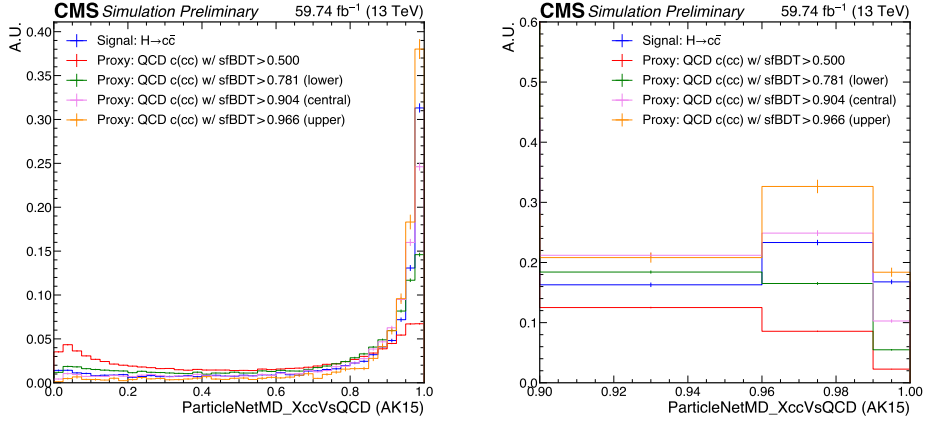


Figure 6.7 The distributions of the tagging discriminant (ParticleNet-MD $X \rightarrow c\bar{c}$ developed for AK15 jets)^[125]. Both equal-width binning (left) and analysis-defined working points^[112] (right) are shown. The histograms demonstrate that the sfBDT is capable of tuning the tagger shape and improving the similarity between the signal and proxy shape.

degree of physical meaning. Second, while gluon contamination is one metric to remove QCD-like jets, it is possible that the additional ISR or FSR partons can not only be gluons, but also quarks. This is illustrated in the right diagram of Fig. 6.8. It implies that one can measure the substructure (e.g. multiprongness) of the jet to accommodate both cases.

In this regard, a metric to evaluate the multiprongness of the jet using the generator-level

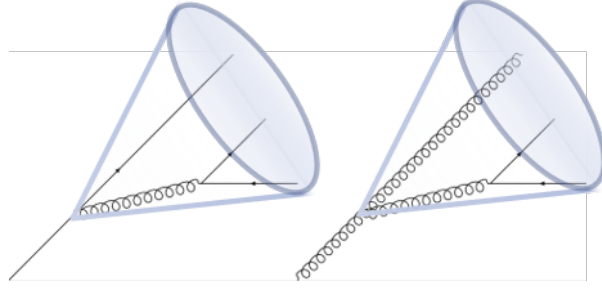


Figure 6.8 The cartoon of a $g \rightarrow b\bar{b}$ ($c\bar{c}$) jet with an additional parton emission in the format of Feynman diagrams.

hadrons is used. We draw to the N -subjettiness variable^[69] of a jet, defined on the first-generation hadrons. Here, the first-generation hadrons are hadrons in the PYTHIA list, which is produced directly from the partons before the sequential decay. After an optimization procedure, the τ_{31} of hadrons is found the optimal choice amongst all N -subjettiness variables and is used as a replacement of metric κ_g . The variable is denoted as τ_{31}^h :

$$\tau_{31}^h = \frac{\sum_{i \in \{\text{had.}\}} p_{T,i} \min[\Delta R_{i, \hat{n}_{3,1}}, \Delta R_{i, \hat{n}_{3,2}}, \Delta R_{i, \hat{n}_{3,3}}]}{\sum_{i \in \{\text{had.}\}} p_{T,i} \Delta R_{i, \hat{n}_{1,1}}}. \quad (6.3)$$

Here, $\hat{n}_{1,1}$ is the direction of the vector sum of all input hadron momenta, and $\hat{n}_{3,j}$ ($j = 1, 2, 3$) are the three subjet axes of the hadrons, obtained by performing the exclusive k_T algorithm^[142, 143] on the hadron list.

For the training of the new sfBDT, the same event and jet selections on the simulated QCD events are adopted. The resulting jets are all $b\bar{b}$ - and $c\bar{c}$ -type jets. However, the QCD events are not taken from the standard QCD simulation in the experiment but produced in a customized form such that we can retrieve the full hadron list by PYTHIA simulation. To improve the efficiency for jet selection, the customized QCD sample is configured to be highly enriched in b and c quarks (with up to three b or three c quarks in an event). Jets that survived the selections are then labelled as the signal-like jet if $\tau_{31}^h < 0.1$ or background-like jet if $\tau_{31}^h > 0.1$. The details of the training are the same as those of the earlier sfBDT development.

6.3.4 Selections via the sfBDT discriminant

With the sfBDT variable, it is crucial to determine the sfBDT selection threshold to draw a better similarity between proxy and signal jets. In Ref.^[125] for the early sfBDT calibration method, the selection is considered either a 1D selection on the sfBDT variable itself or a 2D selection on the sfBDT and the tagging discriminant plane, as demonstrated in the dashed straight line and the curved line in Fig. 6.6. For the latter, the sfBDT threshold is a function

of the tagging discriminant value, where a higher discriminant score results in a looser sfBDT selection to preserve more events in this region. Later development aims to solve the arbitrary in defining the sfBDT selection, introducing an automatic procedure to determine the suitable sfBDT selections to be applied for deriving the SF. Regardless of the methods adopted to determine the sfBDT selection, the primary goal maintains, i.e. when the choices of selections vary, the dynamic range of the proxy-jet tagging discriminant shape should cover the shape for the signal jets.

The following subsections describe details of the later method in defining the sfBDT selections.

6.3.4.1 Automatic determination of sfBDT coastlines

The automatic process is presented as follows. First, to ensure the consistency of the method adopted for various taggers, a tagging discriminant transformation is applied so as to transform the signal tagger shape into a uniform distribution on $(0, 1)$. It is straightforward to see that the transformation function $T(x)$ has the following form:

$$\tilde{x} = T(x) = \int_0^x f(x') dx', \quad (6.4)$$

where x and $\tilde{x}$ are the original and transformed discriminant values, $f(x)$ is the probabilistic distribution function (PDF) of the original signal tagger shape. Figure 6.9 provides an example of the transformation map for the ParticleNet-MD and DeepDoubleX $X \rightarrow b\bar{b}$ discriminants. Also shown are the discriminant distributions before and after the transformation is applied. The following steps in this method will be based on the transformed tagging discriminant instead of the original one.

Then, selections on the sfBDT as a function of the tagging discriminant score are defined. They include a series of such selections, and they are presented by a bunch of curves when demonstrated on the 2D tagger–sfBDT axes. These curves are referred to as “sfBDT coastlines”. An example of sfBDT coastlines is illustrated in Fig. 6.10. They are determined in an automatic way that the remaining proxy jet collection (with its definition summarized in Section 6.3.2 as a reminder) that survives the selection will have the same tagging discriminant shape as the signal jets, i.e. a uniform distribution on $(0, 1)$. Mathematically, denoting the tagging discriminant as x and sfBDT value as y , the sfBDT coastlines are found to be contours of the accumulative 2D PDF

$$F_y(x, y) = \int_y^1 f(x, y') dy', \quad (6.5)$$

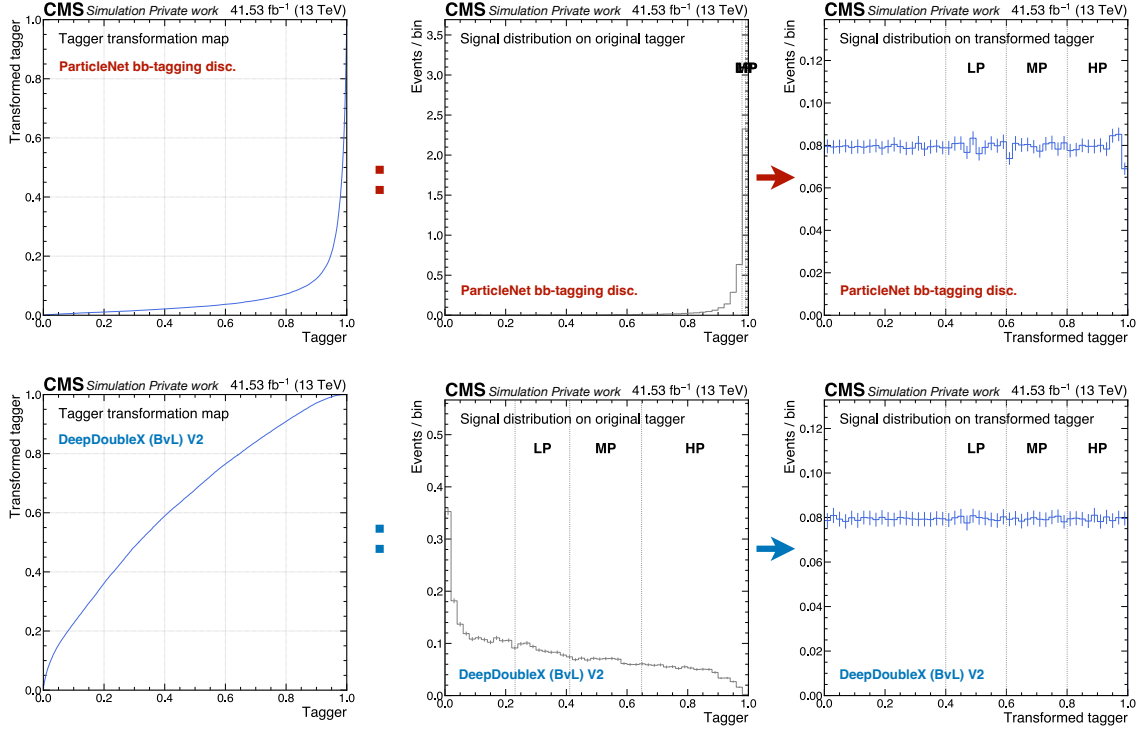


Figure 6.9 The example plots for the tagger transformation map, the original tagger shape and the transformed tagger shape for the ParticleNet-MD (top) and DeepDoubleX (bottom) $X \rightarrow b\bar{b}$ discriminants.

where $f(x, y)$ are the 2D PDF on the tagger–sfBDT plane, which is illustrated in a discrete form by the 2D coloured pixels in Fig. 6.10.

Here, a special note is that when the coastline reaches the bottom axis, it means the events from the high tagging discriminant region are not adequate enough to yield a uniform distribution after selection. We will still stick to the mathematical coastline definition in such a case.

A total number of 9 coastlines is selected. These correspond to 9 selections on the contour lines, determined by $F_y(x, y) = \frac{k}{N} F_y(0.6, 0)$, where $N = 12$ and $k = 3, \dots, 11$. Figuratively (see Fig. 6.10), it means that the loosest coastline selection should always pass (0.6, 0) on the 2D tagger–sfBDT plane. The choice of N and the options of the index k are optimized so that the selected coastlines are representative enough, covering as fully as possible the region where any coastline can lie. Various choices on N , k and the boundary point have been studied and found to have a small impact on the result.

Figures 6.11 and 6.12 show the derived sfBDT coastlines in three exclusive jet p_T bins in the calibration of ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants as an example.

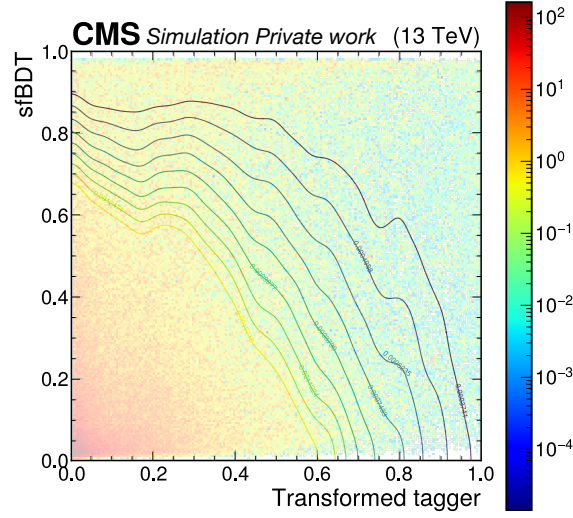


Figure 6.10 The 2D histogram of the tagger–sfBDT variable (the coloured pixels), and illustration of the “sfBDT coastlines” (the curves). The 2D histogram showcases the 2D PDF $f(x, y)$, and the coastlines are mathematically defined as the contour of the accumulative PDF $F_y(x, y) = \int_y^1 f(x, y') dy'$.

To validate if selecting the sfBDT coastlines results in an exact proxy-to-signal similarity on the tagging discriminant, Fig. 6.13 shows the proxy tagging discriminant distribution, taking ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant as an example. When selecting the coastlines with varying indices, the normalized proxy shape matches well with the signal.

6.3.4.2 Similarity of the proxy and signal

In addition to comparing the $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets and proxy jets after the sfBDT coastline selection on the tagging discriminant distribution, it is also necessary to compare

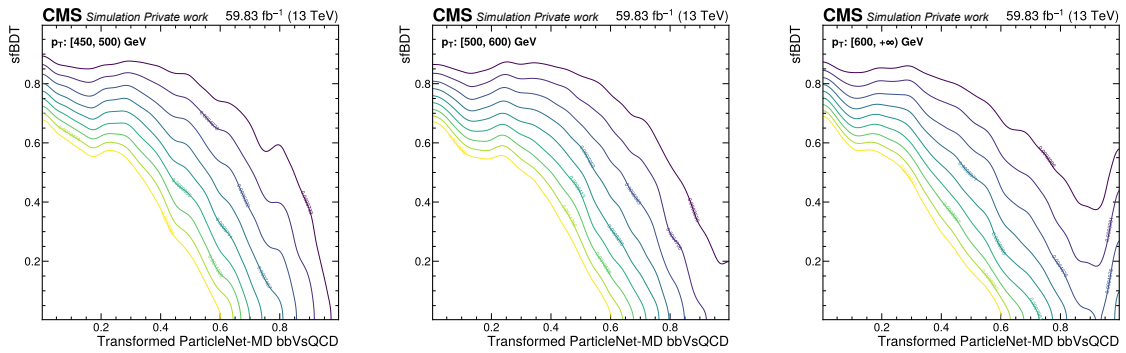


Figure 6.11 The sfBDT coastlines derived in the context of calibrating the ParticleNet-MD $X \rightarrow b\bar{b}$ discriminants in three exclusive p_T bins.

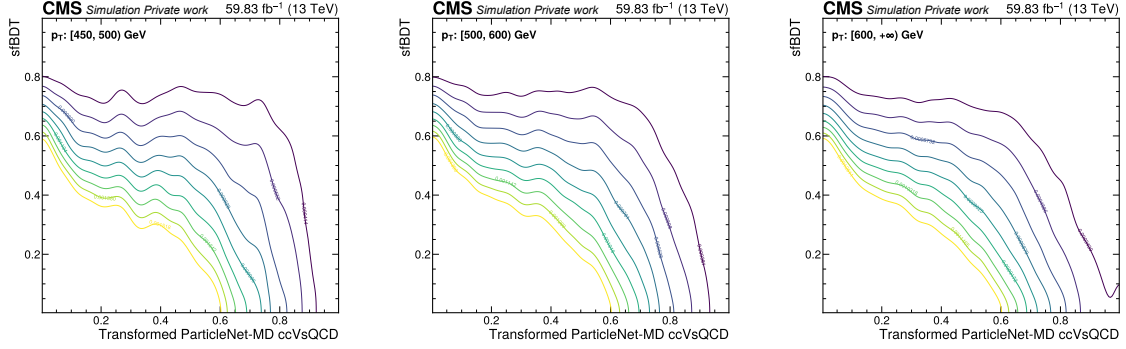


Figure 6.12 The sfBDT coastlines derived in the context of calibrating the ParticleNet-MD $X \rightarrow c\bar{c}$ discriminants in three exclusive p_T bins.

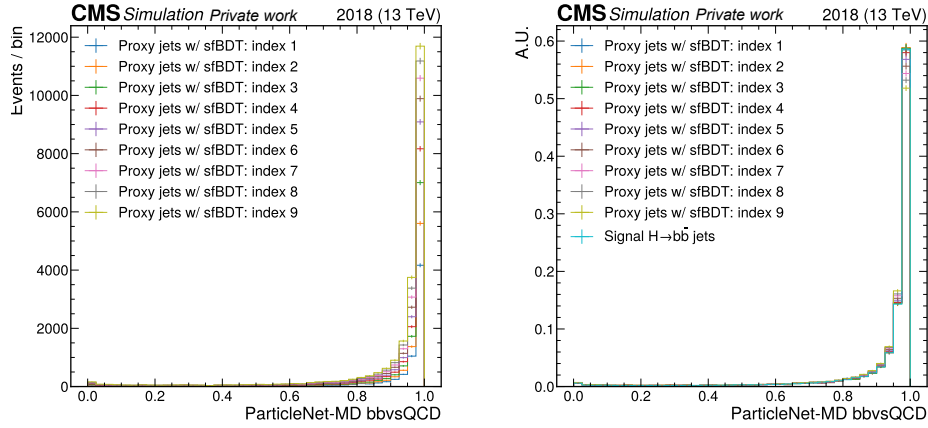


Figure 6.13 The distribution of the tagging discriminant of proxy jets with (left) and without (right) normalized to unity, in the case of calibrating ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant at HP and after applying the sfBDT coastline selection in nine options. In the normalized plot, the signal $H \rightarrow b\bar{b}$ jet distribution is also included for comparison.

them on a variety of jet observables to claim the proxy–signal similarity. We study a series of observables related to high-level properties of the jet, subjets, tracks, and SVs. Specifically, they include jet p_T , p_T for the leading and subleading subjets (sj_1 and sj_2 in the figure), m_{SD} , τ_{21} , ΔR between two subjets, the number of SVs, the number of tracks, and several SV-related features for the SV that is associated to the leading subjet, namely, the corrected mass, impact parameter d_{xy} , d_{xy} significance, and χ^2 over the number of degree of freedom in the SV fit. The histograms are shown in Fig. 6.14.

Comparing the signal jets (blue) with the proxy jets after the sfBDT selection (red) in Fig. 6.14, a very reasonable agreement is seen in most of the variables, except for a few cases, including m_{SD} , τ_{21} , and $\Delta R_{sj_1, sj_2}$ which we discuss later. Here, we should highlight that the

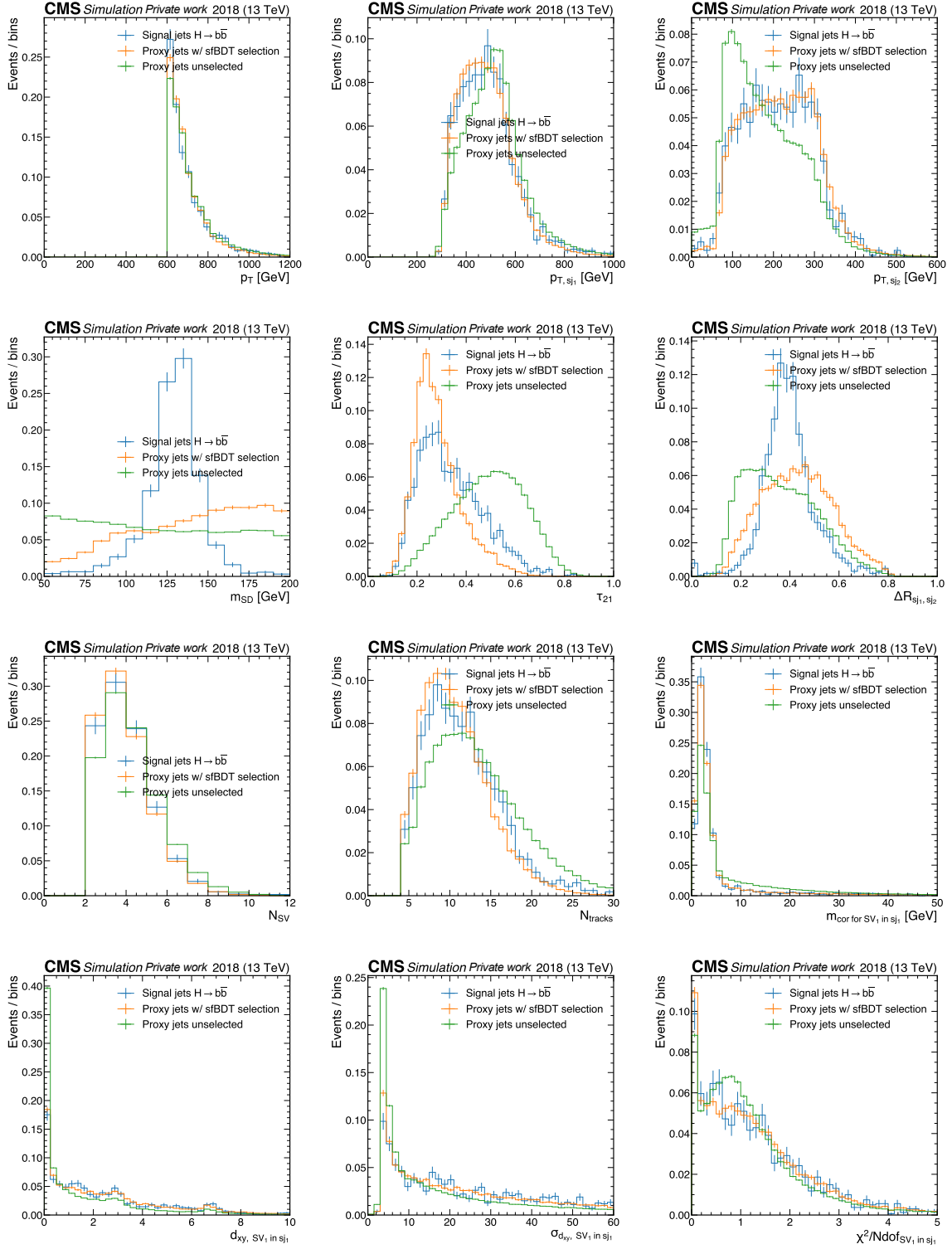


Figure 6.14 Comparison of the signal and proxy jets on a collection of jet observables. The observables are related to high-level properties of the jet, subjets, tracks, and SVs. See the detailed observable list in the text.

sfBDT is not trained to separate the signal $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets versus the QCD proxy jets but trained to separate a more signal-like part in the QCD jets from the rest. Hence, the proxy–signal similarity proves this spirit of BDT design is capable of isolating a subregion from QCD jets that represent the signal $H \rightarrow b\bar{b}$ ($c\bar{c}$) jets well.

For the residual discrepancy in m_{SD} , τ_{21} , and $\Delta R_{s_{j1}, s_{j2}}$, we first note that $\Delta R_{s_{j1}, s_{j2}}$ is correlated with m_{SD} . A mismatch on m_{SD} is expected, however, because the $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal jets always have a resonance peak structure, but surely not in the QCD proxy jets. The effect of proxy–signal discrepancy on m_{SD} and τ_{21} is then studied by performing an additional proxy-to-signal reweighting on either of the two variables on top of the current workflows to build the proxy jets. The additional reweighting is found to have a negligible impact on the calibration result, which addresses the concern.

6.3.4.3 The sfBDT selections for use

Each sfBDT selection will result in an individual calibration result in terms of the SF. The full suite of sfBDT selections is expected to sculpt the proxy jet shape in different ways so that all variations of the shape can cover the signal jet shape. However, selecting simply on those sfBDT coastlines cannot guarantee so because the sculpted proxy shape will just be the same as the signal shape according to the definition of the coastlines. Given this fact, we opt for different sfBDT coastlines for the regions that pass or fail the tagger discriminant selection. This allows us to separately control the level of proxy-to-signal similarity for the “pass” and “fail” regions. Specifically, the proxy jet selection defined by setting the i -th coastline in the “pass” region and the j -th coastline in the “fail” region is defined by: This way of selection is illustrated in Fig. 6.15.

```
selection = ((tagger_value in WP) && (contour_level < level[i]))
           || (!(tagger_value in WP) && (contour_level < level[j]))
```

Figure 6.15 illustrates the selection on the 2D tagger–sfBDT plane.

Therefore, with $i = 1, \dots, 9$ and $j = 1, \dots, 9$, 81 types of coastline selections are used for deriving the calibration result. They correspond to 81 proxy-jet collections with (a) different tightness of the sfBDT selection and (b) different similarity of proxy-to-signal in the “pass” and “fail” regions. Here, signal–proxy similarity can be showcased by an ROC curve for separating the signal and proxy. As a standard for comparison, the perfect proxy should be non-distinguishable from the signal jets; therefore, its ROC curve should be a random guess line. The ROC curves for the 81 selection conditions are shown in Fig. 6.16, and they exhibit

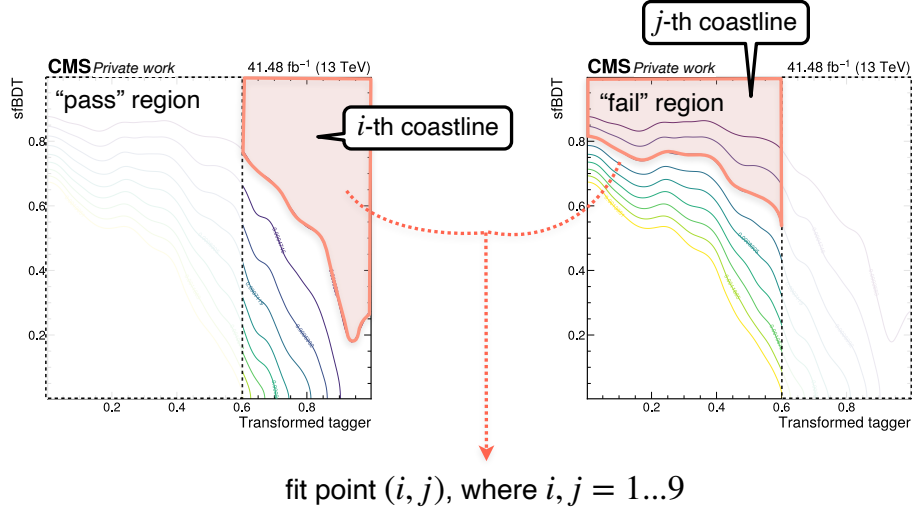


Figure 6.15 Illustration on the choices of “sfBDT coastline” selections applied for an individual fit.

varying degrees of fluctuation around the random-guess line. This illustrates that the 81 sfBDT selections create a variety of proxy phase spaces that are in different vicinity to the signal.

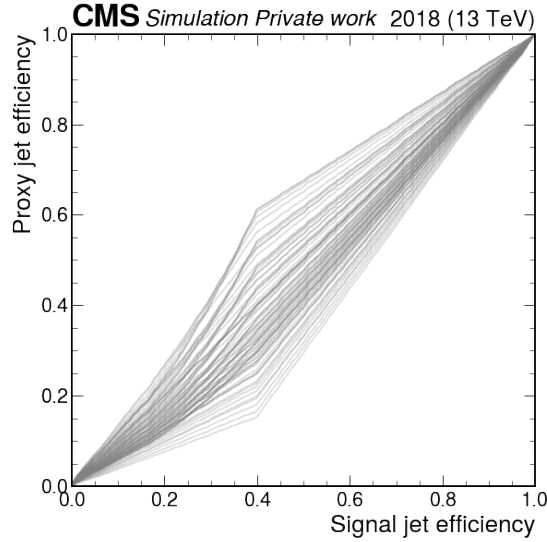


Figure 6.16 The separation power of signal and proxy jets illustrated in terms of the ROC curve, where the proxy jet collection is defined by a selection on the sfBDT coastline over all 81 possible options. The plot corresponds to the calibration of the ParticleNet-MD $X \rightarrow b\bar{b}$ discriminant at HP.

6.3.5 Data and simulation comparison

Data and simulation distributions are compared at two stages. First, after the MC-to-data reweighting, we compare the agreement between data and MC on basic jet kinematics variables and the sfBDT variable. Figure 6.17 shows the histograms in three exclusive jet p_T bins in the 2018 data-taking condition. All simulated processes are merged and then split into three categories: flvB, flvC, and flvL. The jet kinematics agree reasonably well, while the high-sfBDT-score region shows a slight deviation. To study the sfBDT mismodelling between data and MC, a specific uncertainty source is included when deriving the SF.

Figure 6.18 further shows the data-simulation comparison on the transformed ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants, still before applying sfBDT selections. As a comparison, the same distribution when the sfBDT selection is applied specifically on the leftmost, middle and rightmost coastline position is shown in Figs. 6.19 and 6.20.

From Figs. 6.19 and 6.20, one may notice that it is already possible to estimate the SF by looking at the data-simulation difference in the high tagging discriminant region because jets in these regions are dominated by the targeted flavour, i.e. jets with high $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminant values should be dominant by the flvB (flvC) type of jets. The following section introduces a more precise and reliable SF derivation approach based on the template fit.

6.4 Scale factor derivation

This section introduces details of the fit in the SF derivation.

6.4.1 Fit template and variable

The MC simulation has been divided into three flavour categories, flvB, flvC, and flvL. Each component is assigned a free-floating factor to scale its event yield in the region passing the tagging discriminant selection. The three factors are estimated through a template fit. For the calibration of the $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminant, the proxy jet collection is the flvB (flvC) proportion among all simulated jets. Therefore, the factor for the flvB (flvC) is taken as the targeted SF.

The SFs are derived individually for the four data-taking conditions, three regions of the $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminant that are inclusively defined, and three exclusive p_T bins. The regions correspond to the HP, HP, and LP region, defined by the discriminant larger than the corresponding WP defined in Table 6.4. The p_T ranges are (450, 500), (500, 600), and (600, ∞) GeV.

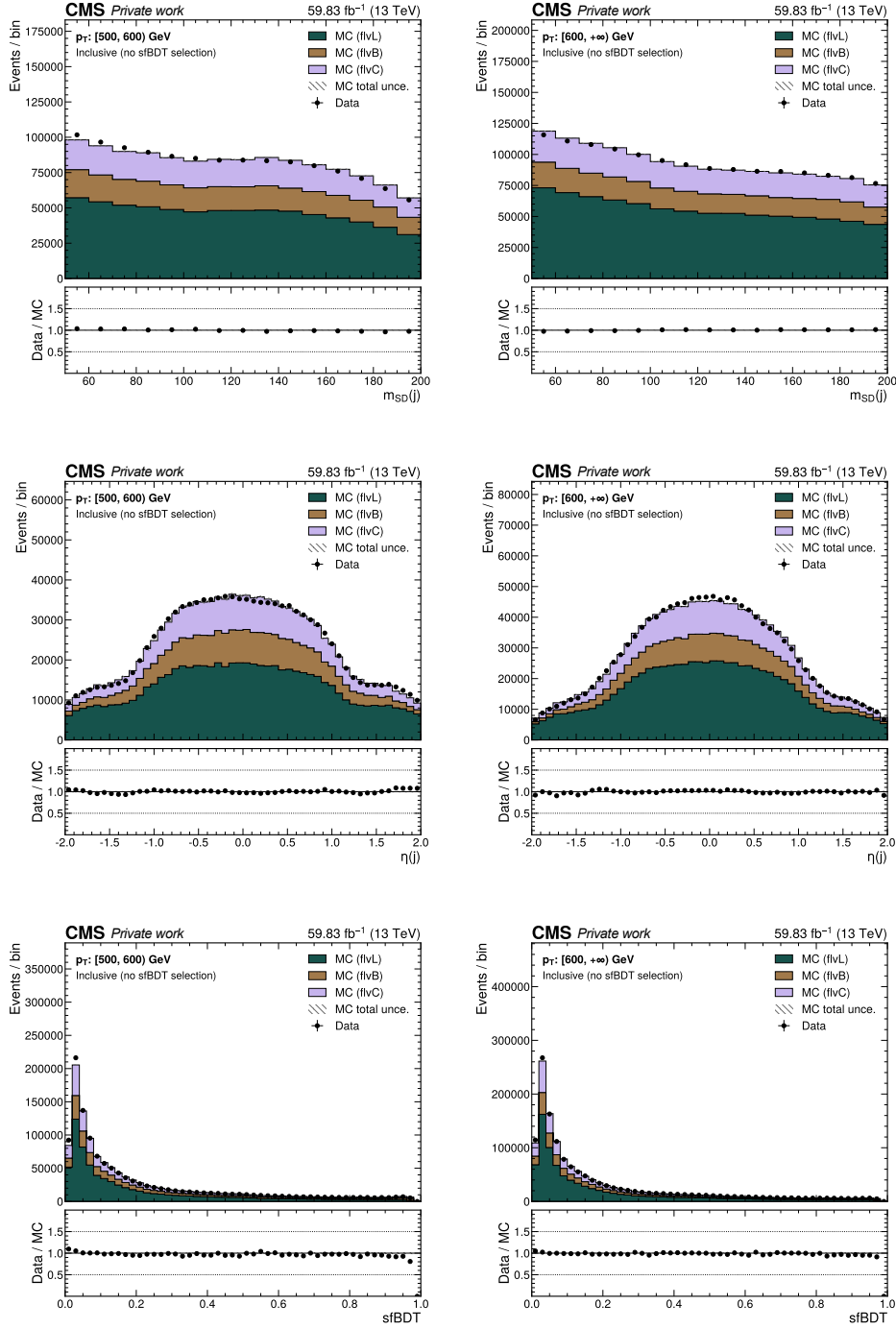


Figure 6.17 Data and MC comparison on various jet kinematics and sfBDT in 2018 data-taking condition, before the sfBDT selection is applied. The distributions for the jet variable m_{SD} , η , and sfBDT discriminant is shown, under two example p_T ranges.

The fit variable is chosen as $\log(M_{SV_1 \text{ for } \max(\sigma_{d_{xy}})}/\text{GeV})$, where the “ $SV_1 \text{ for } \max(\sigma_{d_{xy}})$ ” stands for the leading SV with maximum impact parameter d_{xy} significance. The corrected

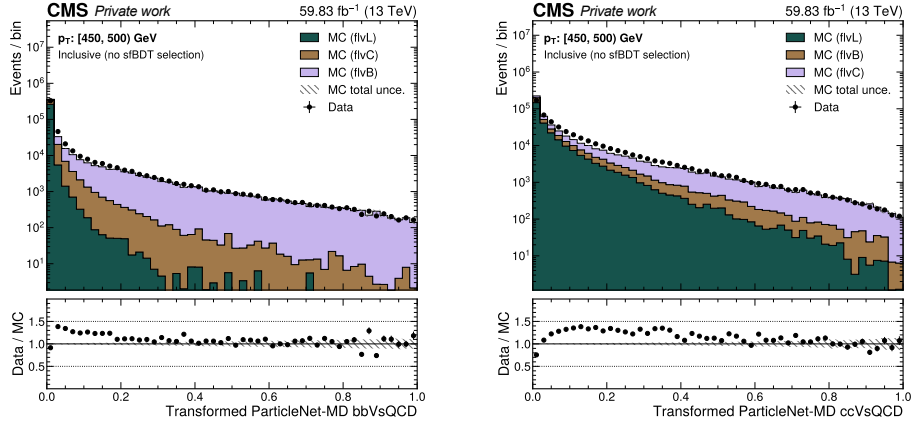


Figure 6.18 Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow b\bar{b}$ (left) and $X \rightarrow c\bar{c}$ (right) tagging discriminants in 2018 data-taking condition, before the sfBDT selection is applied.

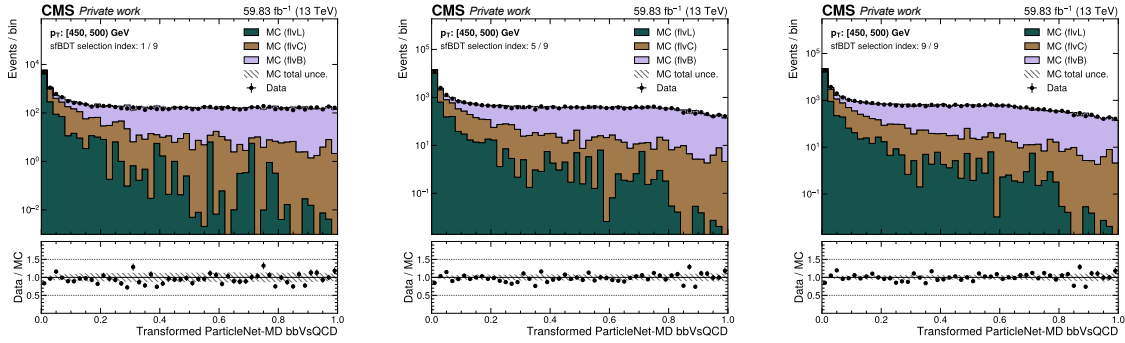


Figure 6.19 Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminants in 2018 data-taking condition, after the sfBDT coastline selection applied on the leftmost (left), the middle (middle), and the rightmost (right) positions.

SV mass is defined as^[75]

$$M_{SV}^{\text{corr}} = \sqrt{M_{SV}^2 + p^2 \sin^2 \theta} + p \sin \theta, \quad (6.6)$$

where M_{SV} is the invariant mass of the tracks associated with the SV, p is the SV momentum obtained from associated tracks, and θ the angle between the p and the vector pointing from the PV to the SV.

The motivation for using this variable is to separate the shape of the three flavour templates using native jet-related features, and here, we rely on the corrected mass of the SV, which manifests the types of heavy-flavour hadrons that initiate the SV. Specifically, the flvB template tends to peak at the B meson mass at around 5 GeV, and the flvC template peaks at the

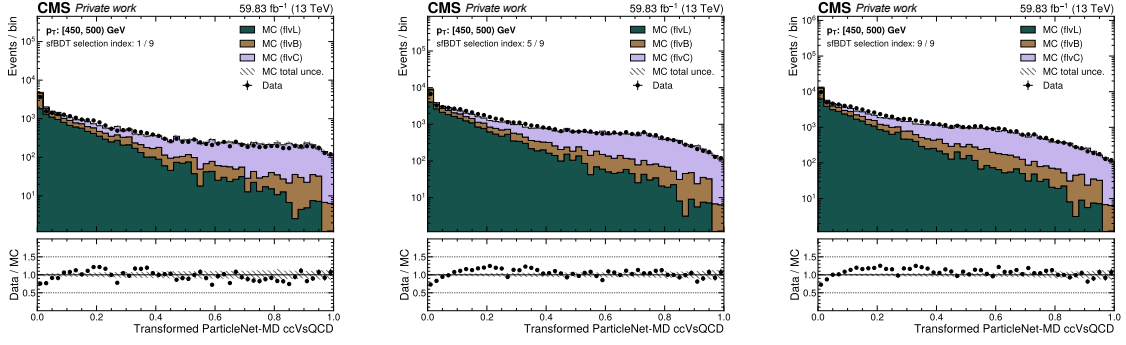


Figure 6.20 Data and MC comparison on the transformed ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminants in 2018 data-taking condition, after the sfBDT coastline selection applied on the leftmost (left), the middle (middle), and the rightmost (right) positions.

D meson mass at around 2 GeV. The spectrum of the corrected SV mass, however, has a long tail due to limited resolution. Therefore, the natural logarithm is taken, and the wide overflow and underflow bins are gathered to only highlight the fit region, where M_{SV_1} is around (1.0, 6.0) GeV. As an example, Fig. 6.21 shows the fit variable distribution of data and MC after passing the middle sfBDT coastline selection for both “pass” and “fail” regions. The plots are presented in three p_T ranges and made under the 2018 data-taking condition.

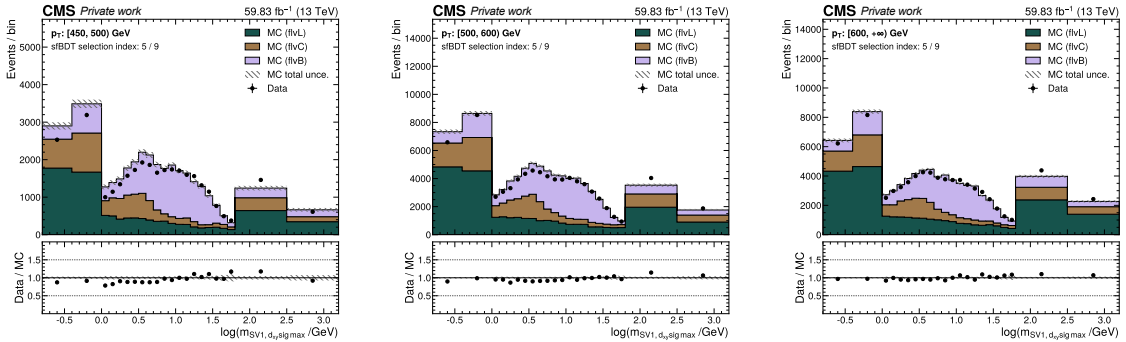


Figure 6.21 Data and MC comparison on the fit variable for the inclusive “pass”+“fail” region, passing the middle sfBDT coastline selection.

6.4.2 Fit strategy

For each SF derivation point, the “pass” and “fail” region is defined by jets passing or failing the specific tagging discriminant WP. The fit is done simultaneously for both the “pass” and “fail” regions. As mentioned, three data-to-MC SFs, one for each flavour template, are

allowed to float individually in each fit. The SFs are determined as

$$\text{SF}_i = \frac{\epsilon_{\text{Data},i}}{\epsilon_{\text{MC},i}}, \quad i \text{ stands for flvB, flvC, and flvL template.} \quad (6.7)$$

The three SF_i parameters are implemented as free-floating parameters, which changes the tagging efficiency while keeping the total yields (pass+fail) unchanged. The SF of the targeted flavour is constrained within the range $[0.1, 10]$, while the other two SFs are constrained within the range $[0.5, 2]$ due to the instability of constraining these SFs.

The three parameters are considered as the parameters of interest (POIs) and they are determined with other nuisance parameters that account for uncertainties in a maximum likelihood fit. The systematic uncertainties are summarized in Section 6.4.3. In addition, since there are 81 choices of selections introduced to deliver calibration at a given discriminant WP, p_T bin, and data condition, a postprocessing procedure is necessary to summarize all 81 SFs, incorporating their variations and forming the final SF. The procedure is detailed in Section 6.4.4.

6.4.3 Systematic uncertainties

The derivation of the SFs considers various systematic uncertainty sources. The theoretical and experimental uncertainties are discussed below.

Theoretical uncertainties

The theoretical uncertainties correct the potential mismodelling in simulation compared with data. For the simulated QCD multijet events, the uncertainty in the fractions of b -, c -, and light-flavour jets are considered. The proportion of heavy flavour content is largely dependent on the phenomenological model within the generator and can potentially lead to mismodelling of the flavour proportions. As a conservative estimate, the study introduces three uncertainty sources accounting for the fractions of b -, c -, and light-flavour jets on the mixture of all simulated events, whereas the ± 1 standard deviation corresponds to varying the yields for that flavour by 20%.

Besides, to account for the inaccuracy of modelling on initial- and final-state radiation (ISR and FSR) in the parton shower, the renormalization scale of QCD emissions in the ISR and FSR in the parton shower simulation is individually varied by 2 and 0.5.

Experimental uncertainty

The experimental uncertainty sources that participate in the fit as individual nuisance parameters are presented below.

- **Jet energy scale:** the uncertainty in the energy scale of large- R jet is considered and propagated to the fit template by varying with ± 1 standard deviation from its nominal value^[34, 144].
- **Jet energy resolution:** the large- R jet energies in the simulation are smeared according to a Gaussian function to accommodate the slightly worse resolution in data in the normal measurement of the SF. The uncertainty in the JER is propagated to the fit template by varying with ± 1 standard deviation of the Gaussian function by its uncertainty^[34, 144].
- **Integrated luminosity:** the integrated luminosity is varied by 1.2–2.5% in the 2016–2018 data-taking eras^[145-147].
- **Pileup reweighting:** the uncertainty in the pileup reweighting procedure is determined by varying the minimum bias cross section used to produce the pileup profile away from the measured central value by 5%^[148].

In addition to these uncertainties contributing to the change of fit templates, two external uncertainty sources are considered.

The first uncertainty accounts for the mismodelling of the sfBDT discriminant score and the fit variable. Increased uncertainty is needed to cover the difference between data and simulated events. A “reweighting” approach is used to account for the mismodelling. The SF derivation procedure is repeated in two additional schemes. In the first scheme, a reweighting on sfBDT discriminant score is performed from the entire simulated event to data before any selection on sfBDT is applied. In the second scheme, the fit variable is reweighted from simulation to data before the template is split into the “pass” and “fail” regions. This results in two additional sets of SFs. The nominal SF is assigned an external uncertainty, determined from the maximum deviation of the central SF values in the new sets from the nominal one. This source of uncertainty is found to be the dominant contributor to the final uncertainty.

The second uncertainty assesses the effect of varying the sfBDT selection thresholds. The dependence of the sfBDT selection on the calibrated SF is crucial to the validity of this approach. Therefore, the 81 choices of selections are adopted, each resulting in an SF, and the results are combined. Unlike the conventional combination of measured results delivered in the same condition, the 81 SF measurements originate from different conditions, which is expected to yield a larger uncertainty to the merged SF than that of the individual uncertainty. The next subsection illustrates the determination of the central value and error bar of the merged results.

6.4.4 Merging of individual measurements

The 81 measured SFs are merged based on their central value and upper and lower bounds in a statistical procedure. The new SF, along with its upper and lower bounds, is determined by randomly selecting one SF from a set of 81 and sampling SF values from the probability density function (PDF) defined by its central value and boundaries. This results in a new PDF with an infinite number of samples, and the central value and the upper (lower) bounds of the new PDF will be used to define the merged SF value and its uncertainty. Mathematically, the cumulative distribution function (CDF) of this process can be calculated from

$$F_{\text{SF}}(x) = \text{mean}_i \left[\Phi \left(\frac{x - c_i}{e_{i,\text{up}}} \eta(x) + \frac{x - c_i}{e_{i,\text{down}}} \eta(-x) \right) \right], \quad (6.8)$$

where x represents the variable for the SF value, c_i , $e_{i,\text{up(down)}}$ ($i = 1, \dots, 81$) denote the central value and the upper (lower) bound of the 81 measured SFs, and $\eta(\cdot)$ symbolises the step function. The final central value c , as well as $c + e_{\text{up}}$ and $c - e_{\text{down}}$ are derived from the CDF at 0 and $\pm 1 \sigma$ respectively.

6.5 Calibration results and discussions

First, to illustrate the SF measurement from a single fitting, Fig. 6.22 pre-fit and post-fit templates for the regions passing and failing the tagger selection. The illustration takes the measurement of ParticleNet $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminant SFs as an example, delivered in the 2018 data-taking conditions and under the tightest working point (HP). The fit displayed in the figure corresponds to applying the middle sfBDT coastline selection in both the “pass” and “fail” regions among the 81 choices. It can be observed that, with a tight $X \rightarrow b\bar{b}$ ($X \rightarrow c\bar{c}$) discriminant, the remaining events passing the tagger selection are almost entirely dominated by the flvB (flvC) subcomponent.

As a summary, the calibration results are derived in terms of SFs, based on seven tagging discriminants (four for $X \rightarrow b\bar{b}$ taggers and three for $X \rightarrow c\bar{c}$ taggers), in three working points (HP, MP, and LP), three exclusive p_{T} bins, and four data-taking eras. Tables 6.8 and 6.9 summarize the SFs for the ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging discriminant SFs. The calibration results for the rest of the five $X \rightarrow b\bar{b}$ or $X \rightarrow c\bar{c}$ discriminants are summarized in Appendix A. These SFs are part of the published results from the CMS experiment in Ref.^[124].

For better illustration, their SFs for ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants are summarized in a plot format in Figs. 6.24 and 6.25. The derived SF is around unity with an

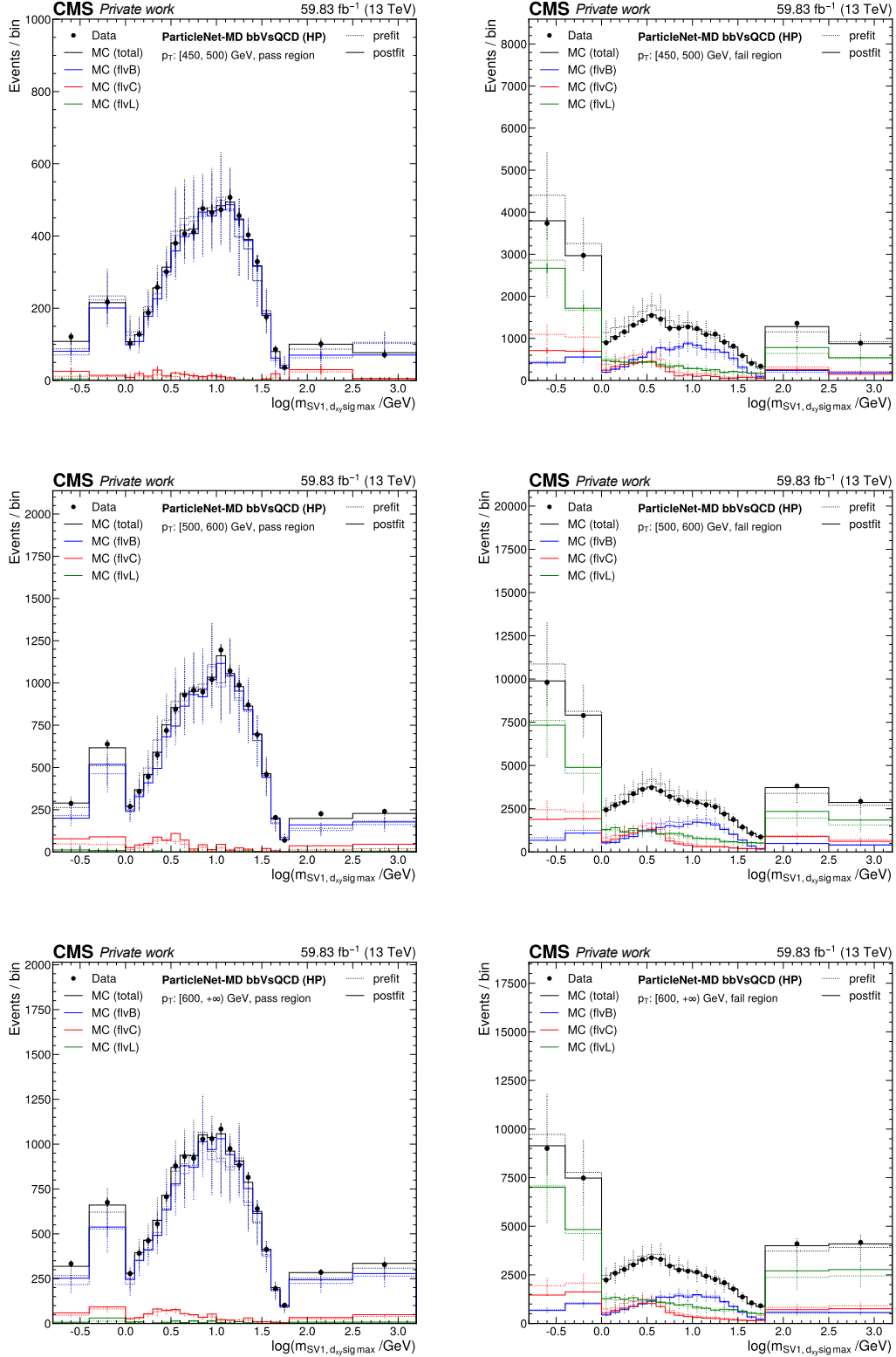


Figure 6.22 Pre-fit and post-fit distributions on the fit variable for the “pass” (left) and “fail” (right) region in the context of calibrating the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant at HP, under the 2018 data-taking conditions. The fit corresponds to applying the middle sfBDT coastline selection and is performed under three jet p_T ranges (from top to bottom).

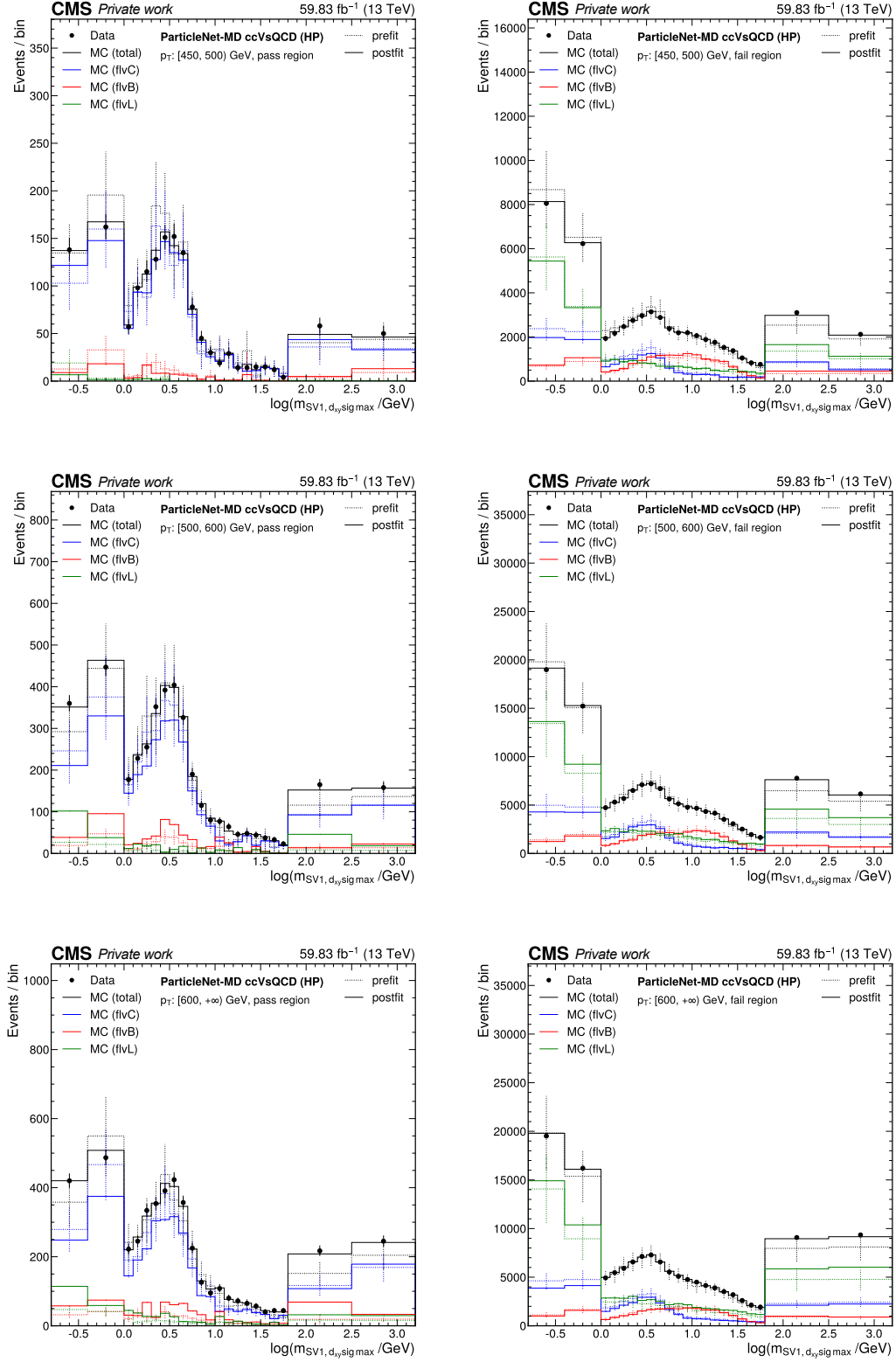


Figure 6.23 Pre-fit and post-fit distributions on the fit variable for the “pass” (left) and “fail” (right) region in the context of calibrating the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant at HP, under the 2018 data-taking conditions. The fit corresponds to applying the middle sFBDT coastline selection and is performed under three jet p_T ranges (from top to bottom).

uncertainty of around 4–8% (10–40%) for $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminants. No obvious discrepancy on the jet p_T is observed from the result.

By comparing the derived ParticleNet-MD $X \rightarrow c\bar{c}$ and DeepDoubleX $X \rightarrow c\bar{c}$ SFs with those presented in Refs.^[113, 114, 125] derived for all detector conditions in Run 2, reasonable agreement is observed regarding the central SFs, the SF dependence on the year of data taking, and the level of uncertainties. This provides a good validation of the sfBDT method.

Overall, SFs derived from the sfBDT method manifest robustness over the 252 set of results. This is the key feature of the sfBDT method, as it has been developed as a unified approach for deriving reliable SF across all detector conditions and tagger discriminants. It is applicable from the early tagger discriminant to the latest ones that have the strongest discrimination power between $X \rightarrow b\bar{b}$ ($c\bar{c}$) signal and backgrounds. In the CMS experiment, this method provides a justified and robust baseline.

Table 6.8 Summary of the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately. These SFs are part of the published results in Ref.^[124].

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.9883, 1.0)	$1.054^{+0.080}_{-0.077}$	$1.139^{+0.083}_{-0.081}$	$1.049^{+0.133}_{-0.130}$
MP: (0.9737, 1.0)	$1.052^{+0.087}_{-0.081}$	$1.068^{+0.078}_{-0.073}$	$0.996^{+0.101}_{-0.097}$
LP: (0.9088, 1.0)	$1.032^{+0.096}_{-0.090}$	$1.062^{+0.092}_{-0.082}$	$1.002^{+0.106}_{-0.101}$
2016 post-VFP			
HP: (0.9883, 1.0)	$1.031^{+0.050}_{-0.046}$	$1.055^{+0.069}_{-0.067}$	$1.088^{+0.076}_{-0.072}$
MP: (0.9735, 1.0)	$1.029^{+0.051}_{-0.045}$	$1.070^{+0.066}_{-0.062}$	$1.077^{+0.067}_{-0.059}$
LP: (0.9137, 1.0)	$1.031^{+0.058}_{-0.050}$	$1.089^{+0.076}_{-0.068}$	$1.057^{+0.077}_{-0.056}$
2017			
HP: (0.987, 1.0)	$1.055^{+0.057}_{-0.054}$	$1.067^{+0.057}_{-0.055}$	$1.045^{+0.045}_{-0.046}$
MP: (0.9714, 1.0)	$1.006^{+0.052}_{-0.052}$	$1.051^{+0.056}_{-0.055}$	$0.991^{+0.038}_{-0.043}$
LP: (0.9105, 1.0)	$0.966^{+0.055}_{-0.057}$	$1.021^{+0.053}_{-0.052}$	$0.979^{+0.035}_{-0.038}$
2018			
HP: (0.988, 1.0)	$0.994^{+0.064}_{-0.064}$	$1.072^{+0.041}_{-0.036}$	$1.046^{+0.038}_{-0.038}$
MP: (0.9734, 1.0)	$0.966^{+0.056}_{-0.057}$	$1.033^{+0.030}_{-0.025}$	$1.010^{+0.030}_{-0.035}$
LP: (0.9172, 1.0)	$0.921^{+0.071}_{-0.077}$	$1.006^{+0.024}_{-0.026}$	$1.001^{+0.035}_{-0.037}$

Table 6.9 Summary of the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately. These SFs are part of the published results in Ref.^[124].

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.9909, 1.0)	$1.375^{+0.244}_{-0.261}$	$1.160^{+0.191}_{-0.163}$	$1.032^{+0.273}_{-0.217}$
MP: (0.9751, 1.0)	$1.252^{+0.191}_{-0.184}$	$0.923^{+0.104}_{-0.098}$	$1.218^{+0.218}_{-0.224}$
LP: (0.9252, 1.0)	$1.168^{+0.230}_{-0.219}$	$0.867^{+0.120}_{-0.089}$	$1.220^{+0.241}_{-0.230}$
2016 post-VFP			
HP: (0.9905, 1.0)	$1.122^{+0.218}_{-0.180}$	$1.127^{+0.279}_{-0.193}$	$1.255^{+0.197}_{-0.184}$
MP: (0.9743, 1.0)	$1.005^{+0.182}_{-0.157}$	$1.130^{+0.185}_{-0.195}$	$0.982^{+0.181}_{-0.148}$
LP: (0.9252, 1.0)	$1.147^{+0.152}_{-0.168}$	$1.069^{+0.192}_{-0.186}$	$1.097^{+0.233}_{-0.271}$
2017			
HP: (0.9909, 1.0)	$1.268^{+0.383}_{-0.357}$	$1.385^{+0.344}_{-0.334}$	$1.225^{+0.209}_{-0.200}$
MP: (0.9765, 1.0)	$1.464^{+0.426}_{-0.422}$	$1.198^{+0.268}_{-0.262}$	$1.203^{+0.230}_{-0.227}$
LP: (0.9347, 1.0)	$1.339^{+0.341}_{-0.333}$	$1.307^{+0.334}_{-0.331}$	$1.126^{+0.210}_{-0.204}$
2018			
HP: (0.9917, 1.0)	$1.204^{+0.334}_{-0.290}$	$1.076^{+0.201}_{-0.177}$	$1.098^{+0.180}_{-0.175}$
MP: (0.9777, 1.0)	$1.098^{+0.234}_{-0.188}$	$1.003^{+0.131}_{-0.119}$	$1.031^{+0.126}_{-0.107}$
LP: (0.9368, 1.0)	$0.958^{+0.189}_{-0.108}$	$0.962^{+0.091}_{-0.086}$	$0.938^{+0.110}_{-0.087}$

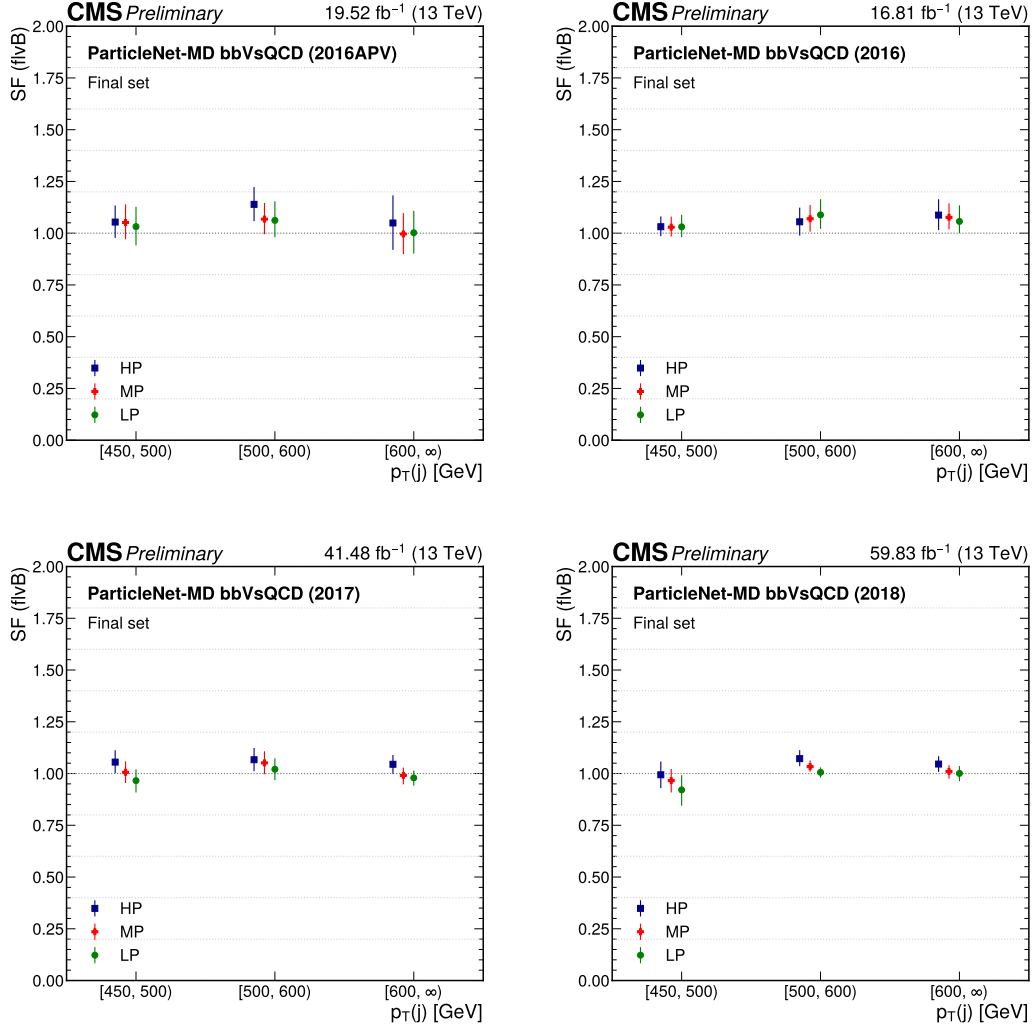


Figure 6.24 Summary of the ParticleNet-MD $X \rightarrow b\bar{b}$ tagging discriminant SFs in the plot format, separately shown for four data-taking conditions. The displayed results are excerpted from those published in Ref.^[124].

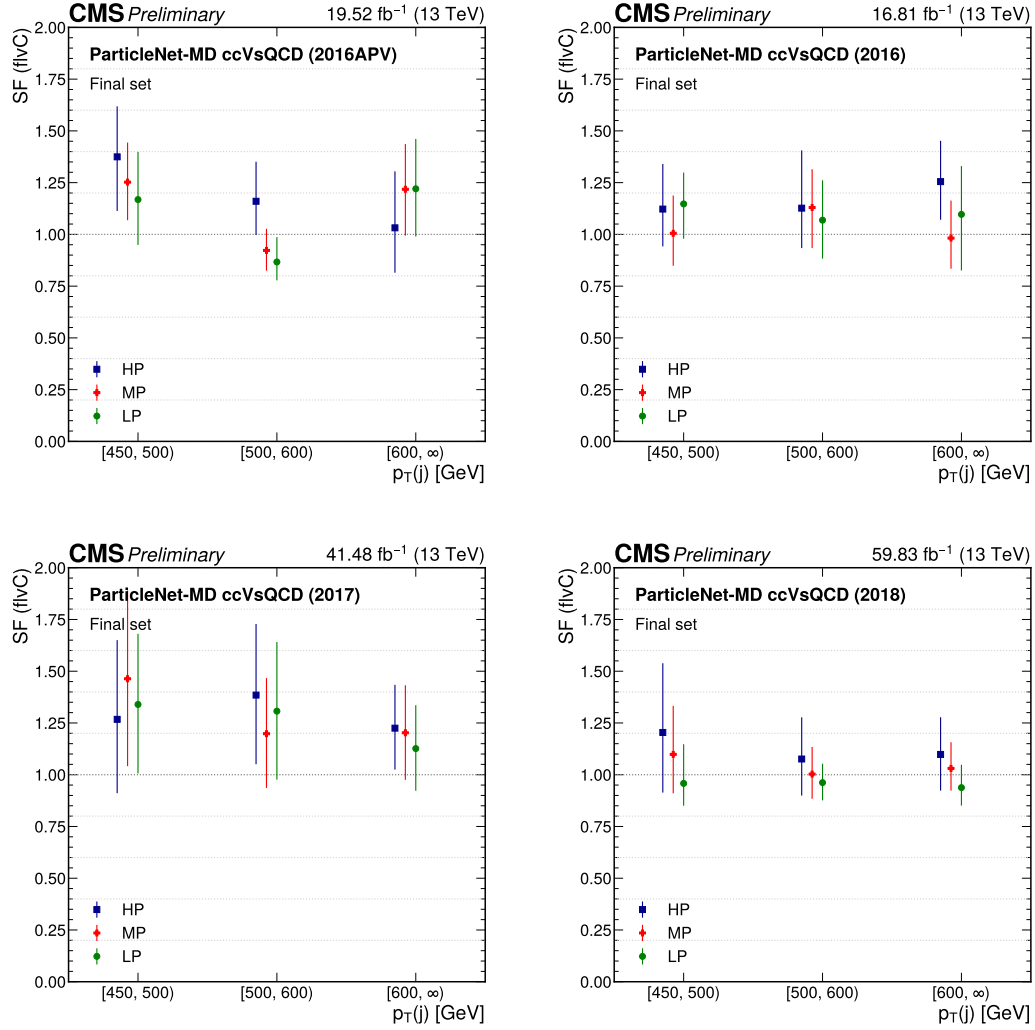


Figure 6.25 Summary of the ParticleNet-MD $X \rightarrow c\bar{c}$ tagging discriminant SFs in the plot format, separately shown for four data-taking conditions. The displayed results are excerpted from those published in Ref.^[124].

Part IV

Application in Higgs-Charms Coupling Measurement

Chapter 7

Introduction to Higgs-Charm Coupling Measurement

7.1 Physics motivation

Since the discovery of the Higgs boson at the ATLAS and the CMS experiment^[149, 150] in 2012, the property measurement of the Higgs boson has been at the core of the LHC physics programmes. One of the most important research targets is to validate whether the Higgs boson couples to the fermion fields with a strength proportional to the corresponding fermion mass, a fundamental feature of Yukawa couplings (see the theoretical background in Section 1.1.3). By far, the ATLAS and CMS experiments have announced the observation of Higgs bosons coupling to third-generation quarks (t and b quarks)^[151-154], and the tau leptons^[155]. Therefore, the next milestone lies in the study of Higgs bosons coupling to the second-generation fermions, including the c and s quarks and muons. The measurement of Higgs bosons coupling to muons has proceeded rapidly, and by far, evidence has been found in the Higgs boson decaying to a $\mu^+\mu^-$ pair^[156]. Figure 7.1 shows the measured reduced Higgs-coupling modifier for fermions and weak bosons in the CMS experiment, indicating that the several Yukawa coupling strengths have been well examined to be proportional to the fermion mass. Meanwhile, it is crucial to push the progress in measuring the coupling of the Higgs boson to the charm quark, an ambitious and challenging subject for the Higgs physics programmes. From the theoretical perspectives, several BSM theories suggest that the Higgs-charm Yukawa coupling can be significantly modified^[157-160]; hence, measuring this coupling constant and validate it with the SM prediction is crucial to finding evidence of new physics.

At the LHC experiments, a direct measurement of the coupling strength is done by choosing a dedicated physics process, where our target interaction between quantum fields plays a role. In processes dominated by the Higgs boson Yukawa coupling to the charm quark, one of the most direct process is the Higgs boson decaying to a $c\bar{c}$ pair (measurements via other channels are introduced in Section 7.2). This is similar to the measurement of the Higgs-bottom coupling done by exploiting the $H \rightarrow b\bar{b}$ decay process, but not to the measurement of Higgs

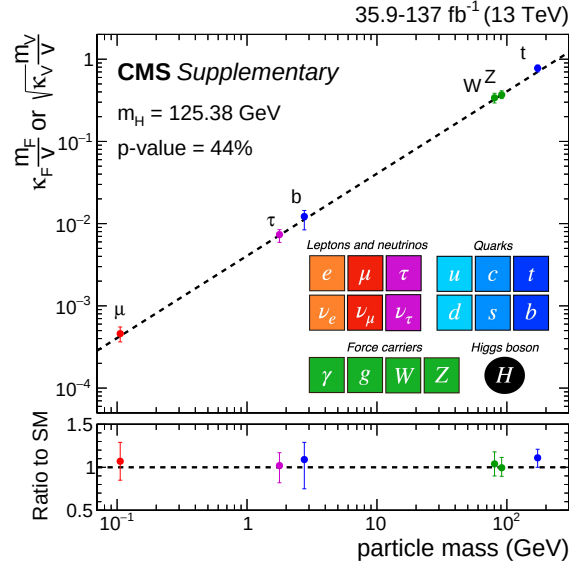


Figure 7.1 The measurement for the reduced Higgs-coupling modifiers extracted for fermions and weak bosons^[156].

boson and top quarks coupling because the heavy top quark mass prevents the process from happening. In the case of $H \rightarrow c\bar{c}$ or $H \rightarrow b\bar{b}$, the measured quantity in the analysis is the signal strength μ of the signal process, defined as the measured cross section of the target process over its SM prediction. The measurement on μ can be converted to the measurement on the charm Yukawa coupling strength, where the directly measured property is the coupling strength modifier κ_c , defined as the ratio of the measured coupling strength over its SM prediction, i.e. $y_c/y_{c,\text{SM}}$.

Despite the clear physics picture in delivering a $H \rightarrow c\bar{c}$ measurement, there are three main challenges to searching for this process at the LHC. First, the identification of the $c\bar{c}$ final state process is difficult due to the intermediate property of jets initiated from a c quarks between b quark and the light quark. Second, the $H \rightarrow c\bar{c}$ process final state is purely hadronic. Hence, its search is especially challenging at the LHC as vast contamination from the QCD multijet process exists. Third, the SM-predicted branching ratio of the $H \rightarrow c\bar{c}$ decay mode is around 2.9%, which is very small (around a factor of 20 smaller than the $H \rightarrow b\bar{b}$ ^[17]) because the c quark mass is much smaller compared to b quark mass in the SM.

These inherent challenges have induced several practical solutions to improve the $H \rightarrow c\bar{c}$ measurement sensitivity. To address the second challenge, the measurement presented in this chapter targets an associated production of the Higgs boson decaying to $c\bar{c}$ with a vector boson, W or Z , with their leptonic decay, which can be used for triggering the event and veto majority of the QCD multijet background. This is commonly recognized as the most sensitive channel

to search for $H \rightarrow c\bar{c}$, abbreviated as $VH \rightarrow c\bar{c}$, in the following text. The Feynman diagrams that initiate the $VH \rightarrow c\bar{c}$ process at the leading order are shown in Fig. 7.2. The first challenge implies that enhanced identification techniques, improving the selection efficiency of the $H \rightarrow c\bar{c}$ signal process while largely vetoing the QCD multijet background, can be especially helpful to this analysis. This is where the modern deep learning method comes into play and releases a significant potential in search of $H \rightarrow c\bar{c}$ at the LHC, which we will highlight in Section 7.3.

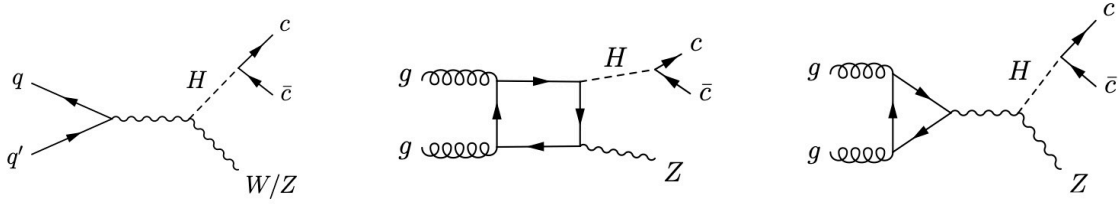


Figure 7.2 The Feynman diagrams of the $VH \rightarrow c\bar{c}$ process at the leading order.

7.2 Overview of relevant measurements

Both the ATLAS and CMS experiments have measured the Higgs-charm coupling via the $VH \rightarrow c\bar{c}$ in the early stage, using partial Run-2 data taken in 2016. The ATLAS analysis targeted the ZH ($Z \rightarrow \ell\ell$) channel, obtaining an observed (expected) upper limit on the signal strength μ at the 95% CL of 110 (150)^[161]. The CMS analysis, on the other hand, exploited multichannels based on the leptonic decay of both W and Z bosons, resulting in tighter constraints on μ corresponding to an observed (expected) upper limit on μ at the 95% CL of 70 (37)^[112].

Recently, the analysis was delivered by ATLAS using the full Run 2 data, corresponding to 139 fb^{-1} of data, exploiting multiple leptonic channels of the W and Z decay, resulting in an upgrade on the observed (expected) upper limit on μ at the 95% CL of 26 (31)^[162]. This corresponds to an observed (expected) constraint on the charm Yukawa coupling modifier κ_c at $|\kappa| < 8.5$ (12.4).

The CMS experiment also explores the search for $H \rightarrow c\bar{c}$ in the ggF Higgs boson production mode instead of the VH mode^[114]. Taking advantage of the recent advanced large- R jet tagger to identify $X \rightarrow c\bar{c}$ signature, the analysis yields competing results on the upper limit μ of the ggF $H \rightarrow c\bar{c}$ signal as 47 (39) at the 95% CL, serving as an important complement to the $VH \rightarrow c\bar{c}$ search.

Besides the conventional $H \rightarrow c\bar{c}$ channel, ATLAS, CMS, and early LHCb experiments also explored the radiative decay of the Higgs boson into charmonium states such as J/ψ . For the Higgs boson decay process to $J/\psi\gamma$, the branching ratio is at the 10^{-6} level, but its width is dependent on κ_c , providing a mean to extract the constraints on the charm Yukawa coupling. Using the partial Run-2 data collected in 2016 (corresponding to around 36 fb^{-1} of data), the measurements yield an observed (expected) upper limit of the measured branching ratio over the SM-predicted value at 95% CL as 260 (170), for the CMS measurement^[163], and around 120 (100), for the ATLAS measurement^[164].

The H - c coupling can also be measured indirectly from the Higgs boson production differential cross section as a function of transverse momentum^[165]. Specifically, the varying κ_b or κ_c constant modifies the quark-initiated Higgs contributions (denoted qqH) initiated by the b or c quarks and also the contributions of the b or c quark within the loop-induced ggF process. This allows the p_T distribution of the Higgs boson to change its shape and normalization. Using the full Run-2 data and exploiting the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^*$ channels, ATLAS and CMS have delivered the differential cross section measurement and set boundaries to κ_c ^[166, 167]. Using both shape and normalization of the p_T spectrum as a constraint, the observed (expected) exclusion limit of κ_c at 95% CL is -2.5 (-3.0) $< \kappa_c < 2.3$ (3.1), for the ATLAS results using the $H \rightarrow \gamma\gamma$ channel^[166], and -5.3 (-5.7) $< \kappa_c < 5.2$ (5.7), for the CMS results using $H \rightarrow ZZ^* \rightarrow 4\ell$ channel^[167]. Figure 7.3 shows the CMS result in the indirect simultaneous constraints on κ_b and κ_c under such conditions.

There are also phenomenological proposals to constrain the charm Yukawa coupling in other channels, e.g. in the Higgs boson production associated with a single c quark^[168], or utilizing the asymmetry of the W^+H and W^-H production^[169]. The H - c Yukawa coupling vertex exists in the diagrams that dominate these processes, which provides a handle for measurements. Experimental realizations of these proposals are future aspects to explore.

7.3 Implications of deep learning

The main challenge of the $H \rightarrow c\bar{c}$ search lies in the difficulty in accurately identifying the complex hadronic final states initiated by c quarks from the vast background events; this implies a complex classification task and thus enters the territory of deep learning.

For a decade, it has been known that supervised machine learning methods are experts at the classification problem, especially since the input features are multi-dimensional. The techniques are often referred to as multivariate analysis approaches. However, major advances

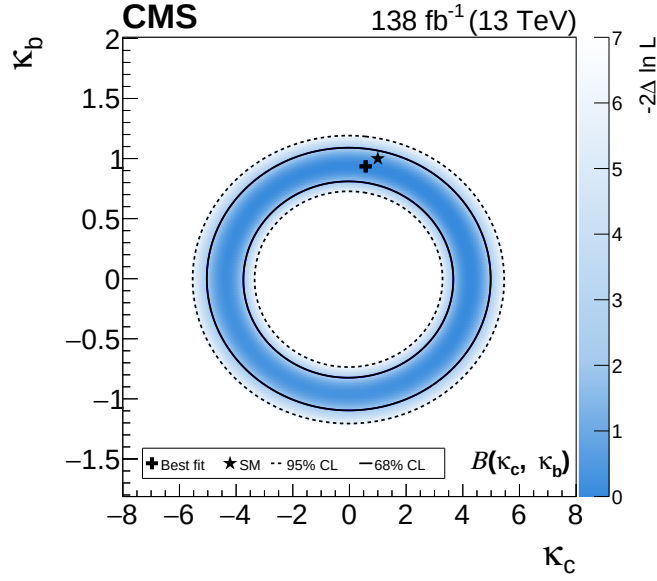


Figure 7.3 The simultaneous fit result of κ_b and κ_c , which are indirectly constraint by the shape and normalization of the differential Higgs boson cross section on p_T ^[167].

in modern deep learning include the ability to deal with data with more complexity and deeper hierarchy. In particular, the more complex the data, the better one can proceed with the means of DNNs. It should be surely noted that this is often accompanied by larger training data size, more sophisticated DNN architecture and training setup, as the deep learning community has revealed (discussed in Section 3.3), but when those advancements are incorporated, the power of DNN in this realm is incomparable.

Therefore, to fully utilize the power of DNN to release the potential of this classification problem, one sets the sight on the large- R jet object as it contains more particles (~ 3 times more than constituents contained in a small- R jet) and the object reveal more complexity that the neural network can explore. The jet data also reveals a hierarchical structure as it is represented by the features of its constituent particles or other associated objects. This implies the potential of using advanced DNNs to tag the large- R jets that cluster all particles as $c\bar{c}$ hadronization products, the case when the searched Higgs boson is highly Lorentz-boosted.

Chapter 8

Analysis Method

In this chapter, we introduce the physics search of the Lorentz-boosted $H \rightarrow c\bar{c}$ in association with a vector boson in the CMS experiment, using the full Run-2 data. We will specifically highlight the development and usage of the advanced large- R jet tagger and its mass regression model based on ParticleNet. This is the main driver of the progress of this search compared to its earlier version and its counterpart searches.

A streamlined introduction to the full picture of the analysis is presented as follows. In search of $VH \rightarrow c\bar{c}$ events where V decays leptonically, this analysis exploits three orthogonal categories defined by the lepton multiplicity, i.e. the 2-lepton (2L) channel for identifying ZH , $Z \rightarrow \ell\ell$, the 1-lepton (1L) channel for identifying WH , $W \rightarrow \ell\nu$, and the 0-lepton (0L) channel for identifying ZH , $Z \rightarrow \nu\bar{\nu}$. The 1L and 2L channels are further divided into the e category and μ category, based on the flavour of leptons. Requirement on the leptons or the missing transverse momentum (p_T^{miss}) is imposed for event triggering and for vetoing the backgrounds. The Higgs boson is constructed by selecting the leading large- R jet with $R = 1.5$ (AK15 jet). It is required to pass high ParticleNet $X \rightarrow c\bar{c}$ discriminant thresholds as a crucial step to suppress the backgrounds. An event BDT is trained to mildly separate $VH \rightarrow c\bar{c}$ and all types of backgrounds using the event kinematical features. The BDT is designed to decouple with any flavour or the mass property of the jet. Finally, a maximum likelihood template fit is performed to extract the signal strength on the binned variable of regressed AK15 jet mass, which is again trained with ParticleNet. Event control regions are defined by inverting the BDT selection and basic event selections to provide a better background event estimate during the fit. As a validation, the same workflow is performed to extract the $VZ \rightarrow c\bar{c}$ signature and compare it with the SM predictions.

This analysis documented in this chapter is denoted by the “merged-jet” analysis, as a contrarary to the “resolved-jet” analysis in which the Higgs boson is reconstructed by two separate AK4 jets. The two analyses are made orthogonal on their event phase spaces, and the results are combined to set a most stringent limit on κ_c .

8.1 Data and simulated samples

This analysis uses the proton-proton collision data at $\sqrt{s} = 13$ TeV collected from 2016 to 2018 at the LHC Run 2 period. The selected data corresponds to an integrated luminosity of 138 fb^{-1} .

In the 0L channel, data are collected with the HLTs related to p_T^{miss} . In 1L and 2L channels, the HLTs requiring a single electron or a single muon are selected.

For the signal process, the WH and quark-induced ZH signal processes are generated at QCD NLO order accuracy using the POWHEG v2^[26, 27, 129, 130] event generator extended with the multi-scale improved NLO (MiNLO) procedure^[170], and the gluon-induced ZH process is generated at LO with POWHEG v2. The production cross sections of the signal processes are corrected as a function of $p_T(V)$ to next-to-next-to-leading order (NNLO) QCD + NLO EW accuracy combining the VHNNLO^[171, 172], VH@NNLO^[173, 174], and HAWK v2.0^[175] generators.

The V +jets background samples are generated with MG v2.6.0^[25] at NLO with up to two additional partons. The top quark pair ($t\bar{t}$) and single top quark production processes^[132, 133, 135] are generated to NLO accuracy with POWHEG v2. The production cross sections for the $t\bar{t}$ samples are scaled to the NNLO prediction with the next-to-next-to-leading logarithmic resummation result obtained from TOP++ v2.0^[131], and the differential cross sections as a function of top quark p_T are corrected to the NNLO QCD + NLO EW prediction^[176]. Diboson backgrounds are generated at NLO with POWHEG v2 (MG v2.4.2) for the WW (WZ and ZZ) process. Production cross sections of the diboson processes are reweighted as a function of the subleading vector boson p_T to NNLO QCD + NLO EW accuracy^[177].

8.2 Jet tagging and mass regression techniques

8.2.1 The ParticleNet-MD tagger for AK15 jets

The analysis uses a customized version of the ParticleNet-MD tagger, trained on the AK15 jets. A brief description of the ParticleNet-MD tagger development can be found in Section 5.1, and more specific details on the sample production and input features will be the focus of this section. It is also worth noting that the training of the GloParT model, which will be discussed later, shares many similarities in terms of development. One can refer to Section 11.2 for more details on the deep learning specifics.

Training samples and jet labels

The training samples are simulated datasets composed of the QCD multijet process and a dedicated diresonance production process to acquire the resonance jets, including $X \rightarrow c\bar{c}$. Samples are produced under both 2017 and 2018 detector conditions. The QCD multijet sample is generated in exclusive p_T bins in a $2 \rightarrow 2$ QCD process, simulated by PYTHIA v8.230. The diresonance production is initiated by the $\tilde{G} \rightarrow HH$ process with the masses of $\tilde{G}$ and the scalar H varied across a wide range. Here, $\tilde{G}$ is a graviton defined in a BSM model; the specific process is less important because our aim is to collect two large- R jets, each initiated by a H resonance. Specifically, $m_{\tilde{G}}$ is specified within 600–6000 GeV with a step size of 100 GeV, and m_H is set to 15–250 GeV: within the range of 15–30 GeV, the step size is 5 GeV, and within 30–250 GeV, the step size is 10 GeV. The produced jet initiated from H thus spans a wide range in p_T and mass, as the p_T is roughly dependent on $m_{\tilde{G}}$ and its mass on m_H . The H decay is processed in PYTHIA with the default two-body decay configuration. The decay modes only cover $b\bar{b}$, $c\bar{c}$, $s\bar{s}$, $u\bar{u}$, and $d\bar{d}$, and their production rate is set as $1 : 1 : \frac{1}{3} : \frac{1}{3} : \frac{1}{3}$.

The large- R jets initiated from these processes are assigned a dedicated jet label used for training. Jets from the QCD process are assigned one label exclusively from the five categories: “bb”, “b”, “cc”, “c”, “others”, following the same truth-matching definition given in Table 6.6. Jets from the diresonance production process are assigned one of the $X \rightarrow b\bar{b}$, $X \rightarrow c\bar{c}$, and $X \rightarrow q\bar{q}$ categories based on the H decay mode and pass the requirement that both two generator-level quarks are matched within the jet axis by $\Delta R < 1.5$. Jets are discarded if they fail to meet these matching criteria.

Sample reweighting

When proceeding to the training, jets from the $X \rightarrow b\bar{b}$, $X \rightarrow c\bar{c}$, $X \rightarrow q\bar{q}$, and the inclusive QCD categories are sampled according to a predetermined probability binned by p_T and m_{SD} . These probabilities are calculated such that all the above categories result in uniform distributions in the 2D binning, and the total jet counts in the four categories are equal. This is referred to as a “reweighting” procedure, although the weight of each selected jet is not reflected in an actual weight value in training, but in the probability for sampling jets to form the training batch. It guarantees that the resulting tagger discriminant is decorrelated with the p_T and m_{SD} . In other words, a selection on the tagger discriminant will not sculpt the background event m_{SD} distribution.

Training

The inputs to the network include features carried on the constituent PF candidates and the SVs that are ΔR -matched to the jet. A summary of variables used for training can be found in Table 8.1. The final $X \rightarrow c\bar{c}$ tagging discriminant is defined as

$$\text{ParticleNet-MD ccvsQCD disc.} = \frac{p(X \rightarrow c\bar{c})}{p(X \rightarrow c\bar{c}) + p(\text{QCD})}, \quad (8.1)$$

where $p(\text{QCD})$ is the sum of scores for five QCD subcategories.

For the AK15 ParticleNet-MD tagger developed for this analysis, an ensemble of three networks, each trained with the same setup but with different weight initialization, is used. The output score is taken as the mean of the corresponding output node, leading to a $\sim 10\text{--}15\%$ improvement in the signal efficiency compared to a single training. The training specifics are adapted from Refs.^[65, 90] where a detailed description can be found later in Section 11.2 in the training of the GloParT model.

Performance and working points

The performance of the AK15 ParticleNet-MD tagger for identifying $H \rightarrow c\bar{c}$ jets generated from the SM Higgs boson production sample, using the $X \rightarrow c\bar{c}$ discriminant defined above, is displayed in Fig. 8.1 in terms of the ROC curve. Compared with the DeepAK15-MD $X \rightarrow c\bar{c}$ tagging discriminant used for the early $VH \rightarrow c\bar{c}$ analysis using 2016 Run-2 data, the ParticleNet algorithm achieves almost a factor of 5 times reduction in misidentification rates for V +jets and $H \rightarrow b\bar{b}$ backgrounds.

Defined on the tagger discriminant, the analysis selects three exclusive regions with different purity of $X \rightarrow c\bar{c}$ signal jets, as shown in Table 8.2.

Calibration

The calibration of the AK15 ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant in data is based on the sfBDT method, which uses $g \rightarrow c\bar{c}$ jets selected from the QCD multijet events with additional BDT selection to represent a proxy of the $H \rightarrow c\bar{c}$ signal. The early development of the method is used^[125], where the BDT is trained to separate two classes of QCD jets, distinguished by the parton-level variables, κ_g (see Section 6.3.3.1). Besides, the sfBDT selections are determined using a constant threshold on the sfBDT discriminant score, which is explained in Section 6.3.4 as the early adopted method and described by the dashed straight line in Fig. 6.8.

In addition, this analysis adopts exclusive working points. The derived $X \rightarrow c\bar{c}$ discriminant SFs for analysis are summarized in Table 8.3. They are applied to the $H \rightarrow c\bar{c}$ signal and

Table 8.1 The input list of variables for the ParticleNet-MD AK15 tagger training, including features of each jet constituent PF candidate and SVs. The track-based variables are only defined for charged PF candidates, and a value of 0 is assigned to neutral candidates.

Variable	Definition
For particle-flow candidates	
$\log p_T$	logarithm of the particle's p_T
$\log E$	logarithm of the particle's energy
$\Delta\eta(\text{jet})$	difference in pseudorapidity between the particle and the jet axis
$\Delta\phi(\text{jet})$	difference in azimuthal angle between the particle and the jet axis
$ \eta $	absolute value of the particle's pseudorapidity
q	electric charge of the particle
pvAssociationQuality	flag related to the association of the track to the primary vertices
lostInnerHits	quality flag of the track related to missing hits on the pixel layers
χ^2/dof	χ^2 value of the trajectory fit normalized to the number of degrees of freedom
qualityMask	quality flag of the track
d_z	longitudinal impact parameter of the track
d_z/σ_{d_z}	significance of the longitudinal impact parameter
d_{xy}	transverse impact parameter of the track
$d_{xy}/\sigma_{d_{xy}}$	significance of the transverse impact parameter
η_{rel}	pseudorapidity of the track relative to the jet axis
$p_{T,\text{rel}}$ ratio	track momentum perpendicular to the jet axis, divided by the magnitude of the track momentum
$p_{\text{par,rel}}$ ratio	track momentum parallel to the jet axis divided by the magnitude of the track momentum
d_{3D}	signed 3D impact parameter of the track
d_{3D}/σ_{3D}	signed 3D impact parameter significance of the track
trackDistance	distance between the track and the jet axis at their point of closest approach
For SVs within the jet cone	
$\log p_T$	logarithm of the SV's p_T
m_{SV}	mass of the SV
$\Delta\eta(\text{jet})$	difference in pseudorapidity between the SV and the jet axis
$\Delta\phi(\text{jet})$	difference in azimuthal angle between the SV and the jet axis
$ \eta $	absolute value of the SV's pseudorapidity
N_{tracks}	number of tracks associated with the SV
χ^2/dof	χ^2 value of the SV fit normalized to the number of degrees of freedom
d_{2D}	signed 2D impact parameter (i.e., in the transverse plane) of the SV
d_{2D}/σ_{2D}	signed 2D impact parameter significance of the SV
d_{3D}	signed 3D impact parameter of the SV
d_{3D}/σ_{3D}	signed 3D impact parameter significance of the SV

$Z \rightarrow c\bar{c}$ jets to correct their $X \rightarrow c\bar{c}$ tagging efficiencies under the given working points. The measured SFs typically range from 0.9 to 1.3 with around 20–30% uncertainties.

As the $H \rightarrow b\bar{b}$ jets is a main background that contaminates the $H \rightarrow c\bar{c}$ signal, it is also crucial to correct the $X \rightarrow c\bar{c}$ mistagging efficiency for $H \rightarrow b\bar{b}$ jets and assign uncertainties on it. However, the sfBDT and existing calibration methods are not able to derive reliable mistagging SFs for the $H \rightarrow b\bar{b}$ jet because of the extremely low fraction of events in the f1vB category surviving the tight ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant. Therefore, the analysis takes a conservative approach and assigns 100% uncertainty in the $H \rightarrow b\bar{b}$ mistagging rate.

Table 8.2 Definition of the three $X \rightarrow c\bar{c}$ tagging categories for the $VH \rightarrow c\bar{c}$ boosted-jet analysis.

Category	$X \rightarrow c\bar{c}$ tagging discriminant
High purity (HP)	> 0.99
Medium purity (MP)	$(0.96, 0.99)$
Low purity (LP)	$(0.90, 0.96)$

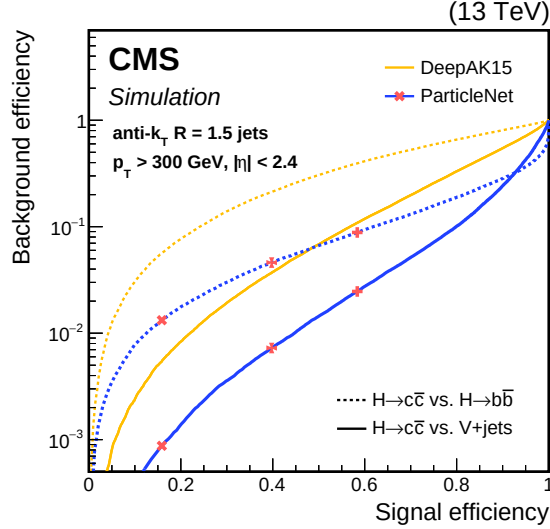


Figure 8.1 Performance of ParticleNet-MD AK15 $X \rightarrow c\bar{c}$ tagging discriminant for identifying $H \rightarrow c\bar{c}$ jets versus the V +jets and the $H \rightarrow b\bar{b}$ backgrounds in terms of the ROC curve.

8.2.2 Mass regression with ParticleNet

An accurate mass reconstruction of the Higgs boson candidate leads to better search sensitivity, as the signal is extracted from the reconstructed Higgs boson mass distribution. The former $VH \rightarrow c\bar{c}$ analysis adopts the soft-drop mass of the Higgs boson, which works effectively to subtract the additional contributions from the soft QCD radiations within the large jet cone. These contributions may come from initial state radiations or pile-up. However, the soft-drop algorithm also often over-subtract genuine hard jet constituents that account for a large fraction of the jet momentum, which can result in $m_{\text{SD}} \gtrsim 0$. Moreover, the application of the soft drop algorithm introduces a long tail at large masses (in particular for highly boosted jets), which results in additional loss of sensitivity. These two effects are illustrated in Figs. 8.2 and 8.3 in $H \rightarrow c\bar{c}$ and QCD jets for various p_{T} ranges.

To improve the mass reconstruction resolution, a regression technique using the advanced ParticleNet neural network architecture is adopted to estimate the generator-level resonance

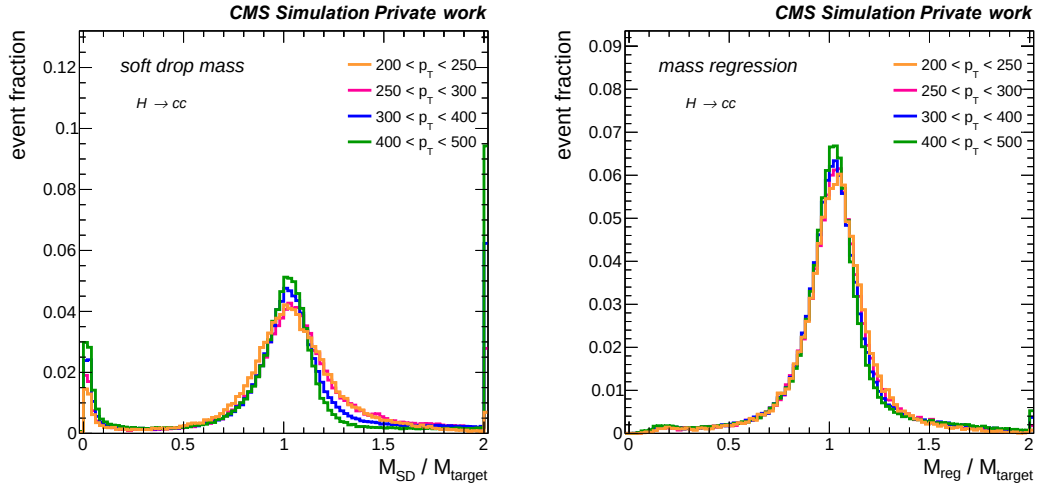


Figure 8.2 The mass resolution obtained with the soft-drop (left) and mass regression (right) algorithms on large- R ($R = 1.5$) $H \rightarrow c\bar{c}$ jets for various p_T ranges.

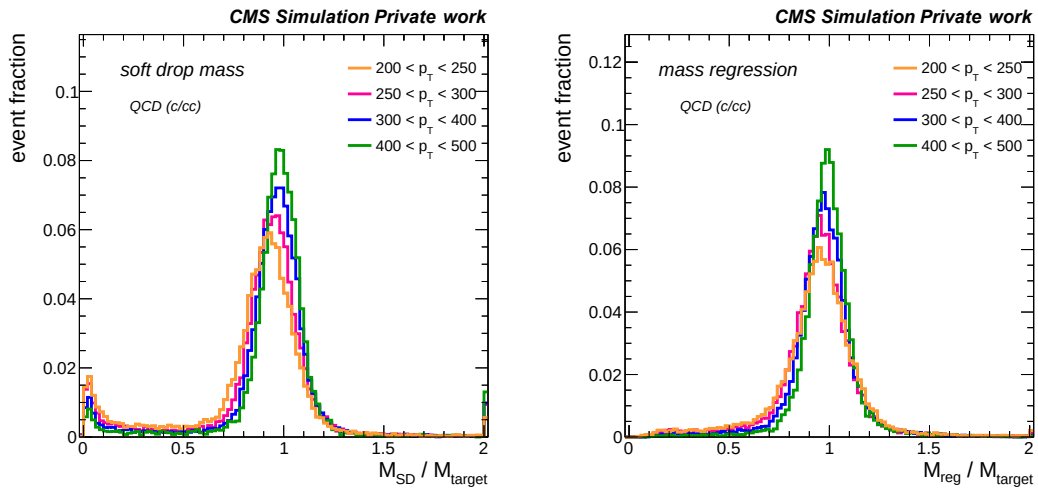


Figure 8.3 The mass resolution obtained with the soft-drop (left) and mass regression (right) algorithms on large- R ($R = 1.5$) QCD jets for various p_T ranges.

Table 8.3 The calibration results of the $X \rightarrow c\bar{c}$ tagging discriminant on large- R ($R = 1.5$) $H \rightarrow c\bar{c}$ signal jets in terms of scale factors. The calibration is dependent on jet p_T and data-taking conditions and is performed using the sfBDT method.

$X \rightarrow c\bar{c}$ tagging category	p_T range in GeV:		
	(200, 350)	(350, 450)	(450, ∞)
2016			
HP	$1.395^{+0.272}_{-0.271}$	$1.258^{+0.269}_{-0.268}$	$1.151^{+0.214}_{-0.197}$
MP	$1.192^{+0.218}_{-0.217}$	$1.125^{+0.201}_{-0.199}$	$0.958^{+0.249}_{-0.196}$
LP	$1.019^{+0.281}_{-0.281}$	$1.062^{+0.310}_{-0.309}$	$1.066^{+0.324}_{-0.336}$
2017			
HP	$1.337^{+0.272}_{-0.271}$	$1.279^{+0.304}_{-0.302}$	$1.245^{+0.244}_{-0.242}$
MP	$1.187^{+0.239}_{-0.239}$	$1.171^{+0.241}_{-0.236}$	$0.954^{+0.433}_{-0.420}$
LP	$1.122^{+0.226}_{-0.226}$	$1.016^{+0.209}_{-0.206}$	$1.180^{+0.448}_{-0.443}$
2018			
HP	$1.001^{+0.315}_{-0.315}$	$1.078^{+0.313}_{-0.305}$	$1.299^{+0.293}_{-0.294}$
MP	$0.996^{+0.182}_{-0.181}$	$1.176^{+0.473}_{-0.476}$	$1.081^{+0.392}_{-0.388}$
LP	$0.952^{+0.231}_{-0.231}$	$1.037^{+0.286}_{-0.281}$	$1.147^{+0.419}_{-0.419}$

mass. Unlike the ParticleNet-MD tagger, there is only one output node, and the softmax unit is removed after the last layer to assist the single-output regression task. Specific studies were done for the regression task, including the training samples and the training objective, as documented below.

Samples and regression target

The signal jets in this analysis are matched to a genuine $H \rightarrow c\bar{c}$ signature with m_H at the SM value. Although it is expected that the regression model will predict the m_H value right at 125 GeV for the signal, one cannot use pure signal jets for training the model because the other model will only learn to regress towards 125 GeV. Therefore, the training jets should include $H \rightarrow c\bar{c}$ signature with m_H spanning from a wide range. The same training sample resulting from the $\tilde{G} \rightarrow HH$ process is exploited, where H decays into $b\bar{b}$, $c\bar{c}$, $s\bar{s}$, $u\bar{u}$, and $d\bar{d}$, and initiate resonant large- R jet with a wide spread of resonance masses.

It is also found important to include the QCD jets in training, as they contribute signifi-

cantly in the analysis search of $H \rightarrow c\bar{c}$, hence their regressed mass distribution can be made predictable if the model is also trained on their phase spaces and regress on proper values consistent with the GEN-level resonance mass for the resonant jets. A viable solution is to choose a continuous target for these QCD jets, such as the generated soft-drop mass m_{SD} . This is defined as the m_{SD} computed using hadron-level constituents of the GEN-jet (defined in Section 1.5.7). The following mass target is selected:

$$m_{\text{target}} = \begin{cases} \text{GEN-level resonance mass} & \text{for resonant jets,} \\ \text{GEN-jet } m_{\text{SD}} & \text{for QCD jets.} \end{cases} \quad (8.2)$$

8.2.2.1 Training

Similar to the ParticleNet-MD tagger training, the sampling-based reweighting is applied on the fly to yield smooth and similar distributions on p_{T} and m_{SD} , for QCD jets and the resonant jets. The input variables are also the same, including features from the associated PF candidates and matched SVs, with a full feature list provided in Table 8.1.

The training is performed by minimizing a log-cosh loss for every batch of training. The log-cosh loss function is defined by

$$L = \log(\cosh(\hat{\theta} - \theta)). \quad (8.3)$$

The log-cosh loss can be viewed as a combination of L2 loss ($L_{\text{reg,L2}} = (\hat{\theta} - \theta)^2$) when the estimated $\hat{\theta}$ is in the vicinity of θ , and the L1 loss ($L_{\text{reg,L1}} = |\hat{\theta} - \theta|$) when the estimated value is further. This can be seen from the shape of the log-cosh function in Fig. 8.4.

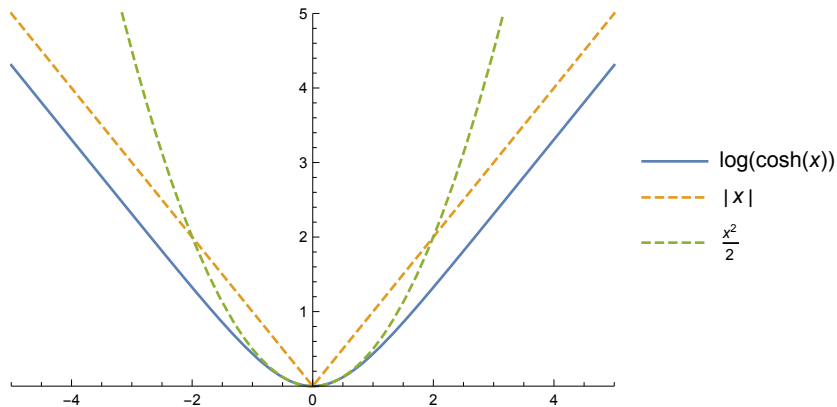


Figure 8.4 Comparison of the log-cosh loss function shape with the L1 and L2 loss function shapes.

Performance

The performance of the mass regression is shown in Figs 8.2 and 8.3 on the right plots in $H \rightarrow c\bar{c}$ and QCD jets for various p_T ranges. The effective mass resolution as a function of p_T , for the soft-drop and the mass regression algorithms is shown in Fig. 8.5 on the left plot, for Higgs and QCD jets. The effective mass resolution $\sigma_{\text{eff}}(m)/m$ is defined as the half of the smallest interval containing 68% of the area under the curve. The mass regression substantially improves the resolution over the full kinematic range compared to the soft drop algorithm. The improvement is larger for $H \rightarrow b\bar{b}$ jets compared to $H \rightarrow c\bar{c}$ due to the larger missing energy carried out from the neutrinos. It is also observed that the improvement in resolution is mild at best for QCD jets, which is most likely due to the target mass being relatively similar to the reconstructed soft drop mass. In Fig. 8.5 on the right plot, the efficiency of a hypothetical event selection $|m - m_H| < 2\sigma_{\text{eff}}$ is shown. This also shows a relative improvement, as the regressed mass is able to recover signal events lost in the tails of the jet mass distribution after applying the soft drop algorithm. Besides, it is interesting to see that the overall efficiency is smaller for QCD jets. This indicates a gain both in the signal yield and in the background rejection by using this technique. The total expected gain in sensitivity is shown in Fig. 8.6 for $H \rightarrow b\bar{b}$, $H \rightarrow c\bar{c}$ and $H \rightarrow q\bar{q}$ jets. In the $VH \rightarrow c\bar{c}$ analysis, the improvement in the significance is expected to be around 20% by applying the mass regression to replace the soft-drop mass.

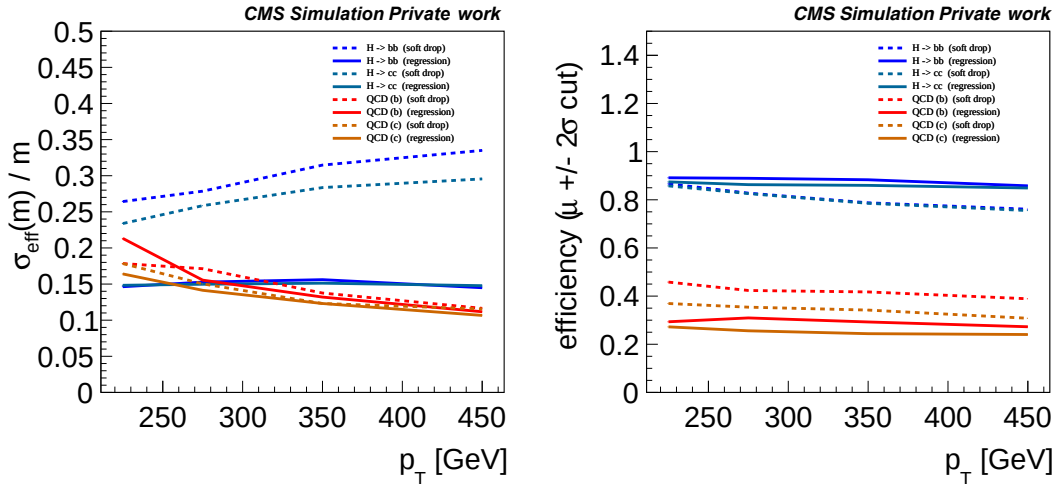


Figure 8.5 The mass resolution as a function of jet p_T obtained with the soft drop and the mass regression algorithms (left); the efficiency of an event selection $|m - m_H| < 2\sigma_{\text{eff}}$ as a function of jet p_T (right).

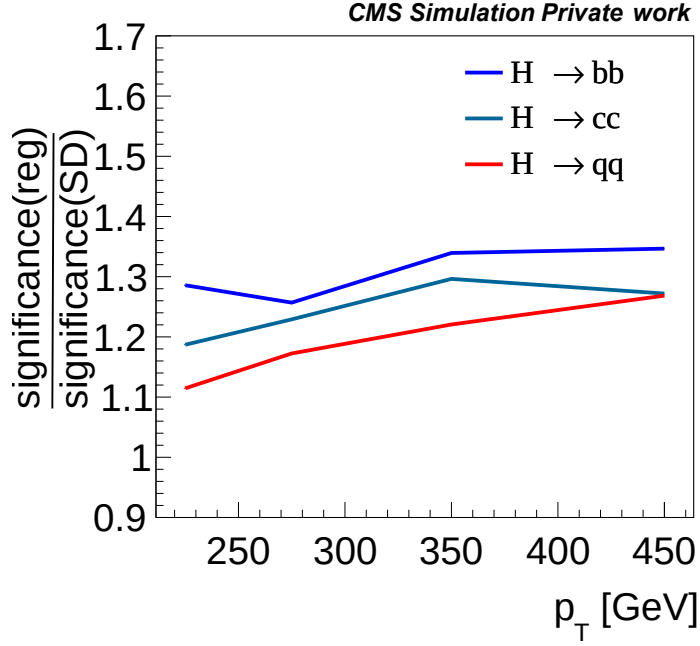


Figure 8.6 The expected gain in significance by using the regressed mass compared with using the soft-drop mass as a function of p_T .

Calibration

The performance of the mass regression is validated in data using a boosted semi-leptonic $t\bar{t}$ control sample (single- μ). The goal is to measure the regressed mass scale (JMS) and resolution (JMR) on the $W \rightarrow cx$ component of the top quark jet, that we identify as a proxy for the $H \rightarrow c\bar{c}$ jet.

Such a sample is selected by first requiring at least one large- R jet (the hadronic top candidate) with $p_T > 200$ GeV. The leptonic W , constructed by combining the momenta of the missing energy and the muon, is required to have a $p_T > 150$ GeV.

In simulation, the template distribution of the large- R jet regressed mass is constructed for each background process, and the $t\bar{t}$ contribution in three components:

- Top-merged: all three quarks from the hadronic top quark decay are contained in the large- R jet, i.e. satisfying $\Delta R(\text{large-}R \text{ jet}, q) < 1.5$.
- W -merged: only the two quarks from the hadronic W boson decay are contained in the large- R jet, i.e., satisfying $\Delta R(\text{large-}R \text{ jet}, q) < 1.5$, while the b quark from the top decay is outside the jet cone.
- Non-merged: if the large- R jet is neither top-merged nor W -merged.

The JMS and JMR are constructed as two parameters of interest, which contribute to the

template change by shifting or smearing the W peak on the regressed mass distribution. They are extracted by performing the maximum likelihood fit, where three $t\bar{t}$ components can vary simultaneously, for events passing and failing the low purity ParticleNet-cc tagger score. The post-fit regressed mass distributions are shown in Fig. 8.7.

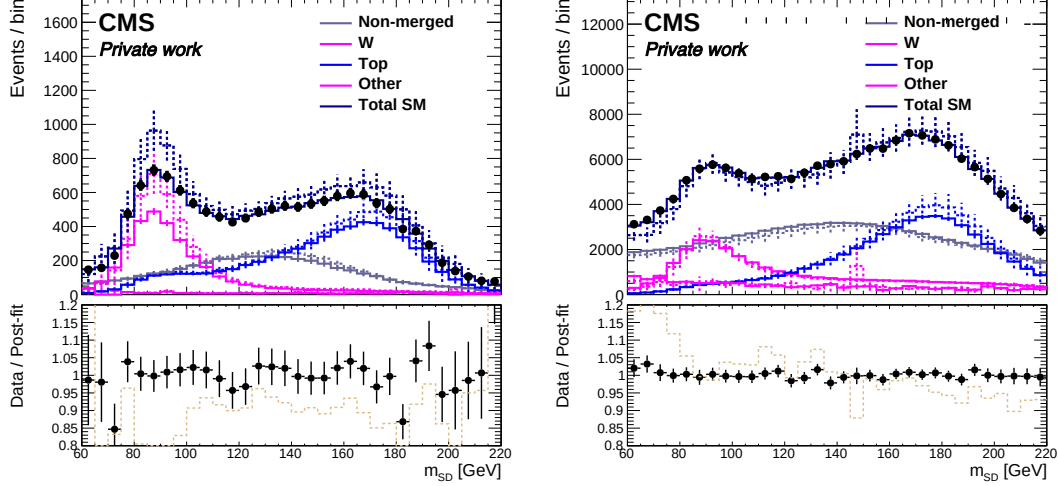


Figure 8.7 The post-fit distributions and data–simulation agreement in the single-muon $t\bar{t}$ control sample as a function of leading large- R jet regressed mass after passing (top) and failing (bottom) the ParticleNet $X \rightarrow c\bar{c}$ tagging discriminant at LP.

The fit results are summarized in Table 8.4. While the JMS is compatible with 1 across all datasets, the fit is able to constrain the JMR factor only to approximately 5% (i.e., to 50% of its initial value). This implies that, to match the resolution observed in the data, the regressed mass in simulation needs to be smeared by an additional 5%.

Table 8.4 The regressed jet mass JMS and JMR (relative to the nominal 5% and 10%) obtained from the template fit of the $W \rightarrow c\bar{c}$ component.

Parameters	2016	2017	2018
JMS	0.01 ± 0.08	-0.07 ± 0.08	-0.14 ± 0.07
JMR	0.44 ± 0.11	0.39 ± 0.11	0.48 ± 0.11

8.3 Reconstruction of objects and events

In the $VH \rightarrow c\bar{c}$ analysis, the reconstruction of PF candidates and SVs follows the description in Section 1.5. These objects are crucial to the training of the ParticleNet tagging and regression model.

8.3.1 Jets

The large- R jet collection used in this analysis is clustered from PF candidates and reconstructed with the anti- k_T algorithm^[33] using a distance parameter $R = 1.5$. As described in Section 1.5, the PUPPI algorithm^[36] is used to correct pileup effects, and the soft drop is applied as a grooming algorithm to acquire the soft-drop mass. The jet energy scale (JEC) and resolution (JER) corrections derived within CMS for AK8 jets are applied to the described AK15 jets. The AK15 jets considered in this analysis have $p_T > 200$ GeV, $|\eta| < 2.4$. These jets are also required to satisfy the tight jet identification criteria recommended within CMS. The AK15 jets overlapping with any loose electrons or muons, i.e., $\Delta R(\text{large-}R \text{ jet}, \ell) < 1.5$, are discarded.

An additional jet collection, the small- R jet collection, is used in the event selection, clustered from PF candidates using anti- k_T algorithm with $R = 0.4$. Jet energy and resolution corrections supplied within the CMS group are applied. The selected jets are required to have $p_T > 25$ GeV and to be within $|\eta| < 2.4$. The small- R jets are also required to satisfy the tight lepton-veto identification criteria and the tight pileup jet identification criteria within CMS. The small- R jets overlapping with any loose electrons or muons, i.e., $\Delta R(\text{small-}R \text{ jet}, \ell) < 0.4$, are discarded.

8.3.2 Leptons

Two sets of selection criteria are used in the analysis of electrons and muons. A set of loose criteria (“loose” leptons) is used to categorize events into the 0L, 1L, and 2L channels and also for jet-lepton cross-cleaning. More stringent requirements (“tight” leptons) are imposed in the 1L channel to select high-purity samples of isolated leptons.

Electron candidates are preselected by requiring $p_T > 20$ GeV and $|\eta| < 2.5$. Additional identification criteria are imposed to suppress background. Electron candidates are identified via a set of selection criteria based on the “MVARun2” identification method within CMS. The Electron MVARun2 approach forms a single discriminator variable that is computed based on multiple parameters of the electron object and provides the best separation between the signal and backgrounds. The loose and tight electron definitions are then defined based on the selection on the Electron MVARun2 discriminant. The loose electrons are required to pass the working point with a signal efficiency of around 90%. For the tight electrons, a tighter working point with a signal efficiency of around 80% is chosen.

Muon candidates are preselected by requiring $p_T > 20$ GeV and $|\eta| < 2.4$. In addition,

we require the muon candidates to pass the cut-based loose ID criterion defined within CMS and have a relative PF-isolation smaller than 0.25. These define the loose muon selection criteria. For the 1-lepton channel, tight muons are required to satisfy a more stringent isolation criterion, with a relative isolation of less than 0.06, and additional requirements on the impact parameters, $|d_{xy}| < 0.05$ cm and $|d_z| < 0.2$ cm. In addition, the muon momentum in data and simulation is corrected with the Rochester corrections.

8.4 Analysis selection

A typical $VH \rightarrow c\bar{c}$ signal event has a clear signature of a vector boson recoiling against a Higgs boson. This analysis focuses on the leptonic decays of the vector boson. By selecting events with one or more well-identified leptons or high p_T^{miss} originating from the neutrinos, the large QCD multijet background can be effectively suppressed. This analysis defines three analysis channels, the 2L, 1L, and 0L channels, based on the number of leptons passing the loose lepton selection criteria. The event selection criteria are described separately for each channel in Sections 8.4.1 to 8.4.3. Finally, the event selection criteria are summarized in Section 8.4.4, Table 8.6.

8.4.1 The 2-lepton channel

The 2L channel targets ZH events with the Z boson decaying into a pair of electrons or muons. Events are selected with at least two electrons or muons passing the loose lepton selection criteria described in Section 8.3.2. The leading two leptons are required to have opposite electric charges and are of the same flavour, and the invariant mass of the lepton pair is compatible with the Z boson mass, with $75 < M(\ell\ell) < 105$ GeV. The Z boson is thus reconstructed from the two leading in p_T leptons.

In this boosted analysis, only events with $p_T(Z) > 150$ GeV are selected. Events are required to have at least one large- R jet with $p_T > 200$ GeV, and the leading large- R jet is selected as the Higgs boson candidate. In $VH \rightarrow c\bar{c}$ signal events, the vector boson and the Higgs boson are typically back-to-back with each other. Therefore, the difference in azimuthal angle between the reconstructed Z and H is imposed, i.e. $\Delta\phi(V, H_{\text{cand}}) > 2.5$. To avoid double-counting, small- R jets are removed from the event if they overlap H_{cand} with $\Delta R(\text{jet}, H_{\text{cand}}) < 1.5$. The number of small- R jets is required to be less than 3 in the 2L channel.

After the baseline selection, the dominant background process is Z +jets, with a small

contribution from diboson production. The $t\bar{t}$ and single top contributions are suppressed to negligible levels due to the requirement of $p_T(Z) > 150$ GeV. The $VH \rightarrow b\bar{b}$ process is a very small but largely irreducible background.

To better distinguish the VH signal process from the main backgrounds, a BDT discriminant, named the “kinematic BDT”, is developed. It is designed using only the kinematics properties of the event, i.e. properties that do not include intrinsic properties like the flavour content and the mass. The goal of the kinematics BDT is to improve the separation between the VH signal and the main backgrounds (i.e., Z +jets in the 2L channel) while remaining largely independent of the $X \rightarrow c\bar{c}$ discriminant and the H_{cand} mass. This is achieved by selecting only kinematic input variables for the BDT training. The input variables used for training the BDT in the 2L channel are displayed in Table 8.5, and the distribution of these variables in the VH signal and in the backgrounds are shown in Fig. 8.8.

The kinematic BDT is trained with XGBoost^[140] on the simulated sample in the 2017 detector conditions. The $VH \rightarrow c\bar{c}$ sample is used as the signal sample, and only the main background, Z +jets, is used as the background sample in the case of the 2L channel. To avoid overfitting, a 10-fold cross-validation^[141] is used in the development of the BDT. Hence, no events are used in both the training and the application of a BDT model. The model is also applied to simulated samples and data with other data-taking conditions without observing a performance loss. The 10-fold cross-validation is not exploited here, but an ensemble using all 10 BDT models trained is used to predict the output score, and the mean is taken. This is found to bring a small improvement in the discrimination power.

Figure 8.9 (top left) shows the distribution of the kinematic BDT after the baseline selection for the VH signal and the backgrounds. A good separation is observed. The BDT discriminant is used to categorize into SR (BDT disc. > 0.55) and CR (BDT disc. < 0.55). To show that the kinematic BDT is largely decorrelated with the $X \rightarrow c\bar{c}$ discriminant and the H_{cand} mass, Fig. 8.9 also shows the BDT distribution in the three $X \rightarrow c\bar{c}$ tagging categories. This is further validated in Fig. 8.10, showing the BDT distributions separately for the $H \rightarrow c\bar{c}$ signal and the main background, Z +jets.

After the selection based on the kinematic BDT, events are required to pass the $X \rightarrow c\bar{c}$ discriminant selections. The distribution of the $X \rightarrow c\bar{c}$ discriminant is shown in Fig. 8.11 for signal and background events in the high BDT region. It shows powerful separation power between $H \rightarrow c\bar{c}$ signal and Z +jets background. Events are categorized into three exclusive $X \rightarrow c\bar{c}$ categories. Figure 8.12 displays the distribution of the mass of the H_{cand} for events

with high kinematic BDT output scores and passing the $X \rightarrow c\bar{c}$ discriminant selections for the three categories. Since both the kinematic BDT and the $X \rightarrow c\bar{c}$ discriminant are designed to be independent of the mass of the large- R jet, the peak structure of the $VH \rightarrow c\bar{c}$ signal and the nature of flat mass spectra of various backgrounds remain largely unchanged. This ensures a signal extraction via a template fit in the H_{cand} mass.

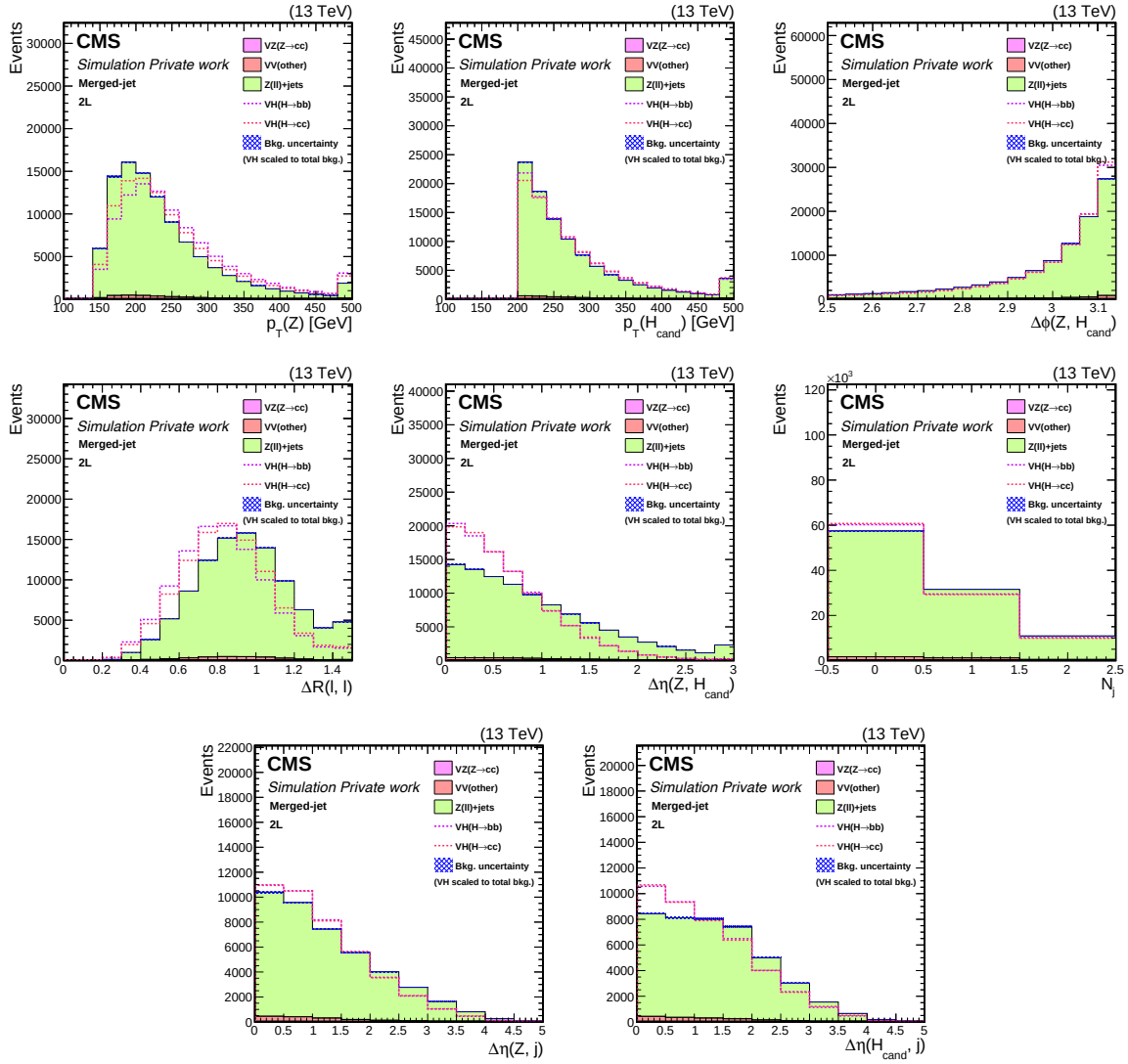


Figure 8.8 The distributions of the input variables used for training the kinematic BDT in the 2L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds. The definition of the variables can be found in Table 8.5.

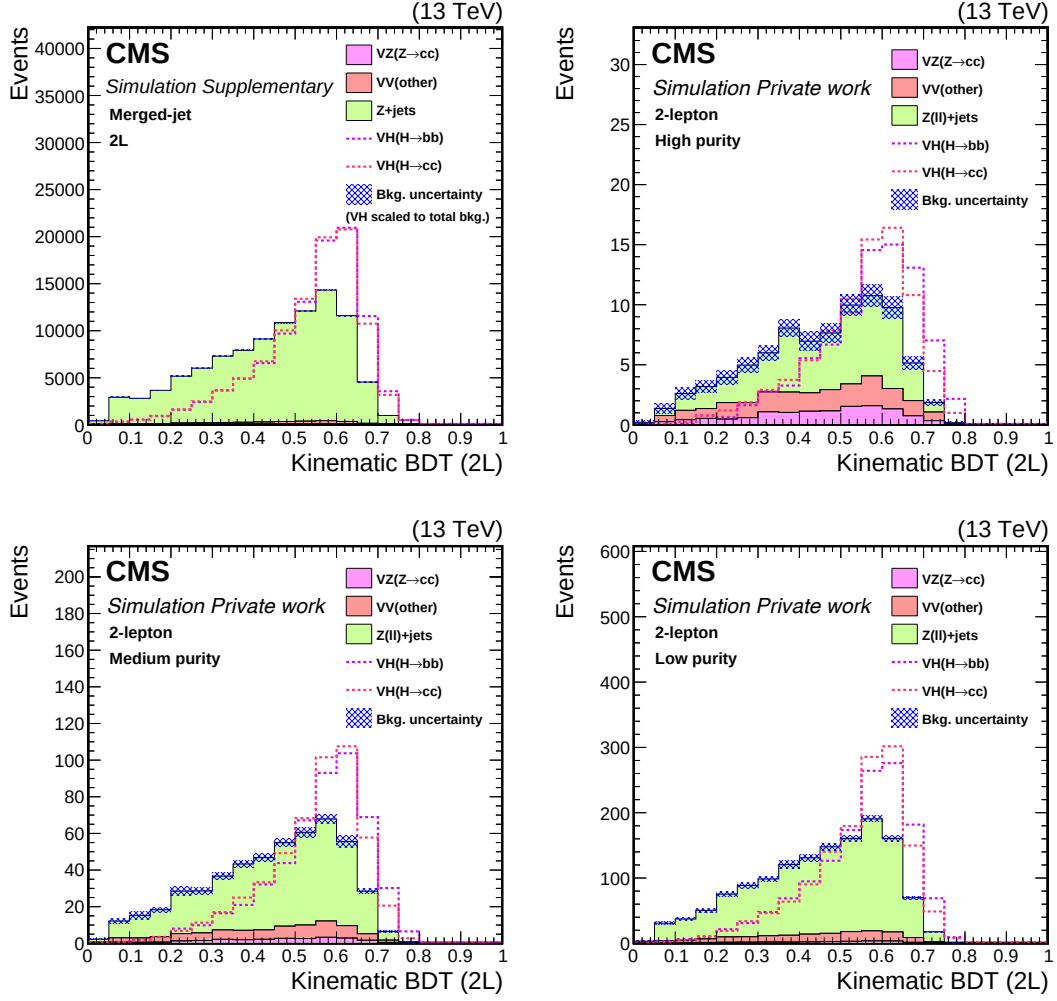


Figure 8.9 The kinematic BDT output distributions after four sets of selections in the 2L channel, i.e. the baseline selection (top left, from Ref.^[113]), the $X \rightarrow c\bar{c}$ tagging category selection at HP (top right), the selection at MP (bottom left), and the selection at LP (bottom right). Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds.

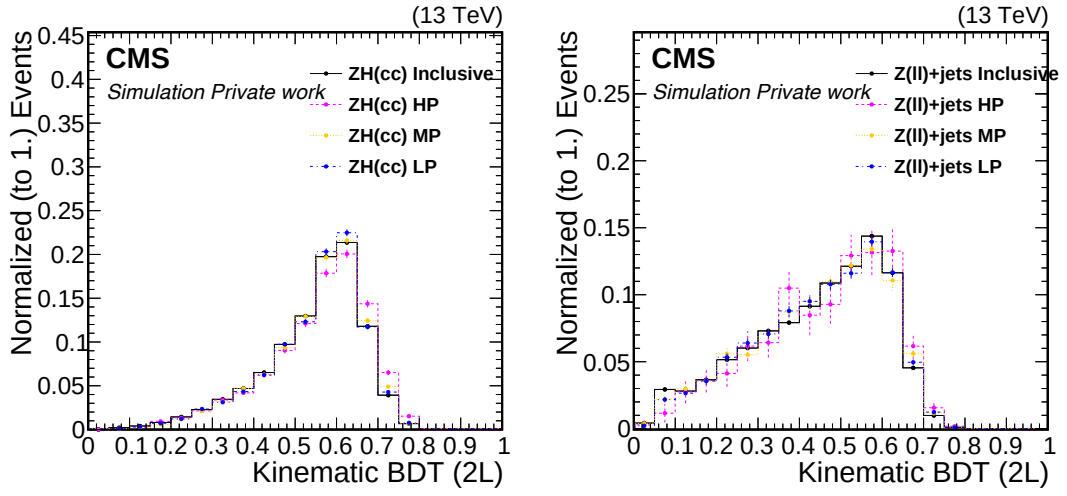


Figure 8.10 Comparison of the kinematic BDT output distribution in the 2L channel before and after the requirement on the $X \rightarrow c\bar{c}$ tagging discriminant.

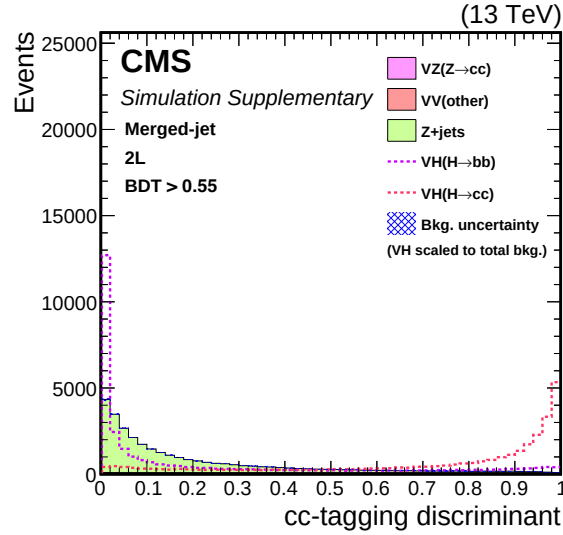


Figure 8.11 Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high kinematic BDT output scores in the 2L channel^[113].

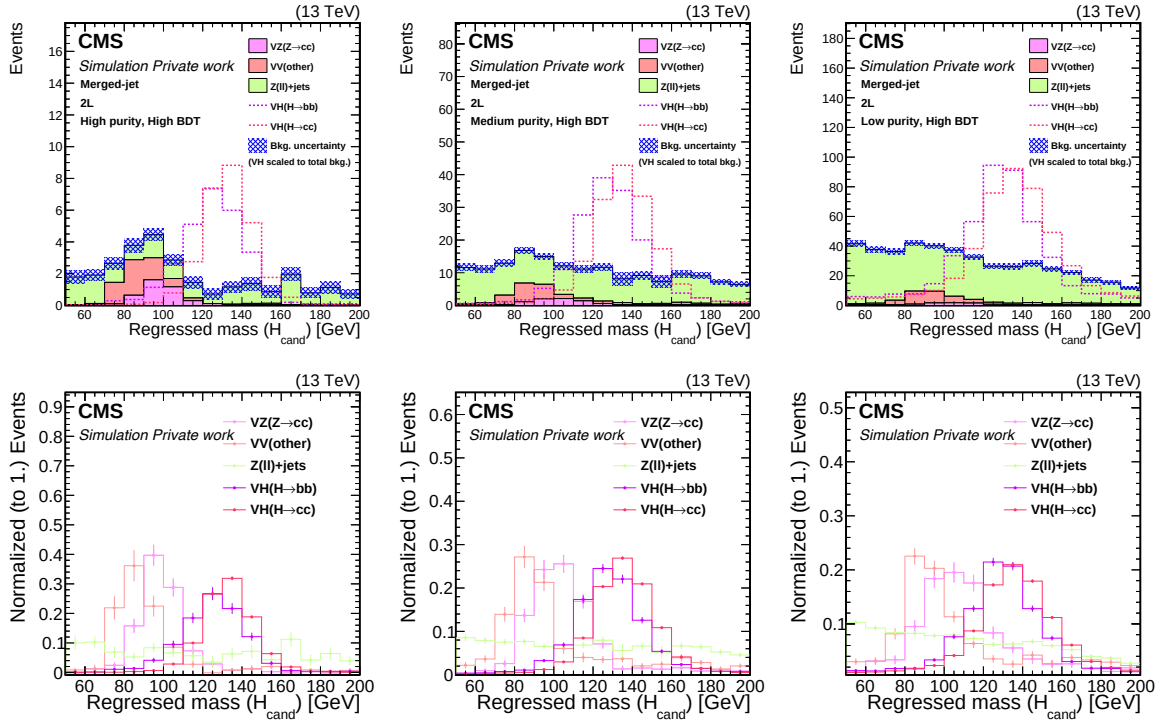


Figure 8.12 The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 1L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.

8.4.2 The 1-lepton channel

The 1L channel targets WH events where the W bosons decay leptonically. Events are selected with exactly one electron or muon passing the loose lepton definition described in Section 8.3.2. Furthermore, the electron and muon identification criteria are tightened to improve the rejection of non-prompt leptons from the QCD multijet background. The p_T of the electron (muon) is required to be above 25 (30) GeV to match the trigger selection threshold. The leptonically decaying W boson is reconstructed as the vector sum of the momenta of the charged lepton and the p_T^{miss} . Besides, a selection on the azimuthal angle between the charged lepton and the missing transverse momentum is required, i.e. $\Delta\phi(\vec{p}_T^{\text{miss}}, \ell) < 1.5$ according to the WH event topology.

For the boosted-jet analysis, the reconstructed W boson is required to satisfy $p_T(W) > 150$ GeV. The leading large- R jet is selected as the H_{cand} with the same $p_T > 200$ GeV requirement imposed. Small- R jets overlapping with H_{cand} are also removed from the events, and the remaining number of small- R jets is required to be less than 2 in the 1L channel to suppress contribution from the $t\bar{t}$ process.

Similar to the 2L channel, a kinematic BDT is also developed to reject the main backgrounds, which are W +jets and $t\bar{t}$, while maintaining the independence of the $X \rightarrow c\bar{c}$ discriminant and the H_{cand} mass. The same training setup is used, but the variables are different from the 2L case and are defined in Table 8.5.

Figure 8.13 shows the distribution of the kinematic BDT after the baseline selection for the VH signal and the backgrounds. For the 1L channel, three categories are defined based on the BDT output. Two of them are considered as SRs: BDT disc. > 0.7 as the high BDT region, and $0.55 < \text{BDT disc.} < 0.7$ as the medium BDT region. The rest of the events enter the low BDT CR. The thresholds are initially optimized to maximize the expected upper limits on the $VH \rightarrow c\bar{c}$ signal. Again, validations show that the BDT discriminant in 1L is found in good independence with the $X \rightarrow c\bar{c}$ discriminant and the H_{cand} mass.

Events selected by the kinematic BDT are then required to pass the $X \rightarrow c\bar{c}$ discriminant selections. The distribution of the $X \rightarrow c\bar{c}$ tagger discriminant is shown in Fig. 8.14 for signal and background events in the high BDT region. The $X \rightarrow c\bar{c}$ tagging discriminant shows a significant separation power between $H \rightarrow c\bar{c}$ signal and W +jets background, which is dominated by light flavour jets. The W +jets background can be highly suppressed by the $X \rightarrow c\bar{c}$ discriminant selection. Figure 8.15 shows the distribution of the H_{cand} mass in the high BDT region and passing three $X \rightarrow c\bar{c}$ discriminant selections. Again, a good mass spectra

separation is observed, which guarantees a sensitive extraction of the signal via the template fit.

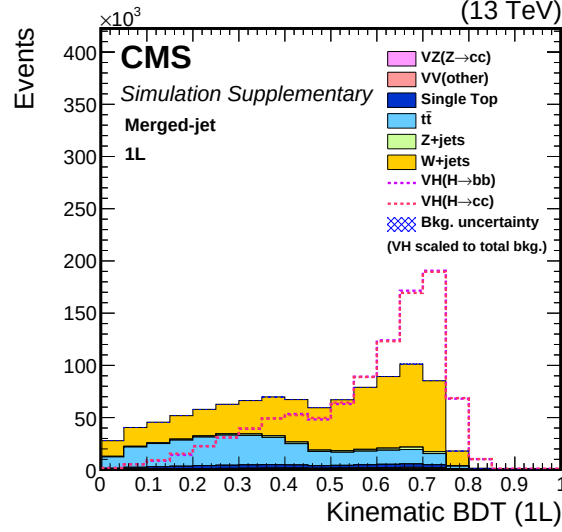


Figure 8.13 The kinematic BDT output distributions after the baseline selection in the 1L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds^[113].

8.4.3 The 0-lepton channel

The 0L channel targets mainly ZH events with the Z boson decaying into a pair of neutrinos, giving rise to large p_T^{miss} . Besides, WH events with W boson decaying leptonically can also enter this channel if the charged lepton misses identification. For the 0L channel, events are selected with no electrons or muons passing the loose lepton definition described in Section 8.3.2. The $\vec{p}_T^{\text{miss}}$ corresponds to the transverse momentum of the vector boson in this channel. The $H \rightarrow c\bar{c}$ candidate is reconstructed in the same way as in the 1L and 2L channels from the leading large- R jet with $p_T > 200$ GeV. Small- R jets overlapping with H_{cand} are removed from the events and the number of remaining jets is required to be less than 2 to suppress contribution from the $t\bar{t}$ process.

As is done in the 1L and 2L channels, a kinematic BDT is developed to reject the main backgrounds, which include Z +jets, W +jets and $t\bar{t}$ for 0L. The same design philosophy and training setup are followed. For signal, both the ZH and the WH samples are used, weighted by their cross sections. The input variables used for training the BDT in 0L channel are defined in Table 8.5.

Figure 8.16 shows the distribution of the kinematic BDT after the baseline selection for

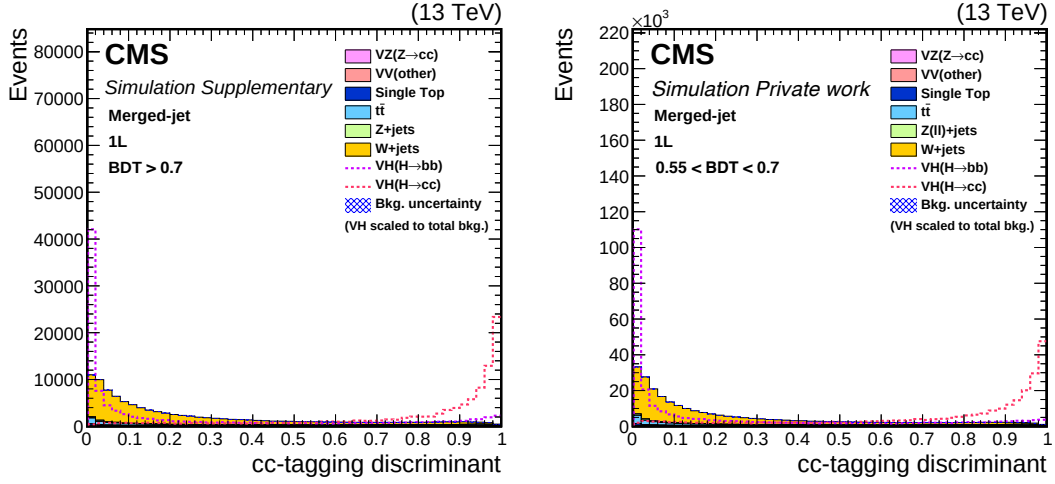


Figure 8.14 Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high (left, from Ref.^[113]) and medium (right) kinematic BDT output scores in the 1L channel.

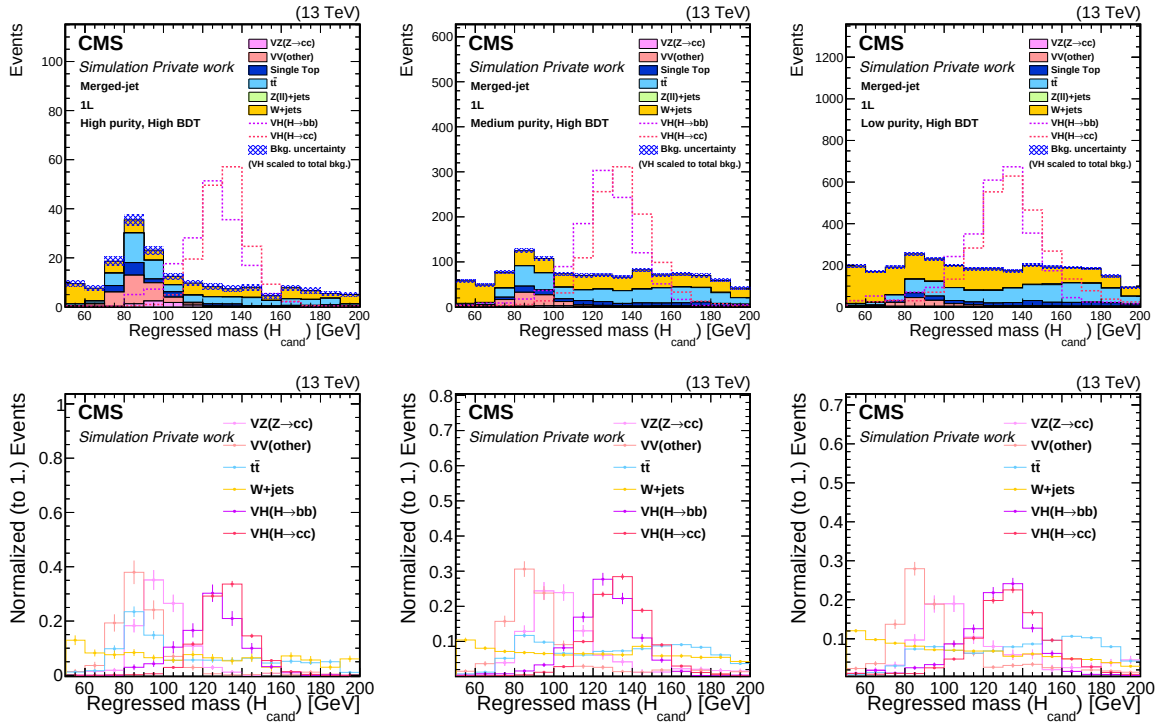


Figure 8.15 The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 1L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.

the VH signal and the backgrounds. The separation power is slightly weaker in 0L than in the 1L and 2L channels, because only the kinematic information of the vector boson in the transverse plane is reconstructed. Events with BDT disc. > 0.55 go to the SR and others go to the CR. It has also been validated that the trained BDT discriminant is largely independent of the $X \rightarrow c\bar{c}$ discriminant and the H_{cand} mass.

Events are required to pass $X \rightarrow c\bar{c}$ discriminant selections on top of the BDT. The distribution of the $X \rightarrow c\bar{c}$ discriminant is shown in Fig. 8.17 for signal and background events in the high BDT region. For 0L, The $X \rightarrow c\bar{c}$ discriminant shows powerful separation power between $H \rightarrow c\bar{c}$ signal and V +jets background, significantly suppressing the V +light flavour backgrounds while leaving mostly V +heavy flavour events and $t\bar{t}$ events. Figure 8.18 displays the H_{cand} mass distribution in the high BDT region and in different $X \rightarrow c\bar{c}$ tagging categories. The spectrum of the signal and background events are in expected shapes.

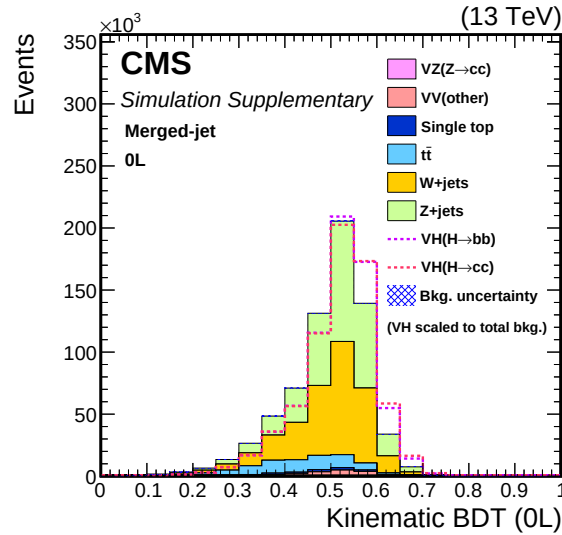


Figure 8.16 The kinematic BDT output distributions after the baseline selection in the 0L channel. Comparison between the shape of the VH signal and the background samples is made, where the VH signal is normalized to the sum of all backgrounds^[113].

8.4.4 Summary of the event selection

As a summary, Table 8.5 provides the list of kinematic input variables for the training of three kinematic BDTs for the three channels. The full event selection for all three channels is concluded in Table 8.6.

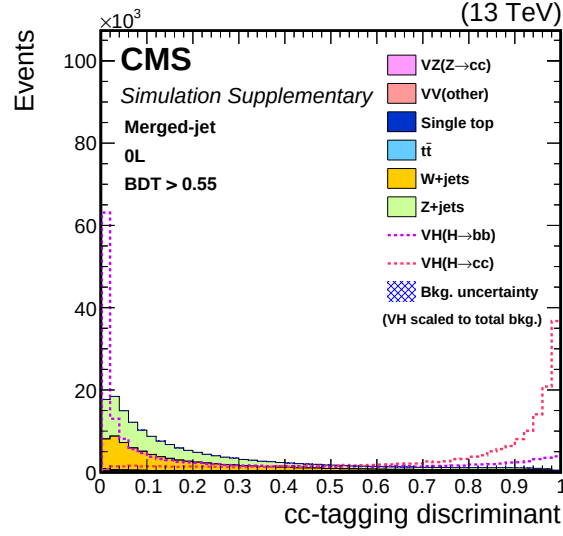


Figure 8.17 Comparison of the $X \rightarrow c\bar{c}$ tagging discriminant between the $VH \rightarrow c\bar{c}$ signal and the backgrounds for events with high kinematic BDT output scores in the 0L channel^[113].

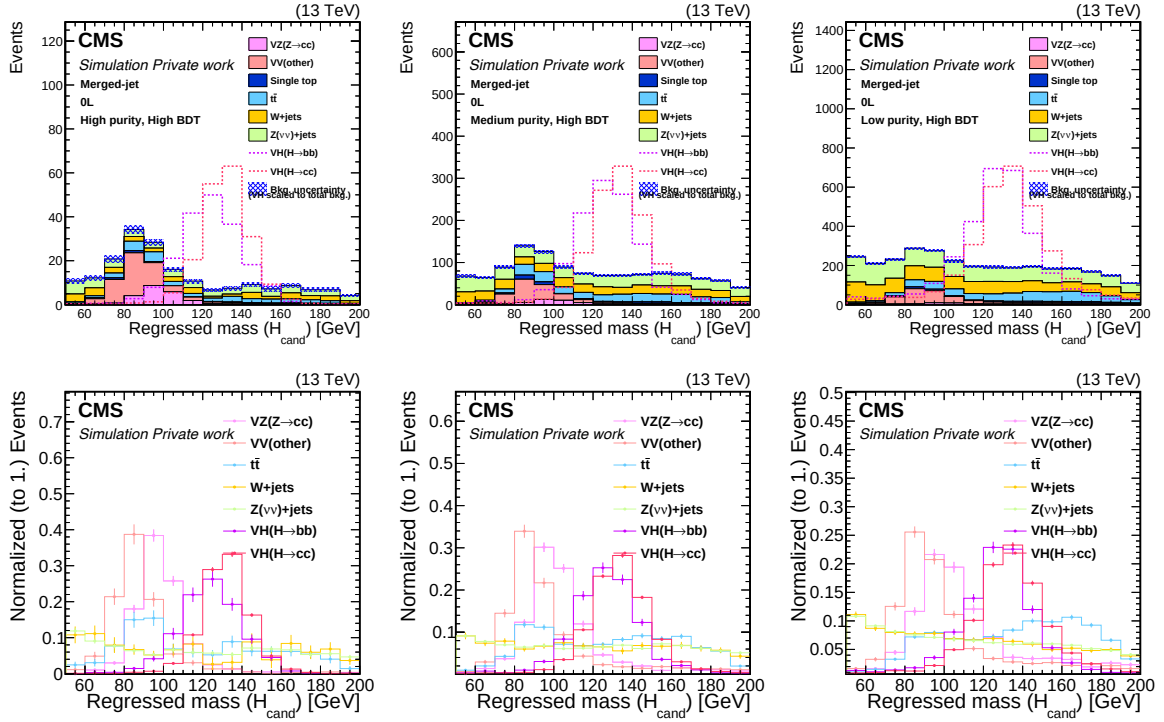


Figure 8.18 The H_{cand} mass distribution for events with high kinematic BDT output scores passing the HP (left), MP (middle) and LP (right) $X \rightarrow c\bar{c}$ tagging categories in the 0L channel. The distribution is made with backgrounds stacked and the $VH \rightarrow c\bar{c}$ signal normalized to the sum of all backgrounds (top), or all the $VH \rightarrow c\bar{c}$ signal and various background processes normalized to unity.

Table 8.5 A list of variables used in the training of the kinematic BDT for the 0L, 1L and 2L channels.

Variable	Description	0L	1L	2L
$p_T(V)$	vector boson transverse momentum	✓	✓	✓
$\Delta R(\ell, \ell)$	angular separation between the two leptons			✓
$p_T(H_{\text{cand}})$	H_{cand} transverse momentum	✓	✓	✓
$ \eta(H_{\text{cand}}) $	absolute value of the H_{cand} pseudorapidity	✓		
$\Delta\phi(V, H_{\text{cand}})$	azimuthal angle between vector boson and H_{cand}	✓	✓	✓
p_T^{miss}	missing transverse momentum		✓	
$\Delta\eta(H_{\text{cand}}, \ell)$	difference in pseudorapidity between H_{cand} and the lepton		✓	
$\Delta\eta(H_{\text{cand}}, V)$	difference in pseudorapidity between H_{cand} and vector boson			✓
$\Delta\eta(H_{\text{cand}}, j)$	min. difference in pseudorapidity between H_{cand} and small- R jets	✓	✓	✓
$\Delta\eta(\ell, j)$	min. difference in pseudorapidity between the lepton and small- R jets		✓	
$\Delta\eta(V, j)$	min. difference in pseudorapidity between vector boson and small- R jets			✓
$\Delta\phi(\vec{p}_T^{\text{miss}}, j)$	azimuthal angle between $\vec{p}_T^{\text{miss}}$ and closest small- R jet	✓		
$\Delta\phi(\vec{p}_T^{\text{miss}}, \ell)$	azimuthal angle between $\vec{p}_T^{\text{miss}}$ and lepton		✓	
m_T	transverse mass of lepton $\vec{p}_T + \vec{p}_T^{\text{miss}}$		✓	
N_j	number of small- R jets	✓	✓	✓

Table 8.6 A summary of event selection criteria for the signal region. Note that the values listed for kinematic variables are in units of GeV, and for angles in units of radians; the selections are listed in the format of (muon, electron) if they are varied between lepton flavours.

Variable	0L	1L	2L
$p_T(V)$	> 200	> 150	> 150
$M(\ell\ell)$	—	—	(75, 105)
p_T^ℓ	—	($> 25, > 30$)	> 20
$p_T(H_{\text{cand}})$	> 200	> 200	> 200
$M_{\text{PNet}}(H_{\text{cand}})$	(50, 200)	(50, 200)	(50, 200)
$\Delta\phi(V, H_{\text{cand}})$	> 2.5	> 2.5	> 2.5
N_j	< 2	< 2	< 3
$N_{a\ell}$	$= 0$	$= 0$	—
p_T^{miss}	> 200	> 60	—
$\Delta\phi(\vec{p}_T^{\text{miss}}, j)$	> 0.5	—	—
$\Delta\phi(\vec{p}_T^{\text{miss}}, \vec{p}_T^{\text{miss}}(\text{trk}))$	< 0.5	—	—
$\Delta\phi(\vec{p}_T^{\text{miss}}, \ell)$	—	< 1.5	—
Lepton isolation	—	($< 0.06, \text{—}$)	($< 0.2, \text{—}$)
Kinematic BDT	> 0.55	(0.55, 0.7), > 0.7	> 0.55
$X \rightarrow c\bar{c}$ tagging discriminant:			
High purity (HP)	> 0.99	> 0.99	> 0.99
Medium purity (MP)	(0.96, 0.99)	(0.96, 0.99)	(0.96, 0.99)
Low purity (MP)	(0.90, 0.96)	(0.90, 0.96)	(0.90, 0.96)

8.5 Signal extraction

The $H \rightarrow c\bar{c}$ signal is extracted from a binned maximum likelihood fit in the H_{cand} mass. This is shown in the previous H_{cand} mass distributions, namely Figs. 8.12, 8.15, and 8.18. The signal characterizes a peak structure around the m_H , while the background is relatively flat within the mass window. This demonstrates how improvements in the ParticleNet $X \rightarrow c\bar{c}$ tagging and mass regression performance enhance the search sensitivity: the former technique enables a significant suppression in the background events, and the latter results in better resolution in signal and a further suppression of the background near the m_H .

It is crucial to calibrate the simulated event yield for all signals and backgrounds. The calibration of the $H \rightarrow c\bar{c}$ signal and $H \rightarrow b\bar{b}$ background jets is introduced in Section 8.2.1. For the rest of the background events, their normalization is constrained in the CR. Overall, the normalization is acquired from a simultaneous fit for SRs and CRs to data, and the normalization factor is shared across dedicated SRs and CRs, dependent on the background types and the channel types. The normalization factor is not shared among the data-taking years. The subtleties, hence, lie in the way of decomposing backgrounds and the determination of SR and CR groups to share the normalization, as introduced below.

8.5.1 Control regions

The CRs designed in the analysis include low-BDT regions and high- N_j regions. The low-BDT regions are mainly composed of V +jets backgrounds, and the high- N_j regions are only defined in 0L and 1L channels to collect a high-purity $t\bar{t}$ sample. In both types of CRs, it is important to note that events are required to satisfy the same $X \rightarrow c\bar{c}$ discriminant criteria as in the SRs. As it is believed that the $X \rightarrow c\bar{c}$ tagging algorithms may be a main source to bring simulation–data discrepancy, it is crucial to keep SRs and CRs defined under the same $X \rightarrow c\bar{c}$ tagger selections, and the splitting of CR and SR is uncorrelated with the tagger. This guarantees that the normalization factor is transferable from corresponding CRs to SRs.

Background decomposition

Within each channel and data-taking year, the V +jets background is considered as a whole. This is because, although a tight $X \rightarrow c\bar{c}$ tagger selection results in flavour content change within each background (e.g. the $V + c/cc$ will be dominant within V +jets process), the flavour content should be similar across the SR and CRs, because they are split by kinematic BDT or N_j variables which are decorrelated with any flavour information. In this case, a single

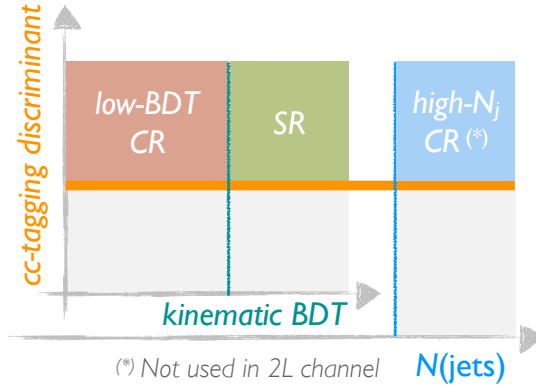


Figure 8.19 Illustration of the setup of the signal region and control regions.

type of background with several flavour components can be safely considered as a whole, with only one normalization factor acting on it and shared across SRs and CRs. Validations have been showcased in Section 8.4, where the $X \rightarrow c\bar{c}$ discriminant shape does not change in the high- or low-BDT region.

For the $t\bar{t}$ background, the estimation relies on a high- N_j control region. The change in the small- R jet multiplicity can affect the composition of different $t\bar{t}$ components, particularly, the relative fraction of the top-matched component (i.e. all three quarks from the top quark decay are contained in a large- R jet) and the W -matched component (i.e. only the two quarks from the W boson decay are contained in the large- R jet). Hence, one cannot treat $t\bar{t}$ backgrounds as a whole. The $t\bar{t}$ process is split into three components:

- $t\bar{t}$ (Top): all three quarks from the hadronic top quark decay are contained in the large- R jet, i.e., satisfying $\Delta R(q, \text{large-}R \text{ jet}) \leq 1.5$.
- $t\bar{t}$ (W): only the two quarks from the hadronic W boson decay are contained in the large- R jet, i.e., satisfying $\Delta R(q, \text{large-}R \text{ jet}) \leq 1.5$, while the b quark from the top decay is outside the jet cone, i.e., $\Delta R(b, \text{large-}R \text{ jet}) > 1.5$.
- $t\bar{t}$ (Other): if the large- R jet is neither top-matched nor W -matched (e.g., the large- R jet can contain the b quark and only one of the two quarks from the W boson decay in this category).

Figure 8.20 shows the comparison between the low- and high- N_j regions for the three $t\bar{t}$ components in the 0L channel, respectively, as an example. The shapes of the $X \rightarrow c\bar{c}$ tagging discriminant and the mass of the H_{cand} are very similar between the low- and high- N_j regions for the top- and W -matched components. Therefore, the normalization scale factors derived from the high- N_j region can be directly applied to the low- N_j region. For the $t\bar{t}$ (Other) component, a larger difference is observed; thus, the normalization scale factors will not be

shared between the low- and high- N_j regions for this component.

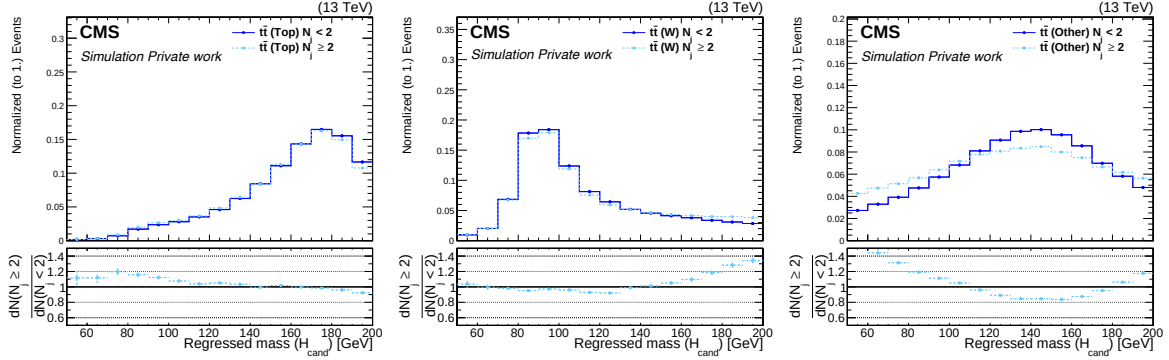


Figure 8.20 Comparison of the H_{cand} mass shapes between events with $N_j < 2$ and $N_j \geq 2$ for the top-matched (left), W -matched (middle) and other (right) components of the $t\bar{t}$ process in the 0L channel.

However, to increase the statistical power, the same background components are specified to share the same normalization scale factors in different lepton channels after their similar distributions are verified. Specifically, the Z +jets normalization scale factors are shared between the 0L and 2L channels for each $X \rightarrow c\bar{c}$ tagging category and data-taking year, and the W +jets normalization scale factors are shared between the 0L and 1L channel. To handle it more properly, an additional 20% uncertainty is assigned to account for possible differences in the normalizations between the channels due to small differences in the event kinematics. For each component of $t\bar{t}$, the normalization scale factor is shared between 0L and 1L channels and also with the single top (denoted by ST) process with the same decomposition, as they are expected to have very similar characteristics. The normalization scale factor for $t\bar{t}$ (W) component is also applied to the diboson process with a hadronic W decay (VW). Except for $t\bar{t}$ (Other), the same normalization scale factor is applied to the signal region and the low-BDT and high- N_j control regions, while for $t\bar{t}$ (Other) a separate normalization scale factor is applied to the high- N_j due to possible difference between the high- and low- N_j regions.

As a summary, the decomposition list of all background processes with the specific merging of components among them is shown below.

- Z +jets 0L+2L (with additional 20% uncertainty for 0L/2L difference),
- W +jets 0L+1L (with additional 20% uncertainty for 0L/1L difference),
- $t\bar{t}$ (Top) 0L+1L, with ST (Top) 0L+1L,
- $t\bar{t}$ (W) 0L+1L, with ST (W) 0L+1L, and VW 0L+1L+2L,
- $t\bar{t}$ (Other) 0L+1L (for SR and low-BDT CR), with ST (Other) 0L+1L (for SR and low-BDT CR)

- $t\bar{t}$ (Other) 0L+1L (for high- N_j CR), with ST (Other) 0L+1L (for high- N_j CR)

8.6 Systematic uncertainties

In this section, various categories of systematic uncertainties that affect the analysis are discussed. They contribute to the nuisance parameters that participate in the fit by varying the fitted templates or the overall yields. The contributions from the theoretical and experimental uncertainties are listed separately as follows.

Theoretical uncertainties

- **Parton distribution function (PDF):** effects of the uncertainty in the PDF is assessed by reweighting each event according to the 100 replicas of the NNPDF set, and taking the standard deviation of the 100 variations, following the prescription in^[178]. Only effects on the shape are considered, i.e. the weights are rescaled to preserve the cross section of each process.
- **Renormalization and factorization scales:** uncertainties due to the renormalization scale μ_R and the factorization scale μ_F are assessed by varying the scales independently by a factor of 0.5 or 2. The effects of such variations are implemented as weights in the event generators. Only effects on the shape are considered, while the inclusive cross section remains unchanged.
- **Cross section of the single top and diboson processes:** the single top and diboson backgrounds are predicted directly from simulation. An uncertainty of 15% is assigned for the single top process based on the CMS cross section measurements^[179, 180]. For diboson process, a 5% uncertainty is assigned based on the CMS cross section measurements^[181-183].
- **The p_T spectra of diboson processes:** the simulated diboson samples used in this analysis are generated at NLO QCD precision. Higher order NNLO QCD + NLO EWK corrections are used to reweight the simulated samples. Uncertainties due to this reweighting are taken into account as a shape uncertainty.
- **Cross section of VH processes:** uncertainties on the VH production cross section due to uncertainty of PDF and the choice of the renormalization and factorization scales are taken into account in this analysis.
- **The p_T spectra of VH processes:** the simulated VH samples used in this analysis are generated at NLO QCD precision. Higher order NLO electroweak and NNLO QCD

corrections, parametrized in $p_T(V)$, are used to reweight the simulated samples. Uncertainties due to this reweighting are taken into account as a shape uncertainty.

- **Treatment of the $VH \rightarrow b\bar{b}$ processes:** contribution of the $VH \rightarrow b\bar{b}$ processes is fixed to the SM expectation in this analysis, and theoretical uncertainties on the VH production cross sections and the $H \rightarrow b\bar{b}$ branching fractions are taken into account.

Experimental uncertainties

- **Jet energy scale and jet energy resolution:** two uncertainties from the jet energy scale are assessed by varying up and down the energy scale and resolution scale factor of each jet within one standard deviation. The uncertainties are treated as correlated between the small- R jets and the large- R jets, by varying the energy scales of the small- R jets and the subjects of the large- R jets simultaneously. Effects of the variation are also propagated to p_T^{miss} .
- **The p_T^{miss} unclustered energy:** the uncertainty in p_T^{miss} related to unclustered energy is assessed by varying each particle type by its own resolution.
- **L1 prefiring issue:** the 2016 and 2017 data are affected by the Level 1 ECAL prefiring issue. Therefore, the simulation samples are reweighted according to the predefined recipe, and uncertainties on the reweighting are propagated.
- **HEM15/16 issue:** the 2018 data are partially affected by the failure of HEM15/16. In addition to the mitigation measures, as a general recipe, an additional uncertainty on the jet energy scale is assigned for 2018 by scaling down the jet energy by 20% for jets with $-1.57 < \phi < -0.87$ and $-2.5 < \eta < -1.3$, and 35% for jets with $-1.57 < \phi < -0.87$ and $-3.0 < \eta < -2.5$, and also propagating it to the p_T^{miss} .
- **Lepton efficiency:** the efficiency of the muon and electron trigger, reconstruction, identification and isolation requirements are determined in data using the standard tag-and-probe technique with Z bosons. Simulation-to-data scale factors are applied to correct the efficiency of the simulation. Uncertainties on the measured scale factors are evaluated from the statistical uncertainties of the measurement and efficiency differences using alternative samples and selections and are propagated to the analysis.
- **MET trigger efficiency:** the efficiency of the MET triggers affects the expected yields of signals and of background processes that are predicted directly from simulated samples in the 0L channel. An uncertainty of 2% is estimated as in the $VH \rightarrow b\bar{b}$ analysis.
- **Pileup reweighting:** effects due to the uncertainty in the distribution of the number of

pileup interactions are evaluated by varying the cross section used to predict the number of pileup interactions in simulation by 4.6% from its nominal value.

- **Pileup jet identification efficiency:** the efficiencies and misidentification rates of the pileup jet identification algorithm applied to the small- R jets are corrected by the simulation-to-data scale factors. Uncertainties on scale factors are propagated in this analysis.
- **Size of simulated samples:** uncertainties due to the limited size of the simulated samples are taken into account via the Barlow-Beeston method^[184].
- **Statistical uncertainty of control regions in data:** the limited size of the control regions causes uncertainty in the normalization of the V +jets and $t\bar{t}$ process measured from the control regions.
- **Integrated luminosity:** the uncertainty in the integrated luminosity measured by CMS for the 2016, 2017 and 2018 data-taking period is taken into account.
- **Efficiency of $X \rightarrow c\bar{c}$ tagging with ParticleNet:** efficiency of tagging a pair of c quarks with the ParticleNet algorithm is measured in data. Simulation-to-data correction factors are applied to the $VH \rightarrow c\bar{c}$ signals and the $VZ \rightarrow c\bar{c}$ process.

Chapter 9

Analysis Results

9.1 $VZ \rightarrow c\bar{c}$ validation results

The full procedure of this analysis is validated by measuring the VZ production with $Z \rightarrow c\bar{c}$. The event selection, including the kinematic BDT, the selection on the $X \rightarrow c\bar{c}$ tagging discriminant and the signal extraction procedure, remains unchanged. The only difference is that, instead of treating $VH \rightarrow c\bar{c}$ as the signal process, we treat the $VZ \rightarrow c\bar{c}$ process as the signal, while $VH \rightarrow c\bar{c}$ is treated as a background and fixed to the expectation of the SM prediction.

Figures 9.1–9.2 show the comparison of the expected signal and background and the observed data in the CRs after the simultaneous fit of all CRs and SR, taking the data collected in 2018 and the 1L channel as an example. The corresponding plots for the SRs are shown in Figs. 9.3–9.4. The best-fit $VZ \rightarrow c\bar{c}$ signal strength is $\mu_{VZ \rightarrow c\bar{c}} = 0.87^{+0.25}_{-0.21}$, consistent with SM expectation.

Table 9.1 shows the expected and observed significance on $VZ \rightarrow c\bar{c}$ as well as the best-fit signal strength $\mu_{VZ \rightarrow c\bar{c}}$.

Table 9.1 Expected and observed significance and best-fit signal strength for $VZ \rightarrow c\bar{c}$.

	0-lepton	1-lepton	2-lepton	combined	$\mu_{VZ \rightarrow c\bar{c}}$
2016 (expected, simulation-only)	2.0σ	0.8σ	1.2σ	2.7σ	$1.00^{+0.51}_{-0.40}$
2016 (expected, CR-only fit)	2.1σ	0.8σ	1.3σ	2.8σ	$1.00^{+0.49}_{-0.39}$
2016 (expected, CR+SR fit)	1.9σ	0.9σ	1.1σ	2.8σ	—
2016 (observed, CR+SR fit)	2.8σ	0.9σ	0.0σ	1.6σ	$0.54^{+0.40}_{-0.34}$
2017 (expected, simulation-only)	2.8σ	1.2σ	1.8σ	3.8σ	$1.00^{+0.42}_{-0.31}$
2017 (expected, CR-only fit)	2.7σ	1.2σ	1.7σ	3.8σ	$1.00^{+0.41}_{-0.31}$
2017 (expected, CR+SR fit)	3.0σ	1.2σ	1.7σ	3.8σ	—
2017 (observed, CR+SR fit)	3.4σ	1.1σ	2.2σ	4.0σ	$1.08^{+0.44}_{-0.32}$
2018 (expected, simulation-only)	2.4σ	1.1σ	1.9σ	3.4σ	$1.00^{+0.54}_{-0.35}$
2018 (expected, CR-only fit)	2.6σ	1.3σ	1.9σ	3.6σ	$1.00^{+0.53}_{-0.34}$
2018 (expected, CR+SR fit)	2.1σ	1.3σ	1.7σ	3.3σ	—
2018 (observed, CR+SR fit)	1.6σ	1.6σ	1.6σ	3.1σ	$1.13^{+0.67}_{-0.46}$
Full Run2 (expected, simulation-only)	4.1σ	1.8σ	2.8σ	5.6σ	$1.00^{+0.27}_{-0.22}$
Full Run2 (expected, CR-only fit)	4.3σ	2.0σ	2.9σ	5.9σ	$1.00^{+0.26}_{-0.21}$
Full Run2 (expected, CR+SR fit)	4.2σ	2.0σ	2.6σ	5.7σ	—
Full Run2 (observed, CR+SR fit)	4.7σ	2.1σ	1.6σ	4.9σ	$0.87^{+0.25}_{-0.21}$

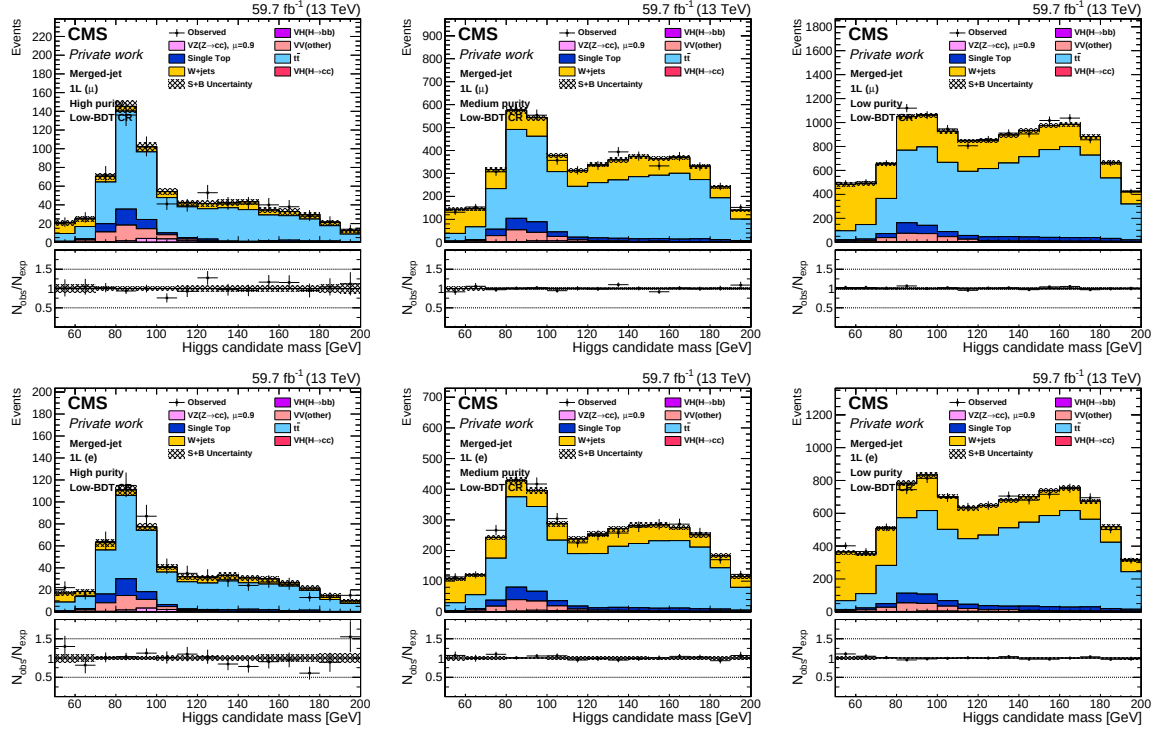


Figure 9.1 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the low-BDT control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

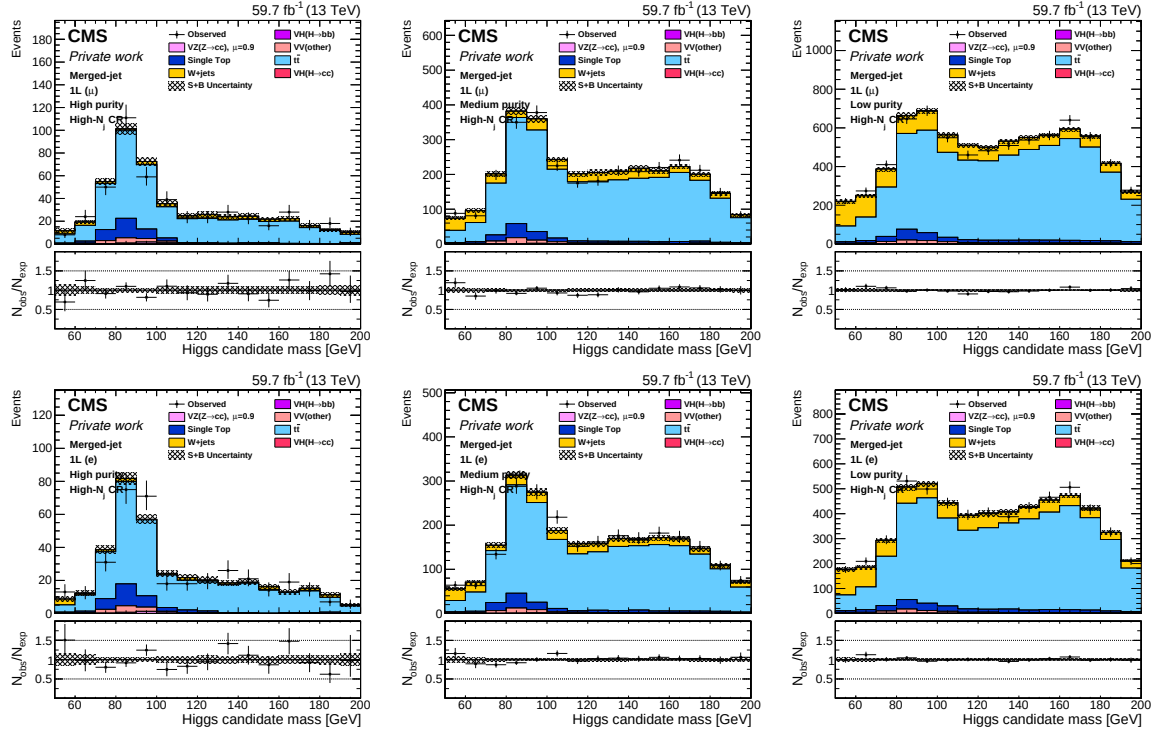


Figure 9.2 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the high N_j control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

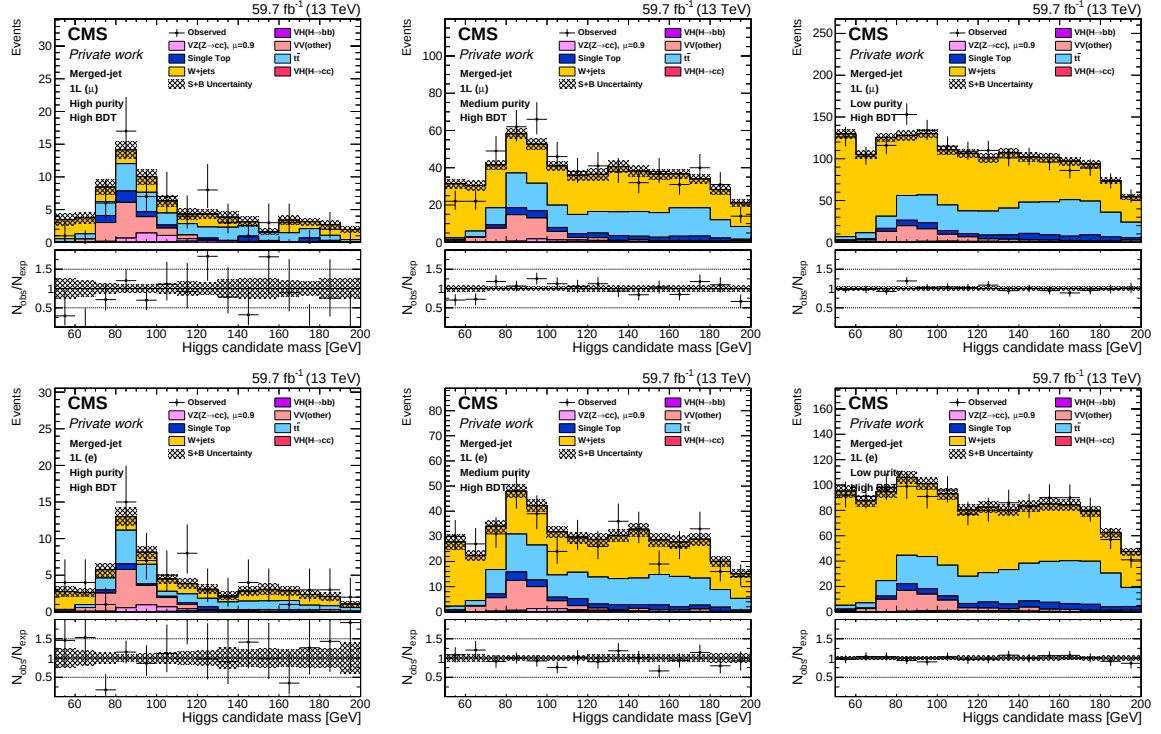


Figure 9.3 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the high-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

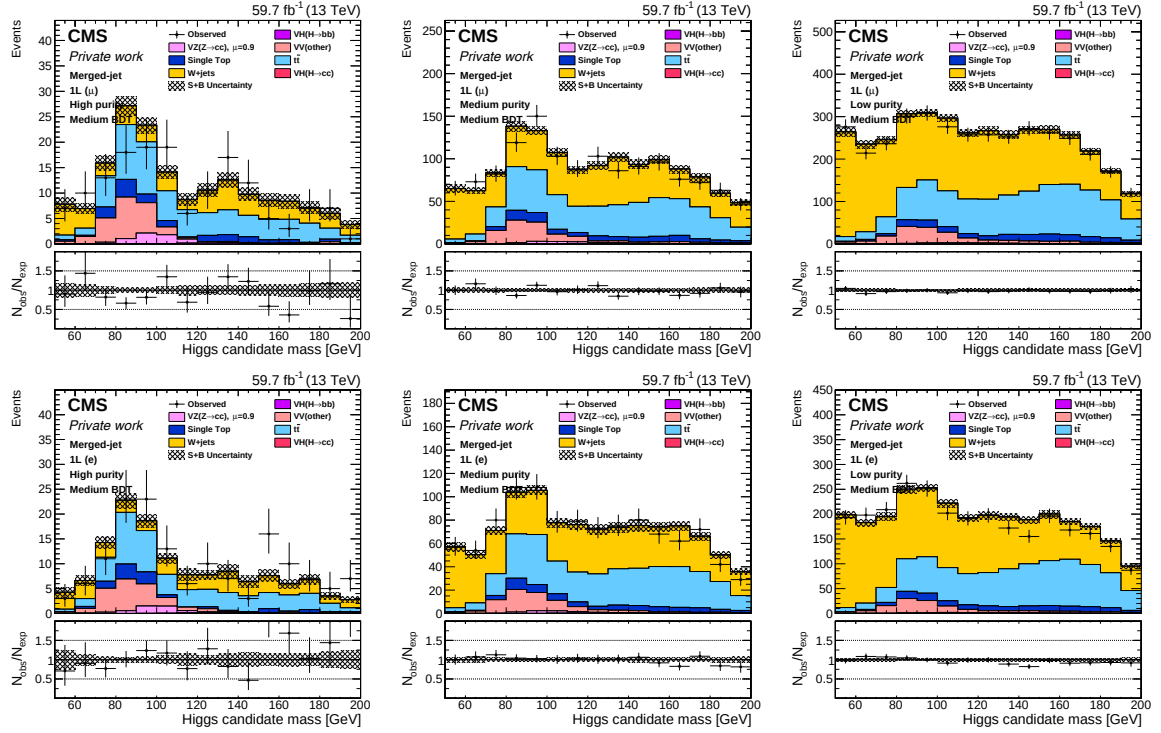


Figure 9.4 Post-fit plots from the $VZ \rightarrow c\bar{c}$ validation fit for the medium-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

9.2 $VH \rightarrow c\bar{c}$ results

The extraction of the $VH \rightarrow c\bar{c}$ signal is performed via a binned maximum likelihood fit on the mass of H_{cand} with the SRs and CRs. Figures 9.5–9.6 show the comparison of the expected signal and background and the observed data in the CRs after a simultaneous fit of all the CRs and SRs, taking the 2018 data and the 1L channel as an example. Data from all three years are fit simultaneously. The corresponding plots for the high-BDT and medium-BDT SRs are shown in Figs. 9.7–9.8.

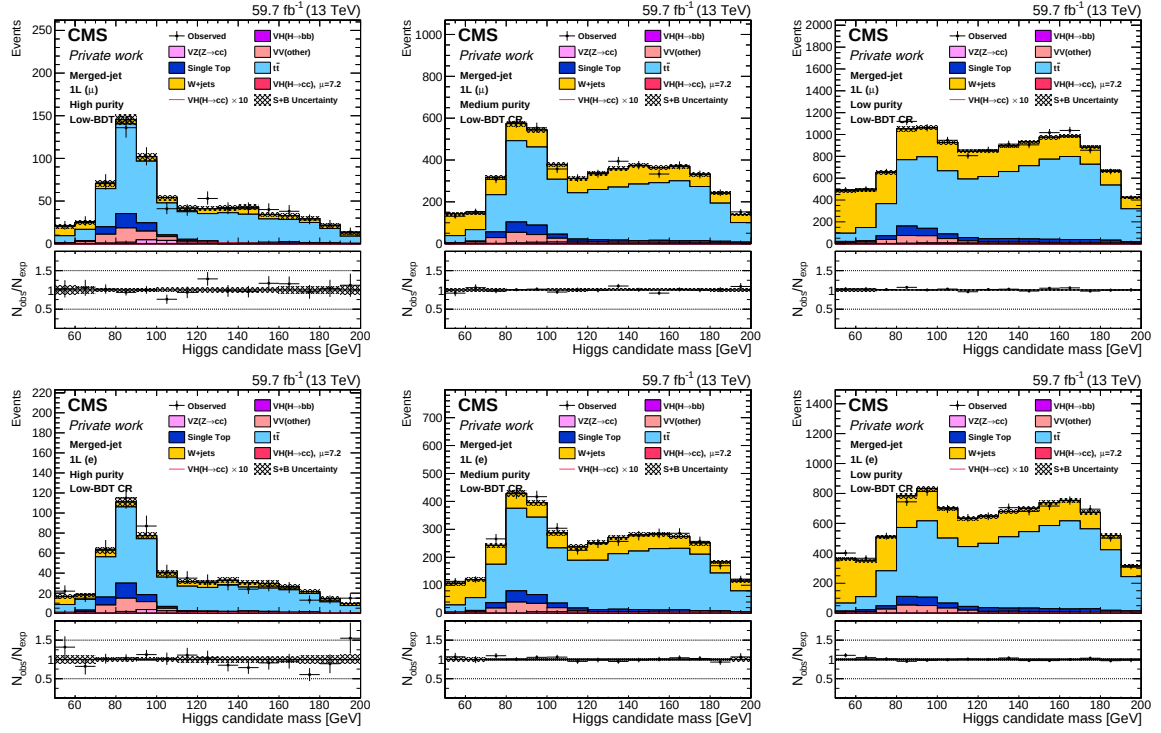


Figure 9.5 Post-fit plots for the low-BDT control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

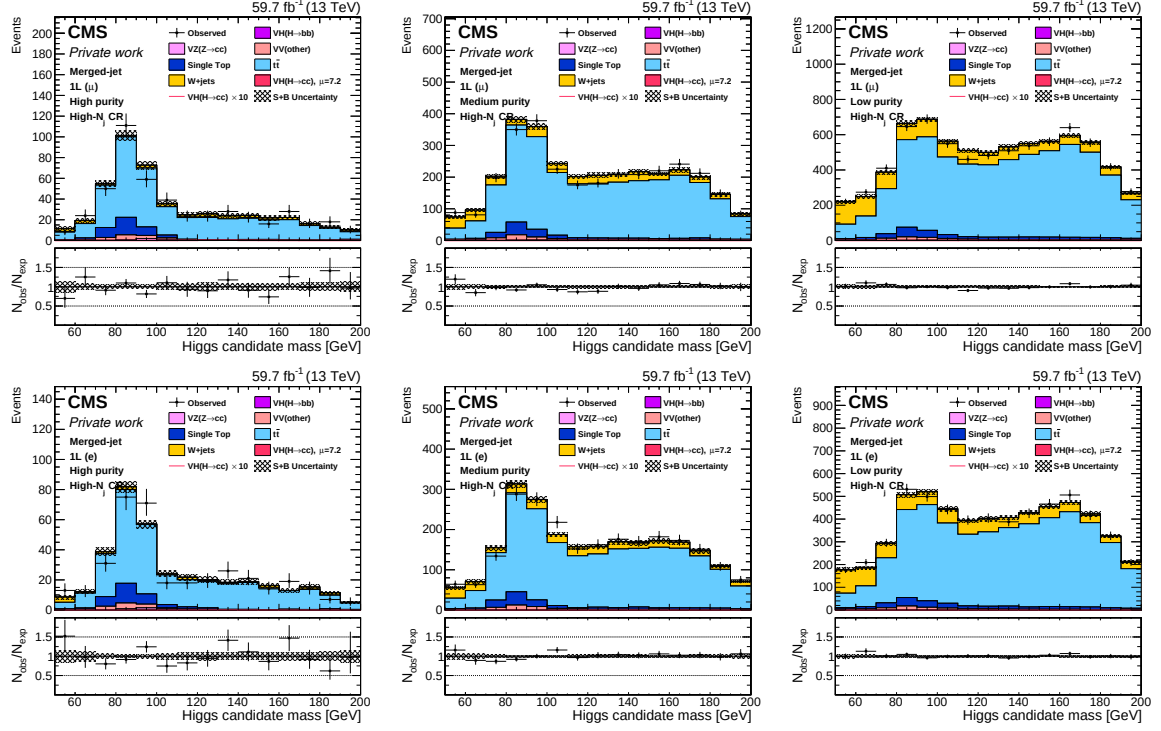


Figure 9.6 Post-fit plots for the high N_j control regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

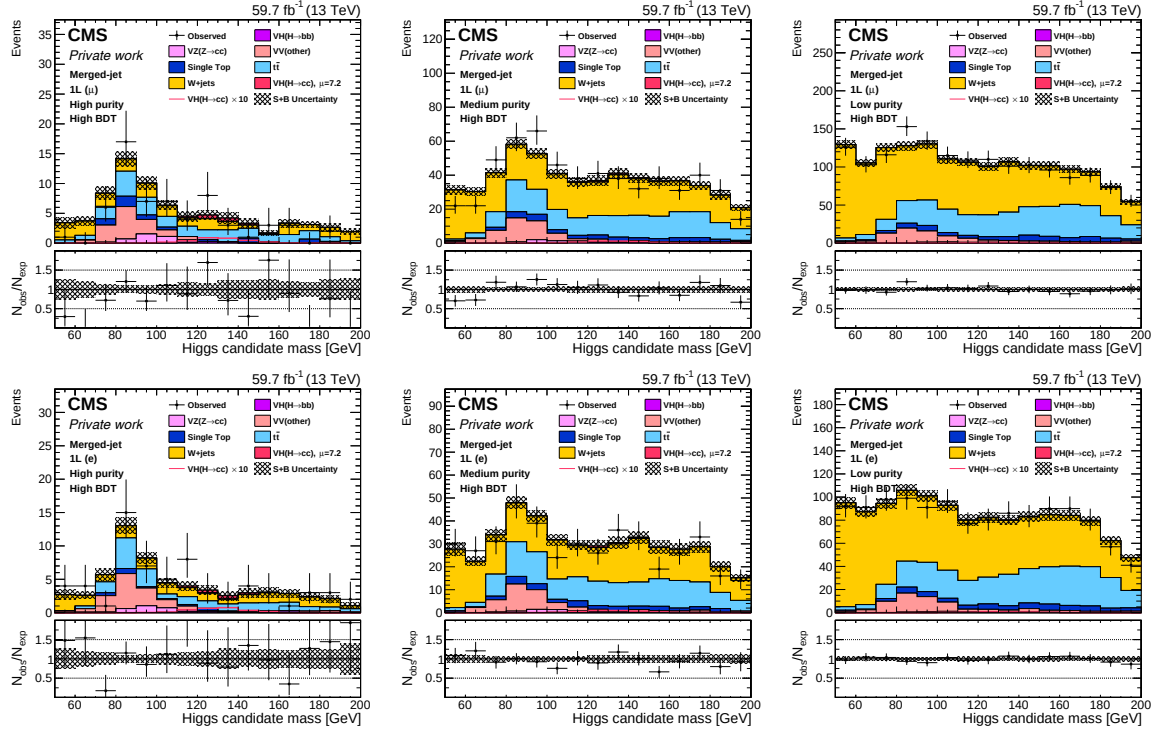


Figure 9.7 Post-fit plots for the high-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

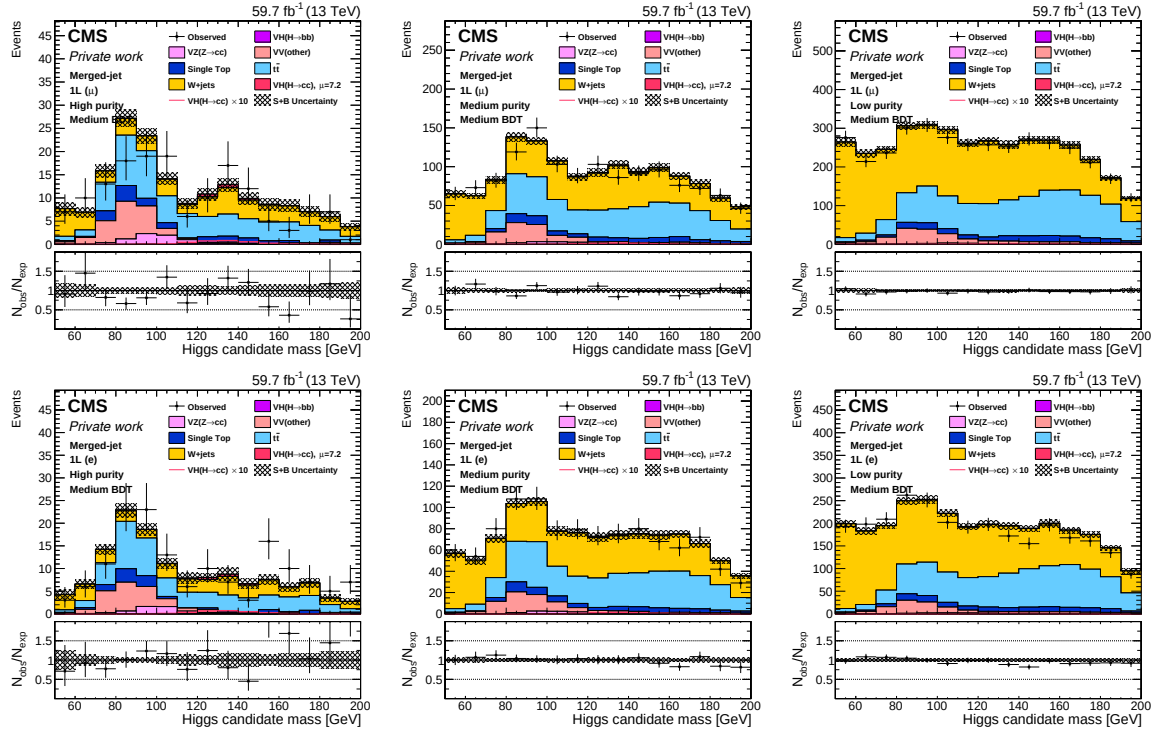


Figure 9.8 Post-fit plots for the medium-BDT signal regions in HP (left), MP (middle) and LP (right) categories of the 1-lepton channel in 2018 data. Top: $W(\mu\nu)$; Bottom: $W(e\nu)$.

The $VH \rightarrow c\bar{c}$ signal yields are scaled by the best-fit signal strength, which is measured to be $\mu_{VH \rightarrow c\bar{c}} = 7.2^{+4.4}_{-4.2}$. The post-fit yields for background and signal processes and the observed data in the signal regions are summarized in Tables 9.2–9.4.

In addition, Fig. 9.9 shows the post-fit values of the nuisance parameters and their impacts on the $VH \rightarrow c\bar{c}$ signal strength. Only the first 30 nuisance parameters with the largest impacts on the $VH \rightarrow c\bar{c}$ signal strength are shown. No significant excess of events is observed beyond the expected contribution from SM processes. The observed (expected) upper limit on the $VH \rightarrow c\bar{c}$ signal strength $\mu_{VH \rightarrow c\bar{c}}$ is 15 (8.5) at 95% confidence level. The results in each category and year are summarized in Table 9.5.

Table 9.2 Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2016 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.

Process	High purity				Medium purity				Low purity			
	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT
V +jets	69.9	55.2	128.2	8.6	792.1	443.8	898.1	119.0	2560.1	1521.1	2981.1	369.6
$t\bar{t}$	18.5	25.4	74.9	0.0	145.0	244.5	631.6	0.0	344.6	565.7	1503.5	0.0
Single top	2.8	7.8	17.2	0.0	14.1	49.0	98.5	0.0	36.7	104.1	227.2	0.0
$VZ \rightarrow c\bar{c}$	16.0	3.9	7.5	2.5	25.3	6.8	12.5	4.2	22.8	5.9	13.8	4.1
$VV(\text{other})$	27.6	19.3	20.8	3.2	104.4	63.1	110.1	13.6	200.8	106.5	221.1	28.3
$VH \rightarrow b\bar{b}$	0.6	0.5	0.4	0.2	1.9	1.5	1.5	0.6	2.9	2.0	2.5	0.9
Total background	135.5	112.0	249.1	14.4	1082.8	808.8	1752.2	137.4	3167.9	2305.3	4949.2	402.9
$VH \rightarrow c\bar{c}$ (SM)	2.6 (0.4)	2.3 (0.3)	2.4 (0.3)	0.9 (0.1)	4.3 (0.6)	3.6 (0.5)	4.3 (0.6)	1.6 (0.2)	3.5 (0.5)	2.7 (0.4)	3.4 (0.5)	1.2 (0.2)
Data	132	113	261	18	1081	846	1710	139	3270	2335	4875	394

Table 9.3 Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2017 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.

Process	High purity				Medium purity				Low purity			
	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	1-lepton Medium BDT	2-lepton High BDT
V +jets	57.2	38.4	80.2	9.1	620.0	372.9	775.3	108.6	1972.3	1232.4	2630.4	333.2
$t\bar{t}$	18.0	28.8	79.2	0.0	144.4	226.5	613.4	0.0	344.5	549.4	1516.3	0.0
Single top	3.4	7.3	16.8	0.0	15.2	42.4	97.1	0.0	34.1	96.6	217.4	0.0
$VZ \rightarrow c\bar{c}$	19.4	5.7	9.6	4.0	33.1	8.2	18.7	7.1	33.2	9.2	20.9	7.1
$VV(\text{other})$	26.3	15.8	30.3	4.0	92.8	66.9	106.1	12.9	182.6	104.6	198.4	25.8
$VH \rightarrow b\bar{b}$	0.6	0.5	0.4	0.2	1.7	1.2	1.3	0.5	2.7	1.8	2.2	0.9
Total background	124.8	96.4	216.6	17.3	907.2	718.1	1611.9	129.2	2569.4	1994.0	4585.6	366.9
$VH \rightarrow c\bar{c}$ (SM)	3.5 (0.5)	3.0 (0.4)	3.3 (0.5)	1.2 (0.2)	5.5 (0.8)	4.7 (0.7)	6.3 (0.9)	2.3 (0.3)	4.7 (0.7)	3.7 (0.5)	5.3 (0.7)	2.0 (0.3)
Data	119	99	214	19	879	764	1582	142	2620	2024	4496	363

Table 9.4 Post-fit yields for background and signal processes and the observed data in the signal regions of each channel for the 2018 dataset. The SM expected $VH \rightarrow c\bar{c}$ signal yields are indicated in parenthesis.

Process	High purity				Medium purity				Low purity			
	0-lepton High BDT	1-lepton High BDT	Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	Medium BDT	2-lepton High BDT	0-lepton High BDT	1-lepton High BDT	Medium BDT	2-lepton High BDT
V +jets	80.3	45.7	94.9	20.9	782.0	545.7	1182.8	141.7	2455.7	1783.8	3883.2	436.1
$t\bar{t}$	22.3	43.0	124.9	0.0	162.9	305.6	879.5	0.0	389.4	750.4	2099.6	0.0
Single top	2.7	9.1	25.4	0.0	19.6	54.3	137.3	0.0	40.8	125.2	296.3	0.0
$VZ \rightarrow c\bar{c}$	20.6	6.9	12.3	4.3	37.1	11.1	23.1	8.7	36.8	10.7	25.9	9.1
$VV(\text{other})$	39.3	26.8	44.0	5.7	133.1	85.4	167.4	20.9	252.6	145.8	300.4	39.8
$VH \rightarrow b\bar{b}$	1.2	1.2	1.1	0.4	3.7	3.4	3.7	1.4	5.6	4.5	5.8	2.2
Total background	166.4	132.7	302.5	31.4	1138.3	1005.4	2393.8	172.7	3181.0	2820.4	6611.2	487.2
$VH \rightarrow c\bar{c}$ (SM)	3.6 (0.5)	3.7 (0.5)	3.9 (0.5)	1.5 (0.2)	6.2 (0.9)	5.9 (0.8)	8.0 (1.1)	2.9 (0.4)	5.1 (0.7)	4.6 (0.6)	6.8 (0.9)	2.6 (0.4)
Data	169	127	304	33	1138	1002	2378	179	3138	2819	6437	467

Table 9.5 Observed and expected upper limits at 95% CL on the $VH \rightarrow c\bar{c}$ signal strength $\mu_{VH \rightarrow c\bar{c}}$ and the best-fit value.

	0-lepton	1-lepton	2-lepton	combined	$\mu_{VH \rightarrow c\bar{c}}$
2016 (expected, simulation-only)	41	30	37	18	—
2016 (expected, CR-only fit)	39	32	38	19	—
2016 (expected, CR+SR fit)	36	31	42	19	—
2016 (observed, CR+SR fit)	39	41	35	21	$2.8^{+8.7}_{-9.3}$
2017 (expected, simulation-only)	25	23	28	13	—
2017 (expected, CR-only fit)	26	24	28	13	—
2017 (expected, CR+SR fit)	25	25	28	13	—
2017 (observed, CR+SR fit)	32	26	31	19	$5.2^{+6.8}_{-6.2}$
2018 (expected, simulation-only)	30	26	28	14	—
2018 (expected, CR-only fit)	32	25	29	14	—
2018 (expected, CR+SR fit)	35	25	33	15	—
2018 (observed, CR+SR fit)	58	28	51	23	$8.5^{+7.8}_{-6.7}$
Full Run2 (expected, simulation-only)	16	14	16	7.9	—
Full Run2 (expected, CR-only fit)	16	14	16	8.0	—
Full Run2 (expected, CR+SR fit)	16	14	17	8.5	—
Full Run2 (observed, CR+SR fit)	26	20	21	15	$7.2^{+4.4}_{-4.2}$

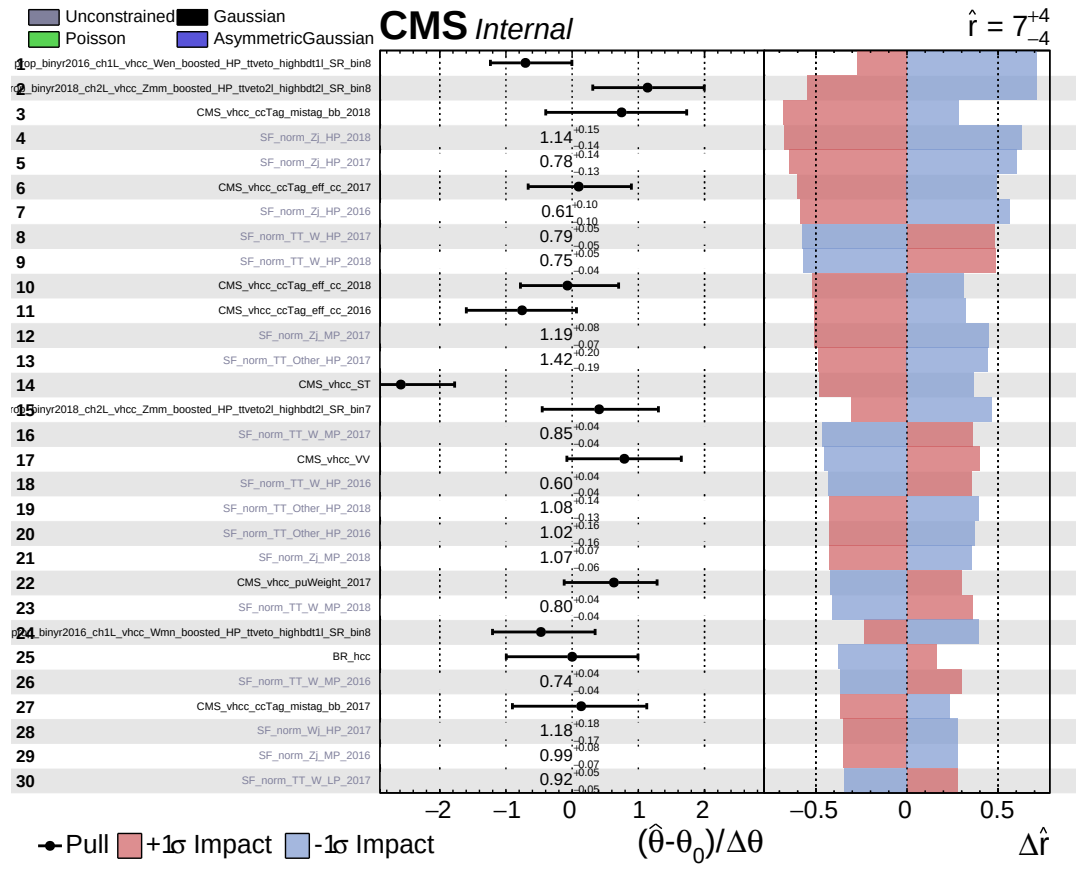


Figure 9.9 Post-fit values and the impacts of the selected nuisance parameters on the $VH \rightarrow c\bar{c}$ signal strength, focusing on those with the largest impacts.

9.3 Combined $VH \rightarrow c\bar{c}$ results with resolved-jet analysis

To further improve the sensitivity of the $VH \rightarrow c\bar{c}$ search, the merged-jet analysis is statistically combined with the resolved-jet approach. To achieve orthogonality between the two analyses, events are categorized based on whether a large- R jet satisfying the requirement described in Section 8.3.1 is present and passing an additional p_T threshold. The p_T threshold of the large- R jet has been optimized to maximize the sensitivity in $VH \rightarrow c\bar{c}$ and $VZ \rightarrow c\bar{c}$ while keeping a robust and stable fit, and a p_T threshold of 300 GeV is found to give the best results.

The correlation scheme followed for the treatment of the systematic uncertainties between the two analyses is summarized below:

- The theoretical uncertainties are treated as fully correlated between the resolved-jet and merged-jet analyses.
- The normalization SFs for $t\bar{t}$ and V +jets backgrounds are treated as uncorrelated between the two analyses. This is because the two analyses probe different phase spaces. In addition, in the merged-jet analysis, the tagger efficiencies for V +jets and $t\bar{t}$ are estimated directly from the CRs and reflected in these SFs; therefore, they cannot be correlated with the SFs in the resolved-jet topology.
- Uncertainties related to the charm-tagging SF are treated as uncorrelated. The procedure followed for the calibration of the two algorithms (DeepJet and ParticleNet) is significantly different and also probes a very different phase space.
- Experimental uncertainties related to JEC, JER, p_T^{miss} , pileup and lepton efficiencies are treated as correlated between the two analyses.

The full analysis procedure in the combination scheme is first validated by measuring the VZ production with $Z \rightarrow c\bar{c}$. This best-fit signal strength for $VZ \rightarrow c\bar{c}$ is found to be $\mu_{VZ \rightarrow c\bar{c}} = 1.01^{+0.23}_{-0.21}$, with an observed (expected) significance of 5.7 (5.9) standard deviations.

For the $VH \rightarrow c\bar{c}$ signal process, no significant excess of events is observed beyond the expected contribution from SM processes. The observed (expected) upper limit on the $VH \rightarrow c\bar{c}$ signal strength $\mu_{VH \rightarrow c\bar{c}}$ is 14 (7.6) at 95% confidence level. The best-fit signal strength is measured to be $\mu_{VH \rightarrow c\bar{c}} = 7.7^{+3.8}_{-3.5}$. The result is interpreted in the κ -framework^[185], where $\mu_{VH \rightarrow c\bar{c}}$ is reparameterized in terms of the Higgs-charm Yukawa coupling modifier κ_c , assuming that only the Higgs boson decay widths are modified, according to the relation:

$$\mu_{VH \rightarrow c\bar{c}} = \frac{\kappa_c}{1 + (\kappa_c^2 - 1) \cdot \mathcal{B}_{\text{SM}}(H \rightarrow c\bar{c})}. \quad (9.1)$$

The observed 95% CL interval for κ_c is $1.1 < |\kappa_c| < 5.5$, with the expected constraint given by $|\kappa_c| < 3.4$.

9.4 Discussion and outlook

The $VH \rightarrow c\bar{c}$ analysis provides the most direct constraint to $\mu_{VH \rightarrow c\bar{c}}$ and κ_c to data after a combination of the merged-jet and resolved jet topologies. The observed sensitivity is driven by the resolved-jet result, nevertheless, the expected upper limit of $\mu_{VH \rightarrow c\bar{c}}$, extracted from simulated events, are led by the advancement in the merged-jet analysis. The improved sensitivity in the merged-jet topology reveals the benefits of exploiting the advanced deep-learning technology to analyse complex data. The techniques mainly involve the advancement in DNN architecture based on ParticleNet to analyze the information contained in the large- R jet with the large radius $R = 1.5$, and it is assisted with sufficient training samples and appropriate training configurations to make the DNN expressive.

In addition, the ParticleNet models are accommodated with the VHcc analysis workflow to make it valid for data analysis. This includes the tagger's reasonable decorrelation performance with the jet mass to allow signal extraction from the H_{cand} mass spectra, using a sampling-based reweighting as an improved mass-decorrelation technique compared to the predecessor, the DeepAK15-MD tagger^①. Notably, the robust and reasonable calibration of the ParticleNet-MD $X \rightarrow c\bar{c}$ discriminant on the signal and background jets is also vital in the analysis. The signal calibration is done by a novel sfBDT method, and the background is calibrated by dedicated CRs and the shared normalization factors.

Based on the merged-jet analysis workflow, the sensitivity of the $VH \rightarrow c\bar{c}$ search is projected to HL-LHC corresponding to the 3000 fb^{-1} data. From the likelihood scan on $\mu_{VH \rightarrow b\bar{b}}$ and $\mu_{VH \rightarrow c\bar{c}}$, shown in Fig. 9.10, the expected best fit of $\mu_{VH \rightarrow c\bar{c}}$ is roughly 1.0 ± 0.8 , yielding an expected upper limit of $\mu_{VH \rightarrow c\bar{c}}$ at 95% C.L. at around 1.6. The results can be improved with further enhancement of the DNN technique, e.g. by exploiting a more performant network architecture, as discussed in Chapter 4. As a concrete practice within the CMS experiment, Chapter 11 will introduce the next-generation large- R jet tagger, which improves by a factor of 1.5–2 on the $X \rightarrow c\bar{c}$ tagging identification against the major backgrounds. This provides one direction to enhance sensitivity in the merged-jet analysis. For resolved-jet analysis, it can be a viable solution to incorporate signature identification with advanced DNNs using the dijet

① The mass decorrelation of the DeepAK15-MD tagger is achieved by including an adversarial training approach. The technique is discussed in Section 5.1, the introduction of DeepAK8-MD.

or full event-level information with proper analysis workflow designs. Overall, for the future of the $VH \rightarrow c\bar{c}$ analysis, there is more to expect regarding improvement in the analysis techniques combined with the deep learning advances, and it is surely a more charming journey ahead.

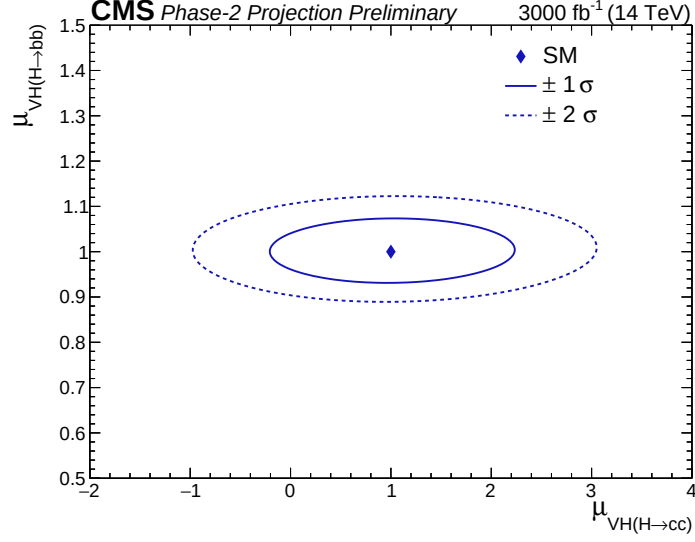


Figure 9.10 The profile likelihood scan on the signal strengths $\mu_{VH \rightarrow b\bar{b}}$ and $\mu_{VH \rightarrow c\bar{c}}$ projected to HL-LHC corresponding to the data of 3000 fb⁻¹. The projection is based on the merged-jet analysis workflow with proper estimation of the statistical uncertainties^[113, 186].

Part V

Application Paradigm for Pre-training and Fine-tuning

Chapter 10

Perspectives on Future Jet Models

The rapid development of deep learning is driving significant advances in AI, sparking a revolution in all aspects of human activity, including scientific research. This progress is expected to become a driving force for the next industrial revolution. Over the past five years, the particle physics community has witnessed the large success of modern deep learning in tasks such as large- R jet tagging. In the coming decades, the basic form of research into fundamental particle physics will still likely depend on accelerators and particle collisions without a major change. In this regard, given the incredible speed of the AI revolution, it is both intriguing and important to explore the potential role of AI in enhancing the processing of data at the LHC and converting this data into meaningful physics results.

This part of the dissertation will extend beyond the specific application of ParticleNet in the CMS experiment and will instead explore how more advanced AI and deep learning techniques could transform future analyses of LHC physics. In particular, it will highlight the potential of these methods in jet tagging and their broader application in physics analyses, focusing on the benefits of pre-training large, comprehensive models and their subsequent fine-tuning.

It is important to note that, at this stage of development, state-of-the-art deep learning tools for analysis still rely heavily on modularized event processing (e.g. models developed to analyze reconstructed jets, leptons, etc.). This is also the basis of the work documented in this chapter. However, in the future, when AI is more advanced, one may foresee more radical paradigm shifts in processing data, potentially moving away from the traditional reliance on jet-based models.

This chapter will begin with a discussion of various perspectives on the future development of jet models.

10.1 Jet tagging problem revisited

As new jet models subsequently surpass their predecessors in tagging performance, a fundamental inquiry arises in the community: where do the performance boundaries of a jet

tagging task lie, and how do we embed the problem into a theory framework to understand the limit? It is worth noting that, the question can be generalized for all classification problems, e.g. the problem to classify signal and background events, which lies at the heart of almost all LHC analyses.

The statistical theory is the answer to the question. For each collision event recorded by the detector (or a jet composed of reconstructed particle candidates), based on the underlying information used to represent it, it can be mapped to a single point on an abstract latent space, although the dimension can be very high. One can imagine that this is achievable if every calorimeter cell or tracker hit is considered one dimension in that space. Then, events in each physics process (or jets from each class) are obtained by sampling points from a prior probability density function (PDF) on that space. If this prior PDF is known, then the best discriminant one can build to separate two classes is simply the ratio of their PDFs, according to the Neyman–Pearson lemma^[187]. This provides a theoretical bound for the classification performance one can reach.

Additionally, it implies that DNNs have provided the best practical solution so far. This is because the aim of DNN training is to manipulate a tunable and complex enough function, our neural network, to estimate the unknown and complex PDF ratio. For example, in the binary classification task to separate signal and background, the most optimal network, with its training loss reaching the theoretical minimum, will output probabilities exactly corresponding to the ratios of signal and background PDFs. This estimation lies within the maximum likelihood estimation problem because the network only sees limited points sampled from the PDF. When the number of data points is huge, and the capacity of the neural network is massive, the optimized neural network will estimate a function closer to the prior PDF ratio, touching the upper bound. This picture, brought from the statistical view, intuitively matches the recent advances in jet tagging, where studies show that a larger dataset size and a more expressive and bigger network are the keys to yield better classification performance.

More importantly, it is not hard to see that this statistical view actually has deep relations with particle physics itself—the aforementioned prior PDF for a given physics process corresponds to the *differential cross sections* based on all constructed dimensions. For instance, in the example of discriminating ggF $H \rightarrow b\bar{b}$ events and QCD multijet events with DNN, the estimation target is the ratio of their PDFs, which also corresponds to the ratio of their fully differential normalized cross sections in every event phase space. This gives an underlying physics meaning when one delivers the DNN classification tasks in the particle physics

scenario. In contrast, in other domains, e.g. the image classification task where one trains an optimal vision network to separate images of cats and dogs, the interests lie merely on the classification performance metric but not in the underlying statistical interpretation. That is to say, less interest is put on the PDF that lives in the dimensions of pixels and colour channels of an image, formed by the entire image class of cats or dogs.

The above discussion introduces a way to interpret the DNN classification task applied to particle physics and reveals its physics meaning. As an inspiration for future analyses, it is then imaginable that, with enhanced use of deep learning tools, one should be capable of reaching near-optimal signal and background discrimination for every analysis. Note that the improvement can be especially sizeable for hadronic channels, which have more complex detector responses. Finding a suitable experiment technique to reach the goal is a pursuit for future analyses delivered at the LHC, and it can imply a revolution in how physicists analyze LHC data. The development may involve various topics, such as the design of the network, the format of input it should proceed with, its training strategy, its integration into the LHC analyses workflow, etc. This part of the dissertation will present a first yet promising exploration of these directions using the fast simulation dataset and its implementation within the CMS experiment.

10.2 Future models

This section explains the design philosophy of several future models from the author's perspective. It focuses on three aspects of the model: its expressiveness, large capacity, and comprehensiveness. Amidst the rapid development of deep learning, experiences can be drawn from CV and NLP when designing the future jet model.

Expressiveness

Progress has been made when it comes to using a DNN to deal with particle-level inputs. Early strategies include using an MLP or nested neural network structure like CNN to deal with particle data^[79-81, 188]; then, the community recognized that the structure for a point cloud or a graph is closer to the essence of particle-format data and improves the expressiveness of the neural network^[87, 88, 90]. Advanced neural network architectures involve GNNs or Transformers^[65, 88-90, 93, 105], and will continue to lead the future direction in this domain. These architectures respect the inherent properties of particle-format data, including their invariance or near-invariance to a set of symmetries, including permutation of particles, Lorentz

transformations or its subsets of the particles. (details can be found in Section 4.1.1). Specifically, experiences in deep learning suggest that Transformers are very suitable for building large models as they scale well with parameter and data sizes^[56, 189], and a similar conclusion is drawn in jet physics domain^[65].

As these advances are made in processing particle-format data, a special note is that the foreseeable ultimate network that processes collider data will likely handle the raw detector information directly. The community has also witnessed the development of such models, such as CNNs to process energy deposition in calorimeters or GNNs to handle raw hits in trackers (taking Refs.^[190, 191] as examples). Presently, the work introduced in this dissertation still focuses on the particle-level input instead of the raw input, for it delivers the jet tagging task where jets are clustered from the reconstructed particle—an intermediate format processed from the raw data.

Large capacity

The large capacity of the network is the key to achieving a closer estimate towards the prior PDF. Recent advances in CV and NLP have witnessed rapid development in large vision models and language models. For example, Fig. 10.1 illustrates the evolution of parameter counts in large models over time. It can be foreseeable that the high-energy physics community will also encounter large collider event models or large jet models. The Transformer developed in this work has reached 10 M parameters, which is smaller than those language or vision models by a factor of thousands, but it has reached 20 times the parameters of ParticleNet developed in CMS. Studies presented in this dissertation reveal that the size of the present Transformer model is compatible with the current amount of data used for training.

Comprehensiveness

The invention of the Transformer^[56] has not only made it easier to train larger models but has also invigorated another subdomain, the “pre-training and fine-tuning” training paradigm (such as Refs.^[50, 51] in the NLP domain). This is a two-step strategy: initially, a general model is trained on a broad task, enabling it to learn a comprehensive latent space to embed inputs; afterwards, this model can be fine-tuned for specific tasks. Given the abundant knowledge acquired during the pre-training process, this knowledge can be transferred to downstream tasks, achieving decent performance with a minimal sample size. The work presented in this dissertation illustrates that pre-training and fine-tuning have a lot of untapped potential in particle physics. They will be elaborated in the following section (Section 10.3) from the fine-tuning

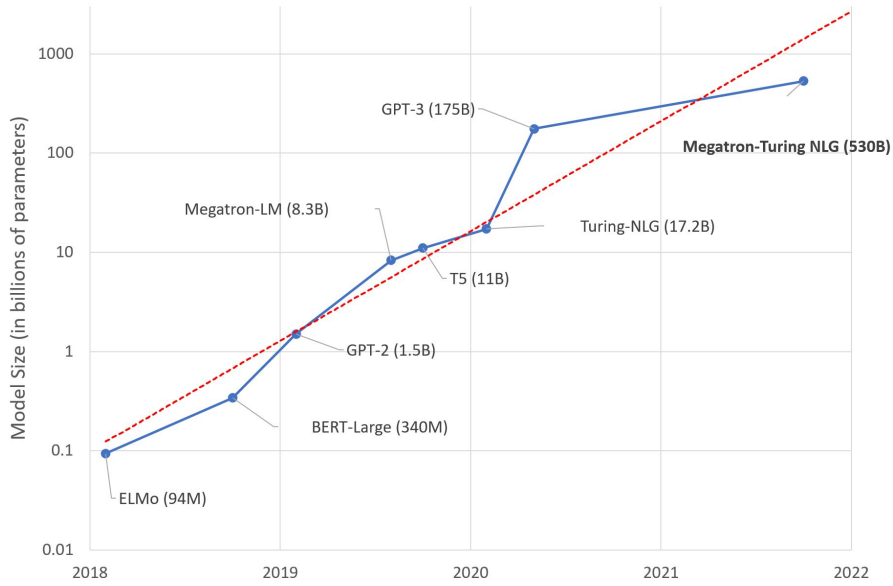


Figure 10.1 The evolution of the large language model size (by 2022). Source: *Hugging face*.

perspective, but a few remarks are presented first on how to make the pre-trained model comprehensive.

By reviewing the historical developments of model pre-training in the CV and NLP domain, one implication is a paradigm shift from supervised to self-supervised training^[50, 51, 192]. This is due to the nature that most image and textual data used for training is unlabelled. At the earlier stage, the model pre-training in a supervised approach relies on the inclusiveness of the training set and a large-scale classification task. Because the jet training data largely relies on simulation, the presence of truth labels is a key feature. As a first approach, this work relies on a supervised training approach. Similar to ImageNet, which contains up to 1000 different labelled images, extending the categories to the hundreds level would be a viable approach for the jet tagging dataset.

It is worth noting that there are two benefits to pre-training a large jet model via the large-scale classification approach to reach comprehensiveness. First of all, as suggested earlier, the classification problem lies at the centre of all physics analyses. An important fact to recognise is that a multiclass classifier can be seen as stacking multiple binary classifiers; that is, the output probabilities can be used to construct optimal discriminants to separate any two classes (or even a mixture of classes) from the initial hundreds of training classes. A theoretical explanation is given in Section 11.2.1 with a detailed validation, and proof can be found in Appendix B. This allows one to extend the usage of one model to a large number of analyses, reaching an optimal discrimination of different types of signals and backgrounds all at once. Secondly,

as model-agnostic searches are gaining momentum due to the lack of new observations of a BSM particle, models resulting from large-scale classification can be an approach to detect potentially unseen signatures. As the model has seen a variety of simulated jets and learned their pattern, it will ensure an easily reached generalization ability.

10.3 Fine-tuning techniques

This section aims to explain why “pre-training and fine-tuning” paradigm has a large potential in LHC experiments. An intriguing fact is that, although there have not been many discussions on this paradigm in the LHC analyses, it has subtly integrated into the design logic of some analyses and is poised to emerge.

As a thought experiment, we consider an analysis using a neural network-based AK4 jet b and c tagger. The traditional approach involves applying a specific working point of the tagger, but when the number of required b and c jets is large, selecting the tagging discriminant might drastically reduce signal efficiency. A common practice is to employ machine learning tools (a BDT or a neural network) to deliver a multivariate analysis, taking the tagger score of each jet as input. This is essentially an event classifier for signal versus background discrimination. It is intriguing to see that the paradigm is much akin to performing transfer learning on the output score of the jet model—jet models (the jet taggers) tend to be more complex due to the need to handle low-level input features and require larger training datasets; in contrast, the analysis-level neural network is typically an MLP containing fewer parameters. Hence, it can be regarded as transferring the knowledge from the more complex model to the event level.

To expand this application scenario to a real transfer learning scene, replacing the input tagger scores with the hidden layer neuron values read from the pre-trained model is simply possible. It means that one can discard the last 1–2 layers of the jet model and connect the hidden neurons directly to the subsequent neural network. This is depicted in Fig. 10.2. In typical CMS analyses, this approach suits just well: one treats the last hidden layer neurons as new jet features, storing them to the “analysis object data” used for delivering analyses, then uses them as inputs for the event-level neural network training.

The hidden layer neurons should provide a comprehensive jet representation as long as the jet model is pre-trained comprehensively on a large variety of data. Utilizing them as input provides powerful ability in the classification task. This is because the pre-trained model essentially finds a way to map the abstract jet phase space comprised of features of all particles (with varying numbers) to a fixed-dimensional phase space. This space is sufficiently versatile

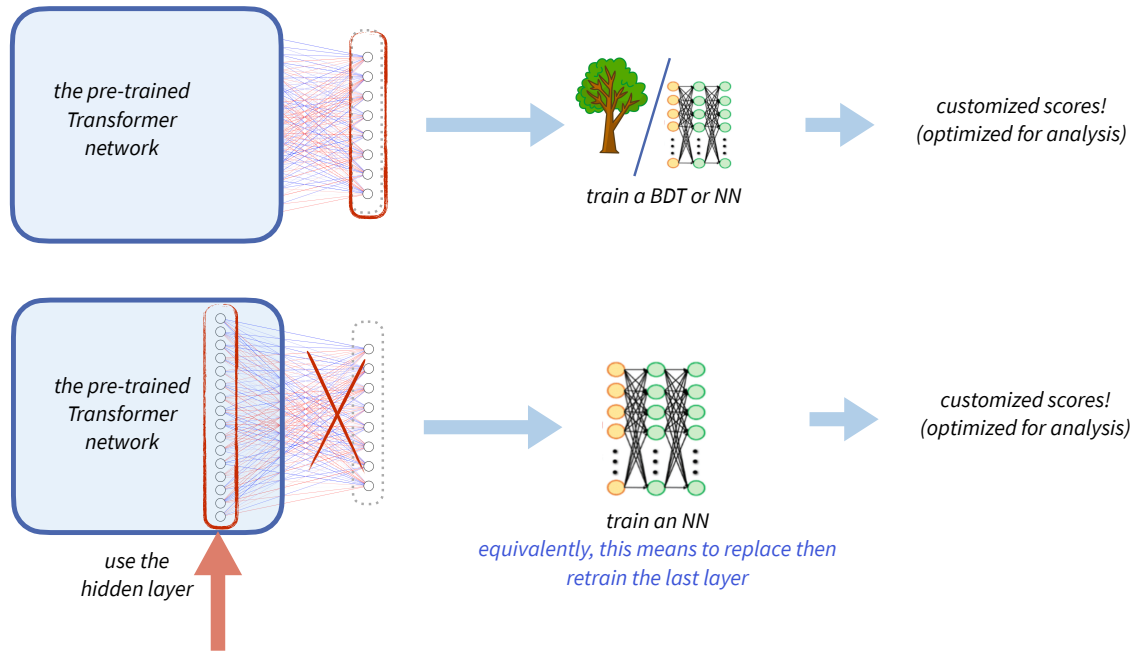


Figure 10.2 Illustration on the connection from the conventional analysis-level neural network training strategy to the concept of transfer learning.

and has strong discrimination power over essential jet attributes, as a few simple MLP layers in the original model can transform this latent representation into optimal discriminants capable of distinguishing any two kinds of signatures.

The following chapters in this dissertation will showcase the new developments following this spirit of design, inside and outside of the CMS experiment. In Chapter 11, a pre-trained model is developed as a large-scale classifier over hundreds of jet classes in CMS. The model is dubbed Global Particle Transformer (GloParT). Then, in Chapter 12, two fine-tuning application schemes within the CMS experiment are studied on GloParT using the simulated event to illustrate its superiority compared to existing approaches. The philosophy of GloParT training is also studied using the fast-simulation dataset fast-simulated produced via DELPHES, as detailed in Chapter 13. This chapter introduces the JETCLASS-II dataset, along with the “Sophon model”, which is close to the philosophy of the GloParT training dataset and the GloParT model. The development process will be detailed, demonstrating the vast potential of the Sophon model to boost the resonance search programmes at the LHC, which inspires the future meaning of the GloParT model in the CMS experiment.

Chapter 11

GloParT: Model Pre-training

Pre-training a DNN jet model that is expressive, large, and comprehensive is the key to enable a high performant fine-tuning applications to CMS analyses. Hence, this chapter introduces detailed specifics about model pre-training, which is performed on a large-scale classification and regression and exploits the ParT architecture as the backbone. Due to the comprehensive large- R phase spaces the model sees and learns, it is named the Global ParT (or GloParT). Its development can be regarded as an all-around extension of the ParticleNet tagger and the regression model introduced in Section 8.2.1. This chapter will illustrate the technology being used, but will also attempt to explain the principle of such techniques in the statistical view ^①.

The training samples and details on the categorization are first introduced in the following text. The development details are shown, highlighting several control experiments on the new features incorporated in model pre-training that have not been studied before. The performance on a number of metrics is provided in the last section.

11.1 Generated samples and jet labels

The training of GloParT is a simultaneous classification and regression task. The large-scale classification is delivered by separating 314 different jet final-state topologies. Like the other jet tagging model, it is designed to distinguish the jets initiated from the QCD multijet process (QCD jets) and various types of jets initiated by a resonance, which usually constitute the signal in dedicated analyses. In order to enlarge the classes of the resonance jets, they are expanded in three dimensions: (1) the multiprongness of the jet that is highly related to the resonance decay process, (2) the combinations of specific particle types and flavour that initiate each prong, and (3) the kinematics of the secondary resonance if it is valid. This results

^① Author’s note (edited November, 2024): The model presented in this dissertation is referred to as the “GloParT stage-2” model, which represents the second iteration of the model developed within the CMS collaboration. This model has undergone further refinement, and the “GloParT stage-3” model was integrated into the CMS software in November 2024. This updated model is expected to enhance upcoming CMS analyses using full Run 3 data, as well as joint Run 2 and Run 3 data, in the coming years.

5. The categories for QCD jets are split into five classes based on their flavour composition, shown in the fifth circle.

The generated events contain the QCD multijet events and the customized BSM events to acquire different types of resonant jets. The QCD jets are collected from the p_T -binned QCD multijet events generated by PYTHIA. The initial ME production is a $2 \rightarrow 2$ QCD process, with the p_T of one object configured within the range of (350, 470), (470, 600), (600, 800), (800, 1000), (1000, 1400), (1400, 1800), (1800, 2400), (2400, 3200), and (3200, $+\infty$) GeV.

For the customized BSM events that produce resonance jets, one commonality for samples is their flat control of the jet p_T and mass distribution. Like the previous jet tagger designs, identification of a certain final-state signature is required to be decoupled with the jet mass and p_T such that the analysis selection on the trained tagging discriminant will not sculpt the background distributions on these basic kinematics features. To achieve this, the original resonance is guaranteed to span over a wide range of p_T and mass and yield a flat distribution. This poses a requirement for the event production, which is described as follows.

11.1.1 Generation of two-prong resonant jets

Events that produce resonance jets are initiated by a di- H production process, which effectively doubles the jet sample size. We utilize both resonance and the nonresonance production of the di- H process.

For the first case, the events used for producing the scalar H decaying to two prongs are initiated by a diresonance process, $\tilde{G} \rightarrow HH$, generated by MG at LO, which is similar to the ParticleNet-MD tagger development described in Section 8.2.1. To refine the kinematics of the jet, we use the mass of $\tilde{G}$ to control the H scalar p_T (hence the jet p_T) and the H mass to control the jet m_{SD} . The mass of the resonance $\tilde{G}$ is set within the range of 600–6000 GeV, with a step size of 100 GeV. For mass of H is set to 15–250 GeV. Specifically, within the range of 15–30 GeV, the step size is 5 GeV, and within 30–250 GeV, the step size is 10 GeV.

The decay of the scalar H particle is modelled by PYTHIA with the two-body decay configuration. The decay modes and their branching ratios are configured in PYTHIA, including $b\bar{b}$, $c\bar{c}$, $s\bar{s}$, $u\bar{u}$, $d\bar{d}$, gg , e^+e^- , $\mu^+\mu^-$, $\tau^+\tau^-$, with their production rates being $1 : 1 : 1 : \frac{1}{2} : \frac{1}{2} : 1 : \frac{1}{2} : \frac{1}{2} : 2$. Here, the production rates correspond to the ratio of event numbers for each class used for training to ensure that we can effectively utilize all generated jets. The $u\bar{u}$, $d\bar{d}$ decays will be merged to form the light-flavour $H \rightarrow q\bar{q}$ category. The tau leptons can undergo leptonic decay to generate an electron or a muon together with three neutrinos (the tau lepton

is denoted as τ_e or τ_μ), or hadronic decay (denoted as τ_h). This is processed in PYTHIA with SM setup.

For the second case, the events are produced by a nonresonance di- H production, where the scalar H in this case is a neutral or charged Higgs boson in the two-Higgs-doublet model (2HDM). The generated process is $pp \rightarrow H^+ H^- / H^0 H^0$, where the decay of $H^\pm$ and H^0 particles is subsequently specified as $H^{\pm,0} \rightarrow bc/cs/bs$. The introduction of $H^{\pm,0}$ decay modes thus brings three additional categories. Among them, $H^\pm \rightarrow cs$ is a most useful class because the corresponding probability score can be used to identify $W \rightarrow cs$ signatures. This is not included in the ParticleNet-MD tagger. For handling the generated jet p_T span, we use the event H_T to control the jet p_T and configure it the same way as the $\tilde{G}$ mass in the first case. The jet m_{SD} span is controlled by the scalar H mass.

The parton showering of the event is proceeded by PYTHIA using the v8.230^[28], using the CP5 underlying event tune^[137] with the NNPDF3.1 NNLO parton distribution function (PDF) set^[138].

11.1.2 Generation of multipronged resonant jets

The multipronged resonance jets are composed of three different topologies. As indicated in Fig. 11.1, they are jets similar to the topology of a $H \rightarrow W^{(*)} W^{(*)}$ jet, $H \rightarrow Z^{(*)} Z^{(*)}$ jet, and $t \rightarrow bW$ jet. It is clear that they are not exactly like jets initiated from the SM process because all the resonance masses are reconfigured to span over a wide range; however, the decay modes of the resonances inherit the properties of SM W and Z bosons.

The $H \rightarrow W^{(*)} W^{(*)}$ jets

The events are produced with the same $\tilde{G} \rightarrow HH$ process simulated with MG. The sequential decay of $H \rightarrow WW \rightarrow q\bar{q}q\bar{q}/\ell\nu q\bar{q}$ is required, and it is modelled by JHUGEN to retain the spin correlation of two W bosons. Specifically, to configure the fully hadronic or semi-leptonic decay of the di- W , the (DecayMode1, DecayMode2) is set to (5, 5) and (10, 5), respectively. The production rates of fully hadronic over semileptonic modes are manually set as 1 : 2 to emphasise a bit more on the semileptonic jets. Under each main class of fully hadronic and semileptonic cases, the decay into specific quarks follows the branching ratio of SM. The hadronization of events is processed by PYTHIA.

To control the generated jet p_T and m_{SD} , the resonance $\tilde{G}$ mass and H mass adopt the same configuration as in the two-prong case introduced above.

Additionally, an important aspect is to configure the mass of the secondary resonance mass, i.e. the W boson mass, because different configurations in m_W/m_H result in different jet topologies. In application, one may hope that the tagging discriminant is optimal for tagging SM $H \rightarrow WW^*$ jets, but it may also be used in the BSM searches to identify a heavy resonance X decaying to on-shell WW signatures, or a lighter resonance decaying to $W^{(*)}W^*$ signatures but is different from $H \rightarrow WW^*$. It is found in our study that mixing all these phase spaces in one training category actually will result in less optimal performance for all scenarios. Therefore, two different sets of events are simulated:

1. the “fixed m_W/m_H ” sample, where m_W/m_H is always fixed in the SM configuration, i.e. $m_W = \frac{80.4}{125}m_H$.
2. the “spanned m_W/m_H ” sample, where m_W is sampled uniformly within $(0.3, 1.5) \times m_{W,0}$, and $m_{W,0} = \frac{80.4}{125}m_H$.

The two sets of samples are figuratively shown in Fig. 11.2, on the 2D plain of the generator configuration of H and W boson masses.

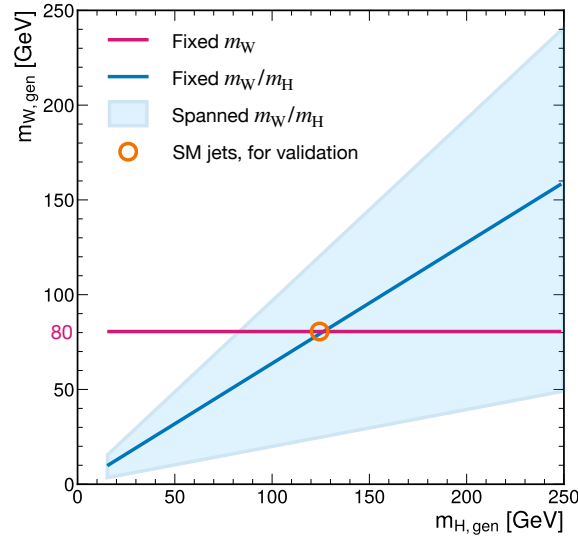


Figure 11.2 The GEN-level m_H and m_W configurations for all produced $H \rightarrow W^{(*)}W^{(*)}$ samples used for training and evaluating the GloParT model.

In the “spanned m_W/m_H ” sample, the resonance jets can be any case within the onshell–onshell, onshell–offshell, and offshell–offshell configuration. Therefore, to enhance the specific ability in tagging $H \rightarrow WW^*$ jets and $H \rightarrow W^{(*)}W^*$ jets, the sample is further split into two. This is done by reading the GEN-level W boson masses from the PYTHIA particle list

during the full event hadronization (denoted as m_{W_1} and m_{W_2}), and defining

$$\frac{|m_{W_1} - m_{W_2}|}{m_{W_1} + m_{W_2}} \begin{cases} < 0.1, & \text{defined as the } H \rightarrow WW \text{ region,} \\ > 0.1, & \text{denoted as the } H \rightarrow W^*W^{(*)} \text{ region.} \end{cases} \quad (11.1)$$

For better explaining this split, the 2D distribution of the variable $\frac{m_H - m_{W_1} - m_{W_2}}{m_H}$ against $\frac{|m_{W_1} - m_{W_2}|}{m_{W_1} + m_{W_2}}$ is shown in Fig. 11.3. Here, m_H is also the actual Higgs boson mass, which is slightly smeared in PYTHIA processing. In Fig. 11.3, the x axis measures how much of a Higgs's mass remains in energy that can be converted into particle kinetic energy after generating the two W bosons. The y axis measures the asymmetry between the two W bosons. These two metrics are the GEN-level indicators affecting the topology of the jet. It can be seen that a split on the y axis at 0.1 roughly separates the onshell–onshell topology versus the other case, and the split roughly yields 1 : 1 in the jet counts.

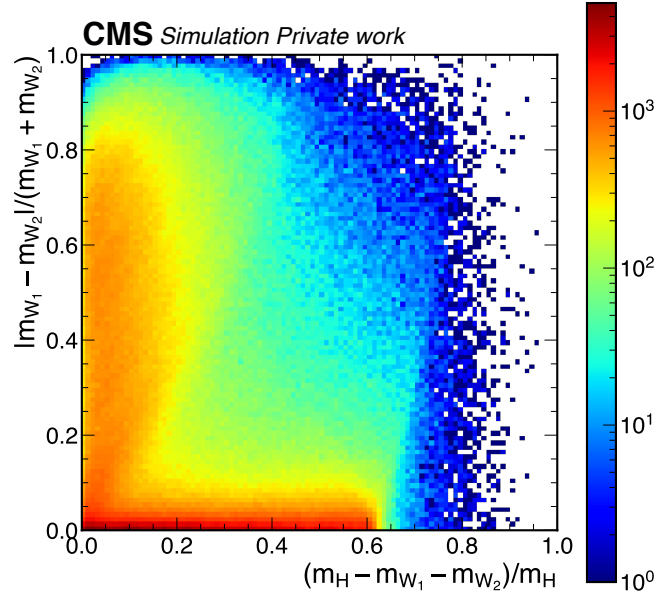


Figure 11.3 The 2D distribution of $(m_H - m_{W_1} - m_{W_2})/m_H$ and $(|m_{W_1} - m_{W_2}|)/(m_{W_1} + m_{W_2})$ for $H \rightarrow W^{(*)}W^{(*)}$ jets in the spanned- m_W/m_H sample.

Overall, the $H \rightarrow W^{(*)}W^{(*)}$ jets encompass three sets of categories, which differ by their event topology. One set targets tagging SM-like $H \rightarrow WW^*$ jets, and the other two aim to tag the signature of $H \rightarrow WW$ or $H \rightarrow W^*W^{(*)}$. Each category has the same fine jet categorization based on the W boson decay mode and the particle contents within the jet cone. They form the $H \rightarrow W^{(*)}W^{(*)}$ categories in Fig. 11.1.

The $H \rightarrow Z^{(*)}Z^{(*)}$ jets

Another bunch of multipronged jet phase spaces is brought by the $H \rightarrow Z^{(*)}Z^{(*)}$ topology, which is similar to the $H \rightarrow W^{(*)}W^{(*)}$ case but is complemented with additional jet content, including more quark flavour types (e.g. $b\bar{b}b\bar{b}$) and the $\ell\ell qq$ signature resulting from the semileptonic channel. As the Z boson has richer decay modes than W boson when they are accurate to the quark flavours, these jets bring a larger number of categories, as shown in Fig. 11.1.

The generation of events that produces $H \rightarrow Z^{(*)}Z^{(*)}$ jets is very similar to that produces $H \rightarrow W^{(*)}W^{(*)}$ except for the change of particle W to Z in the ME simulation step. The $\tilde{G}$ and H mass adopts the same configuration, and the Z boson mass is specified the same with W boson configuration: for the “fixed- m_Z/m_H ” sample, it adopts $m_Z = \frac{80.399}{125}m_H$; for the “spanned- m_Z/m_H ” sample, it is sampled uniformly from the span (0.3, 1.5) times the central value m_Z defined above. In JHUGEN, the fully hadronic or semileptonic decay of the di- Z is specified by setting (DecayMode1, DecayMode2) as (1, 1) and (8, 1), respectively. The rate over the two modes is also configured as 1 : 2. Still, three sets of topologies result from the sample production, which split the full categories threefold.

The $t \rightarrow bW$ jets

Resonant jets close to the topologies initiated by a top quark are generated. There are practical considerations because one can use the corresponding probability scores to construct a discriminant that aims to veto the top-initiated backgrounds (e.g. the $t\bar{t}$ or single top quark processes) in many analyses. However, similar to the general design spirit, the top quark and W boson masses are specified and are all configured as variables.

The events that produce $t \rightarrow bW$ jets is a spin-0 resonance Z' decaying to $t\bar{t}$, generated by MG. In the same configuration of the di-resonance sample production as above, the resonance mass varies between 60 and 6000 GeV; the mass of the top quark varies between 15 to 250 GeV in steps of 10 GeV, with 15–30 GeV in steps of 5 GeV. Also, a crucial aspect is to configure m_W according to m_t . Considering that our motivation for including top-initiated jet topologies is for the purpose of vetoing the real SM top jets, the m_W/m_t is close to the SM configuration but is allowed to be smeared around the centre by specifying the W boson width equal to m_W . The hadronic or leptonic decay of the W boson is specified in the ME level. The rate for producing hadronic over leptonic $t \rightarrow bW$ jet is also configured 1 : 2, similar to the fully hadronic over semileptonic rates when producing $H \rightarrow W^{(*)}W^{(*)}$ and $H \rightarrow Z^{(*)}Z^{(*)}$ jets.

11.1.3 Samples for test

To test the performance of the model, dedicated samples are produced that produce jets closer to the SM property.

For $\tilde{G} \rightarrow HH$ events that generate two-prong resonant jets, the $\tilde{G}$ mass is fixed at 1000 GeV, and the scalar H mass is fixed at 125 GeV such that the resonant jet resembles an SM Higgs jet. Only the $b\bar{b}$, $c\bar{c}$, $s\bar{s}$, and $q\bar{q}$ ($q = u, d$) final states are generated. For $\tilde{G} \rightarrow HH$ events that generate $H \rightarrow W^*W^{(*)}$ jets, two sets of samples are produced, where one specifying both m_H and m_W as the SM value, and then another varying m_H within the range of (15, 250) GeV and keeping m_W as the SM value. These two schemes are illustrated in Fig. 11.2 in the red dot and red line. Lastly, for the $Z' \rightarrow t\bar{t}$ process that produces $t \rightarrow bW$ jets, the initial resonance Z' has its mass sampled from 1200–4500 GeV while both m_t and m_b are fixed at the SM value.

11.1.4 Truth labels of jets

Category labels Resonant jets initiated from the training samples are labelled to one of the classes in Fig. 11.1 based on the matching criteria or discarded. The matching is based on the ΔR distance between the GEN-level quarks or leptons to the jet axis. For large- R jets clustered with the radius $R = 0.8$, a matched GEN-level particle requires $\Delta R < 0.8$. The quarks and leptons are the direct particles produced by the resonance, and they are read from the PYTHIA list. The situation is different from the tau leptons. If it decays hadronically, then one matches the original tau leptons to the jet just like other particles; if it decays leptonically, then one instead uses its decayed e or μ daughter to deliver matching. The neutrinos are excluded from the GEN-level particles because they are invisible to the detector. For a better illustration, Fig. 11.4 shows four examples from the $H \rightarrow W^{(*)}W^{(*)}$ jets that are matched to one category.

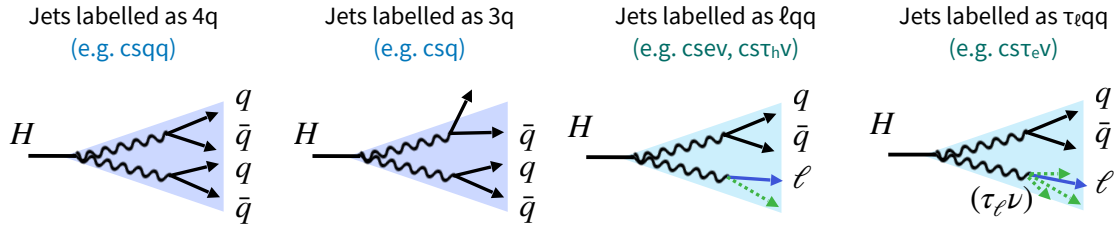


Figure 11.4 A illustration of the matching criteria from the GEN-level quarks or leptons to the jet axis to define the category. The subplots show four possible final states resulting from $H \rightarrow W^{(*)}W^{(*)}$ jets.

For the QCD jets, it is exclusively assigned to the five categories by performing a ghost

matching from the heavy-flavour B or D hadrons to the jet and counting the numbers of the matched hadrons. This follows the same definition as in Table 6.6.

GEN-level masses In addition to the category labels assigned to the jet, two GEN-level masses of a jet are defined, used as the regression target. For resonant jets, these are the “GEN-level resonance mass” and the “GEN-level visible-particle mass”.

The first one is defined as the mass of the resonance (H or t) that initiates the jet; in contrast, the second one is defined as the invariant mass of its composed GEN-level quarks or leptons (excluding neutrinos) that is ΔR -matched to the jet. Again, for the special tau lepton case, we consider its subsequent visible decay products to exclude all neutrinos. Here are a few examples to illustrate the rules.

- for the $H \rightarrow W^{(*)}W^{(*)} \rightarrow e\nu_e cs$ class, the selected visible particles are: e, c, s ;
- for the $H \rightarrow W^{(*)}W^{(*)} \rightarrow \tau_e \nu_\tau cs$ class, where τ_e decays to $e + 2\nu$, the selected visible particles are: e, c, s ;
- for the $H \rightarrow W^{(*)}W^{(*)} \rightarrow \tau_h \nu_\tau cs$ class, if τ_h decays to $\pi^\pm + 2\pi^0 + \nu$, then the selected visible particles are: $\pi^\pm, \pi^0, \pi^0, c, s$.

For QCD jets, the mass labels collected from the matched GEN-jet are assigned to ensure they have roughly consistent regression target definitions. As a correspondence to GEN-level resonance mass, the soft-drop mass of the matched GEN-jet with neutrinos included in the GEN-jet clustering is adopted; as for GEN-level visible mass, we use the soft-drop mass of the matched GEN-jet clustered without including the neutrinos. To sum up, the mass labels assigned to jets are

$$\text{Mass label \#1} = \begin{cases} \text{GEN-level resonance mass} & \text{for resonant jets,} \\ \text{GEN-jet (clustered including } \nu) m_{\text{SD}} & \text{for QCD jets.} \end{cases} \quad (11.2)$$

$$\text{Mass label \#2} = \begin{cases} \text{GEN-level visible particle mass} & \text{for resonant jets,} \\ \text{GEN-jet (clustered excluding } \nu) m_{\text{SD}} & \text{for QCD jets.} \end{cases} \quad (11.3)$$

11.2 Development details

This section introduces details of developing a pre-trained jet model, including the class definition, training objective, input data format, input preprocessing, and training details.

11.2.1 Training objective

The GloParT model outputs 314 classification nodes and 2 regression nodes. The training objective is to minimize a hybrid classification and regression loss function. We provide the details of the objective constructions below, and, more importantly, we will highlight an interpretation of the statistical meaning given the loss function definition.

Classification part

For the classification part, the multiclass cross-entropy loss is used by convention. To be specific, we consider the class $c \in \{1, \dots, C\}$ ($C = 314$) which has one-to-one correspondence to a model's output logit $z_c \in \mathbb{R}$. For an input jet represented as x , the label $y_c \in \{0, 1\}$ indicates whether the jet belongs to class c . The loss function is written as

$$L_{\text{cls}} = - \sum_c y_c \log \frac{\exp(z_c)}{\sum_{c'} \exp(z_{c'})} = - \sum_c y_c \log \hat{y}_c. \quad (11.4)$$

Here,

$$\hat{y}_c = \frac{\exp(z_c)}{\sum_{c'} \exp(z_{c'})} = \text{softmax}(z_c). \quad (11.5)$$

The number is considered the probability score of each jet x classified as the class c , and it satisfies $\sum_i \hat{y}_c = 1$.

From the statistical point of view, jets from each class c are sampled from a PDF denoted by $\rho_c(x)$. It can be proven that (see details in Appendix B), in the limit of infinite data point, the solution that minimizes L_{cls} yield

$$\hat{y}_c(x) = \frac{\rho_c(x)}{\sum_{c'} \rho_{c'}(x)}. \quad (11.6)$$

Therefore, the target of the multiclass neural network is to estimate the ratio of all PDFs. An illustrative explanation is provided in Fig. 11.5, which takes a binary classifier as an example. Note that although training multiclass classifiers has become conventional in particle experiments, the statistical interpretation has often been overlooked by the community in recent years. It is essential to recall the mathematical foundation here because it is the key to the physics analyses in defining the optimal discriminant when the multiclass classifier is used. Specifically, this work introduces a large-scale classifier model and guides how to use the hundreds of probability scores.

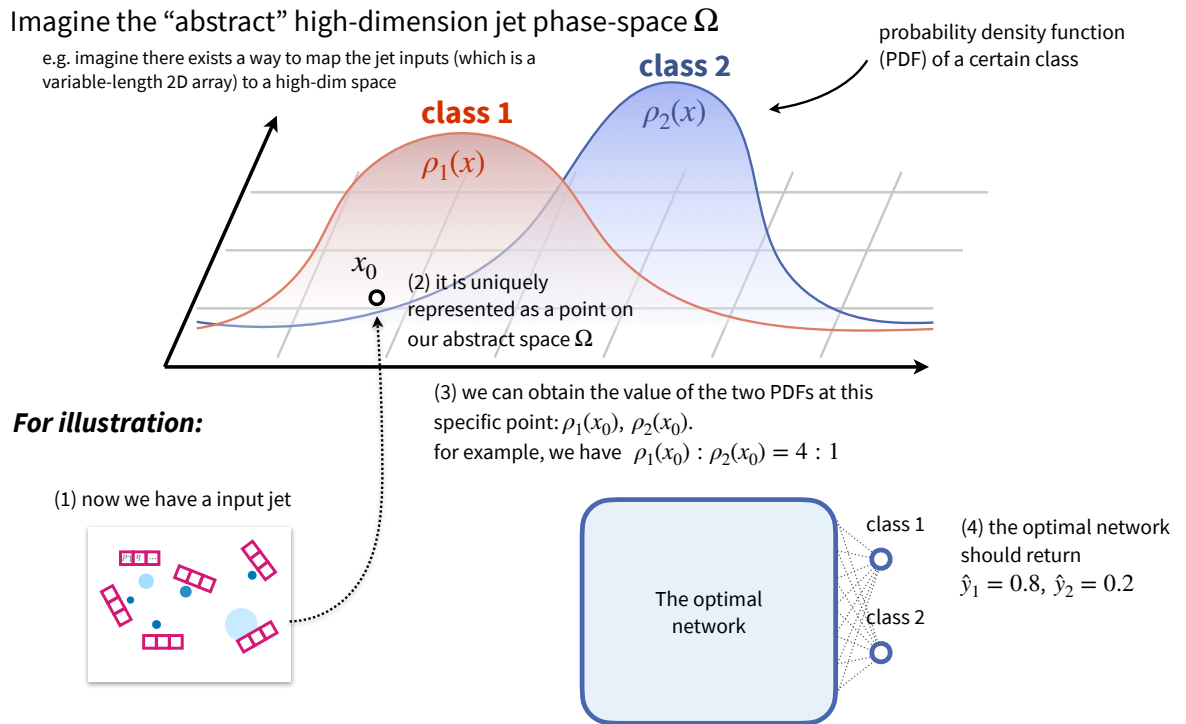


Figure 11.5 Illustration on the fact that each class of jets corresponds to a PDF in an abstract jet phase space Ω , and that a binary or multiclass classifier network trained with the cross-entropy loss is learned to estimate the a normalized PDF ratio.

Regression part

Similar to the mass regression model introduced in Section 8.2.2, the regression still adopts a log-cosh loss, which can be recognized as an L2 loss at the vicinity of the target value and an L1 loss when at a distance. The regression process can also be integrated from a statistical view which can be found in Appendix B.

The final objective is to minimize the loss function L_{tot} which can be written as

$$L_{\text{tot}} = L_{\text{cls}} + \lambda(L_{\text{reg1}} + L_{\text{reg2}}). \quad (11.7)$$

The hyperparameter λ is tuned and adjusted to 5. As there is no entanglement between the terms, the optimal case when the hybrid loss reaches the minimum should result in both optimal performances in classification and regression.

11.2.2 Model architecture

The GloParT model adopts the ParT architecture, introduced in Section 4.2, as the backbone. ParT is a Transformer model that currently achieves state-of-the-art performance on the 100 M JETCLASS dataset. The network proceeds input particles as a permutational-invariant set and updates its features by a stack of multihead self-attention blocks. The pairwise attention in each block is calculated from the particle-level features and the pairwise features between the particles. Two adjustments are made compared to the original architecture, including handling multiple data groups and scaling the model size.

First, the CMS GloParT training involves three types of data, the charged PF candidates, the neutral PF candidates, and the SV. The split of PF candidates is due to the significant difference in their feature space, e.g. the neutral candidates lack features related to the track impact parameters or the pixel hits. Therefore, separate embedding of their feature space is a viable solution. The split has been adopted by several small- R and large- R jet tagging algorithms in the CMS experiments, such as DeepJet^[193], DeepDoubleX^[85, 86], and Robust ParT AK4^[194].

Second, three Transformer sizes are studied. We modified the primary decoder dimension of the Transformer, setting it to either double or half, which roughly changes the number of parameters in the model by a factor of 4 or 1/4, which we refer to as ParT-L (large) and ParT-lite. The original size is referred to as ParT-B (base). The parameter settings for the three sizes of ParT models are listed below. The ParT-L is the final choice for the GloParT model, and a summary of its parameters is presented in Table 11.1. For a clear illustration, the ParT model architecture in the ParT-L size annotated with its main parameters is depicted in Fig. 11.6.

Table 11.1 The model parameters and training configurations for the ParT model in three sizes: ParT-lite, ParT-B and ParT-L (GloParT).

	ParT-lite	ParT-B	ParT-L (GloParT)
Input embed. dim.	(64, 256, 64)	(128, 512, 128)	(256, 1024, 256)
Pairwise feat. embed. dim.	(32, 32, 32, 8)	(64, 64, 64, 8)	(128, 128, 128, 16)
Transformer dim.	64	128	256
Number of heads	8	8	16
Fully-connected layer dim.	(512, 316)	(1024, 316)	(1024, 316)
Initial LR	6.75×10^{-3}	4×10^{-3}	2×10^{-3}
Batch size	768	512	256
Epochs	30	50	50

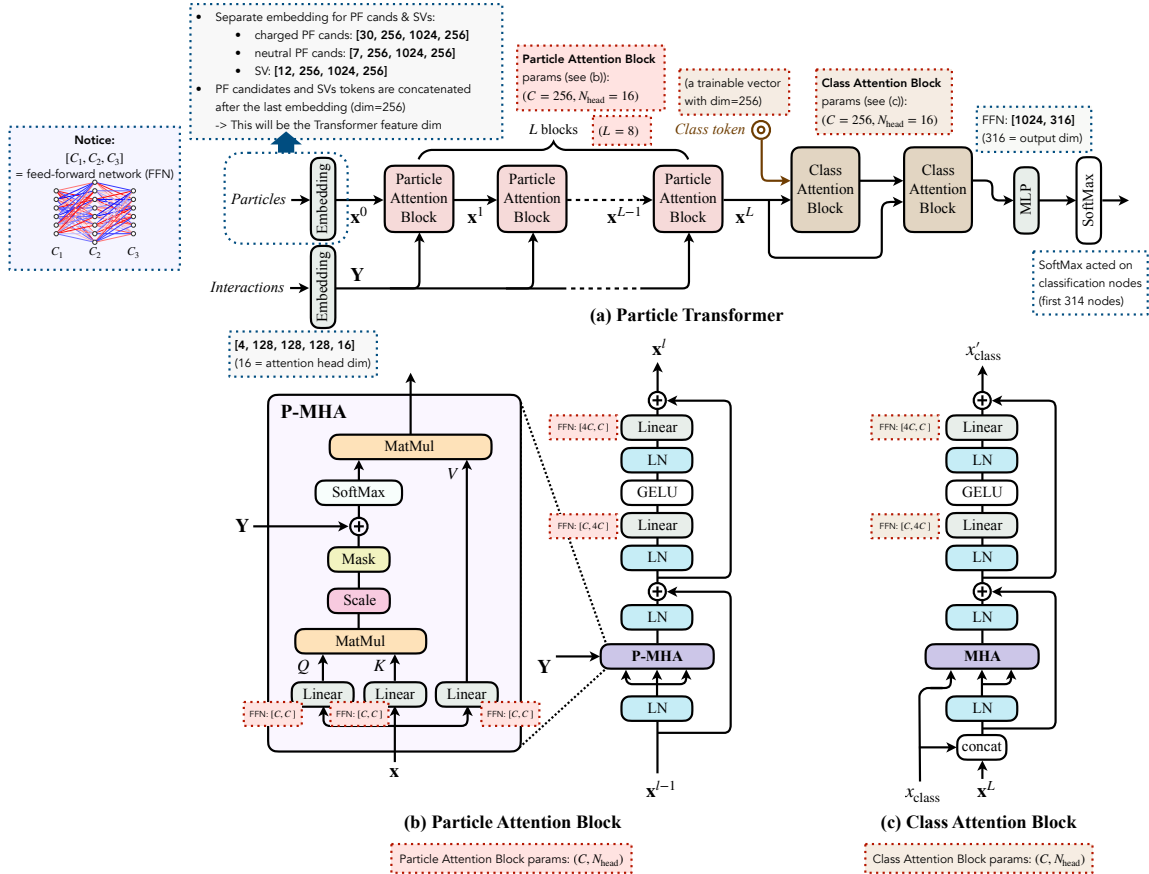


Figure 11.6 Illustration of the parameter details of the ParT-L (GloParT) model. The outline of the ParT architecture is adapted from Ref. [65].

11.2.3 Input

The input features of the GloParT are summarized in Table 11.2. The features include those from the constituent PF candidates and the SVs that are ΔR -matched with the jet, similar to the ParticleNet training in CMS. The difference is that PF candidates are split into charged and neutral ones, and for charged candidates, additional features, including pixel hit patterns in multiple CMS tracker layers, are added. Also, the lost tracks that do not pass the PF algorithm are added back to the charged candidates. A boolean variable (`isLostTrack`) is used to distinguish if the object is a lost track.

To illustrate the complexity difference for PF candidates and SVs across the categories, Fig. 11.7 shows the distributions of the length of charged, neutral PF candidates and SVs. Typical jet categories are shown, including the merged QCD classes, the merged hadronic $H \rightarrow WW^*$ classes, the $H \rightarrow b\bar{b}$ class, the merged hadronic $t \rightarrow bW$ classes, and the merged hadronic $H \rightarrow ZZ^*$ classes, are shown. It can be concluded that multipronged jets generally have larger PF candidates, and jets originating in b -enriched final states have larger SVs.

To provide a full visualization of each input jet, Fig. 11.8 shows the point cloud visualization that plots all constituent PF candidates and matched SVs on the η - ϕ plane. In the plottings, each PF candidate or SV is displayed as a marker with its area proportional to the energy. The circles, triangles (upward- or downward-directed), and pentagons are in the green/blue colour, representing the hadrons, leptons (electrons or muons), and photons in the merged list of PF candidates and lost tracks. The solid (hollow) markers stand for electrically charged (neutral) particles. A marker with a brown edge indicates it is a lost track. The orange/red squares correspond to the SVs. The colour reflects the impact parameter, where a larger displacement with the primary vertex results in more blue (for PF candidates) or more red (for SVs). The category of the plotted jet, its p_T , its GEN-level resonance and visible-particle mass for a resonant jet are displayed in the figure. The visualization reveals the prongness, track displacements, and appearance patterns of various types of PF candidates, and SVs.

11.2.4 Preprocessing

Jets are required to pass preselections of $200 < p_T < 2500$ GeV, $|\eta| < 2.4$, and $20 < m_{SD} < 260$ GeV. Similar to ParticleNet-MD development in CMS discussed in Section 8.2.1, the mass decorrelation is achieved by sampling-based reweighting, where “reweighting” stands for sampling jets with a predefined probability to form the training batch on the fly. The sampling probability is binned on jet p_T and m_{SD} . The fine classes are merged to a level to form

Table 11.2 The list of input variables used in the GloParT model development, grouped into three collections: the charged PF candidates + lost tracks, the neutral PF candidates, and the SVs. The input collections or features are marked as (★) to highlight the difference with the ParticleNet training inputs listed in Table 8.1.

Variable	Definition
For charged PF candidates + lost tracks (★)	
$\log p_T$	logarithm of the particle's p_T
$\log E$	logarithm of the particle's energy
$\Delta\eta(\text{jet})$	difference in pseudorapidity between the particle and the jet axis
$\Delta\phi(\text{jet})$	difference in azimuthal angle between the particle and the jet axis
$ \eta $	absolute value of the particle's pseudorapidity
q	electric charge of the particle
isMuon (★)	if the particle is identified as a muon
isElectron (★)	if the particle is identified as an electron
isChargedHadron (★)	if the particle is identified as a charged hadron
isLostTrack (★)	if the particle a charged PF candidate or a lost track
pvAssociationQuality	flag related to the association of the track to the primary vertices
lostInnerHits	quality flag of the track related to missing hits on the pixel layers
χ^2/dof	χ^2 value of the trajectory fit normalized to the number of degrees of freedom
qualityMask	quality flag of the track
d_z	longitudinal impact parameter of the track
d_z/σ_{d_z}	significance of the longitudinal impact parameter
d_{xy}	transverse impact parameter of the track
$d_{xy}/\sigma_{d_{xy}}$	significance of the transverse impact parameter
η_{rel}	pseudorapidity of the track relative to the jet axis
$p_{T,\text{rel}}$ ratio	track momentum perpendicular to the jet axis, divided by the magnitude of the track momentum
$p_{\text{par,rel}}$ ratio	track momentum parallel to the jet axis divided by the magnitude of the track momentum
d_{3D}	signed 3D impact parameter of the track
d_{3D}/σ_{3D}	signed 3D impact parameter significance of the track
trackDistance	distance between the track and the jet axis at their point of closest approach
pixelBarrelLayersWithMeasurement (★)	number of valid hits in the pixel barrel layers
pixelEndcapLayersWithMeasurement (★)	number of valid hits in the pixel endcap layers
stripTIBLayersWithMeasurement (★)	number of valid hits in the strip TIB layers
stripTIDLayersWithMeasurement (★)	number of valid hits in the strip TID layers
stripTOBLayersWithMeasurement (★)	number of valid hits in the strip TOB layers
stripTECLayersWithMeasurement (★)	number of valid hits in the strip TEC layers
For neutral PF candidates	
$\log p_T$	logarithm of the particle's p_T
$\log E$	logarithm of the particle's energy
$\Delta\eta(\text{jet})$	difference in pseudorapidity between the particle and the jet axis
$\Delta\phi(\text{jet})$	difference in azimuthal angle between the particle and the jet axis
$ \eta $	absolute value of the particle's pseudorapidity
isPhoton (★)	if the particle is identified as a photon
isNeutralHadron (★)	if the particle is identified as a neutral hadron
For SVs within the jet cone	
$\log p_T$	logarithm of the SV's p_T
m_{SV}	mass of the SV
$\Delta\eta(\text{jet})$	difference in pseudorapidity between the SV and the jet axis
$\Delta\phi(\text{jet})$	difference in azimuthal angle between the SV and the jet axis
$ \eta $	absolute value of the SV's pseudorapidity
N_{tracks}	number of tracks associated with the SV
χ^2/dof	χ^2 value of the SV fit normalized to the number of degrees of freedom
d_{2D}	signed 2D impact parameter (i.e., in the transverse plane) of the SV
d_{2D}/σ_{2D}	signed 2D impact parameter significance of the SV
d_{3D}	signed 3D impact parameter of the SV
d_{3D}/σ_{3D}	signed 3D impact parameter significance of the SV

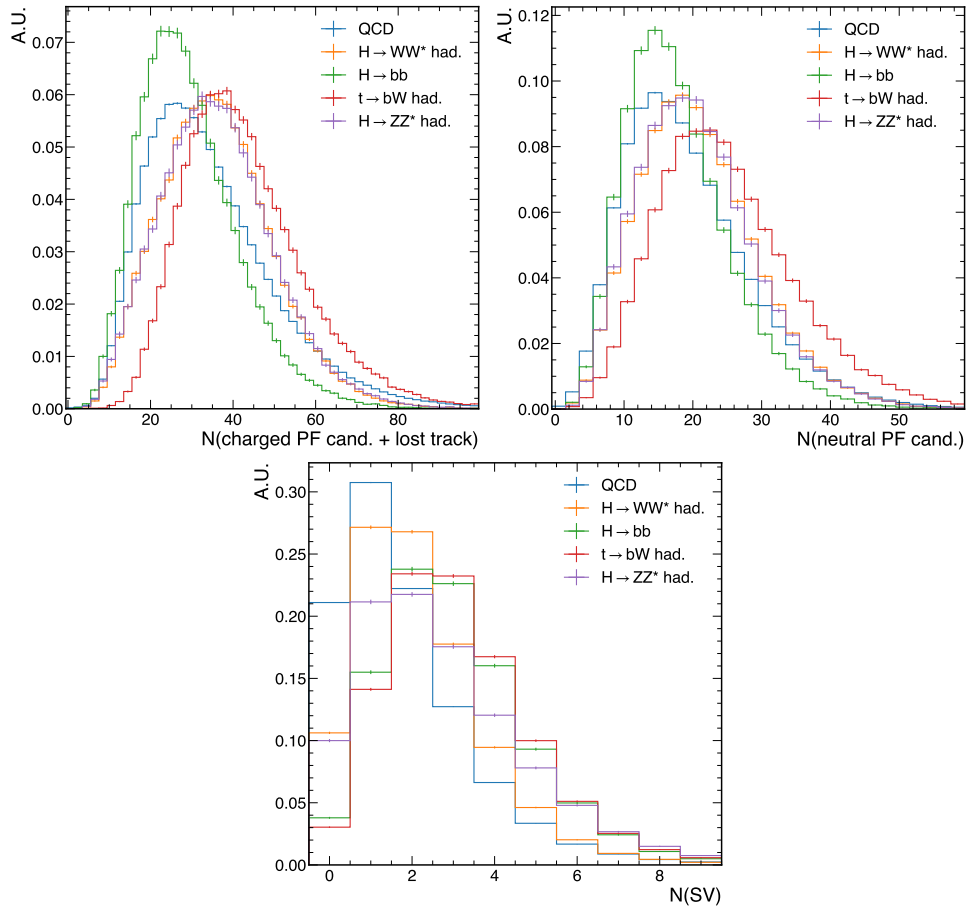


Figure 11.7 The normalized distributions of the number of charged PF candidates + lost tracks, the neutral PF candidates, and the SVs for several typical jet categories.

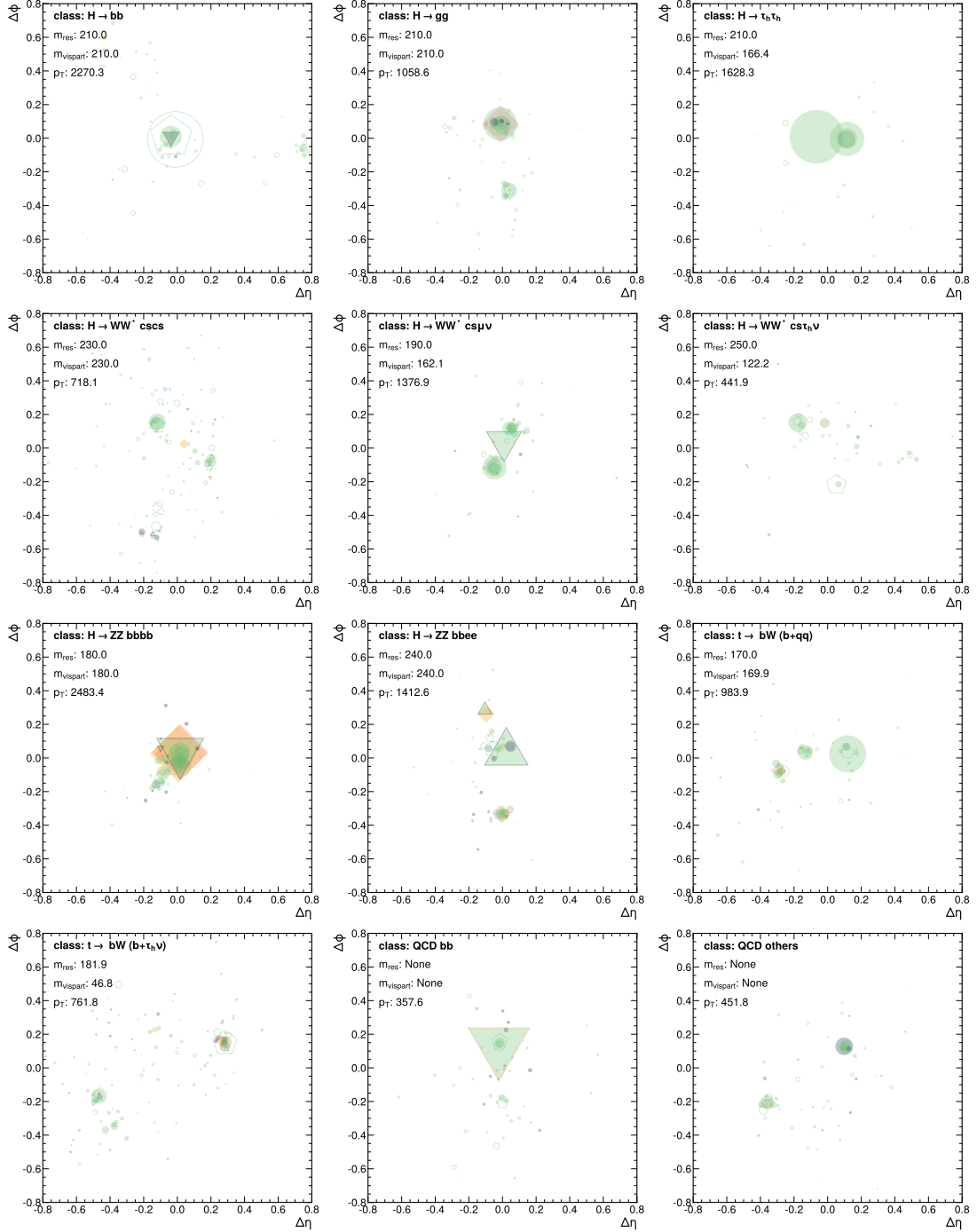


Figure 11.8 The point-cloud visualization of selected jets in different categories. In each visualization, the position of all PF candidates and SVs is marked on the η - ϕ plane, displayed with their energies or other properties (see text for details). The GEN-level properties of the jet are also annotated in the figure.

the basis of the reweighting class. This is shown in Fig. 11.9, where “Q” in the name of the reweighting class means by merging corresponding classes a quark of any flavour. The class weights of each reweighting class are also indicated in the figure, which is tuned based on the importance one hopes to assign in each class and the relative proportions of data size between classes.

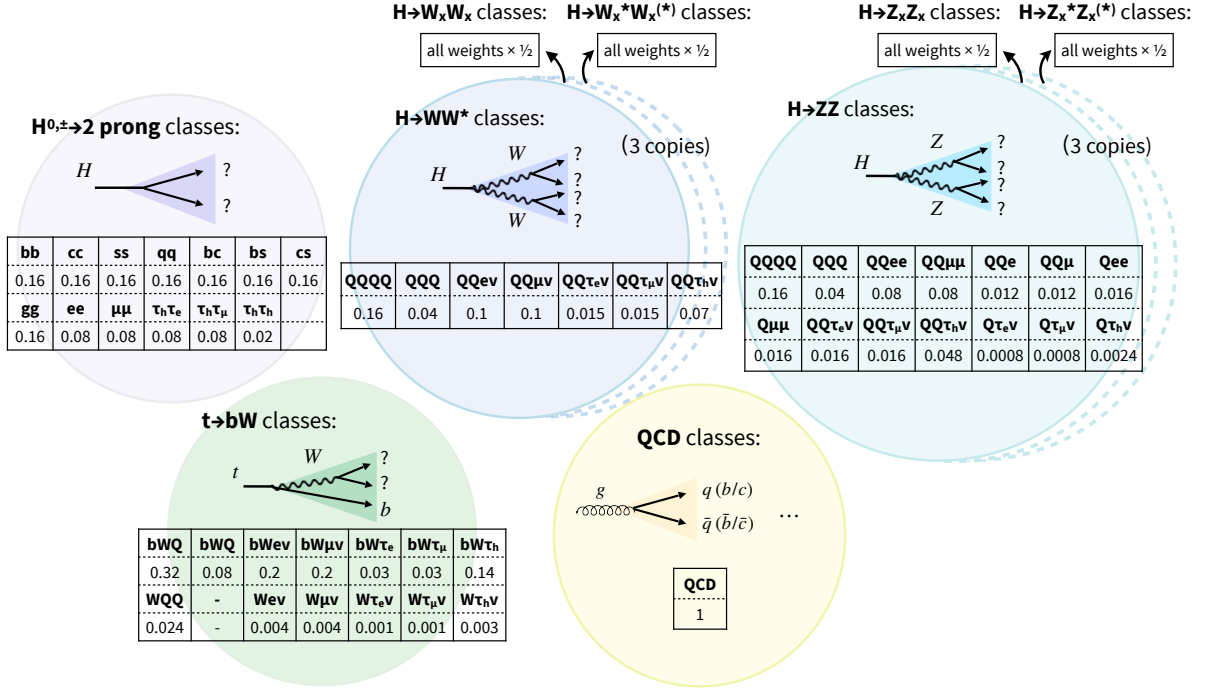


Figure 11.9 A summary of the reweighting classes and class weights used in the GloParT model development.

11.2.5 Training setup

The training is performed using the WEAVER framework, a research and development platform tailored for high-energy physics applications, powered by the PYTORCH backend. A key advantage is its ability to dynamically sample inputs according to predetermined probabilities and its data loader’s capability to handle large-scale data in ROOT format.

The training is performed on 3 Nvidia GTX 3090 graphics cards, using the data distributed parallel (DDP) mechanism to distribute the workload across multiple GPUs. The training batch size, LR and number of epochs are indicated in Table 11.1. Each epoch is defined by iterating over a fixed number of 7.68 M jets for training, using the first 80% of the dataset, then iterating over 0.512 M jets for validation, using the last 20% of the dataset. The iterator keeps looping over the entire dataset used for training and validation. A snapshot of the model is saved at the

end of each epoch. The model snapshot showing the lowest loss on the validation dataset is selected for the final evaluation. The training adopts the LOOKAHEAD optimizer^[101] with $k = 6$ and $\alpha = 0.5$ and the inner optimizer is RADAM^[102] with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-5}$. The LR, set according to Table 11.1, remains constant for the first 70% of the iterations, and then decays exponentially, changing at the beginning of every following epoch, down to 1% of the initial value at the end of the training.

Figure 11.10 shows the training and validation metrics for the three sizes of the ParT model. Performance leap from ParT-lite (orange) to ParT-B (blue), and then to ParT-L (red) can be seen. By comparing training and validation curves, both ParT-lite and ParT-B training maintain similar trends, but for ParT-L one can see that the performance on the validation dataset noticeably lags behind at the later stages of training. This indicates a potential upper limit of the model size that the current data size can accommodate.

11.3 Overall performance

Although the physics goal for GloParT development is to pre-train the model on comprehensive data such that it learns a latent jet representation with excellent generalization ability, The nature of the training task ensures it is suitable for direct classification and regression purposes, simply using its output nodes.

11.3.1 Classification performance

For classification, it has been shown in Eq. (11.6) that the optimal multiclass classifier learns a probability score for each class directly corresponding to the PDF ratio. Thus, taking the scores for class A and B , namely p_A and p_B , and using the notion in Eq. (11.6) for PDFs, one has

$$A \text{ vs. } B \text{ disc.} := \frac{p_A(x)}{p_A(x) + p_B(x)} = \frac{\rho_A(x)}{\rho_A(x) + \rho_B(x)} \quad (11.8)$$

for every input x , which is the optimal discriminant defined to distinguish class A from class B . This implies that no matter how many classes one adds to train the classifier, the optimal discriminant for class A and B can be constructed the same way, and it is irrelevant to all other classes. This is a good property of a fully optimized neural network classifier trained with the multiclass cross-entropy loss we can make use of. Figure 11.11 gives a visual illustration of this aspect. Practically, this has been successfully validated in the jet tagging experiment—with adequate input data size and large enough capacity of the model, such as the GloParT model,

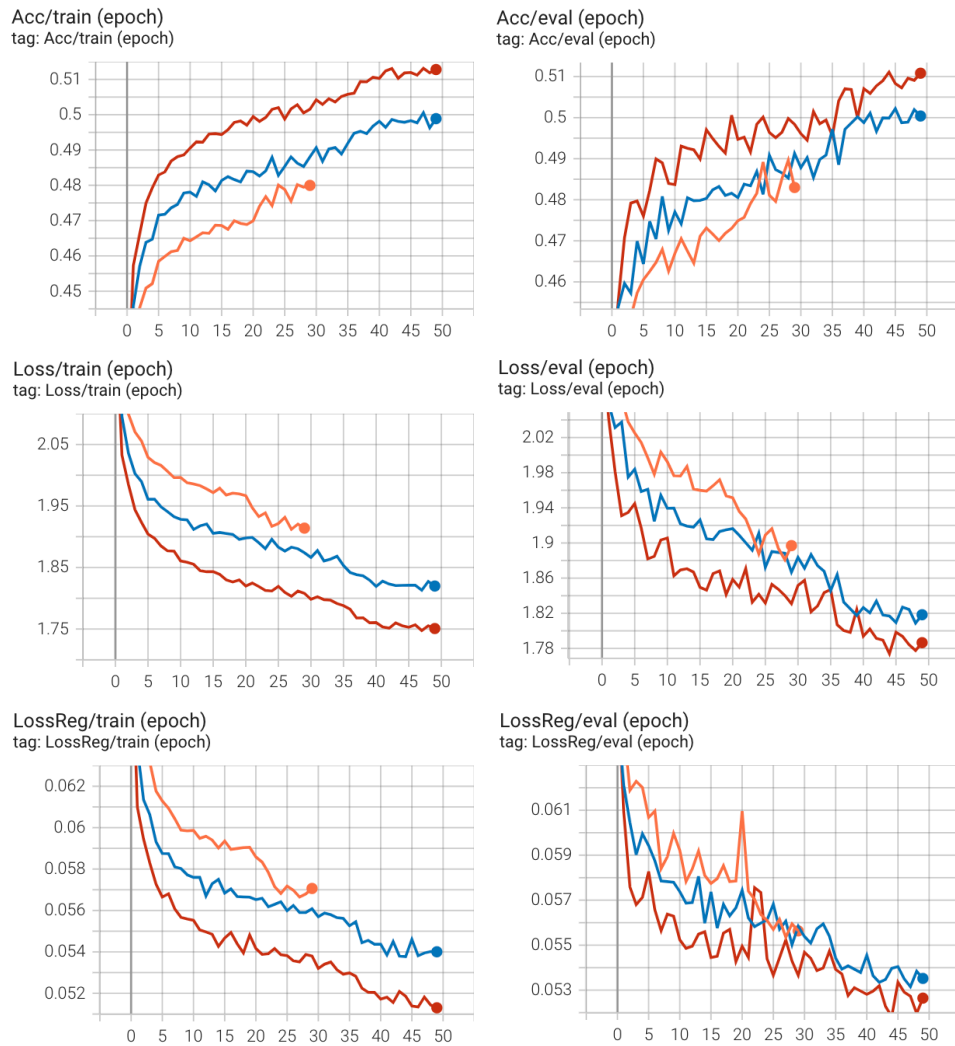


Figure 11.10 The metrics monitored during the training of the ParT-lite (orange), ParT-B (blue), and ParT-L (red) model, including the classification accuracy (top), the classification loss (middle), and the regression loss (bottom), for each epoch of the training (left) and evaluation (right) procedure.

the training can result in nearly identical separating power for given two classes regardless of additional classes added in training.

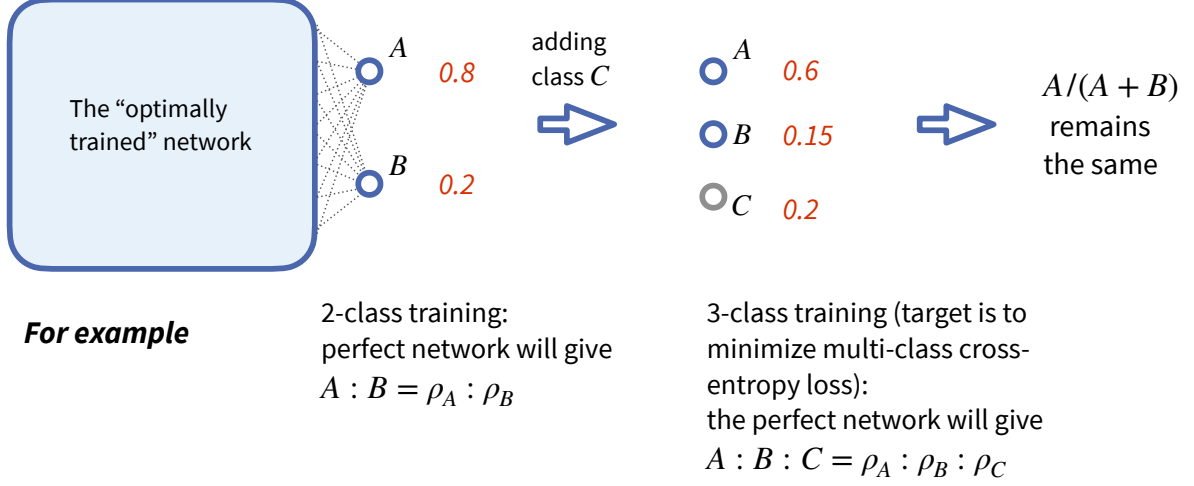


Figure 11.11 Illustration on the fact that the binary score $A/(A+B)$ constructed from the model should represent an estimate of $\rho_A(x)/(\rho_A(x)+\rho_B(x))$ regardless of additional categories included in the training.

Therefore, despite that there are 314 classes in GloParT, one constructs $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants by

$$\begin{aligned} \text{GloParT bbvsQCD disc.} &= \frac{p(H \rightarrow b\bar{b})}{p(H \rightarrow b\bar{b}) + p(\text{QCD})}, \\ \text{GloParT ccvsQCD disc.} &= \frac{p(H \rightarrow c\bar{c})}{p(H \rightarrow c\bar{c}) + p(\text{QCD})}, \end{aligned} \quad (11.9)$$

which only exploits the output nodes corresponding to the $H \rightarrow b\bar{b}$, $H \rightarrow c\bar{c}$, and QCD classes^①.

Figures 11.12–11.14 compare the performance of GloParT with ParticleNet-MD regarding the discrimination ability among $H \rightarrow b\bar{b}$, $H \rightarrow c\bar{c}$ and QCD jets, in terms of ROC curves. The horizontal line indicates the signal efficiency ϵ_S when background efficiency $\epsilon_B = 0.01$. The $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets are originated from the SM Higgs boson and collected from the test data. For both GloParT and ParticleNet-MD, the discriminant used to distinguish $H \rightarrow c\bar{c}$ from $H \rightarrow b\bar{b}$ jet is defined as $p(H \rightarrow c\bar{c})/(p(H \rightarrow c\bar{c}) + p(H \rightarrow b\bar{b}))$. It is clear that GloParT reaches superior results, with QCD background rejection improved by a factor of approximately 1.5 in terms of $H \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging, and around 2 times $H \rightarrow b\bar{b}$ background

① Note that, compared with Eq. (5.2), the notion is changed from $X \rightarrow b\bar{b}$ ($c\bar{c}$) to $H \rightarrow b\bar{b}$ ($c\bar{c}$) but they are in the same meaning.

rejection against $H \rightarrow c\bar{c}$ jets. This demonstrates that expanding the training categories on a large scale is viable. At the same time, the discrimination ability for certain classes is not diminished, making it a very promising option for the future of the tagging model.

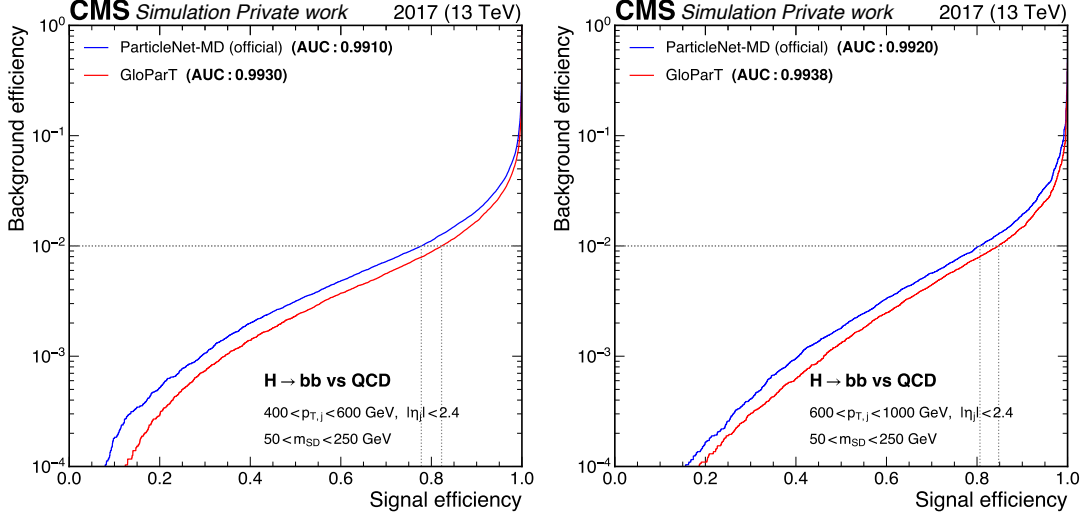


Figure 11.12 The GloParT's tagging performance for distinguishing $H \rightarrow b\bar{b}$ jets from QCD jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.

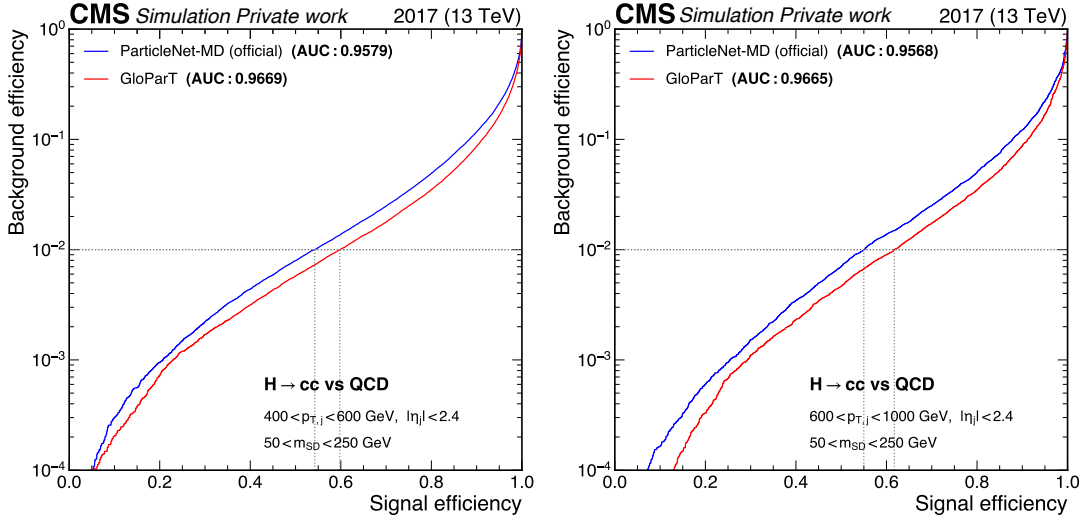


Figure 11.13 The GloParT's tagging performance for distinguishing $H \rightarrow c\bar{c}$ jets from QCD jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.

The mass decorrelation performance of the GloParT and ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ discriminants is shown in Figs. 11.15–11.16. By comparison, the GloParT discriminants still maintain a good decorrelation with m_{SD} , making it suitable for applications to future analyses relevant to Higgs and BSM physics. The calibration of $H \rightarrow b\bar{b}$ ($c\bar{c}$) signal selection efficiencies for GloParT $X \rightarrow b\bar{b}$ ($c\bar{c}$) discriminants results is also preliminarily checked using

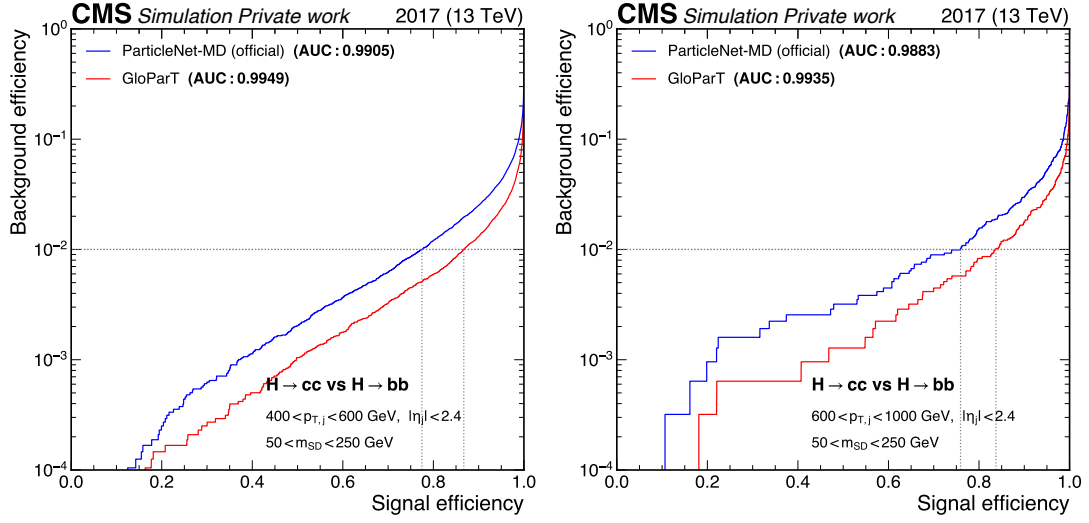


Figure 11.14 The GloParT’s tagging performance for distinguishing $H \rightarrow c\bar{c}$ jets from $H \rightarrow b\bar{b}$ jets in terms of the ROC curve, compared with the ParticleNet-MD tagger.

the sfBDT method, introduced in Chapter 6. It is reassuring to see a reasonable agreement between the performance in data and simulation. The calibrated SFs are around 1 with an uncertainty of 4–8% (15–30%) for $H \rightarrow b\bar{b}$ ($c\bar{c}$) discriminants, similar to the calibration results of ParticleNet-MD.

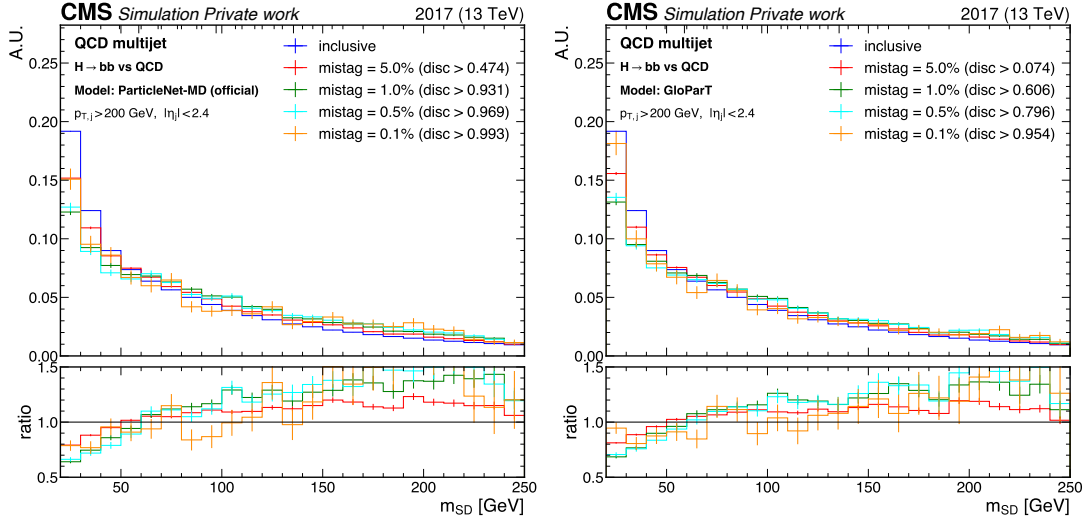


Figure 11.15 The mass decorrelation performance of the $X \rightarrow b\bar{b}$ tagging discriminant for the CMS ParticleNet-MD tagger (left) and the GloParT tagger (right). normalized distributions of the QCD jet m_{SD} on the test dataset, passing multiple thresholds of the tagging discriminant, are shown.

A final remark on the direct classification application of GloParT is that it brings many more opportunities to tag different types of final states, extending the potential for future anal-

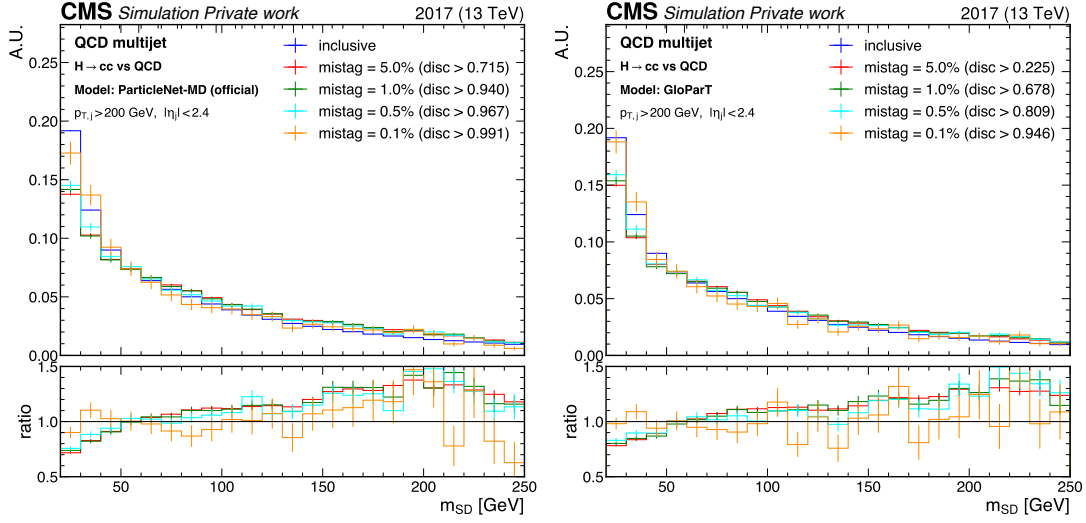


Figure 11.16 The mass decorrelation performance of the $X \rightarrow c\bar{c}$ tagging discriminant for the CMS ParticleNet-MD tagger (left) and the GloParT tagger (right). normalized distributions of the QCD jet m_{SD} on the test dataset, passing multiple thresholds of the tagging discriminant, are shown.

yses in the boosted topology. For two-prong final states, it covers novel flavour contents such as bc , bs . These signatures can be induced from a resonance predicted by many BSM models, such as 2HDM. The inclusion of cs final state also provides a unique handle to tag SM W jets in a mass-decorrelated manner, which is not covered in existing mass-decorrelated taggers. For multipronged final states, a straightforward usage is to exploit scores corresponding to $H \rightarrow WW^* \rightarrow 4q/\ell\nu qq$ final states that can be used to identify the boosted Higgs boson in the $H \rightarrow WW^*$ decay mode. The score corresponding to the top-like categories can be put to the denominator when constructing the discriminants, providing a solution to the veto $t\bar{t}$ background. Appropriate calibration of these discriminants will be an important and open project for future study.

11.3.2 Regression performance

The performance of large- R jet mass reconstruction can be studied utilizing the regression node of GloParT. Figures 11.17–11.18 show the mass regression performance for reconstructing the Higgs boson mass in the SM $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ jets. Both the regression nodes with GEN-level resonance mass and visible-particle mass as targets are shown and compared. Also compared is the ParticleNet regressed mass, a standard tool for CMS analyses, developed with the same strategy as explained in Section 8.2.2 but for AK8 jets. Results show that GloParT regressed mass with GEN-level visible-particle mass target performs slightly better

than the ParticleNet benchmark, while the other GloParT regressed mass reveals a displaced peak structure with worsened resolution.

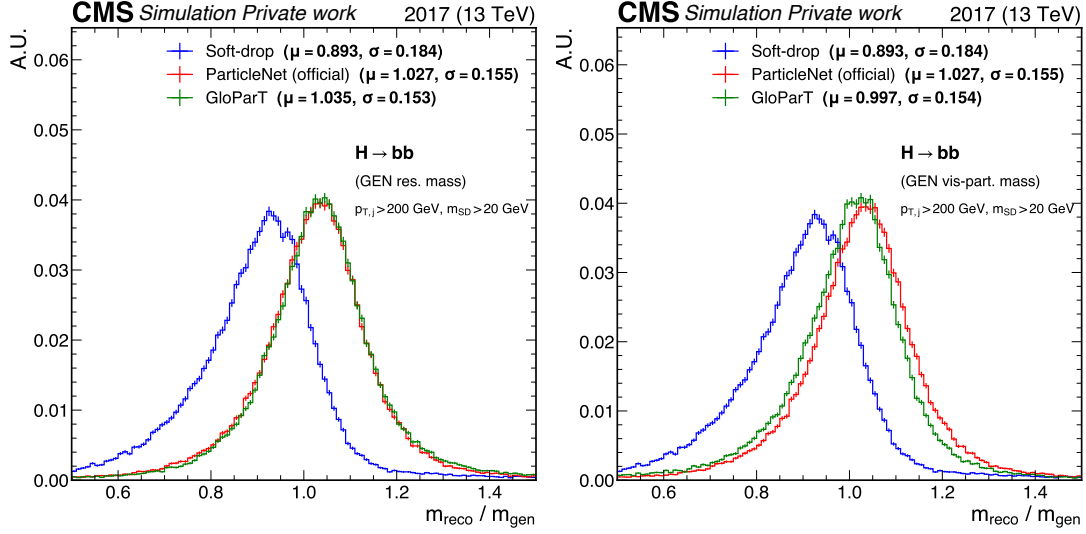


Figure 11.17 A comparison of reconstructed mass resolution in $H \rightarrow b\bar{b}$ jets among three algorithms: the soft-drop, the CMS ParticleNet regression algorithm, and the GloParT model with the regression target being the GEN-level resonance (left) and the visible-particle mass (right).

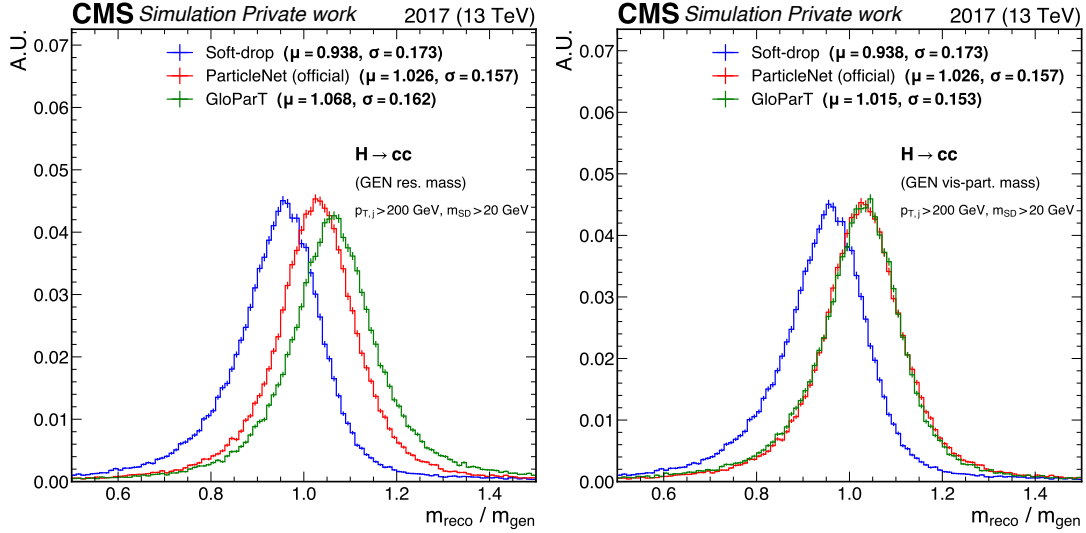


Figure 11.18 A comparison of reconstructed mass resolution in $H \rightarrow c\bar{c}$ jets among three algorithms: the soft-drop, the CMS ParticleNet regression algorithm, and the GloParT model with the regression target being the GEN-level resonance (left) and the visible-particle mass (right).

To explain the performance, one can draw from statistical theory (see Appendix B) that the regression performance cannot be maintained when additional classes of jets are introduced,

because the theoretically predicted optimal regression value is dependent on all regression targets from all classes. This explains the rightward shift in the left plots of Figs. 11.17 and 11.18 because the additional multipronged jets, with their regression target relatively larger than the real jet mass due to the missing neutrino effect, may affect the overall optimal regression result. The problem is alleviated when using the GEN-level visible-particle mass as a regression target.

11.4 Ablation experiments

There are several key features of the GloParT model that one would expect to see the exact influence on the model performance when they are brought in. This section focuses on three aspects: the impact on certain tagging performance when the number of classes increases, the effect of introducing classing splitting within the $H \rightarrow WW^*$ major category according to the mass configuration of m_H and m_W , and the improvement brought by enlarging the scale of the Transformer.

11.4.1 Increasing categories

The GloParT model covers major categories, including the $H \rightarrow WW^*$ categories, the 2-prong resonance categories, and the top-jet categories. Figure 11.19 shows the dedicated tagging performances for distinguishing three $H \rightarrow WW^*$ classes against QCD jets compared over three models, the one with only $H \rightarrow WW^*$ and QCD categories included, and the other two sequentially adding 2-prong categories and top categories. The baseline model is based on the ParticleNet architecture developed at an earlier stage and is denoted as “cat. study baseline” in the figure, and the other two models are annotated as “+2-prong” and “+2-prong +top”. For the tagging performance study for class signal A versus background B , the tagging discriminant is defined according to Eq. (11.8) such that the scores on the numerator and denominator depend on the A and B classes. Also, the signal and background jets used for making the ROC curve correspond exactly to the A and B classes. It is evident that the inclusion of additional, irrelevant classes has a minor impact on the tagging performance in our studied case. This provides a major source of confidence for upgrading the model towards a large-scale classifier.

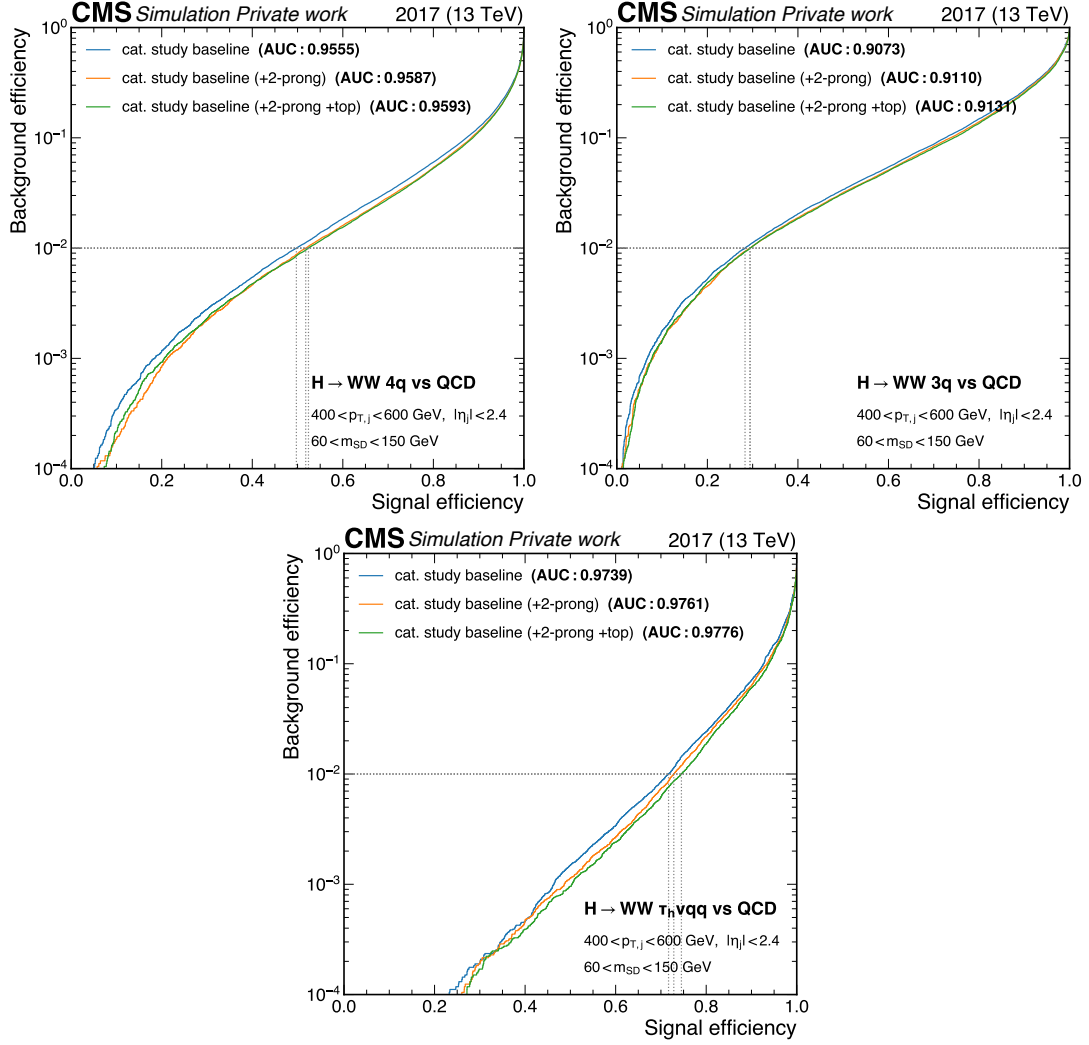


Figure 11.19 A comparison of tagging performance when additional irrelevant categories are added to the classifier model. The performance is compared for the hadronically decayed $H \rightarrow WW^* \rightarrow 4q$ (upper left), hadronically decayed $H \rightarrow WW^*$ with a quark outside of the jet cone (denoted as $3q$) (upper right), and $H \rightarrow WW^* \rightarrow \tau_h \nu qq$ against the QCD jets.

11.4.2 Class splitting via kinematics

A major feature of GloParT is its splitting of $H \rightarrow W^{(*)}W^{(*)}$ categories into three sets with a kinematics difference in the m_H and m_W configurations. In this design, the discriminants to tag $H \rightarrow W^{(*)}W^{(*)}$ can adopt three definitions: one can choose the score corresponding to the $H \rightarrow WW^*$ class, or two BSM classes $H \rightarrow WW$ and $H \rightarrow W^*W^{(*)}$ (according to Eq. (11.1)). This results in three different discriminants for tagging $H \rightarrow W^{(*)}W^{(*)}$ jets, and their performances may differ when tagging SM $H \rightarrow WW^*$ jets or BSM $H \rightarrow W^{(*)}W^{(*)}$ signatures with different m_H .

To demonstrate that this design is effective, the Higgs boson mass is configured as 50, 100, and 200 GeV such that the $H \rightarrow W^{(*)}W^{(*)}$ jets appear to be offshell–offshell, onshell–offshell, and onshell–onshell. Figure 11.20 compares the performance of three discriminant definitions in the three m_H configurations. The model used to deliver this comparison is still based on ParticleNet as of early development, denoted as “kin. study baseline” in the figure. It can be concluded that the $H \rightarrow W^*W^{(*)}$ discriminant performs better for $m_H = 50, 100$ GeV, while the $H \rightarrow WW^*$ discriminant performs better for $m_H = 200$ GeV. The conclusion is that, while it is confident that the SM $H \rightarrow WW^*$ discriminant is optimized only for SM m_H and m_W configurations and hence should perform the best for that case, it is no longer optimal for other kinematics configurations, and these dedicated BSM classes bring additional improvements.

11.4.3 Transformer scales

A comparison of the performance of models in three Transformer sizes, namely, ParT-lite, ParT-B, and ParT-L, explained in Section 11.2.5, is shown in Fig. 11.21. The comparison spans from different types of jet tagging to the mass regression performance.

The jet tagging performance indicates that overall performance improvements can be reached when the Transformer scales up. The improvement from ParT-B to ParT-L is less substantial, which is potentially due to the limited data size when training ParT-L, as can be seen from the loss curve for validation (see explanation in Section 11.2.5). Besides, by studying various tagging tasks, it is found that the progress is more significant in specific tagging modes, such as $H \rightarrow c\bar{c}$ versus $H \rightarrow b\bar{b}$, or $H \rightarrow WW^*$ versus $t \rightarrow bW$ jet tasks (lower plots in Fig. 11.21) compared with discriminating jets versus the QCD jets. The findings imply that DNN’s capability to discriminate against QCD jets has become saturated. Tasks that potentially rely more, for instance, on track displacement features, still have room for further improvements.

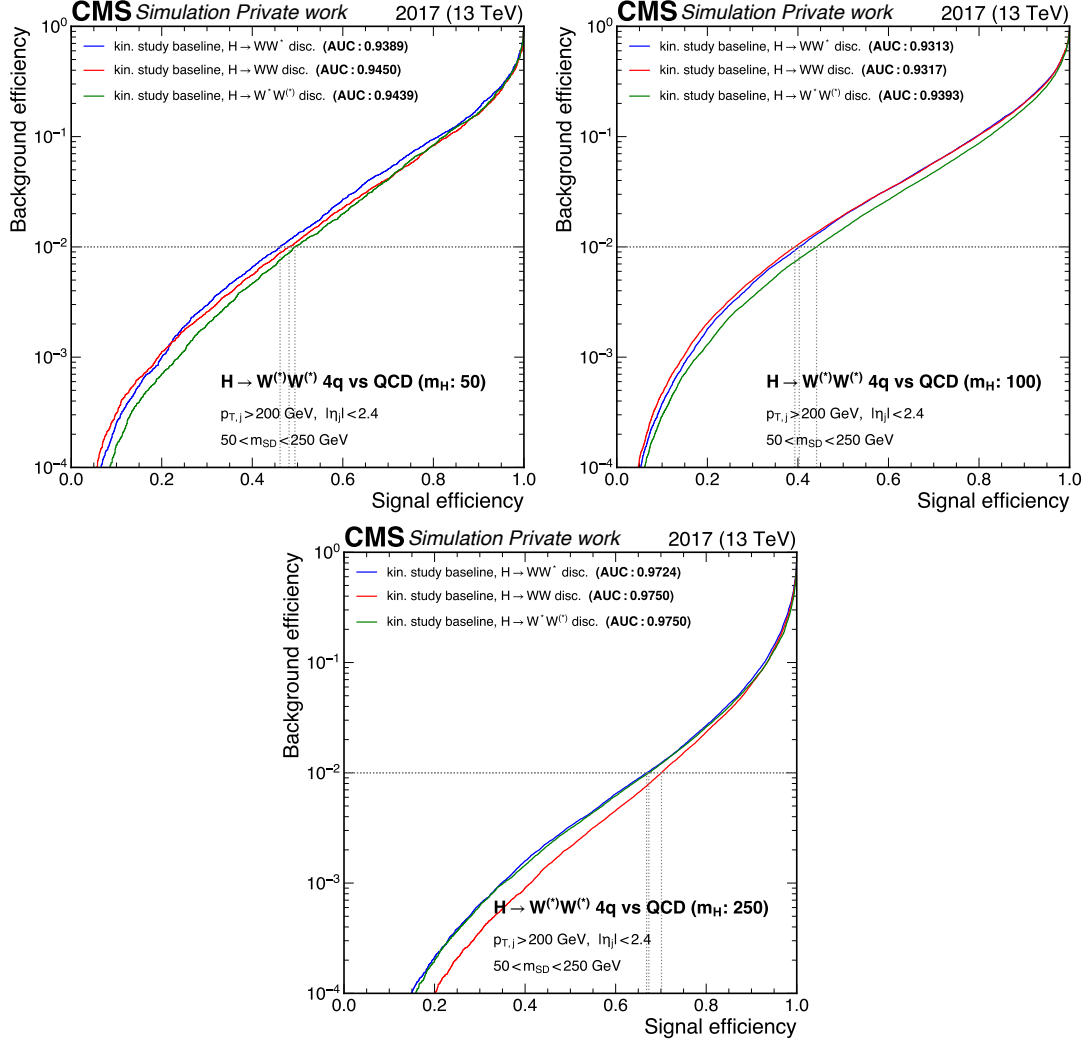


Figure 11.20 A comparison of $H \rightarrow WW^* \rightarrow 4q$ versus QCD jets tagging performance for using the three discriminants defined on different truth $H \rightarrow W^{(*)}W^{(*)}$ labels, i.e. the SM $H \rightarrow WW^*$ class and two BSM classes $H \rightarrow WW$ and $H \rightarrow WW^*$. The generator-level m_H is configured as 50 (upper left), 100 (upper right), and 200 (lower) GeV to generate different kinematics of $H \rightarrow W^{(*)}W^{(*)}$ decays.

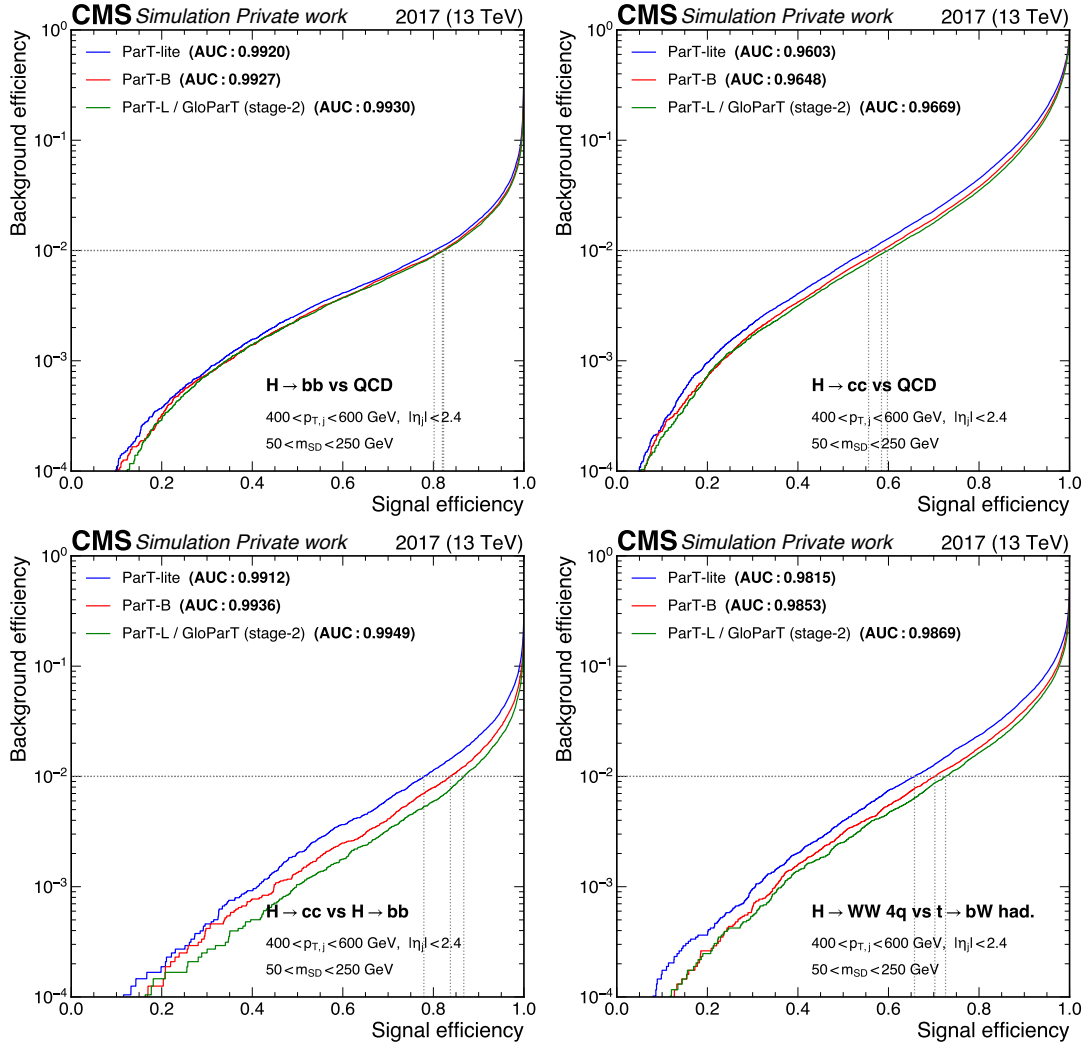


Figure 11.21 A comparison of tagging performances on multiple tasks when the model scales from ParT-lite, to ParT-B, then to ParT-L (GloParT).

In addition, Fig. 11.22 shows the regressed mass resolution for three sizes of Transformers, and it is clear that performance for regression has saturated as well.

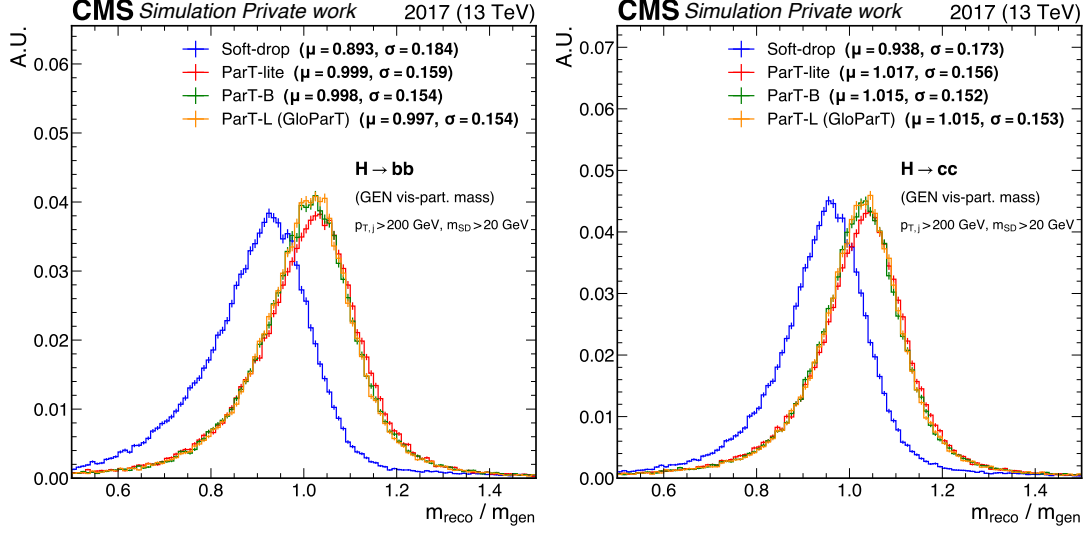


Figure 11.22 A comparison of reconstructed mass resolution on $H \rightarrow b\bar{b}$ (left) and $H \rightarrow c\bar{c}$ (right) jets when the model scales from ParT-lite, to ParT-B, then to ParT-L (GloParT). The regression target is the GEN-level visible-particle mass.

Chapter 12

GloParT: Model Fine-tuning

This chapter will focus on the important implication of building a large, comprehensive GloParT model: its subsequent fine-tuning to downstream tasks. The design of GloParT as a “large model for large-scale classification” ensures it learns a profound jet representation covering a wide range of phase space; thus, one would expect it to generalize to downstream tasks, the ones that dedicated analysis may explore. It is worth noting that this study only considers transfer learning, a specific type of fine-tuning that freezes all model layers before the targeted latent space. This method is most suitable for LHC experiment applications because the large model inference will be delivered in central computing, and the targeted latent space features will be stored in the analysis object data, which is then passed on to downstream analyses.

As a proof-of-concept illustration of its fine-tuning applications, this chapter specifically focuses on two schemes.

In the first scenario, introduced in Section 12.1, the model is fine-tuned to tag a new category of jet that is not included in the training sample. This is most likely met in analyses because the signature of the signal may not necessarily be included in the model’s direct classification nodes. This is analogous to the deep learning vision task, where large vision models pre-trained on ImageNet can be fine-tuned to other image domains. In the study, in order to benchmark the fine-tuning ability, the tagging performance is compared with the maximum ability in this task, given the model’s original capacity.

In the second scenario, introduced in Section 12.2, the model is fine-tuned to derive other jet models within the CMS experiment. A study is performed to fine-tune this mass-decorrelated GloParT model to perform top jet tagging tasks without imposing requirements to decorrelate with mass. During fine-tuning, the jet-level variables m_{SD} and p_T are passed to the network ^①. The study will benchmark the performance with the ParticleNet top tagger as an official tool in CMS, and perform a theoretical calculation on this problem to explain the

^① Author’s note (edited November, 2024): This model fine-tuning scenario has been integrated into the CMS software alongside the “GloParT stage-3” model to infer mass-correlated top tagging discriminant. The base GloParT stage-3 model is first run to extract and store the hidden-layer neuron values. The m_{SD} and p_T variables are subsequently fed into the fine-tuned MLP layers, where they are processed with the stored neuron values to produce the final discriminant scores.

fine-tuning network designs.

12.1 Application on novel tagging

12.1.1 New jet sample and labels

The potential of tagging a leptoquark (LQ) with $LQ \rightarrow b\tau$ decay is studied, where three tau leptons decay channels, τ_e , τ_μ , and τ_h , are considered. The jet is produced in a di-leptoquark event production process. The configuration of event H_T and the leptoquark mass is the same as nonresonance di- H production used to generate a part of the two-prong jet (see Section 11.1.1). This provides a similar control on jet p_T and m_{SD} .

To assign a truth label to the generated jet, the same matching criteria are adopted for the b quark and tau lepton. This produces three new jet labels for three τ decay modes.

12.1.2 Fine-tuning details

The third-to-last layer of the original GloParT model, with a dimension of 256, is used as a latent space feature. The transfer learning of GloParT from these features is then equivalent to treating them as 128 jet features and training a simple MLP. This is exactly the fine-tuning picture to foresee for future analysis applications. The training of MLP is delivered in a multiclass classification task. Besides the three novel classes for $LQ \rightarrow b\tau$, the remaining ones are 13 two-prong categories, 17 $t \rightarrow bW$ classes, and the 5 QCD categories, resulting in 38 classes in total. These categories are brought in such that discriminants can be defined to separate the LQ signature against the two-prong classes (such as $H \rightarrow b\bar{b}$), the $t \rightarrow bW$ classes (such as $t \rightarrow b\tau\nu$), and QCD classes. The MLP comprises two layers, with dimensions of 1024 and 38. Figure 12.1 gives an intuitive illustration of the network setup for this fine-tuning task.

The training of the MLP adopts the same input preprocessing procedure as in the original GloParT training. A sampling-based reweighting on the jet (p_T , m_{SD}) bins is performed on the fly of training. The training procedure is very lightweight. Each epoch is defined by training on 256 k jets, substantially smaller than 7.68 M jets used in the original training. The training process consists of 15 epochs. The 1CYCLE LR policy^[195] for fine-tuning, where the LR changes every batch instead of every epoch. It is raised from an initial LR to a maximum LR, designed as 2×10^{-3} , then dropped to some minimum LR much lower than the initial LR. The same LOOKAHEAD optimizer with the inner optimizer RADAM is adopted. Figure 12.2 shows several metrics for training and validation, together with the LR. It is clear that the training converges very quickly.

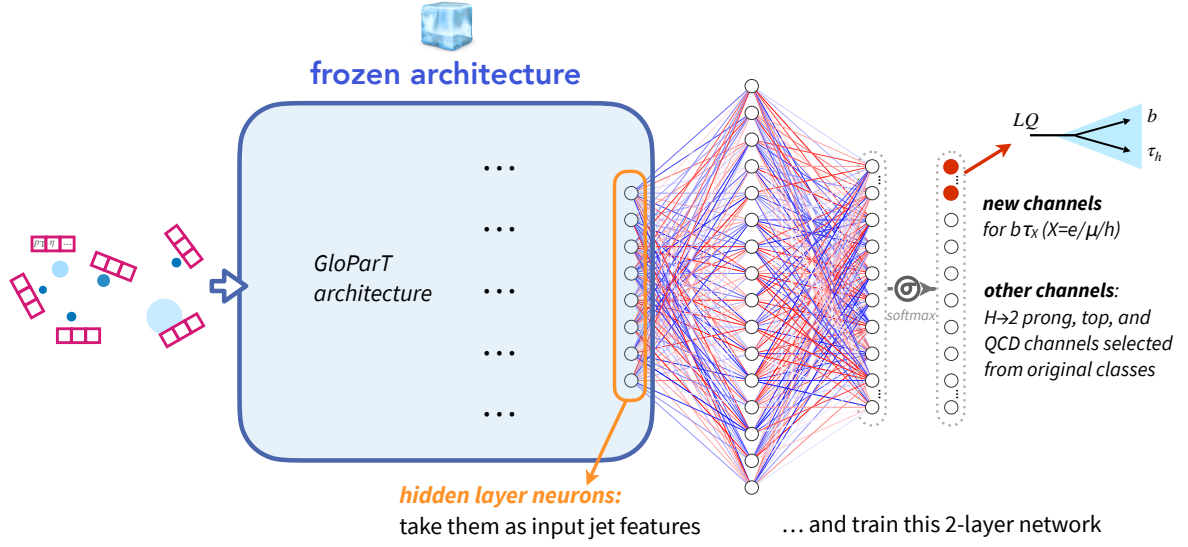


Figure 12.1 Illustration on the fine-tuning experiment of GloParT to perform jet tagging on novel jet categories.

12.1.3 Fine-tuning performance

Figure 12.3 shows the performance for identifying the $LQ \rightarrow b\tau_h$ signal against the QCD, $H \rightarrow b\bar{b}$, $H \rightarrow \tau_h\tau_h$, or $t \rightarrow bW \rightarrow \tau_h\nu$ backgrounds, comparing with a training-from-scratch GloParT model with the three classes included initially. The latter signifies the best performance the present model's capacity can afford.

The results imply that the GloParT model illustrates good transfer learning ability, and it has achieved superior performance to veto against QCD, $H \rightarrow b\bar{b}$, $H \rightarrow \tau_h\tau_h$ backgrounds. Separating the signal and $t \rightarrow b\tau_h\nu$ is the most challenging. The background rejection is around ~ 2 less than the optimal tagging performance. Overall, the result is satisfying because the fine-tuning step is very lightweight compared to training everything from scratch, and the performance is competitive to a certain extent.

12.2 Application on deriving new models

As a comprehensively trained model, it is straightforward to consider it as a base jet model in data analysis such that other jet models can be fine-tuned on top of it. Therefore, the second application targets a derivation of the new model from the pre-trained GloParT. The CMS experiment has developed a ParticleNet model to identify large- R jet initiated from H , W , Z , or top quark without decorrelating with the jet mass. This study aims to derive the same model, specifically reproducing the separating power between top and QCD jets, using the

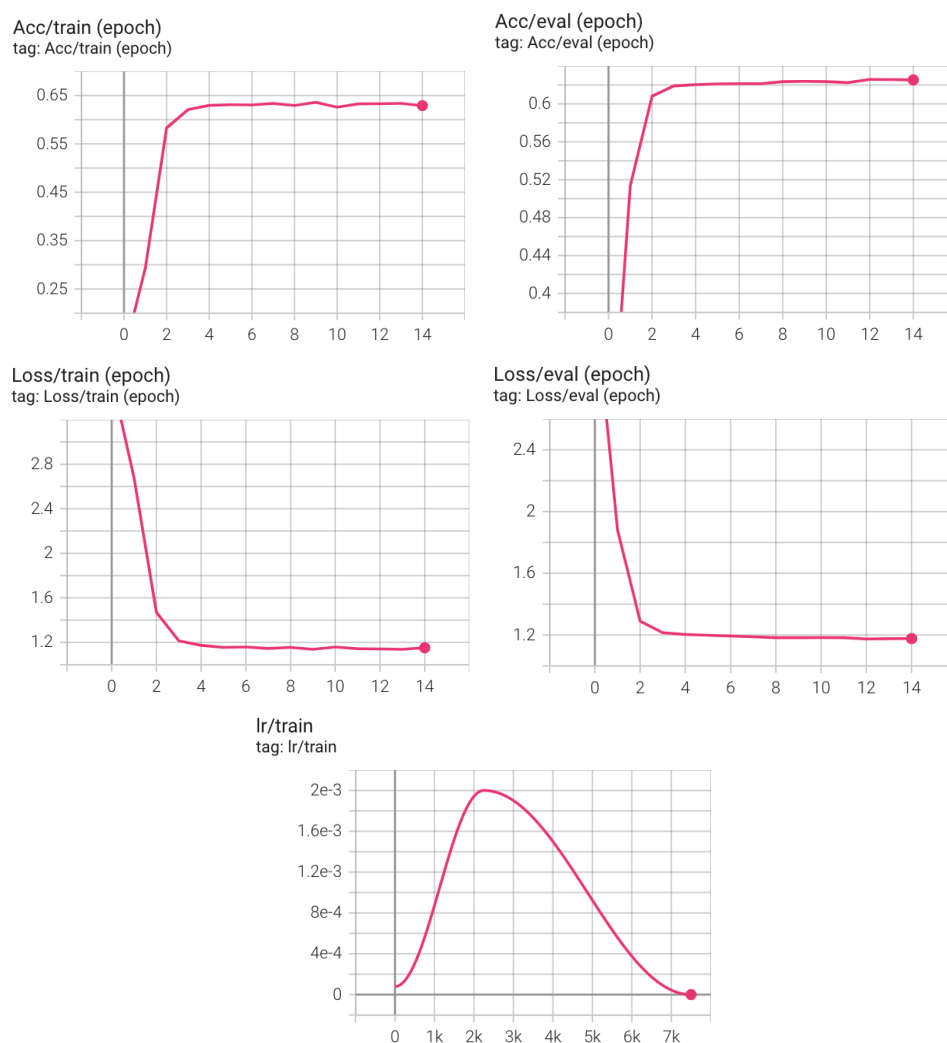


Figure 12.2 Metrics and the per-batch learning rate (bottom) monitored during the fine-tuning experiment of GloParT to perform tagging on novel jet categories. The metrics include the per-epoch classification accuracy (top) and the per-epoch loss (middle) for the training (left) and evaluation (right) procedure.

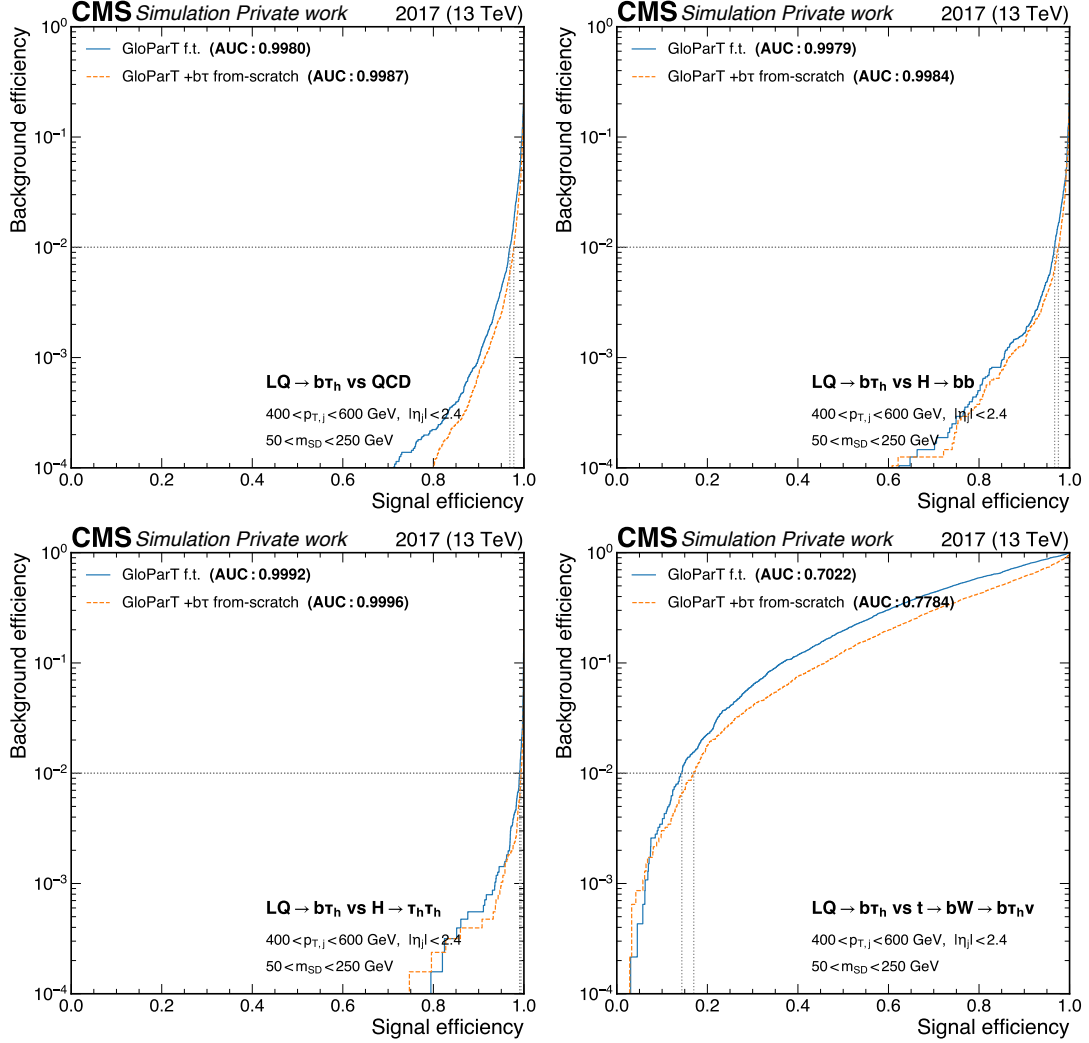


Figure 12.3 The performance on tagging a new class of jet, $LQ \rightarrow b\tau_h$, versus the QCD, $H \rightarrow b\bar{b}$, $H \rightarrow \tau_h\tau_h$, or $t \rightarrow bW \rightarrow \tau_h\nu$ jets as backgrounds. The performance is compared with a training-from-scratch GloParT model with the new classes initially added.

original GloParT but incorporating auxiliary features jet m_{SD} and p_{T} .

Essentially, this fine-tuning experiment attempts to recover the model’s ability to discriminate top and QCD jets optimally in the dependence of m_{SD} . It can be shown from the statistical theory, however, that the optimal classifier in both the mass-decorrelated and mass-correlated cases can be converted to each other by taking m_{SD} as a dependent variable. In this sense, it is reasonable to envision that by passing in the m_{SD} variable and allowing it to entangle with the existing GloParT jet representation within the fine-tuning network, the network will learn such dependency of m_{SD} and hence recover the ability to do optimal mass-correlated tagging.

12.2.1 Experiment setup

Jet classes, input and output

The entire $t \rightarrow bW$ and QCD categories of the GloParT model are used to derive the new model. That means, for the last fully-connected layer in GloParT, only connections to these corresponding output nodes are preserved while the others are removed. This results in 22 output nodes.

To maintain consistency with the training of the original ParticleNet tagger in CMS, i.e. a similar training dataset is adopted. They include top jets produced from $Z' \rightarrow t\bar{t}$ process with varying Z' mass but the top quark mass fixed to the SM value (the same with the test dataset, in Section 11.1.3), and the original QCD sample for GloParT training. Jets are assigned top and QCD labels using the same matching criteria as in the GloParT training.

The fine-tuning input involves two parts: (1) the latent features with dimension 256 from the original GloParT model, and (2) the jet m_{SD} and p_{T} in addition.

Network design

The network design is illustrated in Fig. 12.4. The original two-layer MLP that converts the latent features to the final QCD and $t \rightarrow bW$ logits are preserved; in addition, an auxiliary three-layer MLP network is introduced, which encodes the features of the latent features together with the jet p_{T} and m_{SD} variables. The layer dimensions of the auxiliary network are (258, 256, 256, 22).

The output nodes of the auxiliary path are added back to the original output logits, then proceed with the softmax. The auxiliary network can be considered to provide “residues” to correct the original logits. The design is reasonable from a theoretical standpoint, as will be explained in Section 12.2.2.

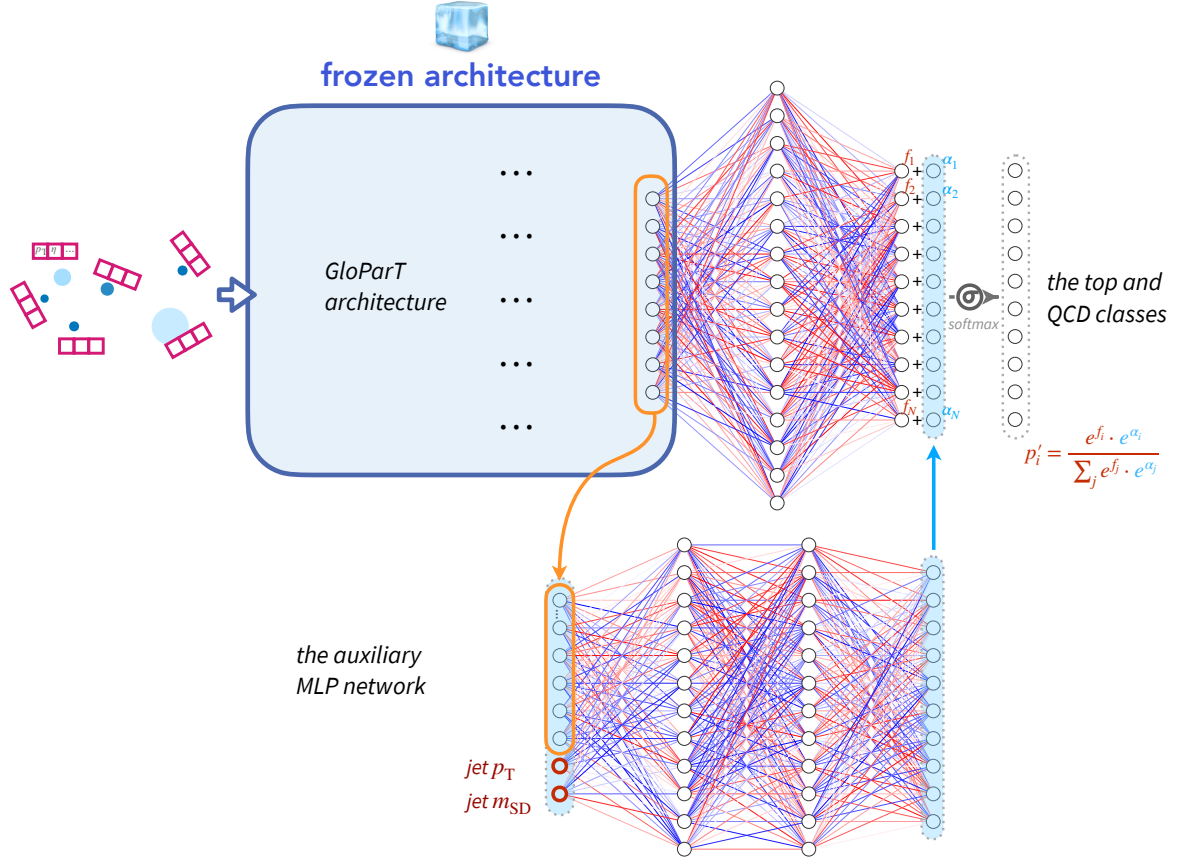


Figure 12.4 Illustration on the fine-tuning experiment of GloParT to derive the mass-correlated top-jet tagging model.

Preprocessing and training The preprocessing of input data adopts the development of the ParticleNet mass-correlated tagger in CMS. In this case, the sampling-based reweighting is performed on the 1D jet p_T template but not on the 2D (p_T, m_{SD}) template so that the mass decorrelation requirement is discarded.

The detailed training setup is similar to the first experiment. The training is performed on 15 epochs, each undergoing a sample size of 256 k. The LR is adjusted based on the 1CYCLE policy. The LOOKAHEAD optimizer with the inner optimizer RADAM is employed.

12.2.2 A guide in theory

It is worth noting that one can find a statistical foundation in this fine-tuning network design. A theoretical relation between the optimal mass-correlated and the mass-decorrelated classifier model can be found, and they differ only on a dependency of m_{SD} .

Denote an input jet as x and a jet class with class index c represented as a probability density function $\rho_c(x)$. Given that the mass-decorrelated training is achieved via sampling-

based reweighting on m_{SD} , it changes the prior PDF to $\rho'_c(x)$ which is dependent on these variables,

$$\rho'_c(x) = \rho_c(x) \cdot \lambda_c(m_{\text{SD}}(x), p_{\text{T}}(x)). \quad (12.1)$$

The $\lambda_c(\cdot)$ is thus the reweighting factor for the class c , which is a function of m_{SD} .

For a normal training without sampling-based reweighting on m_{SD} , denote the output probabilities of an optimal network as p_c ($c = 1, \dots, C$), then for any class A and B , one has

$$\frac{p_A(x)}{p_A(x) + p_B(x)} = \frac{\rho_A(x)}{\rho_A(x) + \rho_B(x)}. \quad (12.2)$$

As an analogue, in the mass-decorrelated network training, the output probabilities p'_c should satisfy

$$\frac{p'_A(x)}{p'_A(x) + p'_B(x)} = \frac{\rho_A(x) \cdot \lambda_A(m_{\text{SD}}(x), p_{\text{T}}(x))}{\rho_A(x) \cdot \lambda_A(m_{\text{SD}}(x), p_{\text{T}}(x)) + \rho_B(x) \cdot \lambda_B(m_{\text{SD}}(x), p_{\text{T}}(x))}. \quad (12.3)$$

Looking back at the network design in Fig. 12.4, denoting the original network's output logits (before softmax) as f_i , the probabilities should be $p_i = e^{f_i} / \sum_j e^{f_j}$. Hence, the binary discriminant we defined on class A vs B should be

$$\frac{p_A}{p_A + p_B} = \frac{e^{f_A}}{e^{f_A} + e^{f_B}}. \quad (12.4)$$

When a residual path is added, an additive bias α_i is provided to each neuron, and the modified probabilities are modified as

$$\frac{p'_A}{p'_A + p'_B} = \frac{e^{f_A} \cdot e^{\alpha_A}}{e^{f_A} \cdot e^{\alpha_A} + e^{f_B} \cdot e^{\alpha_B}}. \quad (12.5)$$

This exactly corresponds to Eq. (12.3) as long as we allow the bias α_i for each class i to encode the m_{SD} variable and estimate the reweighting factor $\log(\lambda_i(m_{\text{SD}}(x), p_{\text{T}}(x)))$. It just implies that the auxiliary network can be an estimate of $\log(\lambda_i(\cdot))$.

The derivation is about adding an auxiliary path to the normal mass-correlated network to learn a reweighting function and make it mass-decorrelated. Our fine-tuning application in this section is an inverse procedure; therefore, it learns a $\log(1/\lambda_i(\cdot))$ function that converts $\rho'_c(x)$ back to $\rho_c(x)$.

As an experimental validation of this interpretation, the contribution of the auxiliary path, in terms of $\exp(\alpha_i)$, is visualized in Fig. 12.5. In this validation, a special auxiliary network is employed that only takes the jet p_{T} and m_{SD} as input. After fine-tuning, the learned auxiliary network is studied by feeding the $(p_{\text{T}}, m_{\text{SD}})$ pairs, with p_{T} fixed at 500 GeV and m_{SD} varying from 30–260 GeV. This allows obtaining the values of $\exp(\alpha_i)$ as a function of m_{SD} . In

Fig. 12.5, we present a plot of $\exp(\alpha_i + \Delta)$, where we have shifted all α_i values by a constant Δ to ensure the minimum value is zero. The result aligns with the understanding of the role of m_{SD} in assisting the mass-correlated tagging. For instance, for the class i being the three-prong bcq or bqq case, $\exp(\alpha_i + \Delta)$ has observed a large value around the top mass. An explanation of the performance of $\exp(\alpha_i)$ for different top tagging classes is annotated in the figure.

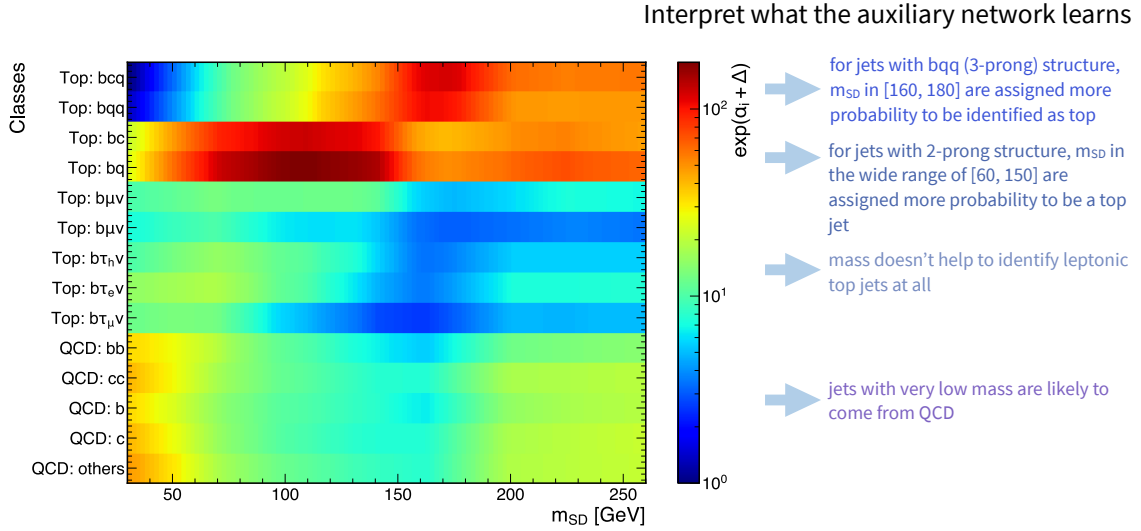


Figure 12.5 A visualization of $\exp(\alpha_i + \Delta)$, where α_i is the output value of the auxiliary network for class i , and it is shifted by a constant Δ to ensure the minimum value is zero. An annotation about how to explain the performance of different top tagging classes is shown in the plot.

As a final remark, although it is adequate to only encode m_{SD} in the auxiliary network, in reality, better performance is achieved if the entire latent features are all encoded into the auxiliary network, and this eventually leads to a network design in Fig. 12.4.

12.2.3 Fine-tuning performance

Figure 12.6 shows the performance of tagging top jets while vetoing against QCD jets in ROC curves. The fine-tuned GloParT model and the original GloParT model are both shown in the plot, illustrating a big leap in performance when the additional mass information is used to distinguish top and QCD jets. Also, their performance is compared with the ParticleNet top tagger developed in CMS to benchmark the performance achieved so far. It is promising to see that the fine-tuned GloParT outperforms the ParticleNet by a factor of ~ 2 in background rejection.

This marks a major success in this fine-tuning application, as one may expect to infer a large jet model and then use it to derive other necessary jet models within an experimental

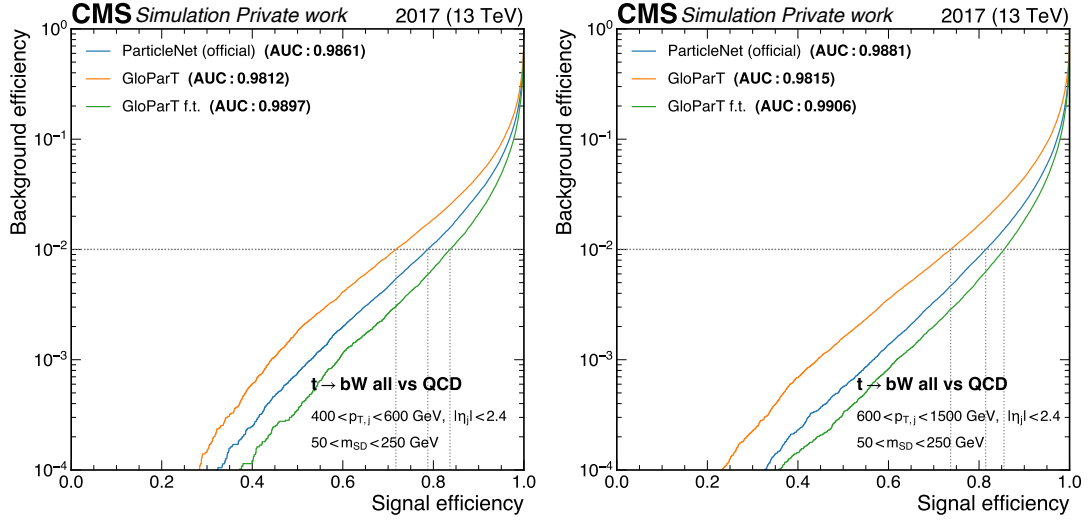


Figure 12.6 A comparison of performance on top versus QCD jet tagging among the CMS ParticleNet top-jet tagging model, the GloParT model without fine-tuning such that the tagging is decorrelated with jet mass, and the GloParT model with fine-tuning (GloParT f.t.). The ROC curve is displayed in both low (left) and high (right) p_T region.

workflow. For example, by deploying GloParT's inference in central CMS computing, the mass-correlated top tagger can be derived by adding the customized MLP after its latent features, effectively saving computing resources.

Chapter 13

The Sophon model

The GloParT model has demonstrated its efficacy as a universal jet model within the CMS experiment. It surpasses previous taggers in existing tagging tasks and also serves as a robust base model for fine-tuning across various downstream analysis scenarios, delivering superior performance.

To extend the design concept of GloParT beyond the CMS experiment, we replicate its capabilities in experimental scenarios external to CMS. This chapter presents the reproduction results of the GloParT model, trained on a DELPHES-based dataset while adhering closely to its original development framework. This effort has led to the creation of the novel JETCLASS-II dataset, characterized by jet kinematics spanning a wide range of p_T and mass values, including diverse jet phase space configurations. Additionally, we introduce the Sophon model, pre-trained as a multiclass classifier on the JETCLASS-II dataset.

The pre-training methodology employed is referred to as *signature-oriented*, where the “signature” corresponds to generator-level jet properties that characterize the unique origin information. This approach proposes a training scheme that enables the model to associate jet input data with their generator-level properties. The simplest form of such a signature is the truth label of a jet under this definition. Consequently, training a multiclass classifier to identify various jet labels becomes the foundational approach for comprehensive pre-training.

Notably, the signature-oriented pre-training paradigm was developed under the concept of *Signature-Oriented Pre-training for Heavy-resonance Observation (Sophon)*^[196], which was primarily aimed at accelerating resonance search programmes at the LHC. The pre-trained model is referred to as the *Sophon model* within this context.

The remainder of this chapter is organized as follows. Section 13.1 provides a detailed overview of the JETCLASS-II dataset. The model pre-training methodology is described in Section 13.2, and its performance benchmarks are presented in Section 13.3. Section 13.4 discusses the implications of the Sophon model for LHC resonance searches. Finally, a discussion and outlook are provided in Section 13.5.

13.1 The JetClass-II dataset

The JETCLASS-II dataset comprises 188 jet classes. Compared to JETCLASS (described in Section 2.3), it features resonant jets with a much more extensive set of final states and broader p_T and mass ranges. In simple terms, a resonant jet may contain two, three, or four prongs, with each prong originating from a quark, gluon, or lepton. These jets are finely categorized based on particle types, quark or lepton flavours, their combinations, as well as the decay modes of tau leptons.

The generator-level simulation of these processes is described in detail below.

13.1.1 Simulation details

The JETCLASS-II dataset includes both resonant jets and QCD jets. Its categorization strategy is generally similar to that of the GloParT training set (illustrated in Fig. 11.1) but incorporates simplifications in the generation of three- and four-prong final states. Specifically, multiple $H \rightarrow W^{(*)}W^{(*)}$ and $H \rightarrow Z^{(*)}Z^{(*)}$ scenarios are merged, focusing solely on $X \rightarrow YY \rightarrow 4$ -prong final states, where both X and Y are spin-0 resonances with variable masses.

In summary, the resonant jets originate from the following processes:

- (1) $X \rightarrow 2$ prong signatures with neutral resonance X ,
- (2) $X \rightarrow 2$ prong signatures with charged resonance X ,
- (3) $X \rightarrow Y^{(*)}Y^{(*)} \rightarrow 4$ prong signatures.

The case (1) is generated using the $pp \rightarrow HH$ process, simulated with MG 2.9.18^[25] generator at LO accuracy, employing the HEFT model. A total of 16 M events are produced. To control the resonant jet's p_T and mass, the minimum p_T of H at the hard-scattering level is sampled at 50 logarithmically spaced points within the range (100, 2500) GeV. Additionally, the H mass is uniformly sampled from (15, 500) GeV in steps of 5 GeV. The decay of the H resonance and subsequent parton showering are simulated using PYTHIA 8.3^[28]. The decay modes and corresponding branching ratios are as follows: bb ($\frac{1}{8}$), cc ($\frac{1}{8}$), ss ($\frac{1}{8}$), dd ($\frac{1}{16}$), uu ($\frac{1}{16}$), gg ($\frac{1}{8}$), ee ($\frac{1}{16}$), $\mu\mu$ ($\frac{1}{16}$), and $\tau\tau$ ($\frac{1}{4}$). Subsequent τ decay follows the SM configuration.

The case (2) is generated using the $pp \rightarrow H^+H^-$ process, simulated with MG at LO accuracy with the 2HDM model. A total of 12 M events are produced. The configuration for the minimum p_T of $H^\pm$ and its mass is identical to that used in the case (1). The decay of the $H^\pm$ resonance and subsequent parton showering are handled by PYTHIA 8.3. The decay modes and corresponding branching ratios are as follows: du ($\frac{1}{6}$), su ($\frac{1}{6}$), bu ($\frac{1}{6}$), cd ($\frac{1}{6}$), cs ($\frac{1}{6}$), and bc ($\frac{1}{6}$).

The case (3) is generated using the $pp \rightarrow hh$ process, simulated with MG at LO accuracy with the 2HDM model. A total of 120 M events are produced. The configuration for the minimum p_T of h and its mass is identical to that used in previous cases. The decay processes are simulated using PYTHIA 8.3. In this case, the resonance h decays into HH and H^+H^- , each with a branching ratio of $\frac{1}{2}$. Here, h , H , and $H^\pm$ represent the two CP -even Higgs bosons and the charged Higgs boson, respectively, as defined in the 2HDM model. The masses of H and $H^\pm$ are parameterized as λm_h , where λ is sampled uniformly within the range (0.2, 0.8). The decay of H follows the same configuration as in the case (1), while the decay of $H^\pm$ is similar to that in the case (2), with the addition of decay channels involving neutrino final states. The branching ratios for $H^\pm$ decay modes are as follows: du ($\frac{5}{48}$), su ($\frac{5}{48}$), bu ($\frac{5}{48}$), cd ($\frac{5}{48}$), cs ($\frac{5}{48}$), bc ($\frac{5}{48}$), $e\nu$ ($\frac{1}{16}$), $\mu\nu$ ($\frac{1}{16}$), and $\tau\nu$ ($\frac{1}{4}$). Subsequent τ decays follow the SM configuration.

QCD jets are initiated by simulating a $2 \rightarrow 2$ QCD multijet process using the PYTHIA 8.3 generator, producing a total of 20 M events. The minimum p_T of the hard-scattering process is sampled logarithmically within the range (100, 5000) GeV to ensure extensive coverage of jet p_T and mass.

The simulated events from all the aforementioned processes are proceeded via DELPHES 3^[63] for fast simulation of the detector response and object reconstruction. The DELPHES simulation card is adapted from the CMS detector configuration, with modifications to the impact parameter resolution of charged particles to align with the CMS tracker resolution, similar to the JETCLASS simulation^[65]. In addition, pileup effects with an average of 50 pileup interactions per event are included, adapted from the CMS detector configuration with pileup^[63]. The PUPPI algorithm^[36] is applied for pileup removal, adapted from the CMS detector configuration in Phase II^[63] with additional parameter modifications based on the Phase-I CMS detector condition and taking the CMS experimental configuration as a reference^[35]. The PUPPI algorithm assigns a value between 0 and 1 to each E-flow object, representing the probability that the object originates from the primary interaction vertex. The four-momentum of each object is scaled accordingly. The processed E-flow objects are used to cluster large- R jets using the anti- k_T algorithm^[33] with $R_0 = 0.8$. The modified DELPHES card is a major highlight of the JETCLASS-II dataset, as it creates detector conditions and a particle reconstruction workflow closely resembling those of the CMS experiment during the Run 2 data-taking period.

After DELPHES processing, jets are required to satisfy $200 < p_T < 2500$ GeV, $20 < m_{SD} < 500$ GeV, and $|\eta| < 2.5$ to be included in the JETCLASS-II dataset.

Table 13.1 Summary of the 188 jet labels in the JETCLASS-II dataset.

Major types	Index range	Label names
Resonant jets: $X \rightarrow 2$ prong	0–14	$bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, \mu\mu, \tau_h\tau_e, \tau_h\tau_\mu, \tau_h\tau_h$
Resonant jets: $X \rightarrow 3$ or 4 prong	15–160	$bbbb, bbcc, bbss, bbqq, bbgg, bbee, bb\mu\mu, bb\tau_h\tau_e, bb\tau_h\tau_\mu, bb\tau_h\tau_h, bbb, bbc, bbs, bbq, bbg, bbe, bb\mu, cccc, ccss, ccqq, ccgg, ccee, cc\mu\mu, cc\tau_h\tau_e, cc\tau_h\tau_\mu, cc\tau_h\tau_h, ccb, ccc, ccs, ccq, ccg, cce, cc\mu, ssss, ssqq, ssgg, ssee, ss\mu\mu, ss\tau_h\tau_e, ss\tau_h\tau_\mu, ss\tau_h\tau_h, ssb, ssc, sss, ssq, ssg, sse, ss\mu, qqqq, qqgg, qqee, qq\mu\mu, qq\tau_h\tau_e, qq\tau_h\tau_\mu, qq\tau_h\tau_h, qqb, qqc, qqs, qqg, qqe, qq\mu, gggg, ggee, gg\mu\mu, gg\tau_h\tau_e, gg\tau_h\tau_\mu, gg\tau_h\tau_h, ggb, ggc, ggs, ggq, ggg, gge, gg\mu, bee, cee, see, qee, gee, b\mu\mu, c\mu\mu, s\mu\mu, q\mu\mu, g\mu\mu, b\tau_h\tau_e, c\tau_h\tau_e, s\tau_h\tau_e, q\tau_h\tau_e, g\tau_h\tau_e, b\tau_h\tau_\mu, c\tau_h\tau_\mu, s\tau_h\tau_\mu, q\tau_h\tau_\mu, g\tau_h\tau_\mu, b\tau_h\tau_h, c\tau_h\tau_h, s\tau_h\tau_h, q\tau_h\tau_h, g\tau_h\tau_h, qqqb, qqqc, qqqs, bbcq, ccbs, ccbq, ccsq, sscq, qqbc, qqbs, qqcs, bcsq, bcs, bcq, bsq, csq, bcev, csev, bqev, cqev, sqev, qqev, bc\mu\nu, cs\mu\nu, bq\mu\nu, cq\mu\nu, sq\mu\nu, qq\mu\nu, bc\tau_e\nu, cs\tau_e\nu, bq\tau_e\nu, cq\tau_e\nu, sq\tau_e\nu, qq\tau_e\nu, bc\tau_\mu\nu, cs\tau_\mu\nu, bq\tau_\mu\nu, cq\tau_\mu\nu, sq\tau_\mu\nu, qq\tau_\mu\nu, bc\tau_h\nu, cs\tau_h\nu, bq\tau_h\nu, cq\tau_h\nu, sq\tau_h\nu, qq\tau_h\nu$
QCD jets	161–187	$bbccss, bbccs, bbcc, bbcss, bbcs, bbc, bbss, bbs, bb, bccss, bccs, bcc, bcss, bcs, bc, bss, bs, b, ccss, ccs, cc, css, cs, c, ss, s, \text{others}$

13.1.2 Jet labelling

This subsection introduces the assignment of 188 jet labels to simulated jets from the various processes. The labels, summarized in Table 13.1, fall into three major categories. Resonance jets produced in the aforementioned cases (1) and (2) are assigned a label corresponding to the $X \rightarrow 2$ -prong scenario or discarded if unmatched. Jets from the case (3) are categorized under $X \rightarrow 3$ -prong or 4-prong labels, while QCD jets are assigned QCD-specific labels.

The label assignment follows a detailed matching criterion, described below.

For a resonant jet, the first-generation decay products of the resonance X or $Y^{(*)}Y^{(*)}$ are counted. These include two particles in cases (1) and (2), and four particles in case (3). The decay products are categorized as follows: u and d quarks are grouped and represented as q , while the tau lepton is further classified into three exclusive subclasses: τ_e , τ_μ , and τ_h , corresponding to sequential decays into an electron, muon, or hadrons, respectively. This classification results in the following truth particle types that may appear in the final state: $b/c/s/q/g/e/\mu/\tau_e/\tau_\mu/\tau_h$, excluding neutrinos. To assign a truth label to a jet, the matching rule requires all truth particles specified in the label to satisfy the following conditions.

1. For all particle types except τ_e and τ_μ , the angular distance between the particle and the

jet must satisfy $\Delta R(\text{particle}, \text{jet}) < R_0$.

2. For τ_e and τ_μ , the matching condition applies to their decay daughters (e or μ), requiring $\Delta R(\text{daughter}, \text{jet}) < R_0$.

For a QCD jet, a label is exclusively assigned based on the number of b , c , and s quarks identified from the PYTHIA parton list prior to hadronization. These quarks must satisfy $p_T > 10 \text{ GeV}$ and be matched to the jet axis within $\Delta R < R_0$. For each quark flavour, the count is categorized into three cases: 0, 1, or ≥ 2 . This classification results in a total of 27 QCD classes.

The JETCLASS-II dataset contains approximately 139 M labelled jets, in its training and validation set. This includes 22 M resonant jets with 2 prongs, 99 M resonant jets with 3 or 4 prongs, and 18 M QCD jets.

13.1.3 Variables in JetClass-II

The jet constituent features serve as input for neural network training. These features are derived from E-flow objects and include kinematic properties, particle identification variables, and impact parameter features, closely following the setup of the JETCLASS dataset. The dataset also provides jet-level kinematic variables, which are used to construct the neural network inputs. Additional variables include jet N -subjettiness observables^[69] up to $N = 4$, which represent high-level jet characteristics.

Moreover, several generator-level variables are provided to capture jet signatures. These include the jet label, determined by counting the matched truth particles as described earlier, and a detailed list of matched particles, specifying their types and kinematic properties.

13.2 Pre-training of the Sophon model

The Sophon model is pre-trained as a multiclass classifier to distinguish 188 jet classes. Since its development largely follows the training procedure of GloParT, a brief overview is provided below.

The Sophon model utilizes the Particle Transformer architecture, as described in Ref.^[65], with modifications to the fully-connected MLP, which is extended to two layers with dimensions of (512, 188). The core of the Sophon model comprises 6 particle attention blocks and 2 class attention blocks, featuring an embedding dimension of 128 and 8 attention heads.

The initial particle features are embedded using a 3-layer MLP with dimensions (128,

512, 128), while pairwise particle features are embedded with a 4-layer elementwise MLP with dimensions (64, 64, 64, 8). The GELU activation function is used throughout the architecture. The Sophon model contains a total of 2.3 M parameters, comparable in size to the original Particle Transformer (ParT) introduced in Ref.^[65].

The model processes input derived from all jet constituents, including kinematic features, particle identification variables, and impact parameter information. The input structure is similar to the ParT training on the JETCLASS dataset (see Section 4.2). However, the Sophon model introduces scaled kinematic inputs, where features related to constituent or jet four-momentum are scaled by a coefficient such that the jet p_T is normalized to 500 GeV. This scaling enhances the scale-invariance of the model, improving its tagging performance across a wide range of jet masses.

The training of the Sophon model incorporates a sampling-based reweighting technique as a critical component, following the same methodology employed in the GloParT training process described in Section 11.2.4. The predefined sampling weights are derived from a two-dimensional histogram spanned by (p_T, m_{SD}) , within the range $200 < p_T < 2500$ GeV and $20 < m_{SD} < 500$ GeV. To define the reweighting classes, the original 188 finely classified categories are partially merged: classes differing only by quark flavour have been combined, and all 27 QCD jet classes are grouped into a single class. This approach aligns with the strategy employed in GloParT training, resulting in a total of 30 reweighting classes.

The model is trained with a batch size of 512 and an initial LR of 5×10^{-4} . The full JETCLASS-II dataset is partitioned into 80/20% for training/validation. Training is conducted over 80 epochs, with each epoch processing 10 M samples. The optimizer and LR scheduler configurations remain consistent with those used for GloParT training, as detailed in Section 11.2.5.

13.3 Performance benchmark

Similar to GloParT, we demonstrate the capabilities of the Sophon model in two tasks. The first is a task based on the presence of final states, where we select the $X \rightarrow b\bar{b}$ tagging task to evaluate its direct tagging ability. The second task tests its transfer learning capability. Specifically, we leverage its hidden layers to adapt the model to a new task that was not included during pre-training: $X \rightarrow bs$ tagging, where bs corresponds to either $b\bar{s}$ or $\bar{b}s$. The two application scenarios of the Sophon model are illustrated in Fig. 13.1.

Our test dataset consists of dedicated signal and background samples, which will be de-

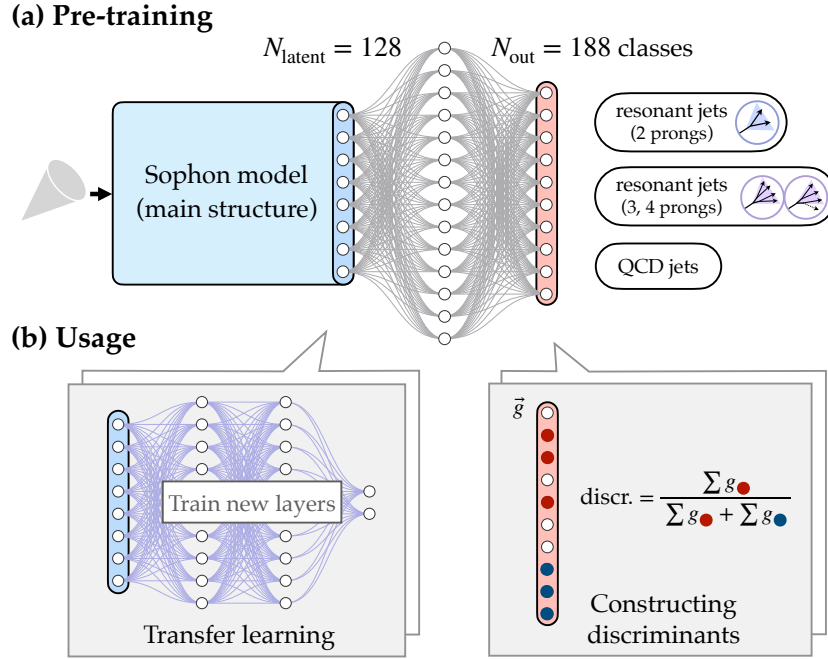


Figure 13.1 (a) Pre-training of the Sophon model as a multiclass classifier. (b) Application of the Sophon model via the construction of discriminants from selected outputs and transfer learning performed using the hidden representations.

scribed in Section 13.3.1. The performance results can be found in Sections 13.3.2 and 13.3.3.

13.3.1 Production of the SM and signal processes

We simulated the complete set of SM processes to emulate actual experimental scenarios, requiring events to pass a boosted-jet-based trigger. This dataset corresponds to 100 fb^{-1} of data collected in LHC pp collisions at $\sqrt{s} = 13 \text{ TeV}$, replicating the Run-2 data-taking conditions. Specifically, the trigger requires the scalar p_T sum of all $R = 0.8$ jets to exceed 800 GeV, with at least one jet having a trimmed mass^[68] greater than 50 GeV.

Events passing the trigger include approximately 5×10^7 events from QCD multijet processes, 9×10^5 $V(= W/Z)$ +jets events, 3×10^5 events from $t\bar{t}$ and single top quark processes, and additional events from processes such as diboson (VV) and Higgs boson production. The signal process originates from a hypothetical heavy spin-0 (Higgs-boson-like) resonance X_0 with a mass of 200 GeV, decaying to $b\bar{b}$ or bs (namely, $b\bar{s}$ or $\bar{b}s$).

All events, except for those from QCD processes, are simulated using MG and PYTHIA for the ME calculation at LO accuracy and hadronization. The QCD process is simulated entirely with PYTHIA. To meet the high- p_T jet requirements, each MG process applies a predefined (highest feasible) selection on the scalar p_T sum of final-state partons to enhance computational

efficiency.

For both signal and SM processes, the leading $R = 0.8$ jet satisfying the trigger criteria is used to evaluate various algorithms. The QCD event yield is significantly higher than that of other SM processes; therefore, the tagger developed for identifying $X \rightarrow b\bar{b}$ or $X \rightarrow bs$ jets focuses on effectively rejecting QCD jets while maximizing the purity of signal jets.

13.3.2 Directly tagging capability

According to the methodology outlined in Section 11.2, the optimal discriminant for $X \rightarrow b\bar{b}$ tagging is designed to reject QCD jets. Specifically, the discriminant is defined as

$$\text{Sophon } X \rightarrow b\bar{b} \text{ vs. QCD disc.} = \frac{p(X \rightarrow b\bar{b})}{p(X \rightarrow b\bar{b}) + \sum_{l=1}^{27} p(\text{QCD}_l)}, \quad (13.1)$$

where $p(X \rightarrow b\bar{b})$ represents the $X \rightarrow b\bar{b}$ output score, and $p(\text{QCD}_l)$ corresponds to the scores of the 27 QCD classes. Ideally, this approach is equivalent to training a binary classifier DNN to distinguish $X \rightarrow b\bar{b}$ jets from undivided QCD jets and then using the $X \rightarrow b\bar{b}$ score as the discriminant.

The top panel of Fig. 13.2 illustrates the tagging performance in terms of the discovery significance, $Z^{[197]}$, for $X \rightarrow b\bar{b}$ jets with $m_X = 200$ GeV. This performance is shown as a function of the SM background selection efficiency, based on 40 fb^{-1} of data and considering events within the mass window $150 < m_{\text{SD}} < 230$ GeV. The search significance is calculated by

$$Z = \sqrt{2((s+b)\log(1+s/b)-s)}, \quad (13.2)$$

where s and b represent the post-selection yields of signal and background events, respectively, within the defined mass window. The curve has an equivalent interpretation to standard ROC curves, as employed in the GloParT studies, to evaluate the performance of the tagger. The bottom panel shows the stacked distribution of all SM processes at a background selection efficiency of $\epsilon_B = 10^{-4}$. The distribution exhibits a smoothly falling background, with the signal appearing as a bump-like structure near the resonance mass value.

Several tagging models are compared in Fig. 13.2. To illustrate the current tagging performance achievable at LHC experiments, we train two dedicated tagging models for both tasks: one using the state-of-the-art ParT^[65] architecture and another using the ParticleNet model^[90]. These architectures are representative of the current tagging capabilities within the CMS experiment^[109, 113, 115]. Both models are trained as binary classifiers to distinguish $X \rightarrow b\bar{b}$ jets

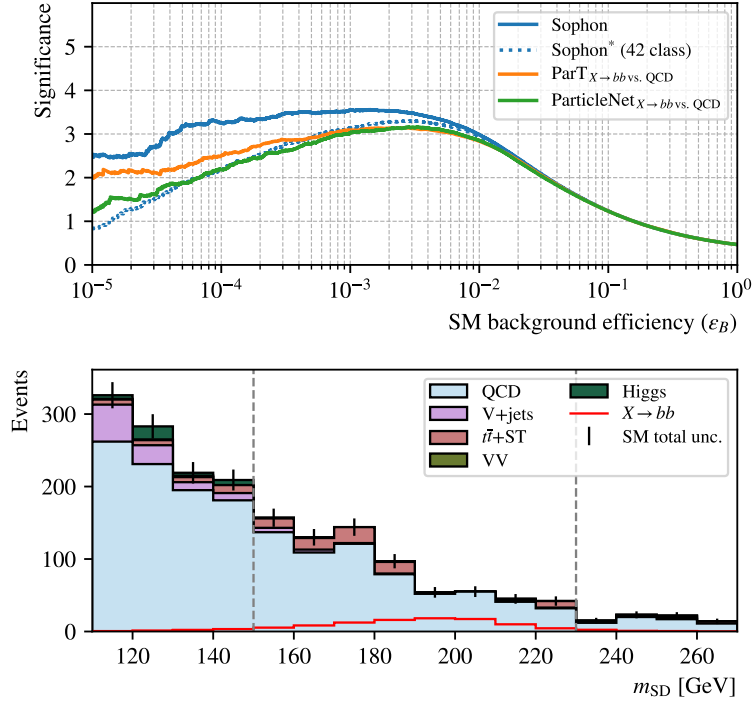


Figure 13.2 Top: Comparison of $X \rightarrow b\bar{b}$ signal discovery significance as a function of the SM background efficiency across different algorithms. Bottom: Stacked distributions of various SM background processes and the $X \rightarrow b\bar{b}$ signal process, selected using the Sophon $X \rightarrow b\bar{b}$ vs. QCD discriminant at $\epsilon_B = 10^{-4}$.

from QCD jets, employing similar training settings. To address the limited number of signal $X \rightarrow b\bar{b}$ jets in the JETCLASS-II dataset, which might constrain the models' performance, we generate an extended set of $X \rightarrow b\bar{b}$ jets with the same configuration. This enlargement creates a dedicated category containing 16 M jets, reaching a comparable size to the QCD jet category. The results indicate that the Sophon model's performance in $X \rightarrow b\bar{b}$ tagging surpasses that of the dedicated ParT and ParticleNet trainings, demonstrating its superior tagging capabilities.

In addition, to investigate whether the Sophon model's superior $X \rightarrow b\bar{b}$ vs. QCD jet tagging performance stems from the large-scale classification task, we perform an ablation study by training a variant of the model on only 42 classes (denoted as Sophon*), including all two-prong resonant-jet classes and the QCD classes. Consequently, the training dataset size is reduced to 48 M jets, compared to the 168 M jets in the full JETCLASS-II dataset. The results indicate that the Sophon model trained on 188 classes with the comprehensive dataset achieves significantly better tagging performance. This demonstrates the importance of pre-training on a large and diverse dataset, even when the additional data samples are not directly relevant to the specific task.

13.3.3 Generalization capability

The generalization capability of the Sophon model is evaluated on the novel $X \rightarrow bs$ tagging task. For $X \rightarrow bs$ tagging, transfer learning is applied to the latent features of the Sophon model, with a latent feature dimension of $N_{\text{latent}} = 128$. A two-layer MLP with (512, 2) nodes is used, where the parameters of the first linear layer are initialized from the corresponding layer in the Sophon model to facilitate learning. The MLP has two output nodes, designed to classify $X \rightarrow bs$ jets and QCD jets.

The training sample is specifically generated using a similar mass-decorrelation approach as in the JETCLASS-II dataset. $X \rightarrow bs$ jets are produced with variable m_X while maintaining the same kinematics as jets in JETCLASS-II, and the same mass-decorrelation technique is applied during training. Notably, the transfer learning process is significantly simpler and faster than the original Sophon model training. It requires only a small fraction (1/320) of the total Sophon training dataset, with a computational cost (in terms of floating-point operations per second) of just 1/1 000 000 of the original training.

The top panel of Fig. 13.2 presents the discovery significance of the $X \rightarrow bs$ signal jets with $m_X = 200$ GeV as a function of the background efficiency. The performance is compared across various algorithms. Similar to the $X \rightarrow b\bar{b}$ tagging study, the ParticleNet and ParT classifiers are trained on a dedicated $X \rightarrow bs$ vs. QCD tagging task using an expanded $X \rightarrow bs$ jet dataset (16 M jets) to represent the best achievable capabilities with current technology at the LHC. The bottom panel again displays the stacked distribution of all SM processes and the signal process, following a selection on the $X \rightarrow bs$ tagging discriminant at $\epsilon_B = 10^{-4}$.

The results demonstrate that the Sophon model, after transfer learning, surpasses the performance of the dedicated ParticleNet and ParT models. As a validation check, we examine whether the high performance in $X \rightarrow bs$ vs. QCD jet tagging is due to the transfer of knowledge from the latent space rather than leveraging the tagging ability of existing classification nodes. To this end, we identify the output node for $X \rightarrow bq$ jets that most closely resembles $X \rightarrow bs$ jets and evaluate its performance using Sophon's $X \rightarrow bq$ vs. QCD jet tagging discriminant. The latter performs significantly worse, confirming that the transfer learning process is essential for achieving high performance.

13.4 Implications to LHC resonance search

The outstanding tagging and generalization capabilities enable Sophon to accelerate the resonance search programmes at the LHC. To illustrate this point, this section discusses three

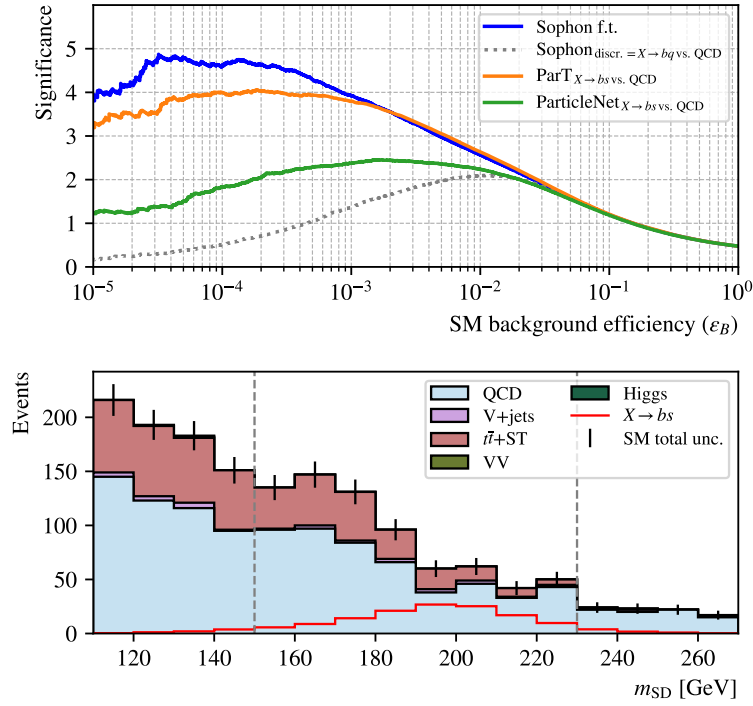


Figure 13.3 Top: Comparison of $X \rightarrow bs$ signal discovery significance as a function of the SM background efficiency across different algorithms. Bottom: Stacked distributions of various SM background processes and the $X \rightarrow bs$ signal process, selected using the fine-tuned Sophon $X \rightarrow bs$ vs. QCD discriminant at $\epsilon_B = 10^{-4}$.

application scenarios. The first two (discussed in Sections 13.4.1 and 13.4.2) focus on global resonance searches in single-jet and dijet final states, while the third (Section 13.4.3) explores anomaly detection through the use of weakly-supervised classifiers to broadly search for anomalous signals exhibiting a dijet resonance structure.

13.4.1 Generic single-jet resonance search

The all-inclusive classification nodes of the Sophon model introduce a novel paradigm of physics search: by constructing custom discriminants from the 188 output scores in specific combinations, where the numerator and denominator are defined as shown in Fig. 13.1 (b), the model can be tailored to focus on specific signal channels. This approach enhances the sensitivity to signals in the targeted channel, pushing the detection capabilities to the forefront.

For illustration, we examine the case where the final-state particles from the resonance decay can be clustered into a single large jet. The following discriminant is proposed to reject QCD jets while enhancing the purity of specific signal processes,

$$\text{Sophon } A \text{ vs. QCD disc.} = \frac{p(A)}{p(A) + \sum_{l=1}^{27} p(\text{QCD}_l)}. \quad (13.3)$$

This study considers three choices for the signature A for illustration:

$$\begin{aligned}
 p(A_1) &= p(X \rightarrow b\bar{b}), \\
 p(A_2) &= p(X \rightarrow cs), \\
 p(A_3) &= p(X \rightarrow bq_{\text{all}}q_{\text{all}}) := \sum_{\substack{\text{sig} \in \{ccb, ssb, \\ qqb, bcs, bcq, bsq\}}} p(X \rightarrow \text{sig}),
 \end{aligned} \tag{13.4}$$

where the $p(A_3)$ term in the discriminant is designed to capture all existing three-prong signatures composed of three quarks, including exactly one b -quark. By applying selections on jets based on three discriminants with progressively tighter thresholds, Fig. 13.4 illustrates the changes in the m_{SD} spectrum of 40 fb^{-1} of SM events.

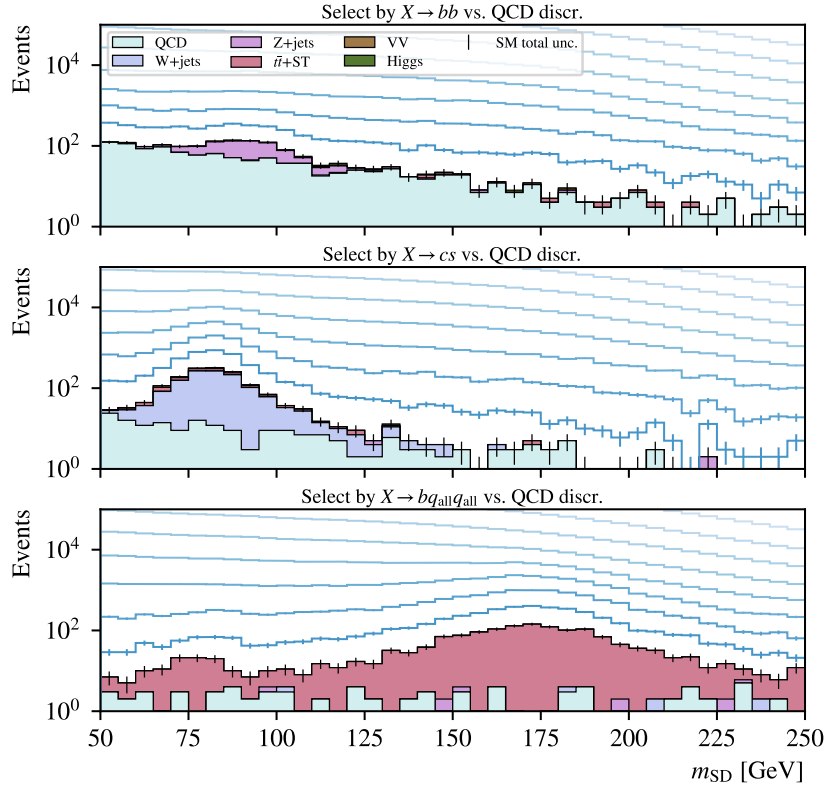


Figure 13.4 Illustration of the application of various Sophon tagging discriminants at different selection efficiencies ($10^{-4} < \epsilon < 10^0$) to unveil the hadronic peak structures of the Z , W bosons, and the t quark.

At the tightest working point, corresponding to a selection efficiency of $\epsilon = 10^{-4}$, the stacked histograms from major SM processes are displayed. Interestingly, the distributions reveal distinct hadronic resonant signatures from the W and Z bosons, as well as the t quark. These signatures correspond to final states identified by the Sophon model, including $W \rightarrow cs$

decays, $Z \rightarrow b\bar{b}$ decays, and $t \rightarrow bW \rightarrow bqq'$ decays.

This example demonstrates a universal approach for using the Sophon model to establish a resonance search programme. Even when the target resonance properties are unknown, it is feasible to conduct a comprehensive study with minimal effort, exploring almost *all* final-state signatures to identify potential peak structures.

13.4.2 Generic dijet resonance search

When the resonance mass extends to hundreds of GeV or even TeV scales, its final-state particles are unlikely to be Lorentz-boosted into a single jet. Instead, it becomes feasible to analyze multiple objects, including combinations of large- R jets, small- R jets, leptons, or photons produced in the event.

In this example, we consider a BSM process with a triboson final state, initiated by $W' \rightarrow W\phi \rightarrow WWW$, where $m_{W'} = 3$ TeV, $m_\phi = 400$ GeV, and all W bosons decaying hadronically. The physics model is described in Refs.^[198, 199] and was explored in an early model-agnostic study^[200]. The simulation of this process follows the same workflow as the earlier examples.

Events are required to pass the same large- R jet-based trigger and contain at least two $R = 0.8$ jets with $p_T > 250$ GeV and $|\eta| < 2.5$. The W' resonance is reconstructed from the two large- R jets with the highest p_T . The dijet invariant mass m_{jj} is used to reconstruct the W' mass. The SR is defined as $m_{jj} \in (2500, 3100)$ GeV.

A discriminant is constructed using the Sophon model's output scores to select the expected signatures for both jets. Ideally, one jet is expected to exhibit a two-prong structure, while the other jet should exhibit a four-prong structure. However, given the large m_ϕ , it is possible for the jet to only capture three quarks within its cone. Therefore, the discriminant selects jets with signatures corresponding to two prongs (initiated by a W boson) or three and four prongs (initiated by WW). The event-selection discriminant is then optimized as the sum of the tagging discriminants for the two jets, expressed as

$$\text{event disc.} = \sum_{\text{jet}=1,2} \frac{p(A, \text{jet})}{p(A, \text{jet}) + \sum_{l=1}^{27} p(\text{QCD}_l, \text{jet})}, \quad (13.5)$$

where $p(A)$ is defined as

$$p(A) := 0.3 p(W^{(2)}) + 0.1 p(WW^{(4)}) + 0.6 p(WW^{(3)}), \quad (13.6)$$

and

$$\begin{aligned}
 p(W^{(2)}) &:= p(X \rightarrow cs) + p(X \rightarrow q\bar{q}), \\
 p(WW^{(4)}) &:= p(X \rightarrow ccss) + p(X \rightarrow qqcs) + p(X \rightarrow qqqq), \\
 p(WW^{(3)}) &:= \sum_{\text{sig} \in \{ccs, ccq, ssc, \\
 &\quad ssq, qqc, qqs, qq\}} p(X \rightarrow \text{sig}).
 \end{aligned} \tag{13.7}$$

As observed, the nodes corresponding to three-prong signatures are assigned a larger weight during optimization. This is partly because three-prong signatures are more likely to appear than the two- or four-prong cases in the actual signal events.

To demonstrate the discovery potential, the significance Z , defined in Eq. (13.2), is evaluated using events in the SR and presented as a function of the SM background efficiency. Compared to the single-jet scenarios shown in Figs. 13.2 and 13.3, the significance continues to increase as the selection becomes tighter (i.e. as the SM background efficiency decreases), illustrating even stronger discrimination power at the most extreme selection working points. This improvement is partly due to applying selection criteria on both large- R jets using the Sophon model. At selection points with SM background efficiencies of $\epsilon_B = 10^{-4}$ and $\epsilon_B = 10^{-5}$, the distributions of the signal and SM processes are shown in Fig. 13.5. Notably, all SM processes are simulated to match the expected event counts for 40 fb^{-1} of data. At $\epsilon_B = 10^{-5}$, it is evident that hardly any background events survive, while the signal events remain prominent. This leads to a high discovery sensitivity.

It is intriguing to note that the trend observed in Fig. 13.5 suggests a promising scenario. Specifically, in LHC searches, where hadronic backgrounds are typically overwhelming, the power of the Sophon model can be leveraged to suppress QCD backgrounds to nearly zero while achieving optimal sensitivity. This represents a novel scheme that has been largely unseen in previous LHC analyses. Interestingly, such a phenomenon was more common in earlier collider experiments. For example, in the 1970s-80s, the discovery of the gluon at the PETRA collider was based on *a single signal event*, where the statistical principle relied on the extremely low likelihood of the background producing a “three-jet process”. Therefore, a key insight from this phenomenological study of the Sophon model at the LHC is that, even data collected at the LHC have immense complexity and a huge volume, particularly in the region of the most common fully-hadronic final states, it remains possible to recreate the aforementioned experimental discovery scenarios, when we exploit advanced background-suppressing AI tools, such of the Sophon model, or the GloParT developed within the CMS experiment.

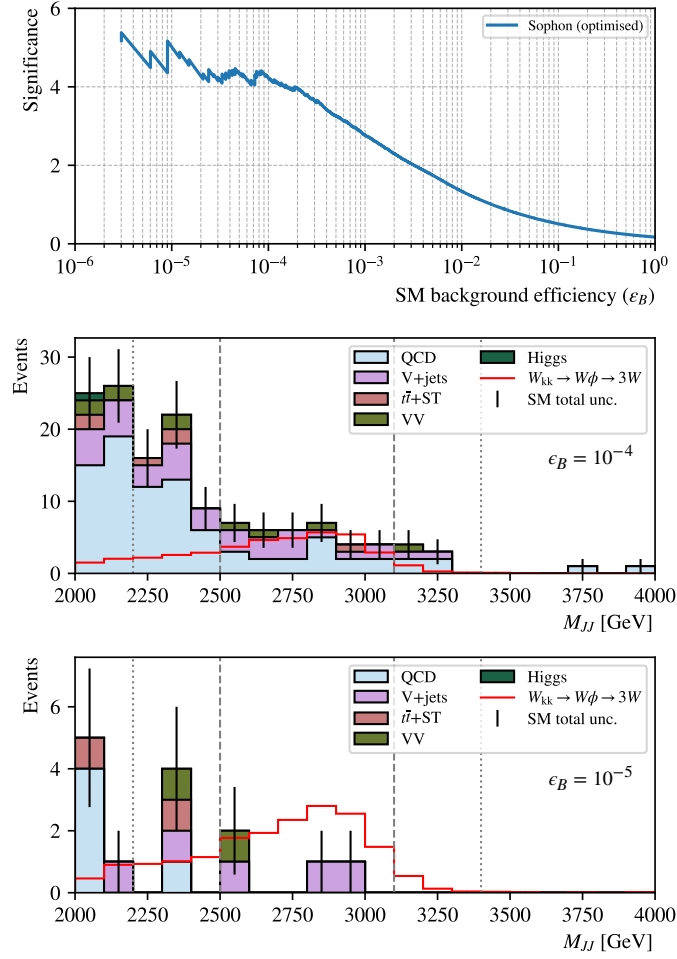


Figure 13.5 Top: Discovery significance for the $W' \rightarrow W\phi \rightarrow WWW$ signal as a function of the SM background efficiency, utilizing an optimized Sophon tagging discriminant. Middle and bottom: Stacked distributions of various SM background processes and the signal process, selected using the discriminant at $\epsilon_B = 10^{-4}$ and 10^{-5} , respectively.

13.4.3 Dijet anomaly detection

As a final experiment, we investigate the potential of employing the Sophon model's fine-tuning for anomaly detection applications. The discussion begins with an introduction to the concept of anomaly detection using weak supervision, followed by an explanation of why combining Sophon's fine-tuning method with this approach can substantially enhance sensitivity. An experiment is then followed to support our statement.

Introduction to anomaly detection via weakly supervised approaches

Anomaly detection using weak supervision originated from the Classification Without Labels (CWoLa) method^[200-202]. This method assumes the search for resonance with unknown properties but relies on knowledge of its mass, which can be reconstructed in a dijet final state within a specified mass window (which is considered the SR). A weak classifier is then trained to distinguish events within the mass window from those in the sideband (SB) region. This enables the differentiation of the signal process from the background. Figure 13.6 visualizes this approach. After applying a tight selection, the presence of new physics can be revealed by identifying a bump structure over a smoothly falling background, as shown in Fig. 13.7.

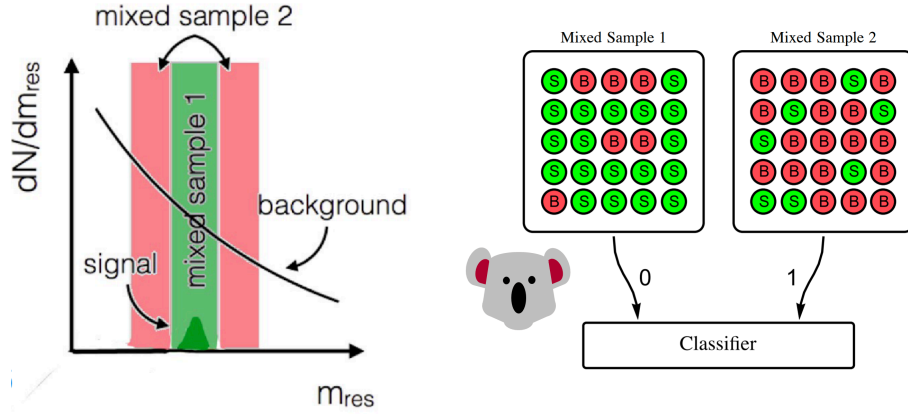


Figure 13.6 Illustration of the anomaly detection method employed in the weakly supervised approach^[201].

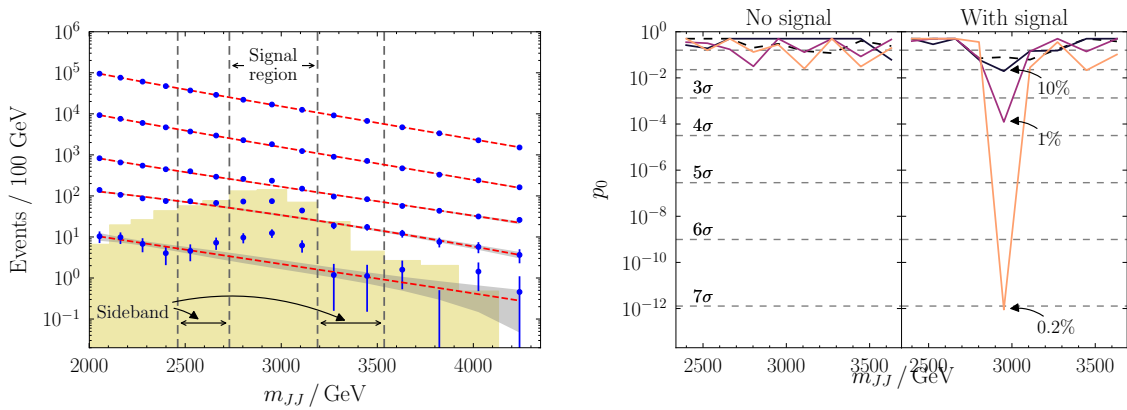


Figure 13.7 Results from Ref.^[200] showcasing the capability to search for $W' \rightarrow W\phi \rightarrow WWW$ within the m_{jj} spectrum using the weakly supervised anomaly detection technique.

This early approach requires the classifier to be uncorrelated with mass information, relying solely on features such as jet substructure to identify the signal. Otherwise, it risks distort-

ing the background’s mass distribution. Subsequent methods have introduced various improvements (see Refs.^[203-207] for a non-exhaustive list), primarily focusing on better defining the background sample for weak classification (e.g. the “mixed sample 2” illustrated in Fig. 13.6). The most idealized approximation assumes that the background sample can be directly taken from the SR.

As a result, Ref.^[205] proposed the idealized case, where the classifier is trained to distinguish data in the SR from the SR background. This scenario is referred to as the idealized anomaly detection (IAD) limit. The IAD limit establishes a baseline for the maximum achievable search sensitivity in the data, independent of the specific anomaly detection method employed.

The background of the anomaly detection described above implies that when the signal injection is minimal, this method becomes insensitive. In such cases, the two classes in the IAD approach become nearly indistinguishable, resulting in a classifier that consistently outputs 50–50% probability split. To improve sensitivity, additional prior knowledge is required to better differentiate between signal and background, so that even a small injection of signal events can be effectively identified by the model.

In this sense, the Sophon model is particularly well-suited for this purpose due to its ability to recognize hundreds of jet topologies with high sensitivity. This capability stems from its pre-training as a classifier, specifically optimized to distinguish *any* two types of jets. Conceptually, the knowledge embedded in the Sophon model can guide the training of weakly-supervised classifiers. From a technical aspect, this can be achieved through transfer learning—a simpler type in the fine-tuning paradigm—where the 128 neuron values from Sophon’s hidden layers are used as inputs for training the classifier.

Experiments

The experiment is conducted as follows. We adopt the same signal and dijet reconstruction strategy described in Section 13.4.2. Notably, the signal process and resonance mass values are identical to those used in the early CWoLa experiment^[200]. To perform the IAD classifier training, the SR is defined as $m_{jj} \in (2500, 3100)$ GeV, the same as in Section 13.4.2, and an additional SB region is defined as $m_{jj} \in (2200, 2500) \cup (3100, 3400)$ GeV. Two IAD models are compared: one leveraging Sophon’s transfer learning approach and the other utilizing high-level jet inputs designed to replicate the results in Ref.^[200].

- To train the IAD classifier using Sophon’s transfer learning approach, we employ a 3-

layer MLP with (512, 64, 2) nodes and ReLU activation functions to construct the classifier. The input comprises two latent feature vectors extracted from the Sophon model, each with a dimensionality of 128, for the two jets. The processing begins with the two input vectors being passed through the first and second layers of the MLP, respectively. The resulting outputs are subsequently summed and forwarded through the third layer. An ensemble of 100 networks is trained, and the average score corresponding to the data node is used as the discriminant.

- To train the IAD classifier using high-level jet observables as input, we consider, for each jet, the jet m_{SD} , the number of constituents, the N -subjettiness variables^[69] τ_1 , τ_2 , τ_3 , and τ_4 . Each variable is logarithmized and standardized to fall within the range $(-2, 2)$. The input comprises two six-dimensional vectors, one for each jet. A 4-layer MLP with (512, 128, 128, 2) nodes and ReLU activation functions is utilized. Similar to the previous case, the input vectors are independently passed through the first and second layers of the MLP. The outputs are then summed and processed through the third and fourth layers.

In both cases, an ensemble of 100 networks is trained, and their averaged score corresponding to the data node is used as the discriminant. This is a common approach to address the instability in the performance of the weakly-supervised classifier.

With the trained IAD classifier discriminant, we evaluate the maximum significance improvement, defined as $\max\{Z|_{\epsilon_B > 10^{-4}}\}/Z|_{\epsilon_B=1}$, under a selection imposed on the discriminant, as a function of the injected signal yield, illustrated in Fig. 13.8. Our results demonstrate that Sophon’s transfer learning approach, combined with the IAD method, not only enhances the maximum significance improvement but also increases sensitivity to the signal at low injection levels. Specifically, the initial significance at which the method becomes sensitive to the presence of a signal is approximately $0.6\text{--}1.0\sigma$.

This result indicates that leveraging the “knowledge” from a pre-trained Sophon model effectively resolves the challenge of achieving higher distinguishing capability from lower-level input features without compromising sensitivity to low signal injections^[208]. Furthermore, the curve corresponding to conventional high-level inputs shows that the sensitivity improvement in our results is approximately consistent with that reported in the original work^[200], where an initial sensitivity of 2σ was shown to improve to roughly 7σ .

Before concluding this section, it is worth including a comparison involving a signal-specific search method by incorporating an optimized Sophon discriminant dedicated to this

signal, which has been investigated in Section 13.4.2. As shown in Fig. 13.8, Sophon’s signal-targeted discriminant significantly accelerates the achievement of a 5σ discovery, requiring approximately 3.5 times fewer signal events than the IAD method based on Sophon’s transfer learning.

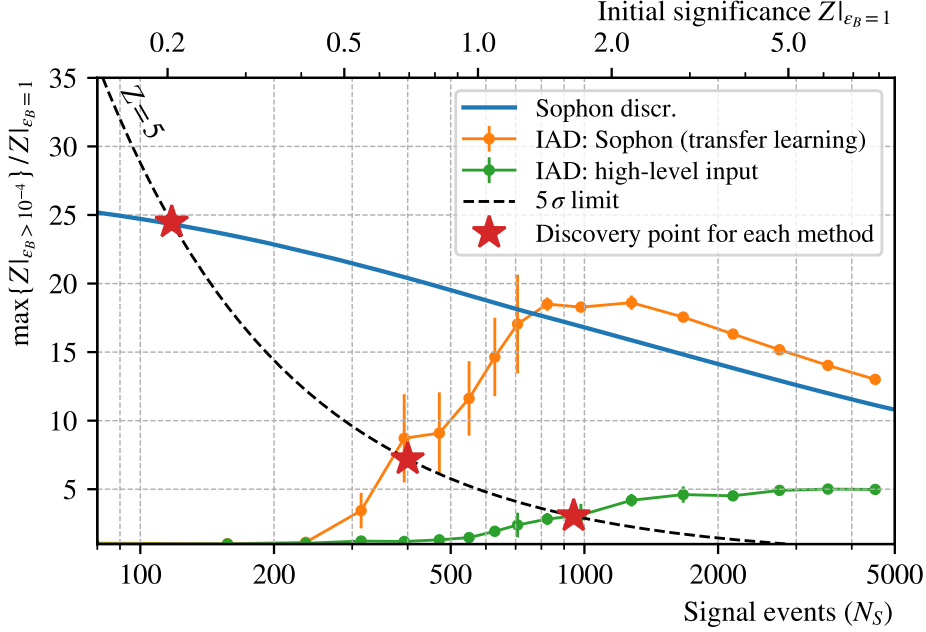


Figure 13.8 Comparison of the Sophon model’s effectiveness in model-specific searches and its model-agnostic capabilities, alongside the conventional method, for probing $W' \rightarrow W\phi \rightarrow WWWW$ within the m_{jj} spectrum.

The advancement shown in this result can be attributed to the inherent characteristics of signal-specific searches, where the discriminant is explicitly optimized for a certain signal hypothesis. Thus, its performance no longer depends on the number of signal events presented in the data. This tailored approach highlights the advantage of designing certain discriminants (can be hundreds of them when leveraging the 188 outputs of the Sophon model) in rapidly reaching discovery thresholds under a wide range of well-defined signal conditions.

13.5 Discussion and outlook

In this section, we discuss the insights gained from the experiments with the Sophon model and JETCLASS-II, as well as the potential implications of GloParT for future CMS analyses.

In the Sophon model experiment, we primarily demonstrated the utility of transfer learning. While its performance may not reach a full-model fine-tuning, its advantage lies in the

opportunity to store hidden-layer neuron representations in the dataset centrally produced in the experimental data workflow, allowing hundreds of downstream analyses directly using them as input to train analysis-level DNNs. This process is essentially similar to traditional DNN training but substitutes the conventional jet input with these hidden-layer “black-box” variables. Yet, this simple change enables rapid and convenient use of the “knowledge” of the pre-trained model to improve the performance in downstream tasks.

In our example, the philosophy of “signature-oriented pre-training” is illustrated with the most straightforward application, where pre-training is performed in a classifier framework utilizing only jet labels as the signatures. As discussed, future research can extend this approach to incorporate a broader range of signatures, including the abundant kinematic properties and PID information of generator-level particles, which can be either the initial resonance or stable particles that initiate the jets. Such studies, combined with novel training techniques, are expected to further enhance the expressiveness and generalization capabilities of the method. It is important to note that these generator-level features have been stored in our JETCLASS-II dataset. This makes it a valuable resource for testing and improving future training strategies.

We would also like to emphasize the significance of JETCLASS-II for subsequent training of foundation models for the broader high-energy physics community. Advancements in large language models in recent years have undoubtedly propelled various fields to leverage their abundant datasets to develop foundation models capable of addressing diverse tasks. However, compared to other fields, where datasets typically cover the scope of all its domain-specific tasks, the unique nature of high-energy physics lies in its goal of potential discoveries of unknown particles. Consequently, the relevant phase space might not be adequately highlighted in the real data collected at the experiment. In contrast to the data or simulated events for SM processes, a simulated dataset encompassing as much of the BSM phase space as possible will likely be of greater help for developing multi-purpose base models. This makes JETCLASS-II particularly significant when compared to its predecessor JETCLASS and other datasets. We believe that JETCLASS-II will make substantial contributions to advancing high-energy physics models in the future.

Finally, the success of the Sophon model implies various promising yet unexplored opportunities for employing the GloParT model in the CMS experiment. This impact can be understood in three key aspects.

1. For the planned analyses to be delivered in Run 3, greater sensitivity improvements are anticipated based on the tagging discriminant of GloParT. Notable examples include

Higgs boson property measurements in several channels, such as $H \rightarrow b\bar{b}$, $c\bar{c}$, WW^* , and $\tau\tau$, as well as studies of Higgs boson self-coupling through di-Higgs searches. Additionally, the model is expected to enhance precision studies in SM physics that require advanced tagging of boosted W , Z , and top jets, among others.

2. The direct tagging capabilities of the GloParT model can be extended to a broader range of final states, thereby expanding the scope of boosted-topology searches. Fine-tuning the model on customized phase space enables generalization to even broader tagging tasks. However, a significant challenge lies in calibrating the dedicated signal processes using the corresponding discriminants, primarily due to difficulties in finding appropriate proxy jets in data. Nevertheless, it is certain that innovative calibration approaches will play a vital role in advancing GloParT applications in previously unexplored signatures.
3. GloParT establishes new paradigms in utilizing boosted jets as a tool to discover new physics. Firstly, the rich final states provided by GloParT offer a straightforward and highly effective approach to searching for new physics. This involves constructing hundreds of discriminants and applying stringent selection criteria to search for potential bumps in the mass spectrum. Invariant masses can be defined from arbitrary combinations of objects, including large- R jets, small- R jets, reconstructed leptons, photons, and p_T^{miss} . This method has the capability to push the sensitivity of final-state-oriented searches to their limits while circumventing the need for signal calibration. Secondly, based on the Sophon anomaly detection experiment, the extensive “knowledge” encoded during the pre-training of the model is anticipated to universally enhance the sensitivity of anomaly detection. That is to say, beyond the weakly supervised approach discussed in this dissertation, this method also provides a new vision for the autoencoder-based approach (e.g. as described in Ref.^[209], with a very recent implementation by the ATLAS experiment reported in Ref.^[210]). Conducting these searches at the LHC under the base model fine-tuning scheme is expected to push various underexplored new physics scenarios to unprecedented levels of sensitivity.

Chapter 14

Outlook

Collider physics plays a crucial role in humanity’s efforts to explore the smallest building blocks of nature and unravel the mysteries of the universe. It is reasonable to believe that even a century from now, particle colliders will continue to be essential tools for investigating fundamental particles and their interactions. These machines allow us to create and study extremely high-energy conditions that are otherwise unattainable. However, the technologies that support collider experiments have evolved significantly over time. Compared to the earliest collider experiments more than 50 years ago, today’s LHC can quickly convert detector signals into digital data and transmit them to sites distributed across the globe for permanent storage. This remarkable capability might have been beyond the imagination of physicists half a century ago, and it is a direct outcome of the technological progress brought about by the third industrial revolution—marked by digitalization and information technology.

We are now in the middle of what many consider the next industrial revolution, driven by deep learning and AI breakthroughs, fuelled by large datasets, powerful models, and substantial computing capabilities. The future shape of collider experiments, empowered by this technological revolution, may be unimaginable to us now, but the present researchers are determining its trajectory. It is a unique opportunity for the author to work at the frontier of collider physics and deep learning, collaborating with scientists to apply state-of-the-art deep learning methods to our experiments. This approach enables physics discoveries that were previously considered out of reach and opens new directions for searches at the LHC.

This dissertation focuses on the study of large- R jets. Within the modular analysis frameworks used by the current ATLAS and CMS experiments, large- R jets are among the most complex objects to analyze. Their complexity makes them a powerful test case for advanced deep learning techniques. Theoretically, large- R jets are valuable tools for studying highly Lorentz-boosted event topologies and high-energy particle resonances. In many analyses, identifying events with signal-like large- R jets is a critical first step, corresponding to the task of large- R jet tagging. This dissertation aims to cover the latest advances in modern deep-learning methods applied to large- R jet tagging.

The dissertation discusses four aspects centred around this technology: the latest developments on DNN algorithms, the new calibration technique critical for reliable physics analyses, a direct application in measuring the Higgs boson properties related to the charm sector, and prospects on how jet tagging may revolutionize its form in the near future, particularly with the “pre-training and fine-tuning” paradigm. All these aspects showcase significant progress in their respective domains. Given the rapid development of deep learning, there are surely new possibilities ahead. While each part of the dissertation includes an outlook for each field, a few selected aspects are highlighted before concluding the dissertation.

On the algorithm side, although Transformer models have been widely adopted across various fields, they have a known deficiency with quadratic computing complexity. As future models handle more complex collider data, the number of tokens will increase, which calls for the development of models with improved computing efficiency. Additionally, each task has its own inductive bias, and this paper focuses on the impact of Lorentz symmetry preservation, an inductive bias that has previously been less systematically considered. Just as permutational symmetry preservation is receiving increasing attention, Lorentz symmetry preservation may also become a standard feature in future network designs. Given the high complexity of current pairwise feature-based methods, finding better model designs that incorporate this symmetry is a viable research direction. Finally, as network design inputs become increasingly low-level, e.g. reaching the level of energy deposits or track hits, it is worth considering how to embed Lorentz symmetry preservation in the lowest-level inputs.

For calibration, we present a new approach that pushes the proxy-based method forward, specifically using a BDT approach to select proxies. However, it is worth considering whether more efficient calibration methods exist. For example, recent methods propose using DNNs to construct full phase space reweighting or morphing, allowing neural networks to learn the corresponding scale factors autonomously. In the future, as jet taggers are applied in more scenarios and proxies become more challenging to identify from data, achieving full phase space calibration will be a promising direction.

For physics applications, this dissertation demonstrates that the integration of our global tagging techniques with the LHC physics programme allows for the exploration of a much broader boosted phase space. This enables searches for new physics through bump hunting without requiring calibration, allowing rapid progress in tightening exclusion limits for various theoretical models. When combined with advanced calibration methods, these techniques can achieve more precise measurements and impose stronger constraints on boosted final states, as

illustrated by the progress in the $VHcc$ search presented in this work.

While there is often an unspoken sentiment among analysts that physics research at the LHC has become less engaging, as a similar experiment is often repeated yearly, yielding improvements under a predictable range, deep learning technologies will bring about one revolution after another, allowing for a large gain previously unimaginable and bringing new searching capabilities at the LHC. Just as the current scale of the LHC was unimaginable by scientists 50 years ago, we are expected to witness and shape the changes in data processing and analysis paradigms. These advancements are likely to redefine what is possible in particle physics at the LHC.

Appendix A

Additional sfBDT Calibration Results

In addition to the calibration results of the ParticleNet-MD $X \rightarrow b\bar{b}$ and $X \rightarrow c\bar{c}$ tagging discriminant discussed in Section 6.5, Tables 6.8–6.9 further summarizes the results for calibrating DeepDoubleX, DeepAK8-MD, double-b $X \rightarrow b\bar{b}$ discriminants, and DeepDoubleX and DeepAK8-MD $X \rightarrow c\bar{c}$ discriminants.

Table A.1 Summary of the DeepDoubleX $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.2739, 1.0)	$1.015^{+0.084}_{-0.083}$	$1.086^{+0.099}_{-0.097}$	$1.056^{+0.154}_{-0.149}$
MP: (0.118, 1.0)	$1.035^{+0.096}_{-0.095}$	$1.097^{+0.111}_{-0.105}$	$1.083^{+0.183}_{-0.178}$
LP: (0.0256, 1.0)	$1.070^{+0.118}_{-0.113}$	$1.111^{+0.137}_{-0.128}$	$1.177^{+0.272}_{-0.261}$
2016 post-VFP			
HP: (0.2786, 1.0)	$1.028^{+0.063}_{-0.060}$	$1.039^{+0.084}_{-0.077}$	$1.135^{+0.143}_{-0.138}$
MP: (0.1213, 1.0)	$1.025^{+0.061}_{-0.059}$	$1.053^{+0.080}_{-0.077}$	$1.164^{+0.168}_{-0.158}$
LP: (0.027, 1.0)	$1.006^{+0.060}_{-0.057}$	$1.114^{+0.106}_{-0.091}$	$1.168^{+0.184}_{-0.163}$
2017			
HP: (0.3154, 1.0)	$0.896^{+0.059}_{-0.059}$	$0.910^{+0.048}_{-0.049}$	$0.926^{+0.034}_{-0.035}$
MP: (0.1566, 1.0)	$0.913^{+0.056}_{-0.058}$	$0.926^{+0.051}_{-0.054}$	$0.906^{+0.035}_{-0.041}$
LP: (0.0404, 1.0)	$0.949^{+0.043}_{-0.047}$	$0.947^{+0.053}_{-0.057}$	$0.935^{+0.033}_{-0.043}$
2018			
HP: (0.314, 1.0)	$0.878^{+0.049}_{-0.045}$	$0.967^{+0.052}_{-0.051}$	$0.928^{+0.039}_{-0.042}$
MP: (0.1566, 1.0)	$0.869^{+0.050}_{-0.051}$	$0.969^{+0.058}_{-0.059}$	$0.939^{+0.034}_{-0.040}$
LP: (0.0399, 1.0)	$0.889^{+0.055}_{-0.064}$	$0.975^{+0.060}_{-0.065}$	$0.961^{+0.029}_{-0.035}$

Table A.2 Summary of the DeepAK8-MD $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.9521, 1.0)	$0.933^{+0.078}_{-0.078}$	$0.946^{+0.079}_{-0.080}$	$0.977^{+0.116}_{-0.113}$
MP: (0.894, 1.0)	$0.990^{+0.084}_{-0.084}$	$1.002^{+0.088}_{-0.088}$	$0.965^{+0.118}_{-0.115}$
LP: (0.7021, 1.0)	$1.013^{+0.102}_{-0.099}$	$1.062^{+0.109}_{-0.103}$	$1.045^{+0.163}_{-0.149}$
2016 post-VFP			
HP: (0.9507, 1.0)	$0.858^{+0.062}_{-0.054}$	$0.895^{+0.063}_{-0.059}$	$0.923^{+0.085}_{-0.083}$
MP: (0.8877, 1.0)	$0.929^{+0.051}_{-0.056}$	$0.971^{+0.059}_{-0.060}$	$1.028^{+0.105}_{-0.099}$
LP: (0.689, 1.0)	$0.979^{+0.048}_{-0.053}$	$1.024^{+0.069}_{-0.068}$	$1.106^{+0.131}_{-0.120}$
2017			
HP: (0.9546, 1.0)	$0.798^{+0.045}_{-0.044}$	$0.837^{+0.043}_{-0.038}$	$0.833^{+0.040}_{-0.041}$
MP: (0.9033, 1.0)	$0.883^{+0.047}_{-0.049}$	$0.896^{+0.048}_{-0.044}$	$0.874^{+0.041}_{-0.049}$
LP: (0.7422, 1.0)	$0.935^{+0.039}_{-0.046}$	$0.955^{+0.034}_{-0.040}$	$0.926^{+0.031}_{-0.044}$
2018			
HP: (0.9551, 1.0)	$0.824^{+0.042}_{-0.048}$	$0.854^{+0.045}_{-0.044}$	$0.877^{+0.048}_{-0.048}$
MP: (0.9058, 1.0)	$0.860^{+0.037}_{-0.043}$	$0.903^{+0.044}_{-0.043}$	$0.915^{+0.046}_{-0.049}$
LP: (0.7432, 1.0)	$0.895^{+0.041}_{-0.048}$	$0.951^{+0.042}_{-0.047}$	$0.958^{+0.029}_{-0.033}$

Table A.3 Summary of the double-b $X \rightarrow b\bar{b}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.8926, 1.0)	$1.030^{+0.054}_{-0.055}$	$1.086^{+0.083}_{-0.082}$	$1.179^{+0.140}_{-0.139}$
MP: (0.7559, 1.0)	$1.045^{+0.061}_{-0.058}$	$1.122^{+0.094}_{-0.085}$	$1.157^{+0.151}_{-0.150}$
LP: (0.3104, 1.0)	$1.099^{+0.091}_{-0.071}$	$1.097^{+0.099}_{-0.083}$	$1.136^{+0.153}_{-0.146}$
2016 post-VFP			
HP: (0.8955, 1.0)	$0.996^{+0.058}_{-0.063}$	$1.057^{+0.048}_{-0.051}$	$1.170^{+0.126}_{-0.122}$
MP: (0.7646, 1.0)	$1.047^{+0.044}_{-0.042}$	$1.080^{+0.050}_{-0.045}$	$1.173^{+0.154}_{-0.133}$
LP: (0.3279, 1.0)	$1.040^{+0.045}_{-0.039}$	$1.085^{+0.060}_{-0.048}$	$1.181^{+0.141}_{-0.119}$
2017			
HP: (0.9185, 1.0)	$0.910^{+0.056}_{-0.057}$	$0.874^{+0.069}_{-0.069}$	$0.939^{+0.041}_{-0.049}$
MP: (0.8237, 1.0)	$0.936^{+0.051}_{-0.055}$	$0.926^{+0.057}_{-0.061}$	$0.953^{+0.030}_{-0.038}$
LP: (0.5068, 1.0)	$0.989^{+0.043}_{-0.045}$	$0.987^{+0.049}_{-0.051}$	$0.994^{+0.029}_{-0.033}$
2018			
HP: (0.9175, 1.0)	$0.915^{+0.044}_{-0.045}$	$0.956^{+0.044}_{-0.046}$	$0.927^{+0.036}_{-0.038}$
MP: (0.8193, 1.0)	$0.915^{+0.033}_{-0.041}$	$0.973^{+0.032}_{-0.037}$	$0.963^{+0.029}_{-0.033}$
LP: (0.4939, 1.0)	$0.976^{+0.032}_{-0.035}$	$1.008^{+0.033}_{-0.036}$	$1.000^{+0.022}_{-0.023}$

Table A.4 Summary of the DeepDoubleX $X \rightarrow c\bar{c}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.2568, 1.0)	$1.203^{+0.243}_{-0.239}$	$1.050^{+0.186}_{-0.179}$	$0.797^{+0.234}_{-0.219}$
MP: (0.1043, 1.0)	$1.221^{+0.241}_{-0.236}$	$1.057^{+0.147}_{-0.136}$	$1.088^{+0.282}_{-0.272}$
LP: (0.0327, 1.0)	$1.293^{+0.289}_{-0.276}$	$0.996^{+0.144}_{-0.138}$	$1.021^{+0.243}_{-0.245}$
2016 post-VFP			
HP: (0.2505, 1.0)	$0.918^{+0.250}_{-0.214}$	$0.980^{+0.181}_{-0.164}$	$0.721^{+0.198}_{-0.175}$
MP: (0.1021, 1.0)	$1.054^{+0.179}_{-0.172}$	$0.951^{+0.179}_{-0.165}$	$1.022^{+0.203}_{-0.190}$
LP: (0.0321, 1.0)	$1.189^{+0.217}_{-0.200}$	$1.000^{+0.172}_{-0.161}$	$1.030^{+0.201}_{-0.191}$
2017			
HP: (0.3013, 1.0)	$0.959^{+0.265}_{-0.249}$	$1.059^{+0.263}_{-0.261}$	$0.885^{+0.134}_{-0.128}$
MP: (0.1361, 1.0)	$0.998^{+0.259}_{-0.241}$	$0.999^{+0.211}_{-0.198}$	$0.918^{+0.133}_{-0.129}$
LP: (0.0457, 1.0)	$0.972^{+0.192}_{-0.178}$	$1.055^{+0.205}_{-0.182}$	$1.002^{+0.156}_{-0.151}$
2018			
HP: (0.3013, 1.0)	$1.027^{+0.272}_{-0.317}$	$0.824^{+0.109}_{-0.106}$	$1.020^{+0.187}_{-0.186}$
MP: (0.1322, 1.0)	$1.001^{+0.127}_{-0.121}$	$0.828^{+0.068}_{-0.072}$	$0.981^{+0.150}_{-0.149}$
LP: (0.0425, 1.0)	$0.831^{+0.096}_{-0.096}$	$0.897^{+0.100}_{-0.109}$	$0.965^{+0.107}_{-0.108}$

 Table A.5 Summary of the DeepAK8-MD $X \rightarrow c\bar{c}$ discriminant SFs derived for the four data-taking year conditions, three exclusive working points (HP, MP, LP) and three p_T ranges separately.

	p_T range in GeV:		
	[450, 500)	[500, 600)	[600, ∞)
2016 pre-VFP			
HP: (0.9326, 1.0)	$0.969^{+0.171}_{-0.168}$	$1.040^{+0.124}_{-0.123}$	$1.105^{+0.276}_{-0.260}$
MP: (0.8496, 1.0)	$1.125^{+0.200}_{-0.198}$	$0.988^{+0.106}_{-0.100}$	$1.081^{+0.257}_{-0.245}$
LP: (0.6733, 1.0)	$1.181^{+0.217}_{-0.204}$	$1.030^{+0.105}_{-0.096}$	$1.022^{+0.221}_{-0.214}$
2016 post-VFP			
HP: (0.9277, 1.0)	$1.043^{+0.159}_{-0.151}$	$1.025^{+0.141}_{-0.140}$	$0.654^{+0.157}_{-0.149}$
MP: (0.8369, 1.0)	$1.177^{+0.173}_{-0.159}$	$1.032^{+0.141}_{-0.137}$	$0.875^{+0.152}_{-0.144}$
LP: (0.6458, 1.0)	$1.147^{+0.185}_{-0.173}$	$0.971^{+0.137}_{-0.144}$	$1.111^{+0.214}_{-0.200}$
2017			
HP: (0.939, 1.0)	$0.928^{+0.190}_{-0.180}$	$1.005^{+0.242}_{-0.239}$	$0.796^{+0.095}_{-0.096}$
MP: (0.8677, 1.0)	$0.936^{+0.164}_{-0.155}$	$0.939^{+0.167}_{-0.169}$	$0.906^{+0.113}_{-0.113}$
LP: (0.71, 1.0)	$0.946^{+0.154}_{-0.152}$	$0.973^{+0.133}_{-0.135}$	$0.954^{+0.127}_{-0.124}$
2018			
HP: (0.939, 1.0)	$0.818^{+0.181}_{-0.181}$	$0.776^{+0.074}_{-0.069}$	$0.909^{+0.128}_{-0.124}$
MP: (0.8662, 1.0)	$0.820^{+0.117}_{-0.096}$	$0.910^{+0.121}_{-0.120}$	$0.946^{+0.121}_{-0.119}$
LP: (0.7017, 1.0)	$0.841^{+0.102}_{-0.096}$	$0.950^{+0.109}_{-0.107}$	$0.920^{+0.082}_{-0.083}$

Appendix B

Statistical Framework of Classification and Regression

In this appendix, we derive the functions that minimize the multiclass cross-entropy loss and the regression loss trained with a log-cosh function. The theory provides a statistical meaning to the pre-trained model output scores when fully optimized.

B.1 Multiclass classification

Multiclass classifier neural networks trained with cross-entropy loss learn the ratio of prior PDFs of input classes in the limit of infinite data, and proof is shown here.

Suppose we find a space Ω to represent input data, and each input class owns a PDF denoted by $\rho_c(x)$. Note that the representation is abstract because the input features have variable lengths. We require

$$\sum_c \int_{\Omega} \rho_c(x) dx = 1, \quad (\text{B.1})$$

such that if the entire data has N entries (N is sufficiently large), the number of jets we have for class c , mapping to the point x_0 and its vicinity, will be $(N \cdot \rho_c(x_0) dx)$. We adopt here an unconventional approach that the $\rho_c(x)$ for each class c is not normalized, but the integral for $\int_{\Omega} \rho_c(x) dx$ for class c should be proportional to the number of jets in that class.

From the statistical view, suppose the network outputs the probability scores $\hat{y}_c(x)$, which are the softmax of logits. The sum of cross-entropy loss on all data is rewritten as

$$L_{\text{cls}} = N \cdot \sum_c \int_{\Omega} -\rho_c(x) \log \hat{y}_c(x) dx; \quad (\text{B.2})$$

or conventionally, it can be written as $L_{\text{cls}} = N \cdot \mathbb{E}_{x \sim \rho_c(x)} [-\log \hat{y}_c(x)]$. Notice that $\hat{y}_c(x)$ is constraint by $\sum_c \hat{y}_c(x) = 1$.

By minimizing L_{cls} , we minimize

$$l_{\text{cls}}(x) = - \sum_c \rho_c(x) \log \hat{y}_c(x) \quad (\text{B.3})$$

for every x , so we omit the label x . With the method of Lagrange multipliers, we construct

$$\mathcal{L}(\hat{y}_c, \lambda) = - \sum_c \rho_c \log \hat{y}_c - \lambda \left(\sum_c \hat{y}_c - 1 \right). \quad (\text{B.4})$$

Taking the derivative with respect to $\hat{y}_c$ and λ ,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \hat{y}_c} = -\frac{\rho_c}{\hat{y}_c} - \lambda = 0, & \text{for every } c, \\ \frac{\partial \mathcal{L}}{\partial \lambda} = -\sum_c \hat{y}_c + 1 = 0. \end{cases} \quad (\text{B.5})$$

We obtain

$$\hat{y}_c = \frac{\rho_c}{\sum_{c'} \rho_{c'}}. \quad (\text{B.6})$$

To see that the extreme point is a minimum, consider a fluctuation on $\hat{y}_c \rightarrow \hat{y}_c + \delta \hat{y}_c$ with the constraint $\delta \hat{y}_c = 0$, then calculate the fluctuation of l_{cls} ,

$$\begin{aligned} \delta l_{\text{cls}} &= - \sum_c \rho_c \left(\frac{\delta \hat{y}_c}{\hat{y}_c} - \frac{\delta \hat{y}_c^2}{2 \hat{y}_c^2} + o(\delta \hat{y}_c^2) \right) \\ &= \left(\sum_c \rho_c \right)^2 \cdot \sum_c \frac{\delta \hat{y}_c^2}{2 \rho_c} + o(\delta \hat{y}_c^2) \\ &> 0. \end{aligned} \quad (\text{B.7})$$

As a side note, the derivation is more intuitive for binary classification cases ($c = 1, 2$). In this case, to reach the minimum of

$$L_{\text{cls}}^{2\text{-class}} = \int_{\Omega} - \left(\rho_1(x) \log(p(x)) + \rho_2(x) \log(1 - p(x)) \right) dx, \quad (\text{B.8})$$

by taking its derivative, one gets $p(x) = \rho_1(x)/(\rho_1(x) + \rho_2(x))$.

In conclusion, the proof shows that the function that minimizes cross-entropy loss is learned to estimate the normalized PDF in the latent space. In our words, given a specific jet, x_0 , the learned score $\hat{y}_c(x_0)$ represents the probability that the jet belongs to class c , which explains why it is named the probability score.

B.2 Regression

We then derive the function that minimizes the regression loss function in Eq. (8.3). In the probability language, a regression target is a mapping from the jet $x \in \Omega$ to $\theta \in \mathbb{R}$. Since the regression target is a GEN-level objective, the same jet x can map to different θ since the generator-to-reconstruction process is a smearing process. Hence, the regressed target is

sampled from a PDF on x and class c :

$$\theta(x_0) \sim \mu_c(x_0, \theta). \quad (\text{B.9})$$

We denote the learned output regression value as $\hat{\theta}(x)$. Therefore, with a sufficiently large N , the sum of regression loss across all samples is written as

$$L_{\text{reg}} = N \cdot \sum_c \int_{\Omega} \int_{-\infty}^{+\infty} \rho_c(x) \mu_c(x, \theta) \log \cosh (\hat{\theta}(x) - \theta) dx d\theta, \quad (\text{B.10})$$

or

$$L_{\text{reg}} = N \cdot \sum_c \mathbb{E}_{x \sim \rho_c(x)} \mathbb{E}_{\theta \sim \mu_c(x, \theta)} [\log \cosh (\hat{\theta}(x) - \theta)]. \quad (\text{B.11})$$

By defining $P(x, \theta) \equiv \sum_c \rho_c(x) \mu_c(x, \theta)$, we see

$$l_{\text{reg}}(x) = N \cdot \int_{-\infty}^{+\infty} P(x, \theta) \log \cosh (\hat{\theta}(x) - \theta) d\theta, \quad (\text{B.12})$$

and $L_{\text{reg}} = \int_{\Omega} l_{\text{reg}}(x) dx$.

Therefore, the optimization process is to reach the minimum regression loss for every jet x . This can also be intuitively reached by thoughts: the regression is independent of class c , so one can just mix all classes together. By a given jet x_0 , the GEN-level regression target should be a distribution, and this is $P(x_0, \theta)$. We are therefore minimizing $l_{\text{reg}}(x)$, which is to find a place, $\hat{\theta}$, in the middle of the distribution $P(x_0, \theta)$ according to the regression loss form. For the log-cosh loss, the optimized $\hat{\theta}$ satisfies

$$\frac{dl_{\text{reg}}(x)}{d\hat{\theta}} = \int_{-\infty}^{+\infty} P(x_0, \theta) \tanh (\hat{\theta}(x_0) - \theta) d\theta = 0. \quad (\text{B.13})$$

When training with the regression objective with the log-cosh loss, we are reaching a function that finds solutions of the above equation for every x_0 .

References

- [1] THOMSON J J. Cathode rays[J]. *Phil. Mag. Ser. 5*, 1897, 44: 293-316.
- [2] RUTHERFORD E. The scattering of alpha and beta particles by matter and the structure of the atom[J]. *Phil. Mag. Ser. 6*, 1911, 21: 669-688.
- [3] CHADWICK J. Possible existence of a neutron[J]. *Nature*, 1932, 129: 312.
- [4] TOMONAGA S. On a relativistically invariant formulation of the quantum theory of wave fields[J]. *Prog. Theor. Phys.*, 1946, 1: 27-42.
- [5] SCHWINGER J S. Quantum electrodynamics. I A covariant formulation[J]. *Phys. Rev.*, 1948, 74: 1439.
- [6] FEYNMAN R P. Space-time approach to quantum electrodynamics[J]. *Phys. Rev.*, 1949, 76: 769-789.
- [7] DYSON F J. The Radiation theories of Tomonaga, Schwinger, and Feynman[J]. *Phys. Rev.*, 1949, 75: 486-502.
- [8] YANG C N, MILLS R L. Conservation of isotopic spin and isotopic gauge invariance[J]. *Phys. Rev.*, 1954, 96: 191-195.
- [9] WEINBERG S. A model of leptons[J]. *Phys. Rev. Lett.*, 1967, 19: 1264-1266.
- [10] SALAM A. Weak and electromagnetic interactions[J]. *Conf. Proc. C*, 1968, 680519: 367-377.
- [11] GELL-MANN M. Quarks[J]. *Acta Phys. Austriaca Suppl.*, 1972, 9: 733-761.
- [12] GROSS D J, WILCZEK F. Ultraviolet behavior of nonabelian gauge theories[J]. *Phys. Rev. Lett.*, 1973, 30: 1343-1346.
- [13] POLITZER H D. Reliable perturbative results for strong interactions?[J]. *Phys. Rev. Lett.*, 1973, 30: 1346-1349.
- [14] HIGGS P W. Broken symmetries and the masses of gauge bosons[J]. *Phys. Rev. Lett.*, 1964, 13: 508-509.
- [15] ENGLERT F, BROUT R. Broken symmetry and the mass of gauge vector mesons[J]. *Phys. Rev. Lett.*, 1964, 13: 321-323.
- [16] GURALNIK G S, HAGEN C R, KIBBLE T W B. Global conservation laws and massless particles[J]. *Phys. Rev. Lett.*, 1964, 13: 585-587.
- [17] Particle Data Group. Review of Particle Physics[J]. *PTEP*, 2022, 2022: 083C01.
- [18] QUINNAN M K. Standard model four-top quark production in the all-hadronic final state in proton-proton collisions at 13 TeV with the CMS experiment[D]. University of California, Santa Barbara, 2022.
- [19] CMS Collaboration. A portrait of the Higgs boson by the CMS experiment ten years after the discovery[J]. *Nature*, 2022, 607(7917): 60-68. arXiv: 2207.00043 [hep-ex].

- [20] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector[M/OL]. Geneva: CERN, 2017. <https://cds.cern.ch/record/2227475>. DOI: 10.23731/CYRM-2017-002.
- [21] Service graphique, CERN. Overall view of the LHC. Vue d'ensemble du LHC[EB/OL]. 2014. <https://cds.cern.ch/record/1708849>.
- [22] CMS Collaboration. Cutaway diagrams of CMS detector[EB/OL]. 2019. <https://cds.cern.ch/record/2665537>.
- [23] CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker[J]. JINST, 2014, 9: P10009. arXiv: 1405.6569 [physics.ins-det].
- [24] CMS Collaboration. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. JINST, 2018, 13: P06015. arXiv: 1804.04528 [physics.ins-det].
- [25] ALWALL J, FREDERIX R, FRIXIONE S, et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations[J]. JHEP, 2014, 07: 079. arXiv: 1405.0301 [hep-ph].
- [26] FRIXIONE S, NASON P, OLEARI C. Matching NLO QCD computations with parton shower simulations: the POWHEG method[J]. JHEP, 2007, 11: 070. arXiv: 0709.2092 [hep-ph].
- [27] ALIOLI S, NASON P, OLEARI C, et al. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX[J]. JHEP, 2010, 06: 043. arXiv: 1002.2581 [hep-ph].
- [28] SJÖSTRAND T, ASK S, CHRISTIANSEN J R, et al. An introduction to PYTHIA 8.2[J]. Comput. Phys. Commun., 2015, 191: 159. arXiv: 1410.3012 [hep-ph].
- [29] BELLM J, et al. Herwig 7.0/Herwig++ 3.0 release note[J]. Eur. Phys. J. C, 2016, 76(4): 196. arXiv: 1512.01178 [hep-ph].
- [30] CMS Collaboration. Using the CMS threaded framework in a production environment[J]. J. Phys. Conf. Ser., 2015, 664(7): 072026.
- [31] CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector[J]. JINST, 2017, 12: P10003. arXiv: 1706.04965 [physics.ins-det].
- [32] FRUHWIRTH R, WALTENBERGER W, VANLAER P. Adaptive vertex fitting[J]. J. Phys. G, 2007, 34: N343.
- [33] CACCIARI M, SALAM G P, SOYEZ G. The anti- k_T jet clustering algorithm[J]. JHEP, 2008, 04: 063. arXiv: 0802.1189 [hep-ex].
- [34] CMS Collaboration. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV [J]. JINST, 2017, 12: P02014. arXiv: 1607.03663 [hep-ex].
- [35] CMS Collaboration. Pileup mitigation at CMS in 13 TeV data[J]. JINST, 2020, 15: P09018. arXiv: 2003.00503 [hep-ex].
- [36] BERTOLINI D, HARRIS P, LOW M, et al. Pileup Per Particle Identification[J]. JHEP, 2014, 10: 059. arXiv: 1407.6013 [hep-ph].

-
- [37] DASGUPTA M, FREGOSO A, MARZANI S, et al. Towards an understanding of jet substructure[J]. JHEP, 2013, 09: 029. arXiv: 1307.0007 [hep-ph].
 - [38] LARKOSKI A J, MARZANI S, SOYEZ G, et al. Soft drop[J]. JHEP, 2014, 05: 146. arXiv: 1402.2657 [hep-ph].
 - [39] DREYER F A, NECIB L, SOYEZ G, et al. Recursive soft drop[J]. JHEP, 2018, 06: 093. arXiv: 1804.03657 [hep-ph].
 - [40] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity[J]. Bull. Math. Biol., 1943, 5: 115-133.
 - [41] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychol. Rev., 1958, 65(6): 386.
 - [42] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.
 - [43] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Comput., 1989, 1(4): 541-551.
 - [44] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Comput., 1997, 9(8): 1735-1780.
 - [45] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Comput., 2006, 18(7): 1527-1554.
 - [46] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. Lake Tahoe, NV, USA: Curran Associates, Inc., 2012: 1097-1105.
 - [47] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Miami, FL, USA: IEEE Computer Society, 2009: 248-255.
 - [48] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time series[M]//The Handbook of Brain Theory and Neural Networks. Cambridge, MA, USA: MIT Press, 1998: 255-258.
 - [49] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems: vol. 27. Montreal, Canada: Curran Associates, Inc., 2014.
 - [50] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//North American Chapter of the Association for Computational Linguistics. Minneapolis, MN, USA: Association for Computational Linguistics, 2019.
 - [51] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[R]. San Francisco, CA, USA: OpenAI, 2018.
 - [52] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.
 - [53] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

- [54] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.
- [55] SANCHEZ-LENGELING B, REIF E, PEARCE A, et al. A gentle introduction to graph neural networks[J]. Distill, 2021.
- [56] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems: vol. 30. Long Beach, CA, USA: Curran Associates, Inc., 2017: 5998-6008.
- [57] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. Virtual: OpenReview.net, 2021: 3156-3164.
- [58] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image Transformers & distillation through attention[C]//Proceedings of the 38th International Conference on Machine Learning: vol. 139. Virtual: PMLR, 2021: 10347-10357.
- [59] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 10012-10022.
- [60] YUN S, JEONG M, KIM R, et al. Graph Transformer networks[C]//Advances in Neural Information Processing Systems: vol. 32. Vancouver, Canada: Curran Associates, Inc., 2019.
- [61] GUO M H, CAI J X, LIU Z N, et al. PCT: Point cloud transformer[J]. Computational Visual Media, 2021, 7(2): 187-199.
- [62] BI K, XIE L, ZHANG H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619(7970): 533-538.
- [63] De FAVEREAU J, DELAERE C, DEMIN P, et al. DELPHES 3, A modular framework for fast simulation of a generic collider experiment[J]. JHEP, 2014, 02: 057. arXiv: 1307.6346 [hep-ex].
- [64] SJÖSTRAND T, ASK S, CHRISTIANSEN J R, et al. An introduction to PYTHIA 8.2[J]. Comput. Phys. Commun., 2015, 191: 159-177. arXiv: 1410.3012 [hep-ph].
- [65] QU H, LI C, QIAN S. Particle Transformer for jet tagging[C]//Proceedings of Machine Learning Research: Proceedings of the 39th International Conference on Machine Learning: vol. 162. Baltimore, MD, USA: PMLR, 2022: 18281-18292. arXiv: 2202.03772 [hep-ph].
- [66] BUTTERWORTH J M, DAVISON A R, RUBIN M, et al. Jet substructure as a new Higgs search channel at the LHC[J]. Phys. Rev. Lett., 2008, 100: 242001. arXiv: 0802.2470 [hep-ph].
- [67] ELLIS S D, VERMILION C K, WALSH J R. Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches[J]. Phys. Rev. D, 2010, 81: 094023. arXiv: 0912.0033 [hep-ph].
- [68] KROHN D, THALER J, WANG L T. Jet trimming[J]. JHEP, 2010, 02: 084. arXiv: 0912.1342 [hep-ph].
- [69] THALER J, VAN TILBURG K. Identifying boosted objects with N -subjettiness[J]. JHEP, 2011, 03: 015. arXiv: 1011.2268 [hep-ph].

-
- [70] CMS Collaboration. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques[J]. JINST, 2020, 15: P06005. arXiv: 2004.08262 [hep-ex].
- [71] TKACHOV F V. Measuring multi-jet structure of hadronic energy flow or What is a jet?[J]. Int. J. Mod. Phys. A, 1997, 12: 5411-5529. arXiv: hep-ph/9601308.
- [72] LARKOSKI A J, SALAM G P, THALER J. Energy correlation functions for jet substructure[J]. JHEP, 2013, 06: 108. arXiv: 1305.0007 [hep-ph].
- [73] MOULT I, NECIB L, THALER J. New angles on energy correlation functions[J]. JHEP, 2016, 12: 153. arXiv: 1609.07483 [hep-ph].
- [74] KOMISKE P T, METODIEV E M, THALER J. Energy flow polynomials: A complete linear basis for jet substructure[J]. JHEP, 2018, 04: 013. arXiv: 1712.07124 [hep-ph].
- [75] CMS Collaboration. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV [J]. JINST, 2018, 13: P05011. arXiv: 1712.07158 [physics.ins-det].
- [76] CMS Collaboration. Performance summary of AK4 jet b tagging with data from proton-proton collisions at 13 TeV with the CMS detector[R/OL]. 2023. <https://cds.cern.ch/record/2854609>.
- [77] CMS Collaboration. Performance summary of AK4 jet charm tagging with the CMS Run2 Legacy dataset[R/OL]. 2023. <https://cds.cern.ch/record/2854610>.
- [78] CMS Collaboration. The CMS trigger system[J]. JINST, 2017, 12: P01020. arXiv: 1609.02366 [physics.ins-det].
- [79] COGAN J, KAGAN M, STRAUSS E, et al. Jet-images: Computer vision inspired techniques for jet tagging[J]. JHEP, 2015, 02: 118. arXiv: 1407.5675 [hep-ph].
- [80] De OLIVEIRA L, KAGAN M, MACKEY L, et al. Jet-images — deep learning edition[J]. JHEP, 2016, 07: 069. arXiv: 1511.05190 [hep-ph].
- [81] KOMISKE P T, METODIEV E M, SCHWARTZ M D. Deep learning in color: towards automated quark/gluon jet discrimination[J]. JHEP, 2017, 01: 110. arXiv: 1612.01551 [hep-ph].
- [82] MACALUSO S, SHIH D. Pulling out all the tops with computer vision and deep learning[J]. JHEP, 2018, 10: 121. arXiv: 1803.00107 [hep-ph].
- [83] LOUPPE G, CHO K, BECOT C, et al. QCD-aware recursive neural networks for jet physics[J]. JHEP, 2019, 01: 057. arXiv: 1702.00748 [hep-ph].
- [84] CHENG T. Recursive neural networks in quark/gluon tagging[J]. Comput. Softw. Big Sci., 2018, 2(1): 3. arXiv: 1711.02633 [hep-ph].
- [85] CMS Collaboration. Performance of deep tagging algorithms for boosted double quark jet topology in proton-proton collisions at 13 TeV with the Phase-0 CMS detector[R/OL]. 2018. <https://cds.cern.ch/record/2630438>.
- [86] CMS Collaboration. Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector[R/OL]. 2022. <https://cds.cern.ch/record/2839736>.
- [87] KOMISKE P T, METODIEV E M, THALER J. Energy Flow Networks: Deep sets for particle jets[J]. JHEP, 2019, 01: 121. arXiv: 1810.05165 [hep-ph].

- [88] HENRION I, BREHMER J, BRUNA J, et al. Neural message passing for jet physics[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning for Physical Sciences. Long Beach, CA, USA: [s.n.], 2017.
- [89] MORENO E A, CERRI O, DUARTE J M, et al. JEDI-net: a jet identification algorithm based on interaction networks[J]. Eur. Phys. J. C, 2020, 80(1): 58. arXiv: 1908.05318 [hep-ex].
- [90] QU H, GOUSKOS L. ParticleNet: Jet tagging via particle clouds[J]. Phys. Rev. D, 2020, 101(5): 056019. arXiv: 1902.08570 [hep-ph].
- [91] WANG Y, SUN Y, LIU Z, et al. Dynamic Graph CNN for learning on point clouds[J]. ACM Transactions on Graphics (TOG), 2019, 38(5): 1-12.
- [92] DREYER F A, QU H. Jet tagging in the Lund plane with graph networks[J]. JHEP, 2021, 03: 052. arXiv: 2012.08526 [hep-ph].
- [93] MIKUNI V, CANELLI F. ABCNet: An attention-based method for particle tagging[J]. Eur. Phys. J. Plus, 2020, 135(6): 463. arXiv: 2001.05311 [physics.data-an].
- [94] GONG S, MENG Q, ZHANG J, et al. An efficient Lorentz equivariant graph neural network for jet tagging[J]. JHEP, 2022, 07: 030. arXiv: 2201.08187 [hep-ph].
- [95] BOGATSKIY A, HOFFMAN T, MILLER D W, et al. PELICAN: Permutation Equivariant and Lorentz Invariant or Covariant Aggregator Network for Particle Physics[C]//Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: [s.n.], 2022. arXiv: 2211.00454 [hep-ph].
- [96] QIU S, HAN S, JU X, et al. Holistic approach to predicting top quark kinematic properties with the covariant particle transformer[J]. Phys. Rev. D, 2023, 107(11): 114029. arXiv: 2203.05687 [hep-ph].
- [97] SHIMMIN C. Particle convolution for high energy physics[R]. [S.l. : s.n.], 2021. arXiv: 2107.02908 [hep-ph].
- [98] MUNOZ J M, BATATIA I, ORTNER C. Boost invariant polynomials for efficient jet tagging[J]. Mach. Learn. Sci. Tech., 2022, 3(4): 04LT05. arXiv: 2207.08272 [physics.comp-ph].
- [99] BOGATSKIY A, ANDERSON B, OFFERMANN J, et al. Lorentz group equivariant neural network for particle physics[C]//International Conference on Machine Learning. Virtual: PMLR, 2020: 992-1002. arXiv: 2006.04780 [hep-ph].
- [100] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[C]//5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.
- [101] ZHANG M, LUCAS J, BAI J, et al. Lookahead optimizer: k steps forward, 1 step back[C]//Advances in Neural Information Processing Systems: vol. 32. Vancouver, Canada: Curran Associates, Inc., 2019.
- [102] LIU L, JIANG H, HE P, et al. On the variance of the adaptive learning rate and beyond[C]//International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020.
- [103] HAO Z, KANSAL R, DUARTE J, et al. Lorentz group equivariant autoencoders[J]. Eur. Phys. J. C, 2023, 83(6): 485. arXiv: 2212.07347 [hep-ex].

-
- [104] LI C, QU H, QIAN S, et al. Does Lorentz-symmetric design boost network performance in jet physics?[J]. Phys. Rev. D, 2024, 109(5): 056003. arXiv: 2208.07814 [hep-ph].
 - [105] MIKUNI V, CANELLI F. Point cloud transformers applied to collider physics[J]. Mach. Learn. Sci. Tech., 2021, 2(3): 035027. arXiv: 2102.05073 [physics.data-an].
 - [106] TOUVRON H, CORD M, SABLAYROLLES A, et al. Going deeper with image Transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Computer Society, 2021: 32-42.
 - [107] CMS Collaboration. Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair[J]. Phys. Rev. Lett., 2018, 120: 071802. arXiv: 1709.05543 [hep-ex].
 - [108] CMS Collaboration. Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. JHEP, 2020, 12: 085. arXiv: 2006.13251 [hep-ex].
 - [109] CMS Collaboration. Search for nonresonant pair production of highly energetic Higgs bosons decaying to bottom quarks[J]. Phys. Rev. Lett., 2023, 131: 041803. arXiv: 2205.06667 [hep-ex].
 - [110] ATLAS Collaboration. Measurement of the associated production of a Higgs boson decaying into b -quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector[J]. Phys. Lett. B, 2021, 816: 136204. arXiv: 2008.02508 [hep-ex].
 - [111] ATLAS Collaboration. Constraints on Higgs boson production with large transverse momentum using $H \rightarrow b\bar{b}$ decays in the ATLAS detector[J]. Phys. Rev. D, 2022, 105(9): 092003. arXiv: 2111.08340 [hep-ex].
 - [112] CMS Collaboration. A search for the standard model Higgs boson decaying to charm quarks[J]. JHEP, 2020, 03: 131. arXiv: 1912.01662 [hep-ex].
 - [113] CMS Collaboration. Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. Phys. Rev. Lett., 2023, 131: 061801. arXiv: 2205.05550 [hep-ex].
 - [114] CMS Collaboration. Search for Higgs boson and observation of Z boson through their decay into a charm quark-antiquark pair in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. Phys. Rev. Lett., 2023, 131: 041801. arXiv: 2211.14181 [hep-ex].
 - [115] CMS Collaboration. Search for a massive scalar resonance decaying to a light scalar and a Higgs boson in the four b quarks final state with boosted topology[J]. Phys. Lett. B, 2022, 19. arXiv: 2204.12413 [hep-ex].
 - [116] ATLAS Collaboration. Identification of boosted Higgs bosons decaying into $b\bar{b}$ with neural networks and variable radius jets in ATLAS[R/OL]. 2020. <https://cds.cern.ch/record/2724739>.
 - [117] CMS Collaboration. Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques[R/OL]. 2020. <https://cds.cern.ch/record/2707946>.
 - [118] ATLAS Collaboration. Transformer neural networks for identifying boosted Higgs bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS[R/OL]. Geneva: CERN, 2023. <https://cds.cern.ch/record/2866601>.

- [119] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016: 770.
- [120] CMS Collaboration. Search for resonant Higgs boson pair production in four b quark final state using large-area jets in proton-proton collisions at $\sqrt{s} = 13$ TeV[R/OL]. 2021. <http://cds.cern.ch/record/2777083>.
- [121] CMS Collaboration. W and top tagging scale factors for Run 2 data[R/OL]. 2020. <https://cds.cern.ch/record/2718978>.
- [122] ATLAS Collaboration. Boosted hadronic vector boson and top quark tagging with ATLAS using Run 2 data[R/OL]. Geneva: CERN, 2020. <https://cds.cern.ch/record/2724149>.
- [123] ATLAS Collaboration. Identification of boosted Higgs bosons decaying into b -quark pairs with the ATLAS detector at 13 TeV[J]. Eur. Phys. J. C, 2019, 79: 836. arXiv: 1906.11005 [hep-ex].
- [124] CMS Collaboration. Performance of heavy-flavour jet identification in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV[R/OL]. 2023. <https://cds.cern.ch/record/2866276>.
- [125] CMS Collaboration. Calibration of the mass-decorrelated ParticleNet tagger for boosted $b\bar{b}$ and $c\bar{c}$ jets using LHC Run 2 data[R/OL]. 2022. <https://cds.cern.ch/record/2805611>.
- [126] CMS Collaboration. Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data[R/OL]. 2020. <https://cds.cern.ch/record/2740688>.
- [127] HAMILTON K, NASON P, OLEARI C, et al. Merging $H/W/Z + 0$ and 1 jet at NLO with no merging scale: a path to parton shower + NNLO matching[J]. JHEP, 2013, 05: 082. arXiv: 1212.4504 [hep-ph].
- [128] LUISONI G, NASON P, OLEARI C, et al. $HW^\pm/HZ+0$ and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO[J]. JHEP, 2013, 10: 083. arXiv: 1306.2542 [hep-ph].
- [129] NASON P. A new method for combining NLO QCD with shower Monte Carlo algorithms[J]. JHEP, 2004, 11: 040. arXiv: hep-ph/0409146 [hep-ph].
- [130] CAMPBELL J M, ELLIS R K, NASON P, et al. Top-pair production and decay at NLO matched with parton showers[J]. JHEP, 2015, 04: 114. arXiv: 1412.1828 [hep-ph].
- [131] CZAKON M, MITOV A. Top++: A program for the calculation of the top-pair cross-section at hadron colliders[J]. Comput. Phys. Commun., 2014, 185: 2930. arXiv: 1112.5675 [hep-ph].
- [132] FREDERIX R, RE E, TORRIELLI P. Single-top t -channel hadroproduction in the four-flavour scheme with POWHEG and aMC@NLO[J]. JHEP, 2012, 09: 130. arXiv: 1207.5391 [hep-ph].
- [133] ALIOLI S, NASON P, OLEARI C, et al. NLO single-top production matched with shower in POWHEG: s - and t -channel contributions[J]. JHEP, 2009, 09: 111. arXiv: 0907.4076 [hep-ph].
- [134] FRIXIONE S, LAENEN E, MOTYLINSKI P, et al. Single-top hadroproduction in association with a W boson[J]. JHEP, 2008, 07: 029. arXiv: 0805.3067 [hep-ph].
- [135] RE E. Single-top Wt -channel production matched with parton showers using the POWHEG method[J]. Eur. Phys. J. C, 2011, 71: 1547. arXiv: 1009.2450 [hep-ph].

-
- [136] KIDONAKIS N. NNLL threshold resummation for top-pair and single-top production[J]. Phys. Part. Nucl., 2014, 45: 714. arXiv: 1210.7813 [hep-ph].
- [137] CMS. Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements[J]. Eur. Phys. J. C, 2020, 80: 4. arXiv: 1903.12179 [hep-ex].
- [138] NNPDF. Parton distributions from high-precision collider data[J]. Eur. Phys. J. C, 2017, 77: 663. arXiv: 1706.00428 [hep-ph].
- [139] ALWALL J, HÖCHE S, KRAUSS F, et al. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions[J]. Eur. Phys. J. C, 2008, 53: 473. arXiv: 0706.2569 [hep-ph].
- [140] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: Association for Computing Machinery, 2016: 785-794.
- [141] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The elements of statistical learning: data mining, inference, and prediction[M]. New York, NY, USA: Springer, 2009.
- [142] CATANI S, DOKSHITZER Y L, SEYMOUR M H, et al. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions[J]. Nucl. Phys. B, 1993, 406: 187.
- [143] ELLIS S D, SOPER D E. Successive combination jet algorithm for hadron collisions[J]. Phys. Rev. D, 1993, 48: 3160. arXiv: hep-ph/9305266.
- [144] CMS Collaboration. Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector[J]. JINST, 2019, 14: P07004. arXiv: 1903.06078 [hep-ex].
- [145] CMS Collaboration. Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS[J]. Eur. Phys. J. C, 2021, 81: 800. arXiv: 2104.01927 [hep-ex].
- [146] CMS Collaboration. CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV[R/OL]. 2018. <https://cds.cern.ch/record/2621960/>.
- [147] CMS Collaboration. CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV[R/OL]. 2019. <https://cds.cern.ch/record/2676164/>.
- [148] CMS Collaboration. Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV [J]. JHEP, 2018, 07: 161. arXiv: 1802.02613 [hep-ex].
- [149] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC[J]. Phys. Lett. B, 2012, 716: 1-29. arXiv: 1207.7214 [hep-ex].
- [150] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC[J]. Phys. Lett. B, 2012, 716: 30-61. arXiv: 1207.7235 [hep-ex].
- [151] CMS Collaboration. Observation of $t\bar{t}H$ production[J]. Phys. Rev. Lett., 2018, 120: 231801. arXiv: 1804.02610 [hep-ex].
- [152] ATLAS Collaboration. Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector[J]. Phys. Lett. B, 2018, 784: 173. arXiv: 1806.00425 [hep-ex].

- [153] CMS Collaboration. Observation of Higgs boson decay to bottom quarks[J]. Phys. Rev. Lett., 2018, 121: 121801. arXiv: 1808.08242 [hep-ex].
- [154] ATLAS Collaboration. Observation of $H \rightarrow b\bar{b}$ decays and VH production with the ATLAS detector[J]. Phys. Lett. B, 2018, 786: 59-86. arXiv: 1808.08238 [hep-ex].
- [155] CMS Collaboration. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector[J]. Phys. Lett. B, 2018, 779: 283-316. arXiv: 1708.00373 [hep-ex].
- [156] CMS Collaboration. Evidence for Higgs boson decay to a pair of muons[J]. JHEP, 2021, 01: 148. arXiv: 2009.04363 [hep-ex].
- [157] GHOSH D, GUPTA R S, PEREZ G. Is the Higgs mechanism of fermion mass generation a fact? A Yukawa-less first-two-generation Model[J]. Phys. Lett. B, 2016, 755: 504-508. arXiv: 1508.01501 [hep-ph].
- [158] BOTELLA F J, BRANCO G C, REBELO M N, et al. What if the masses of the first two quark families are not generated by the standard model Higgs boson?[J]. Phys. Rev. D, 2016, 94(11): 115031. arXiv: 1602.08011 [hep-ph].
- [159] HARNIK R, KOPP J, ZUPAN J. Flavor violating Higgs decays[J]. JHEP, 2013, 03: 026. arXiv: 1209.1397 [hep-ph].
- [160] ALTMANNSHOFER W, EBY J, GORI S, et al. Collider signatures of flavorful Higgs bosons[J]. Phys. Rev. D, 2016, 94(11): 115032. arXiv: 1610.02398 [hep-ph].
- [161] ATLAS Collaboration. Search for the decay of the Higgs boson to charm quarks with the ATLAS experiment[J]. Phys. Rev. Lett., 2018, 120(21): 211802. arXiv: 1802.04329 [hep-ex].
- [162] ATLAS Collaboration. Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector[J]. Eur. Phys. J. C, 2022, 82: 717. arXiv: 2201.11428 [hep-ex].
- [163] CMS Collaboration. Search for rare decays of Z and Higgs bosons to J/ψ and a photon in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. Eur. Phys. J. C, 2019, 79(2): 94. arXiv: 1810.10056 [hep-ex].
- [164] ATLAS Collaboration. Searches for exclusive Higgs and Z boson decays into $J/\psi\gamma$, $\psi(2S)\gamma$, and $\Upsilon(nS)\gamma$ at $\sqrt{s} = 13$ TeV with the ATLAS detector[J]. Phys. Lett. B, 2018, 786: 134-155. arXiv: 1807.00802 [hep-ex].
- [165] BISHARA F, HAISCH U, MONNI P F, et al. Constraining light-quark Yukawa couplings from Higgs distributions[J]. Phys. Rev. Lett., 2017, 118(12): 121801. arXiv: 1606.09253 [hep-ph].
- [166] ATLAS Collaboration. Measurements of the Higgs boson inclusive and differential fiducial cross-sections in the diphoton decay channel with pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector[J]. JHEP, 2022, 08: 027. arXiv: 2202.00487 [hep-ex].
- [167] CMS Collaboration. Measurements of inclusive and differential cross sections for the Higgs boson production and decay to four-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. JHEP, 2023, 08: 040. arXiv: 2305.07532 [hep-ex].
- [168] BRIVIO I, GOERTZ F, ISIDORI G. Probing the charm quark Yukawa coupling in Higgs+charm production[J]. Phys. Rev. Lett., 2015, 115(21): 211801. arXiv: 1507.02916 [hep-ph].

-
- [169] YU F. Phenomenology of enhanced light quark Yukawa couplings and the $W^\pm h$ charge asymmetry[J]. JHEP, 2017, 02:083. arXiv: 1609.06592 [hep-ph].
 - [170] HAMILTON K, NASON P, ZANDERIGHI G. MINLO: Multi-Scale Improved NLO[J]. JHEP, 2012, 10: 155. arXiv: 1206.3572 [hep-ph].
 - [171] FERRERA G, GRAZZINI M, TRAMONTANO F. Higher-order QCD effects for associated WH production and decay at the LHC[J]. JHEP, 2014, 04: 039. arXiv: 1312.1669 [hep-ph].
 - [172] FERRERA G, GRAZZINI M, TRAMONTANO F. Associated ZH production at hadron colliders: the fully differential NNLO QCD calculation[J]. Phys. Lett. B, 2015, 740: 51. arXiv: 1407.4747 [hep-ph].
 - [173] BREIN O, HARLANDER R V, ZIRKE T J E. VH@NNLO — Higgs Strahlung at hadron colliders[J]. Comput. Phys. Commun., 2013, 184: 998. arXiv: 1210.5347 [hep-ph].
 - [174] HARLANDER R V, LIEBLER S, ZIRKE T. Higgs Strahlung at the Large Hadron Collider in the 2-Higgs-Doublet Model[J]. JHEP, 2014, 02: 023. arXiv: 1307.8122 [hep-ph].
 - [175] DENNER A, DITTMAIER S, KALLWEIT S, et al. HAWK 2.0: A Monte Carlo program for Higgs production in vector-boson fusion and Higgs strahlung at hadron colliders[J]. Comput. Phys. Commun., 2015, 195: 161. arXiv: 1412.5390 [hep-ph].
 - [176] CZAKON M, HEYMES D, MITOV A, et al. Top-pair production at the LHC through NNLO QCD and NLO EW[J]. JHEP, 2017, 10: 186. arXiv: 1705.04105 [hep-ph].
 - [177] GRAZZINI M, KALLWEIT S, LINDERT J M, et al. NNLO QCD + NLO EW with Matrix+OpenLoops: precise predictions for vector-boson pair production[J]. JHEP, 2020, 02: 087. arXiv: 1912.00068 [hep-ph].
 - [178] BUTTERWORTH J, et al. PDF4LHC recommendations for LHC Run II[J]. J. Phys. G, 2016, 43: 023001. arXiv: 1510.03865 [hep-ph].
 - [179] CMS Collaboration. Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. JHEP, 2018, 10: 117. arXiv: 1805.07399 [hep-ex].
 - [180] SIRUNYAN A M, et al. Measurement of the single top quark and antiquark production cross sections in the t channel and their ratio in proton-proton collisions at $\sqrt{s} = 13$ TeV[J]. Phys. Lett. B, 2020, 800: 135042. arXiv: 1812.10514 [hep-ex].
 - [181] CMS Collaboration. Measurement of the WW cross section pp collisions at $\sqrt{s} = 13$ TeV[R/OL]. 2016. <https://cds.cern.ch/record/2160868>.
 - [182] CMS Collaboration. Measurements of the $pp \rightarrow ZZ$ production cross section and the $Z \rightarrow 4\ell$ branching fraction, and constraints on anomalous triple gauge couplings at $\sqrt{s} = 13$ TeV[J]. Eur. Phys. J. C, 2018, 78: 165. arXiv: 1709.08601 [hep-ex].
 - [183] SIRUNYAN A M, et al. Measurements of the $pp \rightarrow WZ$ inclusive and differential production cross section and constraints on charged anomalous triple gauge couplings at $\sqrt{s} = 13$ TeV[J]. JHEP, 2019, 04: 122. arXiv: 1901.03428 [hep-ex].
 - [184] BARLOW R J, BEESTON C. Fitting using finite Monte Carlo samples[J]. Comput. Phys. Commun., 1993, 77: 219-228.

- [185] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector[J]. CERN, 2016. arXiv: 1610.07922 [hep-ph].
- [186] CMS Collaboration. Direct search for the standard model Higgs boson decaying to a charm quark-antiquark pair[R/OL]. 2021. <https://cds.cern.ch/record/2802742>.
- [187] NEYMAN J, PEARSON E S. IX. On the problem of the most efficient tests of statistical hypotheses[J]. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1933, 231(694-706): 289-337.
- [188] DATTA K, LARKOSKI A. How much information is in a jet?[J]. JHEP, 2017, 06: 073. arXiv: 1704.08249 [hep-ph].
- [189] ZHAI X, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, LA, USA: IEEE, 2022: 12104-12113.
- [190] CMS Collaboration. Reconstruction of decays to merged photons using end-to-end deep learning with domain continuation in the CMS detector[J]. Phys. Rev. D, 2023, 108(5): 052002. arXiv: 2204.12313 [hep-ex].
- [191] DEZOORT G, THAIS S, DUARTE J, et al. Charged particle tracking via edge-classifying interaction networks[J]. Comput. Softw. Big Sci., 2021, 5(1): 26. arXiv: 2103.16701 [hep-ex].
- [192] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, LA, USA: IEEE, 2022: 16000-16009.
- [193] BOLS E, KIESELER J, VERZETTI M, et al. Jet flavour classification using DeepJet[J]. JINST, 2020, 15: P12012. arXiv: 2008.10519 [hep-ex].
- [194] CMS Collaboration. Transformer models for heavy flavor jet identification[R/OL]. 2022. <https://cds.cern.ch/record/2839920>.
- [195] SMITH L N, TOPIN N. Super-convergence: Very fast training of neural networks using large learning rates[C]//Artificial intelligence and machine learning for multi-domain operations applications: vol. 11006. Baltimore, MD, USA: SPIE, 2019: 369-386.
- [196] LI C, et al. Accelerating resonance searches via signature-oriented pre-training[R]. [S.l. : s.n.], 2024. arXiv: 2405.12972 [hep-ph].
- [197] COWAN G, CRANMER K, GROSS E, et al. Asymptotic formulae for likelihood-based tests of new physics[J]. Eur. Phys. J. C, 2011, 71: 1554. arXiv: 1007.1727 [physics.data-an].
- [198] AGASHE K, DU P, HONG S, et al. Flavor universal resonances and warped gravity[J]. JHEP, 2017, 01: 016. arXiv: 1608.00526 [hep-ph].
- [199] AGASHE K, COLLINS J H, DU P, et al. Dedicated Strategies for Triboson Signals from Cascade Decays of Vector Resonances[J]. Phys. Rev. D, 2019, 99(7): 075016. arXiv: 1711.09920 [hep-ph]. DOI: 10.1103/PhysRevD.99.075016.
- [200] COLLINS J H, HOWE K, NACHMAN B. Anomaly detection for resonant new physics with machine learning[J]. Phys. Rev. Lett., 2018, 121(24): 241803. arXiv: 1805.02664 [hep-ph].

- [201] METODIEV E M, NACHMAN B, THALER J. Classification without labels: Learning from mixed samples in high energy physics[J]. JHEP, 2017, 10: 174. arXiv: 1708.02949 [hep-ph].
- [202] COLLINS J H, HOWE K, NACHMAN B. Extending the search for new resonances with machine learning[J]. Phys. Rev. D, 2019, 99(1): 014038. arXiv: 1902.02634 [hep-ph].
- [203] NACHMAN B, SHIH D. Anomaly detection with density estimation[J]. Phys. Rev. D, 2020, 101: 075042. arXiv: 2001.04990 [hep-ph].
- [204] ANDREASSEN A, NACHMAN B, SHIH D. Simulation assisted likelihood-free anomaly detection[J]. Phys. Rev. D, 2020, 101(9): 095004. arXiv: 2001.05001 [hep-ph].
- [205] HALLIN A, ISAACSON J, KASIECZKA G, et al. Classifying anomalies through outer density estimation[J]. Phys. Rev. D, 2022, 106(5): 055006. arXiv: 2109.00546 [hep-ph].
- [206] HALLIN A, KASIECZKA G, QUADFASEL T, et al. Resonant anomaly detection without background sculpting[J]. Phys. Rev. D, 2023, 107(11): 114012. arXiv: 2210.14924 [hep-ph].
- [207] RAINE J A, KLEIN S, SENGUPTA D, et al. CURTAINS for your sliding window: Constructing unobserved regions by transforming adjacent intervals[J]. Front. Big Data, 2023, 6: 899345. arXiv: 2203.09470 [hep-ph].
- [208] BUHMANN E, EWEN C, KASIECZKA G, et al. Full phase space resonant anomaly detection[J]. Phys. Rev. D, 2024, 109(5): 055015. arXiv: 2310.06897 [hep-ph].
- [209] CHENG T, ARGUIN J F, LEISSNER-MARTIN J, et al. Variational autoencoders for anomalous jet tagging[J]. Phys. Rev. D, 2023, 107(1): 016002. arXiv: 2007.01850 [hep-ph].
- [210] ATLAS Collaboration. Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ TeV with the ATLAS detector[J]. Phys. Rev. Lett., 2024, 132(8): 081801. arXiv: 2307.01612 [hep-ex].

作者贡献 Author Contributions

作者在本论文中涉及的第二至第五部分中，具体贡献介绍如下。其中，涉及的相关论文编码对应下节中“论文工作”的编码。

在第二部分所涉及的前沿 DNN 算法的研究中，作者首先在早期参与了与微软亚洲研究院和 CERN 研究团队合作的 LorentzNet [6] 和 ParT [5] 算法开发的项目。这两个算法均对领域的发展起到推动作用。作者随后就 LorentzNet 和 ParT 表现出色的原因进行了细致研究，发现了 ParT 中提出的添加额外的对变量的好处实际是由遵守 Lorentz 对称性带来的；就此，提出了系统性研究 Lorentz 对称性对网络性能影响的思路，用实验揭示了遵守 Lorentz 对称性是一项喷注网络的重要的归纳偏置 [4]。这一发现也实现了将同期的多项与对称性网络有关的结论联系起来，对它们得以系统解释。论文在该部分前沿算法方面的阐述即按照作者的逻辑，综合作者主导、参与的上述进行梳理完成。

在第三部分所涉及的校准方法中，作者开创了利用 BDT 辅助筛选相空间更接近 $H \rightarrow b\bar{b}l c\bar{c}$ 的 $g \rightarrow b\bar{b}l c\bar{c}$ 代理喷注的方法，解决了此前方法无法校准前沿的 ParticleNet 喷注标记器的问题。作者随后继续将这一方法发展为更为通用的方法，可以对 CMS 实验中所有的 $X \rightarrow b\bar{b}l c\bar{c}$ 标记器在任意工作点进行校准，且准确度相较其他方法最高。它已成为 CMS 的标准技术，目前被包含 [1] 在内的近十项 CMS 分析所使用。在作者担任 CMS BTV（底夸克标记与次级顶点重建）组 Level-3 召集人期间，负责了 CMS 实验对 $X \rightarrow b\bar{b}l c\bar{c}$ 标记算法的三种校准技术的总结 [2]，其中包含这一新方法。

在第四部分所涉及的将前沿的 ParticleNet 喷注标记器用于 Higgs 和粲夸克对耦合的测量工作 [1] 中，作者参与了由 CERN 合作者主导的分析，对 ParticleNet $H \rightarrow c\bar{c}$ 标记器的校准做出重要贡献。作者也在分析 cutflow、拟合的优化方面做出贡献。作者在合作组中代表研究团队进行了 Pre-approval 报告。

在第五部分所涉及的未来的喷注标记的发展中，作者开创了“大模型、多分类”的喷注训练方法，引领团队完成了第一个部署于 LHC 实验中的预训练喷注模型，GloParT。这一模型在合作组经过两轮迭代，在本论文中展示了其最新版本。鉴于该喷注标记模型的通识的鉴别能力、最优的鉴别效果，即微调模型所带来的全新可能，它将作为 Run-3 阶段喷注标记的新基准，预期为后续物理分析的敏感度提升带来重要的帮助。作为可以公开参阅的结果，作者也带领团队基于 DELPHES 样本重现了这一方法 [3]，完成了 Sophon 模型并公开了全新的 JETCLASS-II 数据集，论证了它能够加速 LHC 上对新物理共振态的寻找的潜力。

① An English edition of the author contributions follows below.

The author’s detailed contributions in Parts II–V of this dissertation are outlined as follows. Relevant references correspond to the indices provided in the subsequent section, *Research Work*.

In Part II, which focuses on advanced deep neural network algorithms, the author initially participated in the development of the LorentzNet [6] and ParT [5] algorithms through collaborations with Microsoft Research Asia and the CERN research team. These algorithms marked significant advancements in the field. Building on this work, the author conducted an in-depth analysis of the exceptional performance exhibited by LorentzNet and ParT, revealing that the improvements observed in ParT brought by the pairwise features resulted from the enforcement of Lorentz-symmetry preservation. Motivated by this insight, the author proposed a systematic approach to examine the impact of Lorentz symmetry on network performance. Experimental results demonstrated that respecting Lorentz symmetry serves as a critical inductive bias for jet tagging network [4]. This finding provides a unified framework to explain several recent results concerning symmetry networks. The discussion of these state-of-the-art algorithms is structured according to the author’s logical framework, integrating prior works to which the author played a leading role and contributed.

In Part III, which focuses on the calibration methods, the author pioneered the application of a boosted decision tree to select phase-space regions closely resembling $H \rightarrow b\bar{b}/c\bar{c}$ jets using $g \rightarrow b\bar{b}/c\bar{c}$ proxy jets. This approach, named *the sfBDT method*, overcame previous limitations that have impeded the calibration of the then state-of-the-art ParticleNet jet tagger in CMS analyses. Building upon this, the author developed a generalized framework capable of calibrating all $X \rightarrow b\bar{b}/c\bar{c}$ taggers within the CMS experiment, covering arbitrary working points and achieving better accuracy compared to existing methods. This technique has since been adopted widely and implemented in nearly a dozen CMS analyses, including [1]. During the author’s service as the Level-3 convenor of the CMS BTV (b -tagging and vertexing) Performance group, the author leads the effort to present a comprehensive review of three calibration approaches for $X \rightarrow b\bar{b}/c\bar{c}$ tagging algorithms within CMS [2], which included this novel method.

In Part IV, which involves using the then state-of-the-art ParticleNet jet tagger for measurements of Higgs-charm Yukawa coupling [1], the author contributed to the CERN-led analysis. Specifically, the author played a pivotal role in the calibration of the novel ParticleNet $H \rightarrow c\bar{c}$ tagger and made substantial contributions to the optimization of the cutflow and fit procedures. The author delivered the pre-approval presentation on behalf of the research team

within the collaboration.

In Part V, which discusses the research on future jet tagger developments, the author pioneered the “large model for large-scale classification” method for jet tagging. Leading the team, the author developed the first Transformer-based, pre-trained, and globally applicable jet tagging model for the LHC experiments, *GloParT*. Following two iterative improvements within the team, the latest version of this model (the GloParT stage-2 model) is presented in this dissertation. Due to its general discrimination capability, optimal performance, and novel opportunities enabled through fine-tuning, GloParT is poised to serve as the new baseline for jet tagging during the Run-3 era, with significant potential to enhance the sensitivity of future physics analyses. In addition, to promote open access to GloParT’s developing techniques, the author led efforts to reproduce this method using DELPHES samples [3], resulting in the *Sophon model*, and released the *JetClass-II dataset*. This work demonstrates the model’s potential to accelerate searches for new physics resonances at the LHC.

Research Work

Experimental publications with significant contributions

- [1] CMS Collaboration, “Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Phys. Rev. Lett.* **131**, 061801 (2023), arXiv: 2205.05550 [hep-ex].
- [2] CMS Collaboration, “Performance of heavy-flavour jet identification in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV”, CMS Physics Analysis Summary CMS-PAS-BTV-22-001, 2022.

Phenomenological research

- [3] C. Li, *et al.*, “Accelerating resonance searches via signature-oriented pre-training”, arXiv: 2405.12972 [hep-ph].
- [4] C. Li, *et al.*, “Does Lorentz-symmetric design boost network performance in jet physics?”, *Phys. Rev. D* **109**, 056003 (2024), arXiv: 2208.07814 [hep-ph].
- [5] H. Qu, C. Li, and S. Qian, “Particle Transformer for Jet Tagging”, in *International Conference on Machine Learning*, pp. 18281-18292. PMLR, 2022, arXiv: 2202.03772 [hep-ph].
- [6] S. Gong, *et al.*, “An efficient Lorentz equivariant graph neural network for jet tagging”, *JHEP* **07**, 030 (2022), arXiv: 2201.08187 [hep-ph].

- [7] C. Li, *et al.*, “Loop-induced ZZ production at the LHC: An improved description by matrix-element matching”, *Phys. Rev. D* **102**, 116003 (2020), arXiv: 2006.12860 [hep-ph].

Conference Presentations and Posters

Domestic conferences

- **27th Mini-workshop on the frontier of LHC**, Zhuhai Jan. 2024
Presentation: Cutting-edge jet networks and their impact on LHC hadronic-channel searches
- **17th China TeV Physics Seminar**, Nanjing Dec. 2023
Presentation: CMS new physics search highlights—A lens of recent experimental innovations
- **CLHCP 2023**, Shanghai Nov. 2023
Presentation: Performance of heavy flavour jet identification in boosted topologies with full Run-2 data
- **CLHCP 2022**, Nanjing Nov. 2022
Presentation (virtual): Jet tagging in CMS
- **Higgs Potential 2022**, Beijing Jul. 2022
Presentation (virtual): Constraining the Higgs-charm coupling at CMS
- **CLHCP 2021**, Nanjing & Beijing Nov. 2021
Presentation (virtual): Deep tagger and H_{cc} measurement (CMS)
- **CLHCP 2021**, Nanjing & Beijing Nov. 2021
Presentation (virtual): Cocurrent generator support for CMS computing
- **CLHCP 2020**, Beijing Nov. 2020
Presentation (virtual): Vector-boson scattering in the $ZZ(4\ell)jj$ final state with full Run-2 data

International conferences

- **ML4Jets 2023**, Hamburg Nov. 2023
Presentation: Particle Transformer with built-in IRC safety
- **BOOST 2023**, Berkeley, USA Jul. 2023
Presentation: Performance of heavy flavour jet identification in boosted topologies in CMS 13 TeV data

- **LHCC students poster session, Geneva** Nov. 2022
Poster: Search for Higgs boson decay to a charm quark-antiquark pair in p - p collisions at $\sqrt{s} = 13$ TeV with CMS
- **ML4Jets 2022, New Brunswick, USA** Nov. 2022
Presentation (virtual): Does Lorentz-symmetric design boost network performance in jet physics?
- **BOOST 2022, Hamburg** Aug. 2022
Presentation: Jet tagging algorithm respecting Lorentz group symmetry
- **LHCP 2022, Taipei** May 2022
Presentation (virtual): Boosted algorithms for searches
- **ML4Jets 2021, Heidelberg** Jul. 2021
Presentation (virtual): Boosted jet tagging in CMS

Invited talks

- **Experimental Physics Division Seminar, IHEP, Beijing** Sept. 2022
Invited talk (virtual): Jet tagging algorithm respecting Lorentz group symmetry

致谢 Acknowledgement

五年的时光转瞬即逝。蓦然回首，三年半的孜孜不倦，一年半的内心波澜和从心所欲，可以大抵概括这五年的内心写照。与其作别之时，无奈行笔匆匆而成就了拙作，未能精细雕琢，属实遗憾，还望各位读者海涵。

回望五年的求学和科研之路，我想首先向我的导师李强老师表达由衷的感谢。李老师为包括我在内的每一位同学创造了尤为健康、舒适、清朗的科研环境——关注每一位同学的学业、研究乃至生活上遇到的困难，真诚而公正地对待每一位同学的意见和想法，耐心地辅导每一位同学取得学术进步。特别是，李老师为我创造了尤为宽松的学术环境，将我培养为一个有思想、有品味的独立科研工作者。李老师在学术上前瞻的眼光、敏捷的思维和理解力常令我叹服，其尊重个性、挖掘潜力的指导方式也常令我感动和钦佩，为我树立了终生的学术榜样。

我想向所有 CMS 合作组中帮助过我的同事表达由衷的感谢。感谢我一直以来的合作者、朋友兼导师曲慧麟博士：感谢您在我入门 CMS 实验和深度学习技术的诸多方面给予我耐心的学术指导和生活上的帮助。您为我树立了一个真正具有理想主义情怀的科学家榜样。感谢长期指导我的 Loukas Gouskos 博士，在 Higgs 物理的实验分析中给予我莫大的指点。感谢曾在学术上深刻影响了我的同事和伙伴：感谢 Michele Selvaggi 博士、Santeri Laurila 博士在 VHcc 课题上给予我巨大的帮助；感谢 Stephane Cooperstein 博士、Raffaele Gerosa 博士和 Alexandre De Moor、钱思天同学在诸多有关深度学习技术的讨论上提供的指点和建议；感谢 Javier Duarte 老师、Raghav Kansal 博士、Cristina Mantilla Suarez 博士在诸多课题上的合作与讨论、鼓励与支持；感谢组里的师兄师姐和同仁，Antonis Agapitos、Andrew Levin、卢梦、陈城、吕旭东、宋昀轩、安莹、肖杰、郭启隆、彭靖、关喆、钱思天、邓森、赵宇哲、李彦谷、杨天一等等诸位同仁，在诸多过往讨论中予以我莫大的学术帮助和精神鼓励。感谢各位曾经指点过我的老师们：冒亚军老师、孙小虎老师、周辰老师，与您的思维碰撞让我进步良多。

我想对包括我的导师在内的诸多 CMS 合作者表达由衷的感谢，您教会了我如何参与大型实验项目中，如何更高效地合作和沟通，如何融入到集体并接纳彼此。感谢李强老师，曾在 GEN 组合作并指导过我的 Alexander Grohsjean 博士、卢梦博士，曾在 BTV 组共事的邓森、Svenja Diekmann、李明晏同学、Jan van der Linden 博士、Spandan Mondal 博士、Alexandre De Moor、Sebastian Wuchterl 博士：感谢各位在我面对困难时

① An English edition of the acknowledgement follows below.

的无私帮助和支持，更感谢您对我性格的包容和照顾。非常荣幸在 CMS 组中遇见各位，与大家成为一起前进的伙伴。

我想感谢所有在博士期间曾经合作过课题的同事和同学。感谢孟凡强同学，与你早期的合作和激情澎湃的讨论至今仍带给我充满感动的回忆；感谢付大为和邓森同学，与各位合作中学习到的一丝不苟的精神使我充满振奋；感谢李静舒同学在合作中对我的包容。最后也要感谢在博士的末期一起合作的师弟：高乐耘、卢浩楠、张镇轩、寇子勋、赵宇哲同学，与大家的思维碰撞使我进步良多。

我想感谢北京大学这所园子，她以自由的学术氛围和开放包容的胸怀，在过去的九年里始终感化着我，培养了我的学术情操，并小心翼翼地保护着我的学术热忱——感谢在我过去九年的生命里，在这所园子遇到的所有老师们，负责行政的老师、辅导员们，同仁以及亲爱的同学们。感谢国家留学基金委在我读博期间的慷慨的资金支持，资助我在 CERN 完成一年半的学术工作，使得与国外同行的密切交流合作得以实现。

我想感谢身边的老师、同事和同学对我长期以来的呵护、鼓励和包容。感谢林海芑、刘越、应同和张林枫同学，在近五年来，尤其是疫情期间的陪伴和鼓励。感谢张子扬同学长期的支持与远程陪伴。感谢与曾经的室友宋昀轩博士在多个日夜激情的学术讨论和思维碰撞。感谢在 CERN 学习工作期间的室友，韩婧宜博士、刘锦枫博士、秦雪龙博士、张铭滔同学、赵永柯博士，在日常生活中的给予我的照顾和包容。

最后，我想由衷地感谢潘雁楠同学：十年来你的温暖陪伴和鼓励是我能够勇敢且顺利地越过一个个困难的坚强后盾，是我在生活的激流中始终勇往直前的源动力。我还要尤为感谢我的父母长久以来的养育和陪伴，没有您们，我无以在知识的苦海里遨游至此，摘得今天的果实。

我还想藉此机会，感谢自己对理想的执着。这份执着或许早早地扎根于少年时期对物理的热爱，或许诞生自高考后的某天阅读《最后的问题》的强烈使命感，又或许来自《朝闻道》中猛然收获的认同。我必须向这些少年时期的好友、师长、世界上素未谋面的人们表示谢意。诚然，这世间科研与合作总是充满困难或暗藏荆棘。但是，这份执着一直默默帮助你挺过一切——感谢这个世界，它还在暗中支持你向前。

新的转机和闪闪的星斗，
正在缀满没有遮拦的天空，
那是五千年的象形文字，
那是未来人们凝视的眼睛。（摘自北岛诗《回答》）

继续前进吧。祝你在五十年后回望你的一生，依然能够说出：虽千万人，吾往矣。

李聪乔

2024 年 6 月于燕园

Five years have flown by in the blink of an eye. As I look back, these years can be roughly summed up as three and a half years of relentless dedication, followed by a year and a half of introspection and gradual liberation. As I prepare to move forward, I cannot help but feel a sense of regret that my work, completed under time constraints, may lack the refinement it deserves. I kindly ask for the understanding of my readers.

Reflecting on my five-year journey of study and research, I would like to begin by expressing my deepest gratitude to my advisor, Prof. Qiang Li. Qiang cultivated a supportive, healthy, and inspiring research environment for all his students, including myself. He showed tireless dedication to our academic progress, research challenges, and even personal struggles, always treating our ideas and opinions with respect and fairness. Under his guidance, I have grown into an independent researcher with my own thoughts and aspirations. His visionary academic insights, sharp thinking, and ability to nurture individual potential have always been a source of admiration for me. Prof. Li shows the academic role model I aspire to follow throughout my career.

I am deeply grateful to all my colleagues in the CMS collaboration who have supported me along the way. My heartfelt thanks go to my mentor, collaborator, and friend, Dr. Huilin Qu, who guided me patiently as I started my journey into the CMS experiment and deep learning. You have shown me what it means to be a scientist driven by true ideals and passion. I am profoundly thankful to my mentor during my research at CERN, Dr. Loukas Gouskos, for his invaluable guidance on Higgs physics. My sincere thank goes to Dr. Santeri Laurila and Dr. Michele Selvaggi for their invaluable advice and support in the VHcc research. I would also like to thank Dr. Stephane Cooperstein, Dr. Raffaele Gerosa, Alexandre De Moor, and Sitian Qian, for their guidance and insights during our early collaborations in deep learning topics. My gratitude further extends to Prof. Javier Duarte, Dr. Raghav Kansal, and Dr. Cristina Mantilla Suarez for their collaborations, encouragement, and great support across various research subjects. To my peers and colleagues in the Peking group—Antonis Agapitos, Ying An, Cheng Chen, Sen Deng, Zhe Guan, Qilong Guo, Andrew Levin, Yangu Li, Meng Lu, Xudong Lyu, Jing Peng, Sitian Qian, Yunxuan Song, Jie Xiao, Tianyi Yang, Yuzhe Zhao, and many others—thank you for your academic support and encouragement during countless discussions. I would also like to express my gratitude to Prof. Yajun Mao, Prof. Xiaohu Sun, and Prof. Chen Zhou for inspiring me through thought-provoking conversations.

My gratitude extends to all the CMS collaborators who taught me how to contribute to large-scale experiments, collaborate effectively, and integrate within a team. I want to give

special thanks to Alexander Grohsjean, Meng Lu, and my colleagues from the CMS BTV team—Sen Deng, Svenja Diekmann, Ming-Yan Lee, Jan van der Linden, Spandan Mondal, Alexandre De Moor, and Sebastian Wuchterl—for their generous support and patience. It has been an honour to work alongside all of you and to grow together as a team.

I would also like to thank the colleagues and classmates with whom I collaborated on various projects during my Ph.D. To Fanqiang Meng—our early collaborations and passionate discussions remain a treasured memory. To Sen Deng and Dawei Fu—working with you taught me the value of diligence and precision. To Jingshu Li—thank you for your understanding, patience, and support during our teamwork. I am also grateful to my younger colleagues—Leyun Gao, Zixun Kou, Haonan Lu, Zhenxuan Zhang, and Yuzhe Zhao—for the enriching idea exchanges that pushed me to grow further.

I owe my deepest gratitude to Peking University. Over the past nine years, its open, inclusive, and vibrant academic atmosphere has significantly shaped my character and nurtured my passion for knowledge—I am thankful to all the professors, administrative staff, mentors, and classmates I have had the privilege to meet here. I would also like to express my thanks to the China Scholarship Council for their financial support, which enabled me to conduct my research at CERN and engage closely with international collaborators.

I am immensely grateful to my friends and peers who supported me throughout these years. To Haipeng Lin, Yue Liu, Tong Ying, and Linfeng Zhang—your encouragement, especially during the pandemic, meant so much to me. To Ziyang Zhang—thank you for your unwavering support and care from afar. I also deeply treasure the life spent with my former roommate, Yunxuan Song, and the warmth and companionship of my roommates at CERN—Jingyi Han, Jinfeng Liu, Xuelong Qin, Mingtao Zhang, and Yongke Zhao—who made life easier and brighter with their kindness.

Finally, my heartfelt thanks go to Yannan Pan. Your warmth, encouragement, and unwavering support have been my greatest source of strength over the past ten years, giving me the courage and determination to face every challenge. To my parents, whose unconditional love and nurturing have shaped who I am today, I owe everything—without you, I could never have reached this point in my life.

I would also like to take this opportunity to express my deep gratitude for the longstanding dedication to my ideals. This persistence may have taken root in my youth through a love of physics, emerged on that very day after the Gaokao when I encountered *The Last Question* and felt an overwhelming sense of purpose, or arisen from a profound resonance with *Death*

for Truth (Zhao Wen Dao). I am indebted to my childhood friends, mentors, and strangers I have never met, whose presence has quietly shaped my life. Indeed, the journey of scientific research and collaboration is often fraught with challenges and hidden obstacles. Yet, this very persistence has quietly carried me through adversity. For this, I extend my heartfelt gratitude to the world, which continues to offer its silent support as I press forward.

Forge ahead. May you look back fifty years from now and still declare with the golden conviction.

The vast unobstructed firmament

Is lit with glimmering stars;

They are a new turning:

These hieroglyphs of the ages,

These staring eyes of the future.

(Excerpt from Bei Dao's poem *The Answer*)

Congqiao Li

June 2024, at Yanyuan

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：李聪奇 日期：2024年6月19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；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校☐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：李聪奇 导师签名：李欣

日期：2024年6月19日



学位论文答辩委员会名单

论文题目	大半径喷注标记的现代深度学习技术——算法、校准方法及CMS实验中的应用			
作 者	李聪乔			
专 业	粒子物理与原子核物理			
答辩委员会成员	姓 名	专业技术职称	从事专业	工作单位
主席	何红建	教授	理论物理	清华大学
委员	班勇	教授	粒子物理与原子核物理	北京大学
	陈明水	研究员	粒子物理与原子核物理	中科院高能物理研究所
	陈新	副教授	粒子物理与原子核物理	清华大学
	冒亚军	教授	粒子物理与原子核物理	北京大学
	孙小虎	研究员	粒子物理与原子核物理	北京大学
	周辰	助理教授	粒子物理与原子核物理	北京大学

北京大学博士学位论文答辩委员会决议书

(本表由博士答辩委员会秘书填写,一式两份,一份存学校档案,一份存研究生个人档案)

院、系: 物理学院

专 业: 粒子物理与原子核物理

姓 名: 李聪乔

研究方向: 高能物理

学 号: 1901112050

导师姓名: 李强

学位论文题目: 大半径喷注标记的现代深度学习技术——算法、校准方法及CMS实验中的应

答辩委员会对学位论文和答辩情况的学术评语 (主要就论文选题意义, 创新性成果及学术水平; 论文存在的主要不足之处, 以及博士生答辩情况等方面)。

人工智能技术在高能物理领域的设计和应用是当前前沿热点课题之一。李聪乔的博士论文, 研究了神经网络的深度学习喷注标记技术及其在LHC上希格斯性质测量的应用。选题具有重要物理意义。

该论文在高能物理神经网络技术发展和物理应用上做出了一批原创成果, 包括基于图神经网络的深度学习喷注标记技术和校准方案的开发应用, 验证了洛伦兹对称性在神经网络中的作用, 以及基于多达188喷注分类的大神经网络的迁移学习方法的建立。该论文进而利用CMS探测器所收集的13 TeV质子-质子对撞数据, 通过希格斯与W、Z规范玻色子的联合产生过程, 对希格斯的粲夸克衰变道进行了探寻。利用开发的图神经网络喷注标记技术应用于大动量希格斯的探测, 大幅提高了探测灵敏度。最终, 团队得以在强子对撞机上首次观测到Z玻色子到粲夸克衰变过程, 并获得了希格斯与粲夸克汤川耦合的当前最强实验限制。

该博士论文写作规范, 结构清晰, 研究分析过程严谨, 文字表达准确。作者对粒子物理基础理论知识及实验分析技术掌握扎实, 具备了独立从事科研工作的能力。

在答辩过程中, 答辩人表达清楚, 回答问题正确。答辩委员会一致通过论文答辩, 认为这是一篇优秀的博士论文, 建议授予理学博士学位。

答辩委员会表决结果: 实到答辩委员: 7 人,

同意博士毕业者 7 人, 不同意博士毕业者 0 人, 结论为 准予毕业

同意建议授予博士学位者 7 人, 不同意建议授予博士学位者 0 人, 结论为 建议授予学位

答辩时间: 2024 年 05 月 24 日 8 时

答辩委员会主席(签字): 何红建

委员(签字): 孙晓 陈新 周辰 孙宇 何宇 陈明



提交终版学位论文承诺书

本人郑重承诺，所呈交的学位论文为最终版学位论文。本人知晓，该版学位论文将用于校学位评定委员会审议学位、国家和北京市学位论文抽检。论文版本呈交错误带来的结果将由本人承担。

论文作者签名：李锐乔

日期：2024年 6 月 19 日

