TWO B, AND THEN ANOTHER TWO B... BUT HOW TO MLEARN IT BBBBEST – THAT'S THE QUESTION.

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHYSICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nicole Hartman
September 2022

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Su Dong) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael Kagan)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Patricia Burchat)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael Peskin)

Approved for the Stanford University Committee on Graduate Studies

Preface

There's something special about the human race that drives us not just to exist but to understand the reason for our existence. A part of the mechanistic answer to this question comes from a desire to understand what are the fundamental building blocks of nature. As we moved away from the earth, fire, air, water understanding of the fundamental elements to a more modular undersanding of matter's components, in 400 BC, Democritus thought that he "had" it with the "elements" and he coined the word "atoms" to describe what we now understand as the periodic table of the elements. Atom means "uncuttable" a now hilarious misnomer - as the systematic structure of the periodic table hinted at some underlying components, and the experimentally verified when Thomas X saw the electrons popping off of the atom and we came to understand the The evolution / progression of the scientific revolution has involved us studying ever smaller and smaller distance scales... and discovering a microcosm ever more and more intricate and facinating. In my PhD, I've been working at SLAC the momentous institute where the substructure of the proton was discovered, a discovery which cemented our current understanding of the parton model and the Standard Model of particle phyiscs. We keep reusing the word "elements" "elementary" because we are striving to get to these fundamental distance scales...

Something absolutely facinating to me is the range of masses that we have in the SM that we understand to be entirely pointlike. We currently believe that both the top quark 175 GeV down to the ¡ meV neutrino are entirely point-like objects, and it seems absolutely fascinating to me this 9 decade range of energies can similarly be packed into an infinitesimally small distance. The Higgs is the answer to this mass generation - and therefore studying it and it's properties gives us a clue to this mystery.

This thesis submission marks a decade since the Higgs boson discovery, and we expect to observe the HH process in another 10 years of LHC data taking. And although the Higgs mechaanism is now an old theory, with the massively large datasets we're collecting at the LHC, we can unlock new ways of answering these questions with the big data developments in the burgeoning field of deep learning. This thesis explores the properties of the Higgs through the lens of the Run 2 dataset in this journey of understanding the microscopic realm – a quest that continues to remain interesting, for a whole new generation of scientists to uncover and as we continue to ask the question... what will be next.

Acknowledgments

Every domain of human endeavor is held together by a web of relationships between people. Real people. That web is the fabric that undergirds, contains, and holds together that part of society.

– Bill Burnett & Dave Evans, *Designing Your Life.*

I would like to thank...

- Michael
- Rafael
- Max Swiatlowski
- HH-¿4b squad: Dale Abott, Sean G, Lukas Borgna
- Other NR analyzers: James Grundy, Rui Zhang,
- FTAG / CP people: C Pollard, Franscesco de Bello, Jonathan Shalomi, Manuel Guth, Bing, Dan Guest, Binbin
- Danny Antrim, Kathryn, Johan, Matthew Feickert
- Steve, Jodi, Richard Scaletter, Emily Thompson
- Aviv
- Stats help: Lukas, Giordan
- Valentina and PF
- Jannicke
- SLAC ATLAS mentors: Caterina, Aeriel, Su Dong, Charlie
- Katharine + Liza: Making the best ATLAS sub-group in the world (supposedly a cult)

- HH Party planning committee
- HHonorable mentions: First tequila shot (CV) second tequila shot (MK)
- Amazing group collaborators: Maxime and Yoann
- Reading / exam committee

Contents

List of Tables

List of Figures

1

Introduction

Until taught by prolonged exposure that the universe contained anomalous cards, they saw only the types of cards for which previous experience had equipped them. Yet once experience had provided the requisite additional categories, they were able to see all anomalous cards on the first inspection long enough to permit any identification at all . . .

– Thomas Khun, *The Structure of Scientific Revolutions*

- Why we expect new physics in the SM
- Motivation for HH
- Particulars about the 4b final state
- Connection to ML
- Highlight my work on b-tagging
- Thesis organization

Refs:

- Higgs discovery: [1, 2].

2

Theoretical motivation

Surveying the rich experimental literature from which these examples are drawn makes one suspect that something like a paradigm is prerequisite to perception itself.

– Thomas Khun, *The Structure of Scientific Revolutions*

2.1 The Standard Model

SU(3) x SU(2) x U(1) gauge interactions

$$\mathcal{L}_{SM} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + \left(y_{ij}\bar{\psi}_i\psi_j + h.c.\right) + |D^\mu\phi|^2 - V(\phi) \tag{2.1}$$

2.2 The Higgs mechanism

I heard from Dale that the pdg is a v good ref for this!!

2.3 Effective Field Theories to search for new physics

2.4 Status of the experimental HH results

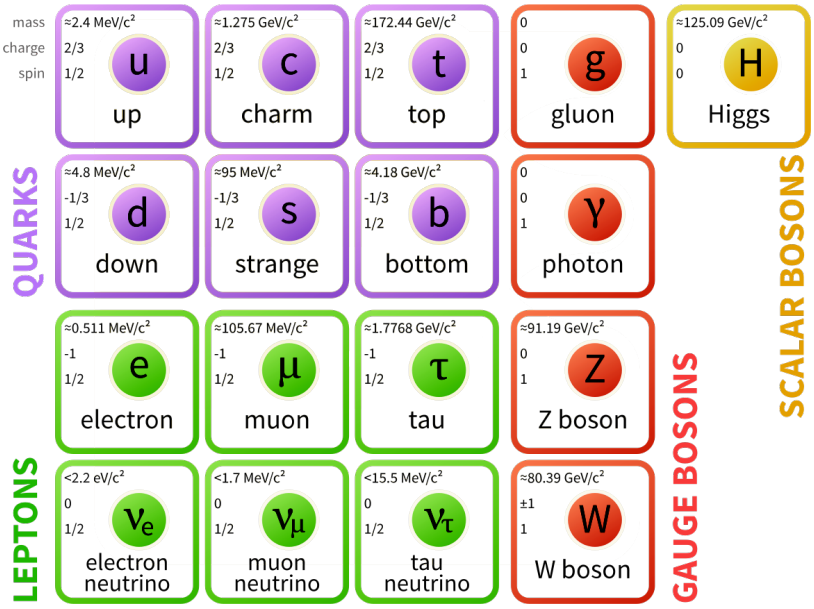


Figure 2.1: [**wiki-sm**]

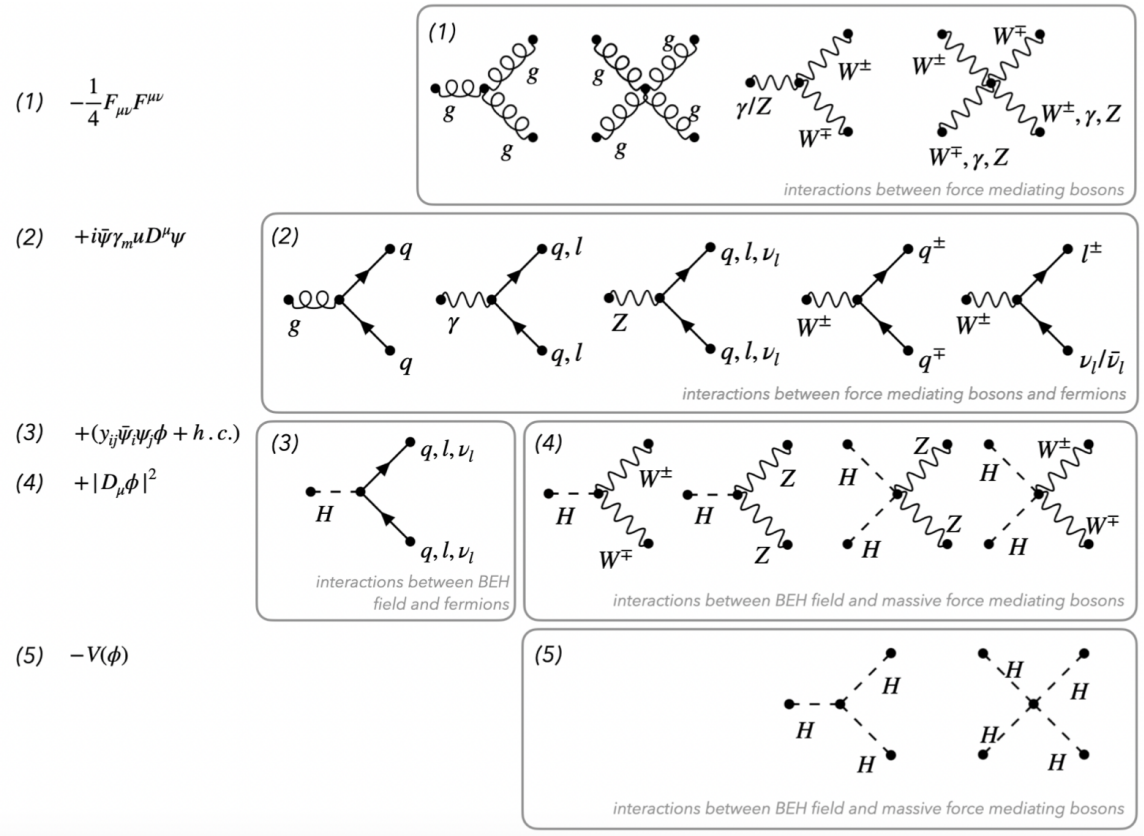


Figure 2.2: [**Melissa-thesis**]

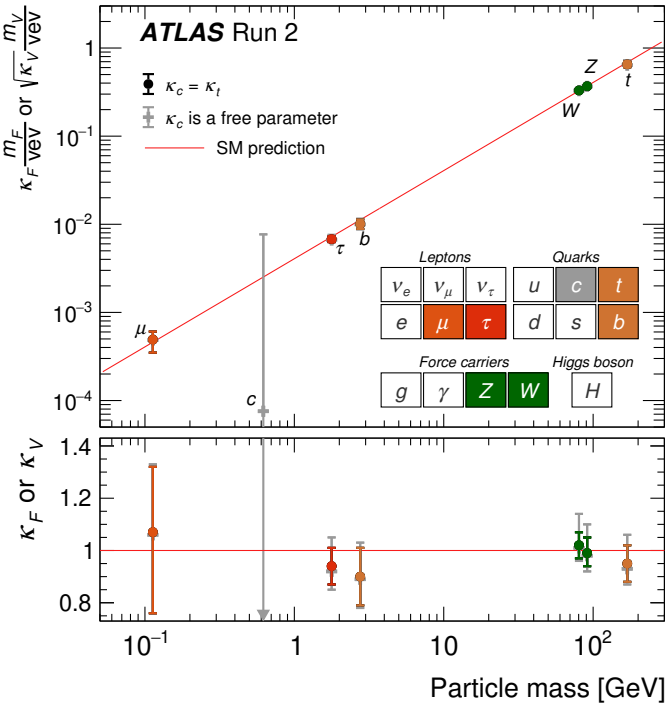


Figure 2.3: [**2207.00092**]

3

The Large Hadron Collider

Every act of creation begins with an act of destruction.

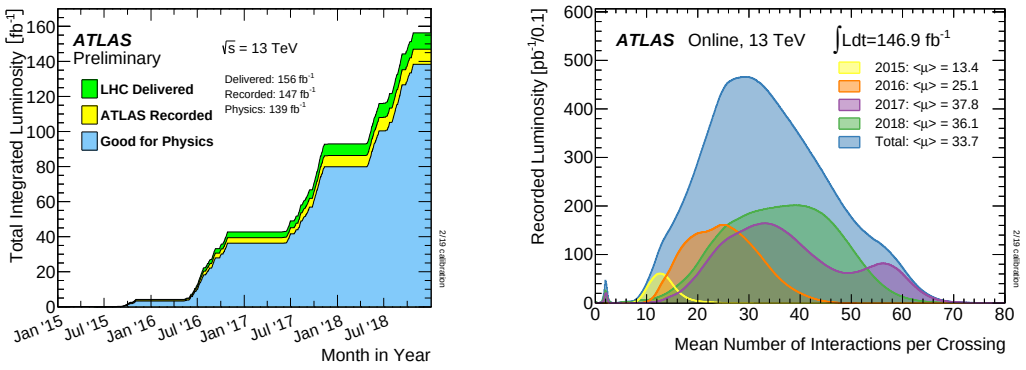
– Pablo Picasso

The LHC is a 27 km circumference particle accelerator straddling the border between France and Switzerland [**LHC**]. It now collides bunches of 10^9 protons every 25 ns at a center of mass energy of 13 TeV [**ATLAS'long**]. Bunches of protons are used since the proton radius is about 1 fm, and thus the probability of a head-on collision between single protons is increased by increasing the number of colliding particles. We expect about 20 collisions per given bunch crossing at the LHC [**PU**].

Since the proton is not an elementary particle, but rather a baryonic resonant QCD bound state, even a head-on collision will not have all of the available energy concentrated at a point. The proton is a composite particle made up of two up and one down valence quarks, along with a sea of virtual quarks and gluons that spontaneously come out of the vacuum because of the uncertainty relation $\Delta x \Delta p \geq \hbar/2$. The 6.5 TeV per proton is distributed among these partons[3]. Therefore, to look at the high-energy regime, we are interested in a "hard scatter" where a single quark or gluon carrying a large fraction of one proton's momentum collides with a high energy constituent of the other proton, lending to the production of one or more Higgs bosons.

- History
- Collider Stats
- I should make a set of slides (or app) making sure that I know what the dipole and focusing magnets do!!

[3] A parton is a constituent of a hadron.



(a) (b)

Figure 3.1: Need to cite: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2#Pileup_Inte

4

The ATLAS detector

I never watch the stars – there's so much down here.

– Lorde "Yellow Flicker Beat"

4.1 Overview

The ATLAS detector [3] at the LHC covers nearly the entire solid angle around the collision point.[1] It consists of an inner tracking detector surrounded by a thin superconducting solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer incorporating three large superconducting toroidal magnets. The inner detector system (ID) is immersed in a 2 T axial magnetic field and provides charged-particle tracking in the range $|\eta| < 2.5$. The high-granularity silicon pixel detector covers the vertex region and typically provides four measurements per track, the first hit being normally in the insertable B-layer (IBL) installed before Run 2 [4, 5]. It is followed by the silicon microstrip tracker (SCT) which usually provides eight measurements per track. These silicon detectors are complemented by the transition radiation tracker (TRT), which enables radially extended track reconstruction up to $|\eta| = 2.0$.[2]

A cutaway view of the ATLAS detector is shown in Figure ATLAS-detector. The inner detector (ID) provides charged particle tracking information closer to the interaction point. The inner detector is immersed in a 2 T solenoidal magnetic field to bend the trajectories of charged particles and allow for momentum measurement. Energy measurement is made in two parts: with an electromagnetic

[1] ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point in the centre of the detector and the z-axis along the beam pipe. The x-axis points from the interaction point to the centre of the LHC ring, and the y-axis points upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z-axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

[2] Text taken from the ATLAS approved detectors text section of the repo, but I'm just including the tracker info since I use hits as my variables.

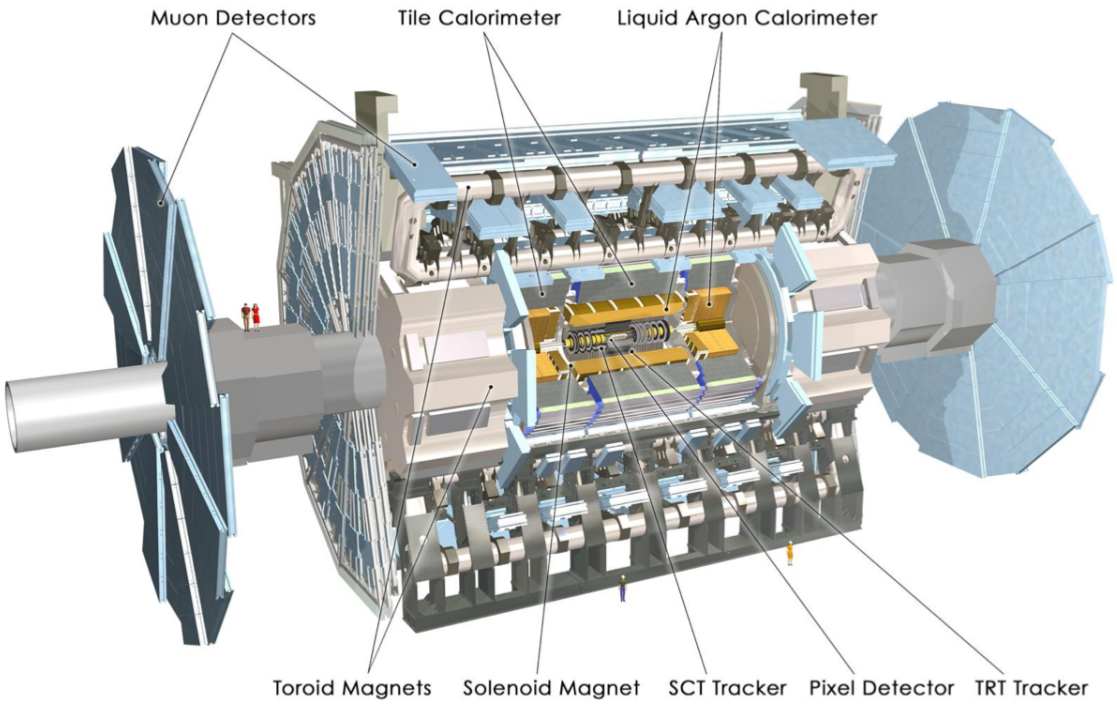


Figure 4.1: Cut away of the ATLAS detector [**ATLAS˙long**]

and hadronic calorimeter. Finally, the outermost layer of the detector is the muon spectrometer, with a 4 T toroidal magnetic field.

4.1.1 ATLAS Coordinate System

At the ATLAS detector, the z-axis is measured along the accelerator beam pipe, the x-axis points into the center of the ring, and y-axis is defined to point vertically up by the properties of a right-handed rectilinear coordinate system. Then a cylindrical coordinate system is used for the ATLAS detector. The azimuthal angle, $\phi = \arctan(\frac{y}{x})$, denotes orientation in the plane transverse to the beam, and the pseudorapidity, η, measures the polar angle inside the detector, where

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) \tag{4.1}$$

and where $\theta \in [0, \pi]$ is the polar angle as measured from the z-axis. The ATLAS detector is forward-backward symmetric to maximize the detector coverage since colliding protons each have the same energy.

From these detector variables, we can get components in the four-momentum using the relations

$$(E, p_x, p_y, p_z) = (E, p_T \cos\phi, p_T \sin\phi, p_T \sinh\eta) \tag{4.2}$$

$$p = p_T \cosh\eta \tag{4.3}$$

We can also get a measurement of the momentum in the calorimeter for relativistic particles. In the relativistic limit, the energy and momentum are the same (in natural units). However, when the momentum is defined from a calorimeter measurement, E_T is used instead of p_T. Equations (4.2) and (4.3) still apply in this case.

4.2 Tracker

4.2.1 Inner Detector

A cutaway view of the inner detector for ATLAS is shown in Figure ATLAS-InnerDetector.

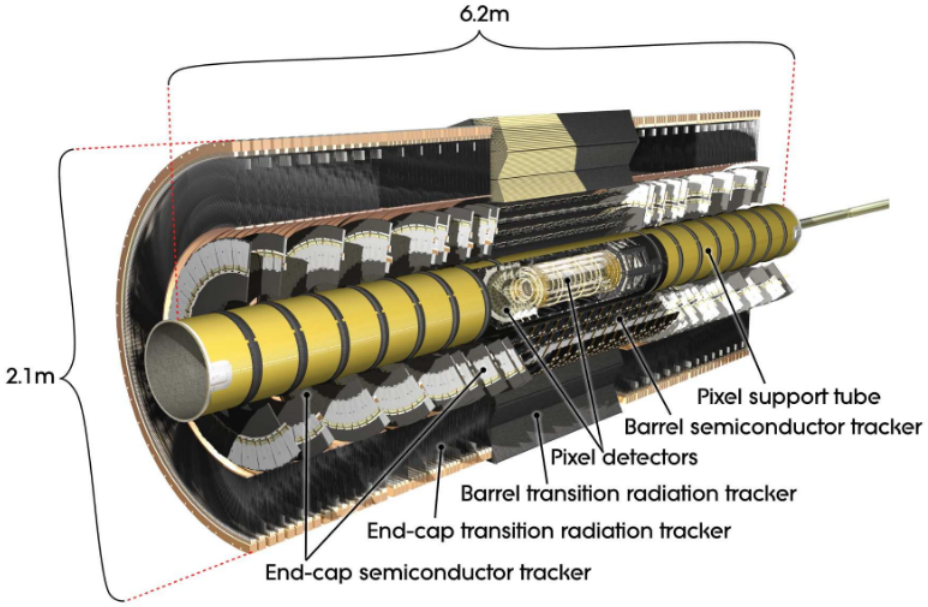


Figure 4.2: Illustration of the orientation of the subsystems inside the ATLAS Inner Detector [**Aad:2008Jinst**]

With approximately 1000 tracks emerging from the interaction point every 25 ns, the ID is divided into three different regions to optimize the pattern recognition and momentum measurement algorithms [**ATLAS˙long**]. The pixel detectors have the most precision, so this layer is closest to the interaction point. The pixel detector also has the highest cost, so the next least expensive tracking option is the silicon microstrip trackers (SCTs), at the next farthest region away from the interaction point. Finally, the Transition Radiation Trackers (TRTs) compose the outermost region of the inner detector. The ID uses pattern recognition to measure transverse momentum as low as 0.5 GeV, and provides electron identification for $|\eta| < 2.0$ for energies up to 150 GeV [**ATLAS˙long**]. Displaced multi-particle vertexes can be resolved by the pixel detector and help identify long-lived B-hadrons.

Item		**Radial extension (mm)**	**Length (mm)**
Overall ID envelope		$0 < R < 1150$	$0 < \|z\| < 3512$
Beam-pipe		$29 < R < 36$	
Pixel	Overall envelope	$45.5 < R < 242$	$0 < \|z\| < 3092$
3 cylindrical layers	Sensitive barrel	$50.5 < R < 122.5$	$0 < \|z\| < 400.5$
2×3 disks	Sensitive end-cap	$88.8 < R < 149.6$	$495 < \|z\| < 650$
SCT	Overall envelope	$255 < R < 549$ (barrel)	$0 < \|z\| < 805$
		$251 < R < 610$ (end-cap)	$810 < \|z\| < 2797$
4 cylindrical layers	Sensitive barrel	$299 < R < 514$	$0 < \|z\| < 749$
2×9 disks	Sensitive end-cap	$275 < R < 560$	$839 < \|z\| < 2735$
TRT	Overall envelope	$554 < R < 1082$ (barrel)	$0 < \|z\| < 780$
		$617 < R < 1106$ (end-cap)	$827 < \|z\| < 2744$
73 straw planes	Sensitive barrel	$563 < R < 1066$	$0 < \|z\| < 712$
160 straw planes	Sensitive end-cap	$644 < R < 1004$	$848 < \|z\| < 2710$

Figure 4.3: List of the dimensions of the subsystems in the ID [**ATLAS'long**].

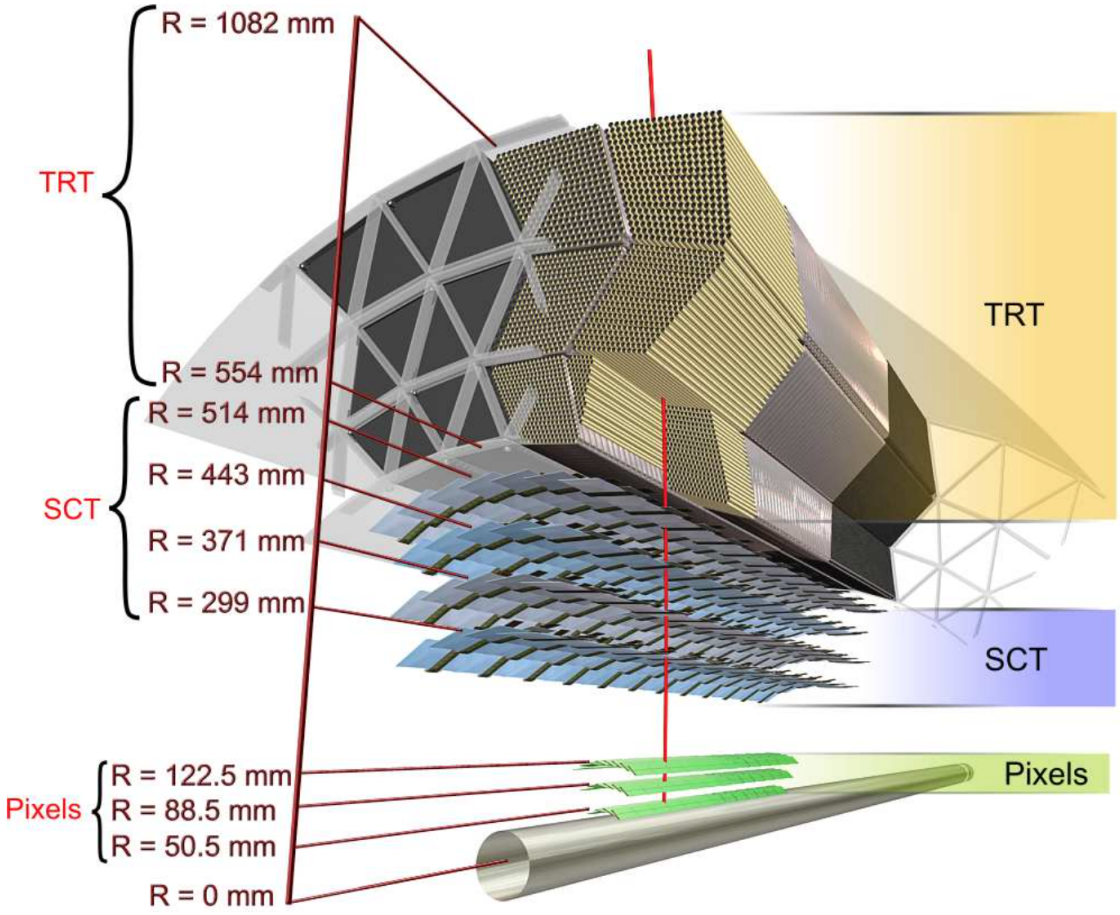


Figure 4.4: The structural elements in the ID at $\eta = 0.3$ [**ATLAS˙long**].

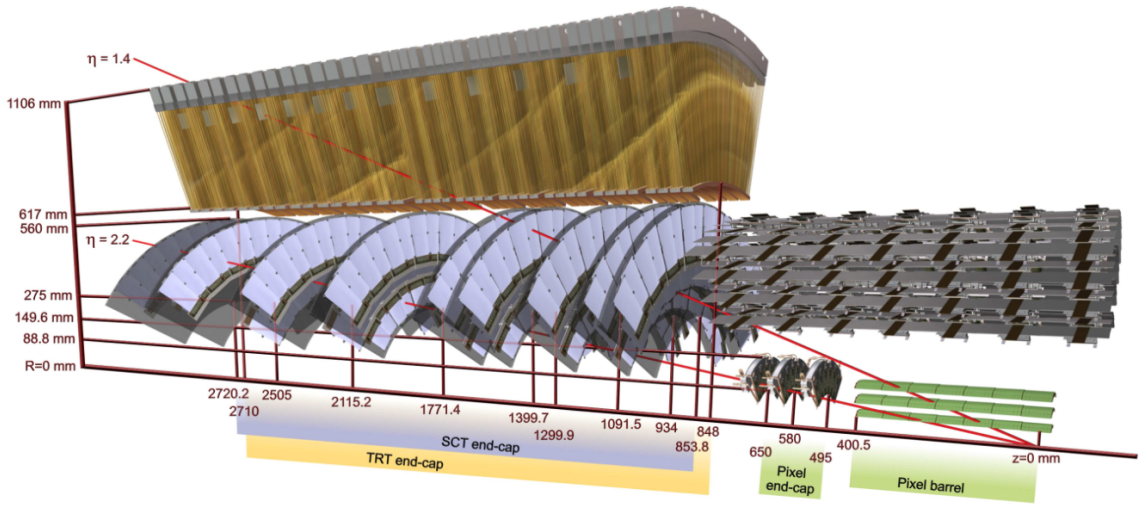


Figure 4.5: The structural elements in the ID at $\eta = 1.4$ [**ATLAS˙long**].

Pixel Tracking System

ATLAS is often referred to as a "giant camera." In a general sense, both a camera and the ATLAS detector provide ways to record specific events, but the parallels run deeper than their purposes, since both cameras and particle detectors can obtain information via pixels. In a camera, an incident photon will go through a silicon diode and knock a valence electron free to generate a few electron-hole pairs. An electric field is set up across the pixel and will then pull the electron and hole apart to the metal contacts where the charge can be read out [6]. The pixel detector at ATLAS works in a similar way, except the energy of the incident particles at ATLAS is orders of magnitude larger: while a photon of visible light has an energy of a few eV, typical "interesting" particles at the LHC are in the MeV to TeV range.

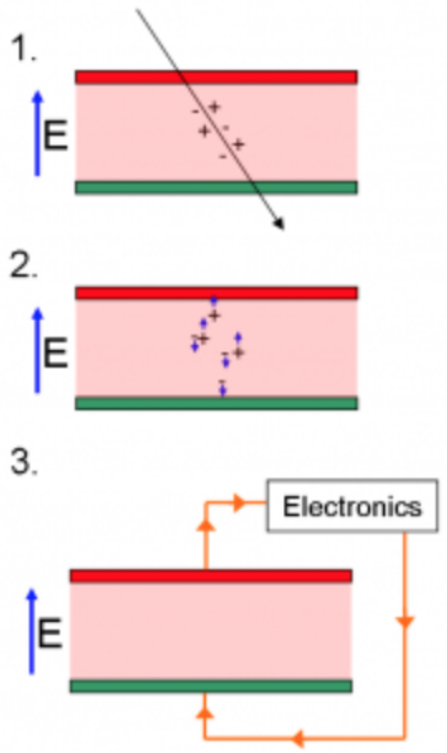


Figure 4.6: Schematic for the general idea for how a pixel detector tracks incident particles. [6]

As a particle passes through the pixel sensor, it creates electron-hole pairs that are separated by the electric field and read out by the electronics, as shown in Figure Pixel-cartoon.

The pixel sensors in ATLAS consist of 80 million [**Detector˙challenges**] rectangular 50 μm$\times$400 μm "n$^+$-in-n" electrodes.[3] The pixel detector's position with respect to the other components of the inner detector is shown in Figure ATLAS-InnerDetector. Zooming in on Figure ATLAS-InnerDetector, the innermost part of the inner detector—the pixel detector—is shown in Figure ATLAS-PixelDetector. The 80 million channels are divided into four cylindrical layers combined in a volume 1442 mm in length with 430 mm radius [**Aad:2008Jinst**]. The pixel detector has a resolution of 15 microns, and this precision is limited by the size of the electronics.

The pixel detector starts 5 cm from the center of

[3] The "n" region is n-type, or doped with atoms that are electron donors; the "in" region is intrinsic, or undoped; and the "n$^+$" region is also n-type, but doped more heavily than other n region.

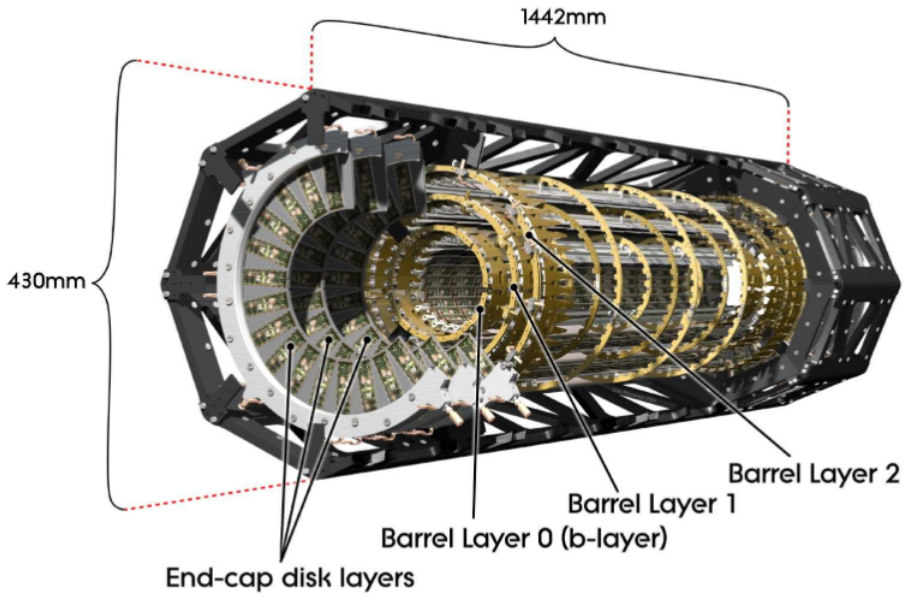


Figure 4.7: Illustration of the active region of the pixel detector with the barrel and endcap layers [**Aad:2008Jinst**]

the beam pipe to gain as much information about the central reaction as possible, and has coverage for $|\eta| < 2.5$ [**Detector˙challenges**]. Since the pixel detector is so close to the interaction point, its electronics must be able to withstand high radiation doses of up to 500 kGy. The p-n diodes can experience a leakage current when an electron-hole pair has enough energy to overcome the potential barrier. The detector minimizes the number of electron-hole pairs spontaneously generated by cooling the system at -6°C.

Silicon Microstrip Trackers

The Silicon Microstrip Trackers (SCTs) operate with a similar principle as the pixel detectors, but have an effective operating voltage determined by the effective doping level, and their leakage current also increases linearly with radiation dose. Initially the operating

voltage was 150 V, but increased up to 250 to 350 V to compensate for the radiation dose after 10 years. The 15912 SCT sensors each have a thickness of 285 ± 15 μm and a pitch of 80 μm.

Transition Radiation Tracker

The Transition Radiation Trackers (TRT) make up the outermost region of the ID, and the dimensions of the TRT are shown in Figure ID-table. "Transition Radiation" is the radiation emitted by a relativistic particle as it traverses an interface between two materials with different permittivities. Each TRT is a 4 mm polyimide drift tube, created by two 35 μm multi-layer films bonded back-to-back [**ATLAS'long**]. A 25 μm thick polyimide film has one side laminated with a 0.2 μm of Al with another 5-6 μm layer of graphite [**ATLAS'long**]. The inside of the tube is filled with a gas composed of 70% Xe, 27% CO_2, and 3% O_2. The anode is composed of 31 μm diameter cylinder of tungsten wire positioned at the center of the drift tube, and coated with 0.5–0.7 μm of gold. After fabrication, the tubes were cut to a 144 cm length for the barrel and 37 cm length for the end-cap region. The barrel straws are read out at each end of the tube, so the middle of the tube is insulated with a 6 mm glass layer which creates a 2 cm spot where the element is not sensitive to incident tracks. The anode is grounded and connected to the front end electronics, while the cathodes are held at -1530V. Minimizing mechanical sag in the straw is crucial for accounting for the error in the position measurements, so the straws are mechanically supported by carbon fibers [**ATLAS'long**].

An incident particle traversing a straw ionizes the gas molecules to free electrons and positive ions. The electrons accelerate through the electric field to the anode, and these electrons in turn will ionize other ions to create a gain of 2.5×10^4. Since the anode is read out at both ends of the wire, this provides a drift

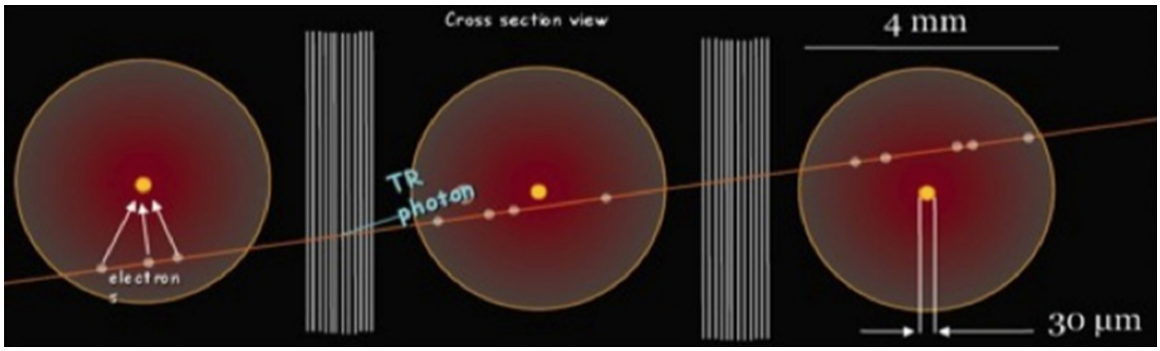


Figure 4.8: Principle of operation for the TRT [7]

time measurement from the difference of arrival times between the two ends of the wire. The time resolution is approximately a nanosecond, which corresponds to a spatial resolution of about 100 microns.

An outgoing particle will hit on average 36 of the TRT tubes; this improves the momentum measurement in the inner detector. During its lifetime, a straw detector can track 10^{15} particles, and a total integrated charge of 1000 C, which corresponds to about 20 years of the LHC's operation. The ATLAS inner detector has 12,000 of these straws in the endcaps and 52,544 straws in the barrel, yielding a total of 351,000 read–out channels. Although the silicon trackers need to be cooled between -5 to -10°C, the TRTs operate at room temperature.

4.3 Calorimeter

4.3.1 ECAL

Calorimetry literally means "heat measurement." The incident particle interacts with the material in the detector to form a shower of particles which are decelerated and absorbed in the detector material to reconstruct the original particle's energy. The ATLAS calorimeter is divided into two parts: the electromagnetic calorimeter (ECAL) and the hadronic

calorimeter (HCAL). The ECAL is for electromagnetic interactions and has higher precision than the HCAL which reconstructs the particles interacting by the strong force.

Sampling vs. Homogenous Calorimeters

For an electromagnetic shower to develop, incident electrons (and positrons) will emit a photon through Bremsstrahlung approximately in through a distance characterized by the radiation length

$$X_0 = \frac{716.4\ \mathrm{g \cdot cm^{-2}} A}{Z(Z+1)\ln\frac{287}{\sqrt{Z}}} \tag{4.4}$$

where A is the number of nucleons and Z is the atomic number (number of protons) in the detector material [**Calorimetry1**]. The units of X_0 means that dividing by the density gives the actual distance traveled by a particle [**Calorimetry2**]. Then the photons will in turn pair-produce electrons and positrons when traveling a distance of $\frac{9}{7}X_0$. This showering phenomena will continue until the average particle energy decreases to below the critical energy, $E_c = \frac{610\ \mathrm{MeV}}{Z+1.24}$. The hadronic showers are characterized by the interaction length, $\lambda = 37.8A^{0.312}\mathrm{g \cdot cm^{-2}}$, the average distance that an incident particle travels before undergoing a nuclear interaction.

Two different calorimeter designs can be used to measure the incident particle's energy. In a sampling calorimeter, the volume is divided into scattering and absorbing slabs, shown in Figure sampling-schematic. The scattering slabs use a high Z material to decrease the radiation length and allow the shower to develop more quickly. The energy deposited in the absorbing region of the calorimeter is measured, and the total energy is the energy deposited in the absorbing region divided by a scale factor, $f_{sampling} = E_{visible}/E_{deposited}$. A sampling calorimeter

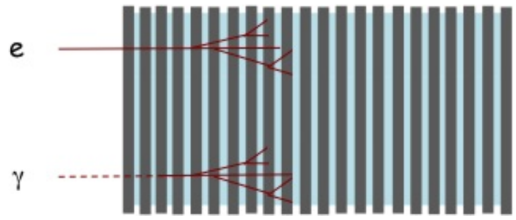


Figure 4.9: Schematic illustrating how a sampling calorimeter causes an incident particle to form a shower more quickly. [**Sampling˙Cal**]

contains the shower in a smaller detector to help minimize the cost of the experiment. One of the downsides of this procedure is that the sampling method is not as precise because only a fraction of the energy deposited is measured, and fluctuations proportional to $\sqrt{E}$ for Poisson statistics will introduce extra errors into the measurements.

A homogenous calorimeter circumvents this problem by using the whole calorimeter as an active volume. A homogenous calorimeter is only useful as an ECAL since hadronic interactions require more material to "contain," and it may not be monetarily feasible to construct such a large volume. To see this, we can look at the approximate formulas for the radiation length and interaction length, $X_0 \sim \frac{A}{Z^2}$, and $\lambda \sim A^{1/3}$. Since Z is approximately $\frac{1}{2}A$, this means $\frac{\lambda}{X_0} \sim A^{4/3}$, a number that can be as large as 30 for high Z materials such as lead, showing that hadronic showers have a larger extent.

Liquid Argon Detector

The ATLAS ECAL is a sampling calorimeter arranged in an accordion structure, as shown in Figure accordian-structure. The high Z material (Z = 82) lead creates the shower, while the energy is measured in liquid argon (LAr), a low Z material (Z = 18). As illustrated in Figure accordian-structure, the folding angle decreases as the radius (measured out from the interaction point) increases. The folding angle varies between 90° and 67° to keep the LAr sampling region width approximately constant at 2.1mm between the absorbers.

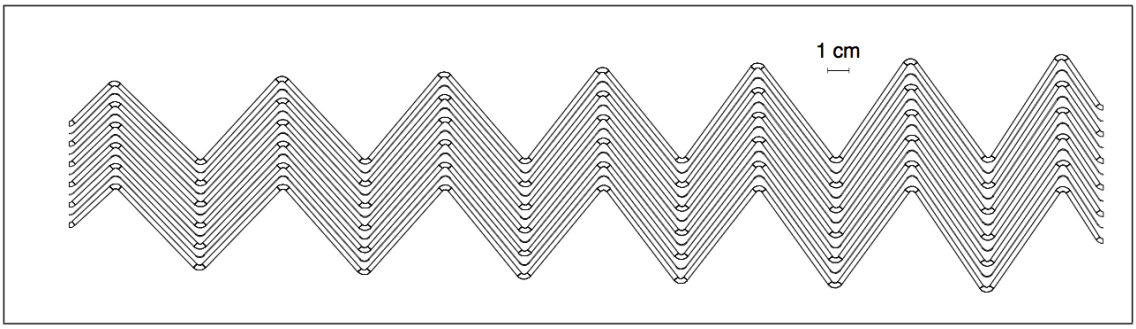


Figure 2-21 Transverse section through the active part of the barrel EM calorimeter (see also Figure 2-iii).

Figure 4.10: Accordian sampling structure for calorimeter. [**LAr**]

The LAr scintillators are read out with wavelength-shifting photosensors. This type of a detector lends itself naturally to a tower structure with the modules forming wedges pointing back to the interaction point, shown in Figure Wedge-LAR-accordian. This makes it easy for a particle's shower to be contained within a few modules or cells. The segmentation in $\Delta\eta \times \Delta\phi$ is 0.025×0.1 in the pre-sampler region, 0.0031×0.1 in the strips, 0.25×0.25 in the main region, and 0.5×0.025 in the back region. This yields an energy resolution of 10-12% $\text{GeV}^{-1/2}$.

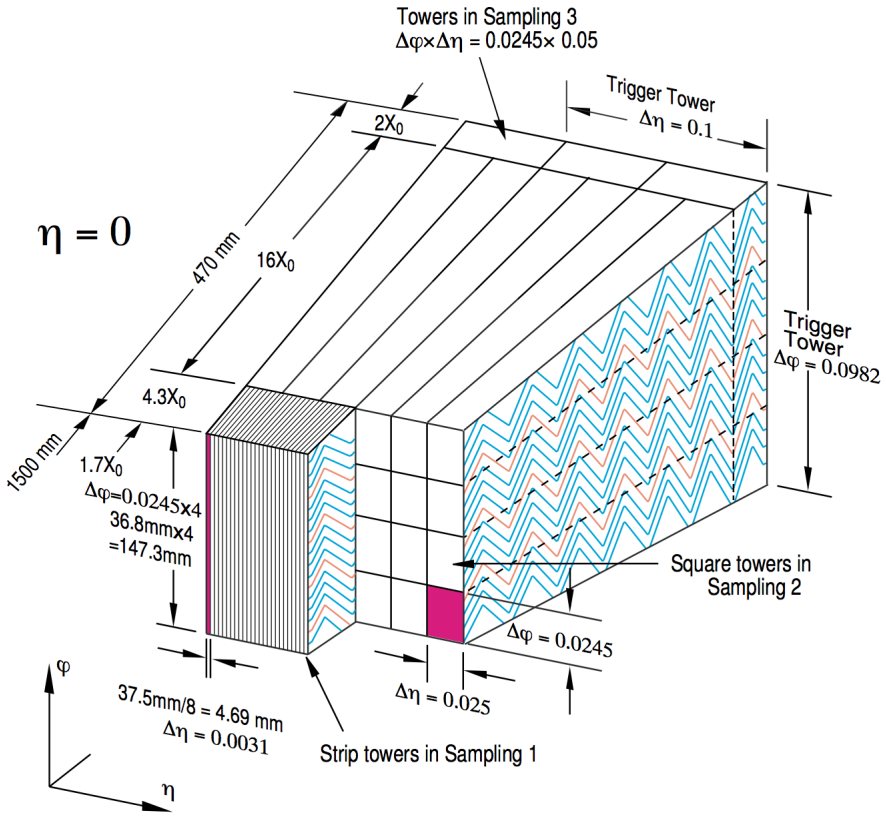


Figure 4.11: Wedge showing the accordian structure for the liquid Argon portion of the calorimeter. [**LAr**]

4.3.2 HCAL

The hadronic calorimeter (HCAL) is the layer just outside of the ECAL, and measures the energy of hadrons that traverse the ECAL without stopping, as well as minimum-ionizing particles like muons. Although the ECAL system could be used to measure the development of the hadronic showers as well, the HCAL system's coarser granularity decreases the monetary cost in covering this larger volume. It is divided into three parts: the tile calorimeter, the LAr hadronic end–cap calorimeter, and the LAr forward calorimeter, as shown in Figure calorimeter.

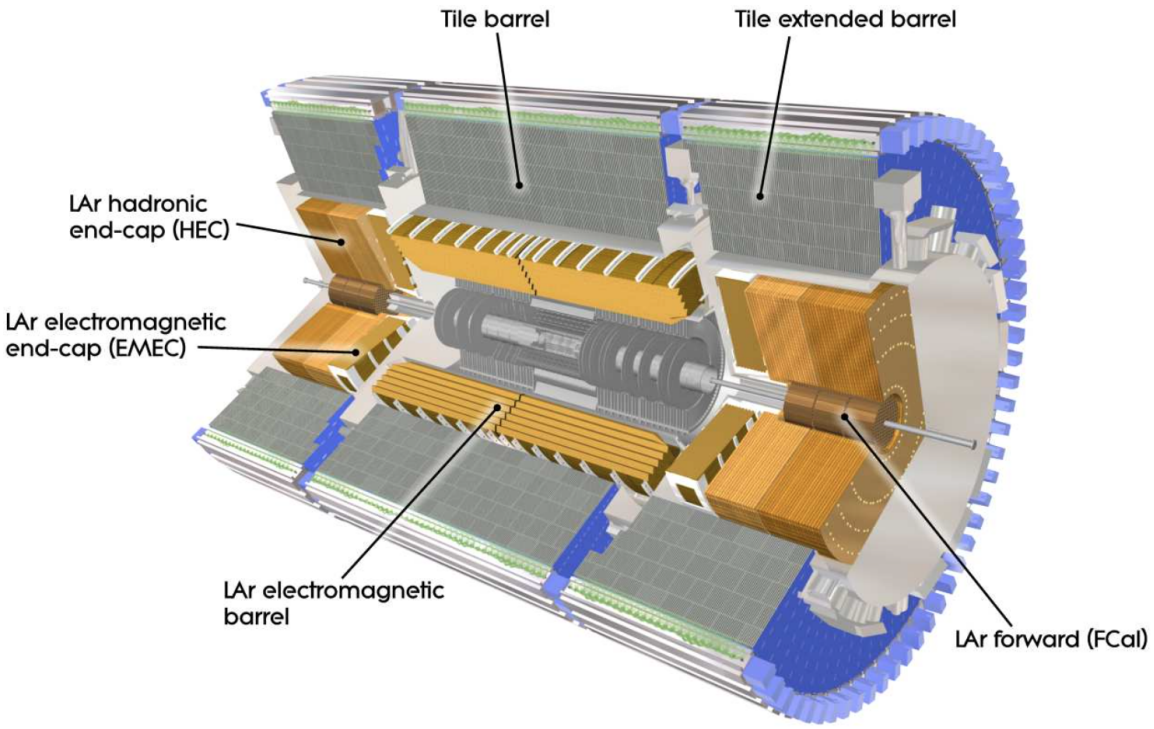


Figure 4.12: Calorimeter system at ATLAS [**ATLAS˙long**].

Tile Calorimeter

The tile calorimeter is just outside ECAL and has an inner radius of 2.28 m and an outer radius 4.45 m. The barrel covers $|\eta| < 1.0$, while the extended barrels cover $0.8 < |\eta| < 1.7$. It is a sampling calorimeter with a steel absorber and scintillating tiles read out with wavelength shifting fibers and photomultiplier tubes. Azimuthally divided into 64 modules, the barrel is segmented into three regions at $1.5\lambda, 1.8\lambda$, and 4.25λ, while the extended barrel has $1.5\lambda, 2.6\lambda$, and 3.3λ depth segmentation. At $\eta = 0$, the HCAL has a 9.7λ depth [**ATLAS˙long**].

LAr hadronic end–cap calorimeter

The LAr hadronic end–cap calorimeter has two wheels per endcap just behind the ECAL endcap calorimeter and housed within the same LAr cryrostats. The wheels have two depth segments, 25 mm parallel copper plates for the wheels closest to the interaction point, and 50 mm plates for the wheels farther from the interaction point. The wheels are divided into 32 wedges, with inner and outer radii of 0.475m and 2.03m, respectively. The copper plates are filled with LAr as the active sampling volume [**ATLAS˙long**].

LAr forward calorimeter

The forward calorimeter covers $\eta > 3.1$ and is approximately 10 interaction lengths deep. Each side has three modules, the first of copper for electronic measurements, and the second two of tungsten for hadronic measurements.

4.4 Muon system

4.4.1 Muon Spectrometer

The tracking system determines the transverse momentum of a charged particle from its radius of curvature in a magnetic field since $R = p_T/(qB)$. A more massive object moving at the same velocity will have a larger transverse momentum, and therefore a larger radius of curvature. Since the muon is 200 times heavier than an electron, it can be difficult to have a tracking system that can accurately measure the momentum for relativistic muons and electrons, so the muon detection system is distinct to measure the p_T for a broad range of muon energies.

The muon detector is a gas detector like the TRT, and operates according to similar principles. The subsections below detail the components of the muon spectrometer, while the parameters for the coverage and number of channels for are listed in Figure Muon-table.

Toroidal Magnets

The muons are detected through the deflection of their tracks in a 4 T magnetic field. There are three large air–core toroids, and each of these 3 toroids has 8 coils. The particles of interest that reach this portion of the detector are muons and neutrinos, but only the muons will be detected because the electrically neutral neutrinos are not deflected by the magnetic field. Backgrounds for the muon spectrometer are photons and neutrons with energies below an MeV and 100 MeV, respectively [**ATLAS˙long**]. The magnetic field is designed to be transverse to the muons' flight direction to minimize multiple scattering. The magnets are housed within cryostats to keep them below the critical temperature for superconductivity. Inside these magnets are the muon chambers divided

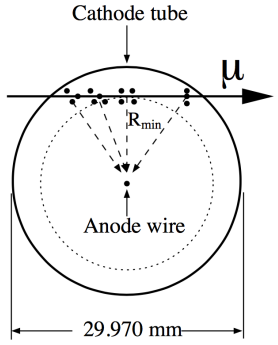
Monitored drift tubes	**MDT**
- Coverage	$\|\eta\| < 2.7$ (innermost layer: $\|\eta\| < 2.0$)
- Number of chambers	1088 (1150)
- Number of channels	339 000 (354 000)
- Function	Precision tracking
Cathode strip chambers	**CSC**
- Coverage	$2.0 < \|\eta\| < 2.7$
- Number of chambers	32
- Number of channels	31 000
- Function	Precision tracking
Resistive plate chambers	**RPC**
- Coverage	$\|\eta\| < 1.05$
- Number of chambers	544 (606)
- Number of channels	359 000 (373 000)
- Function	Triggering, second coordinate
Thin gap chambers	**TGC**
- Coverage	$1.05 < \|\eta\| < 2.7$ (2.4 for triggering)
- Number of chambers	3588
- Number of channels	318 000
- Function	Triggering, second coordinate

Figure 4.13: Main parameters for the Muon Spectrometer. The values in parentheses refer to the configuration from 2009 [**ATLAS'long**].

into three cylindrical chambers about the beam axis in the barrel region, while the transition and end–cap regions arranged as disks, also divided into three chambers [**ATLAS˙long**].

Monitored Drift Tubes

The Monitored Drift Tubes (MDTs) are drift chambers that provide precision measurements. Each tube is made of aluminum with a 3 cm diameter and length between 0.9 and 6.2 m. The tube is filled with a 93% Ar, 7% CO_2 gas mixture, at a pressure of 3-bar [**ATLAS˙long**]. It has a gain of 2×10^4 [**ATLAS˙long**], similar to the TRT drift chambers (see trt). The precision for single muon events is 100 μm, while the precision for multi-muon events is 50 μm [**ATLAS˙long**]. To control the precision, the sag of the wires is minimized by three kinematical mounts placed strategically to minimize distortion due to the support, as shown in Figure MDT-alignment. There are a total of 1174 MDTs positioned in the barrel of the ATLAS detector ($|\eta| < 2$).



(a) Illustration of how an MDT works [**ATLAS˙long**].

figures/Undergrad thesis//MDT-alignment.png

(b) The orientation of the multi-layer MDT chamber with the support structure [**ATLAS˙long**].

Figure 4.14: Monitored Drift Tubes: MDTs

4.4.2 Cathode Strip Chambers

For larger pseudorapidities ($2 < |\eta| < 2.7$), Cathode Strip Chambers (CSCs) are used for their higher granularity in the endcap region [**ATLAS˙long**]. This is a multi-wire proportional chamber with strip read–out with a sense wire pitch of 2.54 mm and a read–out strip pitch of 5.08 mm resulting in a 60 μm track resolution [**ATLAS˙long**].

4.5 Trigger system

The LHC collides proton bunches 40 million times each second – but the vast majority of these events don't contain interesting physics. Since it is infeasible to record every proton-proton collision, ATLAS utilizes a two stage of trigger that determine which events to save and write to disk [**ATLAS'long**], with an overview in Figure 4.15.

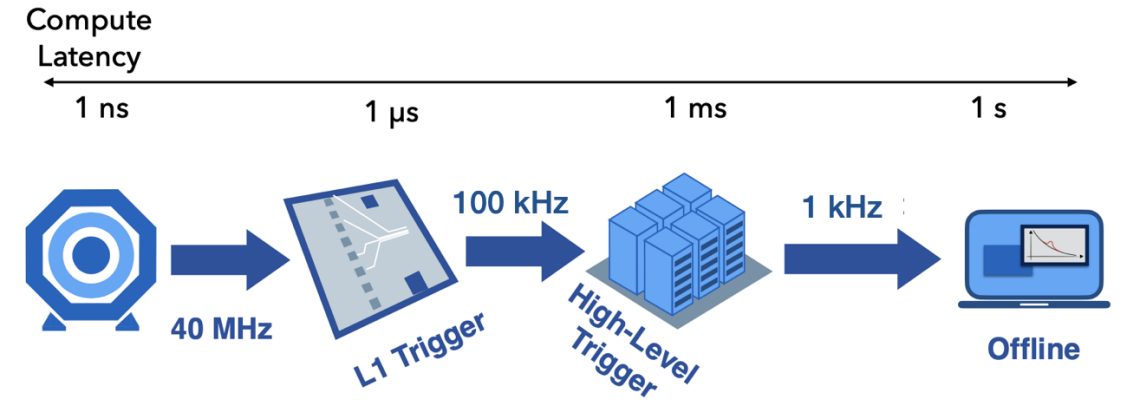


Figure 4.15: Illustration of the data acquisition chain – a sequence of decisions in hardware (L1) and software (HLT) which determines which events get saved [**javier-iaifi-2**] [**ATL-COM-DAQ-2014-054**].

The first stage is the Level I (L1) trigger takes a very coarse view of the detector and has a μs to decide whether to keep the event. This prohibits computationally expensive algorithms (such as tracking) being run, so L1 only looks at trigger objects using information from the calorimeter and muon systems. Since these algorithms need to be very fast, they're implemented in hardware using field programmable gate arrays (FPGAs) before the electronic signals are read off of the detector. It reduces the rate by almost three orders of magnitude (to 100 kHz) for sending the electronic signals off of the detector.

The second stage is High Level Trigger (HLT) is the software based trigger system, that implements an event reconstruction very similar to the offline algorithms. At this stage, tracking information is available, so b-tagging information can be used in the trigger decision (of particular importance for the HH $\rightarrow$ 4b analysis). There are several different "trigger streams" or combinations of selections applied to the trigger objects to collect datasets of interest to the diverse physics program. Events passed from all of these trigger steams get saved at a rate of 1 kHz while data is being taken.

5

Event Reconstruction

The things which are seen are not made of the things which do appear.

– Hebrews 11:3

The task of event reconstruction involves taking electronic read out signals from the 100 million sensors in the detector and to reconstruct objects that can serve as proxies to the physics observables (i.e, b-quarks) that to use in high-level physics analysis. Since the LHC collides protons, we produce lots of quarks and gluons in the final state, with an example of the signature these particles produce in the detector shown in Figure 5.1.

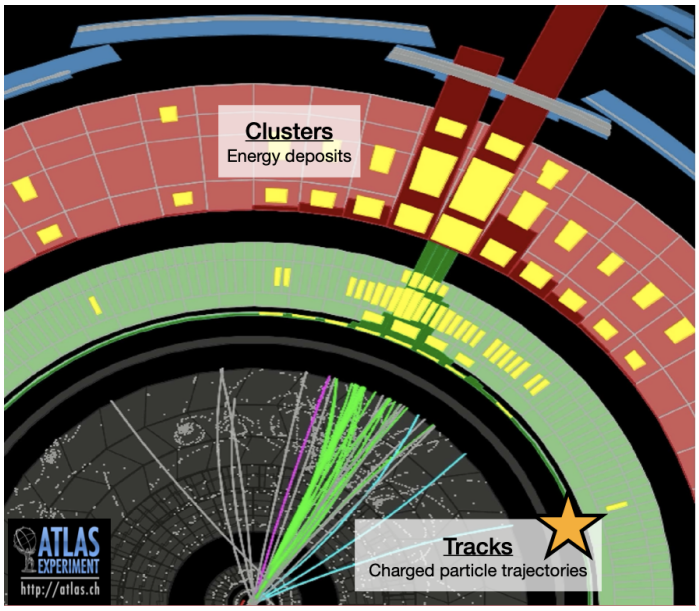


Figure 5.1: Illustration of what a jet looks like in our detector

Since this thesis deals with an analysis with an all-hadronic final states, these are the key input objects for the analysis and for my work in b-tagging, so in this chapter we focus on the reconstruction of these types of events. We include this graphic here, because it nicely demonstrates the outline for this chapter The charged particle trajectories in the inner detector form a set of tracks, and will be described in Section 5.1. The tracks that originate from the same location are then reconstructed into vertices, as described in Section **??**. The unsupervised learning algorithms that get reconstructs the quark and gluon decay products into the "jets" are described in Section **??**. Finally, we conclude with the reconstruction muon tracks in Section 5.4.

Citation: The European Physical Journal C Vol 73 3 (2013) 2304

5.1 Tracks

5.1.1 Track reconstruction

To cite: [**soft-pub-2007-007**]

Inside out:

- Cluster formation
- Seed finding: reco triplets of hits from the pixel detector
- Track fitting with a combinatorial Kalman filter
- Ambiguity solving of which hits with a collection of NNs which decides whether a given cluster is shared between two tracks and how to split the energy depostion between these multiple tracks [**jinst-9-2014-P09009**]
- Extend to the TRT hits

Improve the efficiency due to tracks in that have decays displaced from the original collision point with an **outside in:** track reconstruction alrogihm

- Start with the seeds from the TRT
- Extend to the hits in the silicon detector
- Again use an ambiguity solver.

5.1.2 Challenges in Dense Environments

5.1.3 Discussion of inputs

- R: reference position at which the tracks are defined, i.e, for b-tagging we use the primary vertex of the collision as the reference point.

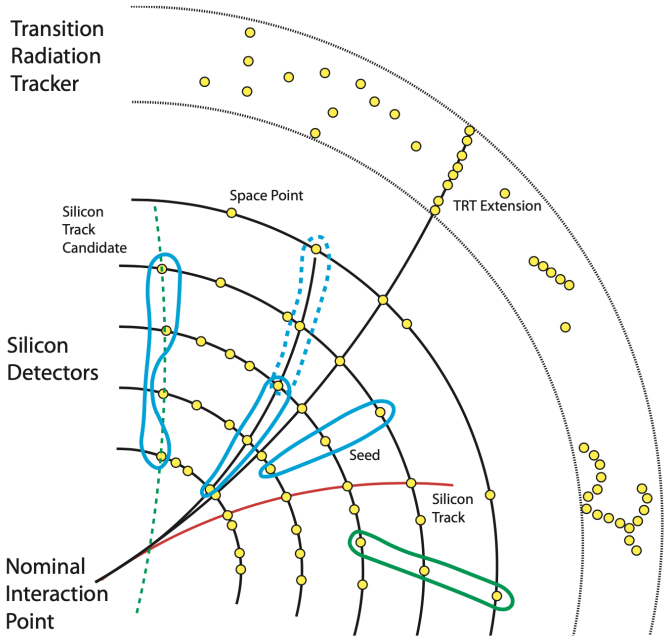
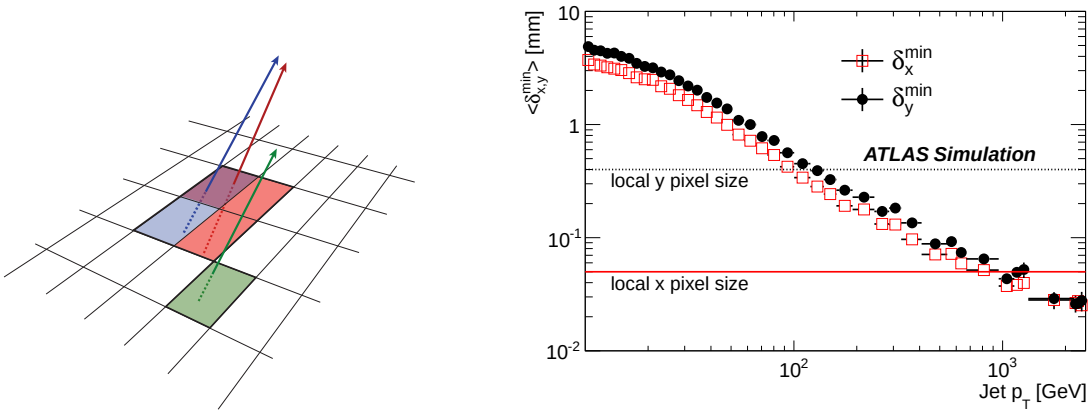


Figure 5.2: need to ask Valentina where this pic was from originally.

- d_0: The transverse impact parameter, point of closest approach (POCA) in the transverse plane with respect to R.
- $z_0 \sin\theta$: The longitudinal impact parameter, or the longitudinal displacement from the POCA (defined in the transverse plane). The multiplication by $\sin\theta$ term is included because it characterizes the 2d distance from z_0 to the closest point along the track trajectory.

$$\begin{aligned} x_V &= x_R + d_0 \cos\left(\phi_p + \frac{\pi}{2}\right) + \rho\left[\cos\left(\phi_V + \frac{\pi}{2} - \cos\left(\phi_p + \frac{\pi}{2}\right)\right)\right] \\ y_V &= y_R + d_0 \sin\left(\phi_p + \frac{\pi}{2}\right) + \rho\left[\sin\left(\phi_V + \frac{\pi}{2} - \sin\left(\phi_p + \frac{\pi}{2}\right)\right)\right] \\ z_V &= z_R + z_0 - \frac{\rho}{\tan(\theta)}\left[\phi_V - \phi_p\right] \end{aligned} \tag{5.1}$$

$$\rho = \frac{\sin\theta}{\frac{q}{p}B_z} \tag{5.2}$$



(a) (b)

Figure 5.3: [**jinst-9-2014-P09009**].

$$d_0 = \text{sign}\,(d_0 - \rho)\sqrt{\left(x_V - x_R - \rho\cos\left(\phi_V + \frac{\pi}{2}\right)\right)^2 + \left(y_V - y_R - \rho\sin\left(\phi_V + \frac{\pi}{2}\right)\right)^2} \tag{5.3}$$

$$\phi_p = \arctan\left(\frac{y_V - y_R - \rho\sin\left(\phi_V + \frac{\pi}{2}\right)}{x_V - x_R - \rho\cos\left(\phi_V + \frac{\pi}{2}\right)}\right) \tag{5.4}$$

$$z_0 = z_R + z_V + \frac{\rho}{\tan\theta}\left[\phi_V - \phi_p(x_V, y_V, z_V, \theta, q/p)\right] \tag{5.5}$$

$$\left(\frac{q}{p}\right)_P = \left(\frac{q}{p}\right)_V \tag{5.6}$$

$$\theta_P = \theta_V \tag{5.7}$$

The way we map from one representation to another is by calculating the Jacobians of the transformations:

$$A = \frac{\partial(d_0, z_0, \phi_P, \theta_P, q/p)}{\partial(x_V, y_V, z_V)} = \begin{bmatrix} -h\frac{X}{S} & -h\frac{Y}{S} & 0 \\ \frac{\rho}{\tan\theta}\frac{Y}{S^2} & -\frac{\rho}{\tan\theta}\frac{X}{S^2} & 1 \\ -\frac{Y}{S^2} & \frac{X}{S^2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{5.8}$$

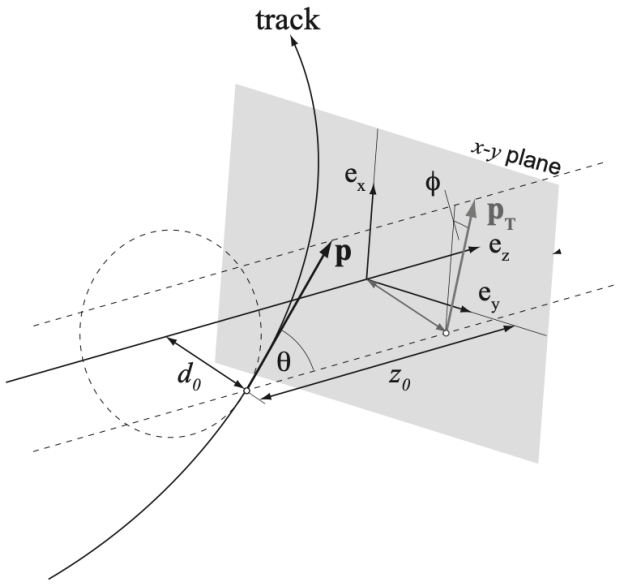


Figure 5.4: [**ATL-SOFT-PUB-2007-003**]

$$B = \frac{\partial(d_0, z_0, \phi_P, \theta_P, q/p)}{\partial(\phi_V, \theta, q/p)} = \begin{bmatrix} -\frac{h\rho}{S}R & \frac{\rho}{\tan\theta}\left[1 - \frac{h}{S}R\right] & -\frac{\rho}{q/p}\left[\Delta\phi - \frac{h}{S}R\right] \\ \frac{\rho}{\tan\theta}\left[1 - \frac{\rho}{S^2}Q\right] & \rho\left[\Delta\phi + \frac{\rho}{S^2\tan^2\theta}R\right] & \frac{\rho}{q/p\tan\theta}\left[\Delta\phi - \frac{\rho}{S^2}R\right] \\ \frac{\rho}{S^2}Q & -\frac{\rho}{S^2\tan\theta}R & \frac{\rho}{S^2 q/p}R \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5.9)$$

where:

$$X = x_V - x_R - \rho\cos\left(\phi_V + \frac{\pi}{2}\right)Y \qquad = y_V - y_R - \rho\sin\left(\phi_V + \frac{\pi}{2}\right)R = Y\sin\phi_V + X\cos\phi_V$$

$$Q = X\sin\phi_V - Y\cos\phi_V$$

$$S = \sqrt{X^2 + Y^2}$$

$$h = \text{sign}\rho$$

$$\phi = \phi_P - \phi_V$$

Q: How do we define the sign of ρ?

5.1.4 The perigee parameters

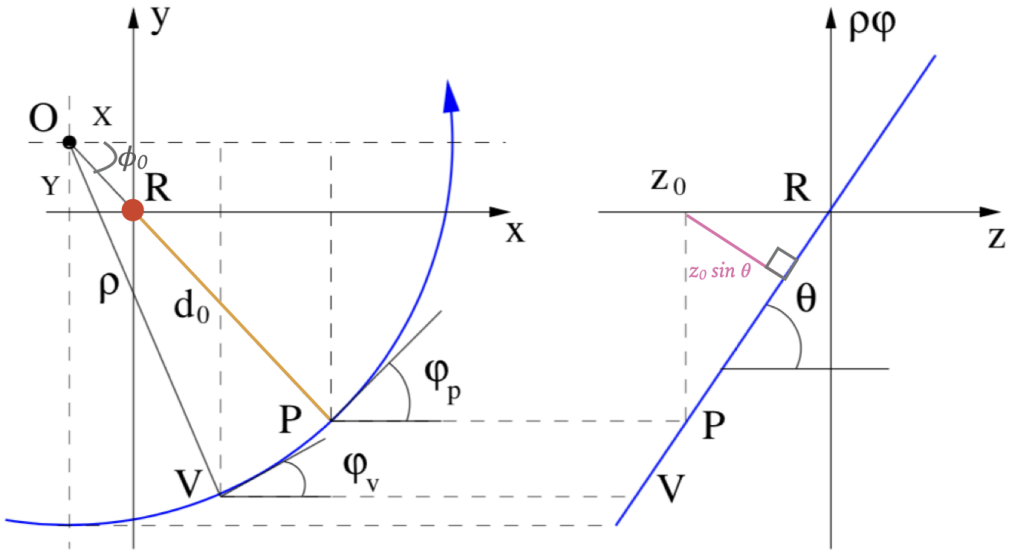


Figure 5.2: Track helix in the transverse (left) and in the $\rho\phi$ plane (right). The track is defined in the *perigee* representation at the point of closest approach P to the reference position R.

Figure 5.5: [**giacinto-thesis**]

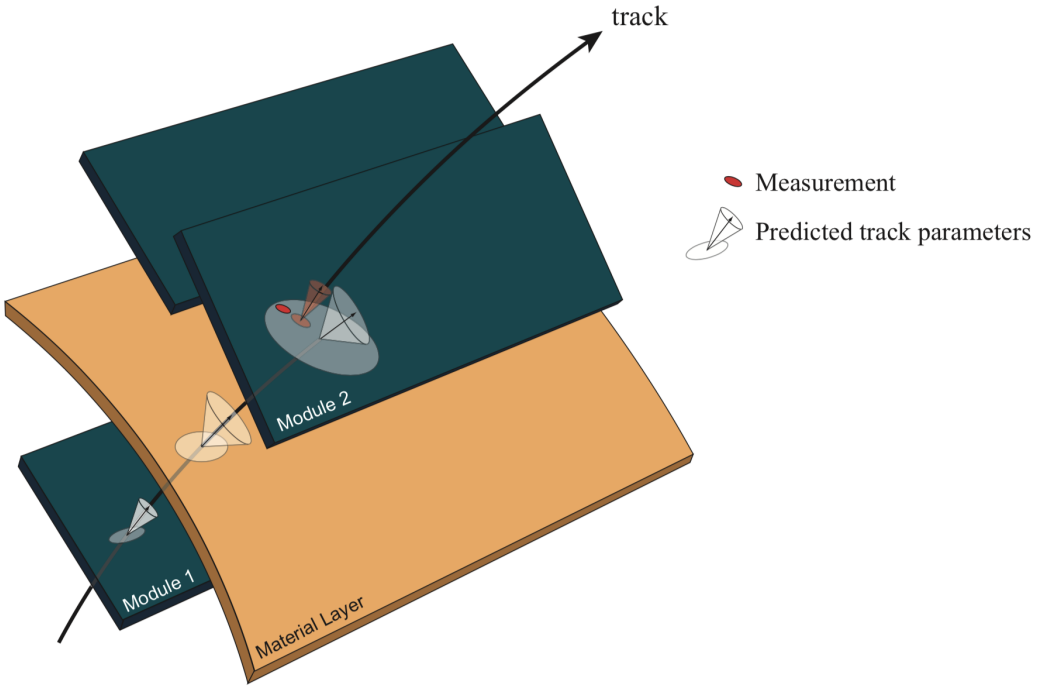
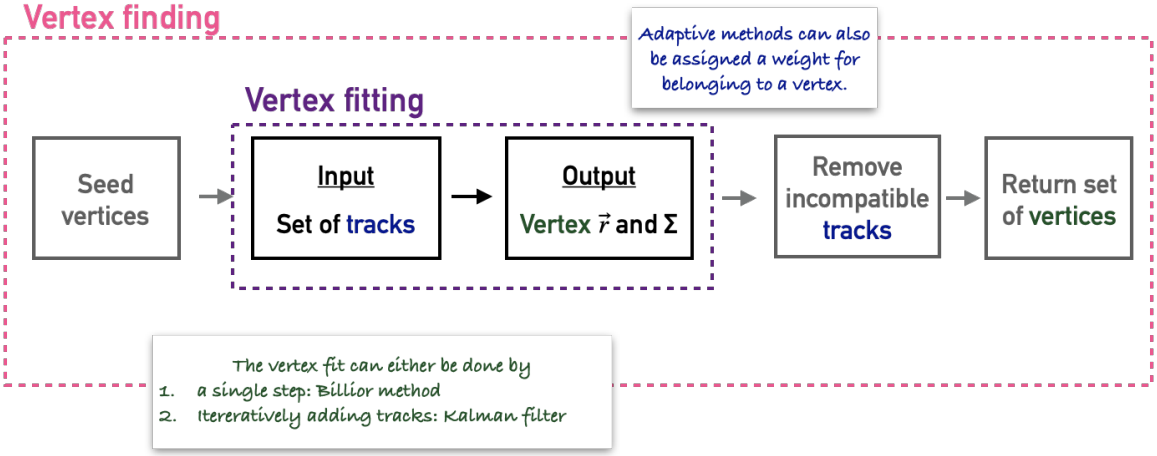


Figure 5.6: Visualization of the track extrapolation with its associated errors [**ATL-SOFT-PUB-2007-005**].

5.2 Vertexing

5.2.1 General problem formulation



The Vertex fit: Position which maximizes the likelihood for the probability density tubes for the tracks to all intersect at this point [**giacinto-thesis**]. This is seen mathematically by Eq. 5.10 below, where $\vec{r}$ is the position of the vertex, and $\vec{r}_i(\phi_{p,i})$ is track i and $\phi_{p,i}$ parametrizes what point of the trajectory that we're on.

$$P(\vec{r}) = \int d\phi_{p,1} d\phi_{p,2} \dots d\phi_{p,n} \prod_{i=1}^{n_{trk}} \exp\left[-\frac{1}{2}\left(\vec{r} - \vec{r}_i(\phi_{i,i})\right)^T COV_{3x3}^{-1}(\phi_{p,i})\left(\vec{r} - \vec{r}_i(\phi_{i,i})\right)\right] \tag{5.10}$$

Is this cov matrix the vertex cov? If so, should I call it Σ instead of Cov_{3x3}?

5.2.2 Primary vertex reconstruction

The event's selected primary vertex (PV) is defined as the reconstructed primary vertex with largest $\sum p_{\mathrm{T}}^2$ of the associated tracks.

5.3 Jets

5.3.1 Jet clustering algorithms

A given pp collision produces many quarks and gluons. However, because of the "confinement principle," no free color charge can exist, which means no isolated quarks or gluons can exist in nature. It becomes energetically favorable for quark / anti-quark pairs to pop out of the vacuum to balance the color charge imbalance by forming color neutral hadrons. This sparks a chain reaction which produces a spray of particles in the detector.

The anti-k$_T$ algorithm provides a way to cluster the energy deposited in the calorimeter to form a "jet," and is the standard algorithm for defining jets at ATLAS [**antiKt**]. First it defines two quantities,

$$d_{ij} = \min(p_{Ti}^{-2}, p_{Tj}^{-2})\frac{\Delta R_{ij}}{R^2} \qquad\qquad d_i = p_{Ti}^{-2} \tag{5.11}$$

where $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, and R is the jet-radius, defined by the user. To find the jets, the algorithm

- Calculates d_{ij} and d_i for all particles in the event, and lets d be the smallest d_i, d_{ij}.
  - If $d = d_{ij}$, then it combines jet i and jet j.
  - If $d = d_i$, then it lets jet i be the final jet.
- It continues the previous steps until all the particles in the event have been accounted for.

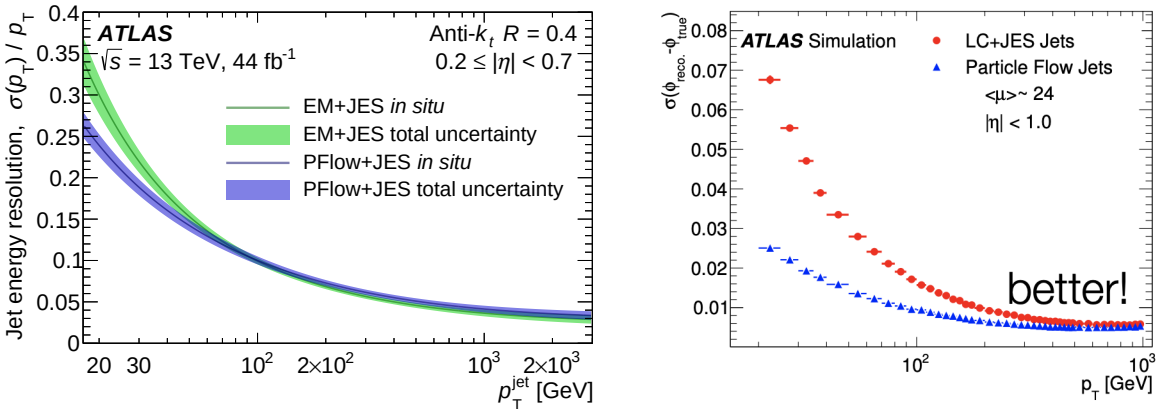
This algorithm clusters high-p_T particles together if they fall within the jet's radius.

The pp collisions at 13 TeV typically will give rise to about 20 moderate p_T jets in the detector, or around $\binom{20}{2} = 190$ viable di-jet candates. So the challenge in elucidating the $H \to b\bar{b}$ or $HH \to 4b$ signals is accurately finding the "correct" pair(s) of jets.

Since the protons are not point-like, the colliding partons may not have equal p_T in the lab frame. By definition, the net momentum will be zero in the center of mass (CM) frame. A heavy (∼TeV) resonance can be created at (or nearly at) rest in the CM frame, but when Lorentz boosting back into the lab frame the decay products can become highly collimated. When the decay products can no longer be resolved individually, we instead search for jets with a large radius parameter ("fat jets"), indicative of merged jets [**SM˙Higgs**, **Merged˙Jets**]. The constituent clusters inside a fat jet are called *subjets*. Large R (about 1.0) correspond to fat jets, and R about 0.4 correspond to the standard resolved jets.

5.3.2 PFlow

Why better?



(a) p_{T} resolution

(b) ϕ resolution

Figure 5.7: Improvement for moving to the PFlow algorithm for jet reconstruction.

To cite for pT plot: 2007.02645

Eur. Phys. J. C 81 (2021) 689

To cite for ϕ plot: https://indico.cern.ch/event/777980/contributions/3236356/subcontributions/271401/attachm

5.3.3 Boosted jets

5.3.4 VR track jets

5.4 Muons

Def useful for muon in jet + pt reco correction!!

6

b-tagging

Technical expertise is the mastery of complexity – creativity is the mastery of simplicity.
– Michael Nielson, *Neural Networks and Deep Learning*

6.1 Introduction

For the physics program of the ATLAS experiment at the Large Hadron Collider (LHC), the identification of jets initiated by b-quarks, or b-tagging, is a fundamental tool. Ensuring its optimal performance is particularly important for the study of the Higgs boson and the top quark [8, 9], as well as many exotic extensions of the Standard Model with resonances preferentially decaying to heavy quarks [10].

The characteristically long lifetime of hadrons containing b-quarks (b-hadrons) of the order of 1.5 ps [11] leads to two classes of b-tagging algorithms: *vertexing* based algorithms which explicitly reconstruct a production point, or vertex, of the b-hadron decay displaced from the primary interaction point, and track based algorithms which exploit the displacement of the reconstructed charged particles trajectories (tracks) produced in b-hadron decays from the primary interaction point.

To further set the stage for the problem of interest in this chapter, in Figure 6.3 motivates what these weakly decaying hadrons look like in simulation that includes the truth information. What we have as inputs to flavor tagging are the set of track features in the perigee representation with the IPs defined with respect to the point of closest approach (POCA). Figure 6.3 shows these track parameters as we extrapolate out from the PV using the extrapolation equations defined in Section 5.2[1]. These representative images illustrate

- ***b*-jet:** There's a characteristic, tertriary decay of both the B and D hadrons.

[1] Many thanks to Jonathan Shalomi for the nice track extrapolator code.

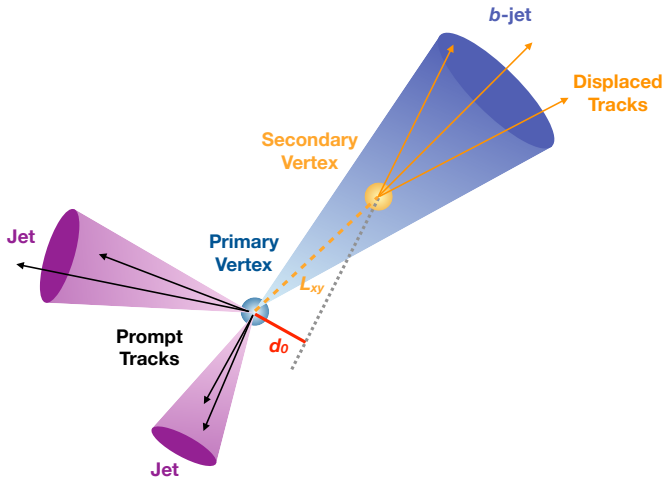


Figure 6.1: Schematic illustration for the characteristic "long" lifetime of a b-hadron [**b-trig-paper**].

- ***c*-jet:** There's a weakly decaying D-hadron, but closer to the PV than the weak decays in the b-jet.
- **light-jet:** Tracks are well collimated with the jet axis and most appear to be originating from the PV, although there are some tracks that that can appear to extrapolate to a point other than the PV.

Table 6.1 shows how often the B and D hadrons decay before the reaching either the first pixel sensor (the IBL) or even the edge of the beam pipe and illustrates that the power of the reconstructing the displaced vertex mostly is coming from the extrapolation.

	B-hadron decays before	**D-hadron decays before**
Beam pipe	2.1 %	0.3 %
IBL	0.94 %	0.1 %

Table 6.1: Decay length of the weakly decaying hadron for jets in from a semi-leptontic $t\bar{t}$ sample.

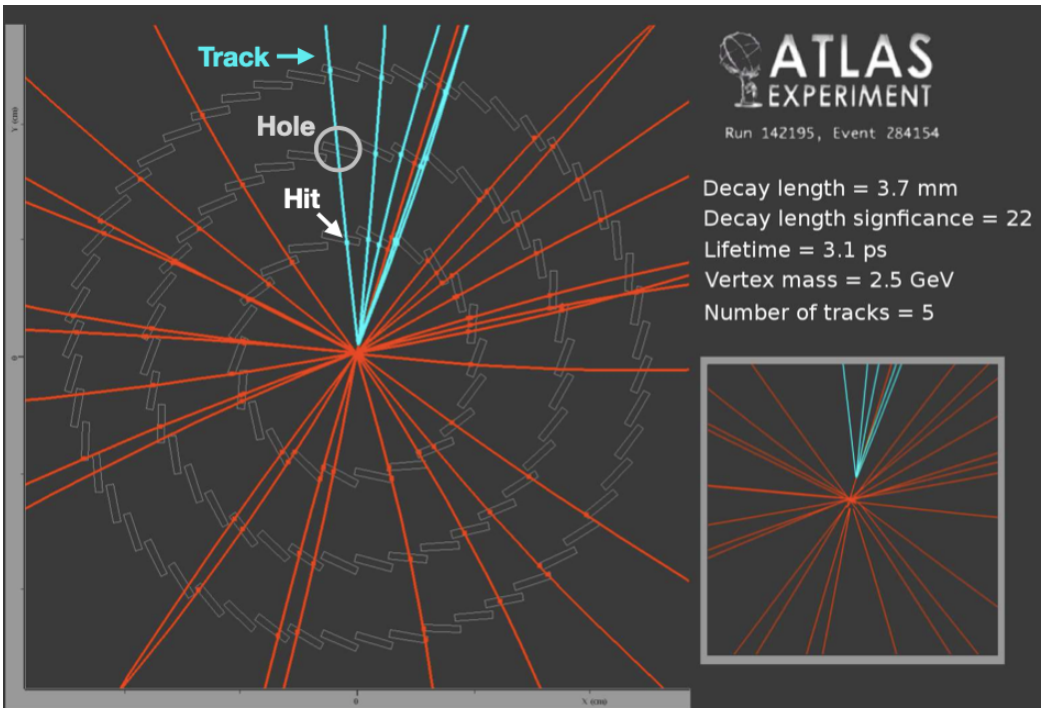


Figure 6.2: Illustration of what a b-decay looks like in the ATLAS detector. The cyan colored lines illustrate the tracks from the b-hadron decay, and in the inset figure you can see the displacement of these tracks from the primary vertex. Only three pixel layers are shown as this is a Run 1 event, and the IBL was not yet installed. Need to revise older notes to find where this event display came from (or ask Su Dong).

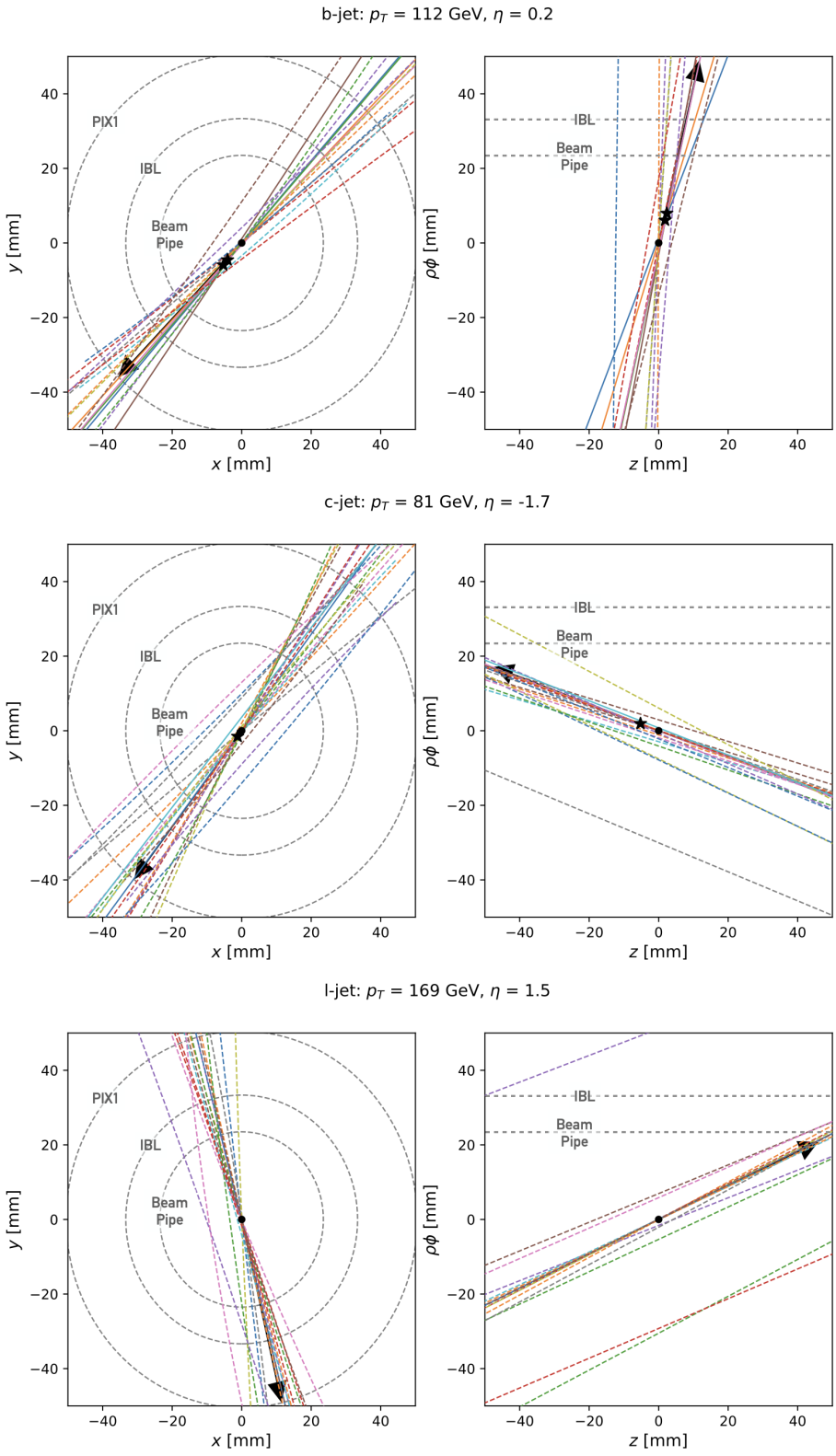


Figure 6.3: Visualization of the (x,y) and $(z, \rho\phi)$ 2d views of the tracks for reconstructing the secondary vertex (or vertices). The arrow on the figure indicates the jet axis, and a $\star$ shows where the weakly decaying hadron decays. The solid lines are tracks from the HF decay, while the dashed lines denote the other tracks associated to the jet.

ATLAS employs several IP-based algorithms which are later combined with vertexing algorithms to produce a "high-level" tagger for general use.

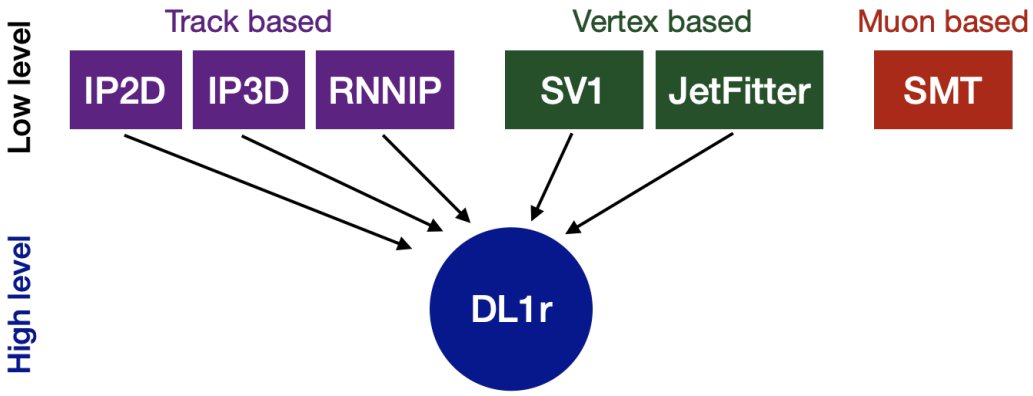


Figure 6.4: Types of b-taggers used on ATLAS

This chapter is organized as follows: Section **??** describes the datasets and selections used to train and evaluate the algorithms, while section **??** details impact parameter based taggers, the Deep Sets algorithm and our specific implementation. Section **??** shows investigations of what the network has learned, results for the timing metrics, discussion on calibrating the algorithm, and the optimization studies conducted. Finally, section **??** summarizes the conclusions.

6.2 Datasets (?)

Algorithm training and evaluation is performed with simulated $t\bar{t}$ events, produced by $\sqrt{s} = 13$ TeV proton-proton collisions, in which at least one of the W bosons, from the top quark decay, decays leptonically. Events are generated using the PowhegBox [12–15] v2 generator at next-to-leading order with the NNPDF3.0NLO [16] parton set of distribution functions (PDF) and the h_{damp} parameter[2] set to 1.5 m_{top} [17], with $m_{\text{top}} = 172.5$ GeV. The events are interfaced to Pythia 8.230 [18] to model the parton shower, hadronisation, and underlying event, with parameters set according to the A14 tune [19] and using the `NNPDF2.3lo` set of PDFs [20]. The decays of b and c-hadrons are performed by EvtGen v1.6.0 [21]. Particles are passed through the ATLAS detector simulation [22] based on GEANT4 [23].

Jets are reconstructed from particle flow objects [24] using the anti-k_T algorithm [25] with $R = 0.4$. The jet energy scale is calibrated according to [26]. Jets used for training and evaluation have $p_{\text{T}} \geq 20$ GeV, $|\eta|$ ¡ 2.5, and are required not to overlap with a generator-level electron or muon from W boson

[2]The h_{damp} parameter is a resummation damping factor and one of the parameters that controls the matching of Powheg matrix elements to the parton shower and thus effectively regulates the high-p_{T} radiation against which the $t\bar{t}$ system recoils.

decays. Additionally, the contamination of jets from other interactions in the beam crossing (pile-up) is surpressed by applying the jet vertex tagger [27] optimized for particle flow jets.

Tracks are associated to jets using a ΔR association cone which decreases as a function of jet p_{T}, with a maximum association ΔR(track, jet) of approximately 0.45 for a jet with $p_{\mathrm{T}} = 20$ GeV and ΔR(track, jet) of approximately 0.25 when the jet $p_{\mathrm{T}} = 200$ GeV. If a track is within the association cones of more than one jet, it is assigned to the jet which has a smaller ΔR(track, jet).

The impact parameter of the track characterises the point-of-closest approach of a track to the PV in the longitudinal ($z_0 \sin\theta$) and transverse (d_0) planes. Of particular use in b-tagging is the IP significance defined as the impact parameter divided by its uncertainty, $s_{d0} = d_0/\sigma_{d0}$ and $s_{z0} = z_0 \sin\theta/\sigma_{z0\sin\theta}$. The track's IP and its significance are signed according to the track's direction with respect to the jet axis and the primary vertex [28]. A positive IP is expected to be consistent with a track produced from a displaced vertex. This procedure is referred to as lifetime signing. The nominal track selection considered in the algorithms to be described requires tracks with $p_{\mathrm{T}} > 1$ GeV, $|d_0| < 1$ mm, and $|z_0 \sin\theta| < 1.5$ mm.

The jets are labelled as b-jets if they are matched to at least one b-hadron having $p_{\mathrm{T}} \geq 5$ GeV within ΔR(b-hadron, jet)< 0.3 of the jet axis. If this condition is not satisfied, then c-hadrons and then τ leptons are searched for, with similar selection criteria. If a jet is matched to a c-hadron (τ-lepton), it is labelled a c-jet (τ-jet). A jet that does not meet any of these conditions is called a light-flavour jet.

6.3 Low level taggers

There are two types of low-level algorithms that are employed by the

While the IP-based algorithms take input a set of tracks and just use the individual track features (such as IPs) to classify the jet, the vertexing-based algorithms (SV1 and JF) involve an iterative approach to reconstruct the topology of the displayed decay.

The outlining of this section is as follows: I first

Although the RNNIP algorithm was proposed before developing my thesis, a key contribution of my thesis work was continuing to optimize this develop

6.3.1 IP2D and IP3D

Lifetime signage

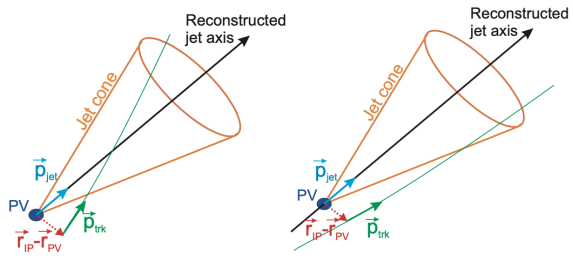


Figure 6.5: Lifetime signage graphic (from [**giacinto-thesis**])

3d sign

$\Delta\vec{r}_{IP} = \vec{r}_{IP} - \vec{r}_{PV}$

$$\text{sign}_{3D} = \text{sign}\left(\left[\vec{p}_{trk} \times \vec{p}_{jet}\right] \cdot \left[\vec{p}_{trk} \times \Delta\vec{r}_{IP}\right]\right) \tag{6.1}$$

2d sign

$$\text{sign}_{r\phi} = \text{sign}\left(\sin\left(\phi_{jet} - \phi_{trk}\right) \cdot d_{0,trk}\right) \tag{6.2}$$

$$\text{sign}_{z} = \text{sign}\left(\left(\eta_{jet} - \eta_{trk}\right) \cdot z_{0,trk}\right) \tag{6.3}$$

(Equations taken from Giacinto's thesis [**giacinto-thesis**].)

How are the IPs signed for the IP2D and IP3D algorithms

- **IP2D:**
  - d_0 signed based on the projection of the vectors in the (x,y) plane

- **IP3D:**
  - d_0 signed with the 3D vectors
  - z_0 signed based on the (r, ϕ) plane

Mathematical motivation

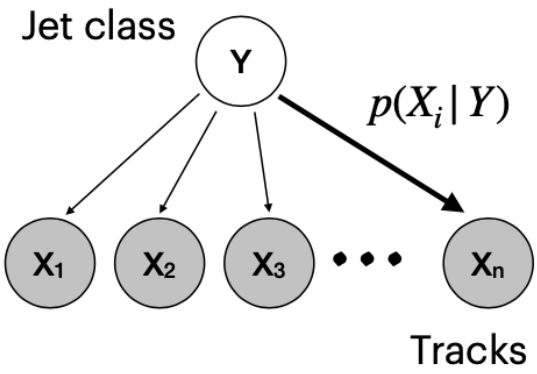
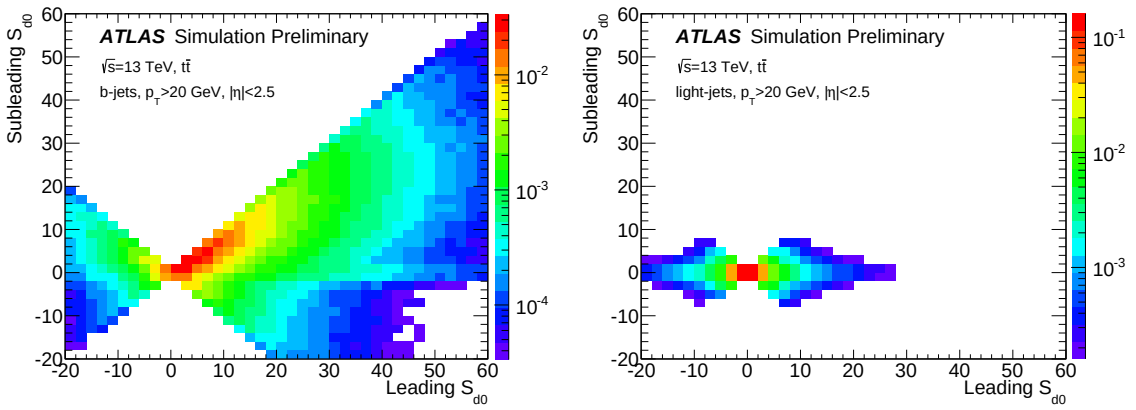


Figure 6.6: Probabilistic graphical model illustration for a Naive Bayes algorithm.



(a) (b)

Figure 6.7: [29]

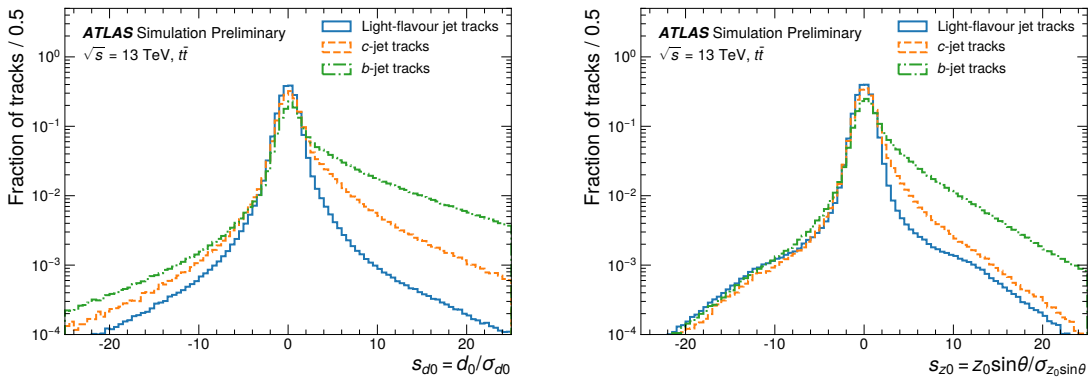
Algorithm description

The IP3D algorithm [30] assigns probabilities to tracks based on two-dimensional likelihood templates, with the tracks' $z_0 \sin\theta$ and d_0 lifetime signed significances, built from simulated jets. These templates are obtained in 14 exclusive categories defined by the hit patterns of the tracks, and separately for

tracks in b-jets, c-jets and light-flavour jets. The inclusive distribution of $z_0 \sin\theta$ and d_0 lifetime signed significances for the different jet flavours are shown in Figure 6.8. By assuming that the track probabilities inside a jet are independent, jet-level likelihoods can be constructed by multiplying the individual probabilities. The IP3D b-tagging discriminants are therefore defined as:

$$D_{\mathrm{IP3D,l}} = \log \prod_{i \in \mathrm{tracks}} \frac{p_b^i}{p_l^i} \qquad D_{\mathrm{IP3D,c}} = \log \prod_{i \in \mathrm{tracks}} \frac{p_b^i}{p_c^i}. \tag{6.4}$$

RNN based IP algorithms aim to overcome this overly simplistic assumption of independence, and offer the possibility to employ more features than only the IP significance in the discriminant [29].



(a) (b)

Figure 6.8: Lifetime signed transverse (a) and longitudinal (b) significances for b-jets, c-jets and light-flavor jets.

6.3.2 RNNIP

Tracks are reconstructed from energy deposits, or hits, in the inner detector system and are required to pass a quality selection: each track must have at least 7 hits in the silicon layers (pixel and SCT, where dead sensors are not penalized), no more than two missing hits where expected in the silicon layers, no more than one hit shared by multiple tracks, at least one hit in the pixel detector, and $|\eta| < 2.5$.

RNNs operate on variable length *sequences* by iterating over the sequence elements, processing them with a neural network, and using previously processed elements when processing new ones. It then outputs a fixed size vector that can be used for classification. The RNNIP algorithm utilizes a Long Short Term Memory (LSTM) cell for the RNN to preserve long range correlations between the elements of the sequence [32]. As shown in [29], the accounting for these correlations allows

#	Category	Fractional contribution [%]		
		b-jets	*c*-jets	light-jets
0	No hits in first two layers; expected hit in IBL and b-layer	1.9	2.0	1.9
1	No hits in first two layers; expected hit in IBL and no expected hit in b-layer	0.1	0.1	0.1
2	No hits in first two layers; no expected hit in IBL and expected hit in b-layer	0.04	0.04	0.04
3	No hits in first two layers; no expected hit in IBL and b-layer	0.03	0.03	0.03
4	No hit in IBL; expected hit in IBL	2.4	2.3	2.1
5	No hit in IBL; no expected hit in IBL	1.0	1.0	0.9
6	No hit in b-layer; expected hit in b-layer	0.5	0.5	0.5
7	No hit in b-layer; no expected hit in b-layer	2.4	2.4	2.2
8	*Shared* hit in both IBL and b-layer	0.01	0.01	0.03
9	At least one *shared* pixel hits	2.0	1.7	1.5
10	Two or more *shared* SCT hits	3.2	3.0	2.7
11	*Split* hits in both IBL and b-layer	1.0	0.87	0.6
12	*Split* pixel hit	1.8	1.4	0.9
13	Good quality	83.6	84.8	86.4

Table 6.2: Categories for defining the IP2D and IP3D templates [31].

the RNN to be more performant than IP3D even when trained on the same inputs. The use of neural networks instead of histograms allows one to avoid the "curse of dimensionality" when using additional variables sensitive to the kinematics of the *b*-hadron decay which significantly improve performance [29].

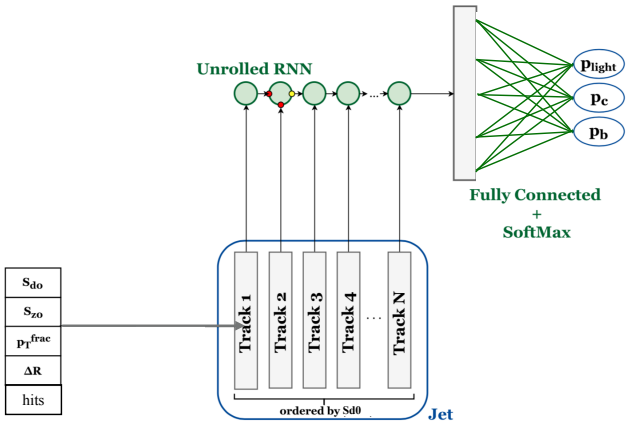


Figure 6.9: RNNIP architecture (modified from **??**).

An implementation of the RNNIP algorithm is used as the baseline for comparison to DIPS, but has further optimisations with respect to [29]. The RNNIP architecture comprises a 100 dimensional LSTM hidden state and a dropout layer, with dropout fraction of 0.2, before a 20 unit fully connected layer for classification, uses track IP significance, kinematics, and the number of hits in the silicon

detectors as features (described in Table 6.3), and orders the tracks by s_{d_0}.

Input	Description
s_{d0}	d_0/σ_{d0}: Transverse IP significance
s_{z0}	$z_0 \sin\theta/\sigma_{z0\sin\theta}$: Longitudinal IP significance
$\log p_{\mathrm{T}}^{frac}$	$\log p_{\mathrm{T}}^{track}/p_{\mathrm{T}}^{jet}$: Logarithm of fraction of the jet p_{T} carried by the track
$\log \Delta R$	Logarithm of opening angle between the track and the jet axis
IBL hits	Number of hits in the IBL: could be { 0, 1, or 2 }
PIX1 hits	Number of hits in the next-to-innermost pixel layer: could be { 0, 1, or 2 }
shared IBL hits	Number of shared hits in the IBL
split IBL hits	Number of split hits in the IBL
nPixHits	Combined number of hits in the pixel layers
shared pixel hits	Number of shared hits in the pixel layers
split pixel hits	Number of split hits in the pixel layers
nSCTHits	Combined number of hits in the SCT layers
shared SCT hits	Number of shared hits in the SCT layers

Table 6.3: Track features used as inputs for RNNIP and DIPS algorithms.

6.3.3 SV1

Philosophy:

SV1 is a vertexing algorithm that reconstructs displaced decays where the set of tracks come from a single secodary vertex. Although strictly speakig this will rarely be

Set of input tracks (preselection):

- Track cuts: $|d_0| < 3.5$mm, $|z_0 \sin\theta| < 5$mm [**giacinto-thesis**] and $p_{\mathrm{T}} > 500$ MeVs.
- Some track cleaning cuts for jets with $\eta > 1.5$ (since these tracks pass through more material).
- Also - only take the 25 highest p_{T} tracks inside of the jet (helped for reducing the reconstruction of vertices from a random crossing of tracks at high jet p_{T} [.]

These track pre-selection cuts are looser than the IPXD (and RIP) algorithms because the next step of selecting the tracks originating from the same point in space acts as another cut on the input tracks. The track ...

- Form all pairs of 2-track vertices satisfying [**giacinto-thesis**]:
  1. Prob($\chi^2_{2-trkvtx}$) > 3.5%
  2. Each track needs to be displaced from the PV with a significance $L_{3D}/\sigma(L_{3D}) > 2\sigma$ and the sum of the two-track significances larger than 6.[3]

[3] L_{3D} is the 3d distance from the tracks POCA to the PV.

3. These tracks need to be "downstream" of the jet axis ($(\vec{r}_{2trk} - \vec{r}_{primary}) \cdot \vec{p}_{jet}$) What is $\vec{r}_{p}rimary$?

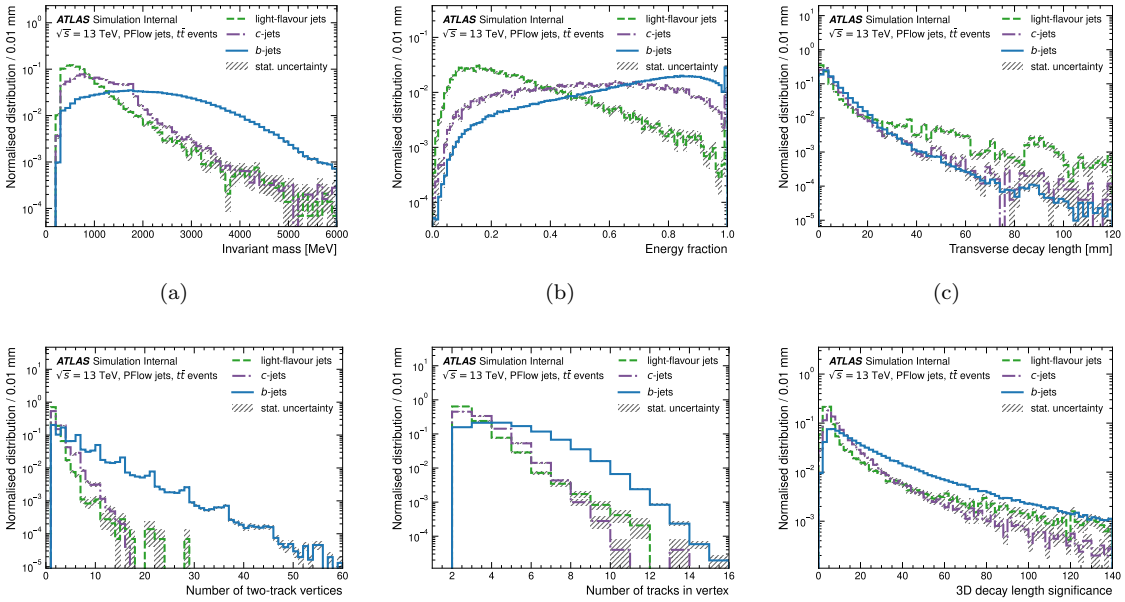
- Veto tracks that form 2-track vertices consistent with[4]:
  1. K_s decays ($|m_{\pi^+\pi^-} - m_{K^0}| < 18$ MeV)
  2. λ decays ($|m_{p\pi^-} - m_{K^0}| < 18$ MeV)
  3. γ conversions ($m_{ee} < 30$ MeV)
  4. hadronic material interactions (veto vertex interactions that overlap with detector material)
- Iterate over this "cleaned" set of tracks fitting a single SV
  - If this vertex fix has a Prob(χ^2_{vtx}) $< 0.1\%$ – or – a vertex mass larger than 6 GeV, remove the track with the largest χ^2 contribution and rise and repeat the fit.

For a true displaced decay, this secondary vertex will have properties consistent with a B or D hadron decay, and some key discriminating variables include the mass of the secondary vertex, the energy fraction, ...

Input	Description
m	Invariant mass of the tracks reconstructed in the secondary vertex
f_E	The energy of the SV over the energy of the jet
$\Delta R(\vec{p}_{jet}, r\vec{_{SV}} - r\vec{_{PV}})$	Opening angle between the jet and the SV flight axis
L_{xy}	SV transverse distance from the PV
L_{xyz}	SV distance from the PV
S_{xyz}	Significance of the displacement of the SV: L_{xyz}/σ_{xyz}
$n_{vtx\ trk}$	Number of tracks in the SV
$n_{2-trkvtx}$	Number of 2 track vertices before the vertex fit

Table 6.4: Features from the SV1 reconstruction that are fed as input to the DL1r tagger.

[4]The tracker only measures the momentum of the particle, so to reconstruct a mass, we need to make an assumption for what the particle ID is for the mass of the track to get the 4-vector. The meson (or lepton) mass used for each track to reconstruct the 2-track vertex invariant mass is denoted by the subscripts. I took these numbers from the JF pub note, will need to confirm is they are the same for SV1.



(a) (b) (c)

(d) (e) (f)

Figure 6.10: The SV1 inputs that (will be) in the FTAG algos paper [**ANA-FTAG-2019-07**].

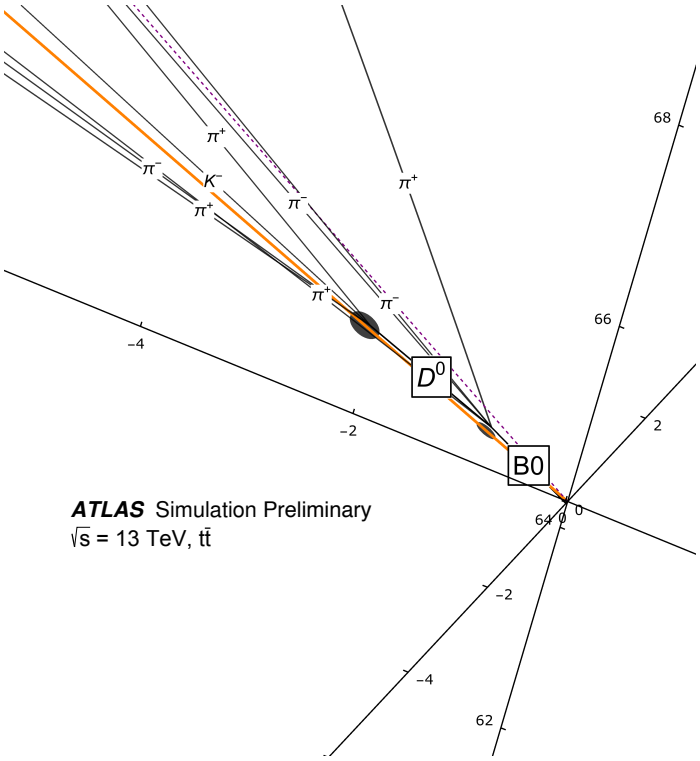


Figure 6.11: [**ATL-PHYS-PUB-2018-025**]

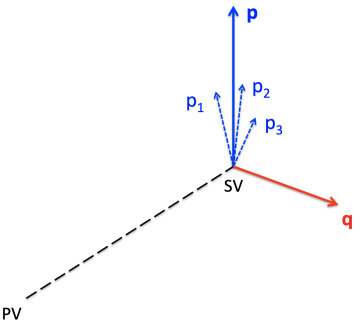
6.3.4 JetFitter

- Motivation for JF: Cascade topology
- Key assumption: Every (well motivated b/c the B-hadron caries a large portion of the initial quarks momentum which causes the B (and subsequent D) hadrons to form a line with respect to the
- Briefly sketch the extension to the KF formalism
- The track selection and JF algorithm
- The final variables defining the JF algorithm and variables targetting c-tagging
- Mass constraint

Mass constraints

I should note the motivation for these checks?

- p : 4-vector of the charged particles
- q: 4-vector for neutral particles
- Let $p_{true} = p + q$ be the true 4-vector of charged and neutral particles.



Then $p = p_{true} - q$, which we can then square. We additionally make a simplifying assumption that the neutral particles are massless:

$$p^2 = p_{true}^2 - 2p \cdot q + \cancelto{0}{q^2} \tag{6.5}$$

Let $p^2 = m_{ch}^2$ the mass from the charged particles, and $p_{true}^2 = m^2$, the weakly decaying hadron mass. Then evaluate the right hand side of Eq. 6.5 in the hadron's rest frame ($\vec{p}_{true}^{CM} = 0$). Then q is perpendicular to the hadron's flight axis $|\vec{q}^{CM}| = q_\perp$, where the "CM" superscript denotes the hadron's center-of-mass frame, and $q_\perp$ is the component of $\vec{q}$ perpendicular to the hadron's flight axis, which is invariant to boosts along the flight axis.

$$m_{ch}^2 = m^2 - 2\left(mq_\perp - \cancelto{0}{\vec{p}_{true}^{CM} \cdot \vec{q}^{CM}}\right) \tag{6.6}$$

$$m_{ch}^2 + p_\perp^2 = m^2 - 2mp_\perp + p_\perp^2 = (m - p_\perp^2)^2 \tag{6.7}$$

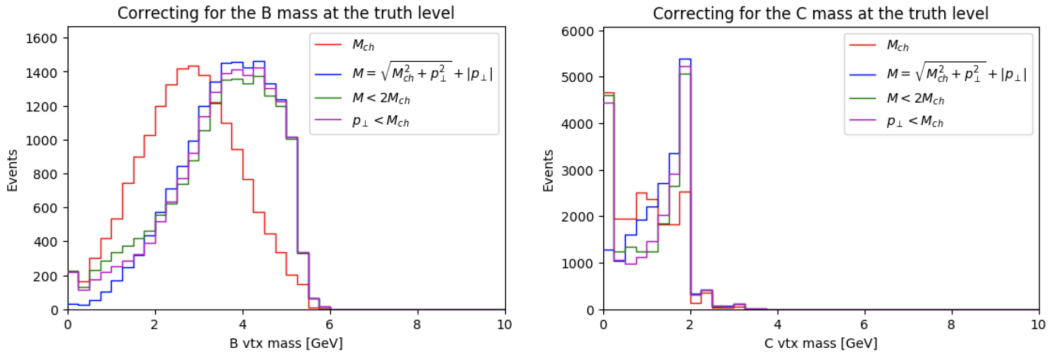
and solve for m:

$$m = \sqrt{m_{ch}^2 + p_\perp^2} + p_\perp \tag{6.8}$$

where we took the + solution of the $\sqrt{}$ for the physical solution of the positive hadron mass.
How JF applies the mass correction:

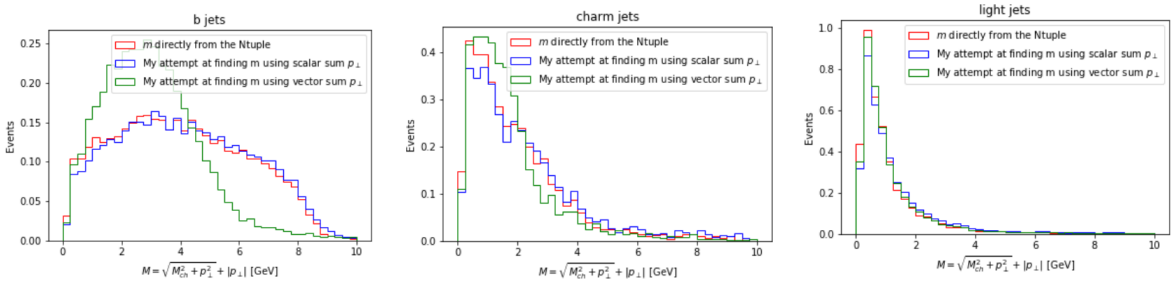
- It uses the scalar sum of the tracks, $p = \sum_i p_\perp^{(i)}$, where i runs over the tracks in the jet. Then just replace the vector sum over the tracks with the scalar sum over the tracks in the formula as: $m = \sqrt{m_{ch}^2 + p_\perp^2} + |p_\perp|$.
- To constrain the tails, if $m > 5$, $m \leftarrow 5\,[1 + 2\arctan(\pi(m-5))]$.

As more tracks are dropped, the scalar sum does increasingly better over the vector sum. Mean number of truth particles not reconstructed: 0.88 Mean truth particles not found by JetFitter: 2.15
The jet input variables that (will be) shown in the FTAG algos paper:



(a) B-hadron reconstruction (b) D-hadron reconstruction

Figure 6.12: Reconstruction of the B (left) and D (right) hadron masses from the truth charged particles.



(a) Vertex reco in b-jets (b) Vertex reco in c-jets (c) Vertex reco in light jets

Figure 6.13

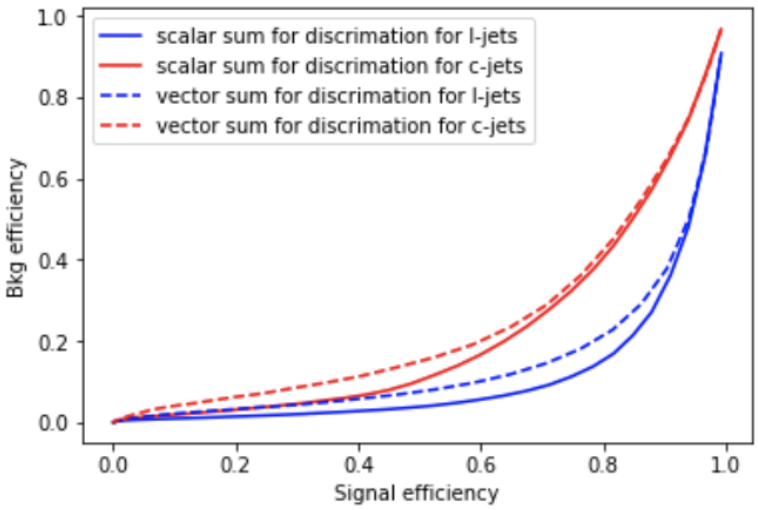
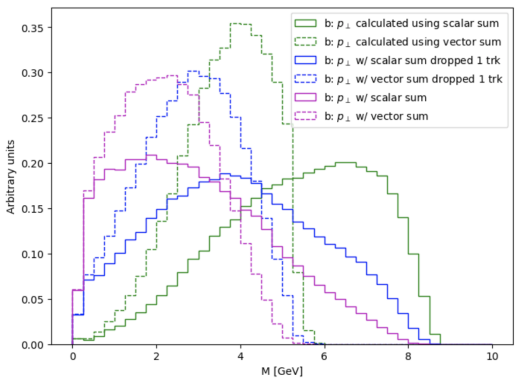


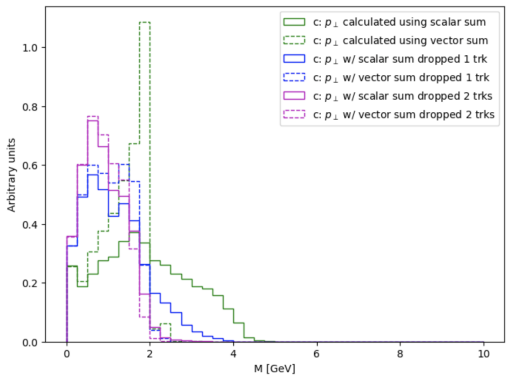
Figure 6.14

Input	Description
m_{JF}	Invariant mass of the tracks attached to displaced vertices
f_E	Fraction of the energy in the displaced vertices compared to the jet's energy
$\Delta R(\vec{p}_{jet}, \vec{p}_{vtx})$	
S_{xyz}	Average significance of all of the displaced vertices
n_{trk}	Number of tracks associated to the fitted displaced vertices along the cascade decay chain
$n_{2-trkvtx}$	Number of 2 track vertices (before the decay chain fit)
$n_{1-trkvtx}$	Number of the single track vertices (after the decay chain fit)
$n_{\geq 2-trkvtx}$	Number of the multi-prong displaced vertices (after the decay chain fit)
L_{xyz}(SV)	3d distance from the first displaced vertex
L_{xy}(SV)	transverse distance from the first displaced vertex
m_{trk}(SV)	Mass of the tracks associated to the first displaced vertex
E_{trk}(SV)	Energy of the tracks in the first displaced vertex
f_E(SV)	Energy fraction of the tracks in the first displaced vertex compared to the energy of the jet
$n_{vtx\ trk}$(SV)	Number of tracks attached to the first displaced vertex

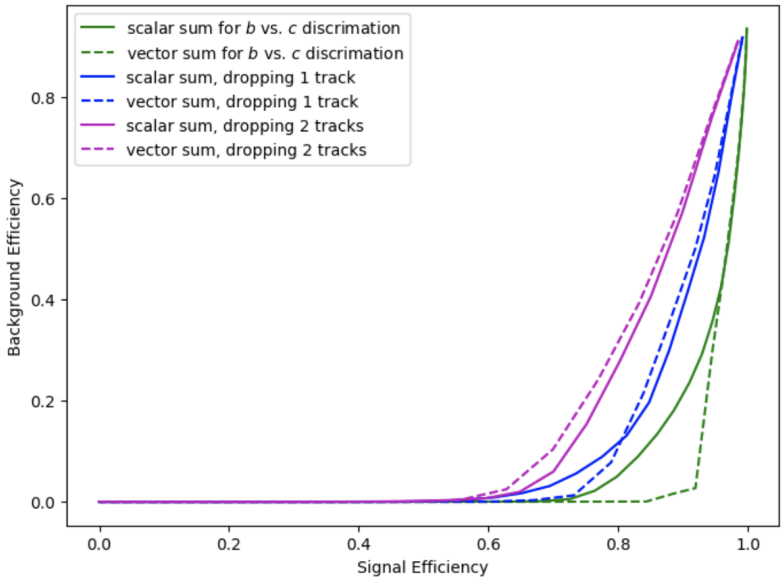
Table 6.5: Features from the JF reconstruction that are fed as input to the DL1r tagger. The first block of variables quantifies the global properties of the displaced vertices and decay topology. The second set of variables just looks at the properties of the first displaced vertex which capture the differences between b-jets and c-jets.



(a) B-hadron reconstruction

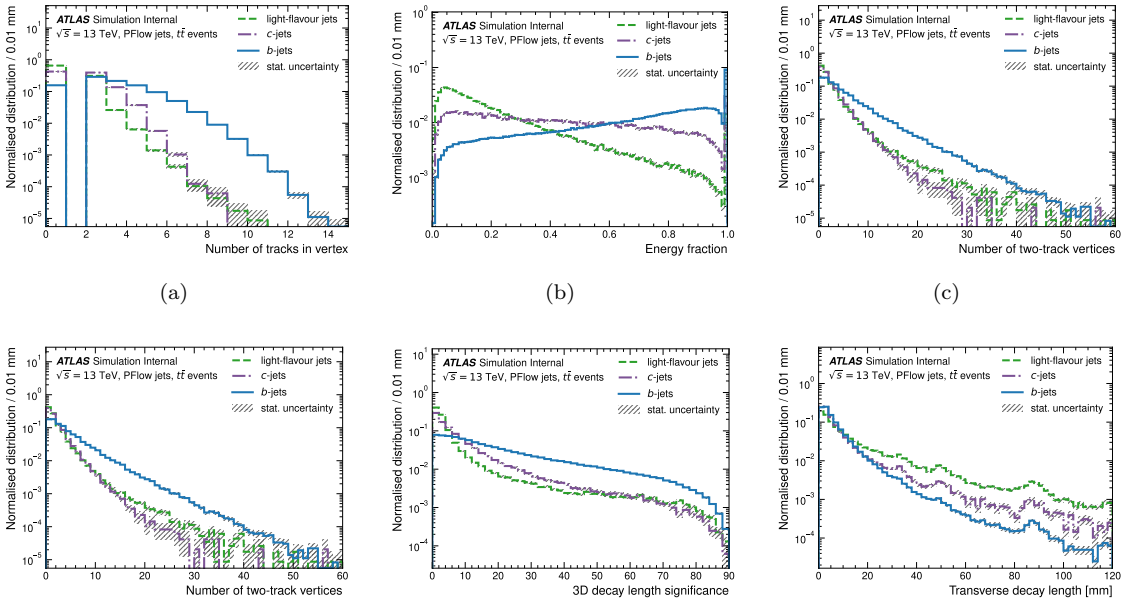


(b) D-hadron reconstruction



(c) roc

Figure 6.15



(a) (b) (c)

(d) (e) (f)

Figure 6.16: The JF inputs that (will be) in the FTAG algos paper [**ANA-FTAG-2019-07**].

6.4 Recommendations for Run 2 b-taggers

6.4.1 High level taggers: DL1 series

We combine the low level taggers into a high-level one, the DL1r algorithm, combining information from the low-level taggers:

- The IP2D and IP3D: $\log p_b$
- The RNNIP outputs: p_b, p_c, and p_l
- The SV1 vertex information (variables in Table **??**)
- The JF displaced vertices characteristics, Table **??**.

The comparisons that we'll show are again

- Explain architecture, optimization
- Explain default values
- pre-processing: mean 0 and standard deviation 1 (except for the binary check variables)

6.4.2 Evaluating tagger performance

6.4.3 PFlow optimization

Hybrid sample definition

In the course of the completion of this thesis, the collaboration switched from using EMTopo jets (which reconstruct a jet from the topo clusters in the calorimeter) to using PFlow jets which cluster a jet using particle flow candidates that takes into account the better resolution of the tracker for reconstructing lower momentum objects. This improves the jet reconstruction algorithm in two ways. (1) We gain a better reconstruction of the jet p_{T} resolution (shown by

- Switching from EMTopo to PFlow
- Main improvement that we expected for b-tagging would be better with the better angular resolution for the jet axis

RNNIP optimization

An illustration of why the task of b-tagging becomes harder at high p_{T} is illustrated in Figure **??**.

Our optimization for the RNNIP training for the PFlow tagger uses 400 hidden units in the ... LSTM cell. This is an increase from the 50 hidden units of [29] since training over the much larger dynamic range needed a more complex architecture. **Do I remember what lr I used?** The

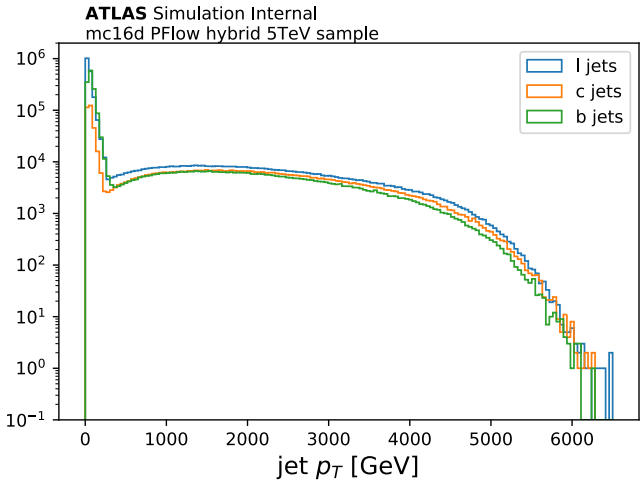


Figure 6.17: The p_{T} spectrum for training the Full Run 2 FTAG recommendations.

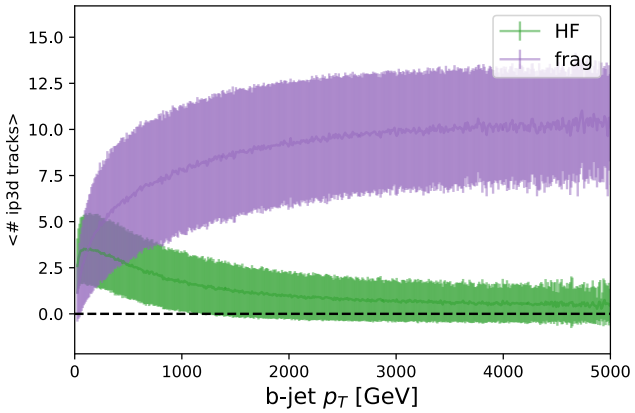
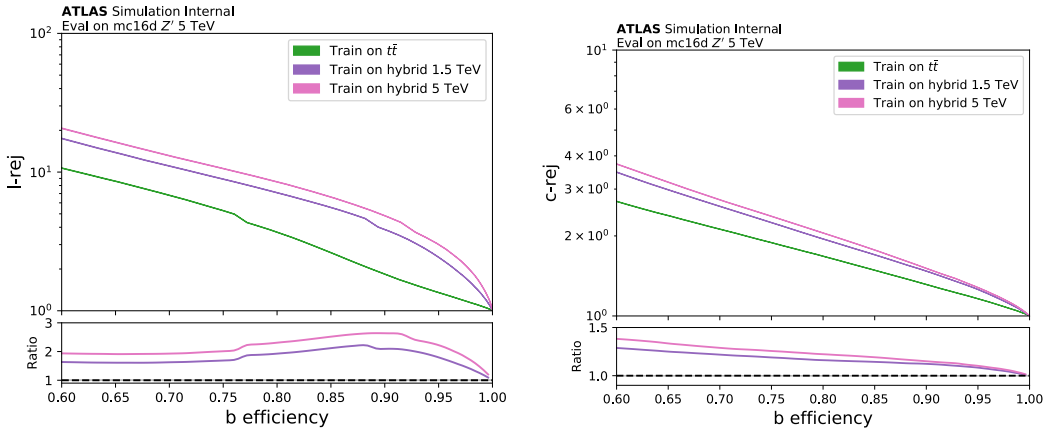
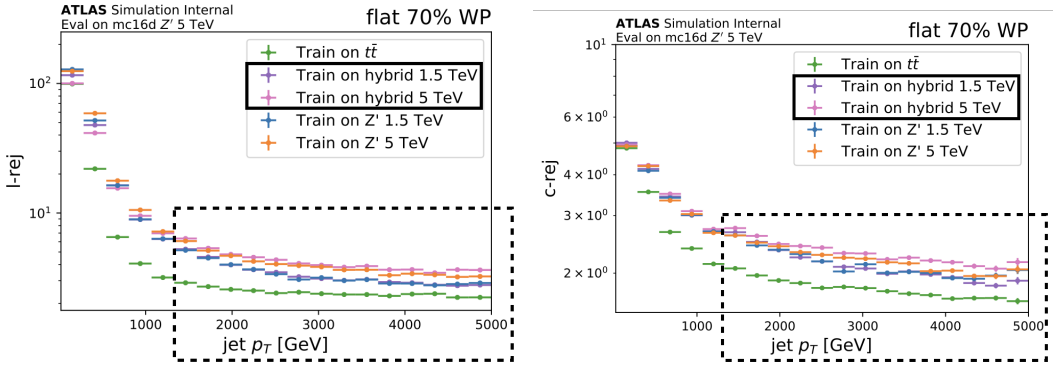


Figure 6.18: The evolution of heavy flavor compared to the number of fragmentation tracks that have $p_{\mathrm{T}} > 1$ GeV, $|d_0| < 1$ mm, $|z_0 \sin\theta < 1.5$ mm, .



(a) light rejection (b) c rejection

Figure 6.19



(a) light rejection (b) c rejection

Figure 6.20

training was done with the adam optimizer, using 5 million jets, and 20% of the dataset was held out as a validation set, and the training was stopped when the validation loss had not improved in the last 20 epochs.

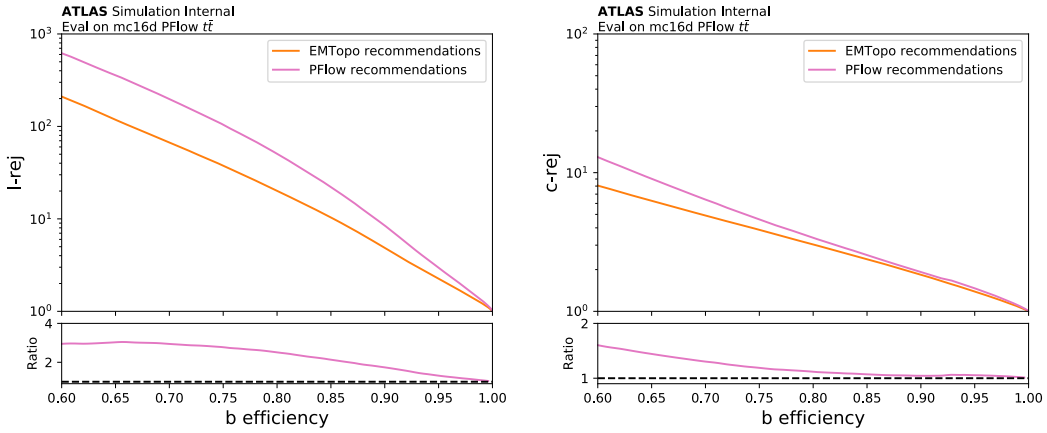
The EMTopo training recommendation was from the 2017 retraining campaign:

Rafael showed we see retraining gains due to the kinematics changes with hdamp from mc15 -¿ mc16

Dedicated retraining on new pflow jet collection

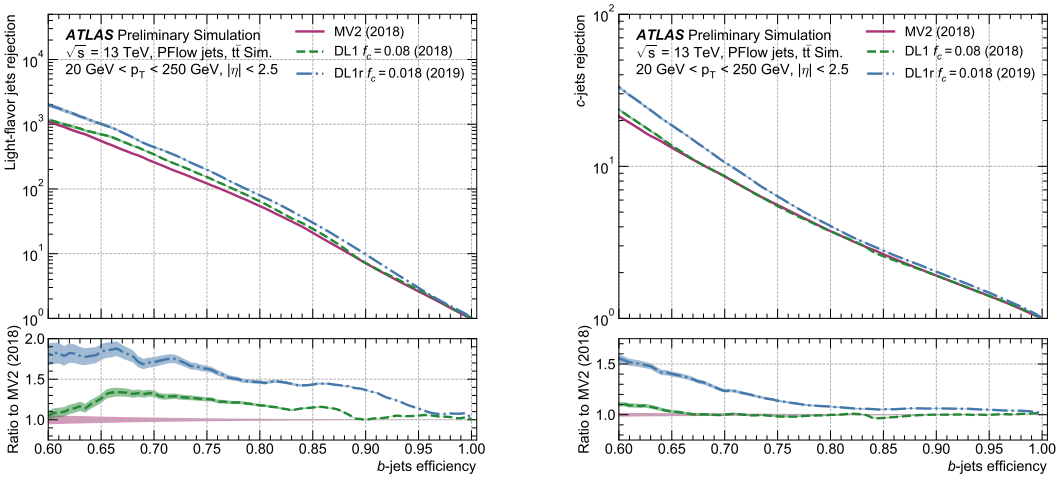
Plus the RNN improvements from this year

DL1r results



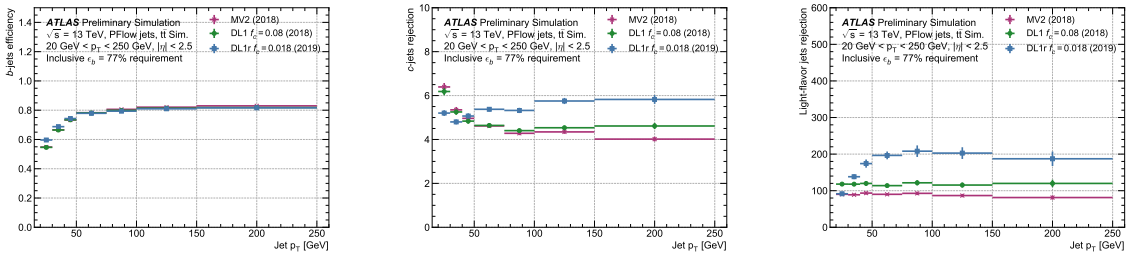
(a) light rejection (b) c rejection

Figure 6.21



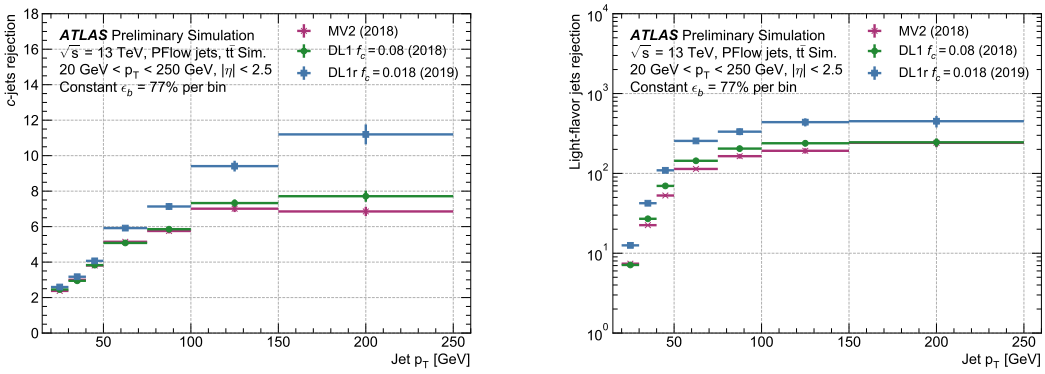
(a) (b)

Figure 6.22



(a) (b) (c)

Figure 6.23



(a) (b)

Figure 6.24

6.4.4 VR track jets optimization

I had to have had a dR matching in this plot . . . let's look up what it was!

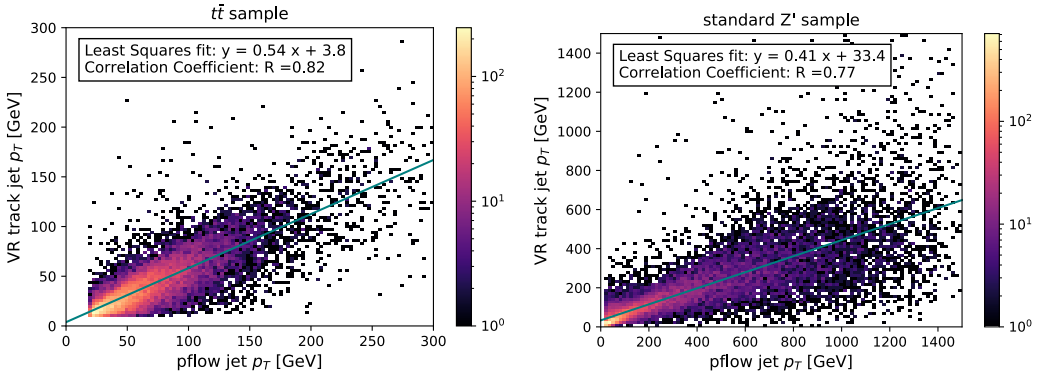


Figure 6.25: Comparison of the PFlow and VR track jet p_{T} for jet reconstruction.

Note: This plot *motivated* us to move the pT cut on the light and c-jets to 125 GeV (although we kept the b-hadron p_{T} cut at 250 GeV).

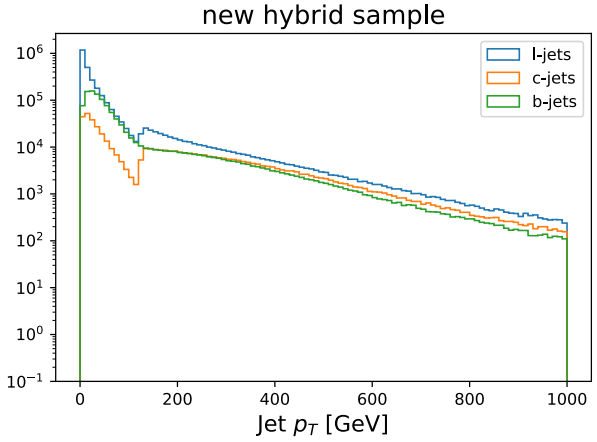
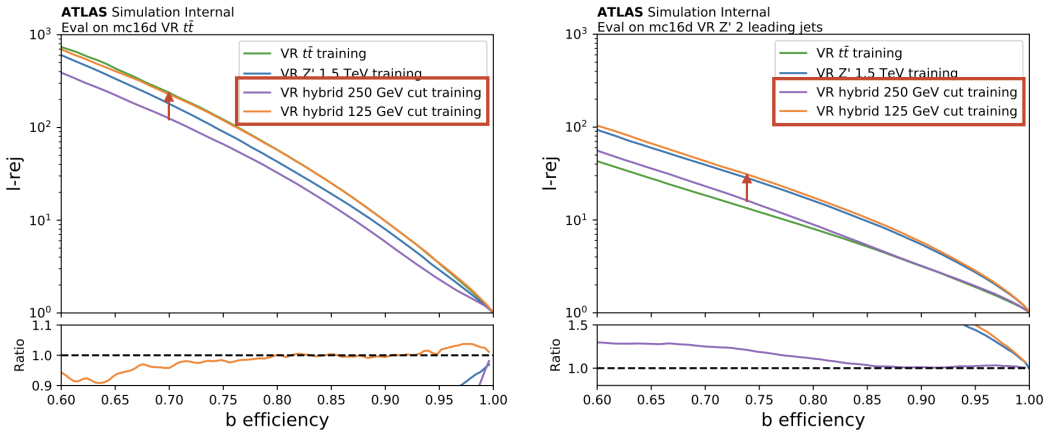


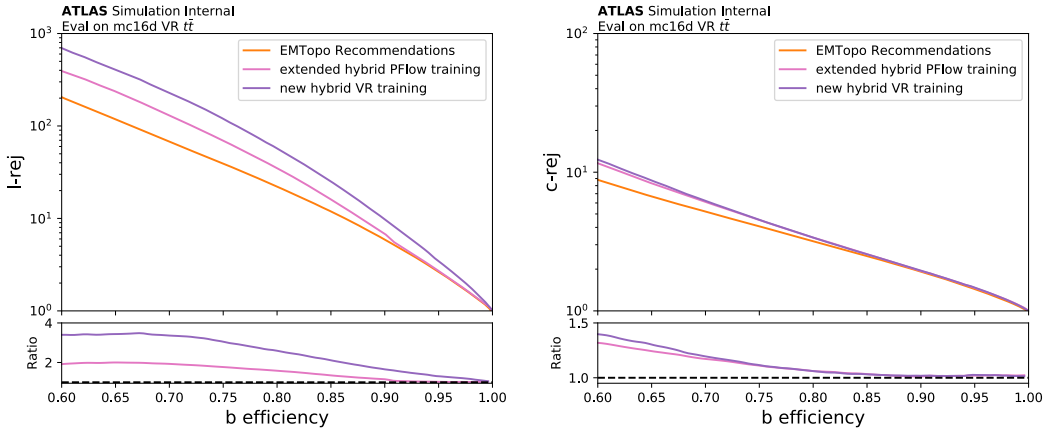
Figure 6.26: The p_{T} spectrum for the VR track jets using the modified sample cut of 125 GeV for the light jet and b-jet p_{T}.

1. EMTopo Rec: What we were applying to VR track jets now
2. Ext Pflow: If we applied the my new pflow training to VR track jets
3. New hybrid VR training: NEW dedicated VR training



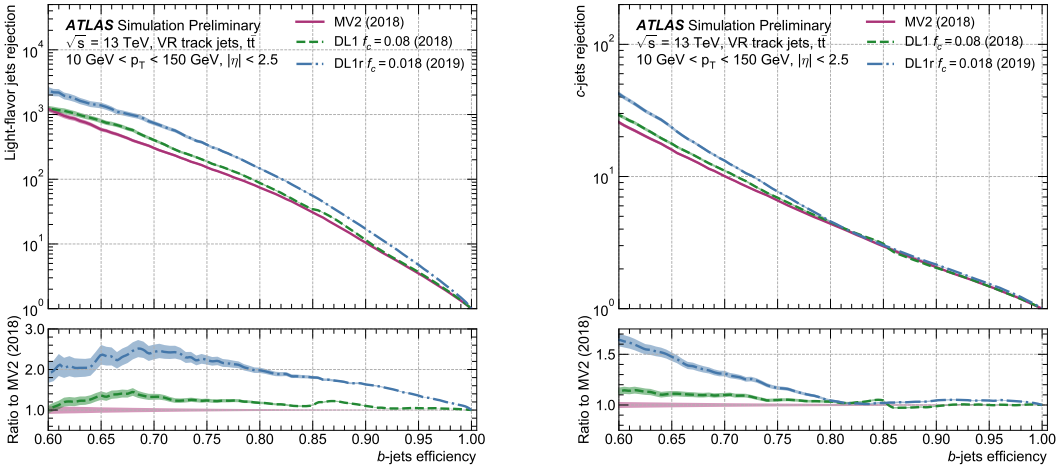
(a) light rejection (b) c rejection

Figure 6.27



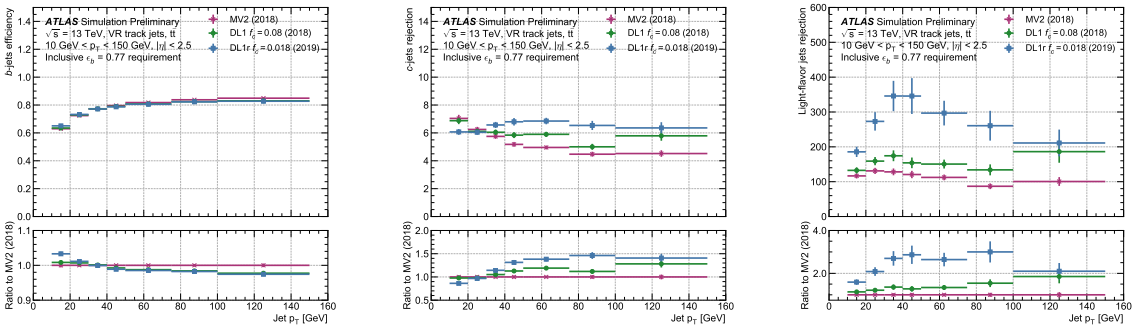
(a) light rejection (b) c rejection

Figure 6.28



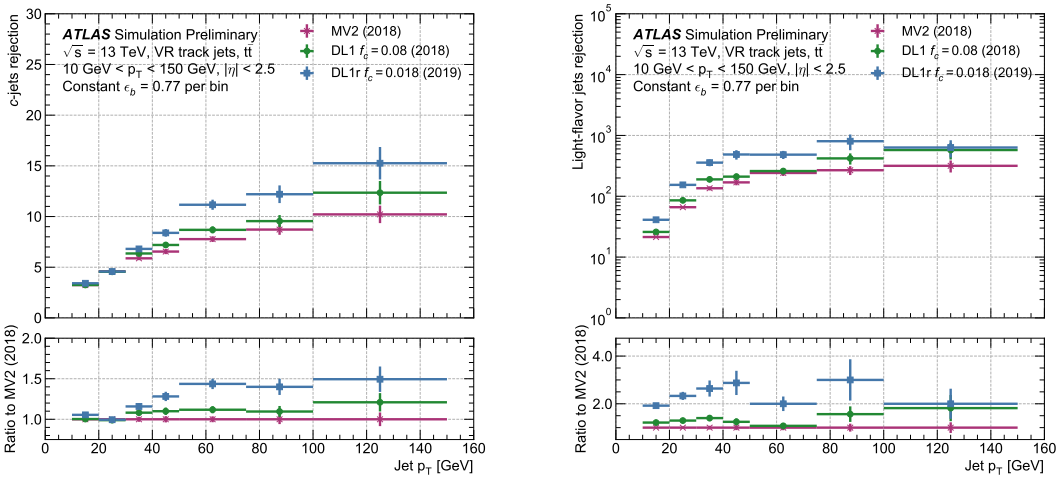
(a) (b)

Figure 6.29



(a) (b) (c)

Figure 6.30



(a) (b)

Figure 6.31

6.4.5 Impact of FTAG improvements on analyses

6.5 RNNIP calibratability

6.5.1 Calibration results

6.5.2 Desiderata for a flipped tagger

6.6 Tagger R&D: DIPS

This work builds on that of the RNNIP algorithm [29], which uses impact parameter information and recurrent neural networks (RNNs) for b-tagging, and provides improvements over other IP-based algorithms by accounting for the correlations between the track features, and the inclusion of additional discriminating variables. tracks in the jet. Here a new algorithm is introduced, Deep Impact Parameter Sets (DIPS), based on the Deep Sets architecture [33] and on the application of the Deep Sets formalism within particle physics known as Energy / Particle Flow Networks [34]. DIPS solves the same task as RNNIP but treats the tracks in the jet as an unordered, variable-sized set rather than as a sequence, avoiding the need to specify a sequence ordering and the slow processing of RNNs. Given that the b-hadron decay products do not exhibit any intrinsic sequential ordering, the Deep Sets architecture is also better physically motivated.

DIPS is demonstrated to be as performant as RNNIP but faster to train, decreasing evaluation

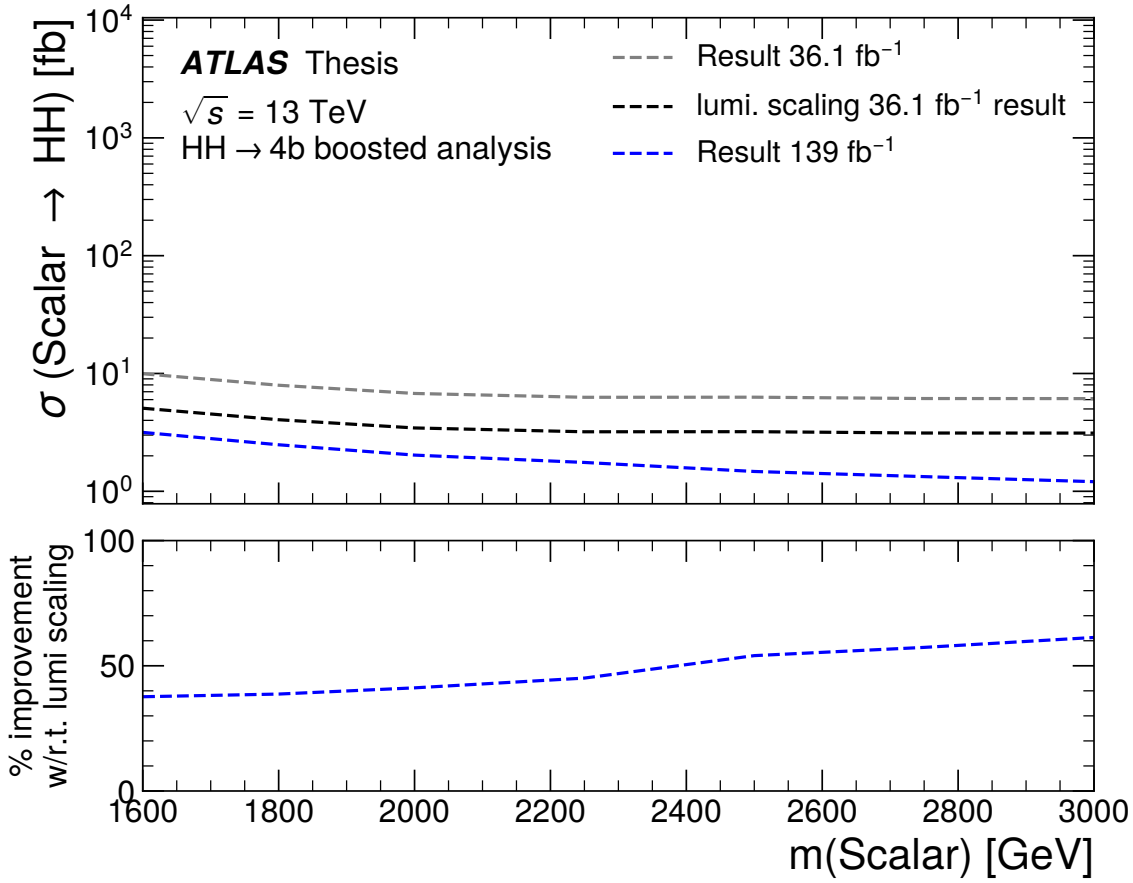
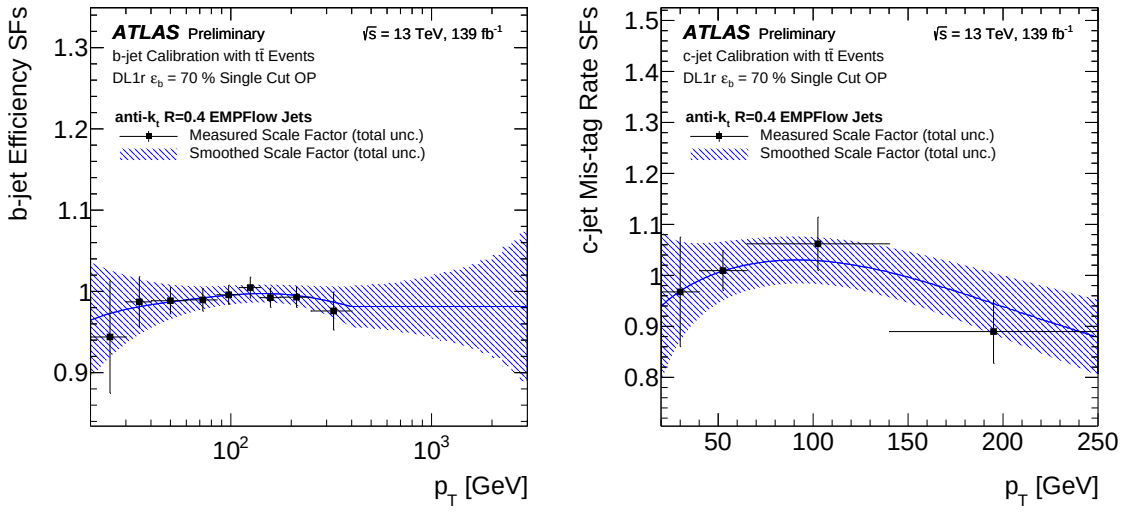


Figure 6.32: Need to cite the 36 ifb and 139 ifb.



(a) (b)

Figure 6.33

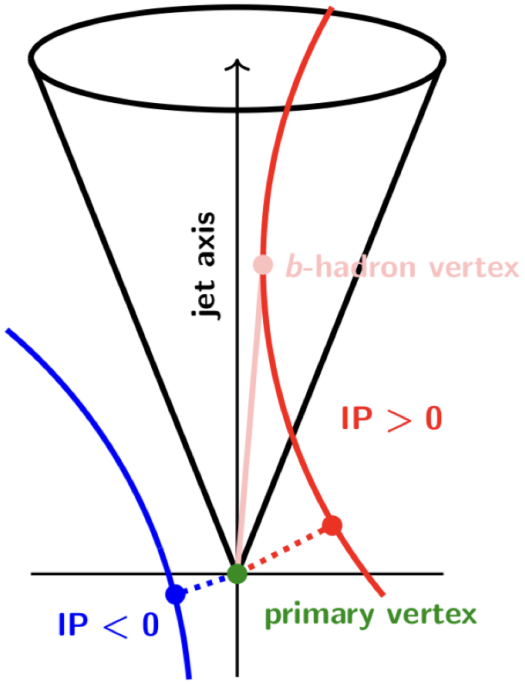


Figure 6.34: Illustration of how the lifetime signage is less likely to be negative for a long lived particle (from Andy Buckley's slides).

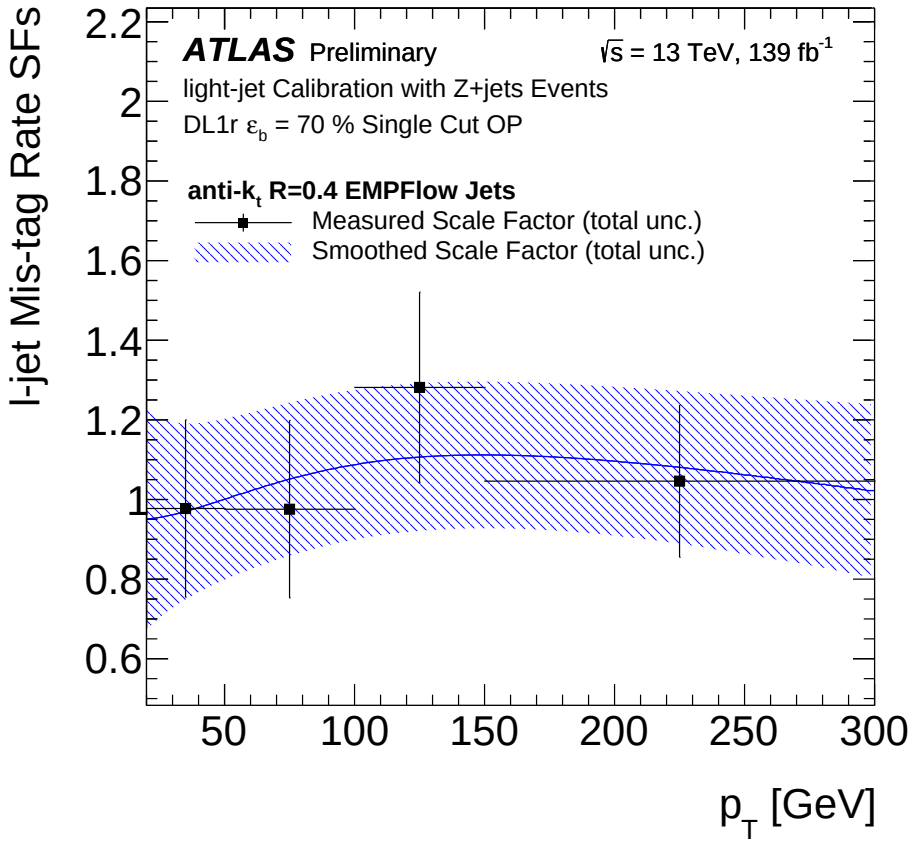

ATLAS Preliminary
√s = 13 TeV, 139 fb-1
light-jet Calibration with Z+jets Events
DL1r εb = 70 % Single Cut OP
anti-kt R=0.4 EMPFlow Jets
Measured Scale Factor (total unc.)
Smoothed Scale Factor (total unc.)
l-jet Mis-tag Rate SFs
pT [GeV]
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
50
100
150
200
250
300

Figure 6.35

time and reducing turn-around time for optimization. Therefore, optimization studies of the track selection criteria and new track features are also included. In addition, a discussion on how to measure the algorithm's efficiency in data, in particular for jets that do not contain a b or a c-hadron, is presented. Finally, one avenue of research in deep learning models is exploring the interpretability of the models, or trying to dissect what information the network is learning. Diagnostic studies from the machine learning literature are presented to demonstrate the well-known characteristics from b-quark fragmentation and hadronization process that the network has gleaned.

6.6.1 Algorithm overview

The Deep Sets architecture [33], which treats the elements as a set without any specific order, maintains the benefits of the RNNIP algorithm, while avoiding the required element ordering (which for b-tagging is empirically driven rather than strictly dictated by the inputs of the problem). This architecture was first employed in particle physics in a phenomenological study on the identification of different types of jets [34]. Adopting the formalism of [34], if p_i is the vector representing the inputs associated with the i^{th} track in the jet, then the Deep Sets architecture applies a neural network (NN) Φ to each track, sums over the tracks, and then applies additional processing on the summed representation with a feed forward NN F, as described in equation 6.9,

$$\mathcal{O}\left(\{p_1, \ldots, p_n\}\right) = F\left(\sum_{i=1}^{n} \Phi\left(p_i\right)\right), \tag{6.9}$$

where $\mathcal{O}\left(\{p_1, \ldots, p_n\}\right)$ represents the b-, c-, and light-flavour class probabilities derived from the inputs for the n tracks in the jet. The architecture bifurcates the problem into operations over inputs and operations over sets, where the track-network Φ extracts the relevant track features, and the jet-network F accounts for the correlations between the tracks. The permutation invariance of the set is encoded with the permutation invariant sum operation, although other permutation invariant operations such as the max or average could be used as well. The presence of this aggregation layer in the architecture encodes information about track multiplicity inside the jet, which is a useful information for identifying b-jets.

This Deep Sets architecture offers the same advantages as RNNIP but encodes permutation invariance between the tracks in the jet, giving a more natural representation of the data and allowing the algorithm to be trained more efficiently with fewer parameters and less data [35]. In addition, Deep Sets offers a major additional advantage over RNNs in that the operation of processing the tracks in the jet with the Φ network can be easily parallelised. This allows training and evaluation to make significantly more efficient use of GPUs over the non-parallelisable iterative processing of the RNN. The timing performance comparison between DIPS and RNNIP is further discussed in Section 6.6.5.

6.6.2 Implementation details

All algorithms are trained with a sample of simulated $t\bar{t}$ events (described in Section **??**) for multi-class classification between b-jets, c-jets and light-flavour jets. To avoid classification based on the differing kinematic spectra of the jet classes, the p_{T} spectra for b-jets and c-jets is reweighted to the light-flavour jet spectra, as described in reference [29].

The class probabilities predicted by the model outputs (p_b, p_c, and p_l), are combined into a b-tagging discriminant:

$$D_b = \log \frac{p_b}{(1 - f_c)p_l + f_c p_c} \tag{6.10}$$

where f_c is a free parameter that balances between the rejection of light-flavour vs c-jets for a given efficiency of selecting b-jets, and can be optimized post-training. A value of $f_c = 0.07$ was used in these studies as this is representative of the fraction of c-jets relative to non b-jets in $t\bar{t}$ events.

For the timing comparisons in Section 6.6.5, the same input features are used for both RNNIP and the DIPS. The features used in each algorithm are described in Table 6.3. The track variables related to the track reconstruction quality focus on the IBL and the next-to-innermost pixel layer (PIX1) due to their strong impact on the IP significance distributions. In particular, the number of split hits, which are hits being created by multiple charged particles [36], is used to help identify dense tracking environments, in which distinguishing tracks from heavy flavour decays is generally more difficult.

After applying the track selections described in Section **??**, the tracks are ordered by decreasing s_{d0}, and the first 15 tracks are kept for processing. The ordering plays a limited role in the algorithm, since typical jets in the topology investigated should have an average number of tracks that is smaller than the maximum allowed number of tracks (see Table 6.8). Since the p_{T}^{frac} and ΔR variables have a tail at larger values, the natural log of the value for these variables is used as the feature in order to improve the convergence time of the training. Variable normalisation to zero mean and unit variance is frequently used for preprocessing of features in ML algorithms. As many of our input variables already have near zero mean, only a subset of the track features are normalised: $\log p_{\mathrm{T}}^{frac}$, $\log \Delta R$, nPixHits, nSCTHits, as well as d_0 and $z_0 \sin\theta$ for the optimised DIPS training.

A simplified scheme of the DIPS architecture is shown in Figure 6.36, which is based on the architecture in reference [34]. A grid search over the hyperparameters including the number of layers in the Φ and F networks, the number of nodes in the Φ and F networks and the dimension of the track latent space revealed similar performance for many different choices of these hyperparameters. Both batch normalisation [37] and dropout [38] were tested, and it was found that batch normalisation was helpful for the DIPS b-tagging performance while dropout was not.

We present a different training for RNNIP than [29] and [39], here training on only $t\bar{t}$ for the timing comparisons. RNNIP processes the inputs using an LSTM cell with 100 hidden units, and after processing the sequence, feeds the vector into a feedforward network with 20 units and a dropout

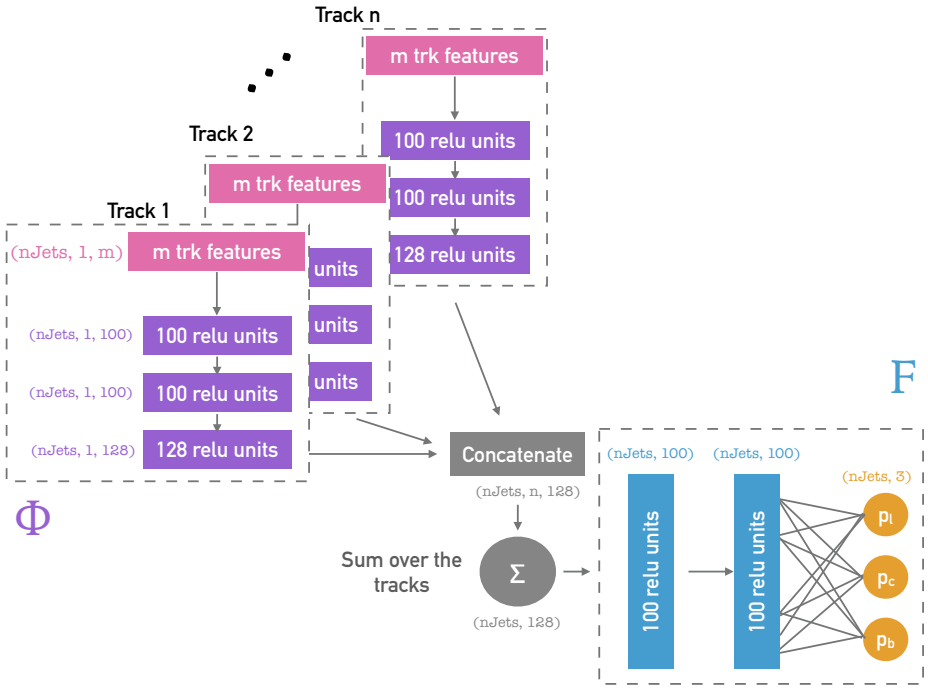


Figure 6.36: Architecture for the DIPS algorithm. The number of hidden units in the different neural network layers correspond to the final optimized architecture.

fraction of 0.2 before a softmax layer for classification.

The RNNIP and DIPS trainings were performed with 3 million jets, with 20% of these jets held out as a validation set to determine when to stop the training. After 10 consecutive training epochs (or iterations through the training dataset) without finding a new validation loss minimum, the training is terminated and the model with the best validation loss was selected. Both the RNNIP and DIPS architectures were implemented in Keras [40] and trained with the TensorFlow backend [41]. Algorithms were trained with the Adam optimizer [42] with a learning rate of 10^{-3} and a batch size of 256. The performance metrics shown in the following sections are obtained with a statistically independent dataset of 3 million jets.

6.6.3 Performance

6.6.4 Baseline Performance

The distribution of the DIPS discriminant D_b (defined in Equation 6.10) for each of the jet flavours is shown in Figure 6.37. The peak at $D_b = -1.3$ is due to jets without any selected tracks. Clear separation between the distribution of b-jets and light-flavour jets can be seen, as well as a strong but smaller separation between b-jet and c-jets as expected due the similarities between b-hadron and c-hadron decays.

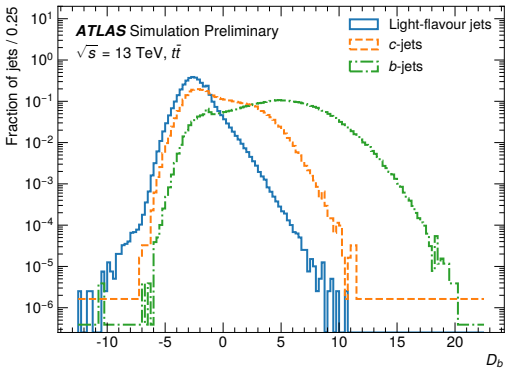
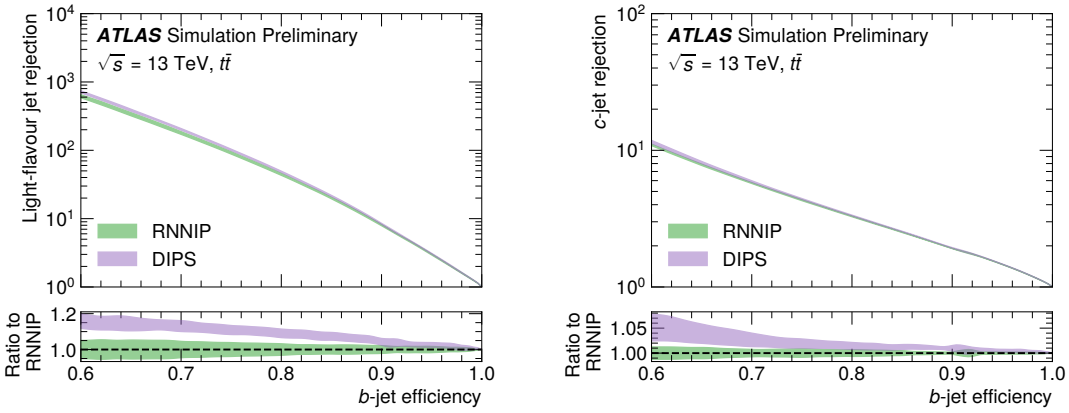


Figure 6.37: Distributions of DIPS b-tagging discriminant, as defined in Equation 6.10, for b-jets, c-jets and light-flavour jets.

The performance of taggers can be examined and compared through a Receiver Operator Characteristic (ROC) curve: a scan is performed for a threshold τ on D_b, and the efficiency for b-jets at each threshold is computed as the fraction of b-jets with $D_b > \tau$, while the rejection of c-jets or light-flavour jets is computed as one over the fraction of c-jets or light-flavour jets (inverse mistag efficiency), respectively, with $D_b > \tau$. The b-jet efficiency and light-flavour (or c) jet rejection

for the same τ are then plotted. Each model is trained five times and for a given b-jet efficiency, the mean of the rejections is used as the nominal value and the standard deviation of the rejections is used for the width of the curve. This ensemble of trainings is known to assess the predictive uncertainty of machine learning-based algorithms [43].

The ROC curves for b-jet efficiency versus light-flavour jet rejection and for b-jet efficiency versus c-jet rejection of the DIPS and RNNIP algorithms are shown in Figure 6.38. The lowest b-jet efficiency displayed corresponds to the lowest efficiency benchmark used in physics analyses within the ATLAS experiment. The DIPS algorithm provides up to a 15% additional light-flavour jet rejection and a 5% additional c-jet rejection at a given b-jet efficiency over the RNNIP algorithm. Notably, as will be discussed in Section 6.6.5, this similar performance comes with a significant decrease in training and evaluation time.



(a) (b)

Figure 6.38: Light-flavour jet rejection as a function of b-jet efficiency (a) and c-jet rejection as a function b-jet efficiency (b) of the RNNIP (green) and DIPS (purple) algorithms. The central curves and error bands show the mean and standard deviation, respectively, of the rejection at each b-jet efficiency for 5 trainings. The ratios are computed with respect to the RNNIP ROC curve.

In order to explore what DIPS is learning in the correlations between features that aids the classification performance, the average saliency map for b-jets with 8 associated tracks and failing a threshold corresponding to 77% b-tagging efficiency is shown in Figure 6.39. The saliency map is computed as

$$\frac{\partial D_b}{\partial x_{ik}} = \frac{1}{N}\sum_{j=1}^{N}\frac{\partial D_b^{(j)}}{\partial x_{ik}^{(j)}}, \tag{6.11}$$

and is the gradient of the discriminant value $D_b^{(j)}$ with respect to each track feature input $x_{ik}^{(j)}$, averaged over jets (j) in a sample of N jets [44]. In this case, the feature inputs are normalized to

zero mean and unit variance, in a similar way to the training procedure. The saliency map gives a linearised view of how the discriminant value is sensitive to changes in the inputs. Figure 6.39 thus shows how this sample of b-jets which failed tagging could be modified to make them more b-jet like. One can see there is a reasonably strong positive gradients for the significances (s_{d0} and s_{z0}) extending up to 5 tracks, which is the average number of charged particle tracks in a b-hadron decay. Beyond 5 tracks, the gradients for all features are either nearly zero or negative, indicating that either these tracks provide no further information or that tracks with large feature values are more indicative of background. In addition, DIPS is highly sensitive to the $\log p_{\mathrm{T}}^{frac}$ and $\log \Delta R$ of the leading s_{d_0} track, which is consistent with the harder fragmentation of b-quarks with respect to light-flavour and charm jets. Interestingly, this strong correlation with $\log p_{\mathrm{T}}^{frac}$ and $\log \Delta R$ for the highest s_{d_0} track also indicates that simply enlarging the IPs of a track in a jet would not directly lead to a jet passing a tagging threshold, as the track must also be consistent with the kinematic expectations from b-jet fragmentation. The gradients for the shared and split hits of the high s_{d0} tracks are strongly negative since tracks formed from random combinations of hits are more likely in highly dense environments. It can also be seen that the correlation with the overall number of hits in the inner most pixel layers, IBL and PIX1, is positive but small. Such features are of high importance to the estimate of the IP and IP resolution. However such information is also encapsulated in the IP significance features which are strongly correlated with the discriminant. We suspect these correlations are observed to be relatively small due to the discriminator heavily relying on the IP significance for the first order estimate of the quality of the track and the track's utility for classification.

6.6.5 Time comparison

A further key comparison metric between the RNNIP and DIPS algorithms is the time needed for training and evaluation. The training time limits the ability to critically perform optimisation tests and compare model variants, while the evaluation time impacts ATLAS reconstruction time when deployed at scale and the ability to use such algorithms in low-latency environments such as the trigger. The DIPS and RNNIP models with comparable numbers of parameters are compared in terms of their speed of training and evaluation in Tables 6.6 and 6.7, respectively. Training comparisons are done on an NVIDIA 2080 Ti GPU, while evaluation comparisons are performed on an NVIDIA Titan X GPU. Five versions of each model are trained and evaluated, and the mean and standard deviation of the training and evaluation time is reported. A significant speed up of more than a factor of 2 for the DIPS algorithm over RNNIP is observed. As training also involves the early stopping procedure, and thus each algorithm may train for a different number of epochs, the training time per epoch is also reported and shows more than a factor of 3 faster speed for DIPS over RNNIP. This is similar to evaluation time, where DIPS is seen to be nearly a factor of 4 faster than RNNIP.

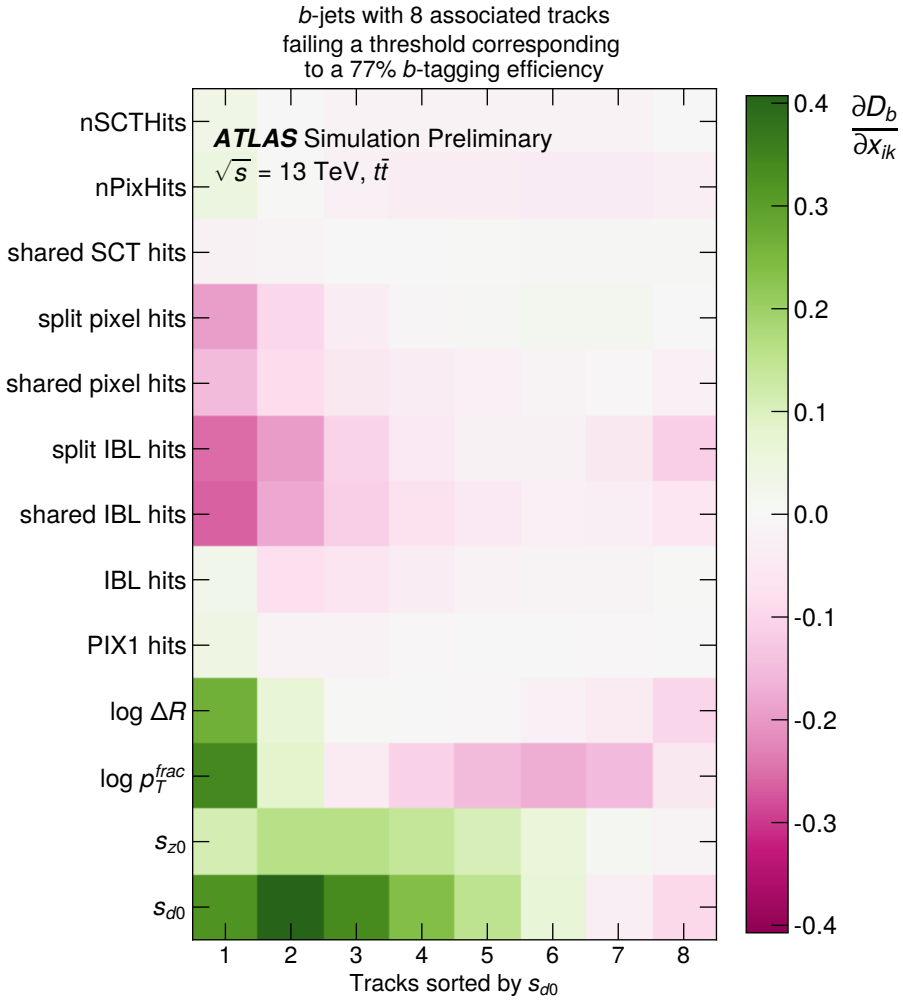


Figure 6.39: Saliency map for b-jets with 8 tracks. The track features are shown on the y-axis, the tracks (ordered by s_{d_0}) are listed on the x-axis. The colors in each pixel represent the gradient defined in Equation 6.11.

Model	**Parameters**	**Training time [min]**	**Time / epoch [s]**
RNNIP	47k	86 ± 13	241 ± 14
DIPS	49k	44 ± 4	78 ± 4

Table 6.6: Timing metrics for trainings performed on Nvidia 2080 Ti GPUs. The nominal value denotes the mean of five independent trainings, while the error bar is the standard deviation.

Model	Parameters	GPU Evaluation time [s]	CPU evaluation time [s]
RNNIP	47k	170 ± 2	685 ± 84
DIPS	49k	46 ± 2	206 ± 98

Table 6.7: Timing metrics for the full test dataset (3 million jets) with GPU evaluations on an NVIDIA Titan X GPU. The nominal value denotes the mean of five independent trainings, while the error bar is the standard deviation.

6.6.6 Calibratability

While performance in simulation gives an important view of an algorithm's performance, ultimately its efficiency must be calibrated to data. This is done using control samples built with specific event selections for each flavour of jet and comparing the observed and simulated efficiency. This is especially challenging for light-flavour jets, as it is difficult to identify a highly pure sample of such jets after the b-tagging requirement.

A large fraction of light-flavour jets are wrongly classified as b-jets due to tracks being on the tail of their IP distribution and are thus mismeasured. This effect is mostly coming from sources, such as detector resolution and pile-up collisions, which have equal probability for mismeasuring a track as having positive or negative lifetime sign, leading to mostly symmetric IP distributions (as seen in Figure 6.8). As such, a data augmentation procedure called *flipping* can be applied whereby the sign of track IPs (and that of secondary vertices) is multiplied by -1, without affecting the overall light-flavour jets IP distributions [45]. The tagger evaluated on flipped inputs, the *flipped* tagger, will then have an approximately equal performance in light-flavour jets as the nominal tagger. However, for b-jets and c-jets with real large IP tracks, the flipping will lead to large changes in their asymmetric IP distribution, with significantly fewer large IP tracks, causing the flipped tagger to be inefficient for identifying these jets. Therefore, applying a b-tagging requirement on the flipped tagger will generate a dataset with a higher fraction of light-flavour jets, when compared to the dataset built with the nominal tagger, such that the light-flavour jet efficiency can be obtained in data. In order for this to succeed, the b-tagging algorithms must uphold this approximate flipping symmetry of the light-flavour jets in their prediction, while reducing b-jets and c-jets tagging efficiencies.

The discriminant distributions of b-jets, c-jets and light-flavour jets with nominal and flipped inputs for the RNNIP and DIPS algorithms are shown in Figure 6.40. The dashed vertical lines represent the discriminant requirement for 85%, 77%, 70% and 60% inclusive b-jet efficiencies, corresponding to the efficiency benchmarks used at analysis level. The desired properties are found for both DIPS and RNNIP, the flipped distribution for light-flavour jets is nearly unchanged, while there is a significant decrease in flipped b-jets and c-jets at high discriminant values. Using these distributions, the efficiencies of the different jet flavours as a function of the RNNIP or DIPS discriminants can be examined, as in Figure 6.41. For both DIPS and RNNIP, one can see the large reduction on the efficiency for selecting b-jets and c-jets for a fixed light-flavour jet rejection as

desired.

6.6.7 Track Selection Optimisation

A major benefit of the reduced training time for DIPS is that it facilitates critical optimisation studies which require retraining the algorithm for each change one would like to examine. Two classes of optimisation are presented here: 1) varying the selection of tracks given to DIPS for processing, and 2) providing additional features per track.

The DIPS implementation described so far relies on the same track selection as the IP3D and RNNIP algorithms. This selection, denoted *nominal*, selects tracks with $p_{\mathrm{T}} > 1$ GeV, $|d_0| < 1$ mm, $|z_0 \sin\theta| < 1.5$ mm. This is a relatively strict selection that is used to keep the number mismeasured and pile-up tracks low, as the IP3D algorithm can be sensitive to such tracks. At the same time, this selection removes some of the key tracks from heavy flavour decays that are vital for classification. With the larger expressive power of the DIPS neural network over the IP3D model, DIPS will have more power to learn which tracks are useful for tagging and thus will potentially be less sensitive to such tracks. As a result, a *loose* selection is examined, defined as $p_{\mathrm{T}} > 0.5$ GeV, $|d_0| < 3.5$ mm, $|z_0 \sin\theta| < 5$ mm, which utilises a lower p_{T} threshold and a wider allowance on the impact parameter thresholds in order to capture more tracks from the heavy flavour decay. In addition, DIPS with the *loose* selection examines up to the 25 highest s_{d_0} tracks, rather than 15 tracks as in the *nominal* selection, to further increase the ability to select tracks from heavy flavour decays.

The average number of tracks of different origin per jet is shown in Table 6.8 for the *nominal* and *loose* selections, and is shown separately per jet flavour. The total number of tracks (n_{trk}), the number of tracks from heavy flavour decays (n_{trk}^{HF}), the number of tracks from hadronisation, excluding those from heavy flavour decays (n_{trk}^{hadr}), and the number of tracks from mismeasurement, material interactions, and pile-up (n_{trk}^{other}), are compared. The *loose* selection increases the average number of tracks per jet from heavy flavour decay by $\approx 15\%$ over the *nominal* selection. However, for all flavours, the *loose* selection also increases the number of fragmentation and other tracks per jet. As can be seen in the ROC curves in Figure 6.42, DIPS with the *loose* selection (shown in pink) outperforms the nominal DIPS (shown in purple) by up to $\approx 40\%$ for light-flavour jet and charm jet rejection.

6.6.8 Optimised DIPS Performance

Beyond the *loose* selection, the impact of adding more per-track features is also examined, namely the impact parameters d_0 and $z_0 \sin\theta$. The DIPS with additional features and *loose* track selection, denoted *Optimised DIPS*, can be seen in orange in the ROC curves in Figure 6.42, compared to a reference of the nominal DIPS or RNNIP trainings, respectively. For the following studies, Optimised DIPS is built with the same architecture described in Section 6.6.2. The Optimised DIPS outperforms

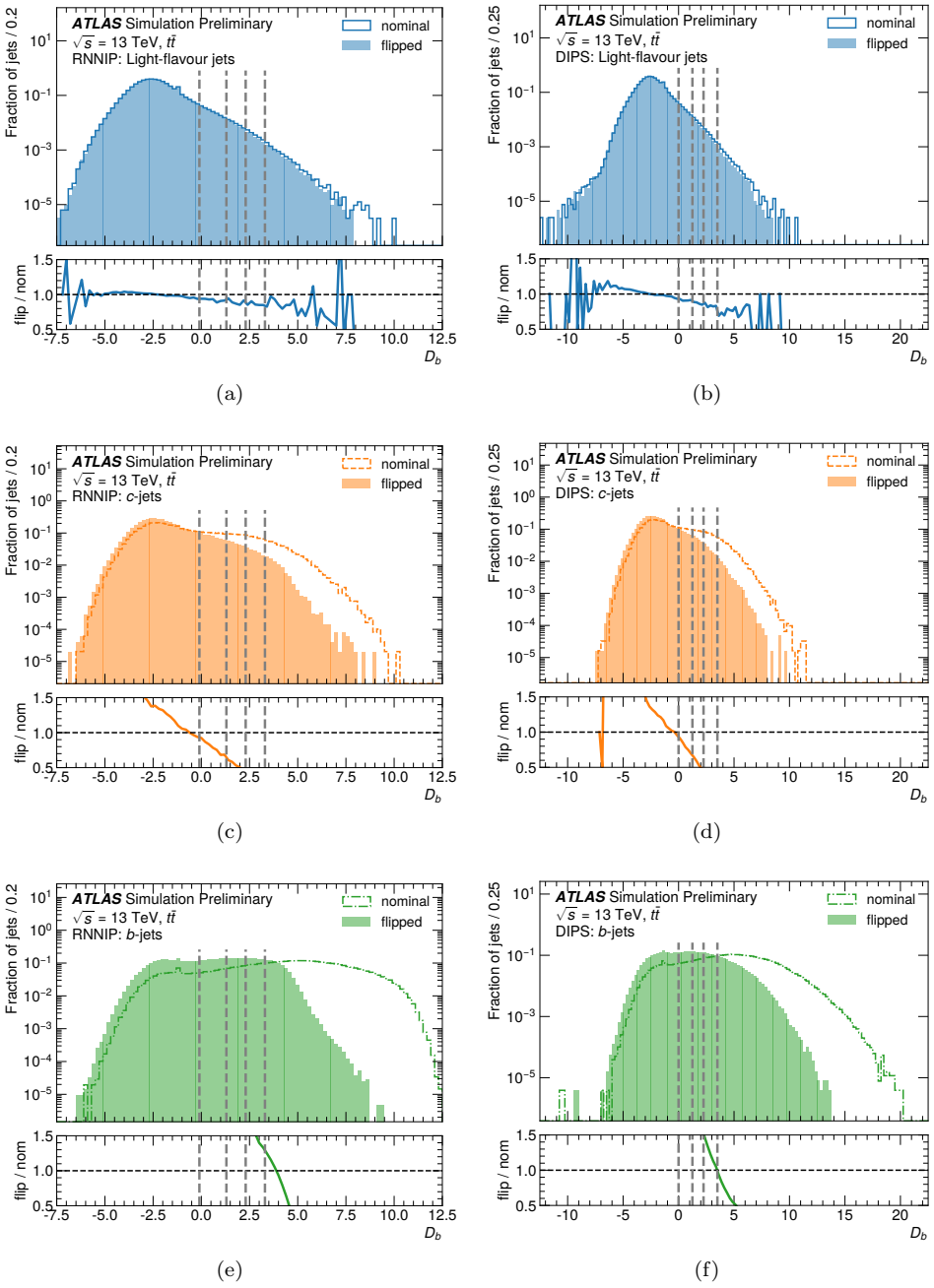
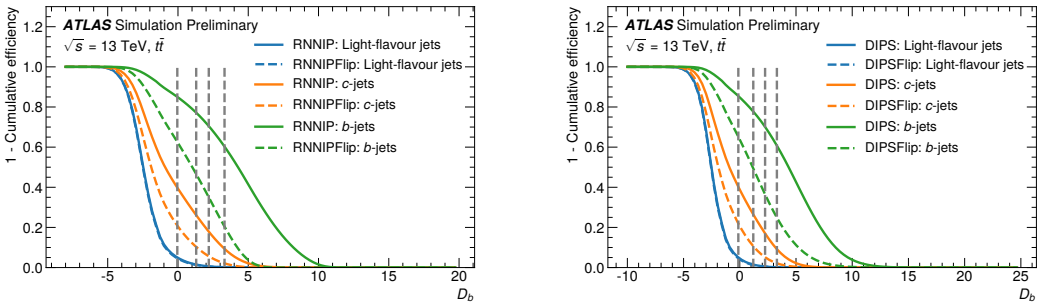


Figure 6.40: D_b discriminant distributions for the nominal and flipped taggers. The vertical dashed lines correspond to the discriminant requirements for 85%, 77%, 70% and 60% inclusive b-jet efficiencies, corresponding to the efficiency benchmarks used at analysis level. Plots (a), (c) and (e) refer to the RNNIP performance, while (b), (d) and (f) refer to DIPS. Plots (a) and (b), (c) and (d), (e) and (f) show light-flavour jets, c-jets and b-jets respectively.

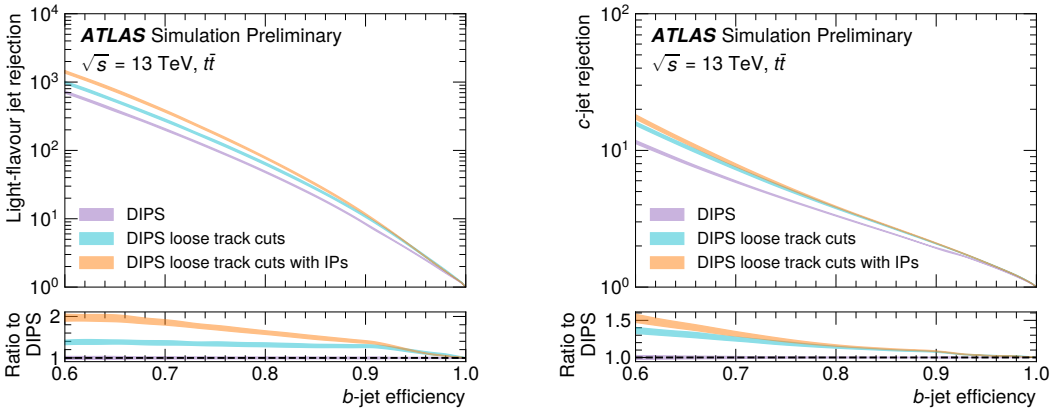


(a) (b)

Figure 6.41: 1 - Cumulative efficiency as a function of b-tagging discriminant for RNNIP (a) and DIPS (b). In both cases, the performance remains nearly unchanged for light-flavour jets when comparing nominal and flipped taggers, while the b-jet and c-jet efficiencies drop.

Jet Flavour	**Track selection**	n_{trk}	n_{trk}^{HF}	n_{trk}^{hadr}	n_{trk}^{other}
b-jets	*nominal*	5.9 ± 2.7	3.4 ± 1.8	2.0 ± 1.9	0.4 ± 0.8
	loose	8.1 ± 3.2	3.9 ± 1.8	2.5 ± 2.1	1.7 ± 1.7
c-jets	*nominal*	5.1 ± 2.5	1.7 ± 1.0	2.9 ± 2.2	0.4 ± 0.8
	loose	7.1 ± 3.1	1.8 ± 1.0	3.6 ± 2.4	1.7 ± 1.7
Light-flavour jets	*nominal*	4.6 ± 2.6	-	4.1 ± 2.5	0.5 ± 0.9
	loose	6.8 ± 3.3	-	5.0 ± 2.7	1.8 ± 2.0

Table 6.8: The average per jet total number of tracks (n_{trk}), the number of tracks from heavy flavour decays (n_{trk}^{HF}), the number of tracks from hadronisation, excluding those from heavy flavour decays (n_{trk}^{hadr}), and the number of tracks from mismeasurement, material interactions, and pile-up (n_{trk}^{other}), are shown for the *nominal* and *loose* selections for each jet flavour.



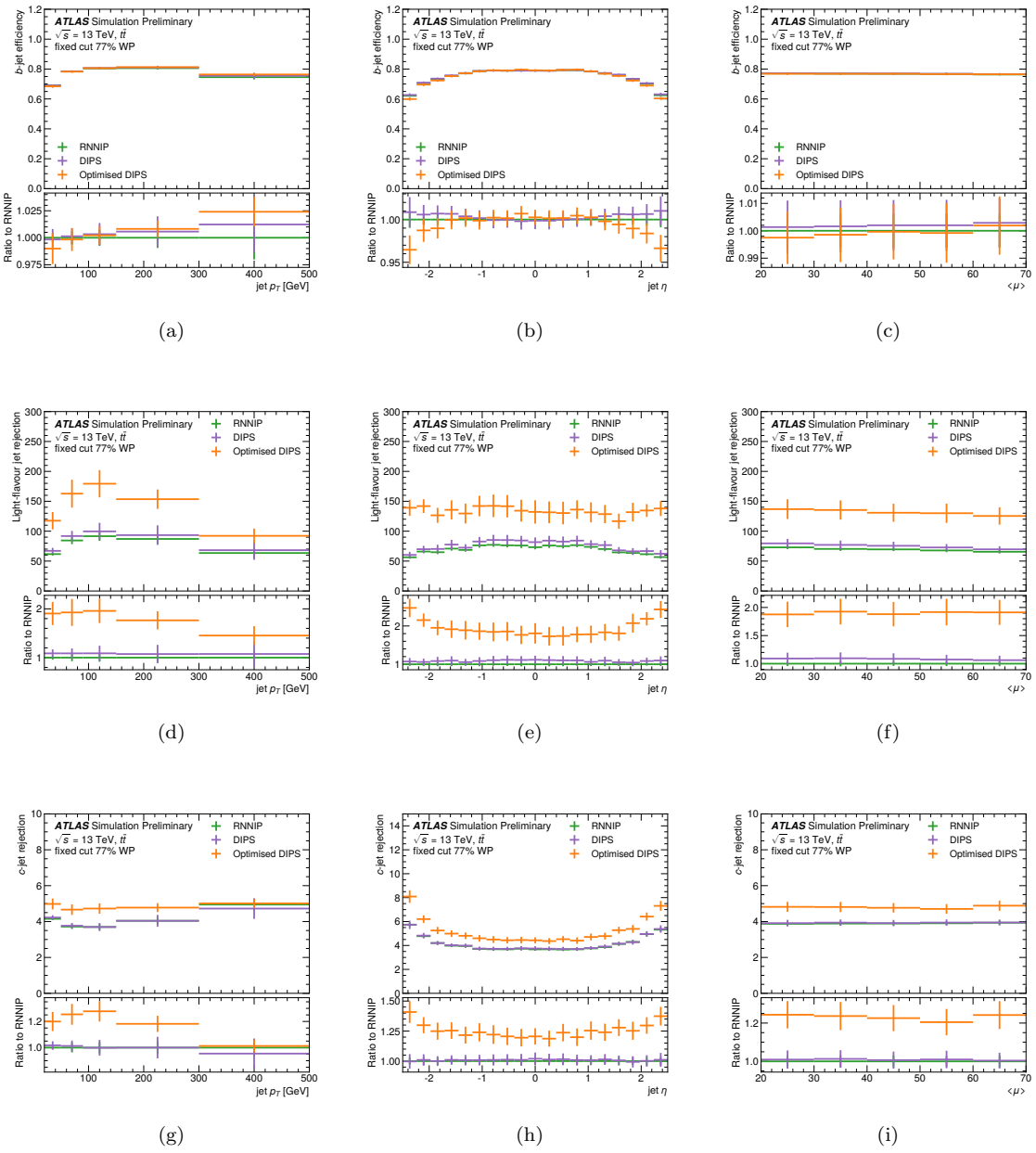
(a) (b)

Figure 6.42: Light-flavour jet rejection as a function of b-jet efficiency (a) and c-jet rejection as a function of b-jet efficiency (b) of the nominal DIPS setup, DIPS with *loose* track selection, and Optimised DIPS with the *loose* track selection and additional IP inputs. The central curves and error bands show the mean and standard deviation, respectively, of the rejection at each b-jet efficiency for 5 trainings. The ratios are computed with respect to the DIPS ROC curve.

the nominal DIPS by up to a factor of 2 in light-flavour jet rejection and a factor of 1.5 in the c-jet rejection.

While ROC curves give a global view of an algorithm's performance, the behavior of the b-tagging efficiency and the background rejection as a function of key kinematic variables is also vital to performance within analyses. To explore this metric, a threshold defining an inclusive 77% b-tagging efficiency for each algorithm is determined, and the b-jet efficiency and background rejections with this fixed threshold are examined as a function of kinematic quantities. The b-jet efficiency as well as the c-jet and light-flavour jet rejections versus jet p_{T} and η are shown in Figure 6.43, for the RNNIP, DIPS, and Optimised DIPS algorithms. The behavior of DIPS and RNNIP are nearly the same across the p_{T} and η range, with DIPS providing a slightly higher light-flavour jet rejection. The Optimised DIPS delivers a factor of 1.5 to 2.5 in additional light-flavour jet rejection and up to $\approx 33\%$ additional charm jet rejection. Loosening the track requirements for Optimised DIPS could potentially have the drawback of increasing the performance dependency on pile-up. We therefore check the b-jet efficiency, c-jet and light-flavour jet rejection as a function of the average number of proton-proton collisions per bunch crossing $\langle\mu\rangle$, also shown in Figure 6.43. The Optimised DIPS performance dependency on $\langle\mu\rangle$ is not found to be significantly stronger than the baseline DIPS or RNNIP.

One challenge in comparing background rejections with a fixed threshold is that the b-tagging efficiency is not the same for each algorithm in each kinematic region. As an alternative, the threshold



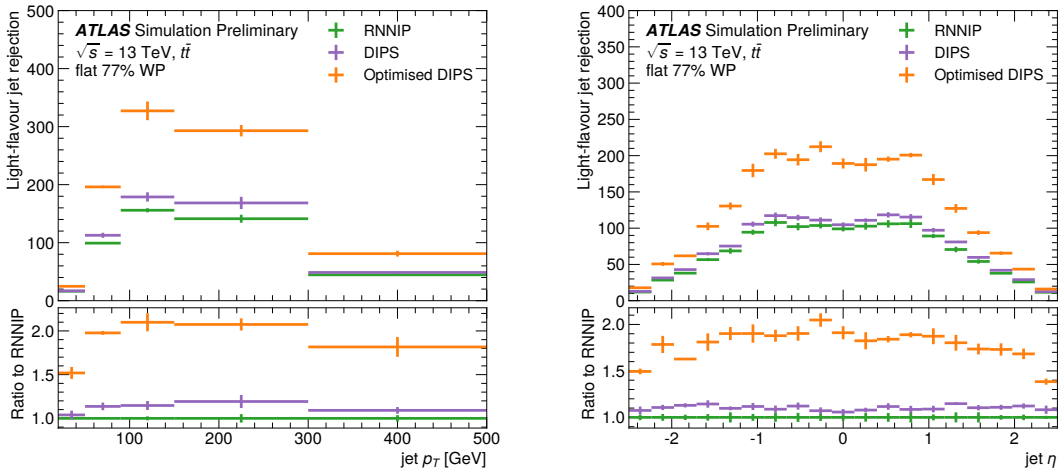
(a) (b) (c)

(d) (e) (f)

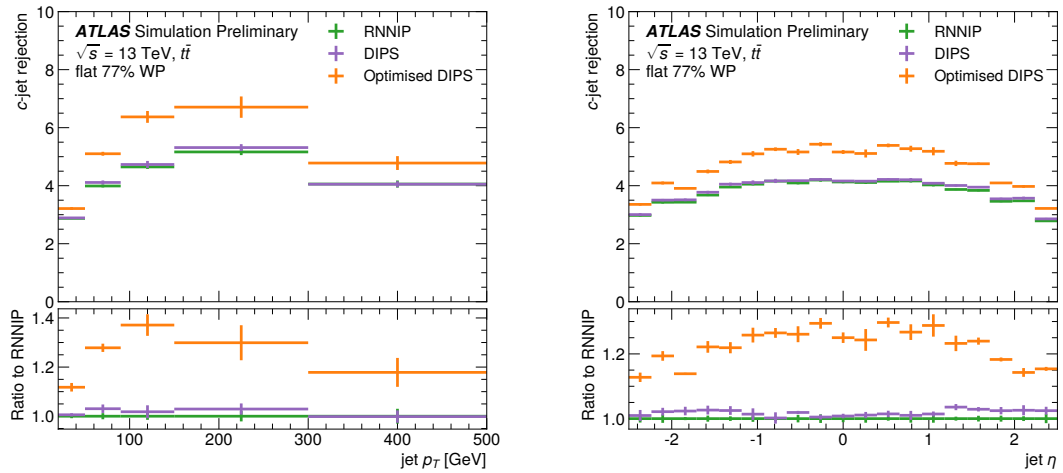
(g) (h) (i)

Figure 6.43: Performance plots using a fixed cut with 77% b-jet efficiency. Plots (a), (b) and (c) show the b-jet efficiency as a function of jet p_{T}, η and average number of proton-proton collisions per bunch crossing $\langle\mu\rangle$. Plots (d), (e) and (f) show the light-flavour rejection as a function of the same quantities, while plots (g), (h) and (i) show the c-jet rejection.

on the b-tagging discriminant can be tuned in each kinematic region to give a constant 77% b-tagging efficiency. A comparison of the c-jet and light-flavour jet rejections as a function of p_{T} and η for the DIPS, RNNIP, and Optimised DIPS algorithms with flat 77% b-tagging efficiency can be seen in Figure 6.44. While DIPS and RNNIP are seen to be quite similar, DIPS provides up to $\approx 20\%$ additional light-flavour jet rejection in some regions of jet p_{T}. The Optimised DIPS shows more than a factor of 2 increase in light-flavour jet rejection and up to $\approx 50\%$ additional charm jet rejection of the DIPS, for jets with p_{T} between 50 and 300 GeV.



(a) (b)



(c) (d)

Figure 6.44: Performance plots using a requirement where the b-jet efficiency is 77% in each bin. Plots (a) and (b) show the light-flavour rejection as a function of jet p_{T} and η, while plots (c) and (d) show the c-jet rejection as a function of the same quantities.

7

Statistical techniques

7.1 Hypothesis testing: a qualitiative introduction

The task of an analysis is to search for a signal in the presence of the plethora of background processes that we also observe at the LHC. In the process of "making a discovery" sort of the bread and butter of day-to-day experimental work is the mathematical rigor of setting up

1. How to characterize the uncertainites of all our experimental parameters that affect our measurement. errors

From statistics, we know that we *make statements* by doing hypothesis testing testing one hypothesis against another.

We define a "null hypothesis" (H0) and an "alternative hypothesis H', and use the *likelihood* to quantify how likely the null is to be true under the likelihood. The only thing we can do in a hypothesis test is reject the null in favor of the alternative, so we set up the null

When observing a physics process (i.e, in the Higgs discovery) the null hypothesis was that the Higgs boson did not exist, and the alternative hypothesis was that the Higgs boson existed, so the process of "making a discovery" meant that we rejected the null and accepted the alternative with "5 σ'' significance, of if the null were true (and the Higgs boson did not exist) the probability that the data could look like this is less than 0.000000001. I need to double check the 5 sigma percent value!

might be fun to have some plots to demonstrate what this looks like?

The SM signal that I searched for in my analysis is fantastically low (35 fb), so we don't expect to see it with our current dataset. To this end, we instead fix the signal shape and set limits on the rate (or overall normalization) of the signal process. In this case, the null hypothesis becomes that the signal *exists* "95%" [1]

[1] I'm giving a general description here I will get specific about what test statistic I'm going to use to define this p-value in section 9.3.

7.2 The likelihood

$$\mathcal{L}(\mu,\theta) = \prod_{j=1}^{N} \frac{\mu s_j + b_j}{n_j!} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k} \tag{7.1}$$

- A.k.a, setting limits verses discerning signal
- The likelihood that we set up for a probability distribution

7.3 Test statistic

$$\tilde{q}_\mu = \begin{cases} -2\ln\tilde{\lambda}(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases} = \begin{cases} -2\ln\frac{\mathcal{L}(\mu,\hat{\hat{\theta}}(\mu))}{\mathcal{L}(0,\hat{\hat{\theta}}(0))} & \hat{\mu} \le 0, \\ -2\ln\frac{\mathcal{L}(\mu,\hat{\hat{\theta}}(\mu))}{\mathcal{L}(\hat{\mu},\hat{\theta})} & 0 \le \hat{\mu} \le \mu, \\ 0 & \hat{\mu} > \mu. \end{cases} \tag{7.2}$$

7.3.1 Asymptotic approximation

7.3.2 Types of Nuisance parameters that we use

7.4

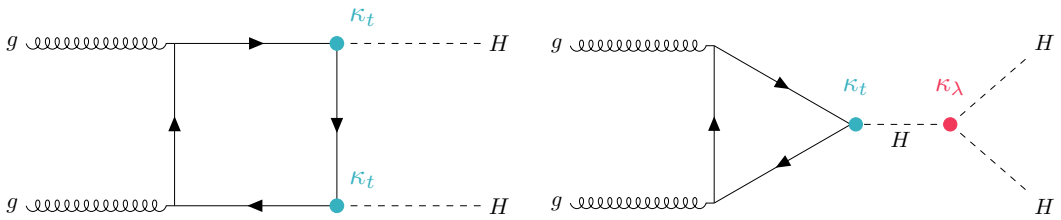
8

HH Physics overview

8.1 HH signatures

Probing the HH self-coupling is extremely interesting, but in the SM, the dominant gluon-gluon-fusion (ggF) cross-section is fantastically low at $\sigma^{\text{SM}}_{ggF\,HH} = 31.05$.[1]. Could I give a rule of thumb for how much smaller this is c.f. other processes at the LHC?

There are two diagrams that contribute to this process at leading order, as shown in Figure 8.1, where there are two diagrams, the box diagram (Figure 8.1) where the top loop radiates two Higgs bosons, and a triangle diagram (Figure 8.1) which is includes the coupling of interest since the Higgs radiated by the top produces another two Higgses by its self-coupling. Although the process is so rare that we won't see it until the HL-LHC [**hh-proj**], we could see it sooner if the Higgs self-coupling deviates from the expectation. In the κ framework, we parametrize the deviations of the couplings from the SM values, i.e, $\kappa_\lambda = \lambda/\lambda_{SM}$, and we can parametrize the deviations of the SM couplings from the other values similarly as well.



(a) The box diagram. (b) The triangle diagram.

Figure 8.1: The leading order gluon-gluon fusion di-Higgs production Feynman diagrams.

In Figure **??**, you can see the contribution from each of the terms individually, and the interference

[1] This includes next-next-to-leading order (NNLO) corrections with an the infinite top limit. The uncertainties of $\sigma^{\text{SM}}_{ggF\,HH} = 31.05\ \pm 3\%\ (\text{PDF}+\alpha_s)\ ^{+6\%}_{-23\%}\ (\text{Scale} + m_{\text{top}})$ fb [**Grazzini'2018**] for a Higgs mass of 125 GeV.

between the two processes. In the SM, this process is suppressed to destructive interference between these two diagrams.

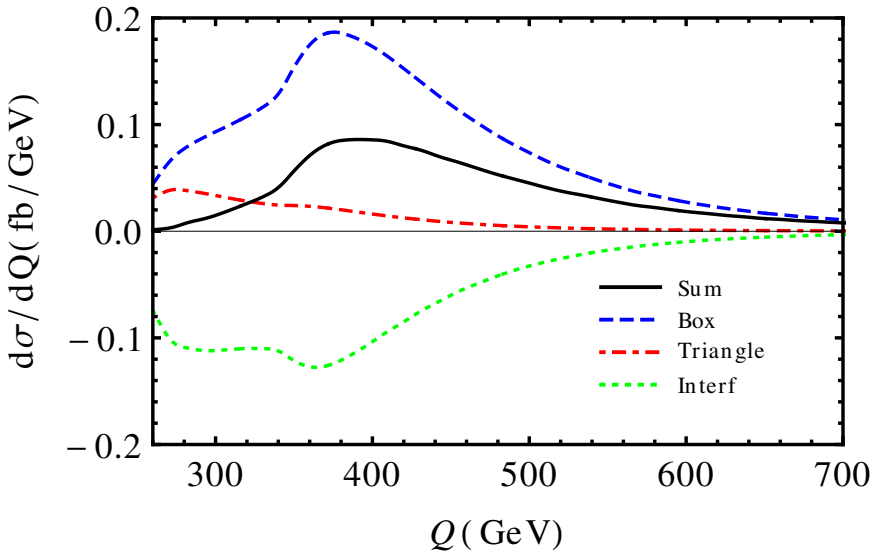


Figure 8.2: Impact of the interference of the box and triangle diagrams for ggF HH production.

from James - need to rephrase sentences The second-leading HH production process is vector boson fusion (VBF), which has a SM cross-section over an order of magnitude smaller than ggF, at $\sigma^{SM}_{VBF\,HH} = 1.726\ \pm 2.1\%\ (\text{PDF}+\alpha_s)\ ^{+0.03\%}_{-0.04\%}(\text{Scale})$ fb [**Dreyer˙2018**] at next-to-next-to-next-to-leading order (N3LO) for a SM Higgs boson with mass of 125 GeV.

8.2 Datasets and signal parametrization

Although it is possible to generate a signal sample for any $\kappa_\lambda$of interest, it is computationally way too expensive to simulate these samples for every signal point in our $\kappa_\lambda$scan, expecially because it isn't necesssary.

The key idea is we can generate the differential cross-section for any process by exchanging the parametrization for the 3 terms (box, triangle, and interference) that characterize the cross-section by using a set of three basis functions by solving a set of 3 linear equations.

We work through the math below for the ggF case Section 8.2.2, and a similar property holds for VBF as well, but involves solving a set of 6 differential equations.

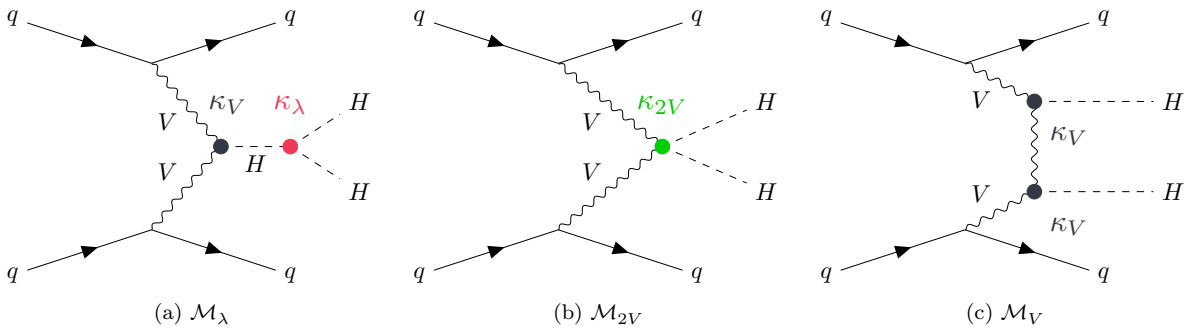


Figure 8.3: The three tree-level vector boson fusion di-Higgs production Feynman diagrams. A convention the matrix elements names is given in the captions of the respective diagrams.

8.2.1 ggF: histogram based reweighting

Although Figure 8.1 shows the LO Feynman diagrams, there is a fully differential next-to-leading-order (NLO) calculation, but the terms in this calculation still break down into diagrams in the box and triangle families. Denoting the sum of the terms in the box and triangle diagrams as B and T, respectively. The impact from a non-SM coupling factors comes as a multiplicative factor for the corresponding vertices, allowing us to write down a combined amplitude (parametrized by κ_t and κ_λ as:

$$\mathcal{A}(\kappa_t, \kappa_\lambda) = \kappa_t^2 B + \kappa_t \kappa_\lambda T. \tag{8.1}$$

The ggF cross section is obtained by squaring the amplitude:

$$\sigma(pp \to HH) = |\mathcal{A}(\kappa_t, \kappa_\lambda)|^2 = \left(\kappa_t^2 B^* + \kappa_t \kappa_\lambda T^*\right)\left(\kappa_t^2 B + \kappa_t \kappa_\lambda T\right) \tag{8.2}$$

$$= \kappa_t^4 \left[|B|^2 + \frac{\kappa_\lambda}{\kappa_t}(B^* T + T^* B) + \left(\frac{\kappa_\lambda}{\kappa_t}\right)^2 |T|^2\right]. \tag{8.3}$$

The κ_t^4 term scales the rate of the process. The 2^{nd} order κ_λ/κ_t polynomial dictates the way these diagrams interfere, so with 3 different $\kappa_\lambda$values, we can get any arbitrary $\kappa_\lambda$value.

To simulate full kinematic sample, consider the basis functions with $\kappa_t = 1$ and $\kappa_\lambda = 0$ (no triangle diagram), $\kappa_\lambda = 1$ and a final (now arbitrary) $\kappa_\lambda = \kappa_0$.

Plugging these basis points into Eq. 8.3 we get three different differential cross-sections:

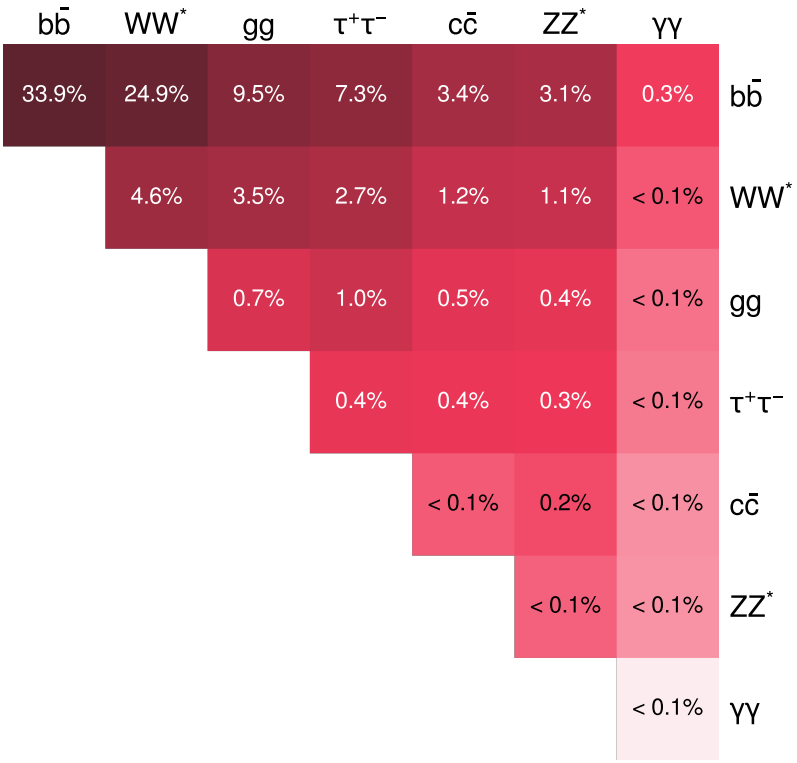


Figure 8.4: Branching ratios of the di-Higgs at the LHC.

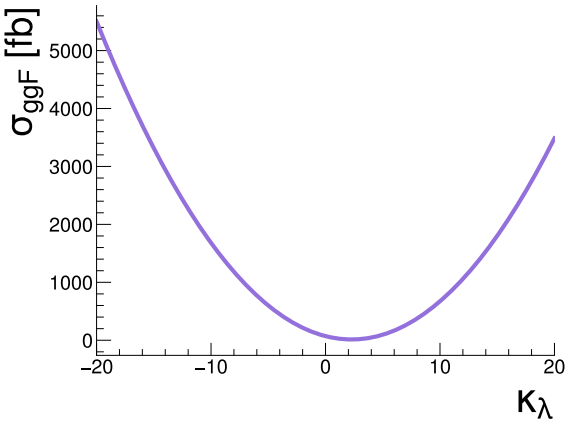


Figure 8.5: Cross-section dependence on κ_λ.

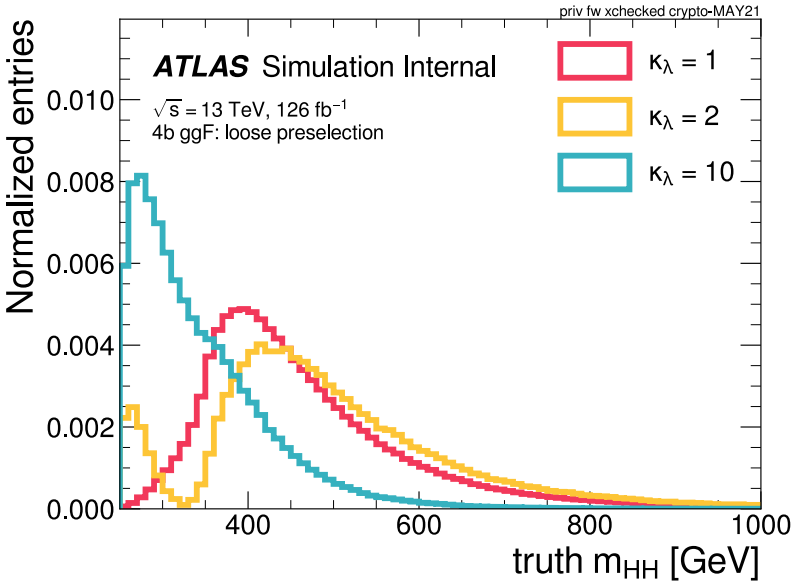


Figure 8.6: Impact of the destructive interference for the $\kappa_\lambda$variations.

$$|\mathcal{A}(1,0)|^2 = |B|^2 \tag{8.4}$$

$$|\mathcal{A}(1,1)|^2 = |B|^2 + (B^*T + T^*B) + |T|^2 \tag{8.5}$$

$$|\mathcal{A}(1,\kappa_0)|^2 = |B|^2 + \kappa_0(B^*T + T^*B) + \kappa_0^2|T|^2 \tag{8.6}$$

Now we solve for $|B|^2$, (B^*T+T^*B), and $|T|^2$ as a function of $|\mathcal{A}(1,0)|^2$, $|\mathcal{A}(1,1)|^2$, and $|\mathcal{A}(1,\kappa_0)|^2$. Define

$$\begin{cases} x = |B|^2 \\ y = (B^*T + T^*B) \\ z = |T|^2 \end{cases} \quad \text{and} \quad \begin{cases} a = |\mathcal{A}(1,0)|^2 \\ b = |\mathcal{A}(1,1)|^2 \\ c = |\mathcal{A}(1,\kappa_0)|^2 \end{cases} \tag{8.7}$$

and write this as a system of linear equations:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & \kappa_0 & \kappa_0^2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad \Longrightarrow \quad a = x \;\text{ and }\; \begin{bmatrix} b-a \\ c-a \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \kappa_0 & \kappa_0^2 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} \tag{8.8}$$

Taking the inverse of the matrix to solve for the remaining two terms:

$$\begin{bmatrix} y \\ z \end{bmatrix} = \frac{1}{\kappa_0(\kappa_0 - 1)} \begin{bmatrix} \kappa_0^2 & -1 \\ -\kappa_0 & 1 \end{bmatrix} \begin{bmatrix} b-a \\ c-a \end{bmatrix} \tag{8.9}$$

$$y = \frac{1}{\kappa_0(\kappa_0-1)}\left[\kappa_0^2(b-a)-(c-a)\right] = -\frac{\kappa_0+1}{\kappa_0}a + \frac{\kappa_0^2}{\kappa_0(\kappa_0-1)}b - \frac{1}{\kappa_0(\kappa_0-1)}c \tag{8.10}$$

$$z = \frac{1}{\kappa_0(\kappa_0-1)}\left[-\kappa_0(b-a)+(c-a)\right] = \frac{1}{\kappa_0}a - \frac{1}{\kappa_0-1}b + \frac{1}{\kappa_0(\kappa_0-1)}c \tag{8.11}$$

Now we are ready to plug these expressions for the box, triangle, and interference terms (x, y, and z) to solve for $|\mathcal{A}(\kappa_t,\kappa_\lambda)|^2$ in terms of Eq. 8.3 in terms of three known cross-sections (a, b, and c):

$$\begin{aligned}|\mathcal{A}(\kappa_t,\kappa_\lambda)|^2 =& \kappa_t^2\Bigg[\kappa_t^2\underline{|\mathcal{A}(1,0)|^2} \\ &+ \kappa_t\kappa_\lambda\left(-\frac{\kappa_0+1}{\kappa_0}\underline{|\mathcal{A}(1,0)|^2} + \frac{\kappa_0^2}{\kappa_0(\kappa_0-1)}\underline{|\mathcal{A}(1,1)|^2} - \frac{1}{\kappa_0(\kappa_0-1)}\underline{|\mathcal{A}(1,\kappa_0)|^2}\right) \\ &+ \kappa_\lambda^2\left(\frac{1}{\kappa_0}\underline{|\mathcal{A}(1,0)|^2} - \frac{1}{\kappa_0-1}\underline{|\mathcal{A}(1,1)|^2} + \frac{1}{\kappa_0(\kappa_0-1)}\underline{|\mathcal{A}(1,\kappa_0)|^2}\right)\Bigg]\end{aligned} \tag{8.12}$$

Shuffling to put the terms in front of the basis functions together:

$$|\mathcal{A}(\kappa_t,\kappa_\lambda)|^2 = \kappa_t^2\left[\left(\kappa_t^2 + \frac{\kappa_\lambda^2}{\kappa_0} - \frac{1+\kappa_0}{\kappa_0}\kappa_t\kappa_\lambda\right)\underline{|\mathcal{A}(1,0)|^2} + \frac{\kappa_0\kappa_t\kappa_\lambda - \kappa_\lambda^2}{\kappa_0-1}\underline{|\mathcal{A}(1,1)|^2} + \frac{\kappa_\lambda^2-\kappa_\lambda\kappa_t}{\kappa_0(\kappa_0-1)}\underline{|\mathcal{A}(1,\kappa_0)|^2}\right]. \tag{8.13}$$

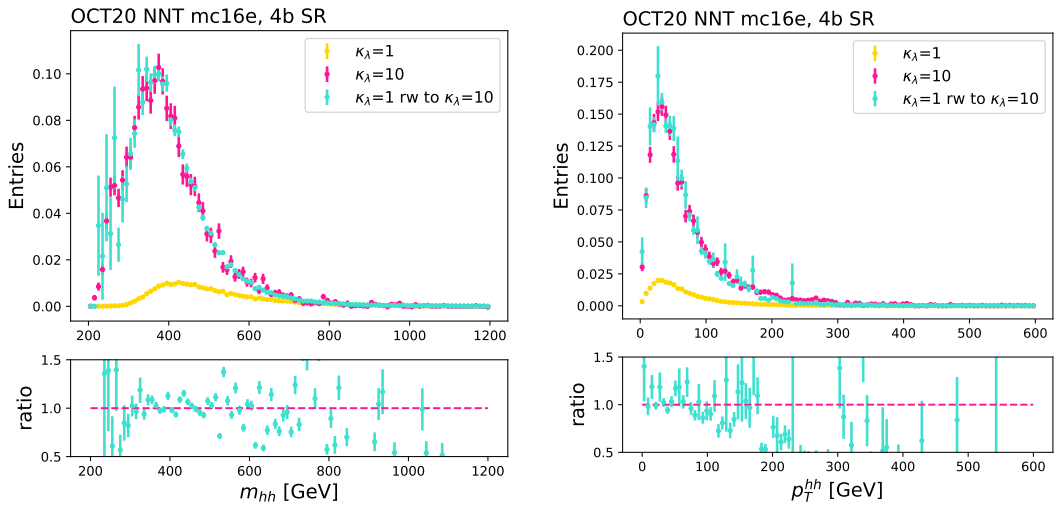
The ATLAS HH analyses use the above prescription with the last basis function as $\kappa_0 = 20$. Also, since we don't constrain κ_t (since we aren't sensitive compared to the ttH analysis), we will set $\kappa_t = 1$ in the following. So simplifying Eq. 8.3 for the combination formula for our implementation:

$$|\mathcal{A}(\kappa_\lambda)|^2 = \left[\left(1 - \frac{21}{20}\kappa_\lambda + \frac{1}{20}\kappa_\lambda^2\right)\underline{|\mathcal{A}(1,0)|^2} + \frac{\kappa_\lambda(20-\kappa_\lambda)}{19}\underline{|\mathcal{A}(1,1)|^2} + \frac{\kappa_\lambda(\kappa_\lambda-1)}{380}\underline{|\mathcal{A}(1,\kappa_0)|^2}\right]. \tag{8.14}$$

Although Eq. 8.14 is fully differential, we just use the m_{HH} differential distribution to derive reweighting functions mapping from the SM ($\kappa_\lambda = 1$) to each other κ_λ signal we test. These signal reweighting functions are derived at the parton level (so before the hadronization, detector effects, and interaction with the detector), and also in the full phase space (or before the analysis cuts). These $\kappa_\lambda$reweighting functions were derived with over a million truth events in the full phase space. This is a simplification to just take into account the m_{HH} performance of the fully differential cross-section, but is an advantageous one because it means we don't need to simulate three large statistics basis samples with the full detector simulation analysis chain, and instead can just use the SM sample (which had 3.8 million simulated events) to get the rest of the $\kappa_\lambda$points. However, to check the veracity of this assumption and the validity in the phase space of the analysis, a κ_λ= 10 sample (with

1.9 million simulated events) was produced and passed through the whole simulation chain, and the modeling of the reconstruction level variables was checked. Figure 8.7 shows a comparison of the reweighting performance in the 4b signal region comparing the actual $\kappa_\lambda = 10$ sample. Since we sew a good closure, we moved forward with this reweighting procedure. The limits for $\kappa_\lambda = 10$ sample versus the reweighted $\kappa_\lambda = 10$ sample, and the agreement was at the %-level, further justifying us moving forward with this.

For completeness, Figure **??** also showed the SM distribution used to reweight these variables, and you can see that the reweighting errors get quite a bit larger in regions of low support for the SM distribution. For the ggF signal, these reweighting errors didn't impact the analysis, but this choice of basis functions to avoid unphysical signal templates was a more challenging task for the VBF analysis.



(a) Reconstructed m_{HH}

(b) Reconstructed HH p_{T}

Figure 8.7: Impact of the signal reweighting for the κ_λ=10 sample. The SM that we reweighted from is shown in yellow, while the reweighted distribution for κ_λ=10 is shown in turquoise and compared to the true κ_λ=10 distribution. Note, I'll want to include the min_dR version of this plot instead of the BDT one.

8.2.2 VBF: event level reweighting

For the VBF differential cross-section at arbitrary coupling values, the methodology is the same as before, except now there are three diagrams to deal with instead of two. These are shown in Figure 8.3, where $|\mathcal{M}_\lambda|$ is the diagram that depends on κ_λ, $|\mathcal{M}_{2V}|$ is the diagram that depends on κ_{2V}, and $|\mathcal{M}_V|$ is the diagram that has two Higgses radiating separately off of the vector boson.

Writing out the differential cross-section and taking into account the scalings with respect to the

couplings:

$$\begin{aligned}\sigma(\kappa_\lambda,\kappa_{2V},\kappa_V) &= \left|\kappa_V\kappa_\lambda\mathcal{M}_\lambda + \kappa_{2V}\mathcal{M}_{2V} + \kappa_V^2\mathcal{M}_V\right|^2 \\ &= \kappa_V^2\kappa_\lambda^2|\mathcal{M}_\lambda|^2 + \kappa_V\kappa_\lambda\kappa_{2V}(\mathcal{M}_\lambda\mathcal{M}_{2V}^* + \mathcal{M}_\lambda^*\mathcal{M}_{2V}) + \kappa_V^3\kappa_\lambda(\mathcal{M}_\lambda\mathcal{M}_V^* + \mathcal{M}_\lambda^*\mathcal{M}_V) \\ &\quad + \kappa_{2V}^2|\mathcal{M}_{2V}|^2 + \kappa_{2V}\kappa_V^2(\mathcal{M}_{2V}\mathcal{M}_V^* + \mathcal{M}_{2V}^*\mathcal{M}_V) + \kappa_V^2|\mathcal{M}_V|^2.\end{aligned} \tag{8.15}$$

There are now 6 terms, so we can pick 6 different choices of the $(\kappa_\lambda,\kappa_{2V},\kappa_V)$ couplings as basis functions to express the kinematics across the full phase space. Although with infinite statistics we would be free to choose any 6 linearly independent couplings to find a basis, in practice with finite statistics, it's important to choose a set of basis functions that is well representative of the BSM phase space to avoid non-physical signal templates, and the corresponding choice basis samples are shown in Table 8.1.

$\boldsymbol{\kappa_\lambda}$	$\boldsymbol{\kappa_{2V}}$	$\boldsymbol{\kappa_V}$
1	1	1
1	1.5	1
2	1	1
10	1	1
1	1	0.5
-5	1	0.5

Table 8.1: Coupling values defining the basis functions for the VBF signal reweighting.

Solving the system of linear equations for these basis points gives the differential cross-section for arbitrary couplings:

$$\begin{aligned}\sigma(\kappa_\lambda,\kappa_{2V},\kappa_V) = &\left(\frac{68}{135}\kappa_{2V}^2 - 4\kappa_{2V}\kappa_V^2 + \frac{20}{27}\kappa_{2V}\kappa_V\kappa_\lambda + \frac{772}{135}\kappa_V^4 - \frac{56}{27}\kappa_V^3\kappa_\lambda + \frac{1}{9}\kappa_V^2\kappa_\lambda^2\right)\sigma(1,1,1) \\ &+ \left(-\frac{4}{5}\kappa_{2V}^2 + 4\kappa_{2V}\kappa_V^2 - \frac{16}{5}\kappa_V^4\right)\sigma(1,1.5,1) \\ &+ \left(\frac{11}{60}\kappa_{2V}^2 + \frac{1}{3}\kappa_{2V}\kappa_V^2 - \frac{19}{24}\kappa_{2V}\kappa_V\kappa_\lambda - \frac{53}{30}\kappa_V^4 + \frac{13}{6}\kappa_V^3\kappa_\lambda - \frac{1}{8}\kappa_V^2\kappa_\lambda^2\right)\sigma(2,1,1) \\ &+ \left(-\frac{11}{540}\kappa_{2V}^2 + \frac{11}{216}\kappa_{2V}\kappa_V\kappa_\lambda + \frac{13}{270}\kappa_V^4 - \frac{5}{54}\kappa_V^3\kappa_\lambda + \frac{1}{72}\kappa_V^2\kappa_\lambda^2\right)\sigma(10,1,1) \\ &+ \left(\frac{88}{45}\kappa_{2V}^2 - \frac{16}{3}\kappa_{2V}\kappa_V^2 + \frac{4}{9}\kappa_{2V}\kappa_V\kappa_\lambda + \frac{152}{45}\kappa_V^4 - \frac{4}{9}\kappa_V^3\kappa_\lambda\right)\sigma(1,1,0.5) \\ &+ \left(\frac{8}{45}\kappa_{2V}^2 - \frac{4}{9}\kappa_{2V}\kappa_V\kappa_\lambda - \frac{8}{45}\kappa_V^4 + \frac{4}{9}\kappa_V^3\kappa_\lambda\right)\sigma(-5,1,0.5)\end{aligned}$$

Although the ggF signal reweighting had good closure for just reweighting based on the truth

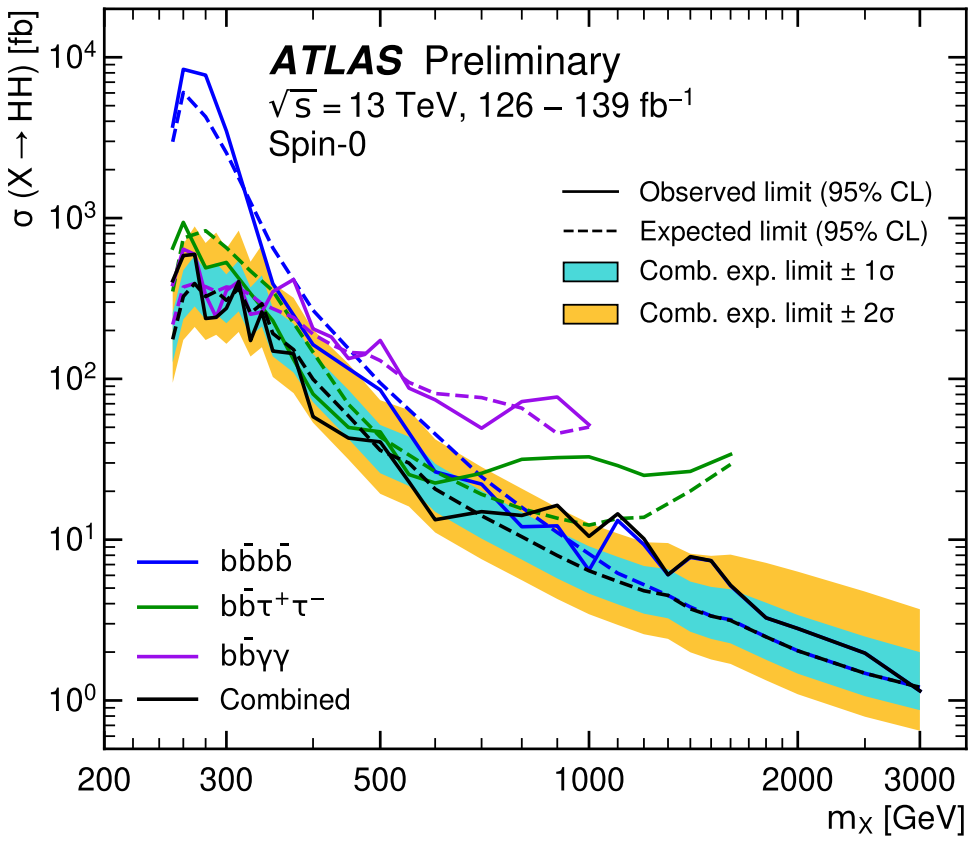


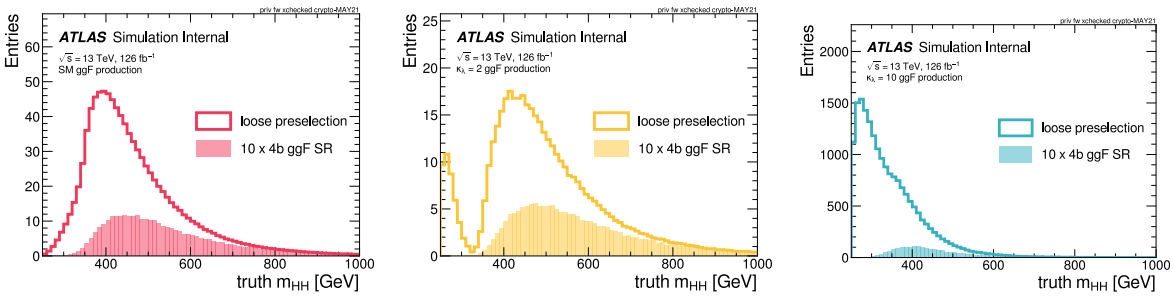
Figure 8.8: Impact of the HH channels in the combination for the resonant scalar mass m_X search [**ATLAS-CONF-2021-052**].

m_{HH} distributions, for the VBF analysis m_{HH} didn't capture the full kinematics of the event that we used to define the analysis cuts and final discriminant, so the VBF signal reweighting is done at the event level using this linear combination of signal samples.

8.2.3 EFTs

8.3 Analysis optimization strategy

- Show how the sensitivity for 4b is *not as great* at low mass



(a) SM (b) $\kappa_\lambda = 2$ (c) $\kappa_\lambda = 10$

Figure 8.9: Impact of the NR analysis selection for selected ggF signals.

9

Analysis selection

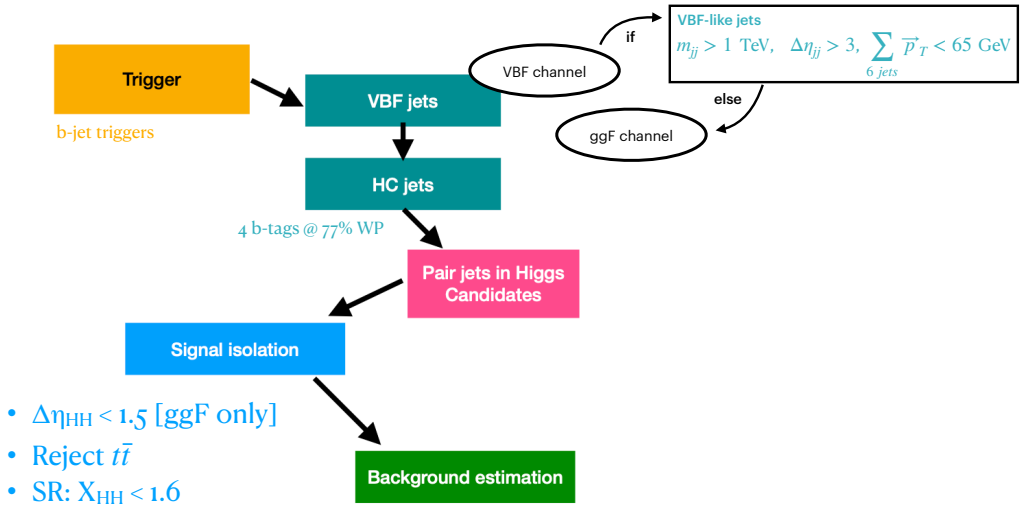


Figure 9.1: Illustration of the high-level analysis strategy.

Jets are separated into two groups based on their kinematics:

***Central* jets:** $|\eta| < 2.5$, $p_{\mathrm{T}} \geq 40$ GeV - these jets are used for triggering and will form Higgs-candidates;

***Forward* jets:** $|\eta| \geq 2.5$, $p_{\mathrm{T}} \geq 30$ GeV - these extra jets are used to improve the acceptance of jets produced in the vector boson fusion production process.

b-jets Selection

In order to maximize sensitivity, events are selected and categorized based on the number of b-tagged *central* jets. The b-tagging algorithm used, DL1r, is described in Section **??**.

Ordering and selection For the jets that form our Higgs Candidates, we take the four leading b-tagged jets. In events with less than four b-jets, the extra jets are selected as the highest p_T jets from the pool of *central* jets which failed the initial b-tag requirement.

***b*-tag Requirements** The b-tag selection for ggF and VBF signals is to require **at least 4 central jets with DL1r 77% WP**. Events with two b-jets are classified as 2b events and are used for deriving the data-driven background estimate described in Section **??**. We also define a systematic on our background estimation using events with 3 b-tags, and the corresponding b-tag categories used in this analysis are summarized in table 10.2.

9.1 Triggers

This analysis uses a combination of multi b-jet triggers. The p_{T} thresholds and b-tagging working points vary slightly by the year of data taking (with the specific cut values delineated in Table 9.1). note - only 2 b-tags are required in the trigger to avoid creating a bias in the control region used in the background estimation that will be described in Section 10.1. The b-tagging SFs are derived for each trigger chain individually required our analysis strategy to specify which trigger stream was considered for the trigger SF application. One other interesting feature of our analysis is our signal is not fully efficient for our analysis, as illustrated by efficiencies that are less than 100% in Fig9.2, and also this efficiency is varying as a function of the reconstructed 4-jet invariant mass.

Trigger Type	Year	HLT thresholds	L1 thresholds
2b1j	2016	$p_{\mathrm{T}} > 100$ GeV jet & two $p_{\mathrm{T}} > 55$ GeV 60% WP b-jets	five $p_{\mathrm{T}} > 15$ GeV jets
	2017	$p_{\mathrm{T}} > 150$ GeV jet & two $p_{\mathrm{T}} > 55$ GeV 70% WP b-jets	$p_{\mathrm{T}} > 85$ GeV jet & two $p_{\mathrm{T}} > 30$ GeV jets
	2018	$p_{\mathrm{T}} > 150$ GeV jet & two $p_{\mathrm{T}} > 55$ GeV 70% WP b-jets	$p_{\mathrm{T}} > 85$ GeV jet & two $p_{\mathrm{T}} > 30$ GeV jets
2b2j	2016	four $p_{\mathrm{T}} > 35$ GeV jets, two 60% WP b-tags	four $p_{\mathrm{T}} > 15$ GeV, $\lvert\eta\rvert < 2.5$ jets
	2017	four $p_{\mathrm{T}} > 35$ GeV jets, two 40% WP b-tags	
	2018	four $p_{\mathrm{T}} > 35$ GeV jets, two 60% WP b-tags	

Table 9.1: Triggers used for non-resonant searches. For b-tagging in the trigger in Run 2, the MV2 version of the b-tagger is used. Also, an L1 $|\eta| < 3.2$ cut is assumed where not specified.

To account for this feature of "operating on the turn on curve" the SF that we apply to account for the trigger effects

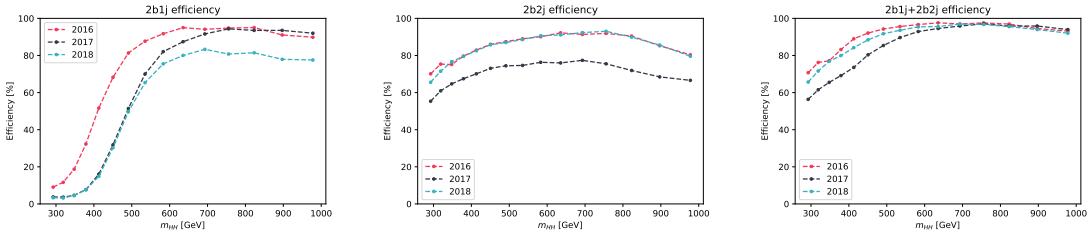


Figure 9.2: Trigger efficiencies of the 2b1j, 2b2j and combined for the MC16a/d/e corresponding to years 2016-2018 for the SM ggF κ_λ=1 signal. Significantly lower efficiency for 2017 2b2j comparing to other years is due to tighter b-tagging requirement (lower efficiency). Is this plot inside of the SR?

9.1.1 Trigger buckets

To distinguish which trigger chain to check, we cut on the offline jets $p_{\mathrm{T},1} > 170$ GeV and $p_{\mathrm{T},3} > 70$ GeV, where the jets are ordered by p_{T}. These jet p_{T} cuts mimic the 2b1j trigger. If the event passes these jet cuts, we put it in trigger **bucket 1**, otherwise it goes in trigger **bucket 2**. In trigger bucket 1, we check the decision of the 2b1j trigger to decide whether to keep the event, and in trigger bucket 2, we check the 2b2j trigger. This procedure is summarized graphically in Figure 9.3.

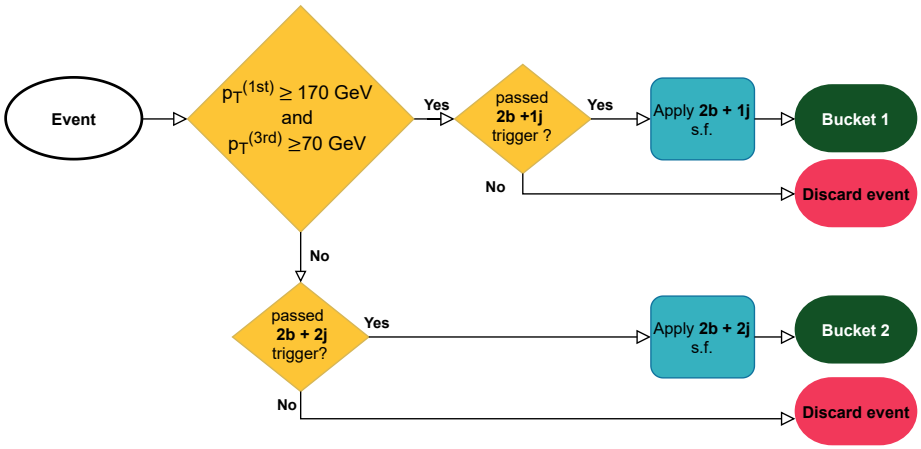


Figure 9.3: Trigger bucket strategy for non-resonant searches.

Figure 9.4 shows how this strategy of using a combination of two triggers gives us sensitivity to complementary phase spaces in the analysis. The 2b1j trigger drives our acceptance for the high m_{HH} events, while the 2b2j trigger provides our low m_{HH} acceptance.

To reconstruct the trigger decision and define the jet level SFs, offline jets are matched to the online jets using a ΔRmatching criterion. These online jets are then checked to pass the (online) thresholds given in Table 2, and if this many jets and b-jets pass this selection, the event passes this trigger. For ease of knowing how to apply the SFs, we will only keep events where the trigger passed

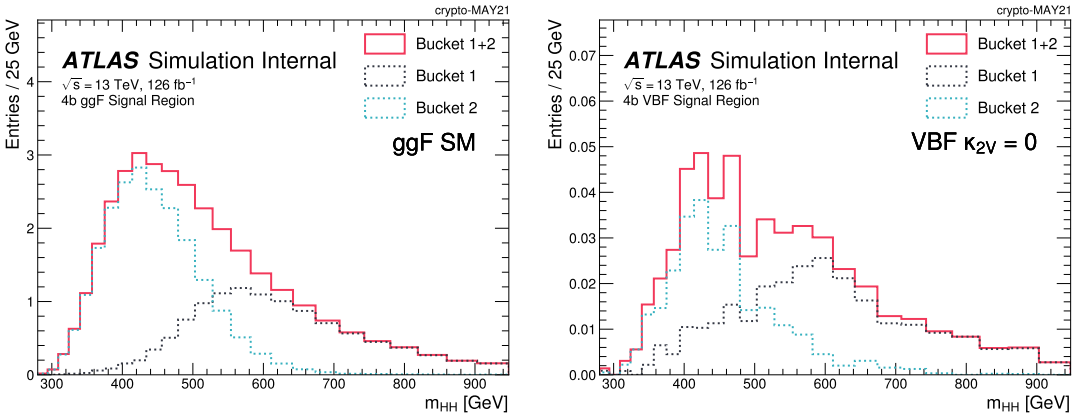


Figure 9.4: The bucket composition of m_{HH} for the SM ggF (left) and $\kappa_{2V}= 0$ VBF (right) HH MC simulation in the 4b Signal Regions. Bucket 1 corresponds to the 2b1j trigger and Bucket 2 corresponds to the 2b2j trigger.

in the relevant bucket, i.e, the 2b1j trigger needs to pass if the event passed the offline cuts in bucket 2, and the 2b2j trigger needs to pass if the event passed the offline cuts in bucket 1. The event level trigger SF is calculated from the jet level SFs (Eq. 9.1) with two contributions:

$$\text{Multi } b\text{-jet trigger SF} = \prod_i SF_{jet}^{b-tag}(i) \times SF_{jet}^{kinematic}(i) \tag{9.1}$$

- b-jet trigger SFs using prescription from by the b-jet trigger group (described in Section 9.1.2)
- Kinematic E_{T} HLT and L1 SFs derived in a custom $t\bar{t}$ analysis (described in Section 9.1.3)

9.1.2 b-jet SF

The offline and online b-tagging decisions are highly correlated, so the online b-tagging SF are derived conditional based on the offline b-tagging decision. Since both the offline and online b-tagging decisions could pass or fail, this gives four cases:

- Case 1: Pass online and offline b-tagging:
$$\varepsilon(\text{on} \wedge \text{off}) = \varepsilon(\text{on}|\text{off})\varepsilon(\text{off})$$
- Case 2: Fail the online b-tag, but pass the offline b-tag:
$$\varepsilon(\overline{\text{on}} \wedge \text{off}) = [1 - \varepsilon(\text{on}|\text{off})]\varepsilon(\text{off})$$
- Case 3: Pass the online b-tag, but fail the offline b-tag:
$$\varepsilon(\text{on} \wedge \overline{\text{off}}) = \varepsilon(\text{on}) - \varepsilon(\text{on}|\text{off})\varepsilon(\text{off})$$
- Case 4: Fail the online and offline b-tagging:
$$\varepsilon(\overline{\text{on}} \wedge \overline{\text{off}}) = 1 - \varepsilon(\text{off}) - \varepsilon(\text{on}) + \varepsilon(\text{on}|\text{off})\varepsilon(\text{off})$$

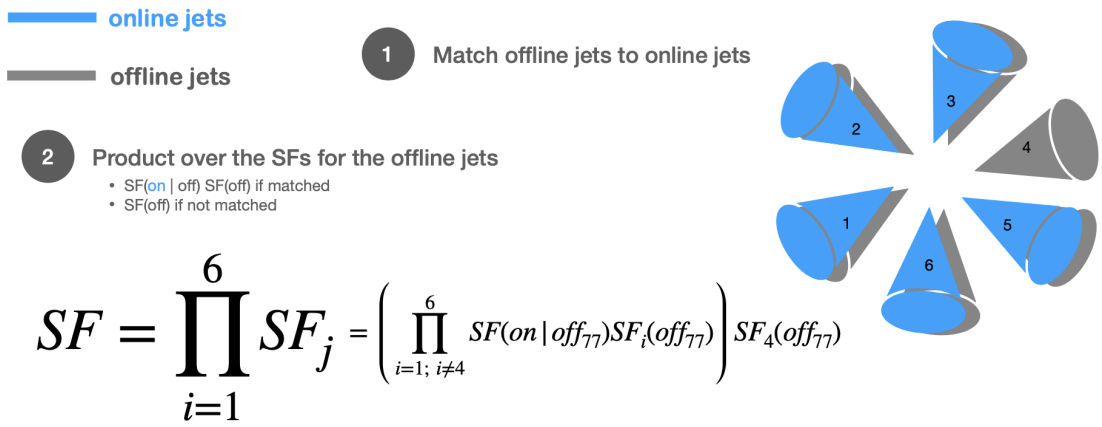


Figure 9.5: Illustration of how the combined offline / online b-tagging SF is calculated.

Then for each efficiency, we still apply $SF = \varepsilon^{data}/\varepsilon^{MC}$. For offline jets that are not matched to a corresponding online HLT jet, just the offline SF is applied, just the offline b-tagging SF is applied, as visualized in Figure 9.5.

Our use of the b-jet triggers dictates SFs dictates how much of the Run 2 dataset we can use.

1. In 2016 there was an issue in the online beam spot calculation, which impacted the primary vertex calculation for the HLT b-tagging. Because of this, we don't use this portion of the data from the 2016 dataset, a loss of 8.3 fb^{-1}from the 32.8 fb^{-1}of the full 2016 dataset.

2. Even for 2017 and 2018, we need to discard the first luminosity blocks of data taking where the beam spot has not yet had time to update. This means analyses with b-jet triggers have $\approx 1.5\%$ lower luminosity in these years than the baseline luminosity [**b-trig-paper**].

3. The astute reader might notice that the 2015 triggers are not included in Table 9.1. As will be explained in Section ch:bkg-est, the background estimate is derived for each year separately to account for the differences in the trigger, and the robustness of the background estimate is partially based on the size of the sample used to derive it. Since it wasn't clear whether the 2015 dataset was large enough to warrant the gains of the additional complexity in the analysis, the 2015 conditional b-jet trigger SFs were never derived with respect to the offline DL1r algorithm, so this year of data is not included.

In summary, when accounting for the above three points, Table 9.2 is the (by year) luminosity for the 4b analysis, with a total luminosity is 126.0 fb^{-1}.

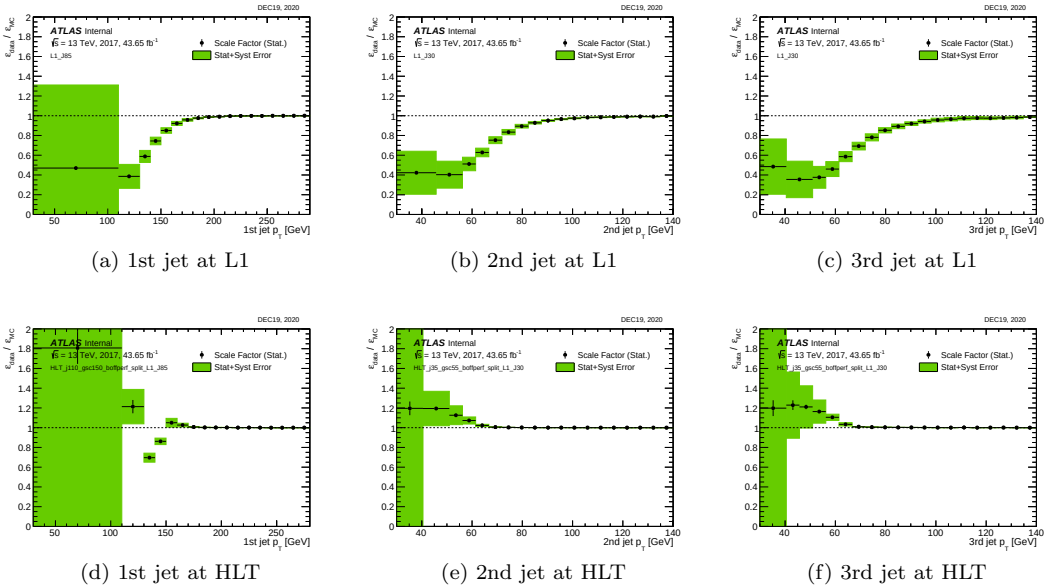
9.1.3 Kinematic SF

Q that I have – how do we apply the jet level SFs? Do we multiply over all of the offline jets in the

Year	Luminosity [fb^{-1}]
2016	24.6
2017	43.7
2018	57.7
all	126.0

Table 9.2: Luminosty (by year) for the 4b analysis.

event, but these are only non-unary for the first N jets ordered by online ET?



(a) 1st jet at L1 (b) 2nd jet at L1 (c) 3rd jet at L1

(d) 1st jet at HLT (e) 2nd jet at HLT (f) 3rd jet at HLT

Figure 9.6: Online jet kinematic scale factors of 2b1j trigger as a function of offline jet p_{T} in 2017. Vertical error bars include statistical uncertainties on the data, while the green bands correspond to the quadrature sum of statistical and systematic uncertainties.

9.2 Muon-in-jet + pt reco

9.2.1 Jets

Jets are clustered using the anti-$\kappa_t algorithm with a distance parameter of R = 0.4$[25]$.This analysis uses jets clustered fr$
$flow (PFlow) objects which improves the jet resolution at low p_{\mathrm{T}} by capitalizing on the excellent resolution of the tracker to$
$up distribution, fluctuations due to the origin of the jet and its stochastic fluctuations, and residual differences between$

`JES_MC16Recommendation_Consolidated_PFlow_Apr2019_Rel21.config`.

To suppress the contribution of jets formed by pile-up processes, jets with p_{T} ¡ 60 GeV and

$|\eta|$ ¡ 2.4 are required to pass a Jet Vertex Tagger (JVT) cut [27]. The (default) tight working point is considered, as it is 96% efficient for hard scatter jets [**jvt-twiki**]. Jets produced by cosmic-rays, beam-induced background, and out-of-time pileup are reduced by imposing a set of quality criteria on variables characterizing the jet profile [46]. Jets considered for these cleaning cuts are clustered from calorimeter clusters (EMTopo jets) and have p_{T} ¿ 20 GeV, since these variables defining the jet cleaning cuts depend on the jet collection, and the recommendation was optimized for EMTopo jets [**jet-cleaning-twiki**]. If a single jet in the event fails the (default) `LooseBad` jet quality criterion, the entire event in vetoed. This recommendation is implemented via the `DFCommonJets_eventClean_LooseBad` variable in the EXOT8 derivation.

This analysis makes use of PFlow jets with $|\eta| < 4.5$ and p_{T} down to 30 GeV. Since the Jet/ETMiss group does provide calibrations down to 20 GeV, jets with a lower p_{T} cutoff were studied, but not found to improve the analysis's sensitivity due the exponential increase of multi-jet background.

9.2.2 *b*-tagging

b-jets are identified by the neural network-based DL1r algorithm with inputs characterizing the displaced tracks and vertices of the weakly decaying *b*-hadron [30]. This newly recommended DL1r algorithm improves on the previously recommended BDT-based algorithm, MV2c10, by including a Recurrent Neural Network to account for the correlation between the tracks in the jet [29].

For the *b*-tagging working-point optimization, both the MV2c10 and DL1r algorithms used dedicated trainings on the newly recommended PFlow jet collection [39]. The higher background rejection of DL1r allowed for a loosening of the b-tagging working point from 70% to 77% with a corresponding 10% improvement in the stat-only ggF SM limits. The decision to move to the 77% working point was made in harmony with the other HH channels for ease in the subsequent combination.

In the context of the future combination, since the majority of the HH analyses include *b*-jets in the final state, other channels are vetoing events with three DL1r *b*-tags at the 77% WP jet in the combination (to conservatively also veto our 4b events, and give us the possibility to explore a 3b analysis category).

9.2.3 *b*-jet corrections

The jet calibrations described in Section 9.2.1 focus on corrections for light-quark and gluon-initiated jets. As such, they systematically underestimate the energy of *b*-jet due to two main effects.

1. When the *b*-hadron decays semi-leptonically with a $W \to \mu\nu_{\mu}$ interaction in the cascade,
    - the neutrino energy is invisible in the jet reconstruction, and
    - the muonic energy is only partially accounted for in the jet's energy estimate since the muon (μ) is not stopped in the calorimeter.

2. The b-jet fragmentation is wider than that of the corresponding light-jets, meaning fewer final state hadrons from the b-quark fragmentation are included in jet clustering reconstruction ("out-of-cone" effect).

To correct for these effects, the HH analyses employ a harmonized μ-in-jet + p_{T}-reco correction to account for this underestimation of the b-jet p_{T}. This centralized b-jet correction is more sophisticated than the previous ggF correction of simply adding back in μ 4-vectors within $\Delta R < 0.4$ of the jet axis, although the previous VBF analysis did have a dedicated BDT-based b-jet energy regression [**HDBS-2018-18-witherratum**]. The μ-in-jet + p_{T}-reco algorithm is implemented centrally by the `BJetCalibrationTool` [**BJetCalibrationTool**], and a brief description is given below.

μ-in-jet

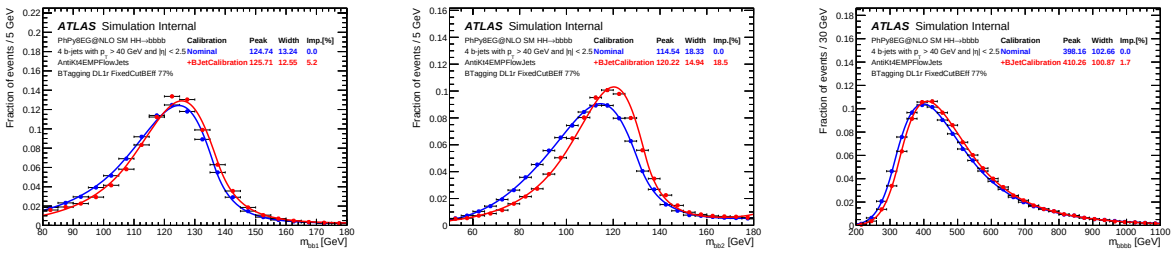
A search for a μ is performed in a variable radius cone $\Delta R(\mu, \mathrm{jet}) < \min\left(0.4, 0.04 + 10/p_T^{\mu}\ \mathrm{GeV}\right)$[1] from the jet axis to account for the increasingly collimated decay products of more energetic jets. If a μ is identified at the medium working point with p_{T} ¿ 4 GeV, $|\eta|$ ¡ 2.5 is within this ΔR cone of the jet axis, its 4-vector is added to that of the jet. If there are multiple μs passing the above criteria, only the μ closest to the jet-axis is added. Then the expected energy that the μ lost in the calorimeter is subtracted since this contribution was already included in the jet energy estimate.

p_{T}-reco

This second step accounts for the missing neutrino energy and out-of-cone effects that Jet/ETMiss calibrations don't capture. This correction factor is derived in $t\bar{t}$ events to correct the reconstructed p_{T} of the b-jets in logarithmic bins of the truth jet p_{T}. Since the correction is larger for b-jets decaying semi-leptonically, these correction factors are derived separately for b-jets with and without a μ.

Figure 9.7 illustrates the improvement achieved by the b-jet corrections in m_{H1}, m_{H2} and m_{HH} resolusion.

[1] The min function selects which of its arguments is smallest, and its use here avoids adding a μ farther away from the jet axis than the jet clustering distance parameter.



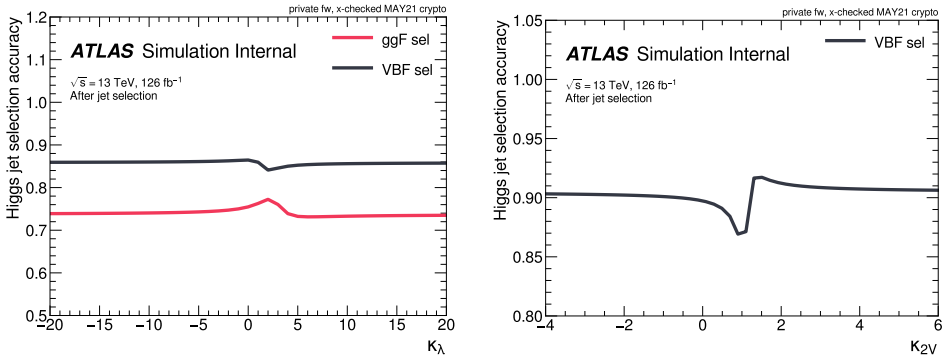
(a) m_{H1} (b) m_{H2} (c) m_{HH}

Figure 9.7: Comparisons of m_{H1}, m_{H2} and m_{HH} distributions before the b-jet corrections (blue) and after the b-jet corrections (red). These distributions are fitted using Bukin function, and the peak, the peak resolusion and the relative improvement are shown in the legend.

9.3 Event selection

9.3.1 Object Selection

Accuracy The Higgs jet selection accuracy is defined as the probability that all four b-jets in the event are matched to the truth b-quarks using a ΔR ¡ 0.3 matching criterion, and is shown is Figure 9.8. The jet selection accuracy is 74% for the ggF selection with %-level variations across the $\kappa_\lambda$values of interest. This means that there is a 74% chance the selected four b-jets are coming from the real Higgs decayed b-quarks. This accuracy loss is dominated by events where one of the b-quarks is out of acceptance. The 4b VBF selection has an average b-quark selection accuracy of 85% and 90% for the respective $\kappa_\lambda$and κ_{2V}signal samples. The dependency on $\kappa_\lambda$and/or $\kappa_{2V}$is likely due to the positive dependence of the b-tagging efficiency on the b-jet p_{T}: harder signals lead to higher b-tagging efficiency therefore higher Higgs jet selection accuracies. Figure 9.9 shows the truth m_{HH} distributions and the reconstructed histograms for the cases where we did or did not selected the correct jets for the ggF and VBF selections at a few signal points. As discussed above, we are less likely to select the correct jets for lower m_{HH}, and also the signal shapes for the $\kappa_\lambda$and κ_{2V}variations are quite different, this is entirely due to the underlying m_{HH} distribution.



(a) Jet selection accuracy vs κ_λ

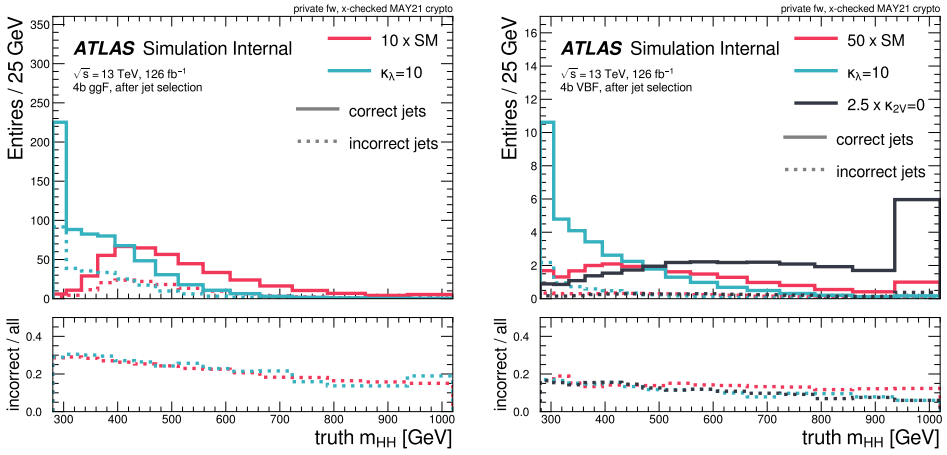
(b) Jet selection accuracy vs κ_{2V}

Figure 9.8: The jet selection accuracy as a function of $\kappa_\lambda$and κ_{2V}.

Definition of ggF and VBF Channels

James wrote this... I need to rephrase

This analysis possesses two channels targeting different HH production processes. One is optimized for gluon-gluon fusion (ggF) and the other for vector boson fusion (VBF). The channel an event is placed in depends on whether it contains the two high energy and well-spaced jets characteristic of the VBF topology. If so, the event is placed in the VBF channel. If not, it is placed in the ggF channel. In this section, for ease of reading, these two jets will be referred to as the *VBF*



(a) ggF signals (b) VBF signals

Figure 9.9: Truth m_{HH} distributions for correctly and incorrectly selected jets, for ggF (a) and VBF (b) signals.

Jets.

First, to belong to the VBF channel, an event must possess a minimum of six jets. The VBF Jets are reconstructed as the two jets with the highest di-jet invariant mass (m_{jj}) from the pool of both central and forward jets that failed the b-tag requirement. If no such pair exists, the event is placed in the ggF channel.

To reduce the number of background events three cuts are then applied. If an event passes all three cuts it is placed in the VBF channel; otherwise, it is placed in the ggF channel. The first two are a cut on the rapidity-gap between the VBF Jets of $|\Delta\eta_{jj}| > 3$ and on their combined invariant mass of $m_{jj} > 1000$ GeV. Finally, the six four-vectors corresponding to the HH and VBF Jets are summed, and a requirement applied that the p_{T} of that combined four-vector be less than 65 GeV. The HH is not reconstructed until after events are placed into VBF and ggF channels. Here, the same method for identifying the HH jets as in the reconstruction is used only to facilitate applying this cut.

If an event passes the VBF channel criteria, the jets used to reconstruct the VBF jets are removed from the pool of jets that the HH can be reconstructed from.

The ggF and VBF channels are designed to be orthogonal in order to facilitate statistically combining them when deriving results. As shown in Section 9.4.2 Table 9.12, a negligible fraction of ggF signal is leaked in the VBF channel even with the priority of the VBF selection. When optimizing the analysis, we saw the impact of the VBF veto for our ggF signals was at the 2% level, and as such, this deemed orthogonalization strategy to be acceptable.

Higgs candidate pairing

The HH system is reconstructed from two *Higgs candidates*, which are themselves reconstructed from two jets each (four *Higgs candidate jets* in total). These jets are selected from the pool of central jets. b-jets are selected first. If the event is a 4b event, the leading four in p_{T} are selected. If it is a 2b event, the remaining places are filled by non-b-tagged jets, which are sorted in p_{T} and the two leading jets taken. For details on the b-tag based selection, see Section 9.

We define **pairing** as the identification of a jet pair as a Higgs candidate. Given the four selected *Higgs candidate jets*, three possible pairings are possible, as sketched in Figure 9.10. We must therefore devise a strategy that accurately predicts which pairing is correct. The **correct pairing** is defined with generator level information. First, b-quarks are matched to b-jets using a $\Delta R < 0.3$ criterion. The correct pair is then defined by the b-quarks which have the same parent barcode ID in the truth record.

The pairing method chosen in this iteration of the analysis is based on the principle that the decay products of the Higgs should show a degree of collimation due to the Higgs's initial momentum. Of the two Higgs boson candidates in a given pairing, the *leading* Higgs candidate is defined as the one with the highest p_{T}. For each of the three pairing options, the leading Higgs candidate is identified and the $\Delta R_{Leading}(jj)$ between its two constituent jets calculated. The pairing option with the smallest $\Delta R_{Leading}(jj)$ is selected.

Pairing accuracy The pairing accuracy is defined as the fraction of correctly paired events among the events where the four Higgs-decayed jets are corrected selected by the jet selection. This definition is selected to decouple the pairing accuracy from the jet selection accuracy, as defined in Section 9. The pairing accuracy is shown in Figure 9.11 as a function of $\kappa_{\lambda}$and κ_{2V}, and in Figure 9.12 as a function of m_{HH}. Signals with harder p_{T} Higgs tend to have more collimated jet pairs, resulting in higher pairing accuracies. This effect leads to a loss of accuracy for low m_{HH} events (i.e., $m_{HH} < 450$ GeV), as also seen a drop of the pairing accuracy with non-SM κ_{λ}values or SM-like κ_{2V}values as they lead to softer kinematics. This is deemed acceptable because most of the analysis background is also located at low m_{HH}, therefore losing these events do not reflect in a loss of performance.

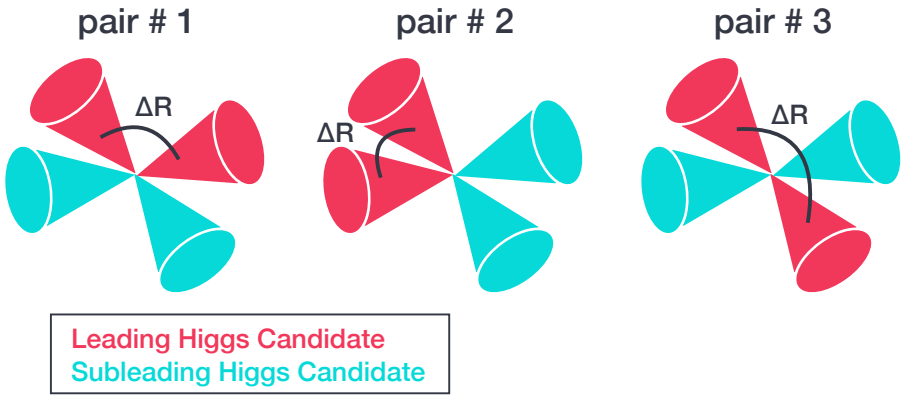
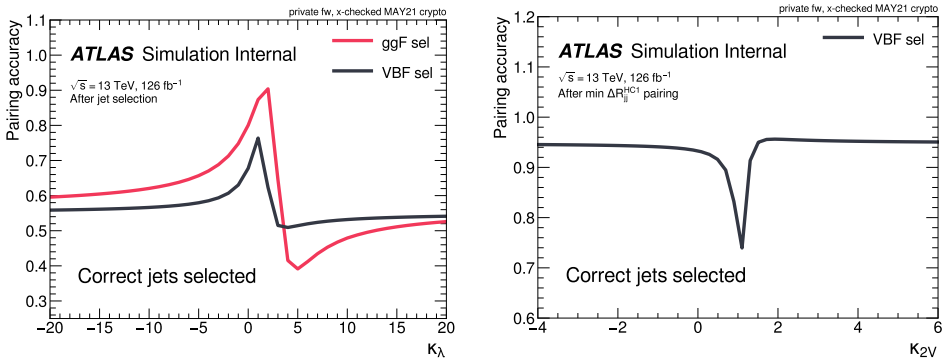


Figure 9.10: The three possible pairing permutations of the four HH jets into the two Higgs candidates. The opening angles between the jets in the leading Higgs Candidate are indicated, so pair number 2 is the selected pairing.



(a) Pairing accuracy vs κ_λ

(b) Pairing accuracy vs κ_{2V}

Figure 9.11: The pairing accuracy as a function of $\kappa_\lambda$and κ_{2V}, given that the correct jets have been selected.

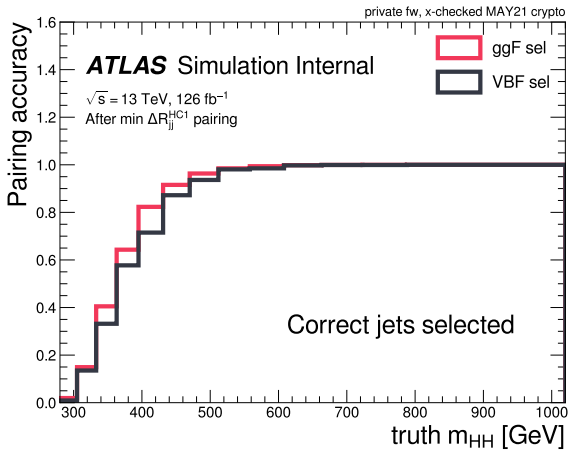


Figure 9.12: Pairing accuracy as a function of truth m_{HH}, given that the correct jets have been selected. The ggF selection accuracy is derived from the ggF SM sample, and the VBF selection accuracy is derived from the VBF κ_{2V}sample.

9.3.2 Background Reduction and $t\bar{t}$ Veto

Rui wrote this - I need to rephrase!

In order to suppress background, a pseudorapidity separation of $|\Delta\eta_{HH}| < 1.5$ is required between the two Higgs candidates in the ggF channel. This cut is not used in the VBF channel as it reduces sensitivity to SM VBF HH production. Figure 9.13 shows the $|\Delta\eta_{HH}|$ distributions for ggF HH signal and blinded data[2] in the ggF channel immediately after the pairing. It demonstrates that the data in the (m_{H1}, m_{H2}) plane, which is a good approximation of the background, tends to have higher values than those of the signal. Therefore, such a cut is applied to improve the signal purity. It is worth mentioning that a cut on the Higgs candidate p_{T} was applied in the previous publication, but was found to not be as powerful in the recent resonant search [**pT˙Cut˙and˙Muon-in-jet˙Correction**]. Therefore, it is dropped in this analysis.

Additionally, a top veto is applied to suppress the background from hadronic top-quark decays. This is applied by cutting on a discriminant, X_{Wt}, that is constructed to measure the compatibility of an event to contain a hadronically decaying top-quark. To construct X_{Wt}, W candidates are formed from any pair of jets with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $|\eta| < 2.5$, including those that were not selected for the Higgs candidates or for the VBF jets. All possible W candidates are considered, and top candidates are built by pairing W candidates with any remaining b-jets that were selected for Higgs candidates. The discriminant X_{Wt} is constructed for each combination, expressed as:

$$X_{Wt} = \sqrt{\left(\frac{m_W - 80.4\,\mathrm{GeV}}{0.1\ m_W}\right)^2 + \left(\frac{m_t - 172.5\,\mathrm{GeV}}{0.1\ m_t}\right)^2} \tag{9.2}$$

Events are then vetoed if the minimum X_{Wt} over all combinations is less than 1.5. Figure 9.14 shows the effectiveness of this cut at reducing the $t\bar{t}$ background while keeping a high efficiency for our signals.

We require the jet from Higgs candidates that is paired with the W candidate to be ***b*-tagged** while in the previous analyses there is no such a requirement.

9.3.3 Kinematic Region Definition

The final cut defining our signal region uses X_{HH} as given by Eq. 9.3. The functional form of this variable is similar to the equation for an ellipse, except that the radius is a function of the Higgs Candidate (HC) masses to allow harsher cuts for higher HC masses where the jets' resolution is better. This is easiest to see by looking at the SR shape in one of the mass planes, i.e, the purple line in Figure 9.15(a), since the "egg shaped" SR has allows for more acceptance at lower HC masses.

[2] By blinded data, we mean we do not show data events that fall in our 4b signal region as defined by Eq. 9.3.

$$X_{HH} = \sqrt{\left(\frac{m_{H1} - 124\,\text{GeV}}{0.1\ m_{H1}}\right)^2 + \left(\frac{m_{H2} - 117\,\text{GeV}}{0.1\ m_{H2}}\right)^2}. \tag{9.3}$$

The values (124, 117) in the X_{HH} definition were chosen to approximately match the centers of the m_{H1} and m_{H2} distributions for correctly paired signal events. The signal region (SR) is defined in Eq. 9.4, as visualized in the solid pink line in the (m_{H1}, m_{H2}) signal mass plane in Figure 9.15. For both the ggF SM signal and the VBF $\kappa_{2V} = 0$, these signal events are nicely peaking inside of this SR, and for the softer $\kappa_\lambda = 10$ spectrum are shown in Figure **??** of Appendix **??**. Figure 9.16 shows X_{HH} for correctly and incorrectly paired signal (SM for the ggF selection and $\kappa_{2V} = 0$ for the VBF selection), and demonstrates that this SR defining cut value of 1.6 has a high purity of correctly paired signal events. The SR center is re-optimized w.r.t. the 4b resonant analysis [**bbbbresolvedNote**], where (120, 110) is used, as the kinematics and mass coverage of both analyses are different. Alternative SR were also tested including a standard ellipse or larger size [**slides:SR-opt**]. No improvement in the background modeling was observed with the alternatives.

$$\text{SR} \quad : \quad X_{HH} < 1.6 \tag{9.4}$$

$$\text{CR Inner Edge} \quad : \quad X_{HH} = 1.6 \tag{9.5}$$

$$\text{CR Outer Edge} \quad : \quad \sqrt{\left(m_{H1} - 1.05 \cdot 124\,\text{GeV}\right)^2 + \left(m_{H2} - 1.05 \cdot 117\,\text{GeV}\right)^2} = 45\,\text{GeV} \tag{9.6}$$

Figure 9.17 shows the blinded 4b data mass planes for the ggF and VBF selections, and the 2b data mass planes for the ggF and VBF selections. Note, the backgrounds for the 4b distributions are built from reweighted 2b data.

The key task for setting limits is correctly predicting the distributions for key discriminating variables in the SR. For this fully-hadronic final state analysis, we have a fully data driven background estimation method derived using events in a kinematically similar control region. We define two control regions: Control Region 1 (CR1) and Control Region 2 (CR2). CR1 is used to derive the data-driven background estimate and CR2 is used to derive a systematic uncertainty associated with our methodology. These points will be expanded on in Section **??**. The region between the closed curves defined by Eq. 9.5 and Eq. 9.6 forms a band, within which CR1 and CR2 are defined. This region is orthogonal to the SR by design. This band is split into quadrants i.e. four sectors of roughly equal area. CR1 and CR2 are each defined as a pair of quadrants, where quadrants are paired such that they are on opposite sides of the band. The boundaries of CR1 and CR2 are shown in Figure 9.17. This is a different choice comparing to the 4b resonant analysis [**bbbbresolvedNote**], where rings of CR (and Validation Region) is used. The new choice reduces potential signal contamination in resonant VR and has a better extrapolation since CR is closer to SR. See Appendix **??** for some

studies.

The four quadrants that define CR1 and CR2 can be orientated in an infinite number of ways. The X_{Wt} cut applied in the selection acts like a W-mass veto for the constructed Higgs Candidates (HCs) causing a distinct drop in the number of events with m_{h1} or m_{h2} equal to ∼80 GeV. In Figure 9.17, this effect can be observed as the two straight light-colored bands centered around ∼80 GeVon the x and y-axes which stretch horizontally and vertically across the plot.[3] The orientation of the quadrants shown in Figure 9.17 was chosen such that these dips in the number of events equally impacted both CR1 and CR2.

Several different orientations were tested. In the studies conducted, different orientations were expressed as the angle between the x-axis and the closest CR1-CR2 boundary above the x-axis. Angles of 0°, 30°, 45° were compared and 45° was found to give better agreement in the 3b + 1 fail validation sample. Further investigation showed that this improvement stems from the X_{Wt} variable similarity, which is discussed in Appendix **??**.

[3]These mass planes before applying the X_{Wt} ¿ 1.5 cut are shown in Figure **??** and Figure **??** for the respective ggF and VBF selections

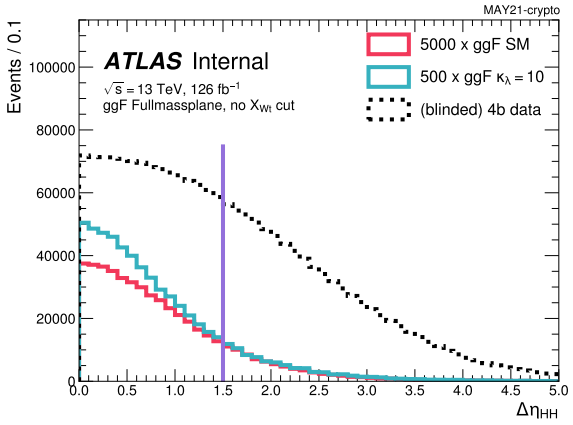
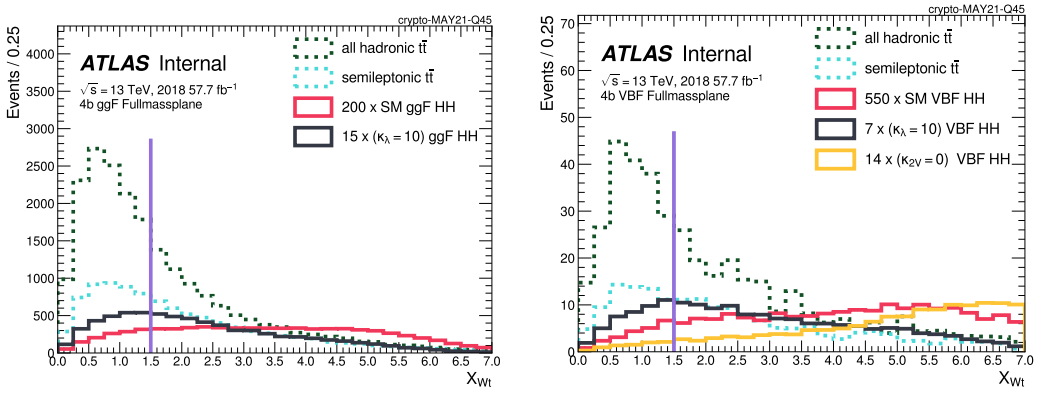


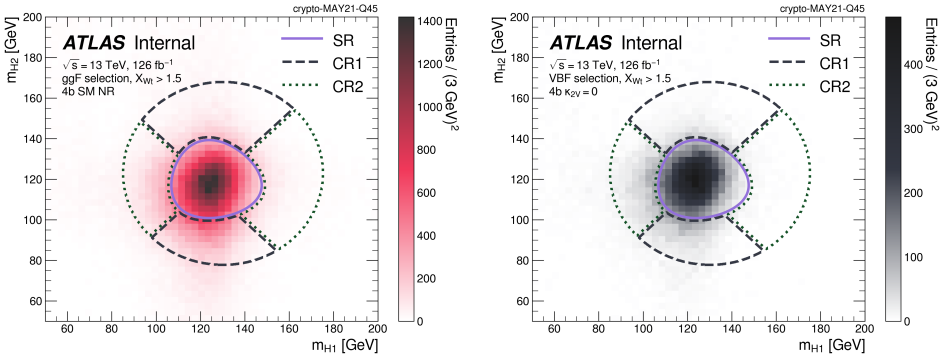
Figure 9.13: The $|\Delta\eta_{HH}|$ distribution for SM ggF HH Monte Carlo simulation and blinded data in the ggF channel. The solid purple line indicates the $|\Delta\eta_{HH}|$ ¡ 1.5 cut that is applied in the ggF selection. Events to the right of this line are discarded.



(a) 4b ggF selection

(b) 4b VBF selection

Figure 9.14: X_{Wt} distributions for our analysis categories in the 2018 dataset. The solid pink line indicates the X_{Wt} ¿ 1.5 cut applied to both the ggF and VBF channels. Event on the left of the line are discarded.



(a) 4b ggF SM signal (b) 4b $\kappa_{2V} = 0$ signal

Figure 9.15: Selected Higgs Candidate signal mass planes.

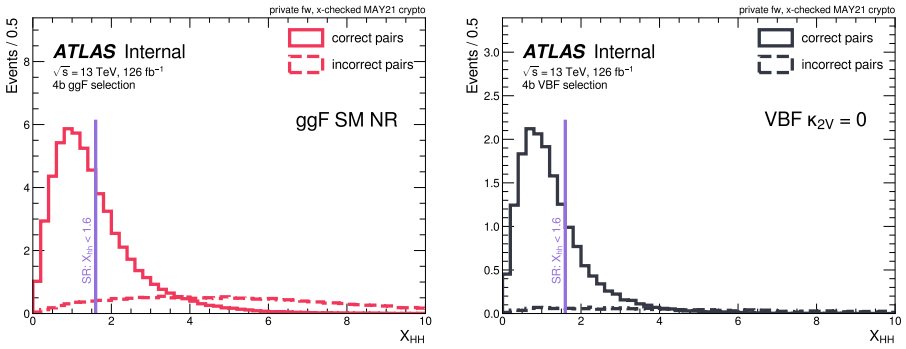
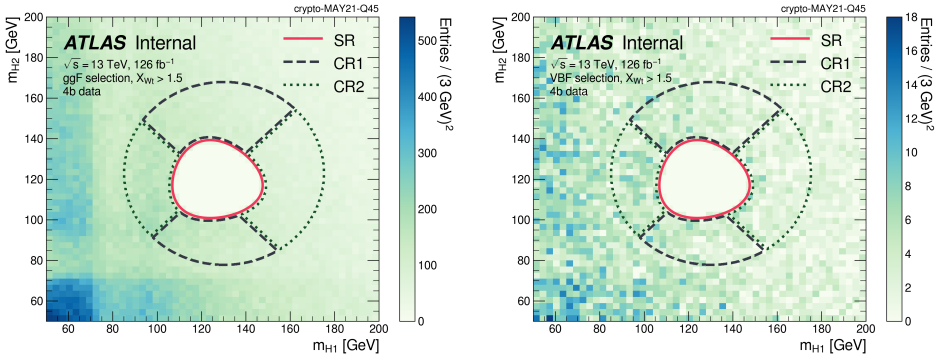


Figure 9.16: Visualization of the X_{HH} distribution for correctly and incorrectly paired events with the ggF (left) and VBF (right) analysis selections. The purple line indicates the SR defining cut.



(a) Blinded 4b data for ggF selection

(b) Blinded 4b data for VBF selection

(c) 2b data for ggF selection

(d) 2b data for VBF selection

Figure 9.17: The Higgs Candidate massplanes for the ggF and VBF analysis selections.

9.4 Analysis Categories

9.4.1 Definition of categories and binning

This analysis, as well as previous iterations, used the invariant mass of the HH system (m_{HH}) as the discriminating variable for the fit [**ATLAS-CONF-2021-035**, **HDBS-2018-18-witherratum**, 47].[4] A multi-variate algorithm (MVA) such as a BDT or NN would provide additional discrimination power; however, given our fully data driven background estimate, using a MVA may affect our ability to validate the modelling of the correlations between the input variables. As a compromise, a number of discriminating variables are used to define extra categories of $S/\sqrt{B}$ purity instead.

As alluded to in Section **??**, $|\Delta\eta_{HH}|$ is a powerful discriminating variable for both the ggF and VBF analyses, and is used as a categorization variable for both these channels (see Figure 9.18 and Figure 9.21). Specifics of the differences between the ggF and VBF categorization will be described in Section 9.4.1 and Section 9.4.1, respectively, along with a visualization of m_{HH} distributions in each category. A crucial step in implementing a categorization is ensuring the background within each category is well-modelled. In Section 10.2.2 are histograms of the background model in the CR1 training region. Good closure is observed.

Further tests of this categorization and validation of the background estimate are given in Section **??**.

Since the m_{HH} distribution is steeply falling, variable width histogram binning – with narrower bins at low m_{HH} and wider bins at high m_{HH} – is used in each of the categories. This allows the analysis to take advantage of the high m_{HH} events by having reasonable statistical uncertainties within these bins.

A logarithmic binning scheme was chosen. After defining the lowest bin edge, the second bin edge is set at (100 + X%) × the lowest bin edge, where X is the specified percentage parameter. Then, the third bin edge is set at (100 + X%) × the second bin edge. Bins of increasing width are added in this way until a upper threshold is surpassed. So, going from the lowest bin edge to the highest, the distance to a bin edge is a constant percentage increase on the previous bin edge. Note, the upper threshold is not the last bin edge. An algorithm is used to calculate the bin edges, and once it calculates a bin edge above this upper threshold, it adds it and stops.

Different logarithmic binning parameters are used for the ggF and VBF channels. These are shown in Table 9.3. These parameters were optimised to keep the relative error on the quadrature sum of the bootstrap and 2b Poisson components of the background model less than 30%, whilst not making the bins so wide as to lose important shape information. This 30% limit was chosen as it corresponds to the relative statistical error on 10 events, the rough threshold at which the asymptotic formulae used in limit setting are valid [48]. The plots demonstrating that the binning parameters

[4] The previous ggF searches actually used corrected m_{HH} to scale the 4-vectors of the HCs to match the Higgs mass of 125 GeV. This modified definition of m_{HH} helped constrain the widths of the signal peaks in resonant searches.

satisfied this are shown in Appendix **??**.

For the ggF channel, the same binning was used across categories (see Section 9.4.1), as the m_{HH} distributions differ little (as can be seen in Figure 9.19 or Figure 9.20). For VBF, the length of the tails of the m_{HH} distribution in the two $|\Delta\eta_{HH}|$ categories differs greatly (Figure 9.22). As such, different parameters were used. The ggF histograms use bin boundaries rounded to the nearest 1 GeV and the VBF histograms use bin boundaries rounded to the nearest 5 GeV. For ggF, the underflow is included in the lowest bin and overflow is included in the highest bin. Whilst, for VBF, the underflow and overflows are defined as additional bins taking all events below and above the nominal binning.

Table 9.3: Parameters used in the m_{HH} logarithmic binning algorithm. *Min* refers to the starting lowest bin edge and *Max* refers to the upper threshold after which the algorithm adds the last bin edge and stops.

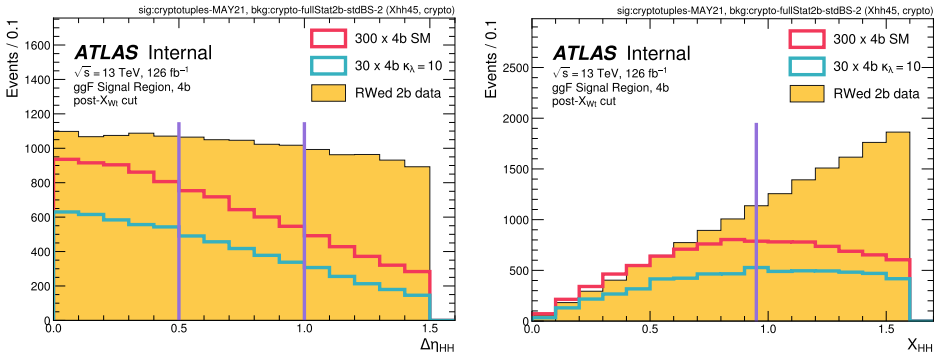
	Min [GeV]	Max [GeV]	Percentage [%]	Rounded to Nearest [GeV]
ggF All Categories	280	950	9	1
VBF low $\|\Delta\eta_{HH}\|$	280	890	10	5
VBF high $\|\Delta\eta_{HH}\|$	290	1470	9	5

Another method for choosing the binning in m_{HH} was tested. This was based on an algorithm that systematically merged bins until the statistical uncertainty was under 30%. This more flexible method resulted in more complicated binning scheme but very close significance and limits, giving us confidence that our simplified category choices are close to optimal.

ggF categories

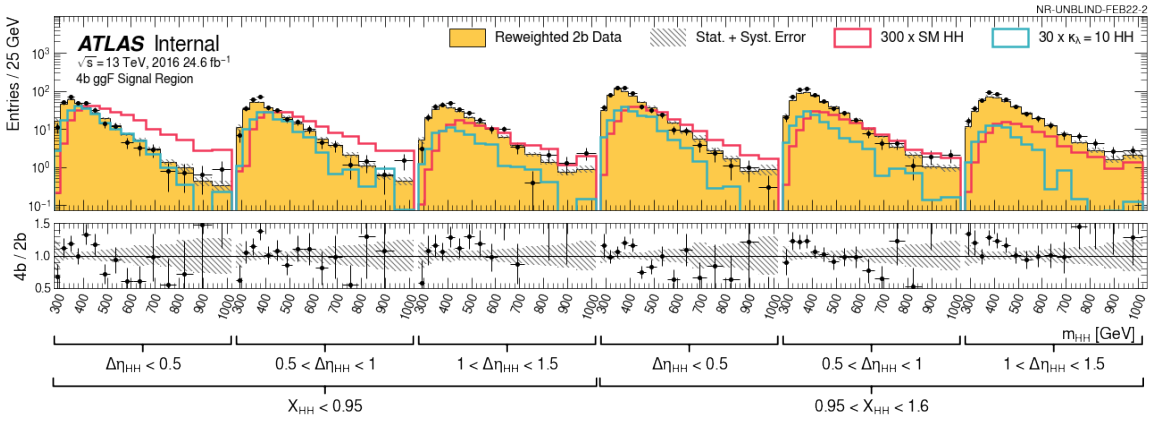
The ggF channels are categorized in two variables – $|\Delta\eta_{HH}|$ and X_{HH}. These variables are already cut on in the ggF channel – $|\Delta\eta_{HH}|$ ¡ 1.5 for the QCD background rejection and X_{HH} ¡ 1.6 for the SR definition. The distributions of these two variables of background prediction in the SR are shown in Figure 9.18, with overlaid the signal shapes. Since the background $|\Delta\eta_{HH}|$ distribution is flat, three equally spaced $|\Delta\eta_{HH}|$ bins were chosen between 0 and 1.5. Additionally, two X_{HH} bins were defined, with the boundary of 0.95 chosen to equally split the correctly paired signal events. This boundary choice also optimized $S/\sqrt{B}$ significance for the SM NR ggF signal.

As discussed in Section **??**, the ggF background estimate is derived separately for the years, so we also fit the years separately as visualized in the 4b ggF histograms in Figure 9.19, with the SM and $\kappa_\lambda = 10$ signals overlaid. The subpanels on these plots show the $S/\sqrt{B}$ significance to visualize which categories drive the sensitivity of the fit. The signal peaks for the lower $|\Delta\eta_{HH}|$, X_{HH} values, so these are the higher purity categories that drive our significance. The same plots with a linear y-axis are shown in Figure 9.20.

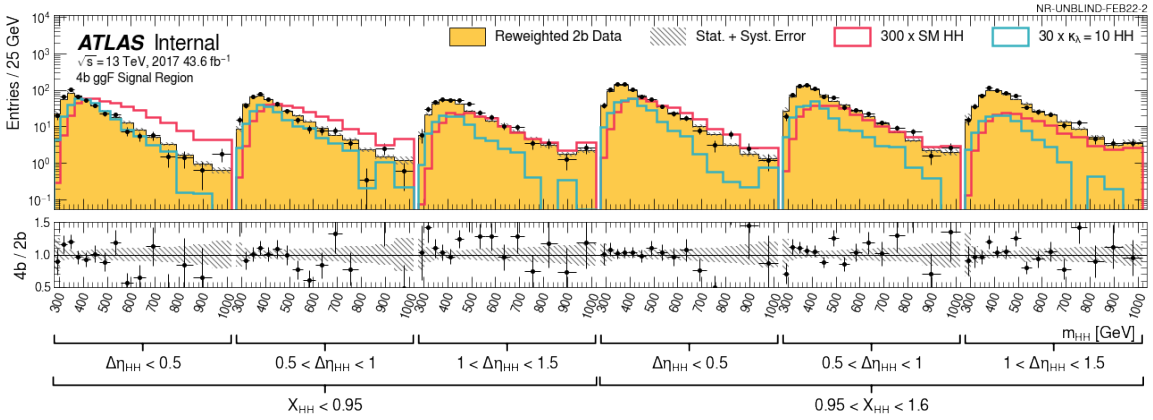


(a) All year merged: 4b ggF $|\Delta\eta_{HH}|$ (b) All year merge: 4b ggF X_{HH}

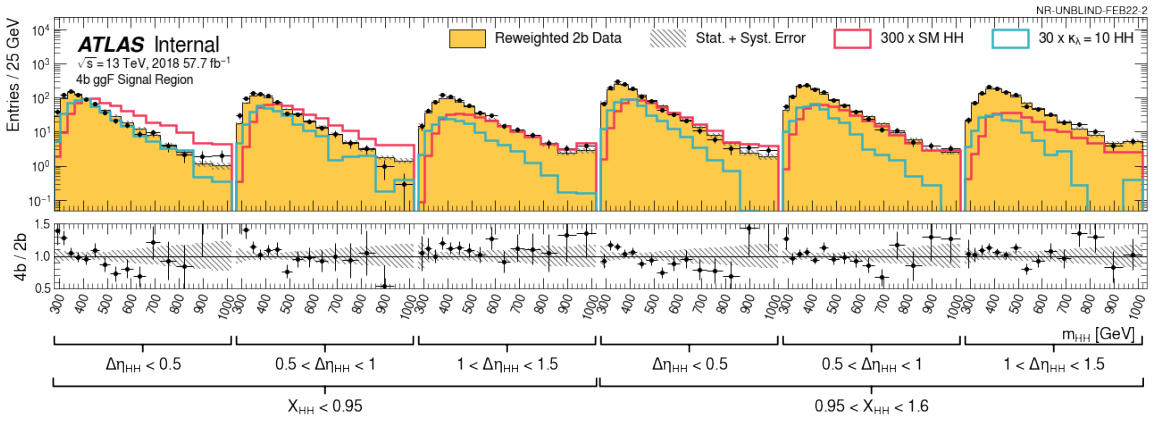
Figure 9.18: Distributions of the variables used for categorization in the ggF channel. Years are merged. To visualize the signals they are scaled by $\alpha = 100$ and 10 for the SM NR and $\kappa_\lambda = 10$ signals, respectively.



(a) 2016: 4b ggF

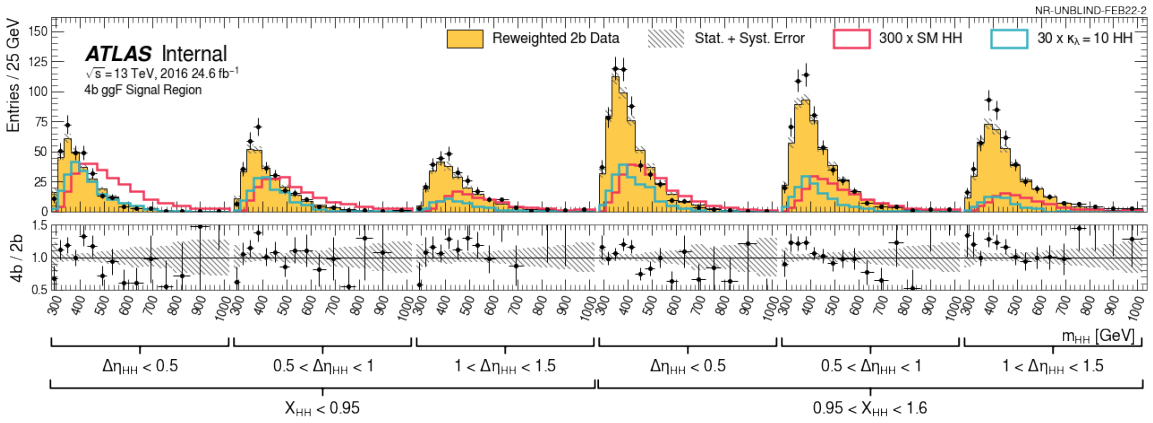


(b) 2017: 4b ggF

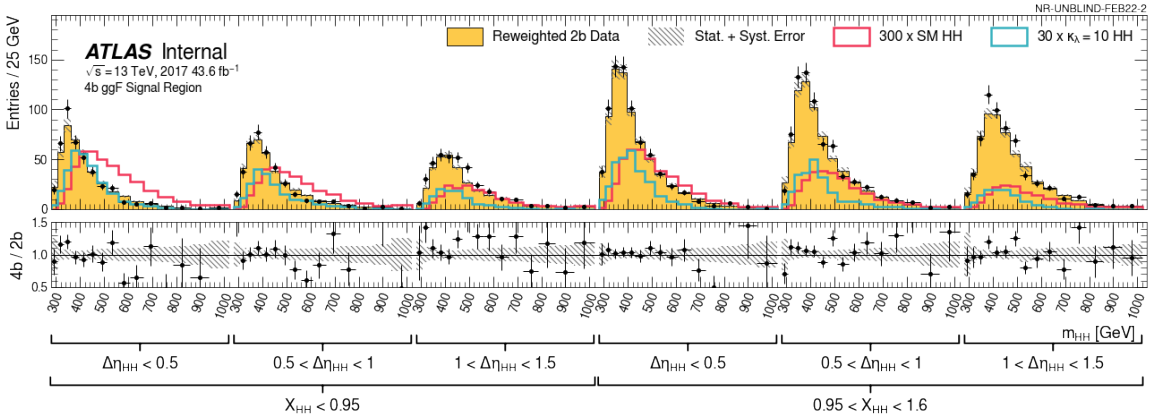


(c) 2018: 4b ggF

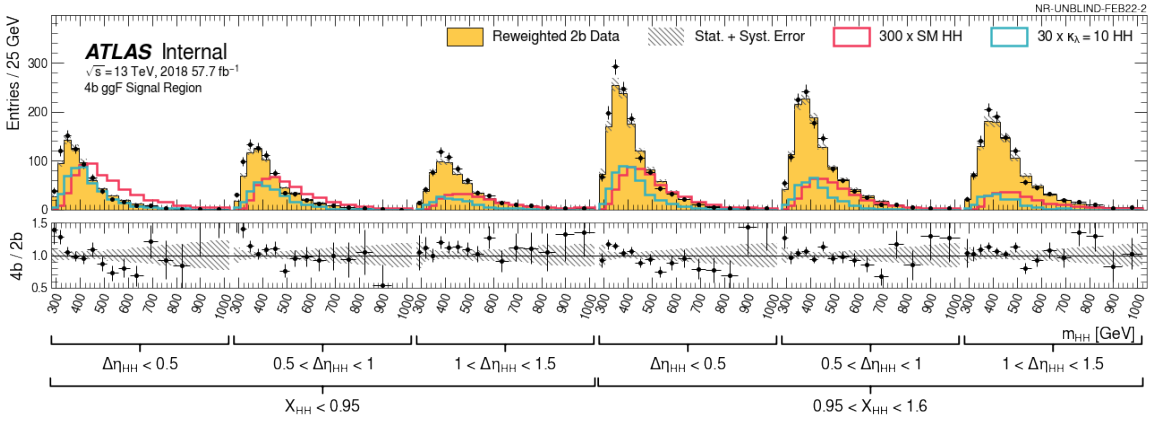
Figure 9.19: 4b ggF background and selected signal histograms for 2016, 2017, and 2018 with the proposed binning and categorization. To visualize the signals they are scaled by $\alpha = 100$ and 10 for the SM NR and $\kappa_\lambda = 10$ signals, respectively.



(a) 2016: 4b ggF



(b) 2017: 4b ggF



(c) 2018: 4b ggF

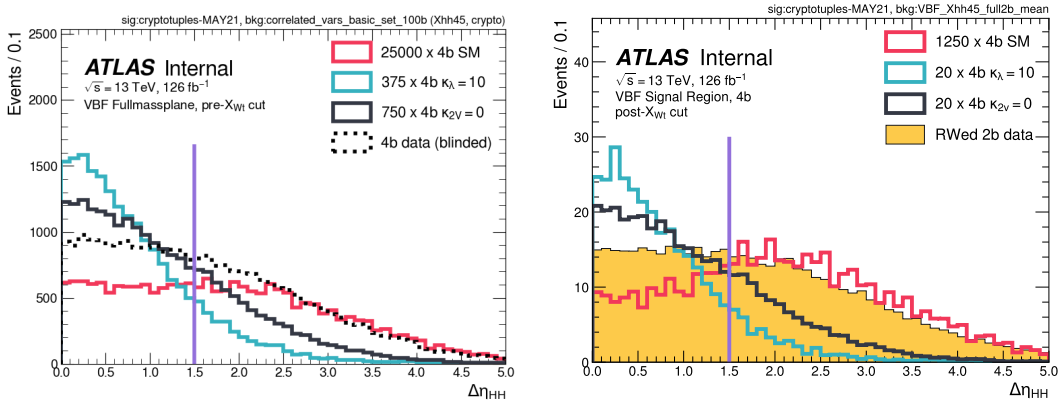
Figure 9.20: 4b ggF background and selected signal histograms for 2016, 2017, and 2018 with the proposed binning and categorization. To visualize the signals they are scaled by $\alpha = 100$ and 10 for the SM NR and $\kappa_\lambda = 10$ signals, respectively.

VBF categories

The VBF analysis has a single categorization based on $|\Delta\eta_{HH}|$. The boundary for the categorization is 1.5, chosen as it satisfied the balance between maximizing significance and maintaining the accuracy of the modeling of the background within the categories.

Figure 9.21 shows the $|\Delta\eta_{HH}|$ distributions before and after the X_{wt} cut for three key couplings – $\kappa_\lambda = 10$, $\kappa_{2V} = 0$ and the Standard Model prediction – alongside 4b data and the background estimate.

As shown in these plots, the $|\Delta\eta_{HH}|$ distribution corresponding to the non-SM couplings peaks close to $|\Delta\eta_{HH}| = 0$. On the other hand, the distribution corresponding to the SM prediction peaks at approximately $|\Delta\eta_{HH}| = 2$. As such, the $|\Delta\eta_{HH}|$ ¡ 1.5 category drives the sensitivity to the non-SM couplings; whereas, the $|\Delta\eta_{HH}| \geq 1.5$ category is more sensitive to the SM prediction.



(a) Full mass plane with SR data blinded.

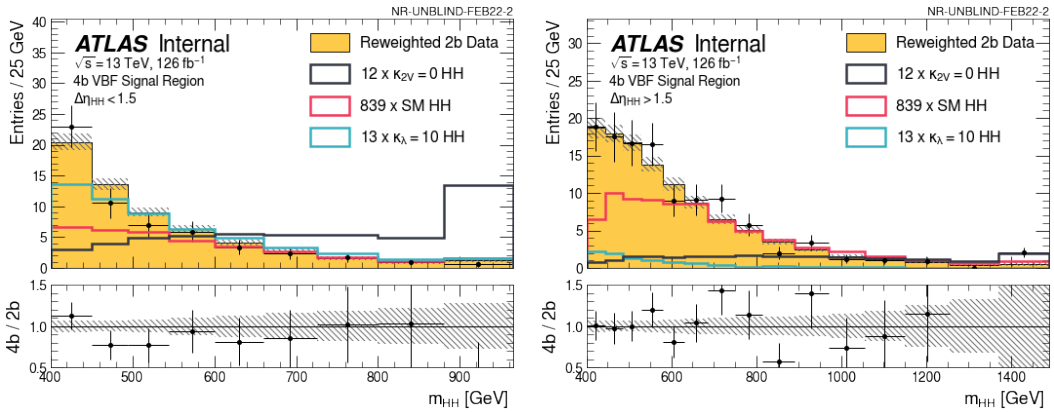
(b) SR with reweighted 2b background estimate.

Figure 9.21: Distributions of the difference in pseudorapidity of the two reconstructed Higgs bosons ($|\Delta\eta_{HH}|$) for signal Monte Carlo simulation, data and the background estimate in the VBF channel. The categorisation boundary is shown as a straight purple line at 1.5. The lefthand plot, Figure 9.21(a), shows the pre-X_{wt} cut distributions for three key couplings – $\kappa_\lambda = 10$, $\kappa_{2V} = 0$ and the Standard Model prediction – alongside the 4b data distribution excluding events in the Signal Region. The righthand plot, Figure 9.21(b), shows the post-X_{wt} cut distributions for the same couplings alongside the reweighted 2b distribution that is used to estimate the background contribution. All signal distributions have been scaled up as to be visible next to data and reweighted data.

Figure 9.22 shows the reconstructed HH mass distributions for the aforementioned three key couplings and the reweighted 2b data used to model the background contribution. Again, the signal distributions are scaled as to be visible next to the background estimate. Additionally, the significance of the scaled signal ($\alpha \times S/\sqrt{B}$) in each of the histogram bins is shown. The same signal scaling is used for each coupling across the two categories. This allows the sensitivity in the two categories to be compared.

Due to the low significance below 400 GeV and poor modelling (Figure **??**), we decided to drop the bins of m_{HH} ¡ 400 GeV in the fit in both categories.

The significance of the $\kappa_{2V} = 0$ signal in the final bin in both plots in Figure 9.22 is far larger than in those preceeding it. The events in the overflow are placed in the final bin for visual purposes, and this is where the increase in signal, and therefore significance, originates. Separating the overflow into finer bins has the potential to improved results by accounting for information on the differing distribution shapes. However, this is a region low in data statistics, particularly 4b events. The binning used was optimized to account for as much shape information as possible whilst ensuring there were enough statistics in each bin for the asymptotic approximation, which is used to derive results, to hold.



(a) $|\Delta\eta_{HH}| < 1.5$ (b) $|\Delta\eta_{HH}| \geq 1.5$

Figure 9.22: Distributions of the reconstructed m_{HH} for signal Monte Carlo simulation and the estimate of the background in each of the two $|\Delta\eta_{HH}|$ categories in the VBF channel. Distributions for three of the key couplings are shown – $\kappa_\lambda = 10$, $\kappa_{2V} = 0$ and the Standard Model prediction. Additionally, the significance of the scaled signal ($\alpha \times S/\sqrt{B}$) in each of the histogram bins is shown. Events in the underflow and overflow bins are counted in the yields of the initial and final bins respectively. The signals distributions are scaled as to be visible on the plot, and the scaling for each coupling is the same across the two categories.

In this section is a breakdown of the yields at the different steps in the analysis event selection (a *cutflow*) for data and important Monte Carlo simulation samples.

Data (2016-18, 126.1 fb^{-1}):

- Table 9.4(a): 4b events in the ggF channel.
- Table 9.4(b): 2b events in the ggF channel.
- Table 9.5(a): 4b events in the VBF channel.
- Table 9.5(b): 2b events in the VBF channel.

ggF HH MC simulation (normalized to 126.1 fb^{-1}):

- Tables 9.6(a): 4b events in the ggF channel for SM ggF HH signal.
- Tables 9.6(b): 4b events in the ggF channel for $\kappa_\lambda = 10$ ggF HH signal.
- Tables 9.7(a): 4b events in the VBF channel for SM ggF HH signal.
- Tables 9.7(b): 4b events in the VBF channel for $\kappa_\lambda = 10$ ggF HH signal.

VBF HH MC simulation (normalized to 126.1 fb^{-1}):

- Tables 9.9(a): 4b events in the VBF channel for SM VBF HH signal.
- Tables 9.9(b): 4b events in the VBF channel for $\kappa_\lambda = 10$ VBF HH signal.
- Tables 9.9(c): 4b events in the VBF channel for $\kappa_{2V} = 0$ VBF HH signal.
- Tables 9.8(a): 4b events in the ggF channel for SM VBF HH signal.
- Tables 9.8(b): 4b events in the ggF channel for $\kappa_\lambda = 10$ VBF HH signal.
- Tables 9.8(c): 4b events in the ggF channel for $\kappa_{2V} = 0$ VBF HH signal.

$t\bar{t}$ MC simulation (normalized to 126.1 fb^{-1}):

- Table 9.10(a): 4b events in the ggF channel in $t\bar{t}$ MC simulation for the non-all hadronic decay mode.
- Table 9.10(b): 4b events in the ggF channel in $t\bar{t}$ MC simulation for the all hadronic decay mode.
- Table 9.11(a): 4b events in the VBF channel in $t\bar{t}$ MC simulation for the non-all hadronic decay mode.
- Table 9.11(b): 4b events in the VBF channel in $t\bar{t}$ MC simulation for the all hadronic decay mode.

Table 9.4: 2016-18 data yields at each step in the analysis event selection for 2b and 4b events in the ggF channel, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [FEB22-unblind production (For data, expect no changes wrt MAR22)]

(a) 4b data (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.59e+10	-	-
Pass NTuple Preselection	5.697e+08	1	-
Trigger	2.807e+08	0.4927	0.4927
Trigger Buckets	2.49e+08	0.4371	0.8873
ggF channel	2.457e+08	0.4314	0.9868
≥ 4 central jets, ≥ 2 b-tags	1.806e+08	0.317	0.7349
≥ 4 b-tags	1.886e+06	0.003311	0.01045
$\lvert\Delta\eta_{hh}\rvert < 1.5$	1.032e+06	0.001811	0.5469
Top Veto	7.506e+05	0.001318	0.7276
Signal Region	1.617e+04	2.839e-05	0.02154
Control Region 2	3.067e+04	5.383e-05	0.04085
Control Region 1	3.204e+04	5.625e-05	0.04268

(b) 2b data (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.59e+10	-	-
Pass NTuple Preselection	5.697e+08	1	-
Trigger	2.807e+08	0.4927	0.4927
Trigger Buckets	2.49e+08	0.4371	0.8873
ggF channel	2.457e+08	0.4314	0.9868
≥ 4 central jets, ≥ 2 b-tags	1.806e+08	0.317	0.7349
2 b-tags	1.579e+08	0.2772	0.8744
$\lvert\Delta\eta_{hh}\rvert < 1.5$	8.27e+07	0.1452	0.5238
Top Veto	7.22e+07	0.1267	0.873
Signal Region	1.553e+06	0.002725	0.02151
Control Region 2	2.913e+06	0.005113	0.04035
Control Region 1	2.983e+06	0.005236	0.04132

Table 9.5: 2016-18 data yields at each step in the analysis event selection for 2b and 4b events in the VBF channel, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [FEB22-unblind production (For data, expect no changes wrt MAR22)]

(a) 4b data (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.59e+10	-	-
Pass NTuple Preselection	5.697e+08	1	-
Trigger	2.807e+08	0.4927	0.4927
Trigger Buckets	2.49e+08	0.4371	0.8873
VBF channel	3.295e+06	0.005784	0.01323
≥ 4 central jets, ≥ 2 b-tags	3.157e+06	0.005543	0.9583
≥ 4 b-tags	2.711e+04	4.759e-05	0.008586
$\lvert\Delta\eta_{hh}\rvert < 1.5$	2.711e+04	4.759e-05	1
Top Veto	2.175e+04	3.818e-05	0.8024
Signal Region	502	8.812e-07	0.02308
Control Region 2	906	1.59e-06	0.04165
Control Region 1	947	1.662e-06	0.04354

(b) 2b data (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.59e+10	-	-
Pass NTuple Preselection	5.697e+08	1	-
Trigger	2.807e+08	0.4927	0.4927
Trigger Buckets	2.49e+08	0.4371	0.8873
VBF channel	3.295e+06	0.005784	0.01323
≥ 4 central jets, ≥ 2 b-tags	3.157e+06	0.005543	0.9583
2 b-tags	2.758e+06	0.004842	0.8736
$\lvert\Delta\eta_{hh}\rvert < 1.5$	2.758e+06	0.004842	1
Top Veto	2.469e+06	0.004334	0.8951
Signal Region	5.873e+04	0.0001031	0.02379
Control Region 2	1.1e+05	0.0001931	0.04454
Control Region 1	1.108e+05	0.0001946	0.04489

Table 9.6: ggF HH MC simulation yields at each step in the analysis event selection for 4b events in the ggF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAR22 production]

(a) 4b SM ggF HH MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	526.6	1	-
Trigger	475	0.9019	0.9019
Trigger Buckets	419	0.7956	0.8821
Multiply FTAG, trig, + JVT SFs	381.8	0.725	0.9112
ggF channel	376.6	0.7151	0.9864
$\geq$ 4 central jets, $\geq$ 2 b-tags	322.4	0.6122	0.8561
$\geq$ 4 b-tags	86	0.1633	0.2668
$\lvert\Delta\eta_{hh}\rvert < 1.5$	71.85	0.1364	0.8355
Top Veto	60.4	0.1147	0.8406
Signal Region	29.1	0.05525	0.4817
Control Region 2	7.137	0.01355	0.1182
Control Region 1	11.41	0.02166	0.1889

(b) 4b κ_λ= 10 ggF HH MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	7338	1	-
Trigger	6917	0.9427	0.9427
Trigger Buckets	6378	0.8692	0.922
Multiply FTAG, trig, + JVT SFs	5279	0.7194	0.8278
ggF channel	5198	0.7084	0.9846
$\geq$ 4 central jets, $\geq$ 2 b-tags	4314	0.588	0.83
$\geq$ 4 b-tags	1002	0.1365	0.2322
$\lvert\Delta\eta_{hh}\rvert < 1.5$	850.6	0.1159	0.8492
Top Veto	569	0.07754	0.6689
Signal Region	182.7	0.0249	0.3211
Control Region 2	66.06	0.009003	0.1161
Control Region 1	86.18	0.01175	0.1515

Table 9.7: ggF HH MC simulation yields at each step in the analysis event selection for 4b events in the VBF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAR22 production]

(a) 4b SM ggF HH MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	526.6	1	-
Trigger	475	0.9019	0.9019
Trigger Buckets	419	0.7956	0.8821
Multiply FTAG, trig, + JVT SFs	381.8	0.725	0.9112
VBF channel	5.208	0.00989	0.01364
$\geq$ 4 central jets, $\geq$ 2 b-tags	5.155	0.009789	0.9898
$\geq$ 4 b-tags	1.14	0.002165	0.2212
$\lvert\Delta\eta_{hh}\rvert < 1.5$	1.14	0.002165	1
Top Veto	1.008	0.001914	0.8838
Signal Region	0.4833	0.0009177	0.4796
Control Region 2	0.123	0.0002336	0.122
Control Region 1	0.1703	0.0003234	0.169

(b) 4b κ_λ= 10 ggF HH MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	7338	1	-
Trigger	6917	0.9427	0.9427
Trigger Buckets	6378	0.8692	0.922
Multiply FTAG, trig, + JVT SFs	5279	0.7194	0.8278
VBF channel	81.14	0.01106	0.01537
$\geq$ 4 central jets, $\geq$ 2 b-tags	80.22	0.01093	0.9888
$\geq$ 4 b-tags	15.29	0.002084	0.1906
$\lvert\Delta\eta_{hh}\rvert < 1.5$	15.29	0.002084	1
Top Veto	11.15	0.001519	0.729
Signal Region	3.099	0.0004223	0.278
Control Region 2	1.449	0.0001975	0.13
Control Region 1	1.851	0.0002522	0.166

Table 9.8: VBF HH MC simulation yields at each step in the analysis event selection for 4b events in the ggF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAR22 production]

(a) 4b SM VBF HH MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	22.26	1	-
Trigger	20.67	0.9287	0.9287
Trigger Buckets	18.39	0.8261	0.8895
Multiply FTAG, trig, + JVT SFs	16.14	0.7251	0.8778
ggF channel	13.9	0.6246	0.8613
≥ 4 central jets, ≥ 2 b-tags	10.75	0.4829	0.7731
≥ 4 b-tags	1.87	0.08402	0.174
$\lvert\Delta\eta_{hh}\rvert < 1.5$	0.9419	0.04232	0.5036
Top Veto	0.736	0.03307	0.7814
Signal Region	0.2352	0.01057	0.3195
Control Region 2	0.07797	0.003503	0.1059
Control Region 1	0.1094	0.004914	0.1486

(b) 4b $\kappa_\lambda = 10$ VBF HH MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	1573	1	-
Trigger	1465	0.9317	0.9317
Trigger Buckets	1303	0.8289	0.8896
Multiply FTAG, trig, + JVT SFs	1092	0.6941	0.8374
ggF channel	955.5	0.6076	0.8754
≥ 4 central jets, ≥ 2 b-tags	756.5	0.481	0.7917
≥ 4 b-tags	134.5	0.08551	0.1778
$\lvert\Delta\eta_{hh}\rvert < 1.5$	109.4	0.06954	0.8132
Top Veto	76.41	0.04859	0.6988
Signal Region	25.19	0.01602	0.3297
Control Region 2	8.892	0.005654	0.1164
Control Region 1	11.42	0.007262	0.1495

(c) 4b $\kappa_{2V} = 0$ VBF HH MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.331e+06	-	-
Pass NTuple Preselection	626.1	1	-
Trigger	513.5	0.8201	0.8201
Trigger Buckets	444.6	0.71	0.8658
Multiply FTAG, trig, + JVT SFs	405.2	0.6471	0.9114
ggF channel	334.4	0.5341	0.8254
≥ 4 central jets, ≥ 2 b-tags	260.5	0.416	0.7788
≥ 4 b-tags	65.23	0.1042	0.2504
$\lvert\Delta\eta_{hh}\rvert < 1.5$	46.37	0.07405	0.7108
Top Veto	43.1	0.06884	0.9296
Signal Region	22.97	0.03669	0.533
Control Region 2	4.854	0.007752	0.1126
Control Region 1	8.26	0.01319	0.1916

Table 9.9: VBF HH MC simulation yields at each step in the analysis event selection for 4b events in the VBF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAR22 production]

(a) 4b SM VBF HH MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	22.26	1	-
Trigger	20.67	0.9287	0.9287
Trigger Buckets	18.39	0.8261	0.8895
Multiply FTAG, trig, + JVT SFs	16.14	0.7251	0.8778
VBF channel	2.238	0.1005	0.1387
$\geq$ 4 central jets, $\geq$ 2 b-tags	2.223	0.09986	0.9931
$\geq$ 4 b-tags	0.7449	0.03347	0.3351
$\lvert\Delta\eta_{hh}\rvert < 1.5$	0.7449	0.03347	1
Top Veto	0.6722	0.0302	0.9024
Signal Region	0.3265	0.01467	0.4857
Control Region 2	0.09092	0.004085	0.1352
Control Region 1	0.1149	0.005164	0.171

(b) 4b κ_λ= 10 VBF HH MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	1573	1	-
Trigger	1465	0.9317	0.9317
Trigger Buckets	1303	0.8289	0.8896
Multiply FTAG, trig, + JVT SFs	1092	0.6941	0.8374
VBF channel	136.1	0.08651	0.1246
$\geq$ 4 central jets, $\geq$ 2 b-tags	135.1	0.08589	0.9929
$\geq$ 4 b-tags	46.14	0.02934	0.3416
$\lvert\Delta\eta_{hh}\rvert < 1.5$	46.14	0.02934	1
Top Veto	34.84	0.02216	0.7551
Signal Region	14.24	0.009055	0.4087
Control Region 2	4.23	0.00269	0.1214
Control Region 1	5.03	0.003198	0.1443

(c) 4b κ_{2V}= 0 VBF HH MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Initial (Unweighted for MC)	1.331e+06	-	-
Pass NTuple Preselection	626.1	1	-
Trigger	513.5	0.8201	0.8201
Trigger Buckets	444.6	0.71	0.8658
Multiply FTAG, trig, + JVT SFs	405.2	0.6471	0.9114
VBF channel	70.74	0.113	0.1746
$\geq$ 4 central jets, $\geq$ 2 b-tags	70.3	0.1123	0.9939
$\geq$ 4 b-tags	27.63	0.04412	0.393
$\lvert\Delta\eta_{hh}\rvert < 1.5$	27.63	0.04412	1
Top Veto	26.52	0.04236	0.9601
Signal Region	17.29	0.02761	0.6517
Control Region 2	3.074	0.00491	0.1159
Control Region 1	3.805	0.006078	0.1435

Table 9.10: $t\bar{t}$ MC simulation yields at each step in the analysis event selection for 4b events in the ggF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAY21-crypto production, missing MC SFs eg trigger SF, expect to be about 10% smaller after applying SFs.]

(a) 4b Non-all hadronic $t\bar{t}$ MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	6.698e+06	1	-
Trigger	6.215e+06	0.9279	0.9279
Trigger Buckets	5.542e+06	0.8274	0.8917
ggF channel	5.488e+06	0.8193	0.9903
≥ 4 central jets, ≥ 2 b-tags	4.313e+06	0.6439	0.7859
≥ 4 b-tags	3.424e+04	0.005112	0.007939
$\lvert\Delta\eta_{hh}\rvert < 1.5$	1.989e+04	0.002969	0.5807
Top Veto	1.07e+04	0.001598	0.5383
Signal Region	403.5	6.024e-05	0.0377
Control Region 2	642.3	9.59e-05	0.06
Control Region 1	734.8	0.0001097	0.06864

(b) 4b All hadronic $t\bar{t}$ MC simulation (ggF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	6.698e+06	1	-
Trigger	6.215e+06	0.9279	0.9279
Trigger Buckets	5.542e+06	0.8274	0.8917
ggF channel	5.488e+06	0.8193	0.9903
≥ 4 central jets, ≥ 2 b-tags	4.313e+06	0.6439	0.7859
≥ 4 b-tags	3.424e+04	0.005112	0.007939
$\lvert\Delta\eta_{hh}\rvert < 1.5$	1.989e+04	0.002969	0.5807
Top Veto	1.07e+04	0.001598	0.5383
Signal Region	403.5	6.024e-05	0.0377
Control Region 2	642.3	9.59e-05	0.06
Control Region 1	734.8	0.0001097	0.06864

Table 9.11: $t\bar{t}$ MC simulation yields at each step in the analysis event selection for 4b events in the VBF channel normalized to 126.1fb^{-1}, alongside the ratio of each yield to the initial yield and to the yield for the previous cut. [MAY21-crypto production, missing MC SFs eg trigger SF, expect to be about 10% smaller after applying SFs.]

(a) 4b Non-all hadronic $t\bar{t}$ MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	6.698e+06	1	-
Trigger	6.215e+06	0.9279	0.9279
Trigger Buckets	5.542e+06	0.8274	0.8917
VBF channel	5.383e+04	0.008036	0.009713
≥ 4 central jets, ≥ 2 b-tags	5.25e+04	0.007839	0.9754
≥ 4 b-tags	351.8	5.252e-05	0.0067
$\lvert\Delta\eta_{hh}\rvert < 1.5$	351.8	5.252e-05	1
Top Veto	219.2	3.272e-05	0.623
Signal Region	8.882	1.326e-06	0.04053
Control Region 2	13.06	1.95e-06	0.05959
Control Region 1	13.07	1.951e-06	0.05964

(b) 4b All hadronic $t\bar{t}$ MC simulation (VBF channel)

	Yield	Yield / Pre-selection	Yield / Prior cut
Pass NTuple Preselection	6.698e+06	1	-
Trigger	6.215e+06	0.9279	0.9279
Trigger Buckets	5.542e+06	0.8274	0.8917
VBF channel	5.383e+04	0.008036	0.009713
≥ 4 central jets, ≥ 2 b-tags	5.25e+04	0.007839	0.9754
≥ 4 b-tags	351.8	5.252e-05	0.0067
$\lvert\Delta\eta_{hh}\rvert < 1.5$	351.8	5.252e-05	1
Top Veto	219.2	3.272e-05	0.623
Signal Region	8.882	1.326e-06	0.04053
Control Region 2	13.06	1.95e-06	0.05959
Control Region 1	13.07	1.951e-06	0.05964

9.4.2 Signal Yields

Tables 9.12 and 9.13 show the 4b SR yields for ggF and VBF HH MC simulation in the ggF and VBF SR. Approximately 1–2% of the overall ggF yield for the various coupling points shown falls in the VBF SR, demonstrating that the additional of the VBF SR does very little to dilute the sensitivity to ggF. A far larger percentage of the overall VBF yield falls in the ggF SR, but the definition of the VBF SR was optimized to maximize significance, resulting in tight cuts that do reject a significant amount signal but reject far more background.

Table 9.12: Yields in the 4b Signal Region for coupling points of interest for ggF HH Monte Carlo simulation normalized to 126.1fb^{-1}.

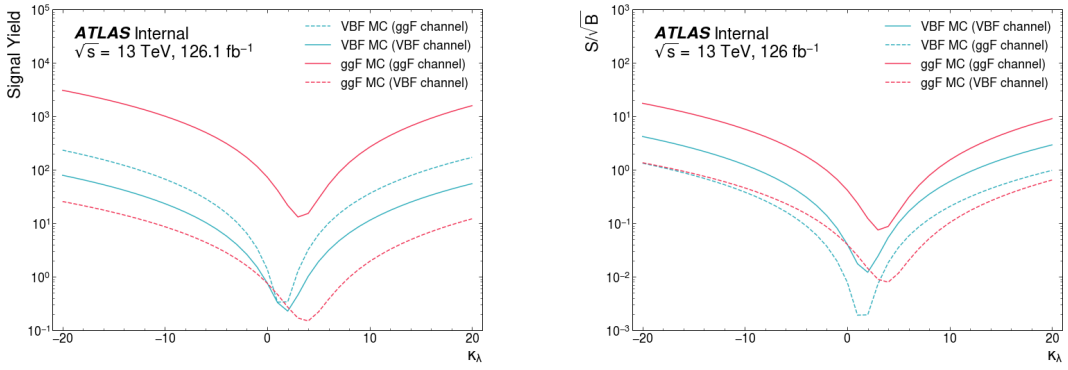
Coupling Point	ggF Channel Yield	VBF Channel Yield
SM	35.2	0.561
$\kappa_\lambda = 10$	237	3.9

Table 9.13: Yields in the 4b Signal Region for coupling points of interest for VBF HH Monte Carlo simulation normalized to 126.1fb^{-1}.

Coupling Point	ggF Channel Yield	VBF Channel Yield
SM	0.276	0.371
$\kappa_\lambda = 10$	30.4	16.4
$\kappa_{2V} = 0$	25.2	18.7

9.4.3 Non-resonant signal acceptance versus $\kappa_\lambda$and κ_{2V}

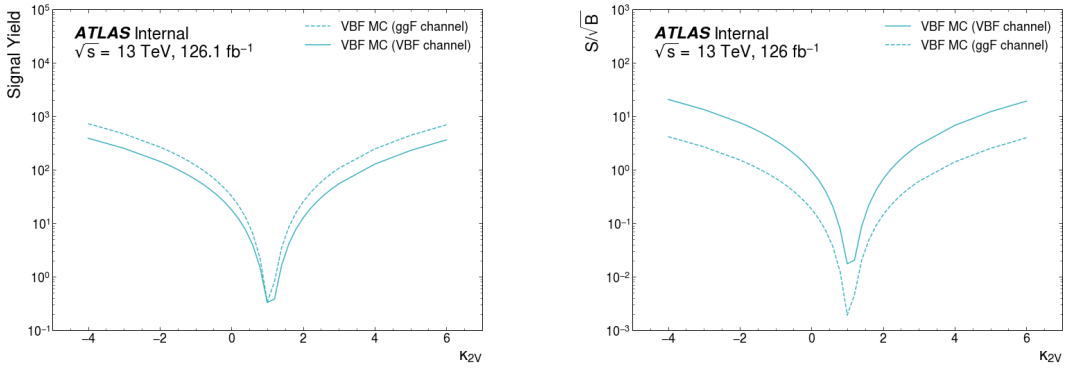
Figures 9.23–9.24 show the variation in the 4b SR signal yields and significance ($S/\sqrt{B}$) versus $\kappa_{2V}$and $\kappa_\lambda$respectively. Even though there is more VBF signal in the ggF SR than the VBF SR for the majority of modifier values, the significance is far larger in the VBF SR.



(a) Signal yield versus κ_λ

(b) Significance versus κ_λ

Figure 9.23: 4b Signal Region yield and statistical significance of the VBF and ggF Monte Carlo simulation in the VBF and ggF SRs versus κ_λ. In the legend, "channel" means SR.

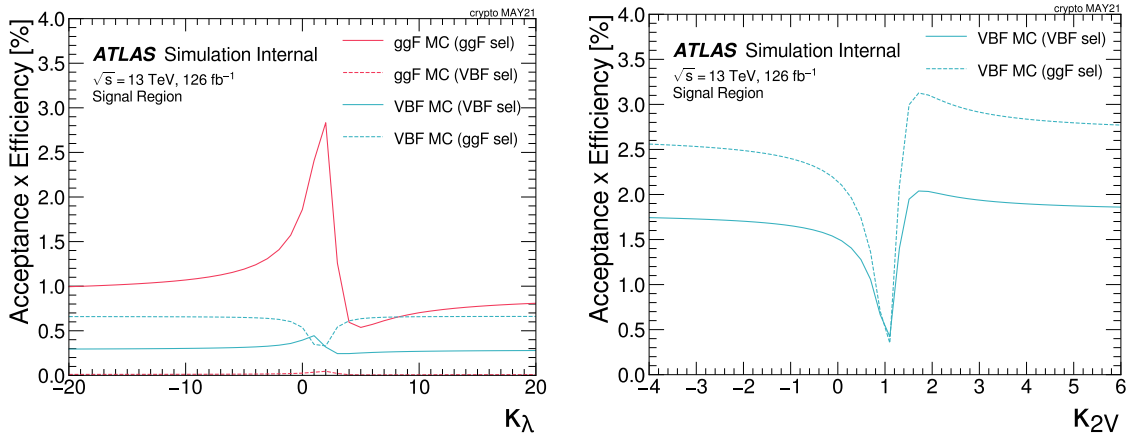


(a) Signal yield versus κ_{2V}

(b) Significance versus κ_{2V}

Figure 9.24: 4b Signal Region yield and statistical significance of the VBF Monte Carlo simulation in the VBF and ggF SRs versus κ_{2V}. In the legend, "channel" means SR.

Figure 9.25 shows the corresponding acceptance times efficiency in 1d as a function of the $\kappa_\lambda$and κ_{2V}variations. Then Figure 9.26 shows the corresponding acceptance times efficiency in 2d as a function of the $\kappa_\lambda$and κ_{2V}variations.

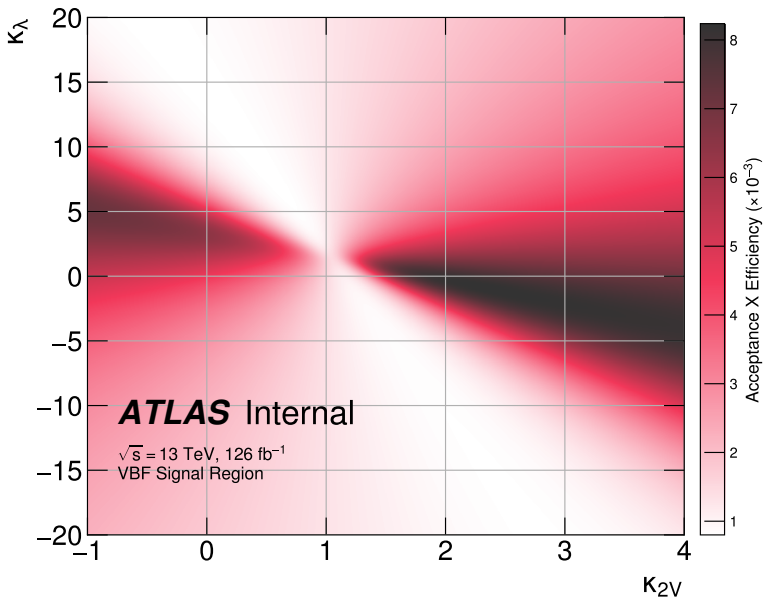


(a) Acceptance times efficiency versus κ_λ

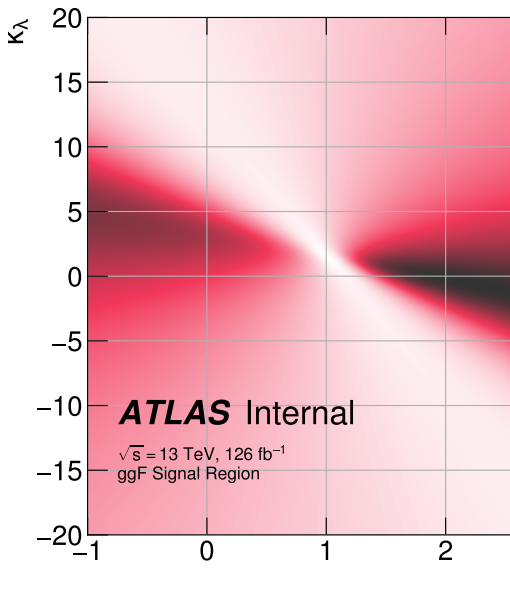
(b) Acceptance times efficiency κ_{2V}

Figure 9.25: 4b Signal Region acceptance times efficiency versus $\kappa_\lambda$and κ_{2V}.

Finally, to visualize our analysis selection as we propagate through the cutflow chain, the plots for acceptance times efficiency as we propagate through the cutflow acceptance are shown in Figure 9.27. Here we show the categories driving our analysis sensitivity the ggF signal in the ggF channel and the VBF signals in the VBF channel. In Appendix **??**, Figure **??** shows the ggF production $\kappa_\lambda cut flows in the VBF channel and the VBF production cut flows for the ggF channel.$

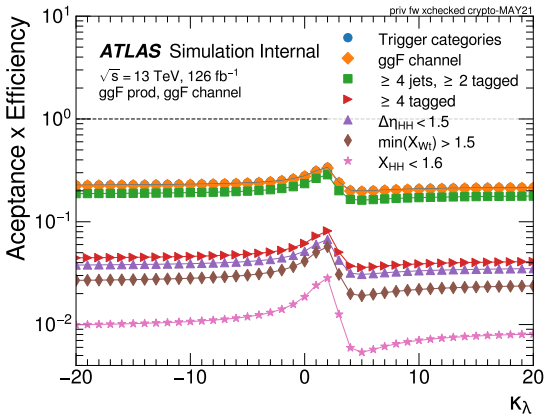


(a) VBF selection

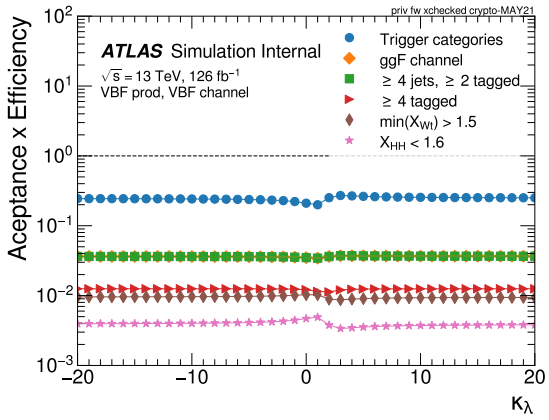


(b) ggF selection

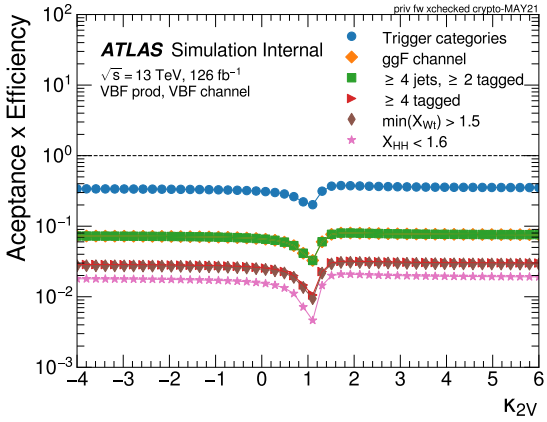
Figure 9.26: 4b Signal Region acceptance times efficiency for the VBF SM Monte Carlo simulation in the VBF or ggF selection in κ_{2V}-$\kappa_\lambda$plane.



(a) ggF MC: $\kappa_\lambda$variation



(b) VBF MC: $\kappa_\lambda$variation



(c) VBF MC: κ_{2V}variation

Figure 9.27: 4b Signal Region acceptance times efficiency.

10

Background estimation

This IS the analysis.

– Georges Aad (editorial board chair for collaboration internal review)

Although the 4b final state is tantalizingly attractive because of the highest branching ratio fraction for the signal, the fully hadronic final state also means that we have larger backgrounds to cope with. QCD backgrounds are notoriously difficult to simulate from first principles, so we need can't rely on simulation to estimate our background estimate, and instead need to rely on data driven methods to form an estimate of the background in a (blinded) signal region.

There are two main proportions of background that we have: QCD and $t\bar{t}$, with the estimated relative proportions shown in Table 10.1(a) and 10.1(b) for ggF and VBF, respectively. Since the $t\bar{t}$ proportion is only 10% of the background, our background estimate is derived inclusively for the QCD and $\bar{t}t$ proportions.

	16	17	18	**All**
pre-X_{Wt}	19.5%	17.6%	18.2%	**18.3%**
post-X_{Wt}	10.4%	9.5%	10.4%	**10.2%**

(a) ggF 4b

	16	17	18	**All**
pre-X_{Wt}	10.9%	11.1%	13.3%	**12.2%**
post-X_{Wt}	7.5%	6.6%	9.2%	**8.1%**

(b) VBF 4b

Table 10.1: Percentage of the data-driven background estimate expected to be composed of $t\bar{t}$ events for the ggF 4b (left) and VBF 4b categories (right).

The key idea of the background estimate is to reweight the distributions in a lower b-tag region into a higher b-tag region by deriving these reweighting maps in dedicated control regions. The task of accurately estimating this background and assigning an appropriate error bar is the key to this analysis – and is the focus of this chapter. Section 10.1 overviews the nominal background estimate strategy, while Section 10.2 shows some of the validation plots that the method is working. the errors

that we assign to the nominal estimate are shown in Section 10.3 and the tests of the methodology in different validation regions are discussed in Section 10.4

Since the different b-tagging regions define different control regions for defining the background estimate or assessing a systematic, in Table tab:b-tag-cat we provide a table of the different b-tag regions that will be discussed here.

Notation	Definition	Usage
2b	Exactly two central jets tagged with DL1r 77% WP	Background estimation
3b1f	Exactly three central jets tagged with DL1r 77% WP and no central jets passing the 85% WP	Background estimate systematic
4b	At least four central jets tagged with DL1r 77% WP	Signal region

Table 10.2: Different analysis definitions based on number of b-tags.

10.1 Reweighting overview

This technique derives a mapping to get an event-by-event weight from a lower tagged region to a higher tagged region (like a generalized ABCD method).

The lower b-tagged region consists of events with exactly two b-tags (which will subsequently be referred to as "2bb" events), and we reweight to match the distributions of events with four or more b-tags ("4bb" events). For "2bb" events, the leading two non-b tagged jets are taken for the other HC jets, and these four jets are still paired into HCs with the minΔR_{jj}^{HC1} pairing algorithm and pass through the rest of the analysis selection.

The reweighting maps are derived in CR 1 (as shown in Need to cite figure) since the events in the 4b SR are not observed until the analysis selection is finalized. The reweighting consists of deriving the maps $w(x)$

$$p_{4b}(x) = w(x) \cdot p_{2b}(x), \tag{10.1}$$

where x is a set of features characterizing the kinematics of the event, specified in Table F.1.

Variable description	**ggF**	**VBF**
$\log(\Delta R_1)$: between the closest two HC jets	✓	
$\log(\Delta R_2)$ between the other two HC jets	✓	
$\log(p_T)$ of the 4th leading HC jet	✓	
$\log(p_T)$ of the 2nd leading HC jet	✓	
$\langle\lvert HC\eta\rvert\rangle$: average absolute value of the HC jets η	✓	✓
$\log(p_{T,HH})$	✓	
ΔR_{HH}	✓	
$\Delta\phi$ between the jets in the leading HC	✓	
$\Delta\phi$ between the jets in the subleading HC	✓	
$\log(X_{Wt})$	✓	✓
Number of jets in the event	✓	
Trigger bucket index	✓	✓
Year index		✓
Second smallest ΔR between the jets in the leading HC (out of the three possible pairings)		✓
Maximum di-jet mass out of the possible pairings of HC jets		✓
Minimum di-jet mass out of the possible pairings of HC jets		✓
Energy of the leading HC		✓
Energy of the subleading HC		✓

Table 10.3: Set of input variables used for the $2b$ to $4b$ reweighting for the ggF and VBF channels. The variables included in the background estimate are denoted with a checkmark.

To enforce positivity of the weights we learn NN $Q(x) = \log w(x)$, minimizing the loss function:

$$\mathcal{L}[Q] = \mathbb{E}_{x\sim p_{2b}}\left[\exp\left(\frac{1}{2}Q(x)\right)\right] + \mathbb{E}_{x\sim p_{4b}}\left[\exp\left(-\frac{1}{2}Q(x)\right)\right] \tag{10.2}$$

results in a model learning $Q^*(x) = \log w^*(x)$ (where the loss function is set up in this way to are positive numbers using the loss function); (See proof in B.)

These mappings are derived in a kinematically similar control region (shown as CR 1 in Figure **??**). Then an error on this method from this choice of training region is taken by using an alternative control region (CR 2) and using the difference of the CR1 and CR2 predictions as an error bar.

Applying these reweighting maps in the SR gives us a background estimate of the observed data in a blinded SR.

In practice, there is some amount of noisiness due to the initializations of the NNs. To this extent, the NN for each background estimate is retrained 100 times and the average of the weights is taken as the nominal weight. Additionally, for each of the NN trainings the CR1 training events have a weight sampled from a Poisson distribution with mean 1 to account for the finite statistics of the training dataset.

- Hyperparameters for the ggF training: 50-50-50
- Hyperparameters for the VBF training: Need to look up.

Since we use different triggers between each of the years, the ggF background estimate has a different background estimate derived for each of the three years. Since VBF has two orders of magnitude less statistics than ggF, it was not as sensitive to this effect, so instead for VBF the training was done inclusively for the years with the year passed as an extra variable.

For ggF, the background estimates were derived before the X_{Wt} cut, while for the VBF the trainings were derived after the X_{Wt} cut.[1]

10.2 Validation plots

What might be good to include here?

- Let's include a *few* key variables for both ggF and VBF (?)
- And also the unrolled plots!!

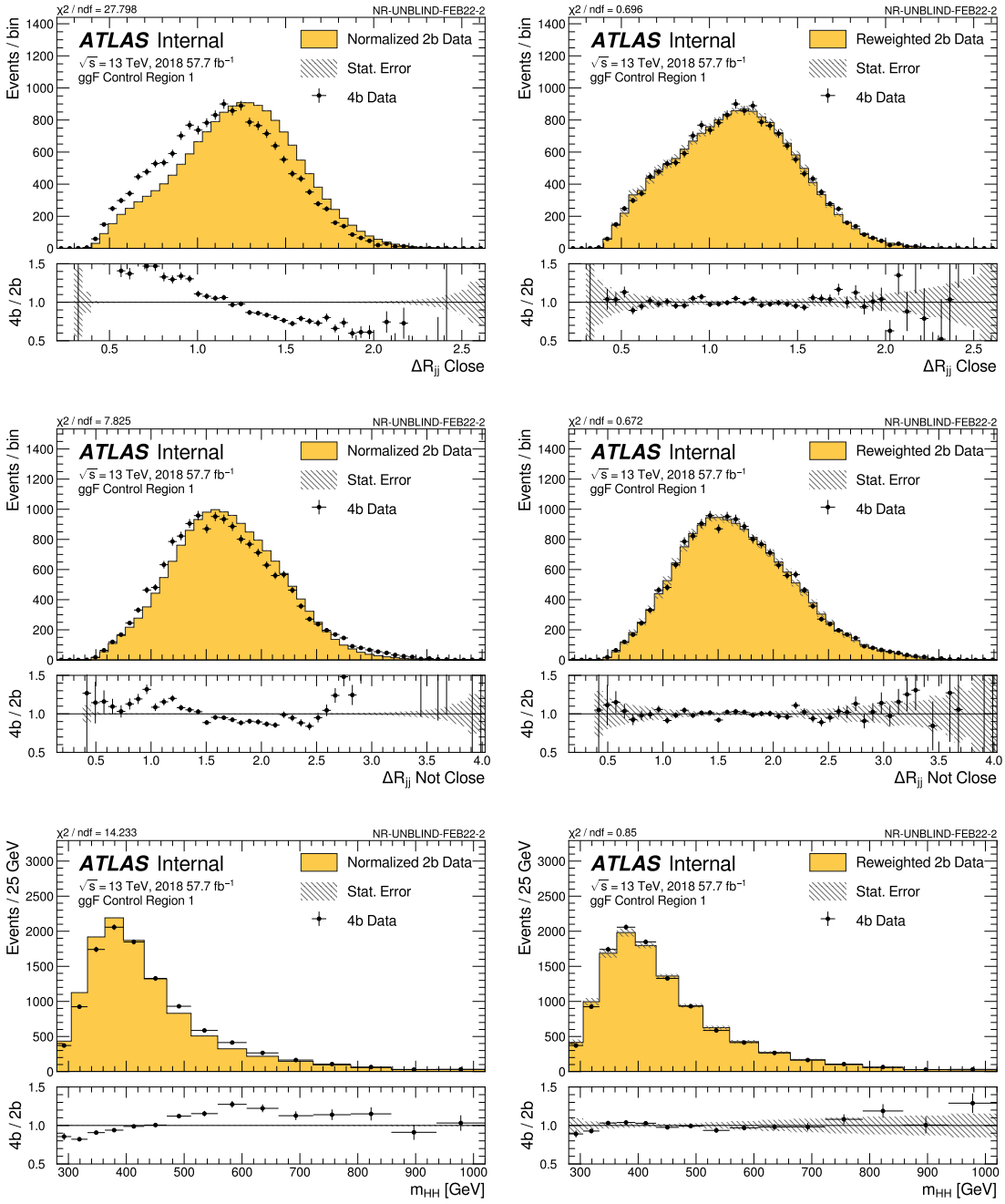
10.2.1 Marginal distributions

VBF plots ...

[1] We saw for ggF that we had the same performance whether we trained before or after the X_{Wt} cut.

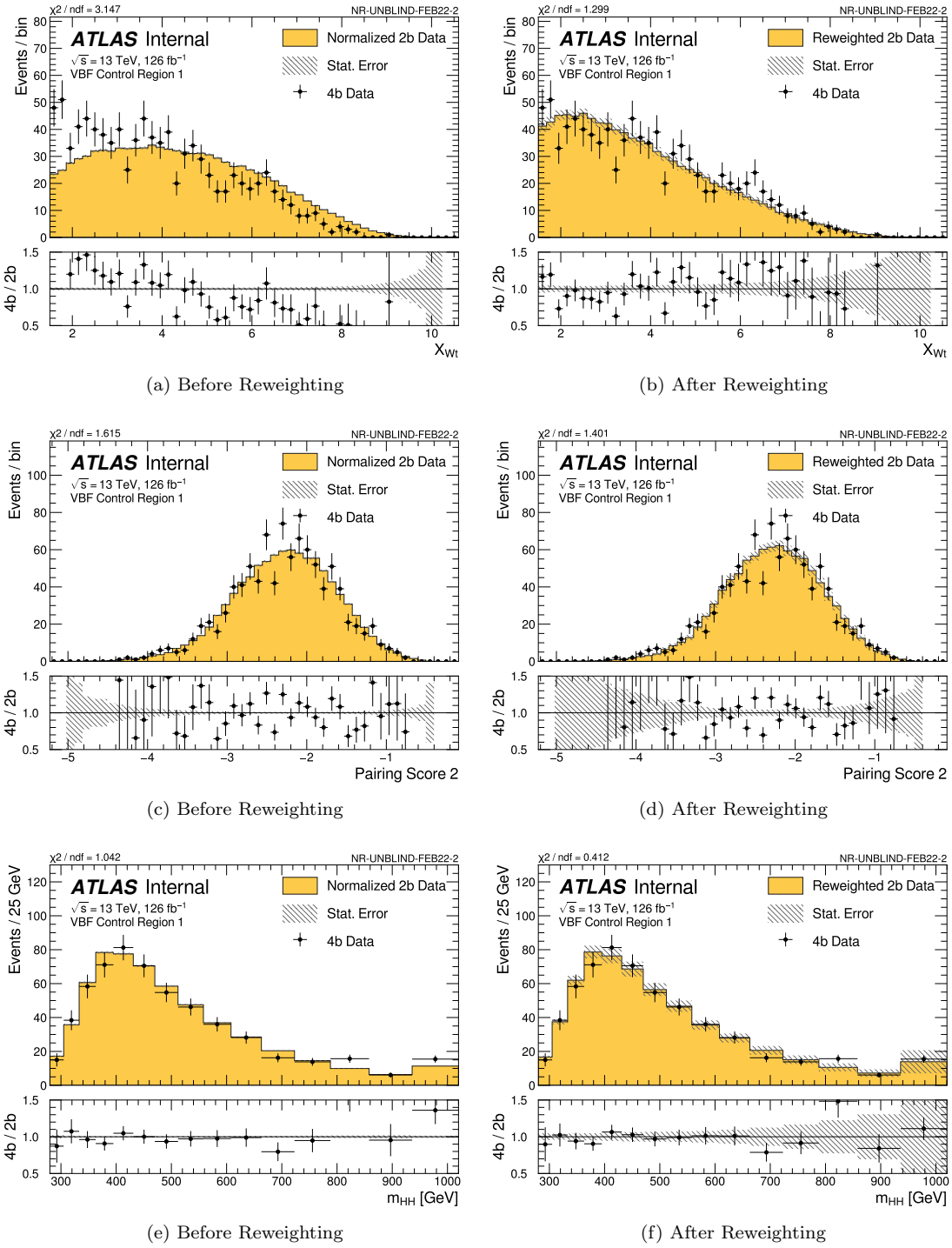
10.2.2 Validation plots in CR1 in categories

Background modeling in the CR1 and CR2 regions are checked for the ggF and VBF channels. Figure 10.3 shows the CR1 distributions for the three channels, taking 2018 for ggF as representatives. The ggF plots for the other years can be found in Appendix **??**, as well as the same set of plots in CR2.



(e) Before Reweighting (f) After Reweighting

Figure 10.1: Distributions of ΔR between the closest Higgs Candidate jets, ΔR between the other two (training variables) and the mass of the di-Higgs system (non-training variable) before and after CR 1 derived reweighting for the 2018 Control Region 1.



(a) Before Reweighting

(b) After Reweighting

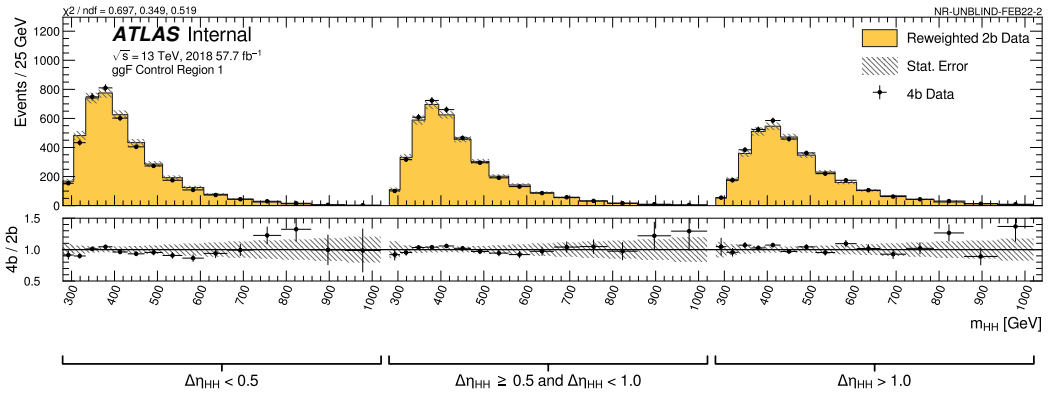
(c) Before Reweighting

(d) After Reweighting

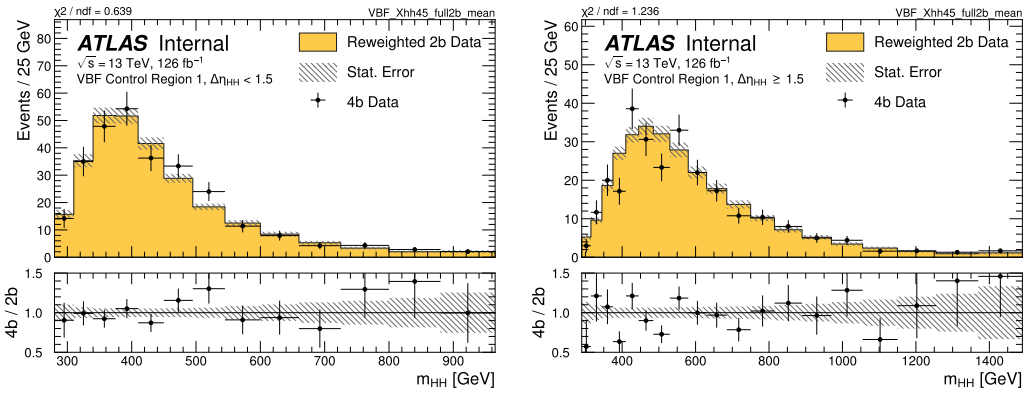
(e) Before Reweighting

(f) After Reweighting

Figure 10.2: Distributions of the top veto variable, X_{Wt}, the second smallest ΔR between the jets in the leading candidate (training variables) and the mass of the leading and subleading Higgs candidates and of the di-Higgs system (non-training variable) before and after CR 1 derived reweighting for the all years inclusive Control Region 1.



(a) 2018: 4b ggF, CR1



(b) Inclusive years: VBF, CR1

Figure 10.3: Distributions in CR1 after reweighting and categorization for ggF 2018 and VBF inclusive years. Bootstrap and Poisson errors are included.

10.3 Background systematics

For our fully data-driven background estimate, we assess several custom background modeling uncertainties:

- NN optimization Section 10.3.1
- choice of CRs for deriving these reweightings Section 10.3.2,
- and (ggF-only) remaining extrapolation uncertainty from testing a reweighting to another (lower) b-tagged target distribution Section 10.3.3.

We present the tables and plots comparing the relative size of each of these contributions (Section 10.3.3), and conclude with a description of how we characterize these uncertainties in the profile likelihood fit in Section subsec:bkgUnc-stats.

10.3.1 Deep ensembles

There are two components to the statistical error for the neural network background estimate. The first is standard Poisson error of the 2b data in the SR, i.e., a given bin, i, in the background histogram has value $n_i = \sum_{j \in i} w_j$, where w_j is the weight for an event j which falls in bin i. Standard techniques then result in statistical error $\delta n_i = \sqrt{\sum_{j \in i} w_j^2}$, which reduces to the familiar $\sqrt{N}$ Poisson error when all w_j are equal to 1.

However, this procedure does not take into account the statistical uncertainty on the w_j due to the finite training dataset. Due to the large size difference between the 2b and 4b datasets, it is the statistical uncertainty due to the 4b training data that dominates that on the background. A standard method for estimating this uncertainty is the bootstrap resampling technique [**Bootstrap**]. Conceptually, a set of statistically equivalent sets is constructed by sampling with replacement from the original training set. The reweighting network is then trained on each of these separately, resulting in a set of statistically equivalent background estimates. Each of these sets is below referred to as a replica.

In practice, as the original training set is large, the resampling procedure is able to be simplified through the relation $\lim_{n\to\infty} \text{Binomial}(n, 1/n) = \text{Poisson}(1)$, which dictates that sampling with replacement is approximately equivalent to applying a randomly distributed integer weight to each event, drawn from a Poisson distribution with a mean of 1.

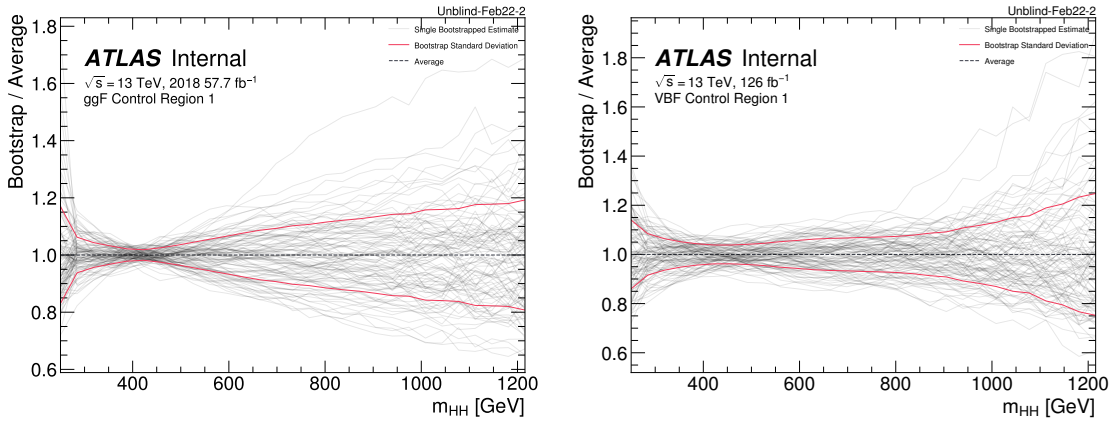
Though the network configuration itself is the same for each bootstrap training, the network initialization is allowed to vary, and this uncertainty It should therefore be noted that the bootstrap uncertainties implicitly capture the uncertainty due to this variation in addition to the previously mentioned training set variation.

The procedure to calculate the bootstrap uncertainty is as follows: first, each network trained on each bootstrap replica dataset is used to produce a histogram in the variable of interest. This

results in a set of replica histograms (e.g. for 100 bootstrap replicas, 100 histograms are created). The nominal estimate is mean of bin values across these replica histograms, with errors set by the corresponding standard deviation.

The variation from this bootstrapping procedure is used to assign a bin-by-bin uncertainty which is treated as a statistical uncertainty in the fit.

Figure 10.4 demonstrates how an uncertainty envelope is calculated from this procedure by visualizing the m_{HH} histogram in the Control Region 1 (where the background estimates are derived) for the 2018 ggF (left) and years inclusive VBF (right) background estimates.



(a) Comparison with histogram bin IQR (b)

Figure 10.4: Illustration of the bootstrap band procedure, shown as a ratio to the nominal estimate. Each grey line is from the m_{HH} prediction for a single bootstrap training, and the solid red line shows the standard deviation of histograms for the 2017 ggF background estimate (left) and VBF years inclusive background estimate (right).

10.3.2 Choice of control region

To account for the systematic bias associated with deriving the reweighting function in CR1, an alternative background model is derived in CR2. To aid in deriving an uncertainty to cover this bias, the Signal Region is split into quadrants - four sectors of approximately equal area - where the angle of alignment is set to match the quadrants that define the control regions, CR1 and CR2. The four SR quadrants are named for their relative positions in the Higgs Candidate mass plane, where Q is short for quadrant - Q_N (north), Q_S (south), Q_E (east) and Q_W (west), and shown in Figure 10.5.

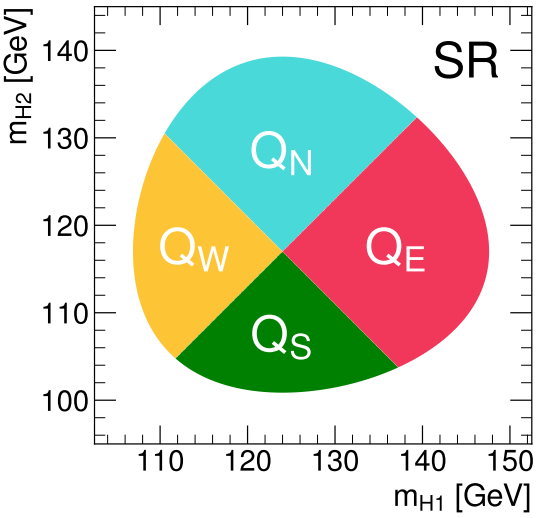


Figure 10.5: SR quadrants chosen to derive the four background variation nuisance parameters.

The nominal background estimate is derived by applying weights derived in CR1 to all four SR quadrants. Four alternative background estimates are derived by applying weights derived in CR1 to three of the SR quadrants and weights derived in the CR2 to the one remaining SR quadrant. For example, one alternative background estimate is derived by applying CR1-derived weights to Q_S, Q_E and Q_W, and CR2-derived weights to Q_N. The procedure is to reduce the constraints of a single nuisance parameter (NP) if we were to take the whole CR2 weights as a single uncertainty. It constraint was seen in previous 36 fb^{-1}paper and the resonant paper, where certain NP decomposition was also applied. As we believe the QCD background contains multiple sources, this choice of splitting NPs is to introduce more degrees of freedom to the fit. This can be done in various ways. The current way of splitting to four quadrants aligns well with the CR1 and CR2 region definitions, therefore is deemed to be a natural choice.

The difference between the baseline and alternative estimates is used to define a shape uncertainty on the m_{HH} spectrum, which is made two-sided by symmetrizing the difference around the baseline. Note that both normalization and shape differences are taken into considered.

These SR quadrants align with the four sectors of CR1 and CR2 to give these shape systematics the

flexibility to naturally follow the kinematic similarity of the adjacent CRs. The symmetrized difference between the nominal and alternative background estimates are shown for the ggF discriminant for each year in Figures 10.6–10.8. Complementary comparisons of the nominal and alternative background estimates for the VBF selection are included in Appendix **??**.

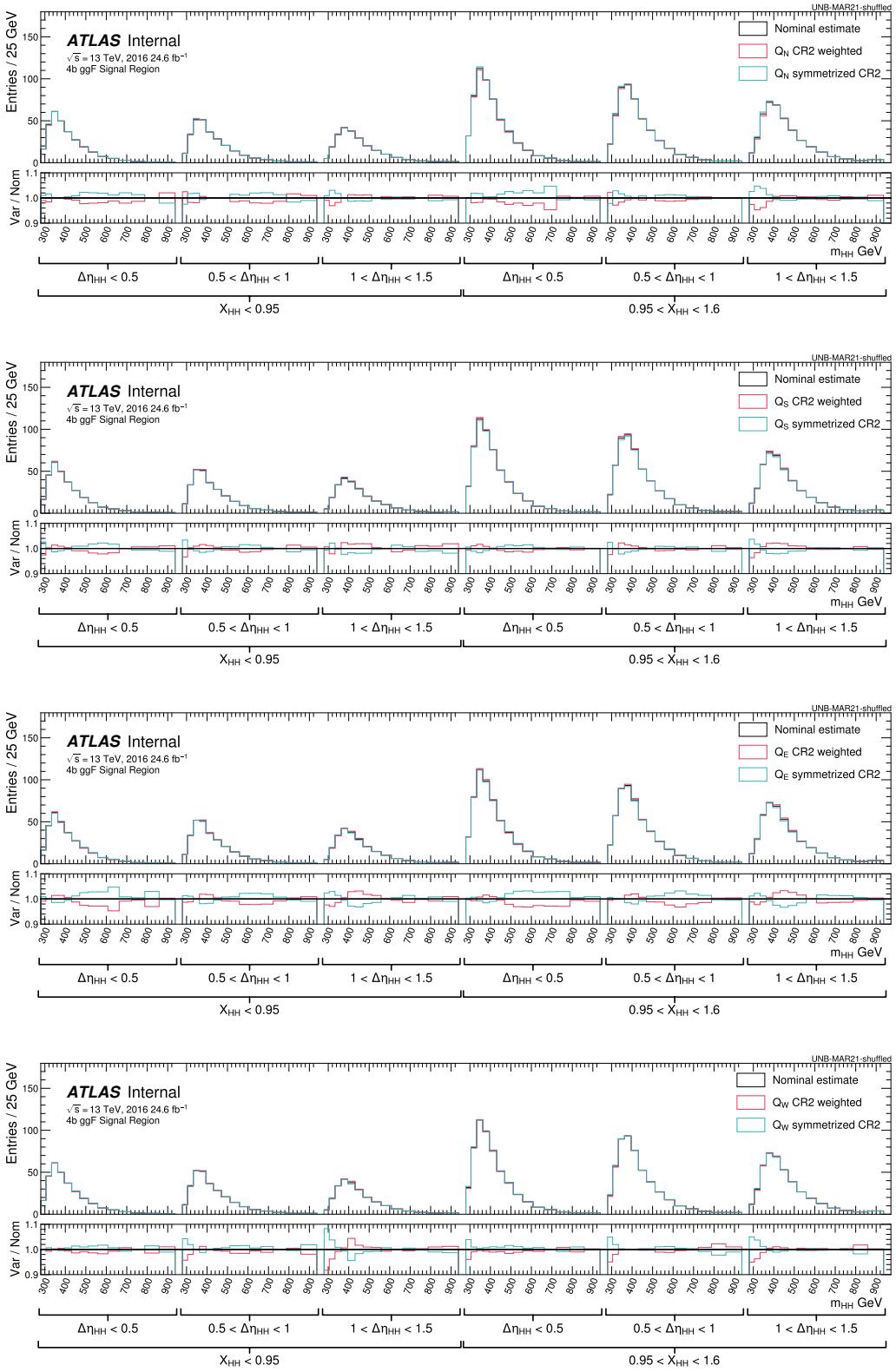


Figure 10.6: Example of variation in the SR NP quadrants for the 2016 ggF discriminant.

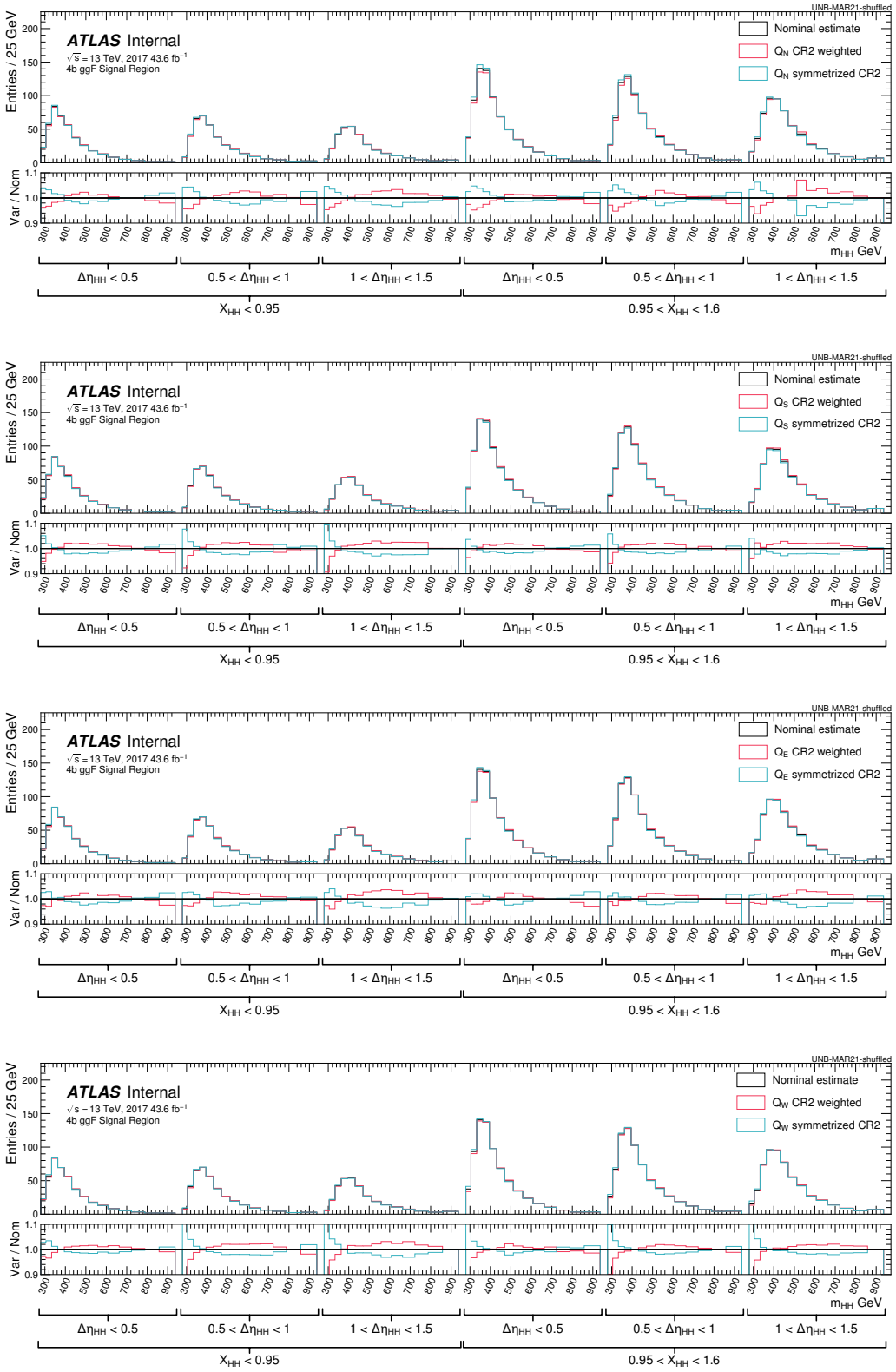


Figure 10.7: Example of variation in the SR NP quadrants for the 2017 ggF discriminant.

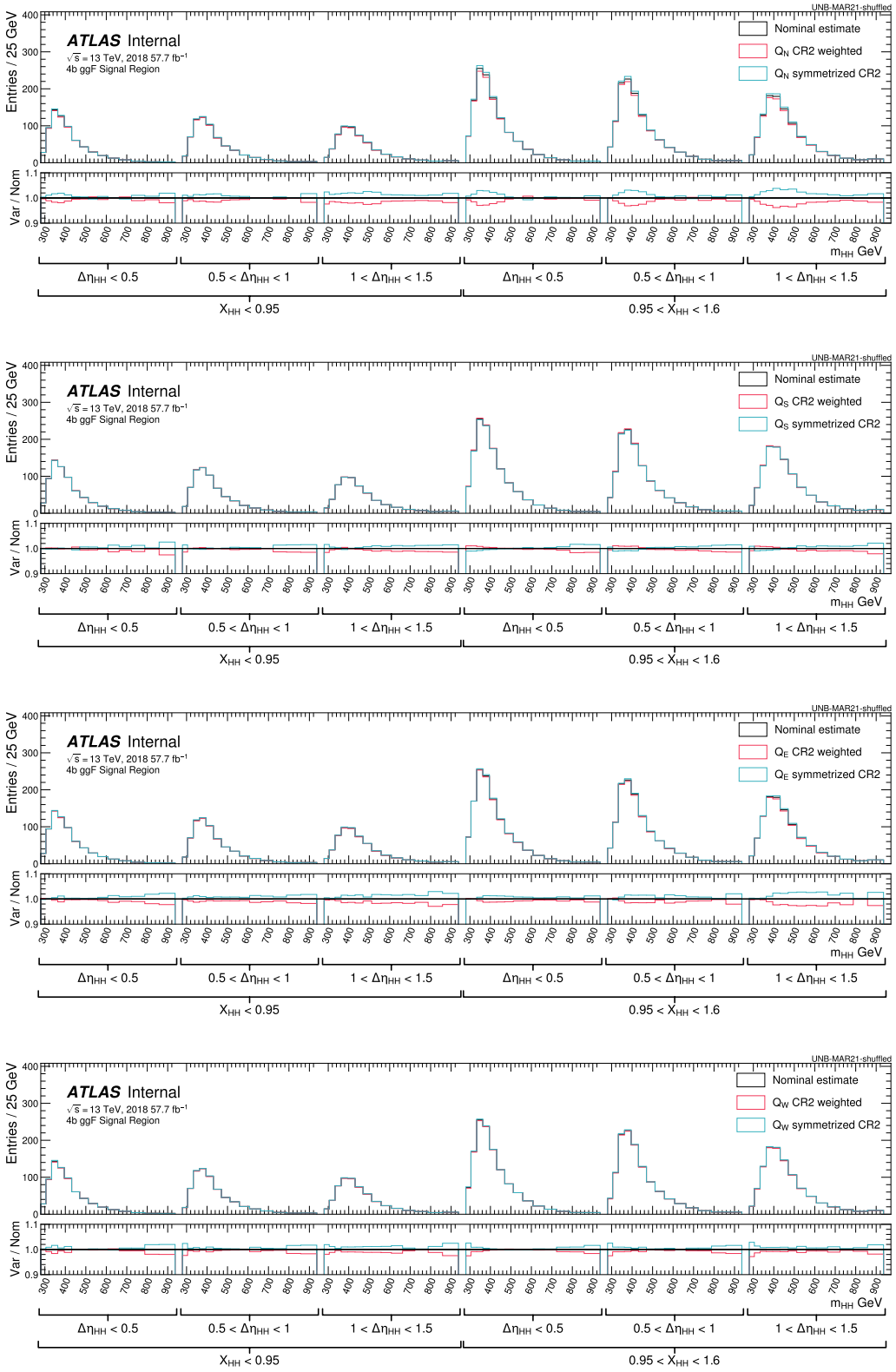


Figure 10.8: Example of variation in the SR NP quadrants for the 2018 ggF discriminant.

10.3.3 3b1f non-closure uncertainty

The 3b1f region is defined such that it contains events with three jets that were b-tagged at the 77% working point, but that all other jets failed to be b-tagged at even the loosest 85% working point. The fourth jet required to reconstruct the di-Higgs is taken as the highest p_{T} non-b-tagged jet. The standard methodology for deriving the background estimation (Section **??**) is performed on this region, with 4b events replaced by 3b1f. The background prediction uncertainties include 2b stat error, bootstrap (Section **??**), and CR1 versus CR2 shape uncertainties (Section **??**). Detailed studies can be found in Appendix **??**. A small deviation between the reweighted 2b and the 3b1f data distributions in the SR was found in the ggF channel but not in the VBF channel. An additional uncertainty due to the 3b1f non-closure is derived for the ggF channel to cover this deviation.

The uncertainty is derived as follows: In each of the analysis categories (Section 9.4.1), the ratio of the 3b1f and the reweighted 2b m_{HH} distributions is computed using the binning described in Table 9.3. The ratio is compared to unity bin by bin. If the difference is not covered by the quadratic sum of the background prediction uncertainties and the 3b1f data statistical uncertainty in that bin, the residual is taken to form a template. To reduce the statistical fluctuations between bins, the template is smoothed by averaging each bin with its two neighboring bins. The smoothed template is used as a non-closure uncertainty in 4b.

Summary of Background Modeling Uncertainties

Relative magnitudes of the uncertainties for each year are shown in Table 10.5 for ggF and Table 10.4 for VBF along with an estimate of the impact of the statistics of 4b data in the signal region on the total error.

Table 10.4: Magnitude of error components for the VBF analysis in the Signal Region. Total statistic error is quadrature of bootstrap and 2b poisson errors. All errors are in the normalisation before categorisation

SR 4b (Bkgd Yield = 354.6) (%)	**Counts**	**RelativeRelative(%)**
Total Statistic Error	32.3	9.1
Shape Q_N	−1.2	−0.35
Shape Q_W	−0.8	−0.23
Shape Q_S	−0.5	−0.15
Shape Q_E	−0.3	−0.095
Total Systematic Error	1.6	0.45
Total Error	32.3	9.1

A visualization of the magnitudes of these templates for the ggF channels is shown in Figure 10.9 and VBF in Figure reffig:bkgRelErr-VBF-4b. The statistical error dominates for high m_{HH} while the shape systematic from the difference in the CR1 and CR2 contributes more in the moderate m_{HH} region that drives our analysis sensitivity. Although the stat error (from the limited 2b data statistics) is negligible relative to the bootstrap error in the bulk of the distribution, it becomes relevant in the high m_{HH} tail. The final statistical uncertainty used for the limit setting is therefore the sum (in quadrature) of these two components.

10.3.4 Choice of background systematic parametrization

OK - I made these plots... I just need to *pop* them in from somewhere else!

The number of categories, fits and all possible background nuisance parameters are described in table 10.6, informed by the fitting strategy described in the previous sub-section. For both VBF and ggF channels, the shape and normalization uncertainties for each category are obtained separately by comparing the CR1 and CR2 models after each category selection.

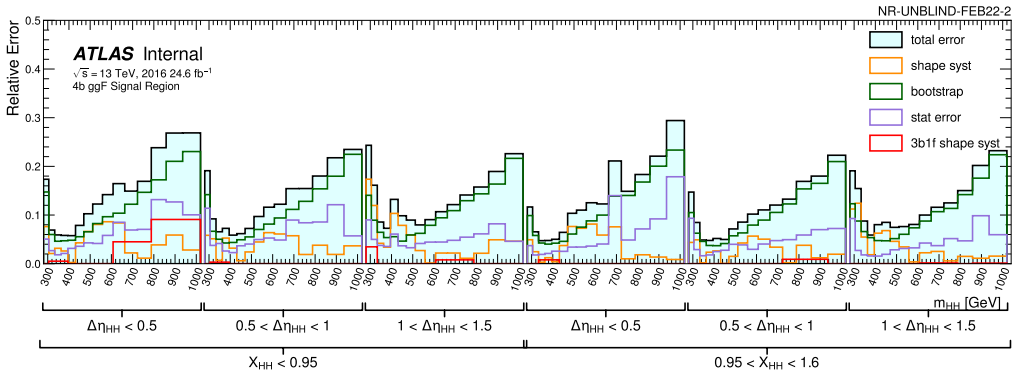
These nuisance parameters are not necessarily completely uncorrelated, and a choice on the correlation scheme must be made to perform the fit. The main argument for correlating the background shape nuisance parameters across kinematic regions is due to the neural network reweighting. The reweighting is trained inclusively with respect to kinematic categories, per year (for the ggF channel) or for all years together (for VBF). Possible mismodelings occuring due to the neural network's

performance, or, more importantly, its ability to interpolate into the signal region, will therefore have the same source. For the VBF channel, one step ahead is taken, and the fit is performed inclusively for all the years for this reason. This approach, however, was observed to be detrimental to the ggF background modeling performance, due to the kinematic differences between the years, and is not done.

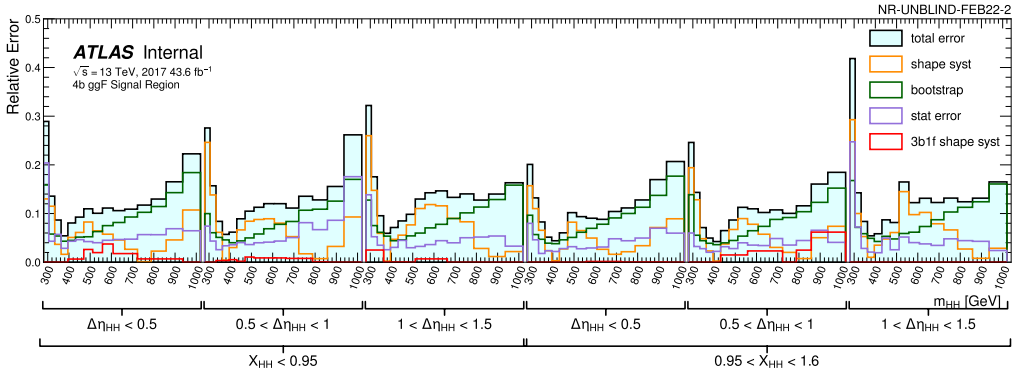
Another important indication that these nuisance parameters should be correlated across kinematic regions is the relative impact they have in the m_{HH} fitting variable. This can be seen on Figure 10.11, which shows the nuisance parameter relative variations within each quadrant and each year of the ggF analysis. Most variations have very similar impacts on m_{HH}, which corroborates the hypothesis of correlated variations.

Table 10.5: Magnitude of error components for the ggF analysis in the Signal Region. Total statistic error is quadrature of bootstrap and 2b poisson errors. All errors are in the normalisation before categorisation

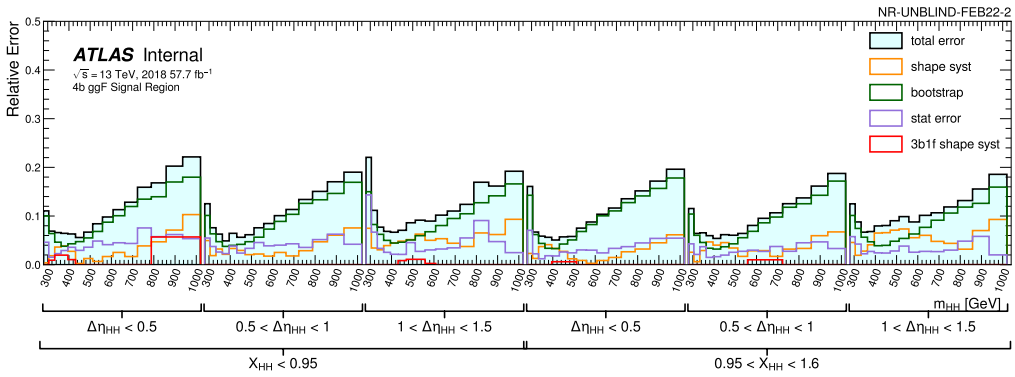
SR 16 4b (Bkgd Yield = 3211.1) (%)	**Counts**	**RelativeRelative(%)**
Total Statistic Error	225.8	7.0
Shape Q_N	−25.3	−0.79
Shape Q_W	−12.1	−0.38
Shape Q_S	16.2	0.50
Shape Q_E	7.0	0.22
3b1f NC	6.8	0.21
Total Systematic Error	33.9	1.1
Total Error	228.3	7.1
SR 17 4b (Bkgd Yield = 4492.7) (%)	**Counts**	**RelativeRelative(%)**
Total Statistic Error	308.9	6.9
Shape Q_N	−24.7	−0.55
Shape Q_W	−4.8	−0.11
Shape Q_S	43.0	0.96
Shape Q_E	14.8	0.33
3b1f NC	17.6	0.39
Total Systematic Error	54.9	1.2
Total Error	313.7	7.0
SR 18 4b (Bkgd Yield = 7720.0) (%)	**Counts**	**RelativeRelative(%)**
Total Statistic Error	483.0	6.3
Shape Q_N	−141.3	−1.8
Shape Q_W	−55.8	−0.72
Shape Q_S	7.0	0.091
Shape Q_E	−87.1	−1.1
3b1f NC	42.6	0.55
Total Systematic Error	180.4	2.3
Total Error	515.6	6.7



(a) 4b 2016 ggF



(b) 4b 2017 ggF



(c) 4b 2018 ggF

Figure 10.9: Relative error contributions of the background for the 4b ggF discriminant.

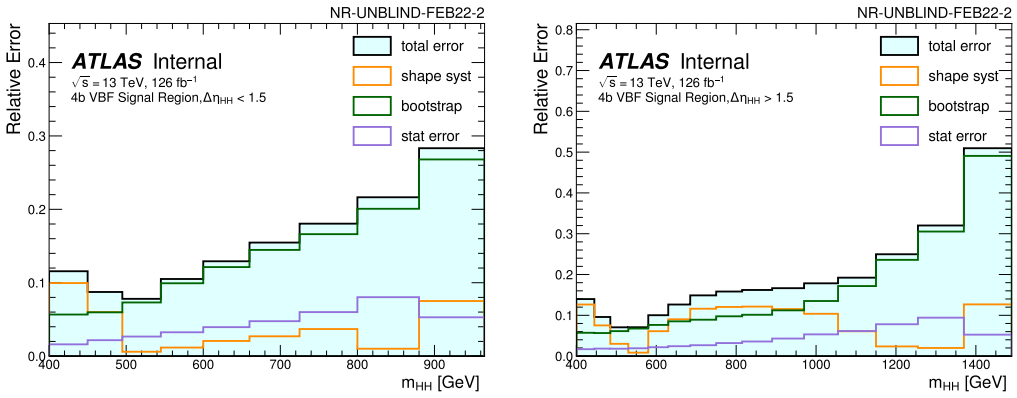


Figure 10.10: Relative error contributions of the background for the 4b VBF discriminant.

	ggF channel	VBF channel
Categories	3 years $\times$ 3 $\lvert\Delta\eta_{HH}\rvert$ $\times$ 2 X_{HH} = 18 categories	2 $\lvert\Delta\eta_{HH}\rvert$ categories
Background shape nuisance parameters	3 years $\times$ $(Q_{\mathrm{N}}, Q_{\mathrm{E}}, Q_{\mathrm{S}}, Q_{\mathrm{W}})$ = 12	$(Q_{\mathrm{N}}, Q_{\mathrm{E}}, Q_{\mathrm{S}}, Q_{\mathrm{W}})$ = 4
3b1f non-closure nuisance parameters	18 (1 per category)	0

Table 10.6: Summary of categorization strategy and background-related nuisance parameters in ggF and VBF analyses.

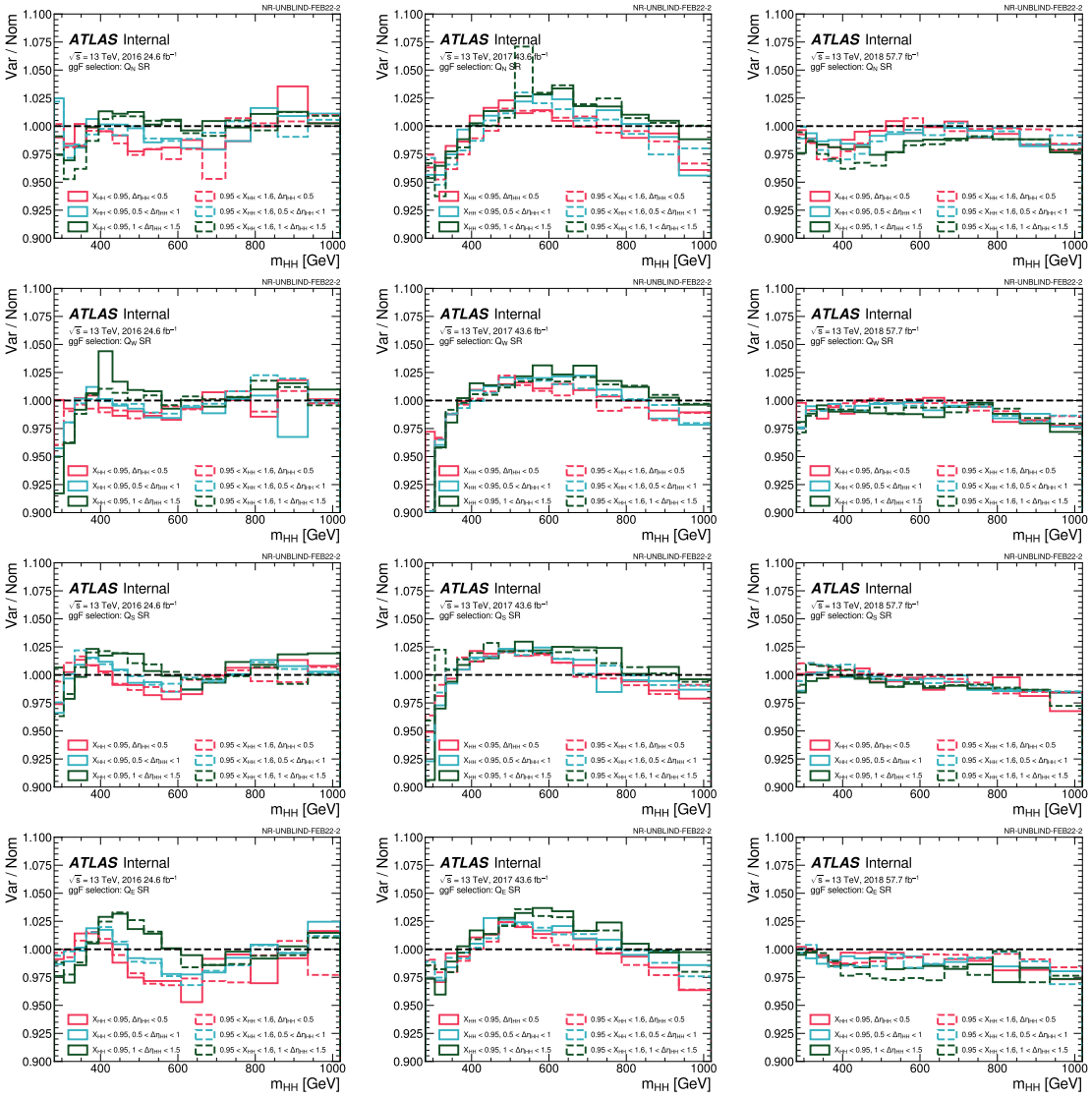


Figure 10.11: Impact of background shape nuisance parameter variation on m_{HH} in different kinematic categories for the ggF channel. Each column is a different year for the ggF channel templates while the rows show the SR NP quadrants.

Given the arguments above, we choose to correlate the background shape nuisance parameters across kinematic categories, within a year and each quadrant. Therefore, the number of free nuisance parameters in the ggF fit related to the background shape is reduced from 72 (3 years $\times$ 6 categories $\times$ $(Q_{\mathrm{N}}, Q_{\mathrm{E}}, Q_{\mathrm{S}}, Q_{\mathrm{W}})$) to 12 (3 years $\times(Q_{\mathrm{N}}, Q_{\mathrm{E}}, Q_{\mathrm{S}}, Q_{\mathrm{W}})$).

For completeness, we also show the templates by category with the different quadrant contributions overlaid in Figure 10.12.

The initial analysis unblinding occurred with a different nuisance parameter correlation scheme, where all parameters were treated as uncorrelated systematics. This choice was made due to the lack of coherent variations in the different validation regions in the analysis, particularly the 3b1f region. Therefore, we believed giving flexibility for the fit to vary in different directions in different categories to be the best option. In addition, it was observed that the expected SR limits with the uncorrelated were more performant across different κ_λ hypotheses. Results pertaining to this unblinding, which show clear indication and preference for the correlated scheme described above, can be seen in Appendix **??**. This appendix also collects extensive studies on understanding these results.

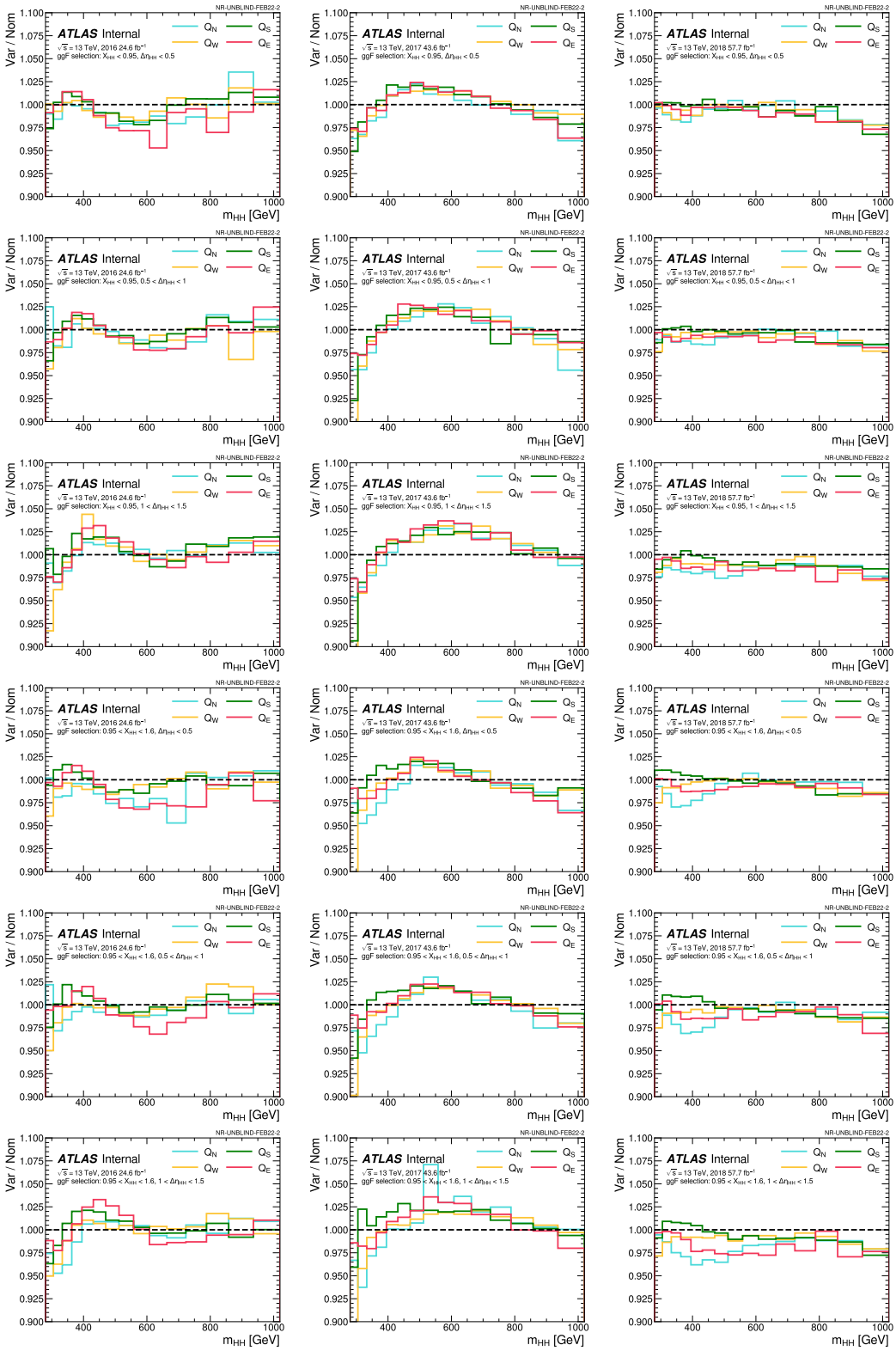


Figure 10.12: Impact of background shape nuisance parameter variation on m_{HH} in different kinematic categories for the ggF channel. Each column is a different years of the ggF channel templates while the rows show each category with the SR NP quadrants overlaid.

10.4 Background validation

- Smth about inverting *all* of the cuts?
- Since the focus of my thesis was the ggF analysis and , I will share the results from that. Also, since the VBF analysis is so statistically dominated, it was easier to see closure just because the stats were so much lower.
- Emphasize how each of these procedures results in retraining all of the NNs for the bootstrap and CR1,2 shape difference
- Maybe emphasize how the 3b1f systematic is not included in these studies

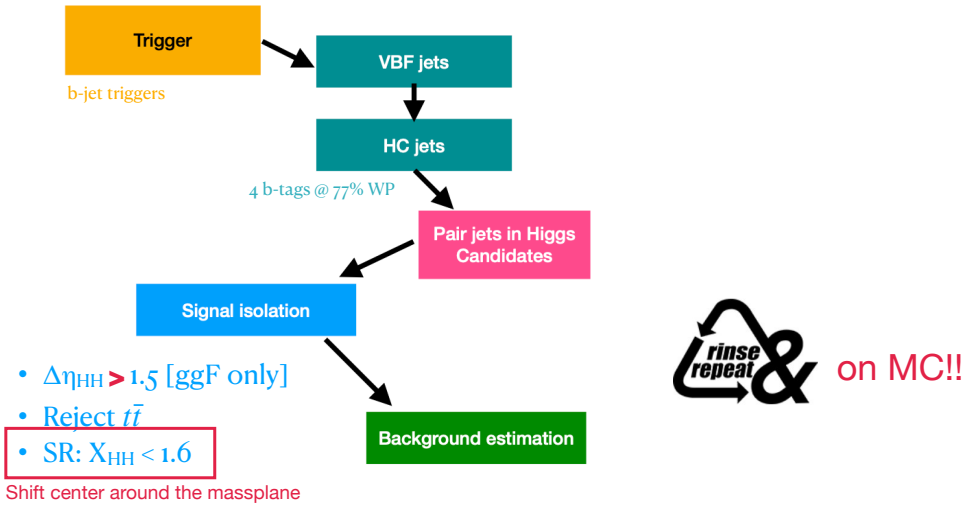
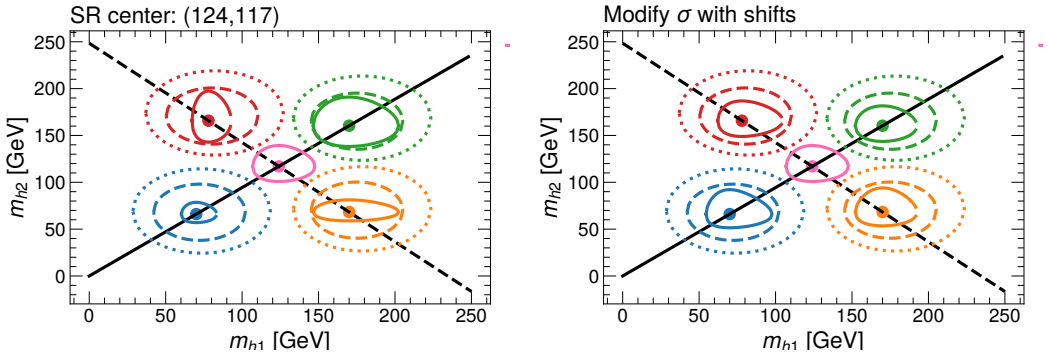


Figure 10.13: Illustration of the modified cuts to test the background estimate strategy.

10.4.1 Reversed $|\Delta\eta_{HH}|$

- $1.5 < |\Delta\eta_{HH}| < 2.5$
- $2.5 < |\Delta\eta_{HH}| < 3.6$
- $|\Delta\eta_{HH}| > 3.6$



(a) Just shift the centers of X_{hh}

(b) X_{hh}^{shift} definition

Figure 10.14: Motivation for choices of shifted SRs.

10.4.2 Shifted regions

Another check that we tried was taking the signal region and moving it around the massplane. Since the equation for the SR defining variable $X_{HH} = \sqrt{\left(\frac{m_{H1}-124\,\text{GeV}}{0.1\ m_{H1}}\right)^2 + \left(\frac{m_{H2}-117\,\text{GeV}}{0.1\ m_{H2}}\right)^2}$ (introduced in Eq. 9.3) has a radius that depends on the m_{H1}, m_{H2}, so just changing the center locations had the effect that the SR area got larger as we moved to larger HC masses, as shown in Figure **??**. To ameliorate this effect, we modified the HC resolutions in the X_{HH} formula so that the SR size would be compatible with the nominal as we moved around the massplane:

$$X_{hh,\text{shift}} = \sqrt{\left(\frac{m_{H1} - m_{H1,\text{center}}}{\sigma_{m_{H1}}\ m_{H1}}\right)^2 + \left(\frac{m_{H2} - m_{H2,\text{center}}}{\sigma_{m_{H2}}\ m_{H2}}\right)^2} \tag{10.3}$$

$$\text{with} \quad \sigma_{m_{H1}} = 0.1 \times \frac{124}{m_{H1,\text{center}}}, \quad \sigma_{m_{H2}} = 0.1 \times \frac{117}{m_{H2,\text{center}}}.$$

This modification allowed the shifted regions to have approximately equal areas with the standard signal region, as seen in Figure **??**. Although we originally tested SRs that translated both with higher and lower HC masses, we did not have good closure for the reweighting with the "lower left" SR. We believed understand this because by examining Figure 10.15, a SR at (70,66) GeV has CRs that are overlapping the kinematic turn on curve in the massplane. Our intuition tells us that it's harder to form a reliable background estimate if the underlying control region is not smoothly varying into the CR, so we believed that failing to reweight in this region was not a methodological issue, but rather a task that was significantly more difficult than our nominal estimate.

The boundary of the CRs just rotates with the center of the SR, and the quadrants are defined in a similar way, as illustrated in Figure **??**. One interesting note is for the lower right SR, the lower quadrant for CR1 has a couple of issues. It (1) intersects with the W-veto from the X_{Wt} cut and

(2) overlaps the kinematic threshold since basically no m_{H2} values below 25 GeV to populate the full lower quadrant. However, the CR2 (left, right) quadrants did not suffer from these two issues (which made the test more challenging than we believed our actual SR is), so in the following the CR2 derived weights are shown as the nominal set. Table 10.7 enumerates of the shifted SR centers and which CR is taken to be the nominal set.

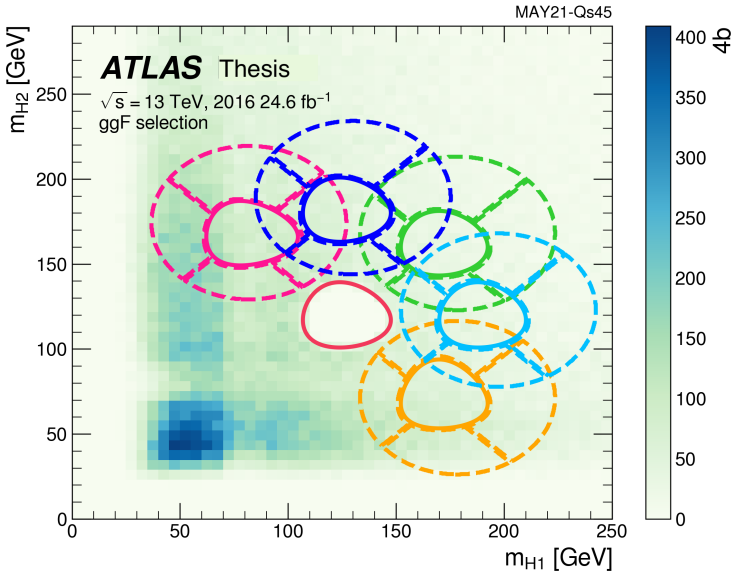


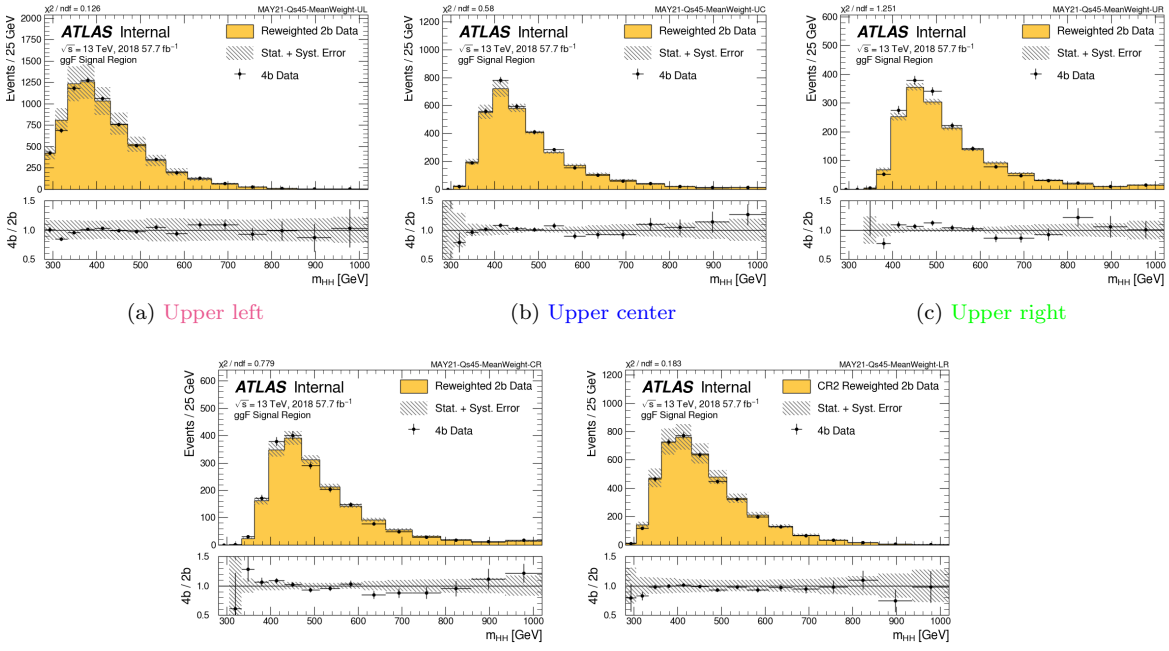
Figure 10.15: The shifted regions for the background validation, with the pink solid curve in the center showing the nominal SR.

(Shifted) Signal Region	m_{H1} center [GeV]	m_{H2} center [GeV]	nominal rw set
upper left	78	166	CR1
upper center	124	180	CR1
upper right	170	166	CR1
center right	188	117	CR1
lower right	170	68	CR2
nominal	124	117	CR1

Table 10.7: Center locations for the shifted SRs validation study. Also included is which quadrants are considered the "nominal" for the background estimate.

The reweighting functions were derived for each of these regions and for all of the year dataset, but to exemplify the agreement we were seeing, in Figure 10.16 we show the m_{HH} predictions compared to the observed 4b data. In the ratio panels, you can see that there is good agreement already with the given error bars, and gave us confidence in the background estimation procedure that we had set up.

To further estimate this, a comparison of the predicted and observed yields was made and showed



(a) Upper left (b) Upper center (c) Upper right

(d) Center right (e) Lower right

Figure 10.16: m_{HH} distributions of reweighted 2b data and 4b data in the shifted SRs for the 2018 background estimates. The background error bar includes the 2b Poisson, deep ensembles, and the CR1 / CR2 shape difference.

in Table 10.8. To quantify the level of agreement between the observation and the predcition,

$$\mu_{\text{norm}} = \frac{(N_{\text{bkgd}} - N_{\text{target}})}{\sigma_{\text{stat}}} \tag{10.4}$$

where σ_{stat} includes the statistical errors from: the 4b Poisson error (on the obs) the 2b Poisson error (on the background template) and the deep ensembles error (on the background template from retraining 100x) – all summed in quadrature. To quantify the difference across the years and regions, Figure **??** shows a histogram of the deviation μ_{norm}, and the Gaussian fit to the μ_{norms} has a mean of -0.08 and a standard deviation of 1.19, which is close to the mean 0 and standard deviation of 1 corresponding to good modeling.

Finally, in Figure 10.18 shows the pull plots for the fit with the background nuisance parameters to the observed data. Since the best fit values are within the 1σ error bands of the inputted templates, this gave us confidence moving forward that our background model well described the shifted SR data, giving us confidence moving forward with unblinding.

Shifted Region	Year	4b Yield	Background Prediction	Deviation (μ_{norm})
upper left	2016	4068 ± 64	4101.6 ± 474.4	0.07
	2017	5586 ± 75	5843.4 ± 514.3	0.50
	2018	9421 ± 97	9559.7 ± 1051.2	0.13
upper center	2016	2197 ± 47	2086.3 ± 52.4	-1.57
	2017	3017 ± 55	2987.9 ± 72.8	-0.32
	2018	5161 ± 72	5058.6 ± 141.6	-0.65
upper right	2016	1182 ± 34	1125.8 ± 38.8	-1.08
	2017	1738 ± 42	1684.4 ± 40.2	-0.93
	2018	2831 ± 53	2732.7 ± 48.4	-1.37
center right	2016	1305 ± 36	1310.2 ± 43.6	0.09
	2017	1922 ± 44	1951.2 ± 73.7	0.34
	2018	3098 ± 56	3108.0 ± 98.6	0.09
lower right	2016	2658 ± 52	2664.2 ± 300.0	0.02
	2017	3635 ± 60	3814.4 ± 326.7	0.54
	2018	6084 ± 78	6241.1 ± 491.1	0.32

Table 10.8: 4b and background prediction in the signal region in the shifted regions in 2016. The error of background prediction includes the 2b poisson statistic error, the bootstrap error and the shape systematic error.

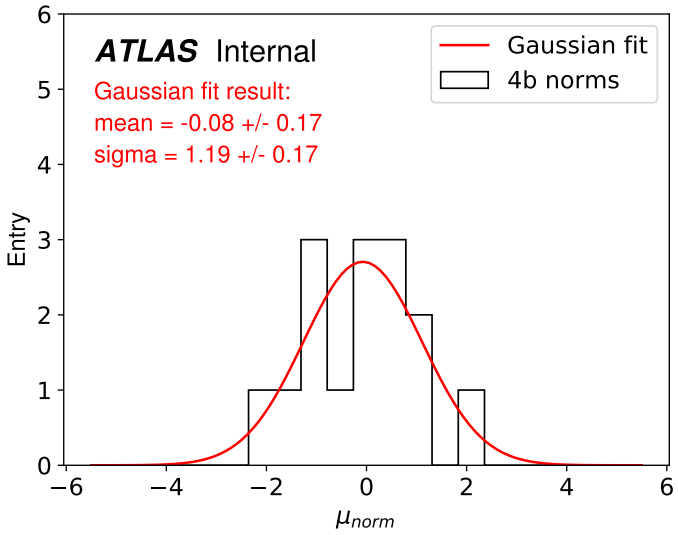
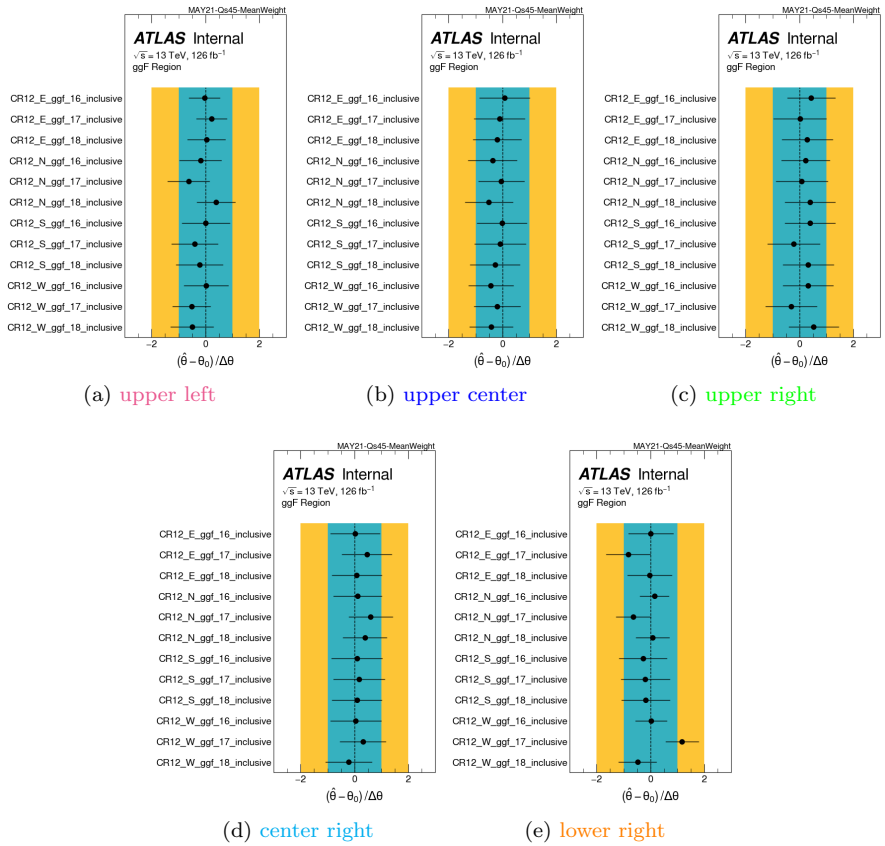


Figure 10.17: μ_{norm} distribution in the shifted regions. 4b normalizations (black) and the gaussian fit (red) are shown.



(a) upper left (b) upper center (c) upper right

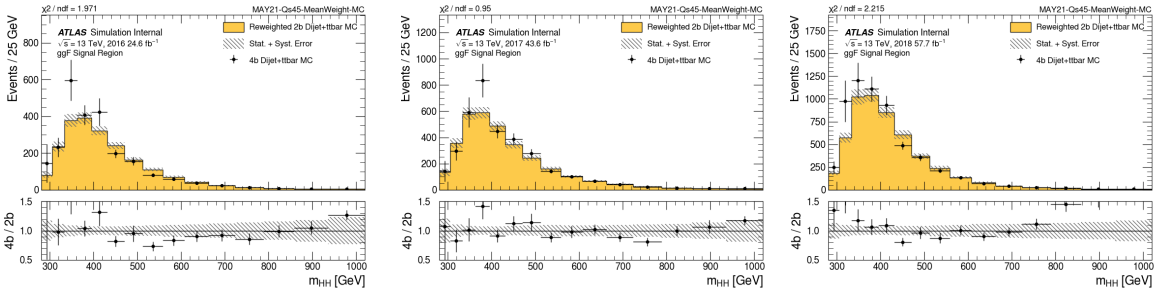
(d) center right (e) lower right

Figure 10.18: Background only pull plots in the shifted regions

10.4.3 MC validation

- A monte carlo sample of QCD (with pythia) and $t\bar{t}$ were also used to test the 4b SR before unblinding.

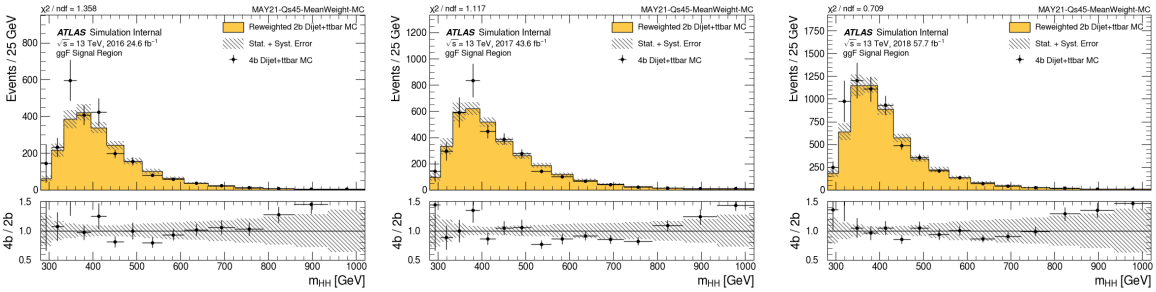
Validate with data trained networks



(a) m_{HH} in 2016 (b) m_{HH} in 2017 (c) m_{HH} in 2018

Figure 10.19: m_{HH} **data reweighted** 2b events and 4b events **evaluated on the QCD and $t\bar{t}$ MC** samples. The background estimate error includes the 2b poisson, deep ensembles, and CR1/CR2 shape systematic errors.

Validate with MC (re)trained networks



(a) m_{HH} in 2016 (b) m_{HH} in 2017 (c) m_{HH} in 2018

Figure 10.20: m_{HH} **MC reweighted** 2b events and 4b events evaluated on the QCD and $t\bar{t}$ MC samples. The background estimate error includes the 2b poisson, deep ensembles, and CR1/CR2 shape systematic errors.

11

Results

11.1 Background modelling

Pre-fit distributions

I think this is a great place to show the unrolled plots year-by-year. (I might try to overlay the postfit plots in the next step though).

- Background systs uncorrelated across the years, but correlated across the X_{HH} and $|\Delta\eta_{HH}|$ cats within a year
- VBF fits all of the years together

11.1.1 B-only fits (?)

Post-fit (ggF) background only fits:

Q: Do I also have the unrolled post-fit plots in the folder?

New organization ideas:

- I want to over pre and post-fit *just with unrolled plots* to quantify the non-closure.
- Then for the plots that also overlay the signals, I will put all of the years on top of each other.

The NP pulls are shown in Figure **??**. No large constraints are observed. Some pulls from some components are expected due to the transition from CR to SR. Especially Q_E and Q_S in 2016 ggF provide shape variations that particularly impact the m_{HH} peak (see top left plot in Figure 10.12), therefore they are the most pulled. For the background estimation uncertainties, the nuisance parameter naming scheme is alpha_{syst}_{quad}_{year}. For example 'alpha_CR12_shape_N_ggf_16' stands for 'CR12 shape uncertainty North quad in 2016'. Note that 'Non-Closure 3b1f' doesn't have a split in quadrants. From a S+B fit, a correlation matrix is derived and shown in Figure **??**.

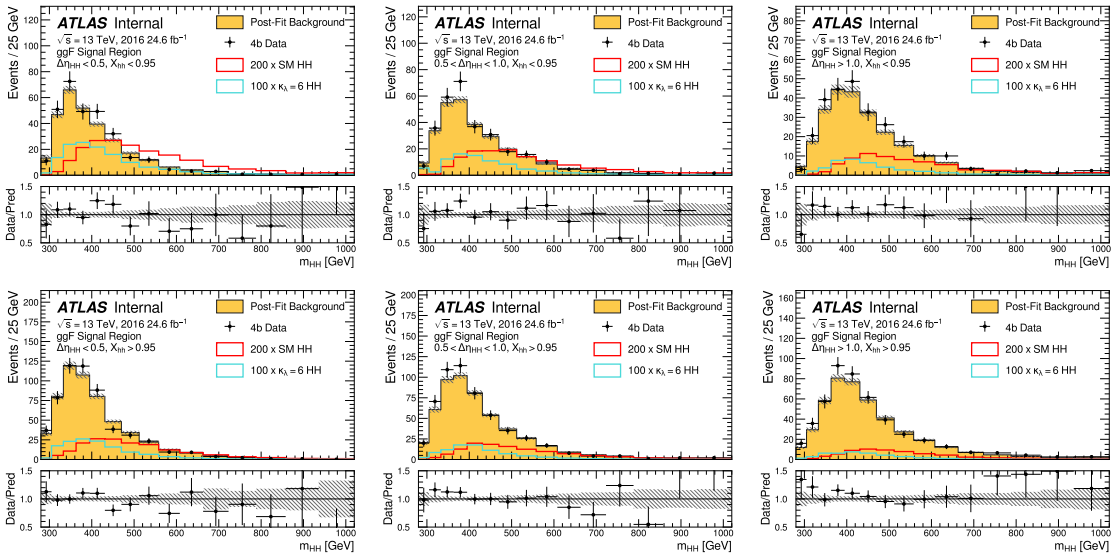


Figure 11.1: ggF 2016 background only post-fit plots.

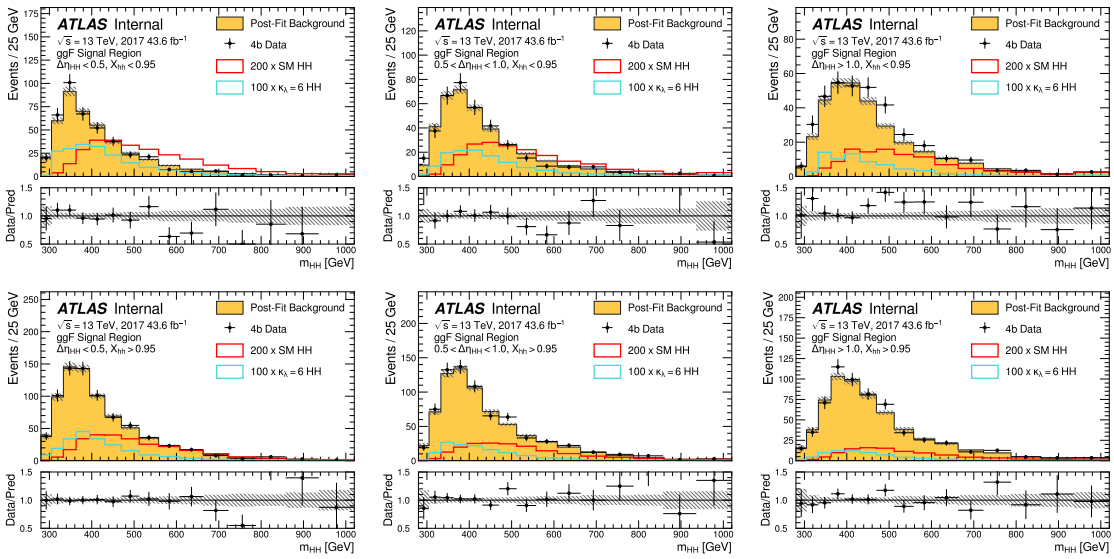


Figure 11.2: ggF 2017 background only post-fit plots.

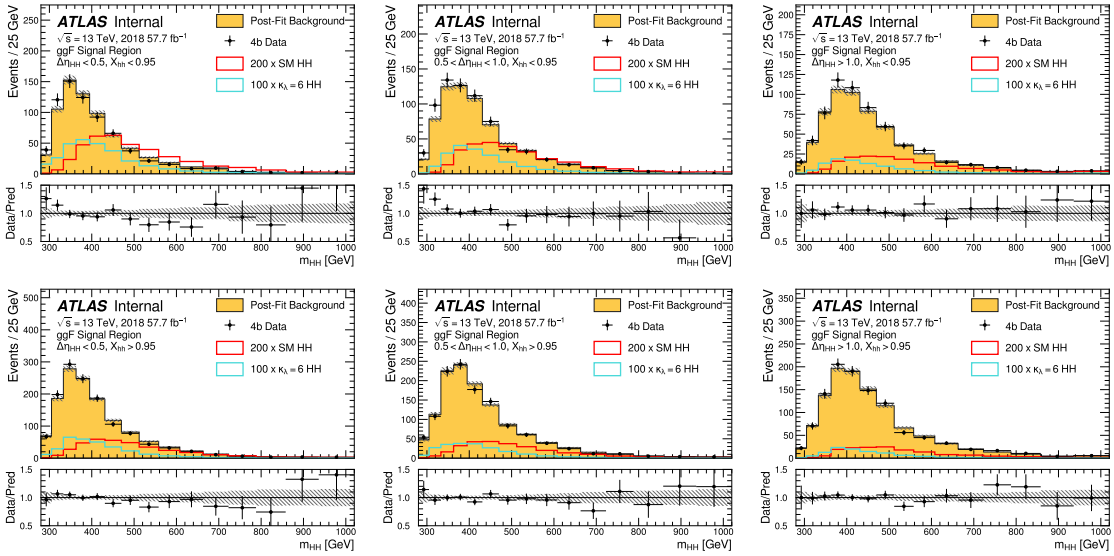


Figure 11.3: ggF 2018 background only post-fit plots.

VBF cat only fits

Similar to the ggF channel, the background shape systematics (CR12) are treated as correlated across $|\Delta\eta_{HH}|$ bins in the VBF channel. Since we derive the VBF background estimate with the years inclusively and do the fits inclusively, we do not have separate background NPs for each year in the VBF channel.

The post-fit VBF plots in the $\Delta\eta_{HH}$ categories are shown in Figure 11.4.

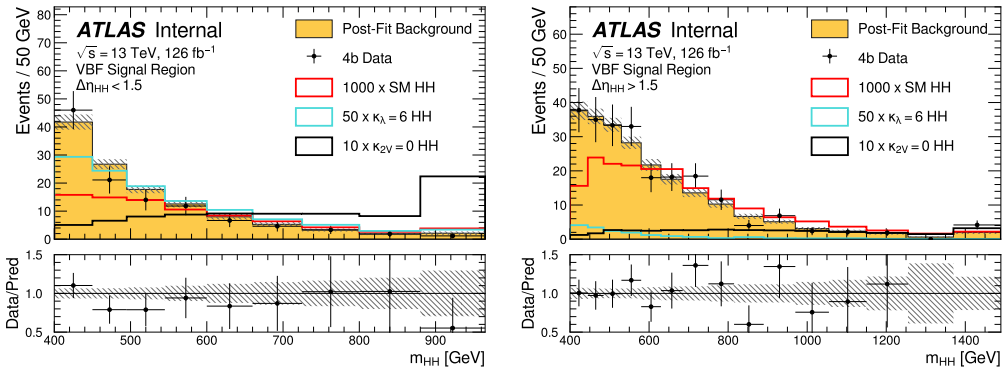


Figure 11.4: VBF background only post-fit plots.

11.2 Overview of signal systematic uncertainties

Although the data-driven background systematics drive the sensitivity of the analysis, we characterize our signals with simulated data, so these samples have additional NPs which account for the known mismodellings of the simulation compared to the data derived in dedicated control regions Except for the E_{T} SFs (which were described in detail in Section 9.1.3), all these are not custom to the analysis strategy, but follow a dedicated prescription from the collaboration. As such, and because they don't drive our analysis sensitivity, we will just give a very brief summary of them here.

Detector modelling uncertainties

To account for the difference in the tagging probability in data, the dedicated FTAG SFs for the b, c, and light-jets are applied using the variations measured by the calibrations (described in Section 6.5). The variation on the jets' energy scale and resolution (JES and JER) are applied, as well as the uncertainty due to the pile-up veto JVT tagger. We apply an uncertainty for the pile-up reweighting factors that are applied to correct the simulation to the pile-up distribution in data. Trigger uncertainties from the HLT b-tag are prescribed by the b-jet trigger group and the dedicated E_{T} SFs derived for this analysis are also applied. How is the uncertainty on ET calculated? The uncertainty on the integrated luminosity is a 1.7% uncertainty on the signal normalization.

Theoretical Uncertainties

Differences between the parton shower (PS) and underlying event (UE) are assessed using the differences between the nominal Pythia 8 sample (which uses the Lund string model for the PS) and an alternative Herwig 7 sample (which uses the cluster hadronization model for the PS) This uncertainty is the largest in the analysis, showing up as an up to 10% impact on the ggF and VBF acceptances for the. The $\pm 1\sigma$ variation templates are derived exclusively for each category, but constrained with a single NP.

The uncertainty in the matrix element is assessed by varying the renormalization and factorization scales (μ_R and μ_F) up and down by a factor of 2. The size of this uncertainty is $\approx$ a 2% difference on the signals, although it gets as large as 6% in some analysis categories. The uncertainties due to the pdfs of the colliding partons are calculated from the 1σ error bar on the pdf replicas.

For the limits on the SM signal strength, we also include the 3.5% normalization uncertainty on the $H \to b\bar{b}$ branching ratio, and the uncertainties on the theoretical cross-section due to the pdf, α_s, renormalization scheme, and m_t scale are accounted for.

11.2.1 S+B fits

From a background-only fit and signal + background fits, the NP pulls are shown in Figure 11.5.

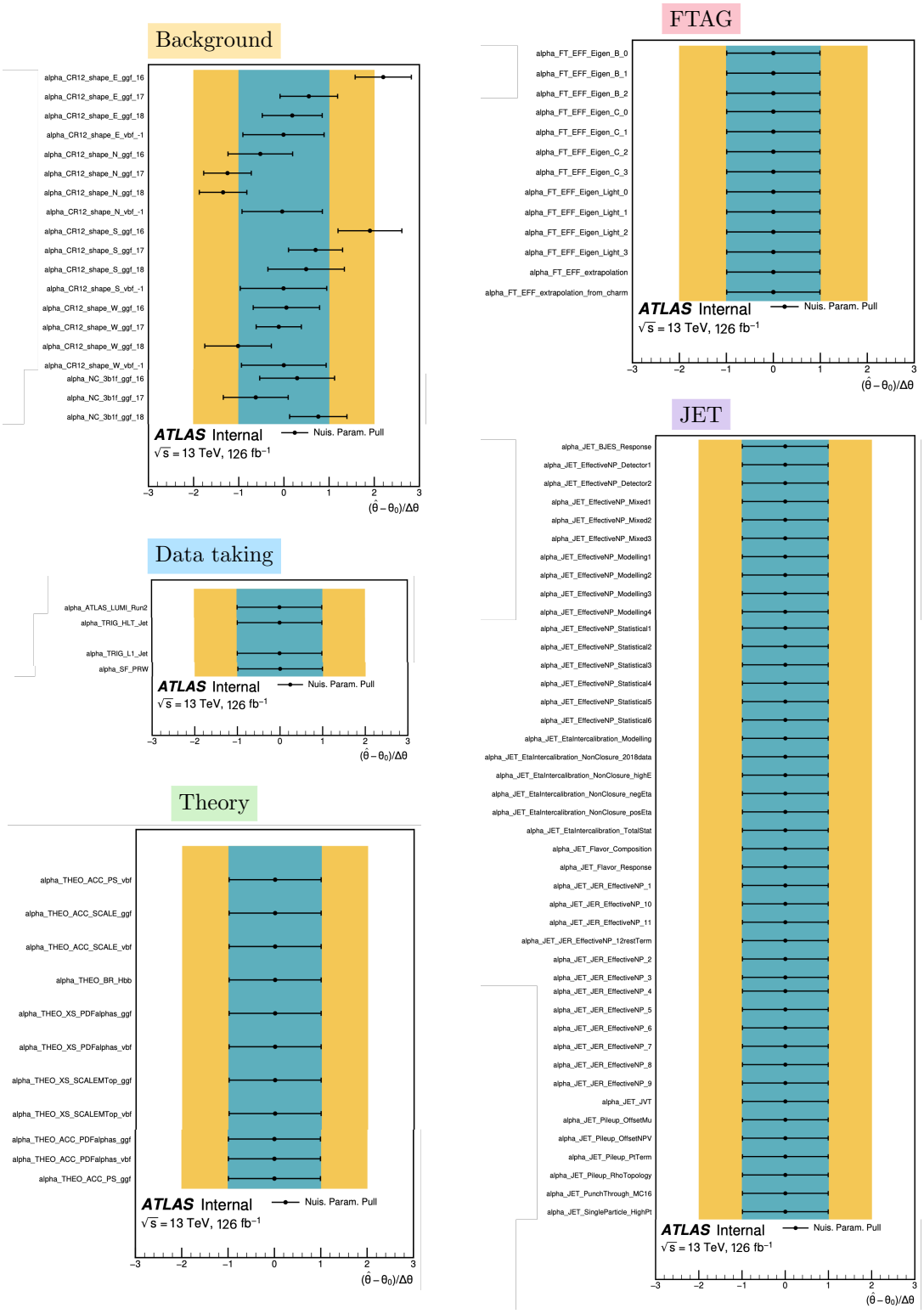


Figure 11.5: Pulls for the fit to the background template.

Systematic identifier	Description	# of NPs
Background	ggF: CR1 / CR2 Shape variations	3 years x 4 SR quadrants = 12
	VBF: CR1 / CR2 Shape variations	4 SR quadrants = 12
	3b1f unc (ggF only)	3 years
	Bootstrap	# of histogram bins
FTAG	b-jet	3 eigenvariations and 1 extrapolation
	c-jet eigendecomp	4 eigenvariations and 1 extrapolation
	light-jet eigendecomp	4 eigenvariations
JET	JES	30
	JER	13
	JVT	1
Data taking	pile-up reweighting	1
	Luminosity	1
	HLT + L1 trigger SFs	2
Theory	PS, pdf, and scale uncertainties	3 x 2 (ggF + VBF)) = 6
	BR $H \to b\bar{b}$ (*)	1
	Cross-section uncertainties (*)	4

Table 11.1: Note: the (*) indicates these are not included in the $\kappa_\lambda$and 2v scans.

11.3 Limit plots

11.3.1 ggF and VBF Channel Combination

The observed (expected) upper limit on the SM ggF+VBF signal strength, including the cross section uncertainty from theory calculations, is 5.45 (8.09) and shown in Table 11.2. The limit set on $\kappa_\lambda$using the combined ggF and VBF channels are the strongest constraints set on this coupling in the $b\bar{b}b\bar{b}$ final state in ATLAS to date.

Expected and observed upper limits on the ggF+VBF cross section are $7.36 \times 32.776 = 241.20$ $5.02 \times 32.776 = 164.49$, where the theory calculation uncertainties on the ggF and VBF cross sections are excluded.

Table 11.2: The observed and expected upper limit on the SM HH production cross-section at the 95% CL. The expected value is shown with corresponding one and two standard deviation error bounds.

	Observed	-2σ	-1σ	Expected	$+1\sigma$	$+2\sigma$
ggF production: ggF channel fit	5.73	4.44	5.97	8.28	12.51	19.71
ggF production: ggF+VBF channels fit	5.51	4.41	5.92	8.22	12.43	19.62
VBF production: VBF channel fit	122.9	72.0	96.7	134.3	194.2	281.6
VBF production: ggF+VBF channels fit	132.3	71.3	95.7	132.8	191.9	277.7
ggF+VBF production: VBF+ggF channels fit	5.45	4.34	5.83	8.09	12.21	19.19

Both the ggF and VBF HH production modes are sensitive to κ_λ, and, as such, are accounted for when setting constraints on this coupling. In order to combine sensitivities from both ggF and VBF, orthogonal channels are defined for each process (see Section **??** for details). Both channels are fit simultaneously, with the same signal strength, μ, associated with both signal processes.

Figure 11.6 shows the upper limit set on the combined ggF and VBF HH production cross-section as a function of κ_λ.

$\kappa_\lambda$values where the limit set on the cross-section is below the theoretical value are excluded at the 95% CL. The constraint set on $\kappa_\lambda$from the combination of the ggF and VBF channels is shown in Table 11.3 alongside the constraints from the individual channels.

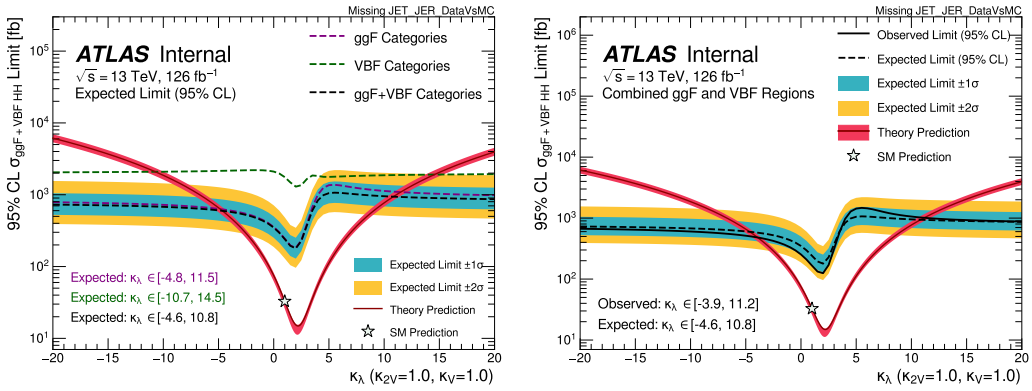


Figure 11.6: The 95% CLs limit on the combined ggF and VBF HH production cross-sections. Left plot is a breakdown of channels.

Table 11.3: In the combined channel, the observed and expected limit intervals on the coupling modifier $\kappa_\lambda$at the 95% CL for the ggF channel, the VBF channel and the combination of the two.

Channel	Observed Interval	Expected Interval
ggF Channel (ggF signal only)	$[-4.5, 13.3]$	$[-5.0, 12.0]$
VBF Channel (VBF signal only)	$[-10.0, 13.2]$	$[-12.3, 15.6]$
Combination (ggF+VBF signals)	$[-3.9, 11.2]$	$[-4.6, 10.8]$

The scan for $\kappa_{2V}$for the VBF production, but combining the ggF and VBF channels is shown in Figure 11.7. The ggF production is accounted for as a (very sub-dominant) background process, and the limit is set on the VBF signal production. However, due to additional contributions of VBF events in the ggF SR, the expected constraint on κ_{2V}[-0.05, 2.12] is slightly tighter than the VBF channel only constraint of [-0.08, 2.16].

The 2D limit on the $\kappa_{2V}$and $\kappa_\lambda$variations for the combined ggF + VBF channels is shown in Figure 11.8 assuming $\kappa_V = 1$.*Limitsarederivedfromtheintersectionofthecrosssectionlimitandthetheorypredictionsa*

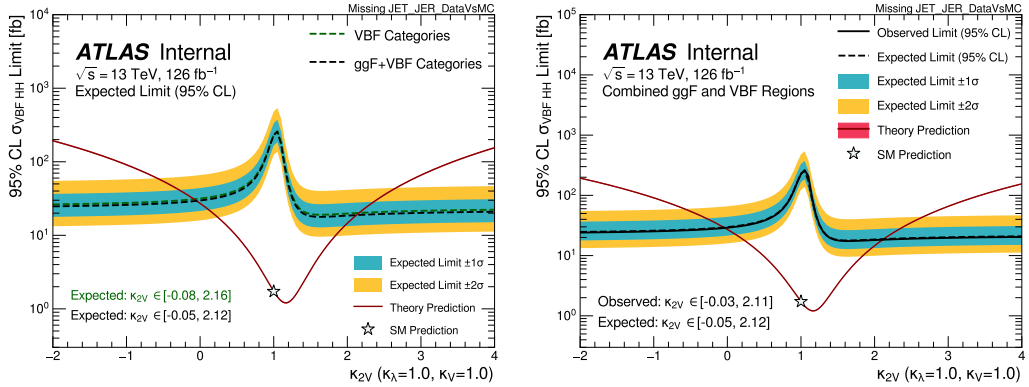


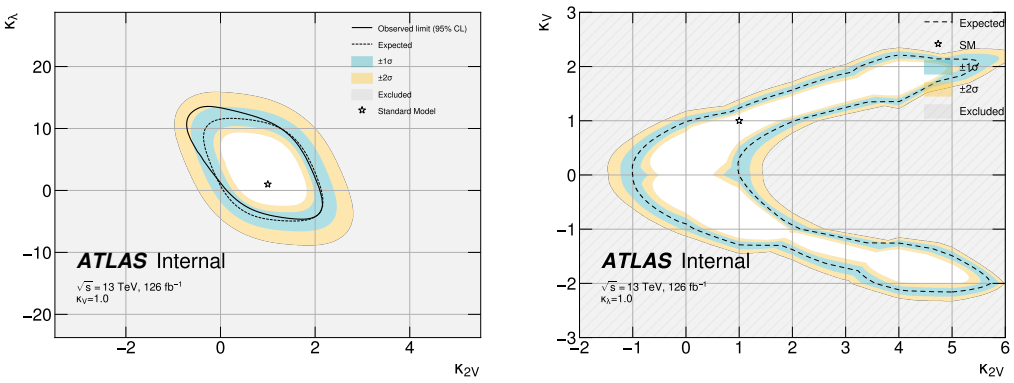
Figure 11.7: The limit intervals on the coupling modifier κ_{2V}at the 95% CL for the combination of the VBF+ggF channels. Left plot is a breakdown of channels.

11.3.2 Improvements from previous analyses

To do: Make the history of the analysis plot!!

To do: Remake the cat improvements plots!!

11.4 Combination result



(a) (b)

Figure 11.8: The obs (solid) and expected (dashed) limit intervals on the coupling modifiers $\kappa_\lambda$vs κ_{2V}(a) and $\kappa_V$vs κ_{2V}(b) at the 95% CL for the combination of the VBF+ggF channels. *TODO: (b) is expected only and with only the background shape uncertainties included. To be updated.*

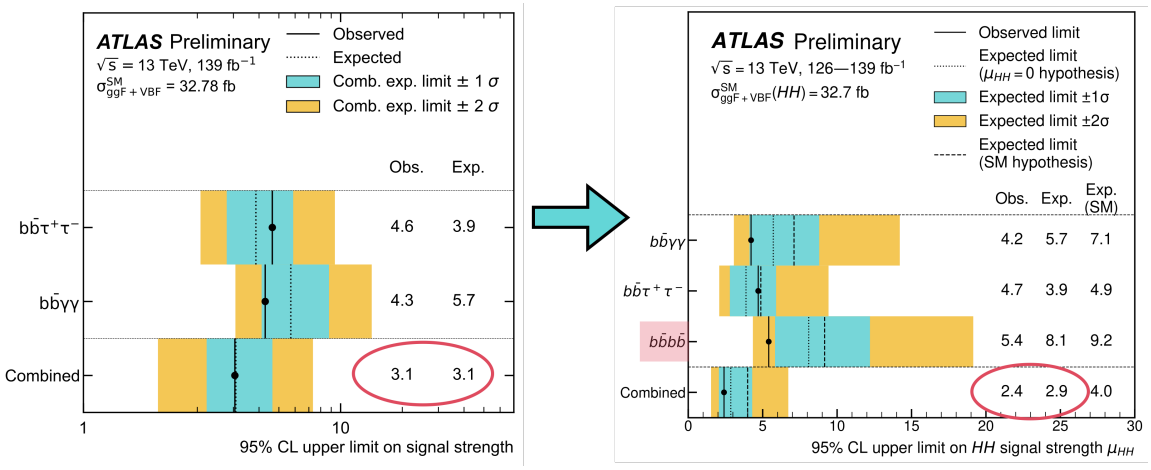


Figure 11.9: [**ATLAS-CONF-2021-052**] [**ATLAS-CONF-2022-050**]

12

Conditional generative models for data-driven background modeling

Then shalt thou see, and flow together, and thine heart shall fear, and be enlarged; because the abundance of the sea shall be converted unto thee.

– Isaiah 60:5

Go with the flow. Force nothing. Let it happen, or not happen..trusting that whichever way it goes, it's for the best.

– Mandy Hale

Do not struggle. Go with the flow of things, and you will find yourself at one with the mysterious unity of the Universe.

– Chuang Tzu, Chinese philosopher (c. 369 B.C. - 286 B.C.)

13

Conclusions

Journey before destination.

– Brandon Sanderson, *The Way of Kings.*

When you dance, your purpose is not to get to a certain place on the floor. It's to enjoy each step along the way

– Wayne Dyer.

DIPS conclusion

DIPS, a new algorithm for identifying heavy flavour jets with impact parameter information and based on the Deep Sets architecture, has been introduced and is shown to be comparable in performance and up to a factor of 3 to 5 faster to train and evaluate over the baseline recurrent neural network based algorithm RNNIP when using the same inputs. The large speed-up of the algorithm facilitates optimisation, and an optimised DIPS with loosened track selections and additional per-track features was shown to improve light-flavour jet rejection by up to a factor of 2.5 and c-jet rejection by up to a factor of 1.5 over the baseline DIPS algorithm, which already outperforms the current RNNIP algorithm by up to 15%. As such, DIPS represents a promising future direction for neural network-based flavour tagging algorithms. Moreover, the parallelisability and increased speed of DIPS not only has the potential to reduce the computational load of the ATLAS reconstruction, but also makes DIPS an excellent candidate for trigger applications where extremely low latency is required.

Appendix A

Tracking optimizations impact on flavor tagging

A.1 Lifetime signage

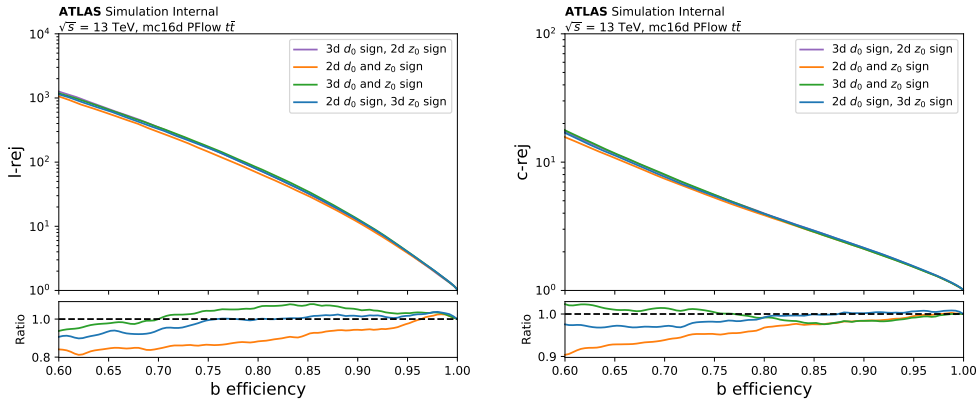
A.2 Different track reconstruction algorithms

A.2.1 Pseudo-tracking

A.2.2 SCT splitting

A.2.3 Looser B-cuts

A.2.4 DIPS retrianing



(a) light rejection

(b) charm rejection

Figure A.1: Retraining DIPS with different specifications for the signs fo d_0 and $z_0 \sin\theta$.

Appendix B

Reweighting loss function

The derivation below follows the proof from [.]

Remember in Section 10.1 we introduced the loss function:

$$\mathcal{L}[Q] = \mathbb{E}_{x\sim p_{2b}}\left[\exp\left(\frac{1}{2}Q(x)\right)\right] + \mathbb{E}_{x\sim p_{4b}}\left[\exp\left(-\frac{1}{2}Q(x)\right)\right] \tag{B.1}$$

We want to show that $Q^*(x) = \log w(x)$, where $w^(x) = p_{4b}(x)/p_{2b}(x)$, minimizes this loss.

Since log is a (monotonically increasing) bijective function, we can just show that the loss function with respect to w is minimized when $w(x) = p_{4b}(x)/p_{2b}(x)$.

$$\mathcal{L}[w] = \mathbb{E}_{x\sim p_{2b}}\left[\sqrt{w(x)}\right] + \mathbb{E}_{x\sim p_{4b}}\left[\frac{1}{\sqrt{w(x)}}\right] \tag{B.2}$$

$$= \int \left[\sqrt{w(x)}p_{2b}(x) + \frac{1}{\sqrt{w(x)}}p_{4b}(x)\right] dx \tag{B.3}$$

This function that is the extrema of this integral can be solved for by the calculus of variations. Let

$$\mathcal{I}(x_1, \ldots, x_n, w, w') = \sqrt{w(x)}p_{2b}(x) + \frac{1}{\sqrt{w(x)}}p_{4b}(x).$$

The solution to the extrema of the loss is given by the Euler-Lagrange equation:

$$\frac{\partial \mathcal{I}}{\partial w} - \frac{d}{dx_i}\frac{\partial \mathcal{I}}{\partial w'} = 0$$

Since there is no explicit dependence on the x_is, we just need to solve for $\frac{\partial \mathcal{I}}{\partial w} = 0$.

$$\frac{\partial \mathcal{I}}{\partial w} = \frac{1}{2} w^{-1/2} p_{2b} - \frac{1}{2} w^{-3/2} p_{4b} = 0$$
$$\implies w(x) = p_{4b}/p_{2b}$$

This is the w^* that we were searching for! To show that it's also a minimum, take the 2^{nd} functional derivative:

$$\frac{\partial^2 \mathcal{I}}{\partial w^2} = -\frac{1}{4} w^{-3/2} p_{2b} + \frac{3}{4} w^{-5/2} p_{4b} = \frac{1}{4} w^{-3/2} \left(3 p_4 b - \frac{1}{4} w p_{2b} \right) \tag{B.4}$$

and evaluate it at the extrema point:

$$\left. \frac{\partial^2 \mathcal{I}}{\partial w^2} \right|_{w = p_{4b}/p_{2b}} = \frac{1}{4} w^{-3/2} \cdot (2.75 p_{4b}) > 0$$

because $w(x)$ and $p_{4b}(x)$ are positive. Therefore, $\mathcal{L}[w]$ is concave up at $w(x) = p_4 b(x)/p_{2b}(x)$ and this w is in fact a minimum.

Appendix C

Further statistics fundamentals

Appendix D

Gaussian Processes

Appendix E

Further statistics details

E.1 Asymptotics approximation

To emphasize some of the fundamentals of the particle physics statistics that we use, in this section I show my own reproduction of one of the experiments in the asymptotic approximation paper need to cite.

Consider a cut-and-count analysis where the signal region (SR) has s signal events, and the background yield is specified by a nuisance parameter (NP) b expected background with b events. Let the number of events that we observe in the SR be n, so then the expected value for the observed event yield is: $\mathbb{E}[n] = \mu s + b$. We additionally consider that we have a measurement in a control region where m events are observed, and this control region where the background yield is scaled by τ relative to the SR, i.e, $\mathbb{E}[m] = \tau b$.

Since this is now just a single bin counting experiment, the likelihood from Eq 7.1 simplifies to the form:

$$\mathcal{L}(\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-\mu s + b} \frac{(\mu s + b)^n}{n!} e^{-\mu s + b} \tag{E.1}$$

In this simple example, we can compare the analytic asymptotic formula for the test statistic's distribution to the result from throwing toys to build our intuition for the problem.

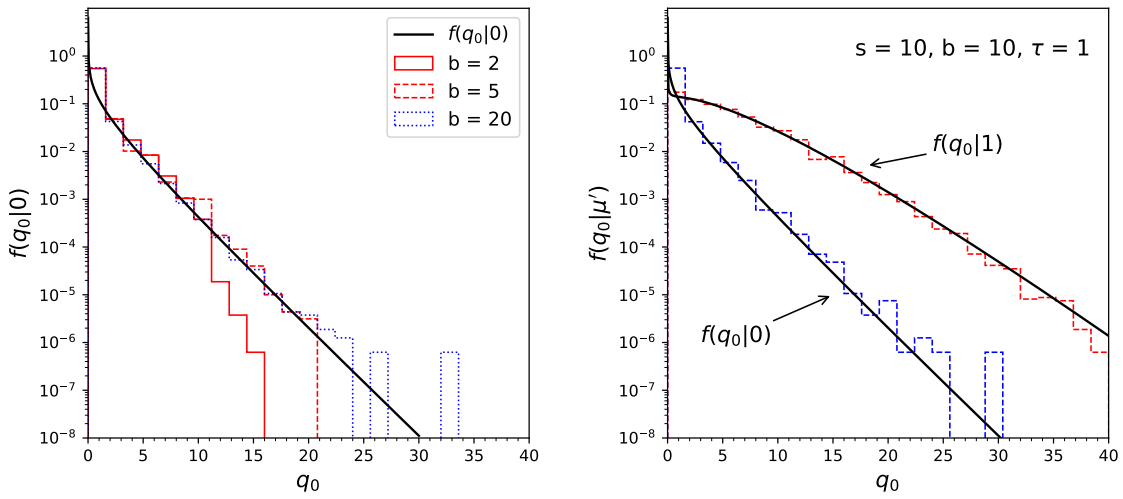


Figure E.1

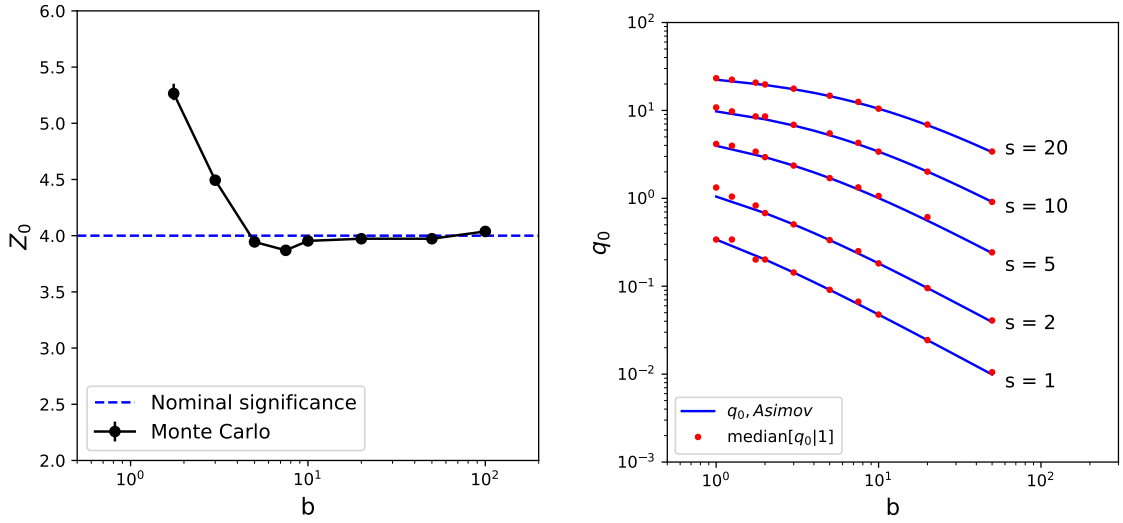


Figure E.2

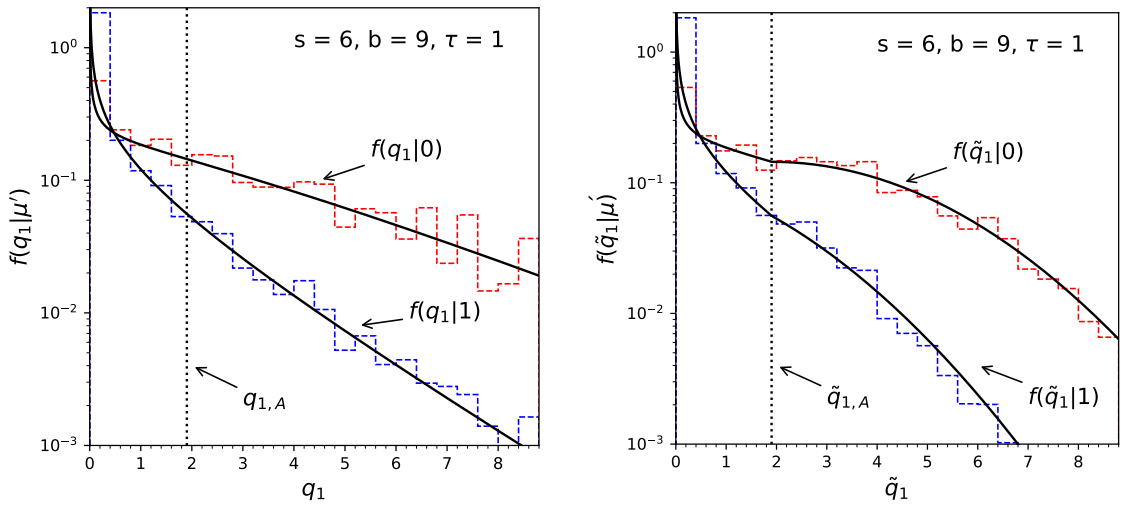


Figure E.3

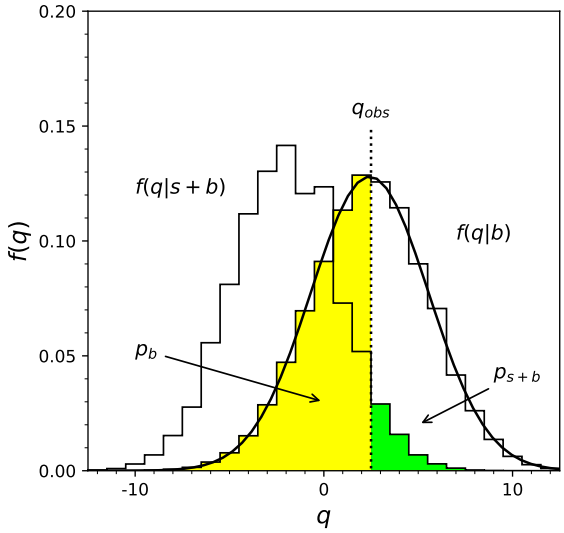


Figure E.4

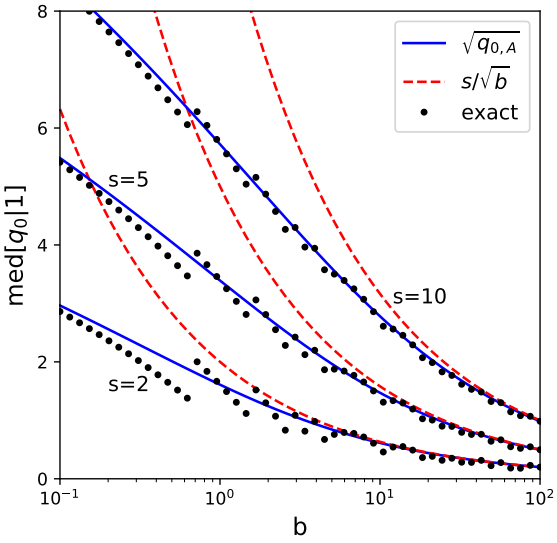


Figure E.5

E.2 Choice of CLs test statistic

E.2.1 A pedagogical example

There can be a problem for the frequentist interpretation if the number of observed events is less than the number of predicted events (in our notation, $n_i < b_i$). As an example of this, consider a Poisson experiment where we expect $b = 8$ background events, but only observe $n = 3$ events (from [**eshep-statsIII**]. Considering the model where $\propto = s + b$, we can calculate the 95% upper limit on the true expected number of events by summing up the poisson probabilities of observing this many events or fewer (which is what is also defined as CL_{s+b}):

$$Prob(\text{observing} \leq 3\text{events}) = CL_{s+b} = \sum_{r=0}^{3} \frac{n^r e^{-n}}{r!}. \tag{E.2}$$

Figure E.6 (blue line) shows the result of CL_{s+b} as we scan over n values. By settig $CL_{s+b} = 0.05$, we find $n = 7.75$. Then solving for an upper limit on the signal yield, we get s, we get $s = n - b = n - 8 = -0.25$, an unphysical result.

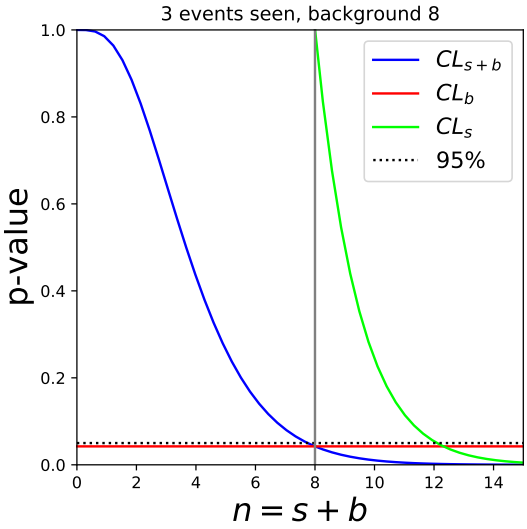


Figure E.6: A comparison of the ...

To avoid the pathologies of the has been a convention that's been used since the Tevatron need citation, and is the convention adopted across the HH physics analyses to use this as how we interpret a 95% upper limit - it corresponds to $CL_s = 0.05$.

You can think of this although this is does mean that we now have a conservative upper limit because

- CL_{s+b}: Probability that we saw this number of events (or fewer) given an expected yield of s+b.
- CL_b: Same as CL_s,but with $b = 0$.

$$CL_s = \frac{CL_{s+b}}{CL_b} \tag{E.3}$$

E.2.2 Intuition building - impact on the 4b analysis

Finally, to illustrate why this was a more conservative choice,

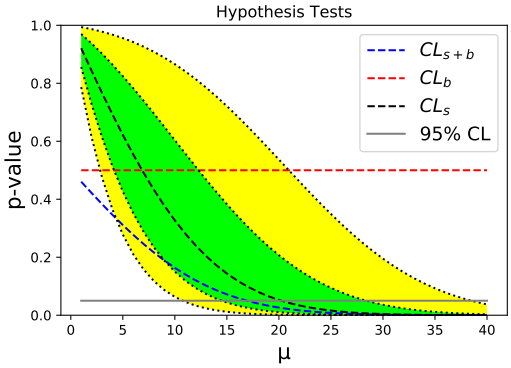


Figure E.7: Impact of using the CL_s verses the CL_{s+b} test statistic.

Appendix F

ML for jet → parton assignment

F.1 Motivation

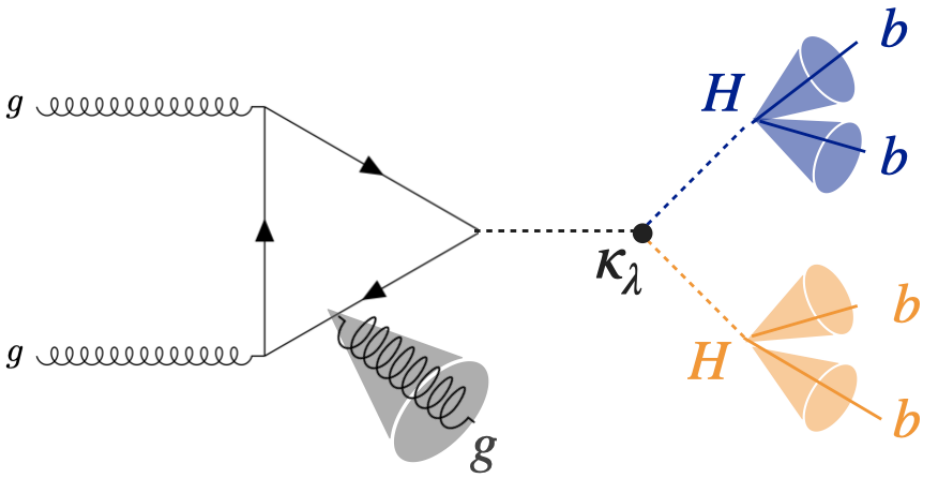


Figure F.1: Illustration of the task for Jet selection and pairing.

I think it would be really nice to include some of the pairing accuracy numbers that were motivating the pairAGraph algorithm

- These studies used the SR definition from the previous analysis
- The signal samples were trained and evaluated using Herwig (instead of Pythia which was later adopted as the nominal signal sample for all of the HH analyses).

F.2 ML based solution

F.2.1 Graph partitioning problem

A graph neural network (GNN) is a deep learning paradigm that has gained a lot of popularity both in the deep learning and particle physics community in recent years [**I should find a few references for this**].

To cast the 4b jet to parton assignment problem in the GNN formalism, consider and pp collision event as a graph where jets are the nodes of the graph, and a set of weighted connections.

The training objective becomes to maximize the edge weights between the jets that come from the same HC, as visualized in Figure F.2.

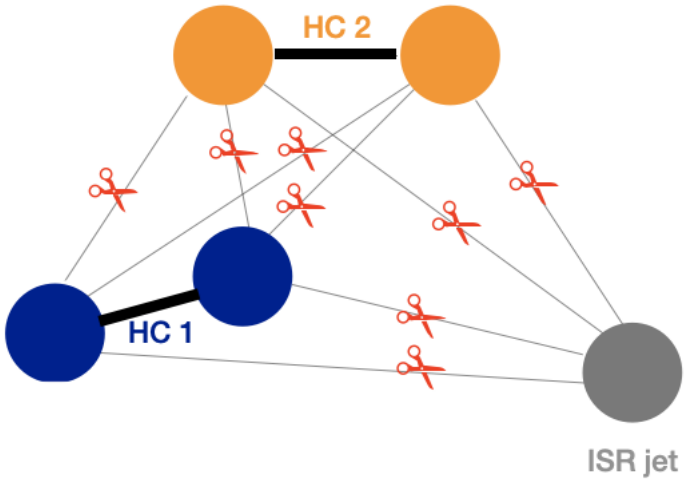


Figure F.2

Describe the space by a jet-similarity metric with similarity matrix W:

$$W_{ij} = \vec{j}_i \cdot \vec{j}_j \tag{F.1}$$

Then for each event we can use this similarity matrix to define a *score* for each possible pairing of jets into HCs.

Score: sum of similarity weights for pairing

- HC1 formed from $\vec{j}_i$, $\vec{j}_j$
- HC2 formed from $\vec{j}_k$, $\vec{j}_l$

$$s = \vec{j}_i \cdot \vec{j}_j + \vec{j}_k \cdot \vec{j}_l \tag{F.2}$$

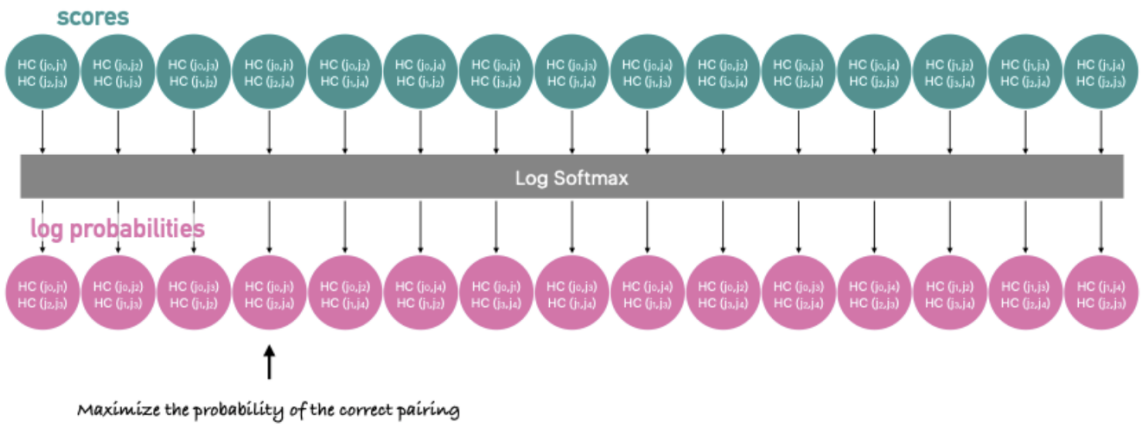


Figure F.3: The loss function for a pairAGraph training event with 5 input jets.

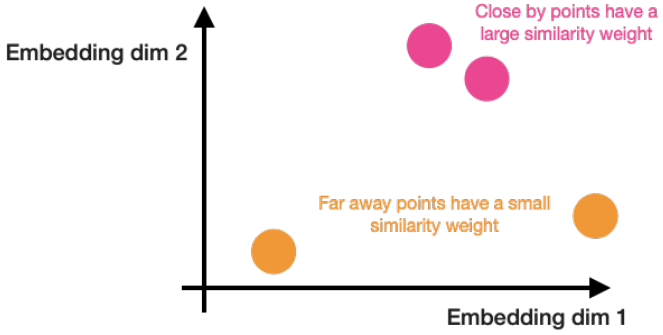


Figure F.4: The jet embedding space.

F.2.2 Transformers

The transformer architecture is now the state of the art in the natural language processing field for reasoning about the variable length sentences that we have today for large language models need GPT3 citation. The model was originally proposed for the task of neural machine translation [**1706.03762**], and used an encoder decoder architecture, as shown in Figure fig:xformer-fig1.

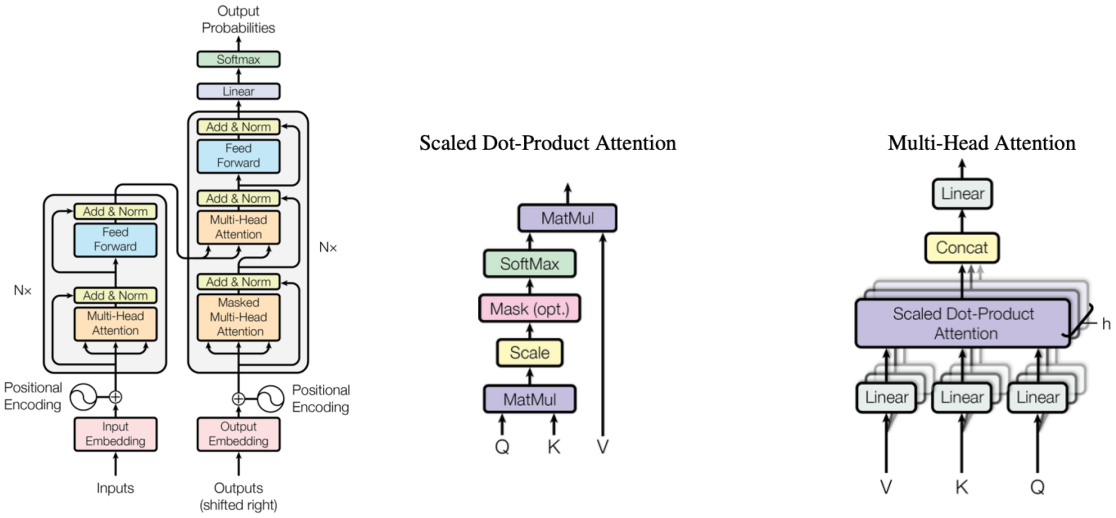


Figure F.5: The transformer architecture (left) and its building blocks: the weighted sum operation (middle) and the more expressive multi-head attention block formed by adding additional channels for the scaled dot product attention blocks (right) T [**1706.03762**].

Set transformer [**1810.00825**].

F.2.3 The 4b implementation: pairAGraph

Compared to the difficulty of the types of problems transformers are usually used for, this was quite a simple problem. We used a small transformer model with only one of these "multi-head attention" blocks. The jet latent dimension was 20, and 20 hidden units were used for computing the query and key weighted attention matrices. Four attention heads were used, and a dropout of rate of 0. 3 was applied to the nodes after the feed forward neural network layer. The loss was minimized with adam [42] using a learning rate of 0.005 was a batch size of 2048 events. A validation set (20% of the training events) were reserved for the training when the loss on this validation set hadn't improve in the last 20 epochs, and we took the model weights that had the best performance on the training dataset..

The training was done on events that passed the trigger and with at least 4 jets with p_{T} ¿ 40 GeVand two b-tags, and we combined events from the simulation corresponding to the 2016, 2017, and 2018 datasets.

We used 2-fold cross validation training two transformers for each physics sample splitting the sample based on the event number. To evaluate the model with the even event numbers, we used the training that had been performed on the odd event numbers so that we could make full use of the final event statistics.

F.2.4 Other baselines we compare to

For my studies comparing the HC pairing accuracy, I compared to both the min dR algorithm (which was adopted in the 4b full Run 2 non-resonant analysis), and also to the baseline "MDR+min(D_{hh})" pairing.

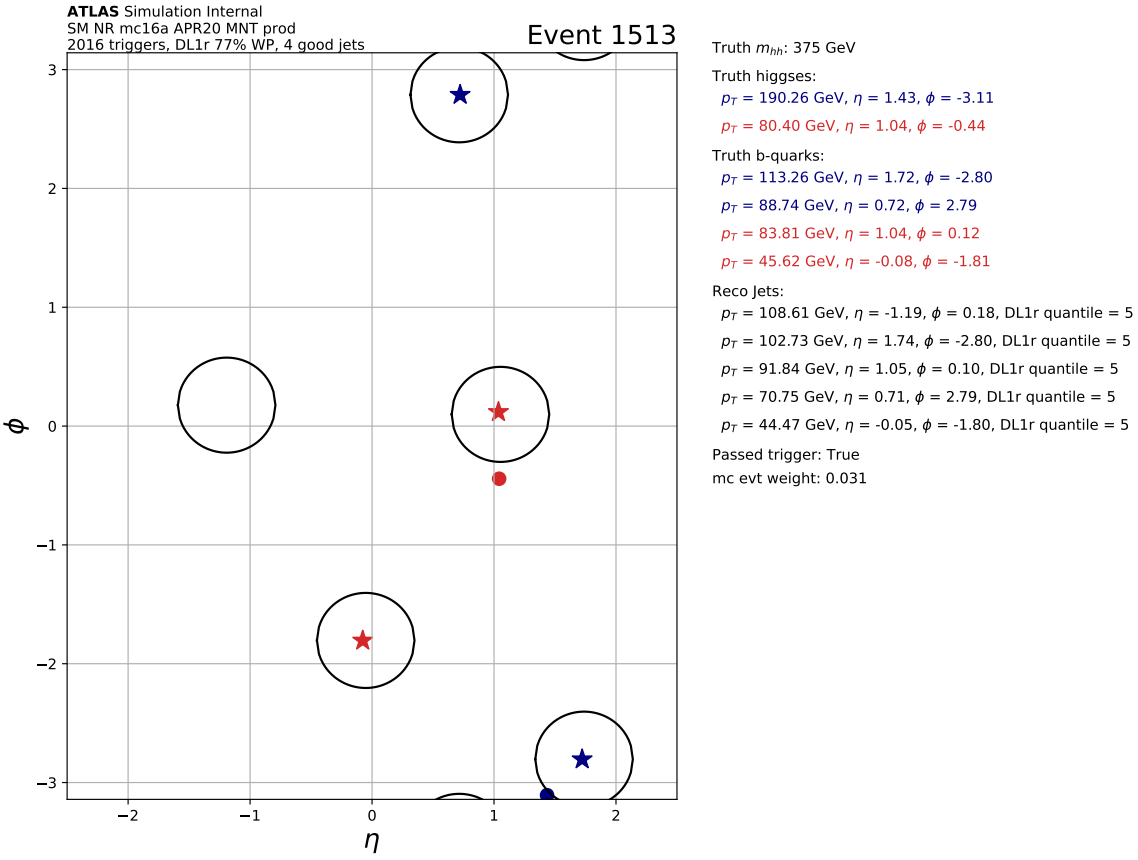
Briefly describe the MDR + min Dhh pairing

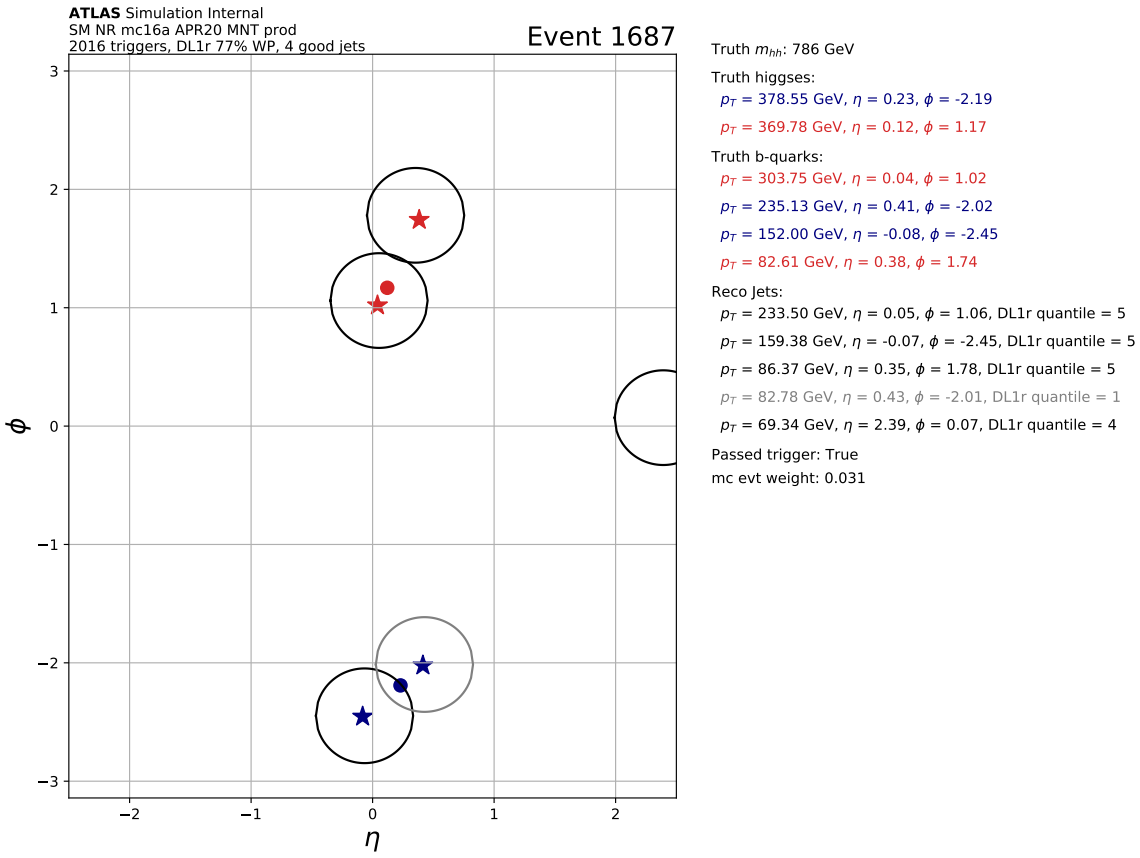
BDT pairing(?)

F.3 Impact on the signal

F.3.1 Jet selection accuracy

F.3.2 Cases where pairAGraph got the correct jets and the baseline selected the wrong jets


ATLAS Simulation Internal
SM NR mc16a APR20 MNT prod
2016 triggers, DL1r 77% WP, 4 good jets
Event 1513
Truth m_{hh}: 375 GeV
Truth higgses:
p_T = 190.26 GeV, η = 1.43, ϕ = -3.11
p_T = 80.40 GeV, η = 1.04, ϕ = -0.44
Truth b-quarks:
p_T = 113.26 GeV, η = 1.72, ϕ = -2.80
p_T = 88.74 GeV, η = 0.72, ϕ = 2.79
p_T = 83.81 GeV, η = 1.04, ϕ = 0.12
p_T = 45.62 GeV, η = -0.08, ϕ = -1.81
Reco Jets:
p_T = 108.61 GeV, η = -1.19, ϕ = 0.18, DL1r quantile = 5
p_T = 102.73 GeV, η = 1.74, ϕ = -2.80, DL1r quantile = 5
p_T = 91.84 GeV, η = 1.05, ϕ = 0.10, DL1r quantile = 5
p_T = 70.75 GeV, η = 0.71, ϕ = 2.79, DL1r quantile = 5
p_T = 44.47 GeV, η = -0.05, ϕ = -1.80, DL1r quantile = 5
Passed trigger: True
mc evt weight: 0.031
ϕ
η


ATLAS Simulation Internal
SM NR mc16a APR20 MNT prod
2016 triggers, DL1r 77% WP, 4 good jets
Event 1687
Truth m_{hh}: 786 GeV
Truth higgses:
p_T = 378.55 GeV, η = 0.23, ϕ = -2.19
p_T = 369.78 GeV, η = 0.12, ϕ = 1.17
Truth b-quarks:
p_T = 303.75 GeV, η = 0.04, ϕ = 1.02
p_T = 235.13 GeV, η = 0.41, ϕ = -2.02
p_T = 152.00 GeV, η = -0.08, ϕ = -2.45
p_T = 82.61 GeV, η = 0.38, ϕ = 1.74
Reco Jets:
p_T = 233.50 GeV, η = 0.05, ϕ = 1.06, DL1r quantile = 5
p_T = 159.38 GeV, η = -0.07, ϕ = -2.45, DL1r quantile = 5
p_T = 86.37 GeV, η = 0.35, ϕ = 1.78, DL1r quantile = 5
p_T = 82.78 GeV, η = 0.43, ϕ = -2.01, DL1r quantile = 1
p_T = 69.34 GeV, η = 2.39, ϕ = 0.07, DL1r quantile = 4
Passed trigger: True
mc evt weight: 0.031
ϕ
η

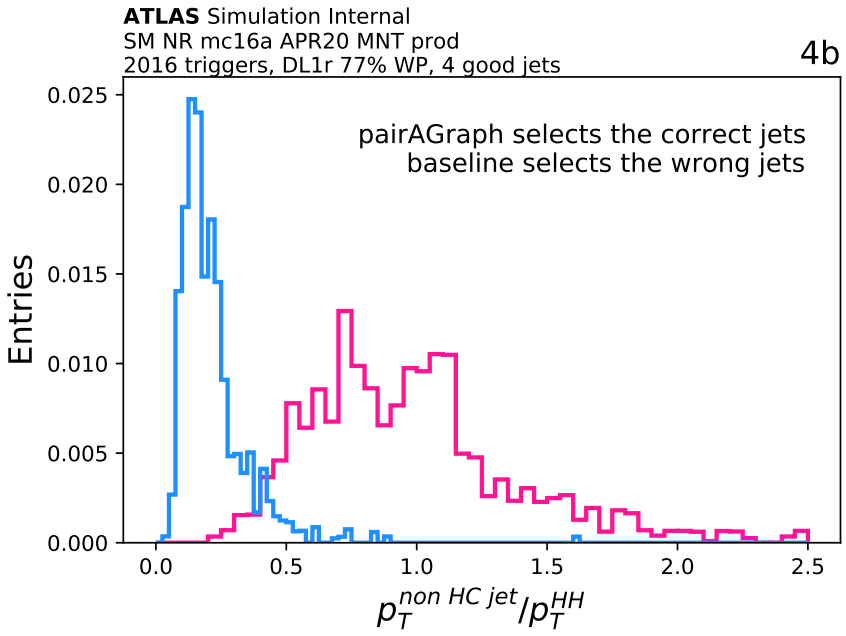


Figure F.6: need to add a legend!!

F.3.3 What extra information was pairAGraph learning?

RECOIL (need plot)

Oh - also I'll need to decide if I just want to show 4b or the 3b cats too!

F.3.4 Pairing accuracy

Pairing algorithm	**Pairing**	**MDR**	**MDpT**	$\Delta\eta_{HH} < 1.5$	$X_{Wt} > 1.5$	**SR**
MDR + min(Dhh)	71.8%	79.7%	79.7%	80.1 %	83.3 %	93.6 %
$min\left(\Delta R_{jj}^{HC\ 1}\right)$	69.7 ˆ	–	73.7 %	74.0 %	78.4%	94.7%
pairAGraph train SM	78.4 %	–	–	79.6 %	82.4 %	94.2%
pairAGraph train $\kappa_\lambda = 10$	76.8%	–	–	78.4 %	81.3 %	94.1 %

Table F.1: Cuts are applied sequentially from the left to the right.

Observation # 1: The "after pairing" column

Observartion # 2: Applying the cuts sequentially

Observation # 3: Inside of the SR

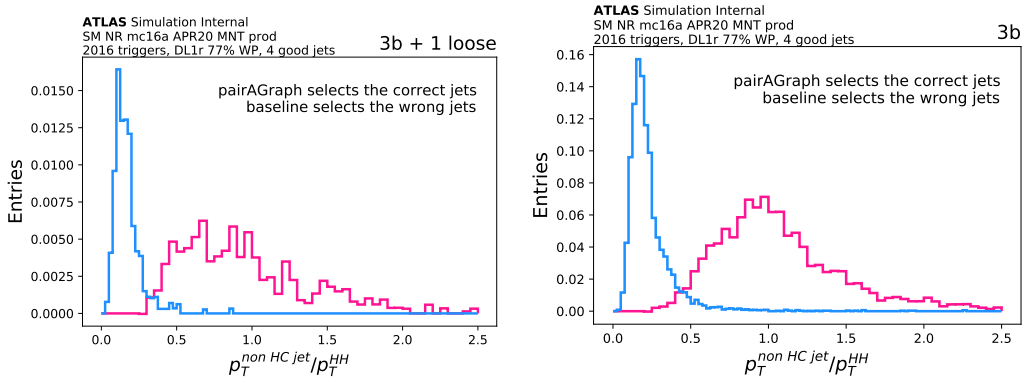


Figure F.7: need to add a legend!!

F.3.5 Visualization of the attention weights

One of the attractive features of using a transformer architecture was that the multi-head attention mechanism in the

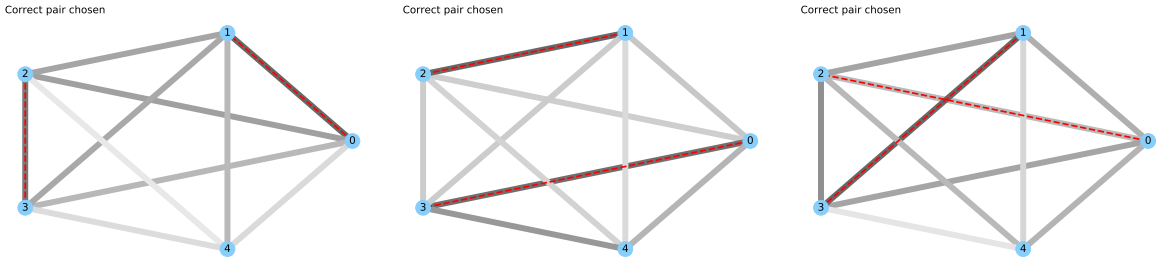


Figure F.8: Visualization of the multi-head attention weights from the transformer. The circles on the graph

F.4 Impact on the background

F.4.1 Impact on the massplanes

F.4.2 Impact on the limits

F.4.3 Impact on the limits with systematics

F.5 Related work and future prospects

People using transformers for other tasks? Top tagging (?), ML4Pion (Muriel's talk).

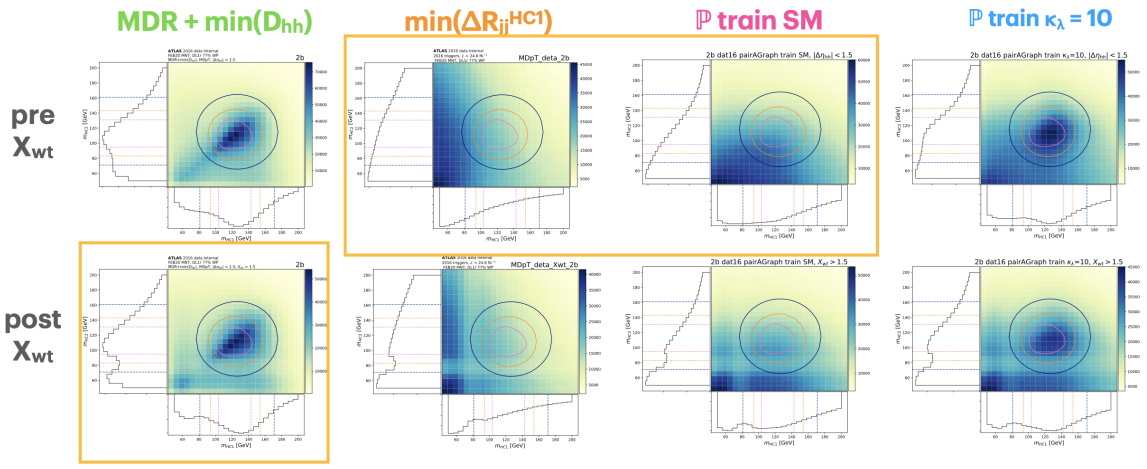


Figure F.9: Massplanes on the 2b data comparing several of the pairing algorithms. The gold line shows the ... where we will derive the reweighting for the limits in ... need to double check which limits I want to show here!

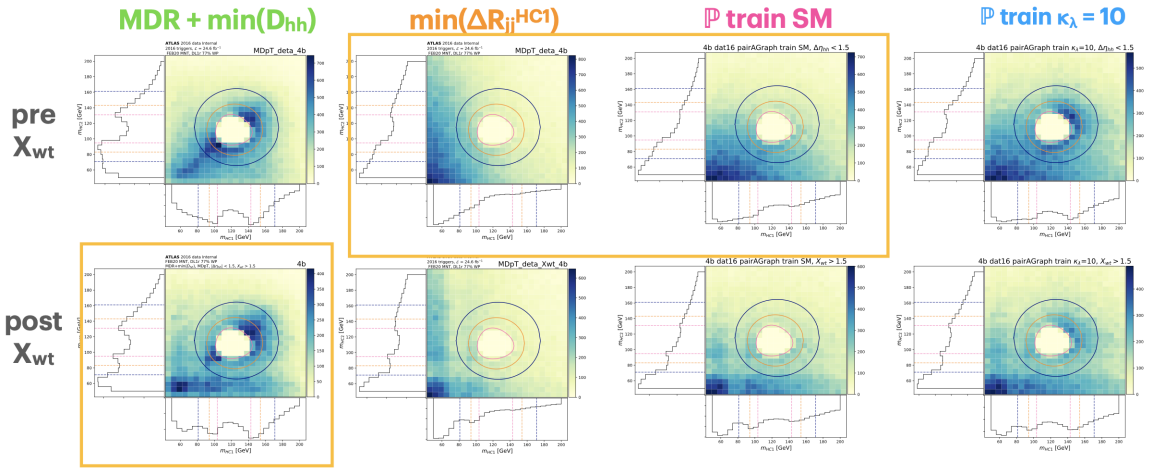


Figure F.10: Blinded massplanes on the 2b data comparing several of the pairing algorithms. The gold line shows the ... where we will derive the reweighting for the limits in ... need to double check which limits I want to show here!

If I had more time, what would I change?

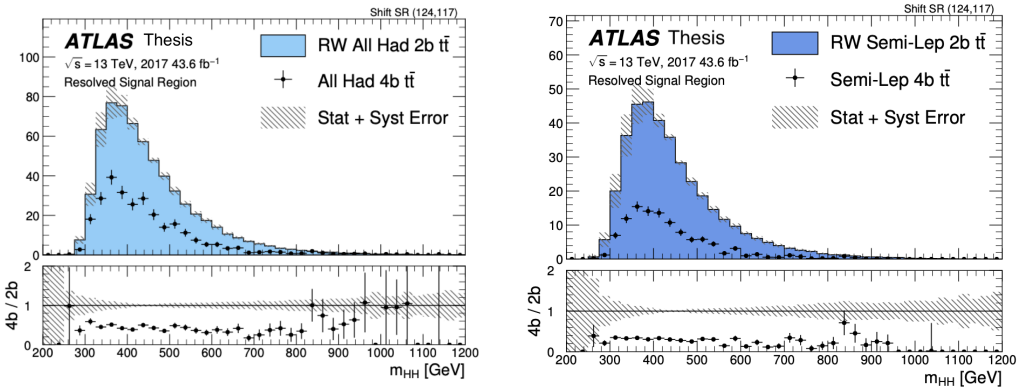
- I really think it would be fun to make pag Lorentz equivariant
- I think maybe what I was missing is unweighting the mHH (or pThh) spectra.

Appendix G

$t\bar{t}$ aware reweighting

G.1 Motivation

As shown in Table 10.1, $t\bar{t}$ is approximately 10% of our background. The background estimate approach inclusively reweights both the QCD and $t\bar{t}$ portions, but we can use the $t\bar{t}$ simulation to check the performance of the data trained reweighting by evaluating it on the $2b$ $t\bar{t}$ simulation and checking the performance compared to $4b$ $t\bar{t}$. Figure G.1 shows this result for the m_{HH} modeling, and the inclusively trained background estimate over-predicts all-had $t\bar{t}$ by **2.28** and semi-leptonic $t\bar{t}$ by **3.4** since the proportion of true $2b$ events is higher in $t\bar{t}$.



(a) all-hadronic $t\bar{t}$

(b) semi-leptonic $t\bar{t}$

Figure G.1: Performance of the inclusively trained reweighting evaluated on the $t\bar{t}$ simulation. The performance of the inclusively trained reweighting is evaluated on $2b$ $t\bar{t}$ simulation and compared to the $4b$ $t\bar{t}$ prediction. The error bar on the background prediction shows the quadrature sum of the 2b Poisson, deep ensembles, and CR1 / CR2 shape systematic error.

- Will decrease the error for negatively correlated variables
- Layout for this chapter

Another reason to Suppose that we have two random variables, X and Y, and let $Z = X + Y$ $\text{Cov}[Z] = \text{Cov}[X] + \text{Cov}[Y] + 2\text{Cov}[X, Y]$

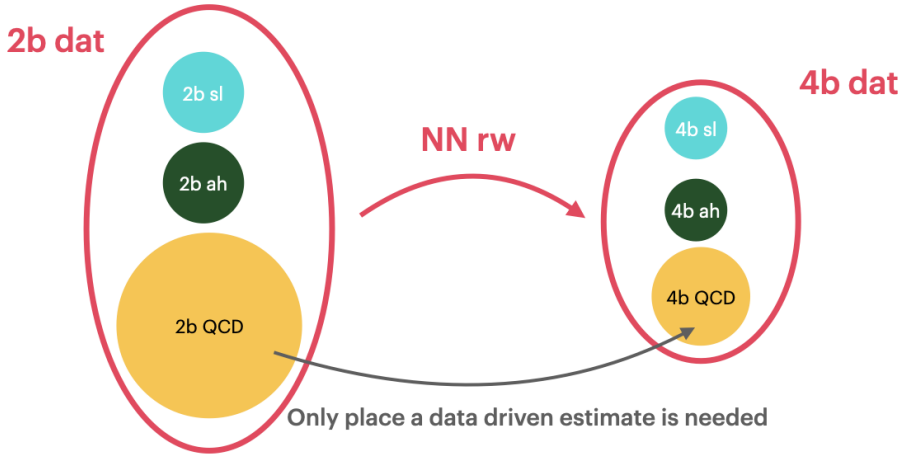


Figure G.2: Illustration of what how inclusively reweighting is doing to the separate components of the background estimate. The only background component that we need to use a data driven technique for is the QCD piece where we don't trust our simulation.

G.2 Fitting $t\bar{t}$ templates

G.2.1 Prescription

Follows closely from the 36 fb^{-1}analysis [**4b-36ifb**].

First we subtract off the 2b $t\bar{t}$ from the data to get the $t\bar{t}$ piece.

$$QCD = NN(2bdat) - NN(2ball - hadt\bar{t}) - \alpha^{sl}_{t\bar{t},2b} \cdot NN(2bsemi - lept\bar{t}) \tag{G.1}$$

$$QCD =$$

Since we will fit the X_{Wt} histogram to extract the template normalizations, here the networks are trained in CR1 *before* applying the $X_{Wt} > 1.5$ cut.

Then the next step is to do a combined fit for 4b considering top sensitive control regions:

- 4b data that includes and isolated μ (helps pin-down the semi-lep $t\bar{t}$ component)

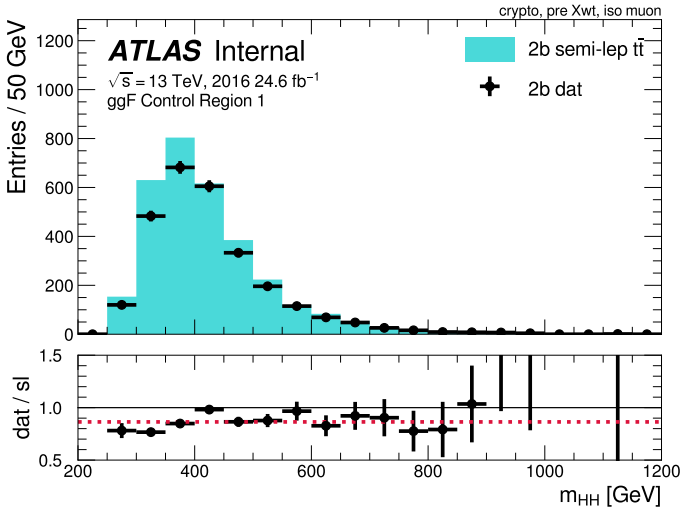


Figure G.3: m_{HH} in $2b$ sample with an isolated muon for 2016 data in CR1. The dashed red line in the subpanel shows the fitted $\alpha^{sl}_{t\bar{t},2b}$

SF	16	17	18
CR1	0.86	0.80	0.83
CR2	0.82	0.81	0.79

Table G.1: Fitted SFs for $\alpha^{sl}_{t\bar{t},2b}$.

- An X_{Wt} shape fit (helps discriminate between the all-had $t\bar{t}$ and qcd templates).

G.2.2 Fit results 4b

Next, we look at a few variables after the $X_{Wt} > 1.5$ cut.

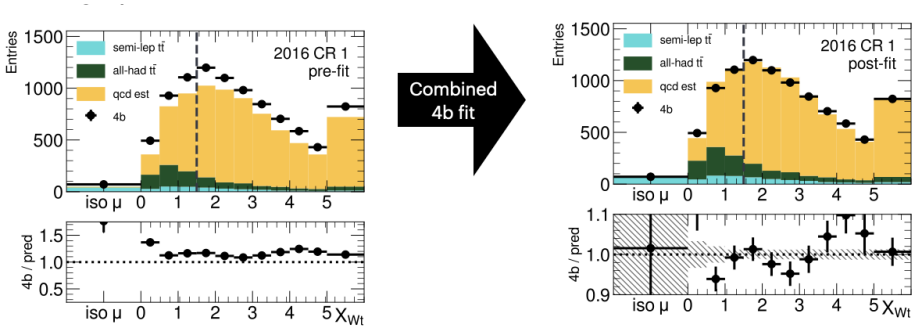


Figure G.4: CR1 fits

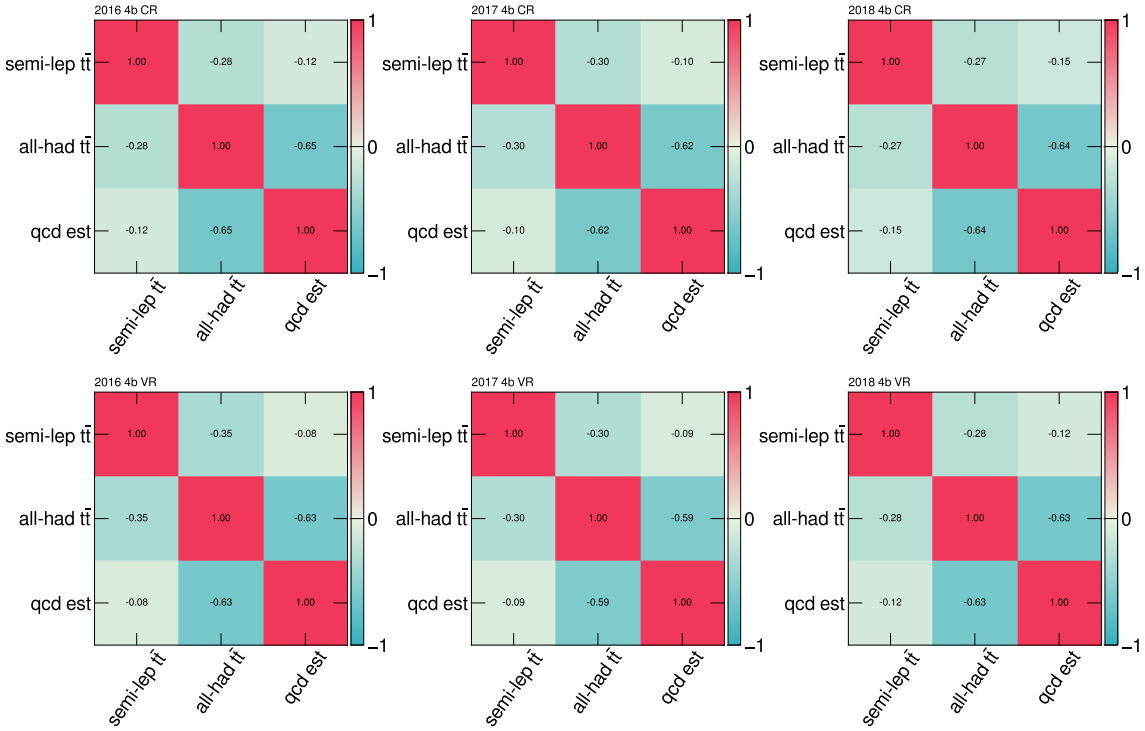


Figure G.5: CR1 fits

	2016			2017			2018		
	semi-lep	all-had	QCD	semi-lep	all-had	QCD	semi-lep	all-had	QCD
CR 1 SF	1.640 ± 0.176	1.322 ± 0.131	1.113 ± 0.020	1.362 ± 0.139	1.740 ± 0.114	1.096 ± 0.016	1.075 ± 0.115	1.673 ± 0.088	1.082 ± 0.013
CR 2 SF	1.841 ± 0.200	1.456 ± 0.135	1.106 ± 0.020	1.540 ± 0.141	1.806 ± 0.110	1.103 ± 0.016	1.232 ± 0.110	1.665 ± 0.083	1.082 ± 0.013
extrap norm CR 2 SF / CR1 SF	1.122	1.101	0.9933	1.130	1.0379	1.0062	1.147	0.995	1.00044
naive extrap norm CR 2 pred / CR1 pred	0.992			0.996			0.968		

We do not include a norm extrapolation NP in the fits.
[Would cause a 50% degradation in limits.]

Sub % level uncertainty for the QCD norm!!
Small c.f. the other unc we assess in the analysis [see table].

Figure G.6

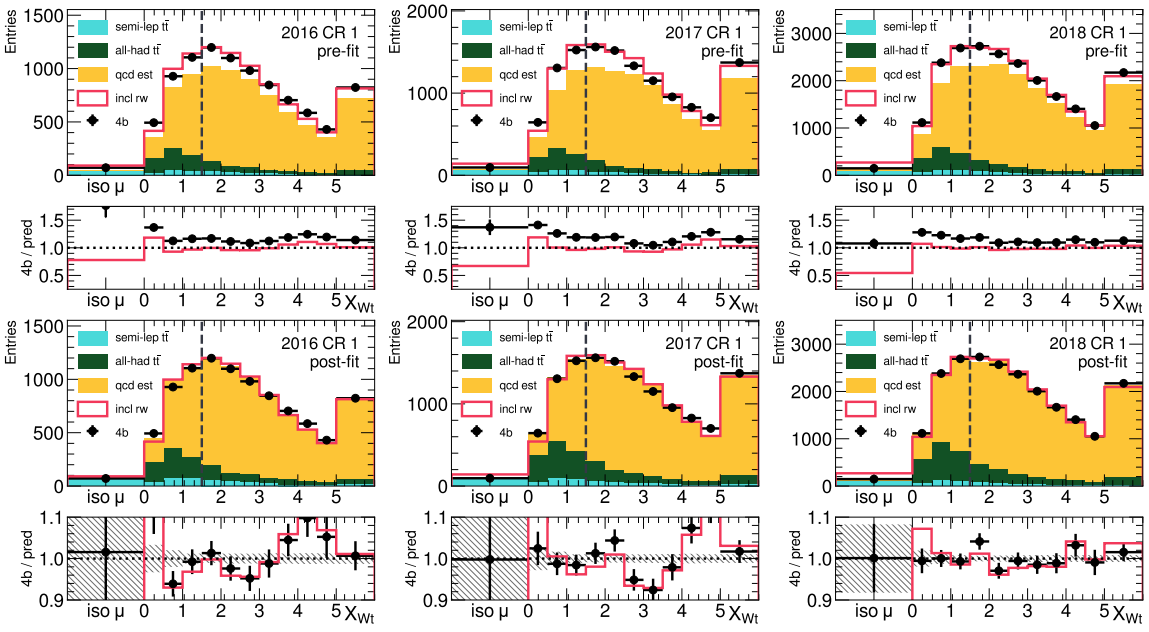
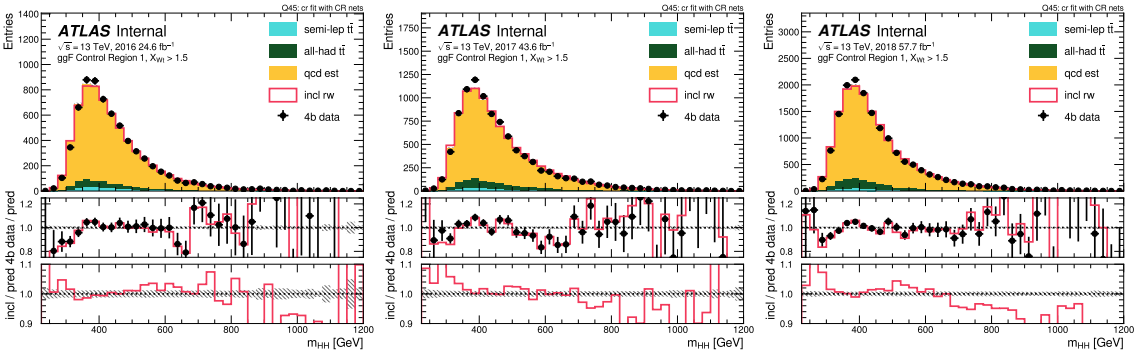
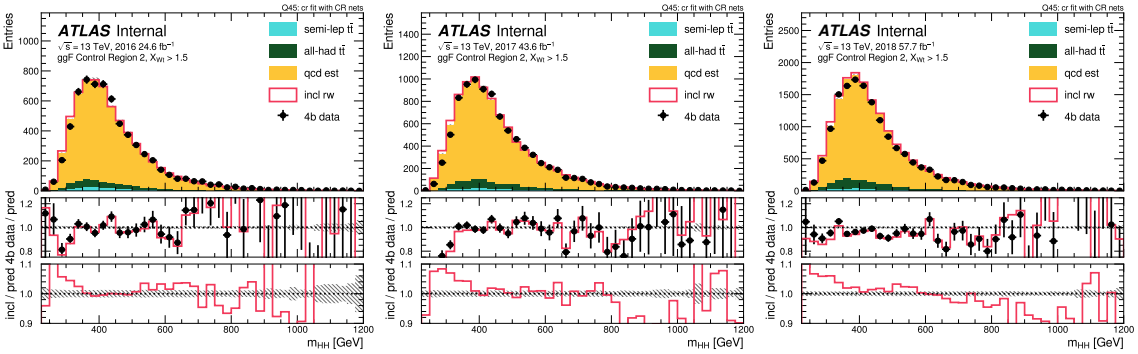


Figure G.7: CR1 fit in CR1



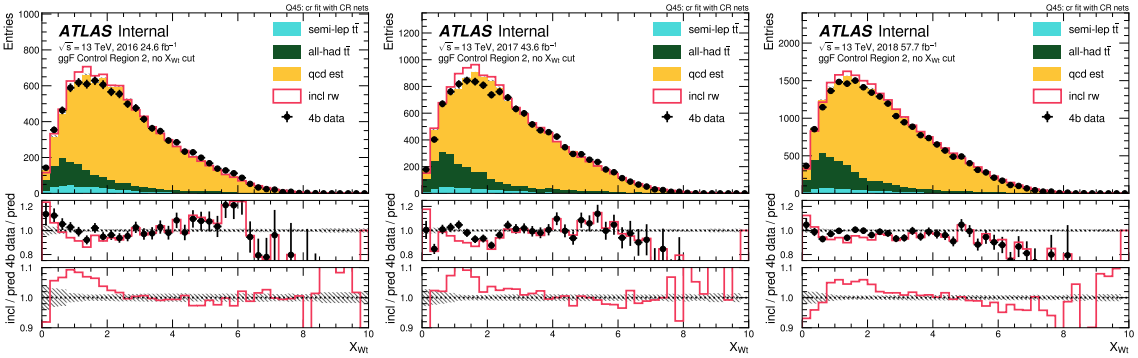
(a) 2016 (b) 2017 (c) 2018

Figure G.8: CR1 evaluation using the CR1 fits



(a) 2016 (b) 2017 (c) 2018

Figure G.9: CR2 evaluation using the CR1 fits



(a) 2016 (b) 2017 (c) 2018

Figure G.10: CR2 evaluation using the CR1 fits

G.2.3 Fit results 3b1f

	2016			2017			2018		
	semi-lep	all-had	QCD	semi-lep	all-had	QCD	semi-lep	all-had	QCD
CR SF	0.995 ± 0.036	1.000 ± 0.033	1.003 ± 0.006	0.916 ± 0.033	1.740 ± 0.114	1.002 ± 0.006	0.929 ± 0.023	1.044 ± 0.022	0.993 ± 0.004
VR SF	0.955 ± 0.034	1.017 ± 0.030	1.004 ± 0.006	0.930 ± 0.032	1.806 ± 0.110	1.009 ± 0.006	0.906 ± 0.022	1.044 ± 0.020	0.994 ± 0.004
extrap norm	0.960	1.018	1.0013	1.015	0.971	1.0069	0.975	0.995	1.0014
extrap norm naive	0.970			0.963			0.971		

Figure G.11

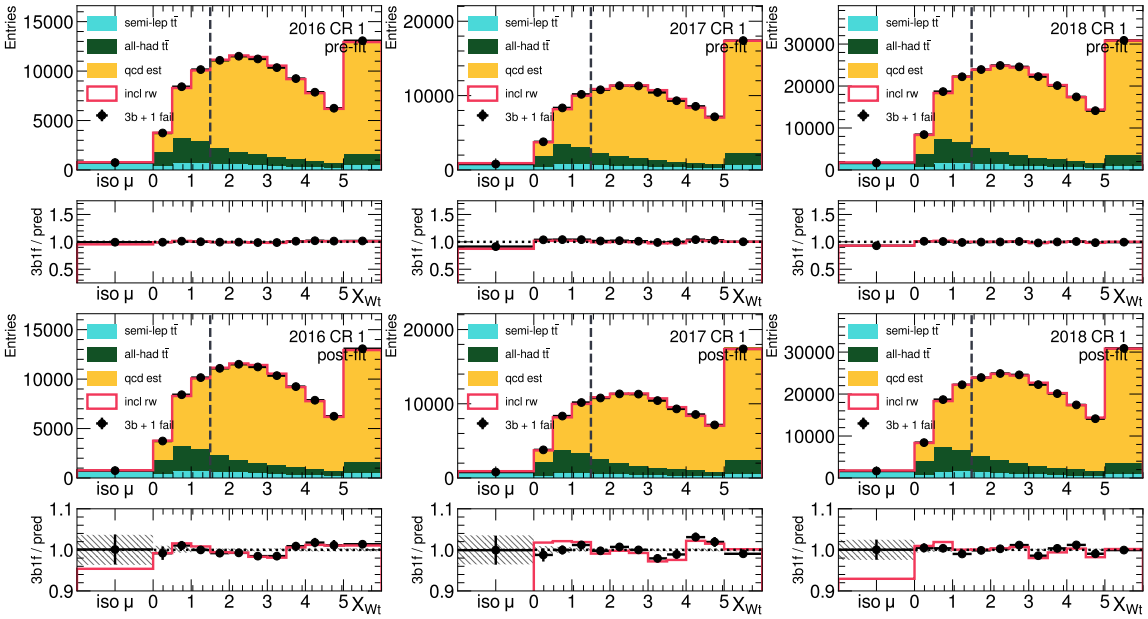
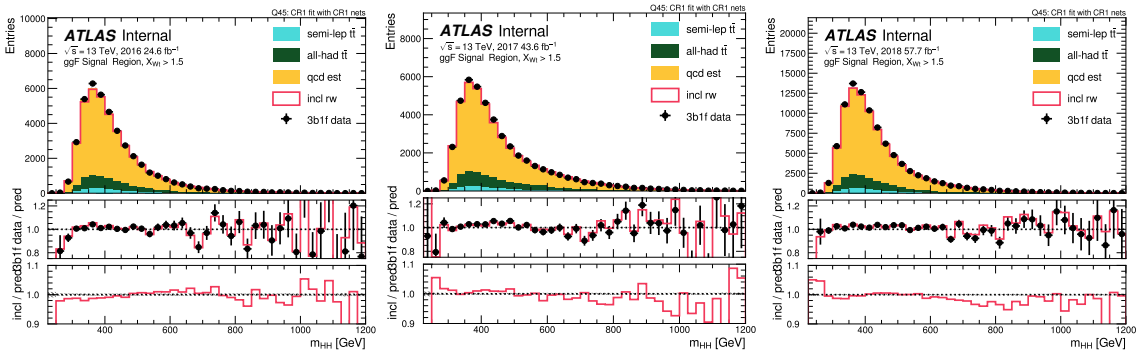
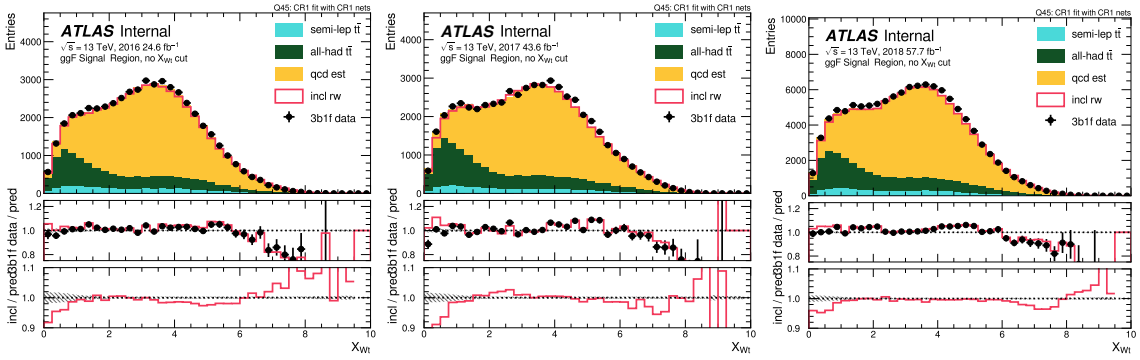


Figure G.12: CR1 fit in CR1



(a) 2016 (b) 2017 (c) 2018

Figure G.13: 3b1f SR evaluation using the CR1 fits



(a) 2016 (b) 2017 (c) 2018

Figure G.14: 3b1f SR evaluation using the CR1 fits

G.3 Pure QCD reweighting

We were a little bit confused why the template fits weren't working as well - and one hypothesis was that we were using the inclusively trained estimate to get the $t\bar{t}$ template. In the next section, we show how to modify the loss function we used to get a pure QCD reweighting.

G.3.1 Mathematical formulation

Rafael Teixeira de Lima developed this formalism, and this write-up closely follows his notes [**pure-qcd-Rafael**].

Let $R_{2b^{all}\to 4b^{all}}$ describe the inclusively trained reweighting. As demonstrated in Figure G.2, this has two

$$R_{2b^{all}\to 4b^{all}} = \frac{\alpha^{t\bar{t}}_{4b} p^{t\bar{t}}_{4b} + \alpha^{Q}_{4b} p^{Q}_{4b}}{\alpha^{t\bar{t}}_{2b} p^{t\bar{t}}_{2b} + \alpha^{Q}_{2b} p^{Q}_{2b}} \tag{G.2}$$

where p denotes the probability densities and α the corresponding normalizations. The subscripts $2b$ and $4b$ indicate the b-tagging requirement and the superscripts indicate the $t\bar{t}$ and QCD (abbreviated "Q") physics samples. Simplify the notation by letting $P^y_x = \alpha^y_x p^y_x$ so that Eq. G.2 becomes:

$$\begin{aligned} R_{2b^{all}\to 4b^{all}} &= \frac{P^{t\bar{t}}_{4b} + P^{Q}_{4b}}{P^{t\bar{t}}_{2b} + P^{Q}_{2b}} \\ &= \frac{P^{t\bar{t}}_{4b}}{P^{t\bar{t}}_{2b} + P^{Q}_{2b}} + \frac{P^{Q}_{4b}}{P^{t\bar{t}}_{2b} + P^{Q}_{2b}} \end{aligned}$$

We want to isolate the individual reweightings that we can derive:

$$\begin{aligned} R_{2b^{all}\to 4b^{all}} &= \frac{P^{t\bar{t}}_{4b}}{P^{t\bar{t}}_{2b}} \cdot \frac{1}{1 + \frac{P^{Q}_{2b}}{P^{t\bar{t}}_{2b}}} + \frac{P^{Q}_{4b}}{P^{Q}_{2b}} \cdot \frac{1}{1 + \frac{P^{t\bar{t}}_{2b}}{P^{Q}_{2b}}} \\ &= \frac{P^{t\bar{t}}_{4b}}{P^{t\bar{t}}_{2b}} \cdot \frac{P^{t\bar{t}}_{2b}}{P^{t\bar{t}}_{2b} + P^{Q}_{2b}} + \frac{P^{Q}_{4b}}{P^{Q}_{2b}} \cdot \frac{P^{Q}_{2b}}{P^{Q}_{2b} + P^{t\bar{t}}_{2b}} \\ &= \frac{P^{t\bar{t}}_{4b}}{P^{t\bar{t}}_{2b}} \cdot \frac{P^{t\bar{t}}_{2b}}{P^{t\bar{t}}_{2b} + P^{Q}_{2b}} + \frac{P^{Q}_{4b}}{P^{Q}_{2b}} \cdot \left(1 - \frac{P^{t\bar{t}}_{2b}}{P^{Q}_{2b} + P^{t\bar{t}}_{2b}}\right) \end{aligned} \tag{G.3}$$

We can now identify:

- $R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}} = \frac{P^{t\bar{t}}_{4b}}{P^{t\bar{t}}_{2b}}$: A MC based $t\bar{t}$ reweighting $2b \to 4b$.
- $R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}} = \frac{P^{t\bar{t}}_{2b}}{P^{Q}_{2b}+P^{t\bar{t}}_{2b}}$: Reweight the $2b$ data into the $2b$ $t\bar{t}$.
- $R_{2b^{Q}\to 4b^{Q}}$: a pure QCD reweighting – what we want to solve for.

The first two reweightings in the list above are ones that we can easily derive. $R_{2b^{all}\to 4b^{all}}$ is the current background estimate, so we can move forward to solve for $R_{2b^{Q}\to 4b^{Q}}$. First, substitute these reweighting definitions into Eq. G.3:

$$R_{2b^{all}\to 4b^{all}} = R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}} \cdot R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}} + R_{2b^{Q}\to 4b^{Q}} \left(1 - R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}\right). \tag{G.4}$$

Rearrange to solve for $R_{2b^{Q}\to 4b^{Q}}$:

$$R_{2b^{Q}\to 4b^{Q}} = \frac{R_{2b^{all}\to 4b^{all}} - R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}} \cdot R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}}{1 - R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}} \tag{G.5}$$

Now that we have a pure QCD reweighting, this gives a prescription for how to apply it to give a prediction for the 4b data:

$$P_{4b}^{all} = R_{2b^{Q}\to 4b^{Q}} \cdot \left(P_{2b}^{all} - P_{2b}^{t\bar{t}}\right) + P_{4b}^{t\bar{t}}. \tag{G.6}$$

The first term $R_{2b^{Q}\to 4b^{Q}} \cdot \left(P_{2b}^{all} - P_{2b}^{t\bar{t}}\right)$ gives the piece for the QCD template, where we use the weights from Eq. G.5. To get just the QCD template, we apply this reweighting formula to the $2b$ data and then subtract off the result of applying the reweighting to the $2b$ $t\bar{t}$ MC. The second term $P_{4b}^{t\bar{t}}$ gives the $t\bar{t}$ template, which we can just get using the $t\bar{t}$ MC.

G.3.2 Experiments

The following reweighting estimates just show a single training (instead of the whole machinery with 100 bootstraps) to get a sense for the scale of the problem and how much this method could help us. We'll want to fit the X_{Wt} template again, so the reweightings are derived in CR1 before the X_{Wt} cut.

Using the neural network parameters from the inclusive reweighting (3 hidden layers with 50 neurons each) for the $R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}$ and $R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}$ reweightings did not give good closure. The hyperparameters were re-optimized and a model with 3 hidden layers of 50, 25, and 10 neurons including batch normalization [**batch-norm**] between the layers was found to work well. The closure on the reweighting variables (and m_{HH}) are shown in Figure **??** and Figure **??** for these two reweightings and demonstrates the confidence that we had that the individual reweightings going into Eq. G.5 were working nicely.

Although it seemed like the new reweighting were optimized well, when I applied the weights from Eq. G.5, the resulting X_{Wt} distribution had a very discontinuous QCD template in the $X_{Wt} < 1.5$ region, motivating us to dive in and understand what the issues were:

- When $R_{2b^{t\bar{t}}\to 4b^{t\bar{t}}}$ is very slightly larger than 1, we get large negative amplification in the weight from the denominator going to zero.

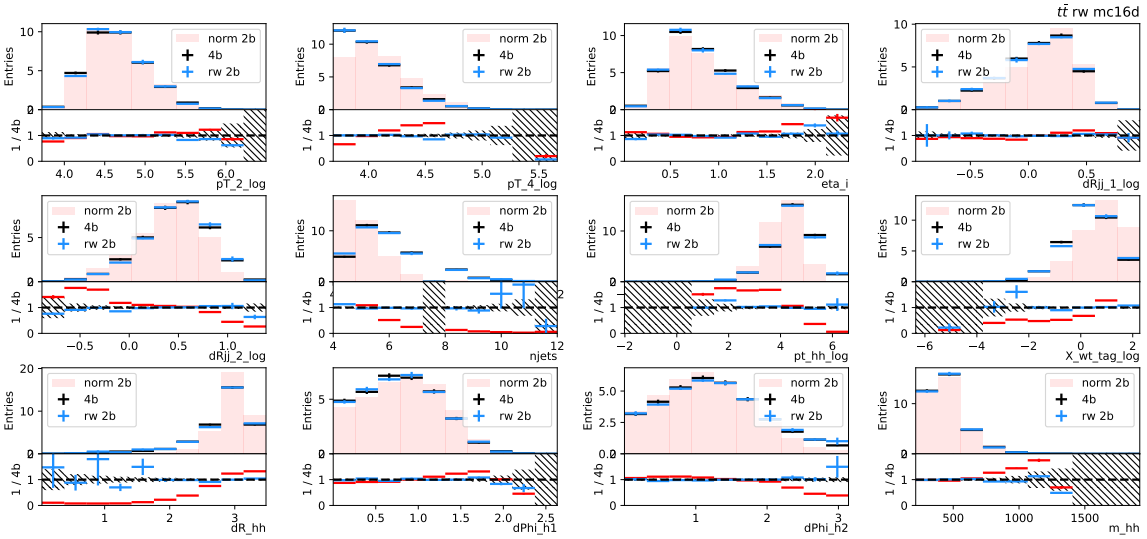


Figure G.15: $R_{2b^{t\bar{t}} \to 4b^{t\bar{t}}}$: The MC based $2b \to 4b$ $t\bar{t}$ reweighting.

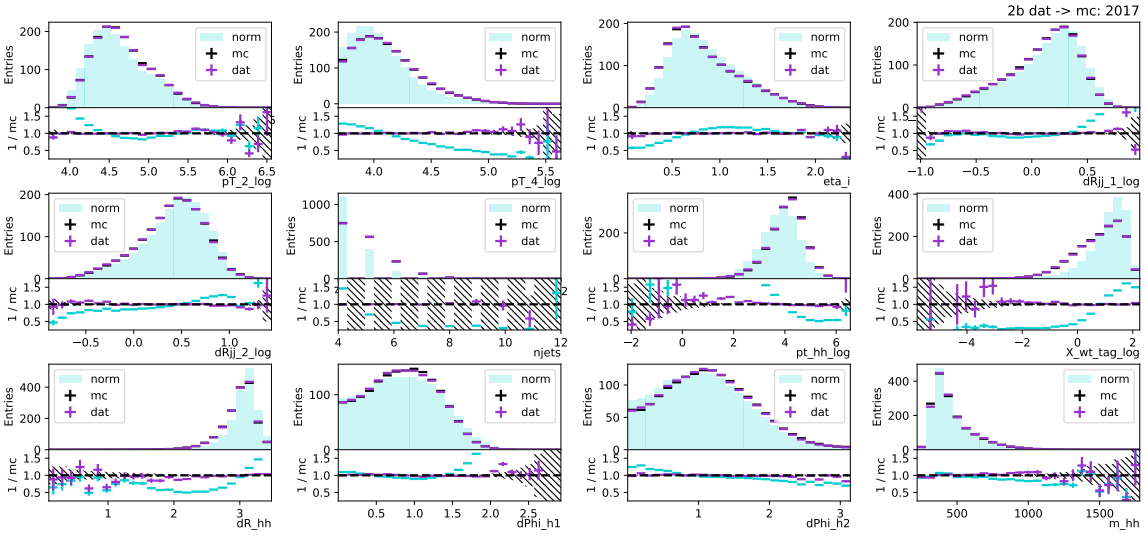


Figure G.16: $R_{2b^{t\bar{t}} \to 4b^{t\bar{t}}}$: Reweighting $2b$ data $\to$ $2b$ $t\bar{t}$.

- Sometimes we also got unphysically large weights for the $R_{2b^{t\bar{t}} \to 4b^{t\bar{t}}}$ reweighting, suggesting that this was a region of low support for the data the 2b distribution)
- Sometimes the subtraction in the numerator caused the weight to become negative and was accompanied by a large amplification from the denominator being small as well.

To fix the odd shaped qcd template issue, we removed the unphysical weights by cutting out the events with a weight more than 10 σ away from the mean. The resulting pre-fit templates for all of the years are shown in the right-hand column of Figure G.17, a by-eye reasonable template for modeling the pure QCD distribution.

We then the 2 components to the X_{Wt} shape inside in CR1 and check the agreement. Table G.2 shows the normalizations for the pure QCD fits, and the left-hand column of Figure G.17 shows the post-fit plots.

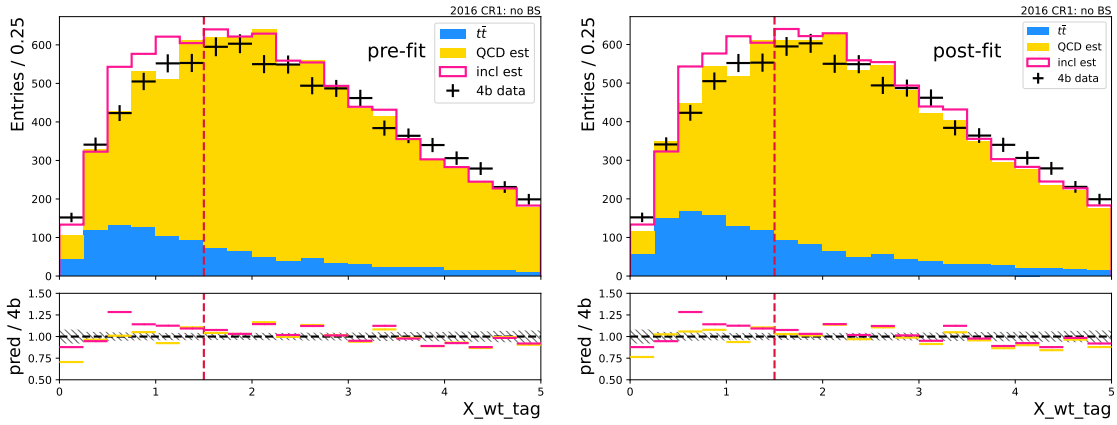
	16	17	18
$\alpha_{t\bar{t}}$	1.27 ± 0.11	1.84 ± 0.09	0.66 ± 0.08
α_{QCD}	0.950 ± 0.018	0.913 ± 0.014	1.012 ± 0.013

Table G.2: Fitted normalizations for the $t\bar{t}$ and QCD templates with the pure QCD reweighting.

I'm presenting the results here because I think it is an interesting study, and in the context of the approval processes for the analyses, how to account for $t\bar{t}$ or single Higgs backgrounds in the reweighting came up a lot. However, in terms of where is most useful to invest time in the next iteration, in Section tbd I will show that accounting for $t\bar{t}$ separately is also possible in other direct density estimation methods, and I personally think these are more promising for pursuing in the future, partially because although the pure QCD reweighting is mathematically elegant, implementing it in practice wasn't quite as trivial since getting a final weight that was reasonable when accounting for the weights of all of these separate pieces was rather non-trivial.

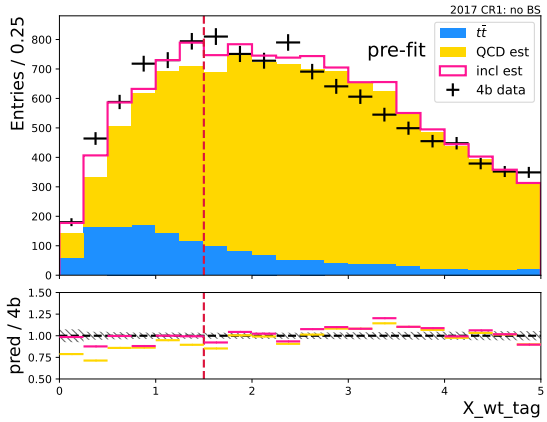
G.3.3 Outlook

After Run 3, our dataset size will *double*, making the intricacies of how we treat the separate components of our background estimate even more relevant. Although we showed several ideas for how to use the power of simulation, it wasn't deemed necessary for this current iteration of the physics analysis in terms of the background modelling, but, as we transition away from setting limits and into the realm of signal extraction, the import of well constraining each component of the background only grows. This chapter focused on $t\bar{t}$ background, and even though it is only 10% of our background, it is 100x as large as our HH signal. Another background we will start to care about more in the future is single Higgs background. Our single Higgs yield in the 4b SR is 3x the SM HH signal. The $bb\gamma\gamma$ analysis already includes the single Higgs background in their background

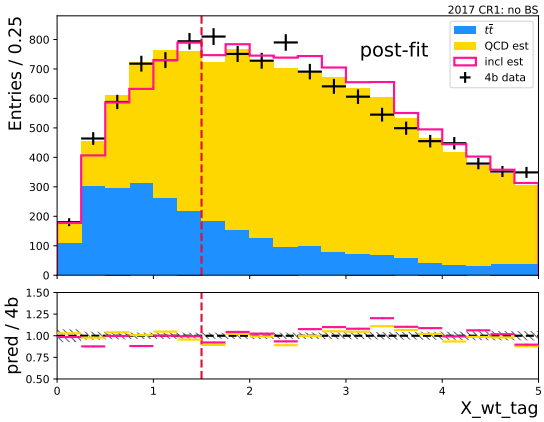


(a) 2016 pre-fit

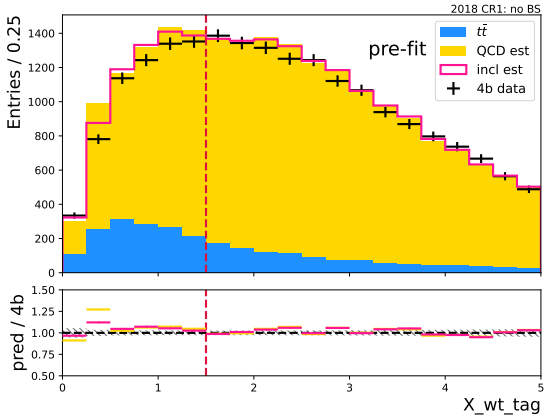
(b) 2016 post-fit



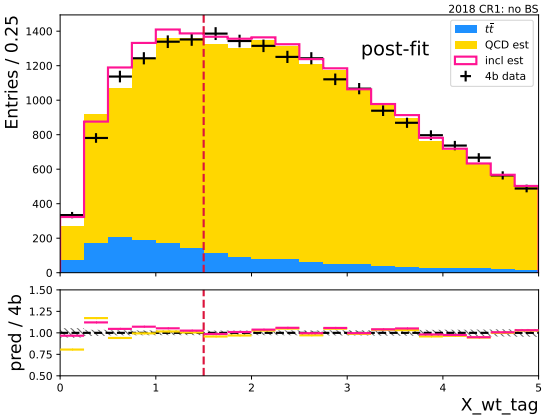
(c) 2017 pre-fit



(d) 2017 post-fit



(e) 2018 pre-fit



(f) 2018 post-fit

Figure G.17: X_{Wt} distributions in CR1 with the pre-fit (left) and post-fit (right) plots, compared to the 4b CR1 data.

modelling for the limit extraction, but the same methods from Section G.3.1 can be used to use MC to separately model the single Higgs background and is good in the future to HHarmonize with the other channels in our quest to extract the HH signal.

mybib