THE UNIVERSITY OF CHICAGO

A SEARCH FOR THE PRODUCTION OF PAIRS OF HIGGS BOSONS DECAYING TO FOUR BOTTOM QUARKS WITH THE ATLAS DETECTOR USING $\sqrt{S} = 13$ TEV PROTON-PROTON COLLISIONS AT THE LARGE HADRON COLLIDER

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
TODD SEISS

CHICAGO, ILLINOIS
JUNE 2021

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my advisor Mel, for being a steady guiding hand throughout the entire PhD experience. Thank you for always being available any time I had questions and for your always-astute observations in just about anything I could bring up. You gave me the freedom to be curious and to never stop learning.

Max, I cannot thank you enough for everything you have done for me. Thank you for answering every little question I had over many years in both FTK and analysis work, and for teaching me the ins-and-outs of ATLAS. I deeply appreciate your patience and all of the time you gave me. I am still impressed by your ability to effectively participate in multiple simultaneous meetings.

I also want to thank the entire hh4b team for making this challenging analysis a reality and for constantly striving for the best. Thank you also to the FTeamK without whom my time at CERN would have been much less fun. To the UChicago ATLAS team, thank you for your guidance, wisdom, and friendship.

Thank you to Henry, Liantao, and Sid for agreeing to be on my thesis committee and taking the time to attend the meetings and read the thesis.

I especially want to thank Henry for your constant support and for being an excellent PHYS 335 advisor, Young-Kee for her boundless energy and efforts to improve the department, and Kyle Cranmer for kindly giving me a home office in New York.

Thank you to all of the friends I made along the way and people who supported me in one way or another. A very incomplete list is

Tova
Simone
Alessandra
Michael H.
Karri
Elliot
Michelle
Liza
Tomoya
Ann-Kathrin
Emily
Amanda S.
Nicole
Nick
Lauren
Jon
Louis
Jan
John M.
Sean
Ashley
David S.
Christos
Michal
Joakim
Walter
Beojan
Victoria
Patrick
Calliope
Bing
Giordon
Jochen
Rachel
Eugene
Rui
Stany
Paolo
Emma
Kate W.
Marco
Tiffany
Rob
Mircea
John A.
Kate P.
Aparajita
Jacqueline
Sang
Tyler
Akis
Doug
Larry
Bri
Matteo
Dawn

Thank you especially to my classmates at Chicago, Mark, Gautam, Lesya, Evan, Bob, Lipi, Karthik, Dani, Ryan, and Alex, for making my time in Chicago such a pleasure.

Of course, I would never be here without the dedicated support of my parents, Brenda and John. You always encouraged me to follow my passions, working to overcome any obstacle I faced in reaching for my goals. I will never be able to fully repay you. Thank you also to my sister, Jenna, for always being a friend, a patient ear, and a cheerful supporting voice.

Lastly, I want to thank Carina. You defined my graduate school experience more than anything else, and I truly cannot wait to see what the future has in store for us. I could not have done this without you.

ABSTRACT

A search for non-resonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state is presented. The analysis uses up to 139 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS detector. The analysis targets Higgs boson pairs produced via gluon-gluon fusion including a diagram involving the Higgs self-coupling and decaying into four resolved b-tagged jets. The observed data are consistent with Standard Model predictions. A 95% confidence upper limit on the standard model cross-section is set at 284 fb, and the trilinear Higgs self-coupling parameter is constrained to $\lambda_{hhh} \in [-5.5\lambda_{\text{SM}}, 12.7\lambda_{\text{SM}}]$ at 95% confidence, assuming Standard Model values for all other couplings.

CHAPTER 1

INTRODUCTION

For millennia, humans have cast their eyes upwards to the stars with a fascination for what lay beyond Earth. Quite early in human history, careful observers noticed regular patterns in the motions of the sun, moon, stars, and planets and were able predict the future with astonishing accuracy, such as the dates of solar eclipses hundreds of years in advance. Through the careful study of these patterns, Isaac Newton was able to extend everyday experiences to explain the motion of celestial bodies through his three laws and the universal law of gravitation, elegantly connecting the motion of planets to arise from the same force that keeps us on the surface of the Earth.

The Newtonian approach to physics, based on intuition, was enormously successful for centuries. However, around the turn of the twentieth century, many physicists turned from studies of outer space to studies of inner space in order to learn how things behave on the smallest scales. Our everyday intuition suggests that the world is continuous. If we divide a material in half, we do not expect the two halves to obey different laws of physics than the whole. However, this scale invariance is dramatically violated in reality, as the basic laws that govern the behavior of small pieces of matter are *not* extensions of our everyday experiences. Scientists throughout the twentieth century were forced to piece together a completely new framework in order to explain their observations, the framework of quantum mechanics, and this framework is not merely an extension of the Newtonian laws of physics. The fundamental laws are completely different at the scale of nanometers (10^{-9} m). Atoms do no behave like tiny billiard balls.

This scientific revolution thrust the microscopic world to the forefront of scientific interest. No longer was outer space the only strange realm to explore – by looking inward at the smallest scales, we found a universe at least equally bizarre and fascinating. Quantum mechanics is the foundation of the modern Information Age. However, as physicists, we are interested in the small not necessarily to revolutionize society, but instead because this tiny universe constantly defies expectations, always begging the question “Why?” While these questions often feel disconnected from the “real world,” like being on another planet, it is in fact out of these bizarre rules that human experience is built. These *are* the fundamental laws of nature. By improving our understanding of the minuscule, we are able to make progress on basic questions about the nature of reality itself.

These concepts do not merely exist in the minds of theorists; they have been tested to extraordinary precision in the laboratory. In the same way the telescope is the tool of choice for exploration of the cosmos, we need a tool to probe the smallest possible distance scales, where interesting new laws of nature are found. The microscope, as the name suggests, probes distances only down to around a micrometer (10^{-6} m), insufficient for our purposes. In fact, the wavelength of visible light is hundreds of nanometers, so anything smaller than this cannot be resolved. Thus the direct use of our eyes to observe this world is impossible.

In fact, this wavelength problem is a challenging barrier. Quantum mechanics says that not only is light a wave, but that everything, even atoms, behaves like a wave. Thus any experiment we do will have some intrinsic maximum distance resolution equal to the distance of one oscillation of the wave. Any physics smaller than this will be “averaged over.” In the case of visible light with a wavelength of hundreds of nanometers, it is not possible to study the behavior of single atom, which is less than a nanometer in diameter, because the light will interact simultaneously with hundreds of atoms, making the small-scale behavior difficult to infer.

Thus, we need to use the smallest possible wavelengths if we want to understand the physics at the smallest scales. Quantum mechanics tells us that the wavelength of any object is inversely proportional to its energy. Double the energy and the wavelength is halved. For this reason, scattering experiments have become the de-facto standard for probing small distances. In a scattering experiment, particles are brought to a high energy and shot against another object. The patterns of deflection and debris that emerge from the collision then contain information about physics down to the distance scale associated with the energy of the incoming particle.

The first modern-style scattering experiment was the Rutherford experiment in 1911. Rutherford shot alpha particles at a gold foil and observed how the alpha particles were deflected. These alpha particles had energies roughly a million times that of visible light, allowing distances down to 10^{-14} m to be probed. He knew that atoms had a size around 10^{-10} m and therefore expected these high energy particles to be able to easily penetrate through. He found unexpectedly that some of the alpha particles deflected straight backwards, indicating that there was some hard object with a size around 10^{-14} m off of which the alpha particles could bounce. This was the discovery of the nucleus of the atom.

Less than 100 years later, the Large Hadron Collider (LHC) would come online, the highest energy (and therefore shortest distance) experiment ever conducted, with an energy scale 10 trillion times that of visible light, around 10 TeV. How this is achieved is infinitely more complex than the Rutherford experiment, but the basic principle of the experiment is the same. At these energies, we are able to probe physics at a distance scales of 10^{-20} m, 10s of zeptometers (1 zm = 10^{-21} m), making the LHC not a microscope, but a zeptoscope.

The LHC was built to discover and study the Higgs boson, which we now know has a mass of 125 GeV (corresponding to a wavelength of 10^{-19} m). The Higgs boson is a fundamental particle that was introduced to solve a problem discovered in the 1960s. While trying to build a consistent theory to explain the behavior of the known fundamental particles, theorists found a mathematical inconsistency. The math needed to describe certain radioactive decays did not allow any particles to have a mass. If particles had mass in this theory, then the probability of certain events happening would add up to more than 1, which is nonsense.

This turned out to be a challenging problem to solve. Theorists were able to prove that this problem would be encountered with almost any theory they could construct to explain these decays. There was, however, one loophole, the Higgs mechanism. Previously, theorists inserted the particle mass “by hand” into the theory as a fundamental property of a particle. However, fundamentally, mass is just an energy associated with the presence of a particle, as originally shown by Einstein in his famous mass-energy equivalence relation, $E = mc^2$.

The Higgs mechanism posits that there is a background Higgs field that carries some energy at every point in space. Particles then interact with this Higgs field, and this interaction carries some energy. The strength of the interaction is directly proportional to the mass of the particle. Any created particle will constantly be interacting with the Higgs field and carry at least this amount of energy, making the interaction energy equivalent to a mass. This mechanism circumvents inserting particle mass by hand and the associated mathematical problems.

In 2012, experiments at the Large Hadron Collider published the first experimental evidence that the Higgs field does exist in nature. Ripples in the Higgs field look like a particle, which we call the Higgs boson, and this is what we are able to observe at the LHC. One of the next major goals of the LHC is to understand all of the properties of the Higgs boson and test whether they agree with our expectations. Should there be a deviation, this would indicate that there is more to learn about the fundamental nature of mass.

One important property of the Higgs boson is how strongly it interacts with other Higgs bosons. We have a precise prediction for the strength of this interaction, and should the interaction be different than expected, this would mean that there are aspects of the Higgs mechanism we do not yet understand, potentially linking it to other open questions in physics.

Studying the strength of this interaction is the goal of this thesis. At the LHC, we collide protons, and if we want to study the interaction of Higgs bosons, we need to study collisions that produce two Higgs bosons. Whether two Higgs bosons are produced is fundamentally random and exceedingly rare. In fact, with the current number of collisions collected, there is no chance of making any good measurement of the interaction strength. However, if the strength of the interaction of Higgs bosons is different than we expect, then the net effect is generally to increase the probability of producing two Higgs bosons in one collision. Thus to study the interaction strength, it suffices to count the number of times that we observe two Higgs bosons produced. If we do not observe any instances of this, then we can rule out possible interaction strengths that would have been large enough that we would have expected to be able to make an observation.

The other major challenge with this measurement is that the Higgs bosons decay immediately into other particles, and in particular, these other particles can be produced via other processes at the LHC. This thesis studies the case when both Higgs bosons decay to pairs of b-quarks, the most probable decay, meaning that we need to look for collisions with four b-quarks. Unfortunately, at the LHC there are copious other ways to produce four b-quarks that do not involve Higgs bosons, so we also have to carefully study the data to understand the other sources as accurately as possible. This allows us to to distinguish a small excess of Higgs boson events over the noise, should such an excess exist.

CHAPTER 2

THEORY AND MOTIVATION: HOW TO MELT MASS

2.1 Fundamentals of Quantum Field Theory

Quantum field theory is one of the most profound theories ever developed when judged by depth of insights between seemingly unrelated phenomena. However, for most students, it is presented initially in quite an *ad hoc* manner, arbitrarily introducing the concept of quantum fields with limited motivation. This is perhaps reflective of the historical development of the theory, but it fails to capture just how remarkably a few assumptions heavily constrain the possible space of theories.

2.1.1 Combining Quantum Mechanics and Special Relativity

Suppose one knows nothing about quantum field theory but wants to develop a theory that is consistent with both special relativity and quantum mechanics. What options are there?

If we want to describe a quantum system, then we will be interested in quantum states $|\psi\rangle$ that are rays in a Hilbert space. Observables are defined by Hermitian operators $\mathcal{O}$ that act on the quantum state as $\mathcal{O}\,|\psi\rangle$. The eigenvectors of $\mathcal{O}$ can be indexed with a such that $|\psi_a\rangle$ has eigenvalues λ_a, $\mathcal{O}\,|\psi_a\rangle = \lambda_a\,|\psi_a\rangle$. According to the Born rule, the probability of observing a general quantum state $|\psi\rangle$ in state $|\psi_a\rangle$ is the squared inner product of the states, $|\,\langle\psi_a|\psi\rangle\,|^2$. Similarly, the expectation value of the operator $\mathcal{O}$ is $\langle\psi|\,\mathcal{O}\,|\psi\rangle$.

A change of coordinates for a quantum system is represented by a unitary transformation U. It is easy to see that a unitary transformation will preserve the probabilities of any observation, since $\langle\psi_a|\psi\rangle \mapsto \langle\psi_a|\,U^\dagger U\,|\psi\rangle = \langle\psi_a|\psi\rangle$.

Now we can begin to incorporate special relativity. The central postulate of special relativity is that the spacetime interval $ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 = \eta_{\mu\nu}dx^\mu dx^\nu$ is invariant under a change of intertial reference frames, which ensures that the speed of light, c, is constant in all reference frames. This means that we will be looking to construct observables that are invariant under the coordinate transformations $x^\mu \mapsto \Lambda^\mu_\nu x^\nu$ where Λ satisfies $\eta = \Lambda^T \eta \Lambda$. The set of these transformations forms the Lorentz group, which in 4 dimensional spacetime forms a 6-dimension Lie group. Three of the group generators correspond to Lorentz boosts (changing the velocity of the system), and the other three correspond to coordinate rotations.

In fact, we are actually interested in a slightly larger group. We would also like to ensure that the results of our experiments are the same no matter where and when the experiments are conducted. We want to ensure that we have four dimensional translation invariance $x^\mu \mapsto x^\mu + a^\mu$ for constant translations a^μ. This is a four dimensional group. When we combine it with the Lorentz-group, we end up with the 10-dimensional Poincaré group.

For our quantum mechanical theory to have observables invariant under general transformations $\mathcal{P}$ in the Poincaré group, then we need to find unitary representations such that if $|\psi\rangle \mapsto \mathcal{P}|\psi\rangle$, then

$$\langle\psi_a| \mathcal{P}^\dagger \mathcal{P} |\psi\rangle = \langle\psi_a|\psi\rangle \tag{2.1}$$

Eugene Wigner showed in 1939 that there are no finite-dimensional unitary representations of the Lorentz group [17]. This means that to satisfy unitarity and Lorentz invariance, our quantum states must have some infinite-dimensional index x. In particular, as first shown by Wigner and carefully proved by Weinberg, the representations are highly constrained [18]. In particular,

1. The representations in four dimensions are indexed by the spacetime coordiantes x^μ.

2. The representations are classified by a non-negative real number m, the mass, and by a positive half-integer, $J = 0, \frac{1}{2}, 1, \frac{3}{2}, ...$

3. If $m > 0$, then there are $2J + 1$ independent states. If $m = 0$ and $J > 0$, then there are exactly 2 independent states, and if $J = 0$, then there is always exactly one state.

This is an extremely profound result that shows that even with minimal physics input, we must be concerned with quantum fields $\psi(x)$ that are defined at every point in spacetime. In particular, we immediately see spin appearing as J, with the expected behavior that spin-1 massless particles (i.e., photons) should have two polarizations.

In particular, for the $J = \frac{1}{2}$ representations, which describe fermions like electrons, there are two similar but independent representations in which $\psi(x)$ is a two-component vector. The two representations do not mix under a Lorentz transformation, but they may be related dynamically. These two different representations will turn out to require the existence of antimatter, another profound result with very little input.

2.1.2 Dynamics

So far, we have not discussed how the quantum system actually evolves in time. We know that time-evolution must be accomplished by a unitary operator $\psi(t) = U(t)\psi(0)$ so that $\langle\psi(t)|\psi(t)\rangle = 1$ for all time. Any unitary operator can be written $U = e^{iH}$ for a Hermitan matrix H, which in this case can be identified as the Hamiltonian of the system (by requiring classical correspondence in the limit $\hbar \to 0$).

To constrain the form of the Hamiltonian, we can introduce a new constraint in addition to unitarity and Lorentz invariance, the principle of locality. Specifically, we want experiments at distant times and places to be uncorrelated. Experiments at Fermilab should not directly and physically impact observations at CERN. Violating this assumption makes following the scientific method extremely difficult, because experiments no longer would be repeatable.

Following the arguments of Weinberg, we write locality as the cluster decomposition principle [18]. Suppose we do several different distant (in space or time) experiments. The initial states of each experiment i are $|\alpha_i\rangle$, and the final states are $|\beta_i\rangle$. Then the set of experiments probes the quantity $\langle\beta_1, \beta_2, ...|\alpha_1, \alpha_2, ...\rangle$. The cluster decomposition principle states that this should factorize

$$\langle \beta_1, \beta_2, ... | \alpha_1, \alpha_2, ... \rangle = \langle \beta_1 | \alpha_1 \rangle \langle \beta_2 | \alpha_2 \rangle ... \tag{2.2}$$

Weinberg discusses that this principle is generally quite difficult to satisfy, and as of the writing of his book, he was aware of only one construction that satisfies the cluster decomposition principle for multiparticle experiments, which is to use a Hamiltonian constructed out of creation and annihilation operators.

Suppose our field has quantum numbers q in addition to the mass, spin, and momentum indices required by Poincaré invariance. Then the creation operator $a^\dagger(q)$ is defined as the operator that adds quantum numbers q to the state, $a^\dagger(q) \left|\psi_{q_1,q_2,...,q_N-1}\right\rangle = \left|\psi_{q_1,q_2,...,q_N}\right\rangle$. Note that any state can then be constructed from the vacuum state with no quantum numbers, $|0\rangle$, by repeated application of creation operators. The annihilation operator, $a(q)$, the adjoint of $a^\dagger(q)$, can be shown to have the opposite effect, of removing quantum numbers q.

A Hamiltonian that satifies the cluster decomposition principle with natural classical correspondence is $H = H_0 + V$ where

$$H_0 = \int dp a^\dagger(p) a(p) E(p) \tag{2.3}$$

with $E(p) = \sqrt{p^2 + m^2}$. V is an interaction term of the fields that can be constructed out of any number (including zero) of creation and annihilation operators but that must include exactly one three-dimensional momentum-conserving delta function. This Hamiltonian shows that in the free-field case of $V = 0$, the energy of the fields comes in quantized packets with minimum energy m.

From this Hamiltonian, one can get the equations of motion $\left(\Box^2 - m^2\right) \psi_l = 0$ for each degree of freedom of the field $\psi = \left(\psi_1 \; \psi_2 \; ... \right)$. Notably, for spin-0 fields ϕ, one gets the

Klein-Gordon equation of motion,

$$(\Box^2 - m^2)\phi = 0 \tag{2.4}$$

This is often written as in a Lagrangian density form,

$$\mathcal{L} = \frac{1}{2}(\partial^\mu \phi)(\partial_\mu \phi) - \frac{1}{2}m^2\phi^2 \tag{2.5}$$

where the first term is the kinetic term and the second term is the mass. Running this Lagrangian through the Euler-Lagrange equations yields the Klein-Gordon equation.

For spin-$\frac{1}{2}$ fields, the internal degrees of freedom can be packaged up into a four-component object ψ called a Dirac spinor that satisfies the Dirac equation

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \tag{2.6}$$

which comes from the Dirac Lagrangian

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi \tag{2.7}$$

where γ^μ are the Dirac matrices defined by the representation of ψ in the Poincaré group (they are the generators of the Dirac algebra).

For spin-1 particles, the representations are more complicated. In particular, there are infinitely many equivalent representations known as gauges that are related to each other by gauge transformations. One can show that the Lagrangian for spin-1 fields is

$$\mathcal{L} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}m^2 A_\mu^2 \tag{2.8}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. When $m = 0$, this Lagrangian leads to Maxwell's equations, and when $m \neq 0$, this Lagrangian is called the Proca Lagrangian.

The same reasoning that leads to the above Lagrangians can also be carried to higher spins. There are no confirmed fundamental spin-$\frac{3}{2}$ particles, but one can write the Lagrangian to describe them, the Rarita-Schwinger Lagrangian

$$\mathcal{L} = -\frac{1}{2}\bar{\psi}_\mu(\epsilon^{\mu\kappa\rho\nu}\gamma_5\gamma_\kappa\partial_\rho - im\sigma^{\mu\nu})\psi_\nu \tag{2.9}$$

The massless spin-2 case is particularly interesting. There turns out to be a single unique Lagrangian that must couple to all the fields present in the theory. This Lagrangian takes the form of the Einstein-Hilbert action and reproduces Einstein's field equations [19]. One therefore observes the surprising fact that gravity naturally emerges from the basic postulates of quantum field theory and is quantized by a massless spin-2 particle, the graviton. In an alternate history without Einstein, general relativity could have been discovered via this route.

Interestingly, it is not possible to construct an interacting theory with massless particles of spin greater than 2, because the gauge invariance requirements are too restrictive. Interacting massive higher spin particles are perfectly allowed (and abundant in nature as composite particles).

2.1.3 Interactions

All of the above discussion has been primarily concerned with free particles. We know that in our universe, particles interact. What possible interactions are allowed? There are two main principles that guide the structure of particle interactions, renormalizibility and local gauge invariance. We will be concerned in this section only with particles of spin-0, spin$\frac{1}{2}$, and spin-1.

Renormalizibility

The procedure of renormalization is one of the most profound insights in quantum field theory and is a purely quantum effect. Originally introduced reluctantly as a way to hide infinities when computing higher order quantum effects, renormalization ended up being a central and defining feature of quantum field theory.

At the most basic level, renormalization handles the fact that the majority of quantum field theories are not scale invariant. Various "fundamental observables" (such as charge and particle mass) actually depend on the energy or distance scale at which they are measured. The example of electric charge is relatively easy to understand conceptually. Suppose one is measuring the fine-structure constant (or equivalently, the charge of an electron). Close to the electron, the electric field is strong enough to polarize the vacuum slightly via the creation of short-lived virtual electron-positron pairs. This implies that far away (at low energy scales), the electron charge is screened by the vacuum polarization, exactly analogous to charge screening in a dielectric. At shorter distances (higher energy), one penetrates this cloud and the charge of the electron is larger. To first order in quantum electrodynamics (with just an electron and photon), the fine-structure constant depends on the energy scale E as

$$\alpha(E) = \frac{\alpha_0}{1 - \frac{2\alpha_0}{3\pi}\ln\left(\frac{E}{m_e}\right)} \tag{2.10}$$

where $m_e = 511$ keV is the mass of the electron, and α_0 is the fine structure constant measured at the scale m_e, $\alpha_0 \approx \frac{1}{137}$. At the energy scales of the LEP collider, $\alpha(E = m_Z \approx 90\text{ GeV}) \approx \frac{1}{127}$, so the charge of electron is measured to be about 7% larger.

In quantum electrodynamics (QED), when the only particles are the electron and photon, there are only two quantities to measure, α and m_e. Given a measurement of these quantities at some energy scale, renormalization provides a procedure to compute the value at any other scale. This is a statement that QED is renormalizable. This is very much not guaranteed for an arbitrary model. A non-renormalizable theory requires more and more parameters to be measured as the energy is increased. The theory is still predictive and often very useful, but will tend to rapidly require an unwieldy number of parameters as the measurement energy is increased.

Proving that a given theory is fully renormalizible can be quite a task. However, as a guide, a useful check is dimensional analysis. The Lagrangian must have mass dimension +4 (ie units of eV^4). Spin-0 and spin-1 fields have mass dimension 1, and spin-1/2 fields have mass dimension 3/2. Couplings will then have the appropriate mass dimension to ensure the Lagrangian is mass dimension +4. The Lagrangian will (usually) be renormalizable as long as the constant parameters in all terms have mass dimensions 0 or larger. For-example, the self-interacting scalar Lagrangian $\mathcal{L} = \frac{1}{2}(\partial^\mu \phi)(\partial_\mu \phi) - \frac{1}{2}m^2\phi^2 - \lambda\phi^4$ is renormalizible because m^2 has mass dimension +2, and λ has mass dimension 0. An interaction term like $\frac{1}{\Lambda}\phi^5$ is non-renormalizible because $\frac{1}{\Lambda}$ has mass dimension -1. This means that to absorb all infinities at sufficiently high energy, one will have to introduce $\phi^6, \phi^7, ...$ terms with their own coupling strengths that can only be determined experimentally.

The additional terms required by a non-renormalizible theory will typically be small at energy scales less than the mass scale of the coupling, $E \ll \Lambda$. An excellent example of how this could be used is in the decay of muons. The muon mass is approximately $m_\mu = 105$ MeV and it decays to electrons and neutrinos whose masses are negligible in comparison. Given no other mass scales in the problem, one might estimate that the muon should decay with a lifetime $\tau_\mu = \frac{h}{m_\mu c^2} \approx 10^{-23}$ sec, assuming a coupling for the decay of order 1. The muon lifetime, however, is dramatically longer, at around 10^{-6} sec.

A direct decay of the muon to 3 other fermions would correspond to a four-fermion term in the Lagrangian, which has a coupling parameter with mass dimension -2. Thus there must actually be another mass scale in this problem, which turns out to be rather large, the W boson mass, $m_W \approx 80$ GeV. Thus the muon decay is suppressed by the ratio of the mass scales $\frac{m_\mu}{m_W}$. The first-order calculation of the muon lifetime gives [14]

$$\tau_\mu = 394\pi \frac{1}{\alpha_w^2} \left(\frac{m_W}{m_\mu}\right)^4 \frac{\hbar}{m_\mu c^2} \tag{2.11}$$

where $\alpha_w \approx \frac{1}{29.5}$ is the weak coupling strength at the muon mass scale. We can see therefore that the muon lifetime is dramatically extended because of the high mass scale involved in the non-renormalizable four-fermion interaction.

Thus the modern interpretation is that renormalizibility is not a necessary requirement of a fully-consistent theory, even though much of modern particle physics was developed with the mindset. The modern approach is to state that requiring renormalizability is equivalent to requiring all non-renormalizible interactions to be strongly suppressed by some large mass scale Λ. This is the effective field theory approach, where we assume that such interactions exist but are small at the energy scales that we can probe. Given that no new physics has been observed above the electroweak scale (≈200 GeV), we can use renormalizibility as a guideline for deciding what interactions we should allow, with non-renormalizible interactions allowing for potential small corrections.

Gauge Theory

Requiring renormalizibility (which is equivalent to saying any unknown physics is at a large energy scale) rules out a huge number of possible interactions. However, the interactions of spin-1 particles are even further constrained by only Lorentz invariance and unitarity. All interactions of spin-1 particles with other particles must be governed by a local gauge symmetry [19].

The simplest example is scalar QED, in which a scalar particle is coupled to a massless photon. Suppose we start with a complex scalar field ϕ and a photon A_μ, then the free Lagrangian is $L_0 = (\partial_\mu \phi^*)(\partial^\mu \phi) - m\phi^*\phi - \frac{1}{4}F^{\mu\nu}$. Note that this Lagrangian is invariant under the transformation $\phi \rightarrow e^{i\alpha}\phi$ for an arbitrary phase α. This is a $U(1)$ symmetry.

The principle of gauge theory is to promote this U(1) transformation to a local transformation that depends on the spacetime coordinate x, $\phi \to e^{iq\alpha(x)}\phi$ for an arbitrary constant q. The Lagrangian is no longer invariant under this transformation; however we can enforce this symmetry by modifying the Lagrangian. In order to ensure local gauge symmetry, we change the derivative to a covariant derivative, $\partial_\mu \to D_\mu = \partial_\mu + ieA_\mu\phi$ and require that A_μ transforms as $A_\mu \to A_\mu + \frac{1}{q}\partial_\mu\alpha(x)$. Thus the Lagrangian becomes

$$\mathcal{L} = (D_\mu\phi)^*(D^\mu\phi) - m|\phi|^2 - \frac{1}{4}F^{\mu\nu} \tag{2.12}$$

Notice that the covariant derivative D_μ induces a coupling between ϕ and A_μ, and in particular ϕ and ϕ^* couple with opposite charge (and are therefore antiparticles).

Similarly, if we have, for example, two particle with identical masses, there is an $SU(2)$ symmetry of the free-particle Lagrangian that rotates them into each other, $\binom{\phi_1}{\phi_2} \to U\binom{\phi_1}{\phi_2}$ for a 2×2 unitary matrix U. Promoting U to a local transformation $U(x)$ and requiring invariance of the Lagrangian then induces coupling between the spin-1 particles and the fields ϕ_1 and ϕ_2, albeit with a more complicated structure due to the non-Abelian nature of the $SU(2)$ group.

While this procedure may seem arbitrary, it is actually the only Lorentz-invariant and unitary way to build interacting spin-1 particles [19]. Thus by categorizing the possible symmetries of a given set of particles, we can enumerate all possible interactions involving spin 1 particles.

Massive Vector Bosons and Chiral Interactions

There are a few problems that the above formulation runs into when confronting the reality of particle interactions. First is the observation that massive gauge bosons exist. This means there must be a term in the Lagrangian like $mA^\mu A_\mu$, which will violate any gauge symmetry, which in turn translates to unitarity violation.

Before resolving this problem, we can examine another complication realized in nature that leads to a similar problem with a similar solution, and that is the problem of chiral interactions.

Spin-1/2 fermions are representations of the Lorentz group, and we often write them as a four component Dirac spinor ψ. However, Dirac spinors are actually a combination of two irreducible representations that transform differently. In the Weyl basis, these two separate two-dimensional representations can be written $\psi = \binom{\psi_L}{\psi_R}$ where ψ_L is the left-handed Weyl spinor, and ψ_R is the right-handed Weyl spinor. In this language, the mass term in the Dirac Lagrangian is $m(\bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R)$.

Now, a bizarre fact of nature is that the symmetry group of the weak force involves only left-chiral particles ψ_L. The group transformation is analogous to $\psi_L \to e^{i\alpha(x)}\psi_L$, $\psi_R \to \psi_R$. This means that the mass term $\bar{\psi}_L\psi_R$ will violate gauge symmetry. Thus we see that in a theory of chiral interactions, massive fermions violate gauge symmetry (and therefore unitarity)!

Spontaneous Symmetry Breaking

Fortunately, both the problem of fermion mass in a chiral theory and the problem of massive vector bosons are solved in the same way, with spontaneous symmetry breaking. The idea is that one writes a Lagrangian that initially has no masses, but the masses are generated dynamically within the theory. The massless fermions and vector bosons couple to a scalar field Φ. The couplings to fermions are simple Yukawa couplings, $y\Phi\bar{\psi}\psi$, and the couplings to the vector bosons requires a gauge symmetry, as discussed above, resulting in terms that look like $g\Phi A^{\mu}A_{\mu}$.

The key trick is that the potential for the ϕ field takes a form like $V(\Phi) = -\mu^2\Phi^2 + \lambda\Phi^4$ that has a nonzero miminum. Particles are excitations about the minimum of the potential, so to do calculations in the standard way, one shifts the field, $\phi = \Phi - v$, where v is the vacuum expectation value of the field, which is the value of Φ that minimizes $V(\Phi)$ ($v = \frac{\mu}{\sqrt{2\lambda}}$

for the simple potential above).

After this field redefinition, the fermions and vector bosons pick up an apparent mass term, $y\Phi\bar{\psi}\psi \to y\phi\bar{\psi}\psi + yv\bar{\psi}\psi$ and $g\Phi A^\mu A_\mu \to g\phi A^\mu A_\mu + gvA^\mu A_\mu$, with masses $m_\psi = yv$ and $m_A = gv$.

This mechanism spontaneously breaks the gauge symmetry to produce particle masses and is called the Higgs mechanism. One of the major accomplishments of the Large Hadron Collider has been confirming that the Higgs field ϕ actually exists and appears to utilize this mechanism.

Summary of Possible Particle Interactions

- Scalars (Spin-0): Can interact with itself up to ϕ^4 terms.
- Fermions (Spin-1/2): Cannot interact directly with other spin-1/2 particles. Can interact with scalars via a Yukawa coupling.
- Vector Bosons (Spin-1): Can only interact with other particles via gauge symmetries. Cannot have mass, except via symmetry breaking mechanisms.

2.1.4 Observables

So far, we haven't discussed how to actually calculate any observable quantity. One of the key observable quantities in quantum field theory is the cross-section, which parameterizes how likely two particles are to interact and produce a specified final state. This quantity is so ubiquitous because it achieves the rare balance of being both easy to measure and easy to calculate (where "easy" is very much a relative term).

Cross Sections

Suppose we run an experiment where two particles come in from far away at $t = -\infty$, interact, and some number of particles travel to far away at $t = +\infty$. Let P be the total quantum mechanical probability that these particles interact, and dP the differential probability of an interaction that results in the final state particles ending in some differential region of phase space. Let Φ be the incoming flux of particles, in number of particles per area per second. We then run the experiment for some time T. We can combine all of these quantities into the definition of the (differential) cross-section,

$$d\sigma = \frac{1}{T}\frac{1}{\Phi}dP. \tag{2.13}$$

With this definition, we can absorb all of the "experiment-dependent" quantities (such as flux, time, number of initial particles) into one number, the luminosity L such that the total number of particles in a differential unit of phase space is

$$dN = Ld\sigma. \tag{2.14}$$

The luminosity is a useful quantity around which an experiment can be engineered and built. It does not depend on any interaction probabilities but only on the incoming particles. Then the experimentalist can directly measure dN for a given process and given L, can directly determine the cross-sections.

Cross-sections are typically measured in barns, where 1 barn = 10^{-24} cm^2 is the order-of-magnitude of the cross-sectional area of a Uranium nucleus (which has a radius of around 11 $\times 10^{-12}$ cm). Luminosity is then measured in inverse barns. At the time of writing this thesis, ATLAS has recorded 139 fb^{-1} of collision data.

If this is how the theory is to be tested, we need to be able to compute cross-sections. We want to know the probability of an initial state $|i\rangle$ at $t = -\infty$ evolving to state $|f\rangle$ at $t = +\infty$. This is typically computed in the Heisenberg picture where the states do not evolve, but the operators do, so we are interested in $\langle f|\, S\, |i\rangle$ where the S-matrix contains all the interaction physics.

The S-matrix contains a trivial component, where $|i\rangle$ and $|f\rangle$ are the same state, so we typically separate this. We also know that because of Poincaré invariance, we will always have a conserved momentum and energy. Therefore, the convention is to write

$$S = \mathbf{1} + i(2\pi)^4 \delta^4(\Sigma p)\mathcal{M} \tag{2.15}$$

where $\mathcal{M}$ is called the matrix element. One can then show that the cross-section for general two-particle input states is

$$d\sigma = \frac{1}{(2E_1)(2E_2)|\vec{v}_1 - \vec{v}_2|}|\mathcal{M}|^2(2\pi)^4\delta^4(\Sigma p) \prod_{\text{final states } j} \frac{d^3p_j}{(2\pi)^3}\frac{1}{2E_j} \tag{2.16}$$

A similar logic works for decay rates, where a one-particle initial state transforms into a multi-particle final state. The probability of this happening over time T is $d\Gamma = \frac{1}{T}dP$. Carrying through the above logic, one finds,

$$d\Gamma = \frac{1}{2E}|\mathcal{M}|^2(2\pi)^4\delta^4(\Sigma p) \prod_{\text{final states } j} \frac{d^3p_j}{(2\pi)^3}\frac{1}{2E_j} \tag{2.17}$$

The total particle lifetime is just $\tau = \frac{1}{\int d\Gamma}$.

Feynman Diagrams

Given these relatively simple formulas for measurable quantities, the goal of the theorist is to then calculate the matrix elements $\mathcal{M}$ in a given model. This is most famously done perturbatively, using a Taylor expansion in the interaction strength. Note that perturbation theory will only be valid for small interaction strengths, which in the real world is true for everything except low-energy quantum chromodynamics, making those calculations notoriously difficult.

To calculate the matrix elements $\mathcal{M}$, Feynman devised a way of representing the rather complicated Taylor expansion algebra pictographically. The incoming particles are represented as edges in a graph, and interactions are represented by vertices. An interaction term in the Lagrangian like $\phi\bar{\psi}\psi$, for example, corresponds to a vertex with one scalar edge and two fermion edges. Each edge is assigned a momentum and a charge, and the vertex must respect momentum and charge conservation. Multiple vertices can then be rearranged and "snapped together" to represent a given physical process. Figure 2.1 shows some example Feynman diagrams in quantum electrodynamics.

From the Feynman diagrams, one follows the Feynman rules to read off exactly what integrals to write down, and then "turn the crank" to compute amplitudes. One challenge is that as one goes to higher order (to diagrams with more vertices/loops), the number of diagrams to include grows factorially. By fourth and fifth order, which is necessary for a handful of precision measurements, there are many thousands of diagrams to include.

2.2 The Standard Model

The above discussion has laid out an overview of the entire framework of quantum field theory, from why fields are necessary to how observable quantities are computed. While the theory space is extremely constrained, we have not yet discussed how this is realized in nature. As far as we currently understand, which particles exist and how they interact is arbitrary, as long as they satisfy the above requirements. The set of particles and interactions realized in nature is called the Standard Model and is one of the most successful scientific theories of all time. While this model has to date successfully passed every terrestrial test, there are still many outstanding issues.

2.2.1 Structure of the Standard Model

Interactions between particles in the Standard Model are almost entirely described by gauge interactions. The fundamental gauge group of the Standard Model is $U(1)_Y \times SU(2)_L \times SU(3)$. The $SU(3)$ group corresponds to the strong force of quantum chromodynamics, and the $U(1)_Y \times SU(2)_L$ component is the electroweak sector. In the electroweak sector, the

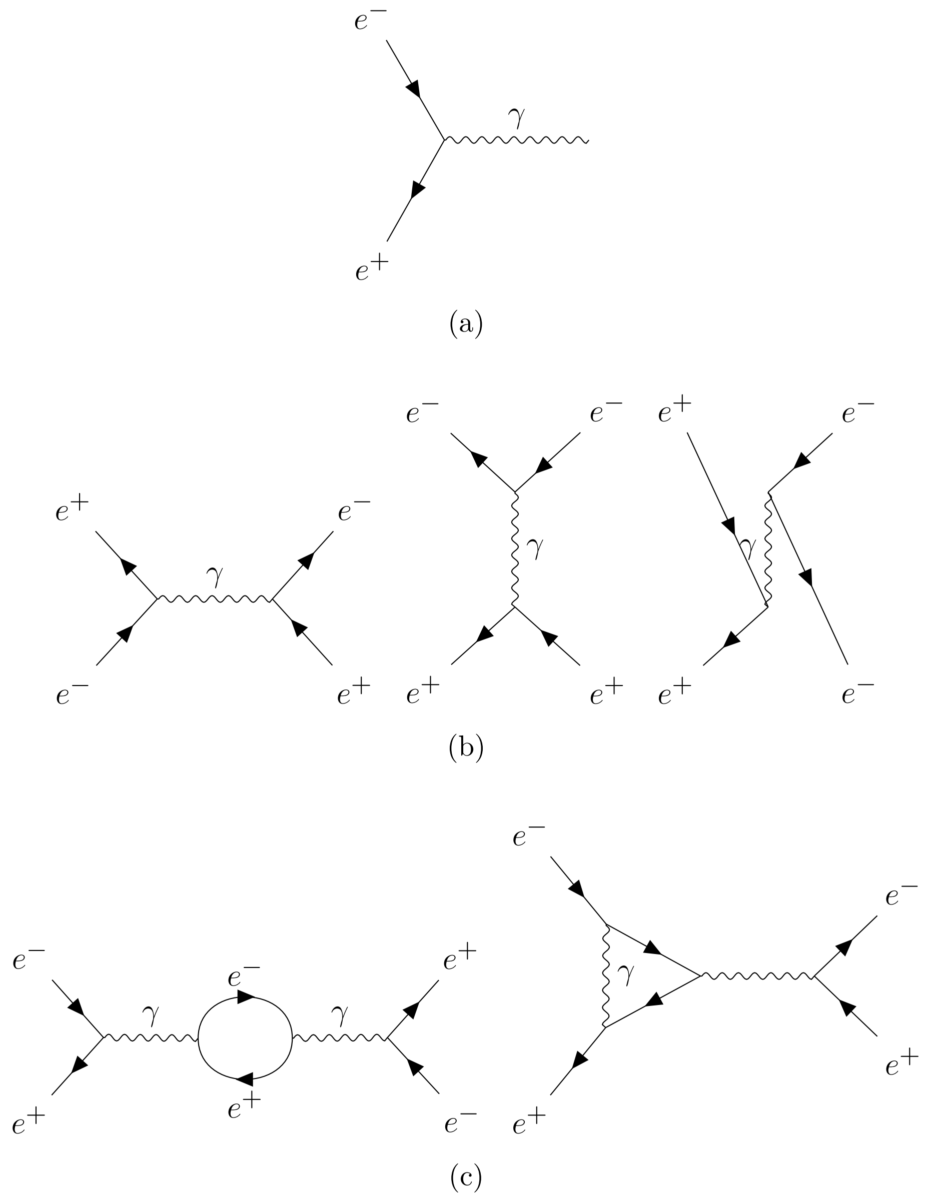


Figure 2.1: (a) The basic vertex in quantum electrodynamics, representing coupling between electrons and photons. The arrows on the electron lines represent electric charge. (b) The three tree-level Feynman diagrams representing elastic electron-positron scattering. Time runs left to right. (c) Two of the many one-loop diagrams for electron-positron scattering. Note that the fermions in the loop in the photon line can by any fermions in the theory, though the lightest fermions will have the most significant contribution.

$SU(2)_L$ transformation couples only to left-chiral fermions, so in order to preserve unitarity, all fermions start massless with Yukawa couplings to a complex scalar Higgs doublet that spontaneously breaks the symmetry, generating fermion mass. After symmetry breaking, there is a single massless boson that couples with a $U(1)$ symmetry group, which is the photon, 3 massive vector bosons, the $W^{\pm}$ and Z bosons, and a single real scalar boson, the Higgs boson. This procedure ultimately requires four independent parameters that have to be measured in experiment, the $U(1)$ coupling strength, the $SU(2)$ coupling strength, and two parameters describing the Higgs potential. In addition, every Higgs Yukawa coupling to the fermions requires a Yukawa coupling strength that is directly proportional to the final fermion mass.

In the fermion sector, there are 12 different particles that are charged under at least one of these groups. The 12 fermions are organized into 3 generations of 4 particles each, two quarks, a charged lepton, and a neutral lepton. The quarks are charged under all three groups, and because they are charged under $SU(3)$, each quark has three possible color charges. The quark pairs form doublets under $SU(2)$, though interestingly, because there are three generations and because all the quarks are massive, the coupling has the opportunity to be non-diagonal with respect to the mass eigenstates. This means that the W boson can (and does) couple quarks from different generations in the same vertex. The matrix associated with this non-diagonality has four parameters, the three CKM mixing angles and one complex phase. Lastly, the "up-type" quarks all have electric charge $+2/3e$, and the "down-type" quarks have charge $-1/3e$.

The gauge group $SU(3)$ has the property that the gauge coupling increases with increasing distance (or decreasing energy). This means that the coupling strength between two distant bare quarks is huge, and it is actually energetically favorable to pull quarks out of the vacuum to pair up with them. This is the principle of confinement – bare quarks do not exist beyond a length scale of around 1 femtometer. They will always join together into colorless composite particles, hadrons, bound together by gluons, the $SU(3)$ gauge boson.

Bound states of a quark-antiquark pair are mesons, and bound states of three quarks is a baryon, of which the proton is the lowest-mass example.

In the lepton sector, each generation has a doublet of one massive charged lepton (the electron, muon, and tau), and one nearly-massless uncharged lepton (the neutrinos). In the standard model, neutrinos are massless and only left-chiral, though the recent observation of neutrino oscillations implies that neutrinos have mass, which complicates this story somewhat, as discussed in section 2.2.2.

All of the free parameters of the Standard Model are listed in Table 2.2.1. This excludes parameters associated with neutrino masses, of which there are up to 9 more. Note that all of the parameters in the table are unitless except for the Higgs vacuum expectation value. This is the electroweak scale, which is the only energy scale that is added by hand to the Standard Model. The rest of the parameters (in particular, the fermion masses) are determined with respect to this scale. Increase v and all of the particle masses increase linearly.

Note that despite this, another scale arises dynamically, the QCD scale $\Lambda_{\mathrm{QCD}} \approx 250$ MeV. This is the scale at which, in perturbation theory, the strong coupling constant diverges (and therefore a scale by which perturbation theory must break down). This scale is what sets the masses of the protons and neutrons (each quark in the proton carries a gluon cloud with energy approximately Λ_{QCD}). The exact value is difficult to calculate, but one can view this scale as determined soley by the short-distance (high-energy) strength of the strong force. The fact that a scale arises out of a theory with no input scale is a purely quantum effect called dimensional transmutation.

Symbol	Description	Value
v	Higgs Vacuum Expectation Value	246 GeV
λ	Higgs Self-Coupling	0.13
y_e	Electron Higgs Yukawa	2.94×10^{-6}
y_μ	Muon Higgs Yukawa	6.07×10^{-4}
y_τ	Tau Higgs Yukawa	1.02×10^{-2}
y_d	Down Quark Higgs Yukawa	4×10^{-5}
y_u	Up Quark Higgs Yukawa	2×10^{-5}
y_s	Strange Quark Higgs Yukawa	6.9×10^{-4}
y_c	Charm Quark Higgs Yukawa	6.9×10^{-3}
y_b	Bottom Quark Higgs Yukawa	2.5×10^{-2}
y_t	Top Quark Higgs Yukawa	0.99
g_e	EM Coupling Strength	0.304
g_w	Weak Coupling Strength	0.630
g_s	Strong Coupling Strength*	1.21
θ_{12}	CKM 12 Mixing Angle	13.0°
θ_{23}	CKM 23 Mixing Angle	2.42°
θ_{13}	CKM 13 Mixing Angle	0.21°
δ_{CP}	CKM CP-Violating Phase	71°

Table 2.1: The 18 free parameters of the standard model, excluding neutrino masses and mixings (which adds 7 or 9 new parameters), which are not yet fully understood, and the QCD vacuum angle [14, 15].
*At the renormalization scale $\mu = M_Z$.

2.2.2 Outstanding Problems with the Standard Model

Particle Content

One of the most immediately striking features of the Standard Model is its seeming arbitrariness. Why are there four particles per generation, and why are there 3 generations? Why is the Standard Model symmetry group what it is, with no apparent other gauge groups? Why is the weak force chiral, coupling only to left-handed particles? What determines the fermion masses, or equivalently, their coupling strengths to the Higgs field? What determines all of the free parameters of the Standard Model?

These and other related questions are difficult to approach, since there is no clear direction for how they should be resolved, experimentally or theoretically. However, there is one interesting nontrivial theoretical statement that can be made about the relationship between the particle content and the gauge groups.

For a given Lagrangian, it is not necessarily true that a given symmetry will hold once quantum corrections are considered. The symmetries may be violated at loop level, which is called an anomaly. When an anomaly is associated with a gauge symmetry, it is called a gauge anomaly, and ultimately will lead to a violation of unitarity. So valid theories must be free of gauge anomalies. It turns out that gauge anomalies can only occur in chiral theories, of which the standard model happens to be one. To ensure that anomalies cancel, there are non-trivial constraints on the possible gauge group charges.

In the Standard Model, one of the most interesting constraints is that the quark electric charge must be exactly $\frac{1}{3}$ or $\frac{2}{3}$ the charge of the electron, in order for the electroweak gauge group to be valid at the quantum level. Thus protons must have exactly the same charge as electrons, so atoms are perfectly neutral.

The constraints in the Standard Model are stronger than this, however. In particular, one can include gravity, since it can also be viewed as a gauge group, and gravity coupling to all the fermions in the Standard Model provides further constraints. In the end, there are two one-parameter families of valid possible charges [19]. The Standard Model falls into one of these families, so it is anomaly free.

A Unified Theory

One of the most famous problems in particle physics is the search for a unified theory. When unifying the electroweak and strong interactions, it is often called a Grand Unified Theory (GUT). When unifying these plus gravity, it is often called a Theory of Everything.

After the unification of the weak and electromagnetic forces, it was natural to attempt to include the strong force as well. It turns out that if one runs the coupling constants of the weak, electromagnetic, and strong forces all to high energies, they nearly intersect at an energy scale around 10^{16} GeV, the GUT scale. All three symmetry groups could even be neatly packaged up into a single gauge group such as $SO(5)$ that has its symmetry broken at the GUT scale. However, Grand Unified theories almost universally predict baryon number violation at the GUT scale, which allows for proton decay into a positron plus other particles, depending on the model. The most recent limits from Super-Kamiokande set limits on the proton lifetime for decay into a positron plus pion at 1.6×10^{34} years [20]. Similar to the muon lifetime problem discussed in section 2.1.3, the proton lifetime is expected to be roughly

$$\tau_p \propto \frac{1}{m_p}\left(\frac{\Lambda_{\mathrm{GUT}}}{m_p}\right)^4 . \tag{2.18}$$

For the proton mass approximately 1 GeV, a limit of 10^{34} years translates to $\Lambda_{\mathrm{GUT}} > 10^{15}$ GeV, tantalizingly close to it's maximum expected value around 10^{16} GeV. We can conclude that there are no baryon-number violating processes below this scale (unless they are specifically suppressed in proton decay). Many naive models, however, actually expect a lifetime several orders of magnitude shorter than this, and these models have therefore been excluded. The fact that these proton decay experiments are able to directly test physics at these mass scales is incredible. This energy corresponds to around 1 nanogram of mass, 10 times the mass of a typical human cell, which would be contained in a quantum fluctuation at lengths scales of the size of the proton. This energy is comparable to the kinetic energy of a small car at highway speeds. In comparison, the highest energy scales currently probed by the LHC are the contact interaction searches that reach an energy scale around 3.5 $\times 10^4$ GeV for very specific interactions [21].

At even higher energies, many people expect that gravity and quantum mechanics could be unified in a theory of everything. The problem is that in quantum field theory, gravity is non-renormalizable. This means that to make predictions at the scale where the quantum effects of gravity should be observable, the Planck scale ($\Lambda_{\text{Pl}} \approx 10^{18}$ GeV), one would need a very large (or inifinte) number of free parameters. Many therefore conclude that some deeper underlying theory must supersede quantum field theory at very short distances, the most well known candidate of which is string theory.

The Hierarchy Problem

Another more subtle theoretical concern with the Standard Model is the so-called hierarchy problem. This particular issue is often considered one of the best motivations for finding new physics at energy scales in the range of a few TeV.

The problem is that when one computes quantum corrections to the mass of the Higgs boson, one finds that due to loop diagrams, the Higgs mass should be roughly the scale of the heaviest particles in the theory. The one-loop correction to the Higgs mass takes the form

$$m_h^2 = -\mu^2 + k\Lambda^2 \tag{2.19}$$

where μ is the bare mass parameter that appears in the Lagrangian, Λ is the highest mass scale up to which the theory is valid, and k is a constant of order 1 that depends on the particle content of the theory.

If there is no new physics between the electroweak scale and the scale of grand unification or quantum gravity, then Λ would be roughly of order 10^{16}-10^{18} GeV. The Higgs mass is 125 GeV, which means that μ would have to be almost identically equal to $k\Lambda$ in order for the correction to cancel out to such a small value. Such a cancellation would require around 30 digits of agreement between μ and Λ, which as far as we understand, are two unrelated parameters. Many physicists find such a degree of fine-tuning to be unnatural and believe that this is evidence of new physics we do not understand.

I am aware of a few primary solutions. The first is the most hopeful one – that Λ is actually not at these extremely high scales, but that there is new physics *just around the corner*. That is, Λ is of the order of 1 TeV, and there is new physics to discover at accessible energy scales, either at the LHC, its successor, or other precision experiments. One of the most popular solutions to the hierarchy problem is supersymmetry. The idea is that every fermion has a bosonic partner, and every boson has a fermionic partner, whose masses are not exactly equal due to a dynamic breaking of this supersymmetry. Because fermions and bosons contribute with opposite signs to the loop diagram corrections to the Higgs mass, they will tend to cancel out, avoiding fine-tuning. Unfortunately, supersymmetry, which in most regions of its large parameter space is rather clearly observable in collider experiments, has not yet been observed. Supersymmetry at larger mass scales than a few TeV starts to require fine-tuning again.

Another TeV-scale solution is to hypothesize that the Higgs is actually a composite particle. The Higgs mechanism and spontaneous symmetry breaking was first understood in the context of condensed matter systems. In these systems, the symmetry-breaking potential has as its independent variable some order parameter, which takes a nonzero value below a critical temperature. For ferromagnets, for example, this is the total magnetization, which is zero above the Curie temperature and which spontaneously picks a direction below. A bizarre feature of the Standard Model Higgs mechanism is that this order parameter is promoted to itself be a fundamental field, the only fundamental scalar particle. However, perhaps as in the ferromagnet case, the Higgs field is actually just a parameter of ignorance that glosses over more fundamental dynamics, and the Higgs is actually a composite state arising out of the microscopic physics of the symmetry breaking. In this case, we would expect to see deviations from the predictions of the Higgs as a fundamental particle at TeV energy scales in order to avoid fine-tuning.

Of course fine-tuning is not necessarily a problem, even if the Higgs is fundamental. It could be that this is indeed a feature of nature. We have however never observed such unmotivated fine-tuning in any other well-understood system, so we have no reason to expect it here. It could also be that there is some unknown Planck-scale physics that drives the cancellation, but this is difficult to observe. Lastly, some consider the anthropic principle to be the solution. The idea is that if the parameters were different from what we observe, then the universe couldn't have existed as we know it for us to observe them. This idea may or may not require a multiverse such that the parameters could be "rolled" many times in order to achieve the ones that work. I find this solution distasteful because it makes no predictions and is only one step removed from simply surrendering.

The Strong CP Problem

In the discussions of the Poincaré group in section 2.1.1, one point that was skipped was that in addition to the continuous transformations of the group (translations, rotations, and boosts), there are three additional binary discrete group transformations. The first is a parity transformation, P, which takes $\vec{x} \to -\vec{x}$. Because a parity transformation flips a particle's chirality, it turns out that the weak force maximally violates parity symmetry. The next possibility is charge conjugation, in which all particles and antiparticles are swapped. Because parity is violated, many physicists had hoped that the combined symmetry CP, in which $\vec{x} \to -\vec{x}$ and particles are replaced with their antiparticles, would be preserved. However, the weak force was also found to also violate CP, though through relatively small and subtle effects. The last symmetry is time reversal, in which $t \to -t$. The combined transformation CPT must be a valid symmetry for Lorentz invariance to be preserved.

So far, CP violation has only been observed in weak interactions. However, there is room in the Standard Model for the strong force to also violate CP. One can add a term to the

Lagrangian,

$$\mathcal{L}_{CP} = \theta_{\text{QCD}} \frac{g_s^2}{32\pi^2} \epsilon^{\mu\nu\alpha\beta} F^a_{\mu\nu} F^a_{\alpha\beta} \tag{2.20}$$

for θ_{QCD} a constant free parameter, and $F^a_{\mu\nu}$ the gluon field strength tensor, analogous to the electromagnetic field strength [19]. While this term is a total derivative and thus always drops out in perturbation theory, it still can have non-perturbative effects that should lead to CP-violating effects at low-energy.

One of the primary effects this term induces is a nonzero electric dipole moment for the neutron. Experimental constraints however are extremely tight, and it is observed that $\theta_{\text{QCD}} < 10^{-10}$. Thus we have another question of fine-tuning. Why is this parameter so close to or equal to zero, when in the Standard Model it is a free parameter?

Many conclude that something must drive it to be small. One of the most popular explanations is to include axions, in which θ_{QCD} is promoted to a field and is driven to zero via a spontaneous symmetry breaking. This mechanism predicts a light, weakly coupled axion particle that has not yet been observed.

Matter-Antimatter Asymmetry

A perhaps surprisingly difficult fact to reconcile with modern particle physics is that the universe appears to be made out of essentially only matter, with very little antimatter. Large quantities of antimatter would lead to gamma-ray annihilation signatures that are not observed. Because we observe this asymmetry throughout the history of the universe, there must be some process in the early universe that allowed matter to dominate over antimatter. In 1967, Andrei Sakharov listed three conditions, the now eponymous Sakharov conditions, that must be satisfied in order to create a matter-antimatter imbalance [22]:

1. Baryon number violation

2. C and CP symmetry violation

3. A departure from thermal equilibrium

Somewhat tantalizingly, the Standard Model allows for all three of these, but not *enough* of any of them. Baryon number violation is allowed in the Standard Model through non-perturbative effects at high temperatures called sphaelerons. C and CP symmetry are both violated by the weak interaction, and the spontaneous symmetry breaking from the Higgs mechanism allows for a discontinuous phase transition, though it is smooth in the Standard Model. Tweaking any one of these can allow for a larger matter-antimatter asymmetry more consistent with what is observed. There are of course many models that satisfy these conditions with physics beyond the Standard Model instead of relying solely on the small Standard-Model-like effects.

Neutrino Mass

A relatively recent problem with the Standard Model was the discovery of neutrino oscillations, the observation that neutrinos can oscillate between flavors. A 100% pure beam of muon neutrinos will turn into some fraction of electron neutrinos over some distance. Because neutrinos are always produced via the weak interaction in flavor eigenstates but propogate through free space in mass eigenstates, this implies that the mass and flavor eigenstates of neutrinos are non-diagonal. This can only be the case if all of the neutrino masses are different, which means they must be nonzero. The Standard Model, however, assumes zero neutrino masses.

The question then is why the neutrino masses are so much smaller than the rest of the other SM particles. The strongest constrains on the neutrino masses come from cosmology where, assuming there are no other new light particles, the sum of the neutrino masses must be less than about 0.17 eV. The direct neutrino mass measurement bounds are about an order of magnitude worse than this. Regardless, there is at lease a six-order of magnitude jump between the electron mass and the neutrino mass, which is much larger than any other mass difference in the Standard Model, and which firmly places neutrinos at their own mass scale.

One very appealing explanation for this is the seesaw mechanism. Because the SM left-chiral neutrinos are charged under the weak interaction (before the Higgs mechanism), they must have a Dirac mass term after spontaneous symmetry breaking, $m(\bar{\nu}_L \nu_R + \bar{\nu}_R \nu_L)$. This means that right-chiral neutrinos must exist. However, they would have no charge under any of the Standard Model forces, which means they would be observable only through their oscillations into left-chiral neutrinos. They also would be allowed to have a Majorana mass term, $M\nu_R^T \sigma_2 \nu_R$, for the Pauli matrix σ_2. Majorana masses are only possible for particles with no charges. If both masses are present, then the mass matrix is non-diagonal, and both chiralities can contribute to the observed neutrino states. In particular, if the Dirac mass m is at the electroweak scale, and the Majorana mass M is near the only other known fundamental scales, the GUT scale or Planck scale, then the mass of the light neutrino eigenstate is $m_\nu \approx \frac{m^2}{M}$, which is in the vicinity of the expectations from experiments. The other mass eigenstate is at the scale M. Thus as M increases, m_ν decreases, and this is called the seesaw mechanism.

Dark Matter and Dark Energy

One of the most pressing open questions in particle physics, astronomy, and cosmology is that of dark matter and dark energy. In the early days of cosmology, Fritz Zwicky discovered that galaxy clusters, if they were bound together gravitationally, were moving too fast compared to the expected gravitational pull from all of the observable mass. Later, in the 1970's, Vera Rubin and many others observed that galaxies were rotating faster than possible given the known amount of mass. At the time, it was assumed there was either diffuse gas or perhaps unknown compact objects that explained this missing mass. However, throughout the intervening decades, each of these sources have been ruled out [23]. Thus some non-baryonic unobserved mass, dark matter, remains the only convincing explanation for this missing 80% of the matter content of the universe. None of the Standard Model particles can explain dark matter, so this is the most direct evidence for physics beyond the standard model.

Despite this, no dark matter particle has yet been observed, despite extensive searches. Generically, dark matter particles are expected to be weakly interacting. Two of the most popular dark matter models at the moment are sypersymmetry and axions. In many supersymmetric models, supersymmetric particles carry an additional quantum number, R-parity, that forbids the decay of the lightest supersymmetric partners to standard model particles. If the lightest supersymmetric partner is weakly interacting, its properties align well with that expected for dark matter. Axions, introduced to solve the strong CP problem, can also be dark matter, and this possibility has spurred a variety of axion dark matter searches.

Another more spectacular failure of modern quantum field theory is that of the problem of dark energy. It has been observed that the expansion of the universe is accelerating, which implies the existence of some background energy, a cosmological constant, that has a constant energy density per unit volume, even as the universe expands. This dark energy must make up around 75% of the energy content of the universe, with the remaining 25% being dark matter and normal baryonic matter combined. Quantum field theory allows for zero-point background energy density, but all estimates at the actual density are spectacularly incorrect, generally overestimating the energy by something like 120 orders of magnitude. It is not clear how this issue will be resolved.

Experimental Anomalies

In addition to these theoretical, astronomical, and cosmological tensions discussed above, there are a handful of terrestrial particle experiments where weak discrepancies have been observed. None of the results are definitive, with follow-up studies either planned or in progress.

One relatively simple experiment to explain is the measurement of the muon g-factor, which measures how strongly the muon interacts with a magnetic field. At lowest order, the expectation is that $g_\mu = 2$. However, one can add loops to the Feynman diagrams, which leads to small corrections, causing g_μ to be about 1% larger than 2. This particular observable posses the rare combination of being precisely calculable and precisely measurable. From the theoretical side, the current state-of-the-art predictions use high orders of perturbation theory and include all Standard Model particles. It is thus an inclusive test of the Standard Model. Any other new particles would appear in the loops and cause slight deviations. The current best calculation for the muon g-factor is

$$a_\mu^{\mathrm{SM}} = \frac{g_\mu^{\mathrm{SM}} - 2}{2} = 116591810(43) \times 10^{-11} \tag{2.21}$$

with the astonishingly small uncertainty of just 368 parts per billion, the result of several years of work from 130 physicists [24].

From the experimental side, the muon g-factor was most accurately measured at Brookhaven Lab by storing muons in a ring in a highly uniform magnetic field, and observing the spin procession over many periods. The observed value is

$$a_\mu^{\mathrm{exp}} = \frac{g_\mu^{\mathrm{exp}} - 2}{2} = 116592089(63) \times 10^{-11} \tag{2.22}$$

which is a 3.7σ discrepancy from the current theoretical estimate [25]. In order to resolve this discrepancy, the muon g-factor is being measured again at Fermilab with reduced experimental uncertainties, while the theory community continues to refine the precise estimate. Results are expected soon.

Another set of anomalies that have been gaining traction are the anomalies in the flavor sector, particularly in B meson decays. Experiments such as BaBar, Belle, and LHCb have been seeing various consistent slight departures from the Standard Model, at the level of a few sigma. One set of deviations is in the decays $B \to D^{(*)}\ell\nu_\ell$. It appears that the τ lepton decays are slightly more likely than expected, relative to the decays into e and μ. Another independent set of anomalies has been observed in $B \to K\mu^+\mu^-$, which is observed to be slightly more likely than expected relative to the analogous decay to electrons [26]. These

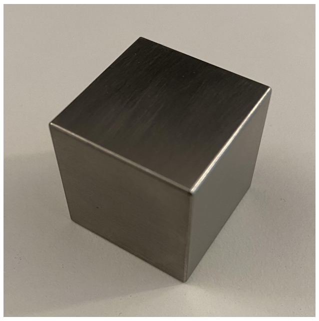


Figure 2.2: A 1” × 1” × 1” cube of 95% pure Tungsten, weighing approximately 0.3 kg (0.65 lb), with a melting point at 3687 K.

observations all point to some level of lepton flavor universality violation, indicating unequal properties of different leptons. However, anomalies in the flavor sector have in the past tended to disappear, due to better understanding of theoretical uncertainties with hadronic form factors or through a better understanding of experimental uncertainties, particularly with respect to systematic differences between measuring different leptons. The results are still inconclusive, but LHCb and Belle II will be collecting much larger datasets that should hopefully resolve this issue.

2.3 Melting Mass: The Electroweak Phase Transition

Now that we have built the theoretical foundations, let us begin to study one of the central questions addressed by the experimental results in this thesis. We can introduce the problem with a thought experiment. Suppose we start with a cube of tungsten (Figure 2.2) at room temperature and heat it up to arbitrarily high temperatures. What happens?

The list of phase transitions is summarized in Table 2.2. The first phase transitions are relatively familiar, first melting into a liquid, then boiling into a gas, then ionizing into a plasma. After this, the typical kinetic energies are large enough to overcome the nuclear coulomb barrier, and the nucleus will begin to disintegrate into a gas of free hadrons (mostly protons, neutrons, and pions). At this point, the element is no longer well defined, and the state is independent of the initial material.

Temperature	Energy	Transition
295 K	25 meV	Room temperature
3687 K	317 meV	Melting point [27]
5829 K	502 meV	Boiling point [27]
9×10^4 K	7.86 eV	Ionization of outermost electron [28]
9×10^8 K	81 keV	Ionization of innermost electron [28]
10^{11} K	10 MeV	Nuclear disintegration to hadron gas [29]
1.7×10^{12} K	150 MeV	Hadron melting to quark gluon plasma [30]
1.9×10^{15} K	160 GeV	Mass melting to electroweak plasma [31]

Table 2.2: The known phase transitions for a block of Tungsten starting at room temperature.

Continuing to increase the temperature beyond this, the interaction energy between particles is so high that the strong coupling constant becomes weaker, and hadrons dissolve into a plasma of quarks and gluons. This is the highest-temperature state of matter that can reach thermal equilibrium in the lab (albeit for extremely short periods of time) by colliding nuclei at high energies at RHIC and the LHC.

Beyond this, it is not possible to probe the next phase transition directly. Instead, we must study the interactions that lead to the phase transition, analogous to studying the attractive potential between Tungsten atoms in a solid lattice in order to understand the solid to liquid transition. The interaction that governs the phase transition is the Higgs potential, which at zero-temperature, takes the form $V(\Phi) = -\mu\Phi^2 + \lambda\Phi^4$. This zero-temperature result can be extended to finite temperatures by considering an effective potential $V^{\text{eff}}(\Phi, T)$ with an effective mass term $m^{\text{eff}}(T)$, analogous to the Landau theory of phase transitions. The phase transition occurs when $m^{\text{eff}}(T_c) = 0$, corresponding to the point at which the potential shifts from a minimum at zero for high temperatures to a nonzero minimum at lower temperatures. Computing this for the Standard Model, the transition temperature is expected to be approximately $T_c = 160$ GeV.

Above this transition temperature, the masses of the fundamental particles drop to zero, and the electromagnetic and weak forces, which are separated at low temperatures, unify into a massless $U(1) \times SU(2)$ gauge force. Thus the electroweak plasma that exists above this temperature is a state of matter composed of massless fermions and bosons interacting under a somewhat different set of forces. This state of matter is expected to have ceased to exist around 1 picosecond (10^{-12} s) after the big bang.

Given the observed Higgs mass and inferred Higgs self-interaction, for the Standard Model Higgs potential, the transition is actually a smooth cross-over, rather than a second or first order phase transition. The Standard Model Higgs potential, however, is merely the simplest possible symmetry-breaking potential. There may be modifications to it that could lead to this transition being first order. Deviations of at least roughly 50% in λ, the Higgs self-coupling parameter, can lead to a strongly first-order phase transition. In this case, the electroweak phase transition in the early universe occurs via a bubbling, like boiling water, and the bubble edge provides an out-of-equilibrium period for the generation of the matter-antimatter asymmetry. Such a mechanism is called electroweak baryogenesis.

There are a few reasons to expect that the Higgs potential might be different from the Standard Model. First, there are a myriad of ways to add new particles that slightly change the low-temperature Higgs potential. It is not very constrained theoretically. Second, the Higgs boson itself may be different than expected. In the Standard Model, there is no microscopic dynamical explanation of the Higgs potential. It is merely introduced as an order parameter to induce a phase transition, and this order parameter is promoted to a fundamental field. The same form of potential describes the phase transition for the loss of magnetism of a ferromagnet through the Curie temperature, but in this case the potential represent spin alignment with local magnetic fields. There is no such analog for the Higgs boson. Perhaps in fact the Higgs is some composite state of other particles that induce this potential dynamically in a way that we do not yet understand. Such a mechanism would be analogous to the generation of the proton mass via QCD's chiral symmetry breaking, which induces massive pions, the analog to the Higgs.

2.3.1 Measuring the Higgs Potential

Constraining and eventually measuring the Higgs potential is one of the main goals of the LHC. The potential is sensitive to new physics, and constraining it can rule out many broad classes of models.

Unfortunately, observing this phase transition directly is (and will remain, for the foreseeable future) impossible. Instead, we must rely on indirect probes.

One indirect probe is actually through graviational waves. If the Higgs self-coupling were large enough to induce a strongly discontinuous phase transition in the early universe, such discontinuities unleash enough energy to produce substantial amounts of gravitational radiation that could potentially be observed with future gravitational wave detectors, such as LISA [32].

The other method to measure the Higgs potential is through scattering experiments, where we search for processes with Feynman diagrams that have a vertex of three or four Higgs, shown in Figure 2.3. Such diagrams correspond directly to the cubic and quartic terms of the Higgs potential. For the SM Higgs potential $V(H) = -\mu^2 H^2 + c_4 H^4$, we can expand around the minimum to get the interactions of the real Higgs particle,

$$V(h) = -\frac{1}{2} m_h^2 h^2 - \lambda v h^3 - \frac{1}{4} \lambda h^4 \tag{2.23}$$

where $v = \frac{\mu}{c_4} = 246$ GeV, $m_h = \sqrt{2}\mu = 125$ GeV, and $\lambda = \frac{1}{2}\frac{m_h^2}{v^2} = 0.13$. Note that v is related to the Fermi constant of weak interactions and can be measured in muon decays, and the Higgs mass m_h has recently been measured by the LHC. These two constants fix λ, which determines the interaction strength of the higher order terms in the potential. In other words, the quadratic term has already been measured at the LHC, and now we need to test the cubic and quartic terms to see if they are consistent with the Standard Model expectations.

Unfortunately, the Higgs quartic interaction is not possible to measure at the LHC.

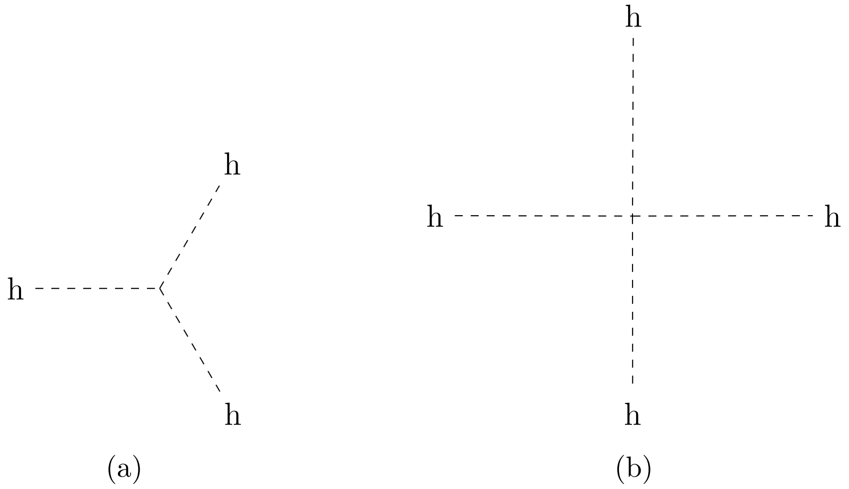


Figure 2.3: (a) The Higgs trilinear self-coupling, with vertex factor λv (b) The Higgs quartic self-coupling, with vertex factor $\frac{1}{4}\lambda$

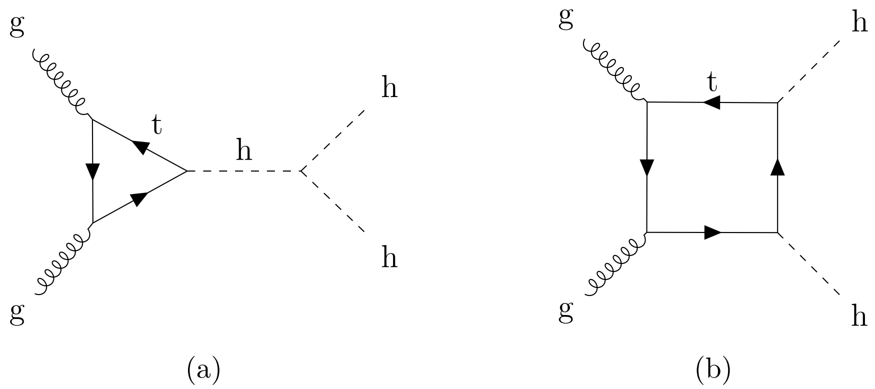


Figure 2.4: The two primary production modes of di-Higgs at the LHC, the triangle diagram (a), and the box diagram (b).

In fact, it is only barely measurable (with roughly 50% uncertainty) at the most ambitious future colliders. We will therefore restrict ourselves to the Higgs trilinear self-coupling, $\lambda v h^3$.

The most direct way to probe the Higgs trilinear self-coupling is through the pair production of Higgs bosons. Two ways this can happen are shown in the Feynman diagrams in Figure 2.4. Both of these diagrams involve initial state gluons interacting through a top loop, and are together called the gluon-gluon fusion production mode, ggF. Unfortunately, these two diagrams contribute with opposite signs to the amplitude and thus exhibit interference, substantially reducing the cross-section.

One way to parameterize deviations from the Standard Model is to measure how different

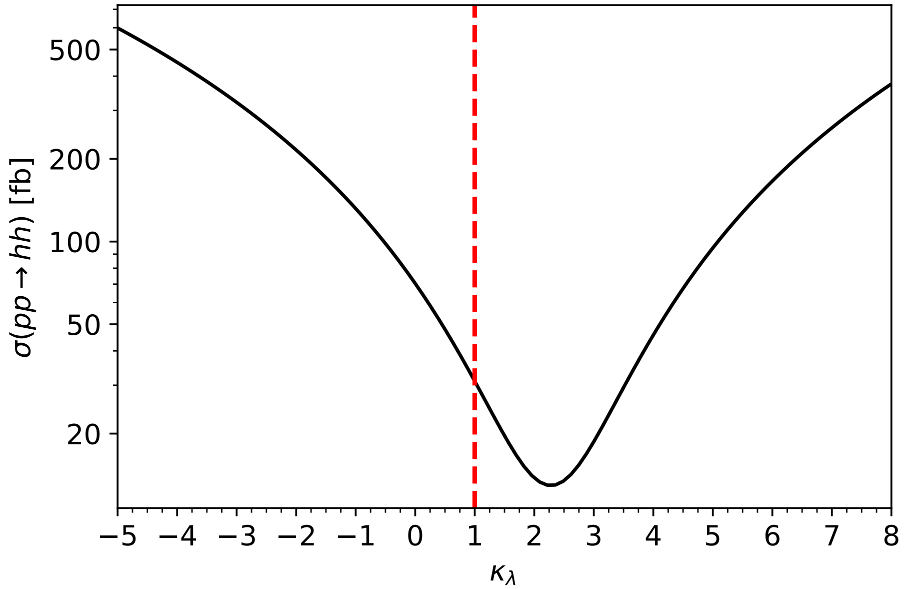


Figure 2.5: The total di-Higgs cross-section as a function of the trilinear Higgs self-coupling parameter, $\kappa_\lambda = \frac{\lambda}{\lambda_{\text{SM}}}$. The red vertical line indicates the Standard Model value.

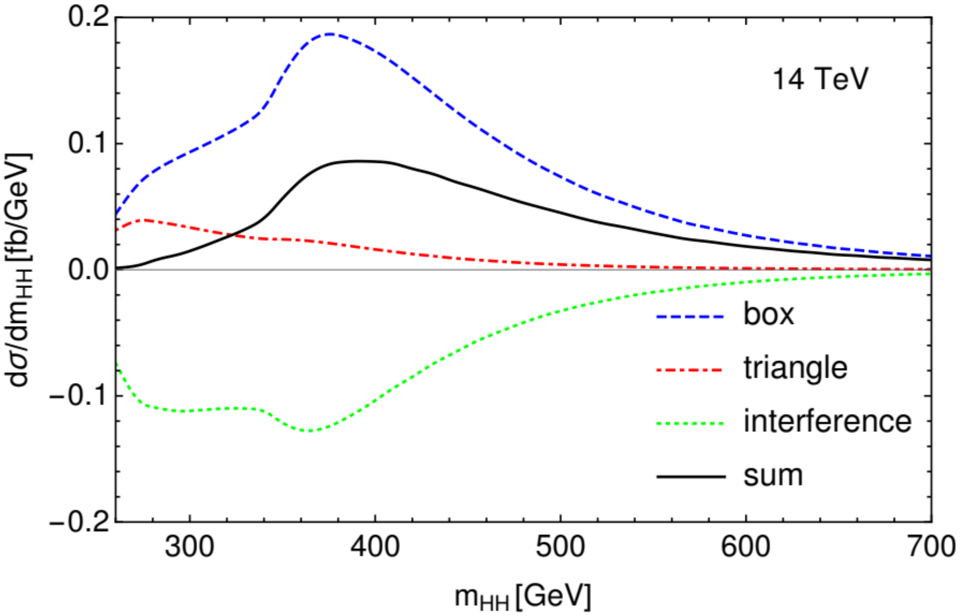


Figure 2.6: The interference in the di-Higgs cross section as a function of the di-Higgs invariant mass m_{hh}, $\frac{d\sigma}{dm_{hh}} \propto |\mathcal{M}_{\text{box}} + \mathcal{M}_{\text{triangle}}|^2$ = (box) + (triangle) + (interference). Since the goal is to be sensitive to κ_λ, we are interested in the triangle diagram, which has the largest cross-section at low values of m_{hh}, even though the overall cross-section peaks at 400 GeV [1].

λ is from the expected value, which will henceforth be parameterized as $\kappa_\lambda = \frac{\lambda}{\lambda_{\text{SM}}}$. The di-Higgs cross-section as a function of κ_λ is shown in Figure 2.5. One can see that in the Standard Model, nature has unfortunately left us with a near-minimal cross-section, making the di-Higgs process difficult to detect. On the other hand, deviations in κ_λ can lead to large enhancements of the cross section, allowing limits to be set by the non-observation of di-Higgs.

The interference in the production diagrams also leads to a kinematic-dependent sensitivity of the cross-section to κ_λ. In particular, as shown in Figure 2.6, the triangle diagram dominates at low m_{hh}, while the box diagram dominates at high m_{hh}, where $m_{hh}^2 = (E_{h1} + E_{h2})^2 - |\vec{p}_{h1} + \vec{p}_{h2}|^2$ is the invariant mass of the di-Higgs system. This means that in general, to maximize sensitivity to κ_λ, we need to focus attention on di-Higgs systems with low invariant mass.

Another important production mode for the di-Higgs system is vector boson fusion (VBF), shown in Figure 2.7. This channel has a cross section around 10 times smaller than the ggF channel and will therefore be very difficult to observe. However, it provides a unique sensitivity to the $VVhh$ vertex, described by the parameter c_{2V}. Due to the interference of the VBF diagrams, small changes in c_{2V} can actually lead to massive enhancements in the cross-section, so while the Standard Model process is currently unobservable, relatively tight limits can be set on c_{2V} already.

One of the biggest experimental challenges in searching for di-Higgs is the very large diversity of final states, which are listed in Table 2.3. One potentially naïve approach is to simply study the state with the largest branching ratio, the $hh \to 4b$ decay, which is the subject of this thesis. Unfortunately this final state has a large and difficult to model background from QCD processes, the modeling of which forms a core part of this analysis.

To get an estimate of sensitivity given the current data collected at the LHC, the total $hh \to 4b$ cross-section at 13 TeV is $\sigma(pp \to hh) \times BR(hh \to 4b) = 10.5$ fb. Given a total (useable) luminosity of 127 fb^{-1} collected in Run 2, this leads to roughly 1300 events. We

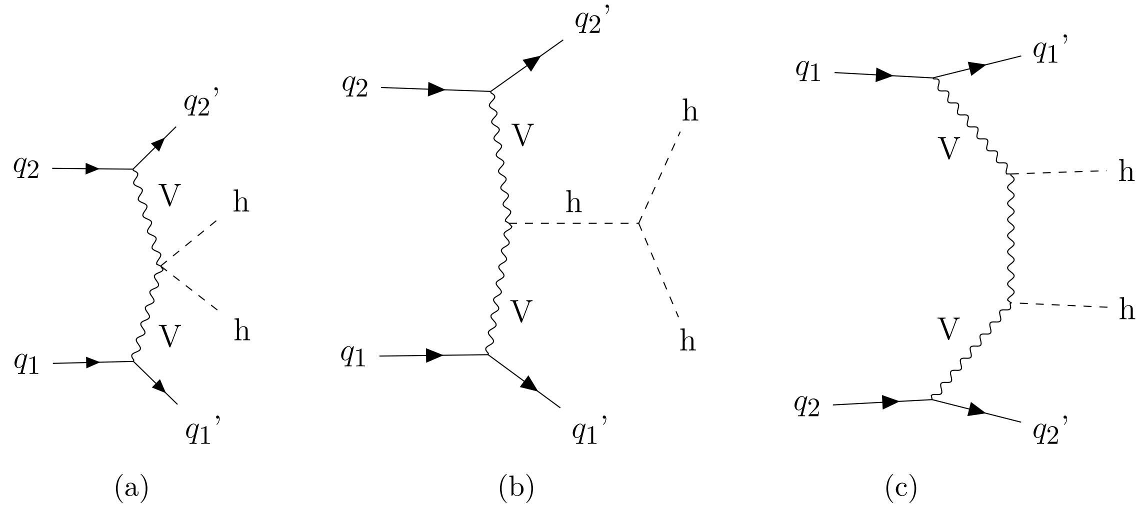


Figure 2.7: The tree-level diagrams for vector boson fusion (VBF) production of di-Higgs. Diagram (a) provides a unique sensitivity to the di-Higgs couplings to the vector bosons, $V = W, Z$. Diagram (b) can provide complimentary sensitivity to the Higgs trilinear self-coupling. The Higgs vertices in diagram (c) are better constrained by single-Higgs measurements, such as Higgs-strahlung.

will see below that the total acceptance is around 1%, which means that we should have around 13 total signal events. This count will be overwhelmed by the large backgrounds, which means that to have any sensitivity, we need high precision estimates of the background.

Regardless, we can still perform a search for an enhanced Standard Model cross section, or an altered Higgs self-coupling κ_λ. Thus, the two major figures of merit will be the 95% confidence exclusion limit on the Standard Model cross-section (the signal cross-section at which we are 95% confident that we would have observed a signal), and the 95% confidence exclusion limit on κ_λ (the range of κ_λ values that we would have been 95% confident that we would have observed, due to the enhanced cross-section). Measuring these is the goal of the rest of this thesis.

Rank	Decay Mode	BR
1	bbbb	33.9 %
2	bbjjjj	12.8 %
3	bbgg	9.53 %
4	bbℓjj + MET	8.63 %
5	bb$\tau_h\ell$ + MET	4.29 %
6	bbcc	3.36 %
7	bb$\tau_h\tau_h$	3.06 %
8	bb$\ell\ell$ + MET	2.65 %
9	bb$\tau_h$jj + MET	2.47 %
10	ggjjjj	1.80 %
11	ℓjjjjjj + MET	1.63 %
12	ggℓjj + MET	1.21 %
13	jjjjjjjj	1.21 %
14	$\tau_h\ell$jjjj + MET	1.13 %
15	$\ell\ell$jjjj + MET	1.05 %
16	bbjj + MET	0.852 %
17	gggg	0.670 %
18	$\tau_h\ell\ell$jj + MET	0.643 %
19	ccjjjj	0.635 %
20	gg$\tau_h\ell$ + MET	0.604 %
21	$\tau_h\tau_h$jjjj	0.579 %
22	$\tau_h\tau_h\ell$jj + MET	0.567 %
23	ccgg	0.473 %
24	$\tau_h$jjjjjj + MET	0.467 %
25	gg$\tau_h\tau_h$	0.431 %
26	ccℓjj + MET	0.428 %
27	gg$\ell\ell$ + MET	0.372 %
28	gg$\tau_h$jj + MET	0.348 %
29	$\ell\ell\ell$jj + MET	0.337 %
30	bb$\ell\ell$jj	0.311 %
31	bb$\gamma\gamma$	0.264 %
32	$\tau_h\tau_h\ell\ell$ + MET	0.262 %
33	cc$\tau_h\ell$ + MET	0.213 %
34	$\tau_h\tau_h\tau_h\ell$ + MET	0.204 %
35	$\tau_h\ell\ell\ell$ + MET	0.168 %
36	jjjjjj + MET	0.161 %
37	bb$\tau_h\tau_h$ + MET	0.154 %
38	cc$\tau_h\tau_h$	0.152 %
39	cc$\ell\ell$ + MET	0.131 %
40	bbγjj	0.125 %
41	cc$\tau_h$jj + MET	0.123 %
42	bb + MET	0.122 %
43	ggjj + MET	0.120 %

Table 2.3: All decay modes of HH with Higgs decays through bb, ZZ, WW, $\tau\tau$, $\gamma\gamma$, cc, gg, Zγ, and $\mu\mu$. ℓ indicates e or μ with equal likelihood. τ_h indicates a hadronically decaying τ. MET indicates at least one neutrino of any flavor. j indicates a jet of any flavor from Z or W decays. g indicates gluons, which always arise from the $h \to gg$ decay and not from the decays of other particles or intitial and final state radiation. Initial state radiation and final state radiation are not considered.

Rank	Decay Mode	BR
44	$\tau_h\tau_h\tau_h$jj + MET	0.117 %
45	ℓjjjj + MET	0.109 %
46	cccc	0.0835 %
47	$\tau_h\tau_h$jjjj + MET	0.0751 %
48	bb$\tau_h\ell$jj + MET	0.0719 %
49	$\tau_h\tau_h\tau_h\tau_h$	0.0693 %
50	bb$\tau_h\tau_h$jj	0.0661 %
51	$\ell\ell$jjjjjj	0.0587 %
52	$\tau_h\ell$jj + MET	0.0542 %
53	$\ell\ell\ell$ + MET	0.0517 %
54	$\gamma\gamma$jjjj	0.0500 %
55	gg$\ell\ell$jj	0.0437 %
56	ccjj + MET	0.0423 %
57	$\ell\ell$jjjj + MET	0.0421 %
58	$\tau_h\tau_h$jj + MET	0.0406 %
59	gg$\gamma\gamma$	0.0372 %
60	bbγ + MET	0.0357 %
61	$\ell\ell$jj + MET	0.0339 %
62	$\gamma\gamma\ell$jj + MET	0.0337 %
63	$\tau_h$jjjj + MET	0.0311 %
64	jjjj + MET	0.0284 %
65	bb$\mu\mu$	0.0253 %
66	$\tau_h\ell\ell\ell$jj + MET	0.0239 %
67	γjjjjjj	0.0236 %
68	gg$\tau_h\tau_h$ + MET	0.0217 %
69	$\tau_h\ell\ell$jjjj + MET	0.0212 %
70	bb$\ell\ell$jj + MET	0.0195 %
71	ggγjj	0.0175 %
72	gg + MET	0.0172 %
73	$\tau_h\gamma\gamma\ell$ + MET	0.0168 %
74	bb$\ell\ell\ell\ell$	0.0162 %
75	$\gamma\ell$jjjj + MET	0.0159 %
76	ℓjj + MET	0.0155 %
77	cc$\ell\ell$jj	0.0154 %
78	$\tau_h\tau_h\ell\ell$jj	0.0141 %
79	$\tau_h\ell$jjjjjj + MET	0.0136 %
80	$\ell\ell\ell$jj + MET	0.0131 %
81	cc$\gamma\gamma$	0.0131 %
82	bb$\gamma\ell\ell$	0.0130 %
83	$\tau_h\tau_h$jjjjjj	0.0125 %
84	$\tau_h\tau_h\gamma\gamma$	0.0119 %
85	$\tau_h\tau_h\ell$jjjj + MET	0.0110 %
86	$\gamma\gamma\ell\ell$ + MET	0.0103 %
87	gg$\tau_h\ell$jj + MET	0.0101 %
88	$\tau_h\gamma\gamma$jj + MET	9.64e-3 %
89	gg$\tau_h\tau_h$jj	9.30e-3 %
90	$\tau_h\tau_h\ell\ell$jj + MET	8.88e-3 %
91	γjjjj + MET	8.31e-3 %
92	$\tau_h\gamma\ell$jj + MET	7.98e-3 %

Table 2.3, continued

Rank	Decay Mode	BR
93	bb$\tau_h\ell\ell\ell$ + MET	7.97e-3 %
94	$\tau_h\ell$ + MET	7.73e-3 %
95	cc$\tau_h\tau_h$ + MET	7.65e-3 %
96	$\tau_h\tau_h\tau_h\ell$jj + MET	7.63e-3 %
97	$\tau_h\tau_h\tau_h\tau_h$ + MET	7.15e-3 %
98	bb$\tau_h\tau_h\ell\ell$	6.90e-3 %
99	ccγjj	6.19e-3 %
100	cc + MET	6.06e-3 %
101	$\tau_h\tau_h$ + MET	5.79e-3 %
102	$\tau_h\tau_h\gamma$jj	5.64e-3 %
103	$\gamma\ell\ell$jj + MET	5.22e-3 %
104	ggγ + MET	5.02e-3 %
105	$\mu\mu$jjjj	4.78e-3 %
106	$\ell\ell$ + MET	4.76e-3 %
107	$\tau_h\gamma$jjjj + MET	4.55e-3 %
108	$\gamma\ell$jj + MET	4.55e-3 %
109	$\tau_h$jj + MET	4.45e-3 %
110	$\ell\ell\ell\ell$jjjj	3.78e-3 %
111	$\ell\ell$jjjjjj + MET	3.69e-3 %
112	cc$\tau_h\ell$jj + MET	3.57e-3 %
113	gg$\mu\mu$	3.56e-3 %
114	$\gamma\gamma$jj + MET	3.39e-3 %
115	cc$\tau_h\tau_h$jj	3.28e-3 %
116	$\ell\mu\mu$jj + MET	3.23e-3 %
117	$\gamma\ell\ell$jjjj	3.03e-3 %
118	bb$\tau_h\gamma\ell$ + MET	3.01e-3 %
119	$\tau_h\tau_h\tau_h\tau_h$jj	2.99e-3 %
120	bb$\tau_h\tau_h\gamma$	2.77e-3 %
121	gg$\ell\ell$jj + MET	2.75e-3 %
122	$\tau_h\tau_h\tau_h$jjjj + MET	2.41e-3 %
123	$\ell\ell\ell\ell$jj + MET	2.33e-3 %
124	gg$\ell\ell\ell\ell$	2.28e-3 %
125	$\tau_h\gamma\ell$ + MET	2.27e-3 %
126	bb$\ell\ell\ell\ell$ + MET	2.10e-3 %
127	$\tau_h\ell\ell\ell$jjjj + MET	1.86e-3 %
128	gg$\gamma\ell\ell$	1.83e-3 %
129	ccγ + MET	1.77e-3 %
130	$\gamma\ell\ell\ell$jj + MET	1.76e-3 %
131	$\tau_h\tau_h\gamma$ + MET	1.70e-3 %
132	$\tau_h\ell\ell\ell$jj + MET	1.68e-3 %
133	$\tau_h\tau_h\ell\ell$jjjj	1.61e-3 %
134	$\tau_h\ell\mu\mu$ + MET	1.61e-3 %
135	bb$\tau_h\tau_h\tau_h\ell$ + MET	1.60e-3 %
136	jj + MET	1.53e-3 %
137	$\tau_h\ell\ell\ell\ell$ + MET	1.47e-3 %
138	$\gamma\ell\ell$ + MET	1.42e-3 %
139	$\tau_h\tau_h\ell\ell\ell$jj + MET	1.34e-3 %
140	$\tau_h\gamma$jj + MET	1.30e-3 %
141	bb$\tau_h\tau_h\ell\ell$ + MET	1.30e-3 %

Table 2.3, continued

Rank	Decay Mode	BR
142	cc$\mu\mu$	1.26e-3 %
143	$\gamma\gamma\ell\ell$jj	1.24e-3 %
144	$\tau_h\tau_h\mu\mu$	1.15e-3 %
145	gg$\tau_h\ell\ell\ell$ + MET	1.12e-3 %
146	$\tau_h\gamma\ell\ell\ell$ + MET	9.99e-4 %
147	$\ell\ell\mu\mu$ + MET	9.89e-4 %
148	gg$\tau_h\tau_h\ell\ell$	9.71e-4 %
149	cc$\ell\ell$jj + MET	9.70e-4 %
150	$\tau_h\tau_h\ell\ell\ell\ell$ + MET	9.62e-4 %
151	$\tau_h\tau_h\tau_h\ell\ell\ell$ + MET	9.61e-4 %
152	$\tau_h\mu\mu$jj + MET	9.24e-4 %
153	$\tau_h\gamma\ell\ell$jj + MET	8.88e-4 %
154	bb$\gamma\ell\ell$ + MET	8.18e-4 %
155	cc$\ell\ell\ell\ell$	8.05e-4 %
156	bb$\tau_h\tau_h\tau_h\tau_h$	7.34e-4 %
157	$\tau_h\tau_h\ell\ell\ell\ell$	7.33e-4 %
158	$\ell\ell\ell\ell\ell\ell$ + MET	7.15e-4 %
159	$\tau_h\gamma\ell$jjjj + MET	7.01e-4 %
160	γjj + MET	6.74e-4 %
161	cc$\gamma\ell\ell$	6.46e-4 %
162	$\tau_h\tau_h\gamma$jjjj	6.45e-4 %
163	$\tau_h\tau_h\gamma\gamma$ + MET	6.03e-4 %
164	$\tau_h\tau_h\gamma\ell\ell$	5.89e-4 %
165	$\gamma\ell\ell\ell$ + MET	5.50e-4 %
166	$\gamma\gamma\gamma\gamma$	5.15e-4 %
167	$\tau_h\tau_h\tau_h\ell\ell$jj + MET	5.02e-4 %
168	$\ell\ell\ell$jjjj + MET	4.89e-4 %
169	$\gamma\gamma\gamma$jj	4.87e-4 %
170	$\gamma\gamma$ + MET	4.85e-4 %
171	$\tau_h\tau_h\gamma\ell$jj + MET	4.62e-4 %
172	gg$\tau_h\gamma\ell$ + MET	4.23e-4 %
173	cc$\tau_h\ell\ell\ell$ + MET	3.96e-4 %
174	gg$\tau_h\tau_h\gamma$	3.89e-4 %
175	$\tau_h\tau_h\gamma\ell\ell$ + MET	3.72e-4 %
176	$\tau_h\tau_h\tau_h\ell$jjjj + MET	3.71e-4 %
177	$\tau_h\tau_h\gamma$jj + MET	3.54e-4 %
178	cc$\tau_h\tau_h\ell\ell$	3.43e-4 %
179	$\tau_h\tau_h\tau_h\gamma\ell$ + MET	3.19e-4 %
180	$\mu\mu$jj + MET	3.19e-4 %
181	$\tau_h\tau_h\tau_h\tau_h\ell\ell$	3.12e-4 %
182	$\tau_h\tau_h\ell\ell$jjjj + MET	3.03e-4 %
183	gg$\ell\ell\ell$ + MET	2.96e-4 %
184	$\tau_h\gamma\gamma\ell$jj + MET	2.86e-4 %
185	$\tau_h\tau_h\gamma\gamma$jj	2.63e-4 %
186	gg$\tau_h\tau_h\tau_h\ell$ + MET	2.24e-4 %
187	$\tau_h\tau_h\tau_h\tau_h\ell\ell$ + MET	2.07e-4 %
188	$\gamma\ell\ell$jjjj + MET	1.90e-4 %
189	gg$\tau_h\tau_h\ell\ell$ + MET	1.83e-4 %
190	$\tau_h\tau_h\tau_h\tau_h$jjjj	1.71e-4 %

Table 2.3, continued

Rank	Decay Mode	BR
191	$\tau_h\tau_h\tau_h\tau_h$jj + MET	1.60e-4 %
192	$\tau_h\tau_h\tau_h\tau_h\ell$jj + MET	1.52e-4 %
193	cc$\tau_h\gamma\ell$ + MET	1.49e-4 %
194	$\gamma\gamma\gamma$ + MET	1.39e-4 %
195	cc$\tau_h\tau_h\gamma$	1.37e-4 %
196	$\tau_h\tau_h\tau_h\tau_h\gamma$	1.25e-4 %
197	$\tau_h\tau_h\tau_h\tau_h\tau_h\ell$ + MET	1.22e-4 %
198	$\ell\ell\mu\mu$jj	1.16e-4 %
199	gg$\gamma\ell\ell$ + MET	1.15e-4 %
200	MET	1.10e-4 %
201	cc$\ell\ell\ell$ + MET	1.04e-4 %
202	gg$\tau_h\tau_h\tau_h\tau_h$	1.03e-4 %
203	$\tau_h\tau_h\tau_h\gamma$jj + MET	1.01e-4 %
204	$\gamma\gamma\mu\mu$	9.88e-5 %
205	$\gamma\ell\ell\ell\ell$jj	8.96e-5 %
206	cc$\tau_h\tau_h\tau_h\ell$ + MET	7.92e-5 %
207	$\gamma\gamma\ell\ell$jj + MET	7.77e-5 %
208	$\ell\ell\ell\ell\ell\ell$jj	7.44e-5 %
209	cc$\tau_h\tau_h\ell\ell$ + MET	6.46e-5 %
210	$\gamma\gamma\ell\ell\ell\ell$	6.45e-5 %
211	γ + MET	6.42e-5 %
212	$\tau_h\ell\ell\ell\ell\ell$jj + MET	5.83e-5 %
213	$\tau_h\tau_h\mu\mu$ + MET	5.76e-5 %
214	$\gamma\gamma\gamma\ell\ell$	5.08e-5 %
215	$\tau_h\tau_h\ell\ell\ell\ell$jj	4.75e-5 %
216	$\gamma\mu\mu$jj	4.66e-5 %
217	$\mu\mu$ + MET	4.56e-5 %
218	$\tau_h\gamma\ell\ell\ell$jj + MET	4.40e-5 %
219	cc$\gamma\ell\ell$ + MET	4.06e-5 %
220	$\tau_h\tau_h\gamma\ell\ell$jj	3.81e-5 %
221	cc$\tau_h\tau_h\tau_h\tau_h$	3.64e-5 %
222	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h$	3.32e-5 %
223	$\tau_h\gamma\gamma\ell\ell\ell$ + MET	3.17e-5 %
224	$\tau_h\tau_h\gamma\gamma\ell\ell$	2.74e-5 %
225	$\tau_h\ell\mu\mu$jj + MET	2.69e-5 %
226	$\tau_h\tau_h\tau_h\tau_h\tau_h$jj + MET	2.68e-5 %
227	$\tau_h\tau_h\mu\mu$jj	2.47e-5 %
228	$\tau_h\tau_h\tau_h\ell\ell\ell$jj + MET	2.43e-5 %
229	$\tau_h\tau_h\ell\ell\ell\ell$jj + MET	1.88e-5 %
230	$\ell\ell\ell\ell\ell$jj + MET	1.49e-5 %
231	$\gamma\mu\mu$ + MET	1.33e-5 %
232	$\tau_h\gamma\gamma\gamma\ell$ + MET	1.17e-5 %
233	$\gamma\ell\ell\ell\ell$jj + MET	1.16e-5 %
234	$\tau_h\tau_h\gamma\gamma\gamma$	1.08e-5 %
235	$\tau_h\tau_h\tau_h\tau_h\ell\ell$jj	1.01e-5 %
236	$\tau_h\tau_h\tau_h\gamma\ell$jj + MET	8.81e-6 %
237	$\gamma\gamma\ell\ell\ell\ell$ + MET	8.36e-6 %
238	$\ell\ell\mu\mu$jj + MET	7.30e-6 %
239	$\tau_h\tau_h\gamma\ell\ell$jj + MET	7.18e-6 %

Table 2.3, continued

Rank	Decay Mode	BR
240	$\tau_h\tau_h\tau_h\tau_h\gamma$ + MET	6.69e-6 %
241	$\tau_h\tau_h\tau_h\gamma\gamma\ell$ + MET	6.34e-6 %
242	$\ell\ell\ell\mu\mu$	6.06e-6 %
243	$\tau_h\tau_h\gamma\gamma\ell\ell$ + MET	5.17e-6 %
244	$\gamma\ell\ell\mu\mu$	4.87e-6 %
245	$\mu\mu\mu\mu$	4.73e-6 %
246	$\tau_h\tau_h\tau_h\tau_h\gamma$jj	4.05e-6 %
247	$\gamma\gamma\gamma\ell\ell$ + MET	3.19e-6 %
248	$\tau_h\tau_h\tau_h\tau_h\ell\ell$jj + MET	3.17e-6 %
249	$\gamma\ell\ell\ell\ell\ell\ell$	3.12e-6 %
250	$\tau_h\ell\ell\ell\mu\mu$ + MET	2.98e-6 %
251	$\tau_h\tau_h\tau_h\tau_h\gamma\gamma$	2.92e-6 %
252	$\tau_h\tau_h\ell\ell\mu\mu$	2.58e-6 %
253	$\tau_h\gamma\ell\ell\ell\ell\ell$ + MET	2.44e-6 %
254	$\tau_h\tau_h\tau_h\tau_h\tau_h\ell$jj + MET	2.34e-6 %
255	$\tau_h\ell\ell\ell\ell\ell\ell\ell$ + MET	2.16e-6 %
256	$\tau_h\tau_h\gamma\ell\ell\ell\ell$	1.99e-6 %
257	$\ell\ell\ell\ell\ell\ell\ell\ell$	1.94e-6 %
258	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h$ + MET	1.67e-6 %
259	$\tau_h\tau_h\ell\ell\ell\ell\ell\ell$	1.65e-6 %
260	$\tau_h\tau_h\tau_h\ell\ell\ell\ell\ell$ + MET	1.40e-6 %
261	$\tau_h\gamma\ell\mu\mu$ + MET	1.13e-6 %
262	$\tau_h\tau_h\gamma\mu\mu$	1.04e-6 %
263	$\tau_h\tau_h\ell\ell\ell\ell\ell\ell$ + MET	1.03e-6 %
264	$\tau_h\tau_h\tau_h\gamma\ell\ell\ell$ + MET	1.02e-6 %
265	$\tau_h\tau_h\gamma\ell\ell\ell\ell$ + MET	7.89e-7 %
266	$\ell\ell\ell\ell\mu\mu$ + MET	7.86e-7 %
267	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h$jj	7.16e-7 %
268	$\gamma\ell\ell\ell\ell\ell\ell$ + MET	6.25e-7 %
269	$\tau_h\tau_h\tau_h\ell\mu\mu$ + MET	5.96e-7 %
270	$\ell\ell\ell\ell\ell\ell\ell\ell$ + MET	5.36e-7 %
271	$\tau_h\tau_h\tau_h\tau_h\ell\ell\ell\ell$	5.27e-7 %
272	$\tau_h\tau_h\ell\ell\mu\mu$ + MET	4.86e-7 %
273	$\tau_h\tau_h\tau_h\tau_h\gamma\ell\ell$	4.23e-7 %
274	$\tau_h\tau_h\tau_h\tau_h\ell\ell\ell\ell$ + MET	3.55e-7 %
275	$\gamma\ell\ell\mu\mu$ + MET	3.06e-7 %
276	$\tau_h\tau_h\tau_h\tau_h\tau_h\ell\ell\ell$ + MET	2.79e-7 %
277	$\tau_h\tau_h\tau_h\tau_h\mu\mu$	2.74e-7 %
278	$\tau_h\tau_h\tau_h\tau_h\gamma\ell\ell$ + MET	1.33e-7 %
279	$\tau_h\tau_h\tau_h\tau_h\tau_h\gamma\ell$ + MET	9.78e-8 %
280	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\ell\ell$	7.47e-8 %
281	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\ell\ell$ + MET	3.29e-8 %
282	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\gamma$	3.00e-8 %
283	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\ell$ + MET	1.73e-8 %
284	$\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h\tau_h$	3.97e-9 %

Table 2.3, continued

CHAPTER 3

COLLIDER PHYSICS: THE LARGE HADRON COLLIDER, THE ATLAS DETECTOR, AND STATISTICS IN PARTICLE PHYSICS

3.1 The Large Hadron Collider

Unfortunately (or fortunately for the existence of life), the natural flux of high energy particles (at the scales of electroweak physics) is small, so before we can study any high-energy collision process, we must first bring particles to the desired energy, align them, and bring them to a small collision point manually. This is an enormous challenge, made possible only through decades of research in accelerator physics.

3.1.1 Accelerator Physics

Acceleration

The very first problem one faces when wanting to build a particle accelerator is how to accelerate the particles. The subatomic force that we have macroscopic control over is electromagnetism, so we are forced to consider acceleration of only charged particles through only electric and magnetic fields.

Originally, particle acceleration was accomplished with DC electric fields, such as in cathode ray tubes, Cockcroft-Walton accelerators, and Van de Graaff generators. The maximum voltage acheivable in a Van de Graaff generator is around 15 MV, correspond to a maximum particle energy around 15 MeV, compared to our target at the TeV scale [3]. One might consider using DC fields to steer a particle repeatedly through a high-voltage DC field, but this will always be impossible since static fields are conservative.

Thus to achieve high energy acceleration, we must use time varying electromagnetic fields. The modern standard of high-energy accelerators is the synchrotron, which sychronizes a

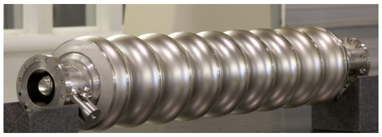


Figure 3.1: A niobium superconducting radiofrequency cavity from Fermilab [2].

radiofrequency electromagnetic wave with a spatially limited bunch of particles such that the particles are always in a region where the electric field is accelerating along the beam direction. The electric field must be parallel to the direction of the beam, which can be achieved in a controlled manner with conducting cavities, such as the one shown in Figure 3.1. Radiofrequency (RF) cavities are designed to operate at resonances with exactly the field directions and wavelengths necessary for acceleration, so the RF resonant standing wave can be setup relatively easily with an RF input current.

To achieve high accelerating gradients, one needs strong electric fields and therefore strong currents in the cavity. Many modern high-energy accelerators do achieve this with superconducting cavities, typically made of niobium alloys that are both workable and superconducting at useful temperatures (liquid Helium temperatures, around 4 Kelvin). One major challenge of superconducting RF cavities is the cryogenics, especially considering the nearby high energy beam, high currents, and large sizes of the devices.

The maximum accelerating gradients in RF cavities are limited by the dielectric breakdown of the cavity walls, at which point the cavity will spark. Thus another major challenge is the material science and engineering required to suppress this breakdown as much as possible.

Future accelerating technologies may include devices such as plasma wakefield accelerators. These use a drive beam (laser or particle beam) to disturb a plasma in such a way that creates a wake behind the beam. Because of the free charges in a plasma, the electric field strength in this wake is enormous, so the idea is to use it to accelerate a second beam to high energies over very short distances. Plasma is notoriously difficult to model and control, so such devices tend to have poorly controlled beams. However, research is progressing steadily, control is improving, and small-scale integration into some accelerators may be possible soon.

Once the accelerating cavities are built, particles must be passed through multiple times in order to reach interesting energies. There are two fundamental architectures to do this. The first is a linear accelerator, in which a large number of accelerating cavities are connected linearly. Thus the maximum collider energy is just the cavity's accelerating gradient times the length of the accelerator. The highest energy linear accelerator ever built is the Stanford Linear Collider, which beginning in 1989, collided electrons and positrons at a center of mass energy of 90 GeV by accelerating them over its 3.2 km length [33].

The other acceleration strategy is to build a circular machine such that the beam is passed multiple times through the same accelerating section. The majority of the machine is then taken up by the dipole magnets used to steer the beam in a circle, and the maximum energy of the beam depends on the magnetic field strength. One of the most important design considerations for a circular accelerator is synchrotron radiation, the radiation emitted by the particles from their centripetal acceleration. The power radiated is given by the relativistic Larmor formula

$$P = \frac{1}{6\pi\epsilon_0}\frac{e^4}{m^4c^5}B^2E^2 \tag{3.1}$$

where E is the particle energy and B is the magnetic field strength [3]. Note that the radiated power depends on the mass to the fourth power, which means that light particles emit dramatically more radiation than heavier particles. For example, protons will emit 10^{13} times less synchrotron radiation than electrons. This places severe limitations on circular electron collider energies, since the energy gained in a pass through the accelerating stage can be completely lost in synchrotron radiation in just one revolution, and large synchrotron radiation can overheat the refrigeration for the superconductors necessary for the large electric and magnetic fields.

Stability

The accelerating cavities and dipole magnets mentioned above will work for an ideal particle traveling at exactly the expected momentum and position. However, any real machine must be stable against perturbations, since the input beam will always contain a spread of momenta and positions. The non-ideal particles must not veer off-course. The total phase space of a particle in an accelerator is 6 dimensional, three dimensions of position and momentum. In the absence of couplings between the x, y, and z directions (a good first approximation), the phase space can be split into 3 two-dimensional spaces, one of which is longitudinal (in the direction of the beam) and two of which are transverse (perpendicular to the beam). In the absence of strong coupling between the different directions, all three directions must be independently stable.

Longitudinal stability of momentum ensures that the energy spread of the beam is small, and longitudinal stability of position ensures that the particles remain in phase with the accelerating portion of the RF fields in the accelerating cavities. Thus the longitudinal phase space is typically measured in terms of ϕ and ΔE with respect to the ideal particle.

It turns out that in RF accelerating cavities, the longitundinal behavior of beams is automatically stable in a region of phase space around the ideal particle. Suppose for example that a particle enters an accelerating cavity with a higher energy than the test particle but the same phase. In the first accelerating cavity, they both receive the same increase in energy. In traversing the distance to the second accelerating cavity (or in traveling one cycle back to the same cavity), the higher energy particle will move slightly ahead of the test particle (assuming a positive slip factor, which means that the energy of the particle is more important than path length in determining transit times – negative slip factors are also stable, though the RF phase has to be different). Because it is slightly ahead in phase, it will not be in phase with the peaks in the electric field, and receive less energy. This will continue to happen on successive passes until the energy drops below the ideal test particle energy, at which point the phase shifts will begin to reverse as the particle loses

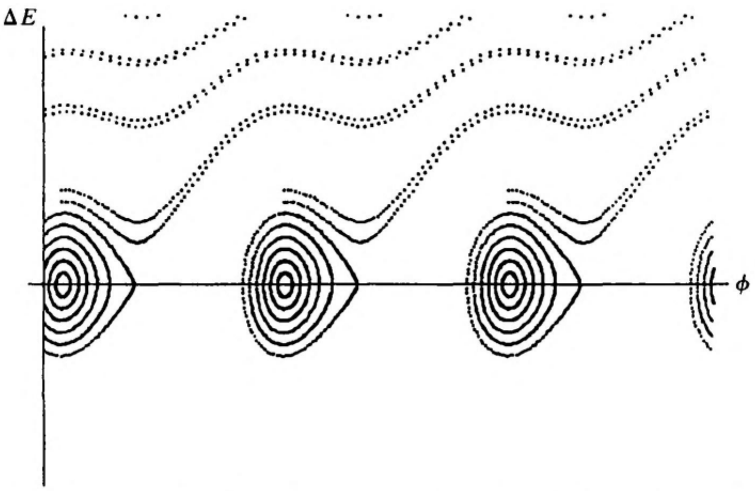


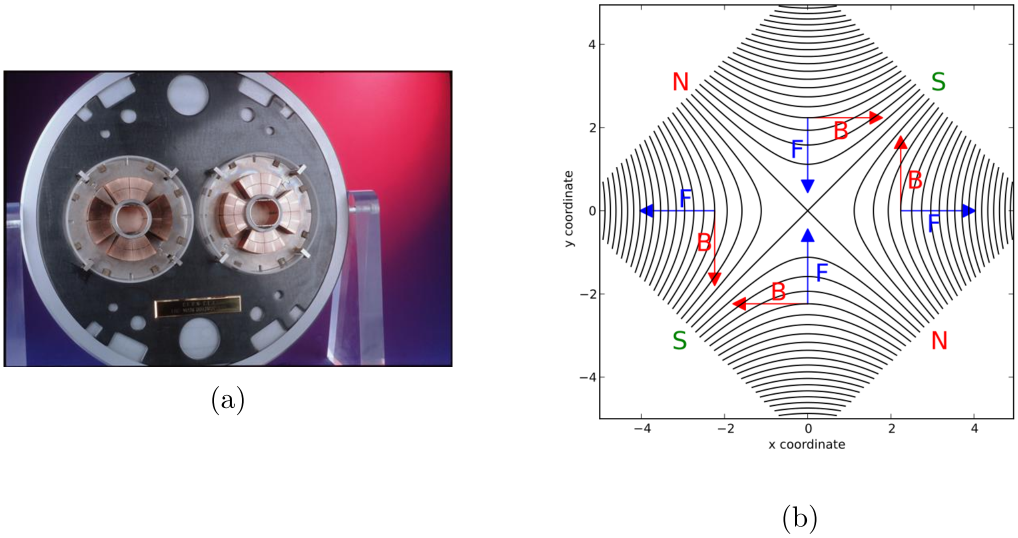
Figure 3.2: Phase space trajectory for longitudinal oscillations, reproduced from [3]. The exact trajectory shapes will of course depend on the specifications of the accelerator.

more energy with respect to the ideal particle. Eventually it will reach a phase-synchronized point and it begins to gain energy again. Thus the phase-space trajectory of this particle forms a closed loop around the ideal particle, and trajectories are stable. These oscillations are called synchrotron oscillations. This is shown diagramatically in Figure 3.2.

Each one of the stable regions aligns with a different peak in the RF field and is known as a bucket. The particles within a bucket are called a bunch. All buckets need not be filled.

Note that in Figure 3.2, particles with sufficiently different phase and energy with respect to the test particle are not stable and will drift. The boundary between stable and unstable motion is called the separatrix. One of the goals of the accelerator design is to ensure that the phase space area of one bunch (called the longitudinal emittance) is less than the phase space area of one bucket, so that all of the particles are kept in stable orbits.

In addition to the longitudinal stability of the beam, we need to ensure transverse stability, which is not at all automatic. One of the most important developments in accelerator physics was the development of strong focusing, which dramatically improves beam stability characteristics. The idea is to use quadrupole magnets which have a magnetic field that increases rapidly away from the central beam axis such that off-axis particles are deflected back toward the center. An example quadrupole magnet is shown in Figure 3.3.



(a)

(b)

Figure 3.3: (a) An example of the LHC quadrupole magnets, one for each beam pipe [4]. (b) An ideal quadrupole field, with forces shown [5].

Notice that it is not possible to focus in both the x and y directions with one magnet (which one can show will be true for any magnet configuration due to $\nabla \times \vec{B} = 0$). However, a sequence of a focusing then a defocusing magnet separated by some drift length will tend to focus in both directions. Such a configuration is called a FODO cell and is the most common strong focusing configuration. When particles pass through many FODO cells, this will lead to transverse oscillations, known as betatron oscillations. Particles that are off-axis in one direction will be deflected toward the beam axis, cross the beam axis, then be deflect back. Such oscillations are stable as long as the spacing between the quadrupoles, L is less than twice the focal length of the quadrupoles f, $L \leq 2f$. This ensures the particles are not lost before they are refocused.

Higher moment magnets are also often used to correct for additional affects related to the non-ideal behavior of various components. The complete set of magnets is called the machine lattice, which is the fundamental design of the accelerator that determines all the beam behavior. In the case of perfect dipoles and quadrupoles, the equations of motion of the particles in the lattice can be solved exactly.

In the approximation of perfect linearity (thin magnets and many periods), the solution to the equations of motion for the transverse oscillations can be written in the form [3]

$$x(s) = A\sqrt{\beta(s)}\cos[\psi(s) + \delta] \tag{3.2}$$

where s is the distance along the ideal trajectory, x is the transverse displacement (in meters), $\beta(s)$ is the amplitude function that depends on the lattice at position s, $\psi(s)$ is the local oscillation wavelength, and A and δ are constants. The equations of motion relate $\beta(s)$ and $\psi(s)$ by $\frac{d\psi}{ds} = \frac{1}{\beta(s)}$. Thus β can be interpreted as proportional to the local wavelength of the betatron oscillations. In a collider, the β function at the interaction point is often called $\beta^{\star}$, due to the fact that the amplitude of the transverse oscillations and therefore the beam size, is related to β. A smaller $\beta^{\star}$ means a narrower beam, and a larger number of collisions.

Another type of betatron oscillations occurs in circular machines due to the fact that particles of different momenta take different paths through the dipole bending magnets. The discussion is analogous to above, and another lattice function, the momentum dispersion function $D(s)$ is introduced in analogy with $\beta(s)$.

Suppose the initial particles are produced with some finite distribution in positions and momenta. In the limit of a large number of particles, this will fill some area (or volume) in phase space. Because of the adiabatic, Hamiltonian nature of the fields, the phase space volume will be conserved by Liouville's theorem. The phase space area in the x, $\frac{dx}{ds}$ plane is called the emittance, ϵ. In terms of the equation of motion, one has that $\epsilon \propto \sqrt{A}$, depending on the exact definition of emittance and phase-space shape of the beam. The beam emittance (in each dimension) is one of the most central design parameters of an accelerator.

All of the discussion so far concerned an ideal machine, neglecting imperfections in fields and interactions between particles. These effects are extremely important to measure and correct, as they will induce instabilities, though they can be quite complex and will not be addressed here.

Collisions

The goal of a collider is to provide as many collisions as possible, usually at the highest energy possible. In a circular collider, the maximum energy is determined by the field strength of the bending dipole magnets and the radius of the machine (and possibly the synchrotron radiation).

The luminosity is determined by the beam size and the frequency of bunch crossings,

$$L \propto \frac{fN^2}{4\epsilon\beta^\star} \tag{3.3}$$

where f is the bunch crossing frequency and N is the number of particles per bunch.

The maximum bunch crossing frequency is determined by the bucket spacing and therefore by the frequency of the RF accelerating cavities. The maximum number of particles per bunch is limited in a complicated manner by the interactions between the particles in the bunch, and by interactions of the bunch charge with the conducting accelerator walls. The emittance ϵ is determined by the beam source (assuming it can be preserved), and $\beta^\star$ is determined by the machine lattice. Thus the luminosity depends on many different aspects of the accelerator design, and maximizing luminosity is a challenge.

3.1.2 The Large Hadron Collider

The Large Hadron Collider (LHC) at CERN is the highest energy particle accelerator ever built, currently reaching 6.5 TeV of energy per proton. It is also the second highest luminosity collider ever built, with a maximum instantaneous luminosity of 2.14×10^{34} cm^{-2} s^{-1} [34] and typical operating luminosity near 1.0×10^{34} cm^{-2} s^{-1}.

Many of the basic LHC machine parameters are summarized in Table 3.1. The LHC uses a FODO-based latice with both the dipoles and quadrupoles incorporated into the same cryostat.

At the current 13 TeV center of mass energy, the total proton-proton cross section is

Parameter	**Value**
Circumference	26659 m
Center of Mass Energy	14 TeV
Luminosity	10^{34} cm^{-2} s^{-1}
Emittance	3.75 μm rad
$\beta^{\star}$	0.55 m
Bunch Spacing	25 ns (7.5 m)
No. of bunches per beam	2808
No. protons per bunch (at start)	$1.5{\times}10^{11}$
Collision crossing angle	300 μrad
Average crossing rate	31.6 MHz
Number of RF Cavities	8
RF Frequency	400.8 MHz
RF Voltage (at 7 TeV)	16 MV
Number of Dipoles	1232
Number of Quadrupoles	858
Dipole operating temperature	1.9 K
Peak Dipole Field Strength	8.33 T
Stored beam energy	360 MJ
Stored energy in magnets	11 GJ
Typical beam lifetime	10 h
Synchrotron radiation per beam	6 kW

Table 3.1: Nominal design machine parameters of the Large Hadron Collider [16].

110 mb ($1.1{\times}10^{-25}$ cm^2) [35], which implies a total collision rate around 10^9 collisions per second.

Note that protons are injected into the LHC in beams at an energy of 450 GeV. This is achieved in the complex, multistage accelerator complex, largely consisting of former state-of-the-art accelerators and colliders. The LHC operates in “fills,” in which the number of bunches in the beam is gradually accumulated from the injector complex at 450 GeV. Once all of the required bunches are present, injection is over and the beams are ramped in energy before collisions.

During collisions, the instantaneous luminosity falls exponentially over time as the protons are taken out of circulation from the collisions. Each fill typically runs for around 10 hours before the collision rate drops enough that it becomes more time-efficient to dump the beam and start a new fill.

Around the LHC ring, there are four interaction points, around which are constructed four large particle detectors, ATLAS, CMS, LHCb, and ALICE. ATLAS and CMS are general-purpose particle detectors built to search for new particles and measure the properties of Standard Model particles. LHCb is more specialized to study the property of particles involving b-quarks, which have various interesting rare and suppressed decay modes in which deviations are strong signs of new physics happening at loop level. ALICE is designed to study the collision of ions, which is an alternate running mode of the LHC, typically receiving around a month of run time every year or two. In ion collisions, fundamental properties of finite-temperature quark matter can be studied.

One challenging aspect of the LHC is that in ATLAS and CMS, the collision luminosity is sufficiently high that many (10's of) proton-proton collisions occur in each bunch crossing. The number of collisions per bunch crossing is known as pileup. No collider before the LHC faced this challenge. The pileup leads to a dramatic increase in the noise of the detector, requiring sophisticated algorithms to filter it out. In addition, the 25 ns bunch spacing is sufficiently short that collision products from different bunches are propagating through the detector at any given time. The short bunch spacing and pileup make detector design extremely challenging.

This thesis analyzes data collected during Run 2 of the LHC, which occured between 2015 and 2018. During Run 2, ATLAS recorded 139 fb^{-1} of data at 13 TeV. The pileup distribution of the data is shown in Figure 3.4.

Run 3 will begin around 2022 and last for around 3 years, doubling the total collected luminosity. Then beginning around 2028, the LHC will be upgraded to the High-Luminosity LHC (HL-LHC), which should reach an instantaneous luminosity around 10^{35} cm^2 s^{-1} and record a total of 3000 fb^{-1} of data per detector. At the HL-LHC, pileup will reach up to 200, a situation that the current detectors are not able to process. Thus between now and then large upgrades will be conducted on all of the detectors in order to handle the more challenging environment. The HL-LHC is expected to then run until the mid-2030's or 2040,

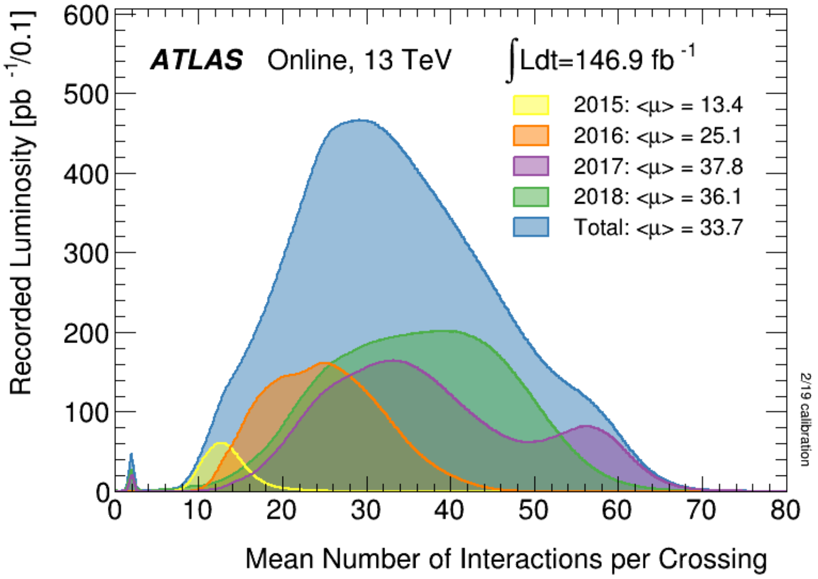


Figure 3.4: The distribution of pileup in ATLAS during Run 2.

at which point it will hopefully be superseded by the next collider.

3.2 ATLAS

3.2.1 Overview

ATLAS is one of the two general purpose detectors built to measure proton-proton collisions at the LHC. Accomplishing this at the LHC energy scales requires an enormous detector, 44 meters long and 25 meters tall, which is shown in Figure 3.5. The total detector weight is around 7000 tons, and it is instrumented with 3000 km of cables that read out 100 million channels. The raw unfiltered data rate is on the order of 100 TB/s, which is then filtered on-the-fly down to 300 MB/s, which still totals 10's of petabytes of data per year produced. Over 3000 scientists and engineers contribute to the ATLAS detector, keeping it operating and implementing upgrades [36].

The goal of the detector is to measure the four-vectors (momentum and energy) of every particle leaving the collision point. Additionally, we want to know the particle charge and

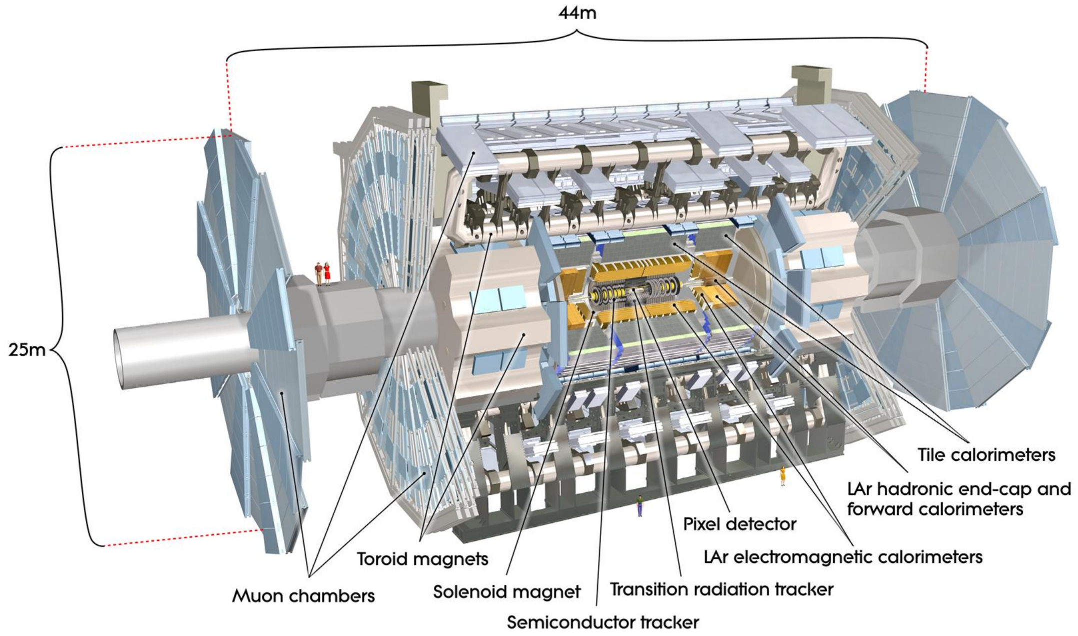

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 3.5: A rendering of the ATLAS detector, shown to scale with two “standard humans” standing on the detector on the left.

identity when possible. There are only a handful of stable particles that need to be measured and distinguished. The particle ID in ATLAS is accomplished largely through the variation in penetration depth. Electron and photons are both absorbed quickly but can be distinguished by the trail of ionization left by the charged electron. Hadrons penetrate much further and are absorbed later. Muons are highly penetrating and escape the detector, but leave a trace of ionization throughout. Neutrinos do not interact measurably with the detector, and are measured by an apparent momentum imbalance. This is summarized in the now-classic diagram, Figure 3.6.

ATLAS cannot distinguish long-lived hadrons (pions, kaons, protons, and neutrons) very well, but this is not usually important to first order, since hadrons typically are grouped into jets with algorithms that are relatively agnostic to particle ID (see Section 4.1).

To take advantage of the different interactions with matter of the different particle types, ATLAS is segmented into layers, each of which is specialized for a specific particle type. The inner-most layer is the tracker, which tracks charged particles as they propagate in

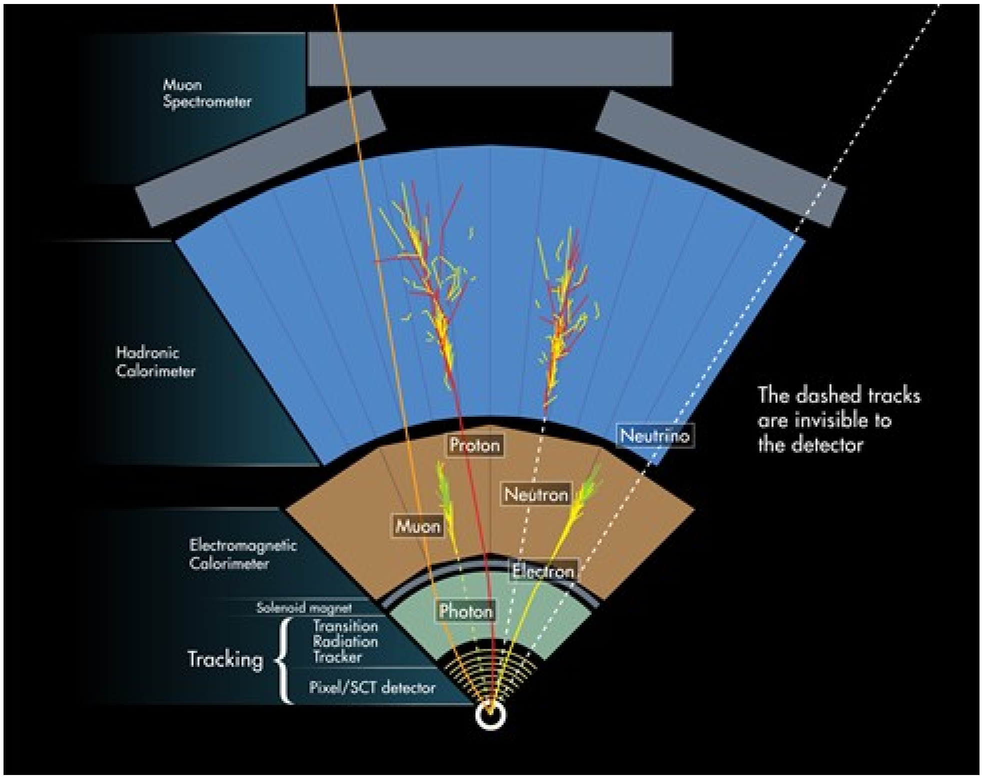
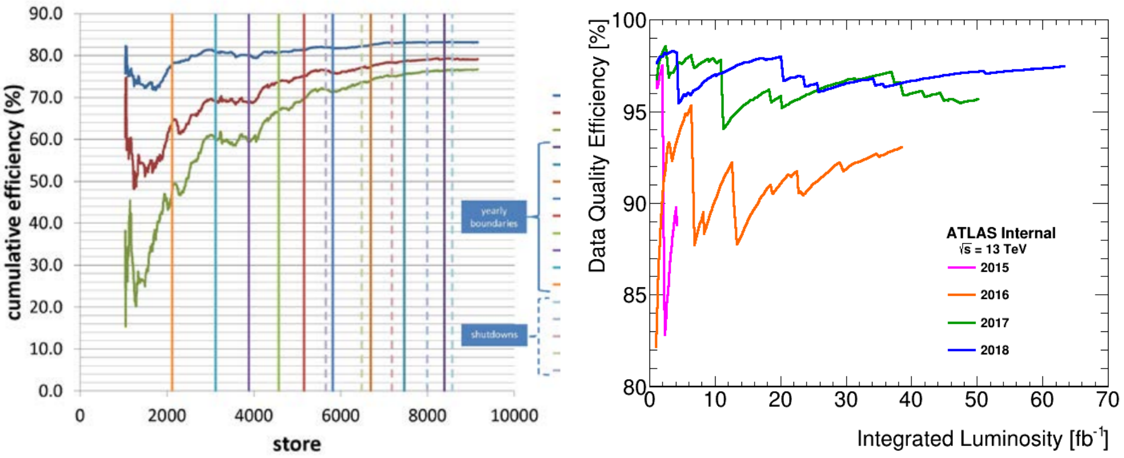


Figure 3.6: A diagram of how different particles interact with the ATLAS detector.

a magnetic field, providing a measurement of momentum, position, and the sign of the charge. The next layer is the electromagnetic calorimeter that entirely absorbs electrons and photons in order to measure their energy and position. The hadronic calorimeter absorbs the remaining hadrons for an energy measurement. Muons are the only interacting particle to penetrate all of these layers, so are tracked again in the Muon Spectrometer, which comes immediately after and just before a toroidal magnetic field in order to provide an independent momentum measurement.

ATLAS operations have been remarkably stable, with 95.6% of the collisions delivered by the LHC being useable for physics analysis. In comparison, at the Tevatron, the spiritual predecessor to the LHC, the CDF experiment data recording efficiency peaked below 80%, as shown in Figure 3.7.

The standard ATLAS coordinate system aligns the x-axis toward the center of the LHC ring, the y-axis is vertically upward, and the z-axis points along the direction of the beam.



(a) CDF

(b) ATLAS

Figure 3.7: Data taking efficiency over time for CDF [6] and ATLAS.

ATLAS typically uses coordinates where the azimuthal coordinate ϕ is the angle in the xy-plane with respect to the x-axis and θ is the polar angle measured with respect to the z-axis. A frequently used variable in collider physics is pseudorapidity, η, which is defined as

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \tag{3.4}$$

Pseudorapidity is a useful quantity because when particle momenta are much greater than their masses (which is usually but not always true at the LHC), then differences in pseudorapidity are invariant under Lorentz boosts in the z-direction. This is useful because the initial z-momentum is unknown in proton-proton collisions since the constituent quarks may be carrying any fraction of the total proton momentum, so the initial z-momentum imbalance can be large. Because ϕ is a transverse variable, it is inherently invariant under boosts along z, so this allows us to construct the invariant difference in position betweeen two particles, $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. Lastly, transverse momentum $\vec{p}_T$ is the 2-component vector of momentum in the xy (transverse) plane and the longitudinal momentum p_L is the component of the momentum along z.

The ATLAS detector is divided in pseudorapidity into two components. Transverse to the beam is the central region, which is covered by the nearly cylindrical “barrel” detector, that extends to roughly $|\eta| < 2.5$, depending on the sub-detector. For angles closer to the beam pipe, in the “forward” region, the detector orientation and technology changes to cope with much higher particle fluxes (since the total particle flux is roughly constant with η and $\eta \to \infty$ near the beam pipe.

3.2.2 Tracker

The innermost layer of the ATLAS detector is the tracker. When charged particles pass through the tracker, they trigger detectors that measure positions in three dimensions, and these points can be connected to form particle tracks. The entire tracker is immersed in a 2 Tesla solenoidal magnetic field along the z-direction, so that the trajectories form helices and the transverse momentum can be measured (assuming unit charge, which is true for all particles stable enough to be tracked).

The tracker is divided into two main components, the inner detector, which is based on silicon diode detectors, and the transition radiation detector, which is based on gaseous ionization chambers.

The Silicon Tracker

Silicon diode detectors are relatively simple in operation principle, though difficult and expensive to build. The principle of operations is based on the fact that as charged particles pass through silicon, they produce electron-hole pairs that in the presence of an applied electric field, drift toward external electrodes and produce a pulse. To apply a large electric field without producing a large background “leakage” current, the silicon is doped to be a diode. Impurities are added to one side of the silicon that donate electrons, making it n-type, while the other side of the silicon has impurities that donate holes, making it p-type. The net effect is that a current can easily flow downhill from the n-type to the p-type semiconductor, but not in reverse, making it a diode. For a detector, the diode is reverse biased so little current flows.

The number of electron-hole pairs produced is determined by their ionization energy, which in silicon is around 3.6 eV. For particles at LHC energy scales (GeV and up), the energy lost in order to produce a measurable signal in the thin silicon will be small. In fact, in ATLAS, much more energy is lost to the readout electrons on the silicon than in the silicon itself. The variation in number of electrons produced is actually not a Poisson distribution, but has a variance 5-10 times less than a Poisson variance. The ratio between the observed and Poisson variances is the Fano factor.

The time resolution of silicon detectors is determined by the thickness and the electron-hole drift velocity. Electrons have a higher mobility and thus travel faster than holes for a given electric field strength, though the difference is only a factor of 2-3. This means incidentally there will be two pulses closely separated in time that are usually integrated together in the readout electronics.

In ATLAS, there are two different readout geometries of silicon detectors used. The first is the pixel detector, which forms the innermost four layers of the detector. The silicon plane is bump-bonded to individual readout pads with dimensions of 50 μm in the $R\phi$ direction, and 300 μm in the z direction. This leads to 140 million readout channels total. A single charged particle will produce a signal in multiple nearby pixels, which are then clustered into a hit. The clustering allows the position resolution to be better than the pixel size, typically around 12 μm in $R\phi$ and 65 μm in z. The innermost pixel layer is placed at 3.3 cm from the interaction point. Such close proximity is necessary for accurate vertexing and b-tagging (see section 4.1).

The second readout geometry is the strips used in the “SCT” layers. In this geometry, the silicon is readout with 80 μm wide strips. Because strips only give one dimension of information, each SCT layer is actually a bilayer of two detector glued together at a relative angle of 40 mrad. This then provides two-dimensional information, with a resolution of 16 μm in $R\phi$ and 600 μm in z. The SCT detector is segmented into 4 layers with a total of 6.2

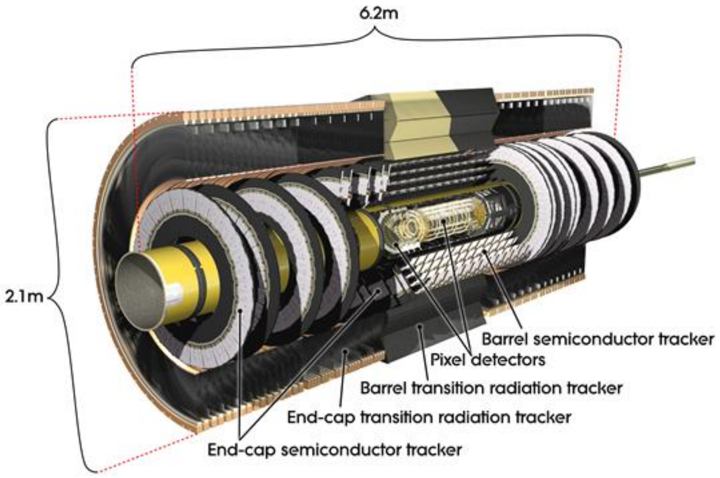

6.2m
2.1m
Barrel semiconductor tracker
Pixel detectors
Barrel transition radiation tracker
End-cap transition radiation tracker
End-cap semiconductor tracker

Figure 3.8: A computer-generated image of the ATLAS Inner Detector.

million readout channels.

Each of the pixel and SCT detectors have two different installation orientations. The barrel detectors, up to approximately $|\eta| < 1$, are cylindrical with horizontal sensors parallel to the beam, while the end-cap detectors, up to approximately $|\eta| < 2.5$, are disks with the sensors oriented vertically. The geometry is shown in figure 3.8.

One of the major challenges of the silicon detector is radiation hardness. As the silicon is exposed to radiation, Silicon nuclei are knocked out of the lattice, disrupting the band structure and degrading performance. Partly for this reason, the ATLAS silicon tracker will be completely replaced for the High-Luminosity LHC.

The Transition Radiation Tracker

The Transition Radiation Tracker (TRT) uses a technology altogether different from the silicon tracker — straw tubes. In a straw tube, a thin wire (30 μm diameter in ATLAS) is threaded through a straw filled with a gas. The wire is held at a large positive voltage with respect to the straw wall, inducing a large radial electric field through the gas. When a charged particle passes through the straw, some of the gas is ionized. The electrons then drift toward the center of the tube. Close to the cathode wire, the electric field is large enough that a free electron can collide with and ionize the gas atoms, producing more elections. This causes a chain reaction that produces an avalanche of electrons that is then collected on the cathode, leading to a current pulse.

The ATLAS TRT is an array of 50000 straws in the barrel, arranged parallel to the beam, and 320000 straws in the end-cap, arranged radially. The straws provide around 36 additional hit coordinates with resolutions perpendicular to the straws around 170 μm and no resolution parallel to the straws.

In addition, between the straw tubes are polymer fibers that induce transition radiation when electrons pass through. Transition radiation is introduced when rapidly moving charged particles cross between the interface of two media with different dielectric coefficients, in this case polymer fibers and CO_2 gas. On either side of the interface, the moving particle electromagnetic fields are solutions of the homogeneous Maxwell equations that differ in the dielectric coefficient. Matching these solutions at the boundary leads to a radiation field that propagates to infinity. The energy emitted in the radiation is proportional to the Lorentz factor γ for relativistic particles. Because $\gamma = \frac{E}{m}$, transition radiation is larger for electrons than pions by a factor of $\frac{m_\pi}{m_e} \approx 300$. The X-rays emitted by electrons from the transition radiation are then absorbed in the straw tubes, producing a significantly larger signal than other charged particles. This provides a mechanism to distinguish electrons from pions (and other heavier hadrons).

One of the challenges with the TRT is its high occupancy, the average number of straws that fire per event, which can reach up to 60% in the LHC Run 2. This can make distinguishing tracks difficult, as well as causing tracks from the next event to be missed due to the wire recharge time, which is greater than the 25 ns event spacing.

3.2.3 Calorimeters

The next layers of the ATLAS detector outside the tracking volume (and outside the 2 T solenoid) are the calorimeters. The goal of these components is to measure the energy of

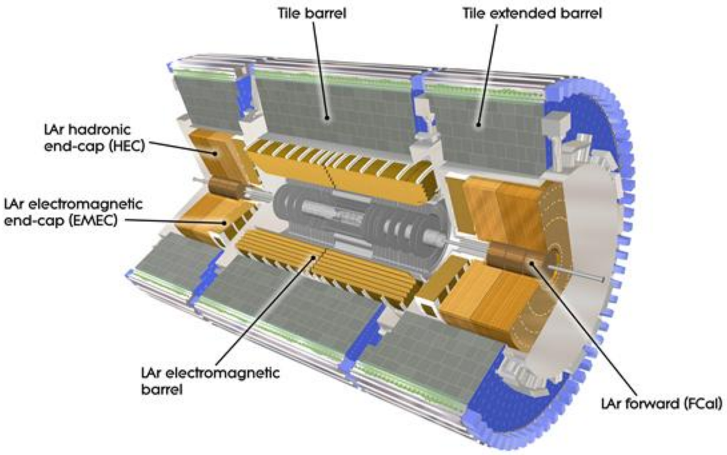

Tile barrel
Tile extended barrel
LAr hadronic end-cap (HEC)
LAr electromagnetic end-cap (EMEC)
LAr electromagnetic barrel
LAr forward (FCal)

Figure 3.9: The ATLAS electromagnetic and hadronic calorimeters.

particles by completely absorbing them.

The Electromagnetic Calorimeter

The electromagnetic calorimeter is designed to absorb photons and electrons (and positrons) in order to measure their energy and position. When an electron or photon enters a material, it induces an electromagnetic shower.

When a high energy electron enters a material, it begins to radiate photons and to ionize atoms, releasing electrons. The electrons can then go on to radiate more photons and ionize more electrons. The photons, depending on their energy, either pair produce into an electron-positron pair, Compton scatter, freeing an electron, or are absorbed via the photo-electric effect. A high-energy photon entering a material will undergo a similar shower, though the starting position depends on when the initial photon converts, which induces a greater variability. Which process dominates depends on the depth and time in the shower, but regardless, most of the measurable energy will be deposited by soft (MeV-scale) particles.

The electromagnetic shower will reach a maximum energy deposit per centimeter at some depth, then taper off. The distance over which an electron looses 63% $(1-e^{-1})$ of its energy is called the radiation length X_0, which depends on the material roughly as A/Z^2. The mean free path of a photon is proportional to the radiation length as $9X_0/7$. Similarly, the typical transverse size of a shower is called the Molière radius and is less material dependent.

Most modern calorimeters are sampling calorimeters that have sensing elements interspersed with dense absorbers that catalyze shower development. The sensing elements can either use an electric field to collect the electrons in the shower or photo-sensors to measure the light in the shower. A common technique is to measure the light of scintillating materials, which emit light from atomic excitation in the presence of charged particles. Regardless of the sensing mechanism, a key aspect of calorimeter design is that the output voltage (pulse height or pulse area) should be directly proportional to the particle energy. The proportionality constant is then calibrated in a test beam or *in situ*.

In a well-designed calorimeter, the resolution is dominated by Poisson statistics, such as the fraction of the shower energy deposited in the sensing components, or the number of scintillation photons reaching the photosensor. Since the number of photons should be proportional to the total energy, calorimeter uncertainties are often parameterized fairly accurately as $\frac{\sigma}{E} = \frac{c}{\sqrt{E}}$. At very large or very low energies, different effects may still dominate, such as electronic noise at low energy where the signals are smaller. Therefore the total energy resolution is often parameterized as

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c \qquad (3.5)$$

where the $\oplus$ conventionally denotes addition in quadrature. The first term is the sampling term, the second is the noise term, and the third is the constant term.

The ATLAS electromagnetic calorimeter uses a lead absorber and a cryogenic liquid argon sampling material. The shower is detected in the argon via ionization, with the ionized electrons drifted in an electric field to an anode, producing a pulse. The calorimeter somewhat uniquely uses an accordion geometry, which features alternating accordion-shaped layers of lead, electrodes, and liquid argon. This shape allows continuous coverage in ϕ

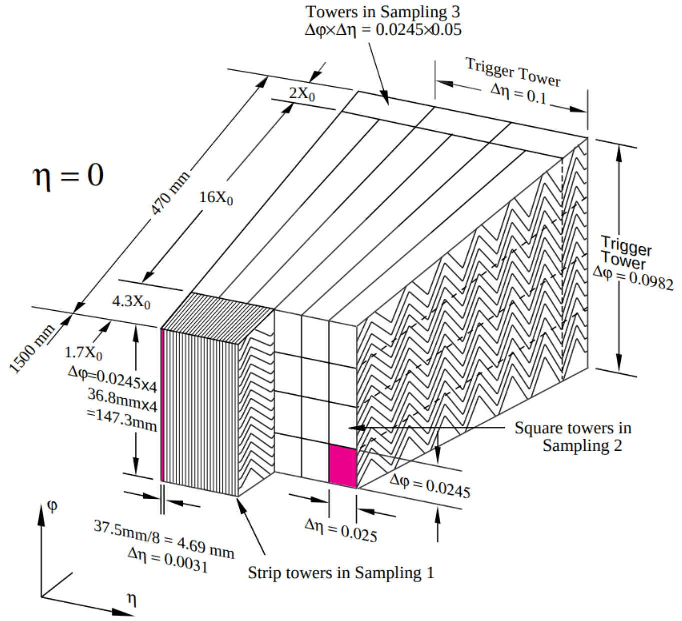


Figure 3.10: The ATLAS Liquid Argon electromagnetic calorimeter geometry [7].

without any gaps or dead zones. In addition, the calorimeter is segmented into three sections longitudinal to the shower development. The innermost layer is finely segmented in η and ϕ to allow electron and photon position resolution. The second layer absorbs the majority of the shower, and the third layer allows for correction of any shower punch-through. The geometry is shown roughly in Figure 3.10. In addition to the main calorimeter layers, there is a thin pre-sampling layer (not shown in the figure) in regions of the detector where there is a large amount of material before the calorimeter. The presampler measures how far the shower has already developed so that if the shower develops early, the energy loss can be corrected. The total longitudinal length of the calorimeter is 20-25 radiation lengths. The energy resolution parameters in equation 3.2.3 are approximately $a = 10\%$, $b = 400$ MeV, $c = 0.2\%$.

The Hadronic Calorimeter

Hadronic calorimetery is very similar to electromagnetic calorimetery except that the showers induced by hadronic particles are significantly more variable. Hadronic showers produce

π^0 particles that promptly decay to photons, inducing electromagnetic showers within the hadronic shower. The fraction of the hadronic shower carried by electromagnetic particles can fluctuate, and the shower profile can have unusual peaks due to stochastic π^0 production. Additionally, nuclear effects further complicate the shower. For example, hadrons can break apart nuclei via spallation, causing the nuclear binding energy to be lost. Then neutrons can then travel a substantial distance and time without producing a signal, potentially exiting the primary shower volume.

The typical length scale in hadronic showers is the nuclear interaction length, the average distance that hadrons travel before a nuclear interaction. This distance is typically much larger than one electromagnetic radiation length, so hadronic showers are much longer than electromagnetic showers. Combined with the fact that the shower shape variability can lead to randomly long showers, calorimeter punch-through is common, where the shower is not entirely absorbed in the calorimeter, even for deep calorimeters.

One challenge of hadronic calorimeters is that without special effort, they are inherently non-linear – the output voltage is not directly proportional to the energy. The primary source is that the fraction of a hadronic shower that is electromagnetic increases (logarithmically) with energy. If the calorimeter has a different response to the hadronic and electromagnetic components, then there is an inherent nonlinearity. Calorimeters that ensure equal electromagnetic and hadronic responses are called compensating calorimeters. The ATLAS hadronic calorimeter is non-compensating, so nonlinearities have to be accounted for in shower reconstruction.

The ATLAS barrel hadronic calorimeter is constructed out of a steel absorber and plastic scintillator sampling tiles, known as the tile calorimeter. The geometry of a segment of the hadronic calorimeter is shown in figure 3.11. The light from the plastic scintillating tiles is collected on both sides via optical fibers that connect to photomultiplier tubes. The calorimeter is around 9 interaction lengths deep. The tile calorimeter is designed to have a energy resolution for jets of $\frac{\sigma}{E} = \frac{50\%}{\sqrt{E}} \oplus 3\%$ and an ability to correct for nonlinearities to a

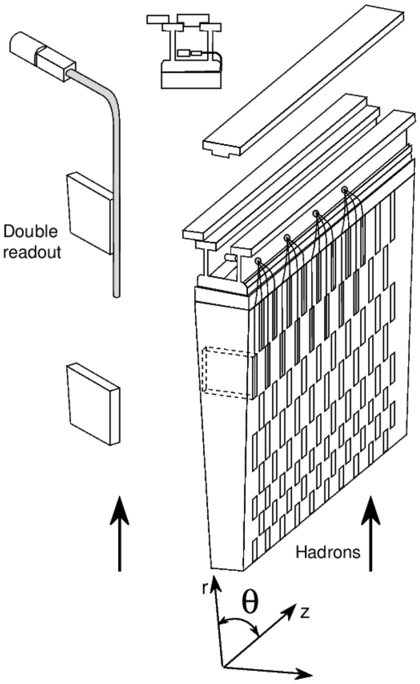


Figure 3.11: The ATLAS hadronic tile calorimeter geometry [8].

level of 1-2%.

In the endcap and forward regions ($\eta > 1.5$), the hadronic calorimetry is provided via a liquid argon and copper (endcap) or tungsten (forward) sampling calorimeter, similar to the electromagnetic calorimeters, albeit with a flat plane instead of an accordion geometry.

3.2.4 Muon Spectrometer

The ATLAS muon system is the outermost layer of the ATLAS detector and is made of four different detector subsystems and two different magnet geometries. The muon system is designed to be able to operate independently of the rest of the detector, with one original design goal of being able to measure $H \to ZZ^* \to 4\mu$ independently.

In the barrel is the main large toroidal magnet that is the namesake of the former ATLAS acronym (A Toroidal LHC ApparatuS). The end-caps feature two smaller magnets. The strong magnetic fields bend muons for a measurement of momentum, hence the name for the muon system, the muon spectrometer. The air-core magnets are superconducting and

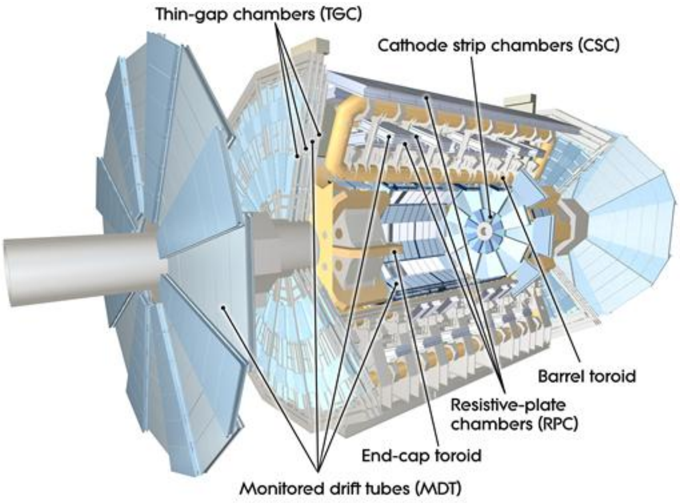

Thin-gap chambers (TGC)
Cathode strip chambers (CSC)
Barrel toroid
Resistive-plate chambers (RPC)
End-cap toroid
Monitored drift tubes (MDT)

Figure 3.12: A rendering of the ATLAS muon subsystem.

produce field strengths in the range 0.5-2 T.

The primary detector in the barrel region is the Monitored Drift Tubes (MDTs), which provide muon coordinate data at three locations along its trajectory. The drift tubes are composed of an aluminum tube at ground filled with Argon and a tungsten-rhenium wire through the middle held at 3.3 kV. The muons ionize the argon, producing an avalanche near the wire that leads to a pulse. The positions of the drift tube modules are carefully monitored with optical lasers to ensure precise coordinate alignment. The MDTs are also used in the large wheels in the end-cap region.

The primary detector in the end-cap region is the cathode strip chambers, which can sustain the higher rate being situated closer to the beamline in the forward region. The cathode strip chambers are multiwire proportional chambers with cathode strip readout. In a multiwire proportional chamber, several anode wires in parallel in a plane are held at high voltage and ionized electrons from charged particle tracks induce a negative pulse on the anode. In the cathode strip chambers, the anode pulse is not measured directly, but rather the postive image charge that is induced on the conductive cathode strips is measured, which can be used to increase the position resolution by measuring the centroid of the pulse.

Muon trigger signals are provided via two other detectors in the barrel and endcaps.

In the barrel, the trigger is provided by the resistive plate chambers. These chambers are constructed by holding two resistive Bakelite sheets 2 mm from one another and producing an electric field strong enough to induce avalanching. The avalance is not readout directly, but rather readout strips are capactively coupled to the Bakelite. Strips on either side of the RPC are oriented orthogonally, providing a two dimensional position measurement with a resolution of a few millimeters. This measurement is in fact the primary η measurement for the barrel, since the monitored drift tubes are one dimensional. The RPC response has a rise time of 5 ns and is thus used as a fast trigger. Three of the four layers of RPCs must fire in coincidence for a muon trigger.

In the endcaps, the muon trigger is provided by the Thin Gap Chambers, which are a type of self-quenched streamer detector, which are geometrically similar to the multiwire proportional chambers used in the cathode strip chambers, though operated somewhat differently. The wires are held at a higher voltage such that the avalanche reaches saturation where the electric field from the positive ions reduces the electric field felt by the electrons and stops further avalanche formation. This large number of ions, however, typically results in the emission of UV photons that can trigger avalanches far away from the initial avalanche. To suppress this, a quenching gas is used that highly absorbs UV photons, in this case a 55-45 mix of CO_2 and n-pentane (n-$C_5H_1$2). This allows the position resolution to be maintained. Similar to the CSCs, the TGCs are readout via capacitively coupled strips rather than via the anode wires directly.

One critical piece of the muon system is the endcap detectors that are placed before the endcap magnet. A precise position measurement is necessary in order to accurately measure the muon deflection angle from the magnetic field. However, the CSC detector in this location can suffer from the high rate, reducing muon efficiency and momentum resolution. Additionally, the end-cap TGC trigger has no tracking before the endcap toroid, which reduces the ability to use the muon trajectory to reduce backgrounds. In fact, 90% of the forward TGC muon triggers are fakes, mostly protons, from material interactions elsewhere.

For these reasons, the inner radii of the end-cap muon system before the magnet is being upgraded to the New Small Wheel (NSW). The NSW installation is currently ongoing, with operation expected starting in Run 3 through the end of the HL-LHC.

The NSW will include strip Thin Gap Chambers for triggering and micromegas for precise position measurements. The sTGCs are similar to the main ATLAS TGCs, except the readout strip width and spacing is reduced by a factor of 10 to reach the necessary position resolution, and the resistance of the cathode has been reduced to allow faster recharge times in the denser environment. Micromegas on the other hand are similar to the resistive plate chambers, except rather than producing a strong electric field across the entire thickness, a micromesh is placed at a small distance from the anode and biased to create a strong electric field in the thin amplification region. The ionized electrons will therefore drift toward and through the micromesh before being amplified onto the anode, which is segmented with strips and provides position resolution around 100 μm. Because of the thin gap between the anode and the mesh, the positive ions from the avalanche are neutralized quickly compared to other gaseous detectors, allowing higher rate operation.

3.2.5 Trigger

Once all of the detectors have been readout and processed on the front-end boards (a difficult task beyond the scope of this discussion), we then need to decide whether a given event should be stored to disk or discarded. It is not possible to store every event given the enormous amount of data that would be produced. The vast majority of proton-proton collisions are "soft" anyway, where the protons either scatter elastically or simply fragment into constituent quarks and gluons. Because the front-end readout boards and trigger system have finite buffer memory, trigger decisions need to happen quickly, which limits the processing and often necessitates using simple algorithms or clever design to compute more complex quantities.

The ATLAS trigger in Run 2 is divided into two levels. The first is the Level 1 trigger, which processes every event using custom hardware and firmware. The Level 1 trigger

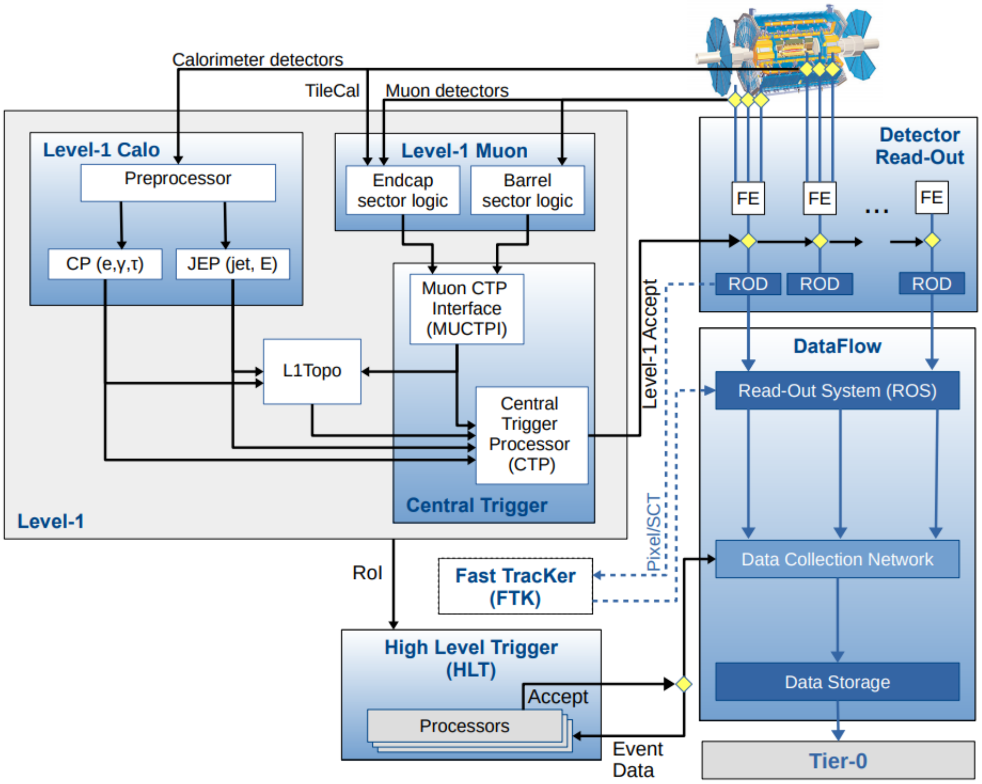


Figure 3.13: The ATLAS Trigger and Data Acquisition (TDAQ) architecture in Run 2 (except for the FTK system, which was tested but unused).

receives data at an event rate of 40 MHz and accepts only 100 kHz, an acceptance of only 0.25%. After an L1 accept, the data are read off of the front-boards and stored in a separate, deeper data buffer. The data are also passed to the software-based High-Level Trigger (HLT) that receives the 100 kHz input event rate and accepts only around 1.2 kHz to write to disk. A diagram of the ATLAS trigger and data acquisition architecture is shown in Figure 3.13. The entire system is located in an underground cavern physically close, though shielded from, the ATLAS detector in order to minimize latency in signal transmission times.

The Level 1 system has two initial processing stages, L1Calo and L1Muon, that do processing of local areas of the detectors, called regions of interest. L1Calo processes the information from the calorimeters, performing basic energy clustering and thresholding to test if particle energies (with rough calibrations) are above a given trigger threshold. L1Calo also performs scalar and vector sums all of the calorimeter energy to trigger on total energy and missing energy. The L1Muon system uses the muon spectrometer trigger detectors to check for muons above the trigger p_T thresholds. Information from regions of interest are combined (mostly in the central trigger processor) to trigger on object multiplicity above specified thresholds. Additionally, Run 2 included L1Topo, which can compute more complex geometric quantities such as invariant mass and angular distance between objects, so that these can be used to suppress rates at Level 1. All of these algorithms are implemented in firmware and run on Field Programmable Gate Arrays (FPGAs).

The High Level Trigger is a complex software system composed of around 40k parallel Processing Units that make decisions within a few hundred milliseconds. Many trigger algorithms first conduct a rough online rejection step, followed by a longer offline-like reconstruction for more precise rejection. Such algorithms can become quite complex, such as for b-jet reconstruction, which features tracking, jet reconstruction, and neural networks.

Tracking using inner detector data is one of the most difficult aspects of the trigger. The large number of particle hits per layer means that the connecting-the-dots track reconstruction naively takes exponential time with increasing occupancy. Tracking cannot currently be used at Level 1, and can only be used to a limited extend in the High Level Trigger (either in regions of interest or very limited full-scan tracking). The Fast Tracker (FTK) was a system that used pattern recognition in ASICs and fast tracking algorithms on FPGAs in order to provide tracks to the HLT with a latency of 100 ms. In practice, the FTK project was extremely complex, ran into delays, and was eventually canceled. Future trigger upgrades, however, will need to use tracking to suppress higher rates. Similar hardware to FTK is under consideration for the HTT (Hardware Tracking for the Trigger), as well as alternative solution using commodity hardware, such as simply smarter algorithms on CPUs, machine learning, or other hardware acceleration, such as GPUs.

3.2.6 Computing

Without the facilities and infrastructure to process it, the data recorded by ATLAS is useless, and given the large volume of data, processing it is a challenge. ATLAS relies on the Worldwide LHC Computing Grid (WLCG), which is a globe-spanning network of computing centers, to store and process data. The ATLAS software infrastructure, Athena, implements common reconstruction algorithms that run on all data to build and calibrate all of the basic particle objects.

From the analyzer point of view, there are only a handful of different object types presented in the software. These include electrons, photons, muons, taus, jets, b-jets, and missing energy. The actual reconstruction algorithms may vary, but the majority of objects will be labeled as one of these.

ATLAS data processing is done in stages, each stage filtering out events and improving the quality of the data passing the filters. The first stage is the central processing and calibration, which is run on all data to produce calibrated basic objects, stored in files called xAODs. The next stage is the derivation, a DAOD file, which makes basic event selections and is run over all of the xAODs. The derivations may be shared by several analyses and are only run over the full dataset a handful of times throughout the several-year lifetime of an analysis. The next stage of processing is to use the DAODs to produce analysis-specific datasets, called NTuples. Depending on the analysis, the physics results and plots may be produced directly from the NTuple, or there may be an additional stage of processing to produce NanoNTuples that contain only the bare-bones information and events necessary to make statements about physics. NanoNTuples may even be storeable on a personal laptop.

Throughout the processing chain, the software becomes progressively less centrally managed. The central processing and derivation steps always directly use the Athena infrastructure. NTuple production is usually a local analysis framework code that makes callbacks to tools available in Athena packages. NanoNTuple production and later processing may not rely on Athena at all.

Data processing software is mostly written in C++ using the ROOT framework, though final-stage processing and plotting has been gradually switching to python for ease of use and integration with industry-standard tools.

3.3 Statistical Analysis

Once the data have been recorded by the detector, the next challenge is determining whether a particular physical process happened in the data and to quantify the sensitivity of the process in order to either determine the uncertainty of the measurement in the case of an observation, or to determine what theory space is ruled out in the case of a non-observation. Fundamentally, this translates to a counting problem, where the number of collisions with certain kinematics and final state particles is observed. Given a luminosity, this can then be translated into a cross-section and compared against theory. Thus in the final statistical analysis, we tend to be concerned with Poisson counting statistics. However, in many intermediate stages, we will make measurements of quantities such as particle momentum, which may have Gaussian or more complicated uncertainties. The uncertainty in every step of the analysis needs to be included and propagated into the final result.

3.3.1 Poisson Statistics

In the ideal limit of no systematic uncertainty, the only statistical analysis concerns counting events, which is governed by the Poisson distribution. The uncertainty associated with random fluctuations in event counts due to the probabilistic nature of quantum mechanics is called the statistical uncertainty and is usually relatively straightforward to quantify.

The Poisson distribution gives the distribution for an integer random variable n and is given by

$$P(n;\nu) = \frac{\nu^n}{n!}e^{-\nu} \tag{3.6}$$

for the parameter ν. The expectation value and the variance of the Poisson distribution is just ν. In the case of event counting, $\nu = \sigma L \epsilon$, where σ is the cross-section, L is the luminosity, and ϵ is the selection efficiency. In the limit of large ν, the Poisson distribution will be approximately Gaussian.

In a typical search or measurement, we will predict the expected number of signal events, S, and the expected number of background events, B. However, the only quantity that is measured is the total count of events, $N = S + B$. The question is then whether N is significantly different from B. We want to measure how confident we are that the observed N is greater than expected from random fluctuations in B. Since B will be distributed according to a Poisson distribution, the standard deviation of B is just $\sigma_B = \sqrt{B}$. Thus to see how our observation compares to typical fluctuations, we compute $(N - B)/\sigma_B = S/\sqrt{B}$. Particle physics measurement significance is often presented in units of σ, the standard deviation, indicating how many standard deviations above the background estimate the observed counts were. Typically, 3σ is considered to be "evidence" for a process, while 5σ is the standard for a discovery. For a Gaussian distribution, 5σ from the mean indicates that the probability that the observation is a fluctuation of the background is just 2.87×10^{-7}. Conversely, for excluding new models, the standard is to use a "95% confidence," corresponding to 1.64σ, a much weaker standard due to the fact that consequences of a false exclusion are much less than a false discovery.

A real data analysis is much more complex than the simple analysis presented above, though the simple analysis provides good intuition. In reality, we often parameterize the total number of events with the signal strength μ such that $N = \mu S + B$. Then $\mu = 1$ usually corresponds to some nominal signal, such as the standard model signal, and $\mu = 0$ is the background-only hypothesis. In a measurement, the goal is to find the best-fit value of μ given the data and to quantify the quality of the fit. In a search that is designed to rule-out theories, the goal is to calculate the maximum μ for which we would have been able to observe a signal at 95% confidence.

In particular, in a real analysis, there are additional systematic errors, such as uncertainty in the selection of background model, the calibration of the detector, the uncertainty in the theoretical predictions, etc. These will increase the total uncertainty and at the LHC are often dominant over the purely statistical error. Thus the statistical analysis needs to account for these errors as well. Generically, systetmatic errors are constrained through the use of nuisance parameters. Each nuisance parameter corresponds to one source of systematic uncertainty. For example, in a background prediction, there may be one nuisance parameter for the background normalization and potentially a few nuisance parameters for the shape of various kinematic distributions (such as p_T's or masses of various objects) that induce correlations between histogram bins. Energy scale calibration is another significant source of nuisance parameters, including for example the uncertainty in efficiency of finding the primary vertex, uncertainty from different simulated event generators, uncertainty in the reconstruction algorithms, etc. There are many sources of systematic uncertainty and a robust analysis needs to account for all of the significant ones.

3.3.2 Likelihoods

In ATLAS, the statistical framework used to analyze signal strengths, nuisance parameters, and confidence levels is the principle of maximum likelihood. Suppose we have some probability density function that is parameterized by a signal strength μ and a set of nuisance parameters $\boldsymbol{\theta}$. For example, in a one-bin histogram, this might be the probability distribution of the possible counts n. Let us write the probability distribution generally as $f(x|\mu;\boldsymbol{\theta})$, where x is some measureable quantity, such as counts in histogram bins. This distribution is a model of the data and does not necessarily reflect the true underlying distribution. The model amounts to a hypothesis for the data.

We are then provided some dataset and we would like to know how consistent the model is with the data. One way to measure this is to compute how likely the observations are given the model. That is, suppose we have N independent observations, $\{x_i\}$. The probability of observing x_1 in the model is $f(x_1|\mu;\boldsymbol{\theta})$, and the probability of several indepedent observations is just the product of the probabilities. Thus we compute the probability of observing the entire dataset, assuming the model, as

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\mu; \boldsymbol{\theta}) \tag{3.7}$$

This function L is called the likelihood. Then to fit the model to the data, one approach we can take is to find the μ and $\boldsymbol{\theta}$ that maximize the likelihood of observing the data. That is, we find the parameters that maximizes the probability of observing the data, assuming the model. This is the method of maximum likelihood, which in many cases can be shown to produce an optimal fit.

When we are fitting counts in a histogram, we know that the underlying distribution is a Poisson distribution, so in this case our model can be very good. Suppose we have N bins of some histogram, indexed by $i = 1, ..., N$. From our Monte Carlo simulation, we can predict a number of signal events in each bin $s_i(\boldsymbol{\theta})$. s_i is a function of the nuisance parameters $\boldsymbol{\theta}$ because it will in general depend on energy scale calibrations, choice of generator software, etc. Similarly, we can make a prediction for the background events in each bin $b_i(\boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ will in general include, for example, the analysis-specific background normalization scale factor and correlations between bins. In addition, we often make measurements in a control region with M bins where we do not expect any signal. This can help to constrain the nuisance parameters. Suppose n_i is the observed counts in each of the signal region bins, m_i are the observed counts in the control region bins, and $u_i(\boldsymbol{\theta})$ is the prediction for the observed counts in the control region. Then we arrive at the baseline likelihood model used in the majority of ATLAS analyses,

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{[\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})]^{n_i}}{n_i!} e^{-[\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})]} \prod_{j=1}^{M} \frac{u_j(\boldsymbol{\theta})^{m_j}}{m_j!} e^{-u_j(\boldsymbol{\theta})} \tag{3.8}$$

Now, given the likelihood model and some histograms of data, we can scan the model parameters μ and $\boldsymbol{\theta}$ to find the values at which the likelihood is maximized. The values at which the likelihood is maximized are often labeled $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$, and these reflect our measurement of the signal strength and nuisance parameters. However, we would also like to quantify the uncertainty in the measurement and test the goodness-of-fit. We want to know for values of μ slightly different from $\hat{\mu}$ whether the data are still consistent, and at what point μ is sufficiently different that the likelihood is too small and the data are inconsistent. This will allow us to set confidence intervals and upper limits on theory parameters.

We will define the p-value, α, as the probability that given the data, the measurement $\hat{\mu}$ could have fluctuated to a new value μ. Typically, the cutoff in limits is made at $\alpha = 0.05$, corresponding to a μ at which, assuming $\hat{\mu}$ is true, we would expect to observe μ (or larger) instead only 5% of the time. In the case of a measurement where μ is bounded on both sides, this is called a 95% confidence interval. In the case of ruling out theory space, we often produce one-sided limits bounding μ from above, called upper limits.

To compute p-values for a given μ, we use the likelihood ratio test for goodness of fit,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{3.9}$$

where the denominator $L(\hat{\mu}, \hat{\boldsymbol{\theta}})$ is the likelihood at the global optimum $(\hat{\mu}, \hat{\boldsymbol{\theta}})$. The μ parameter in the numerator is the independent variable that we are testing. For each value of μ, we will find the nuisance parameters correspond to the maximum likelihood for that μ, denoted $\hat{\hat{\boldsymbol{\theta}}}(\mu)$. One can show that in the asymptotic limit of a large number of data samples, the statistic $t_\mu = -2 \ln \lambda(\mu)$ follows a known distribution, a non-central chi-squared distribution [37]. This distribution can therefore be integrated to find the probability of observing any given μ range given $\hat{\mu}$ measured in the data.

In this analysis, we do not expect to have any statistical power to observe a signal, so we will set an upper limit on μ. In particular, we expect that $\mu > 0$, so the test statistic should include this property. From [37], an appropriate statistic that accomplishes this,

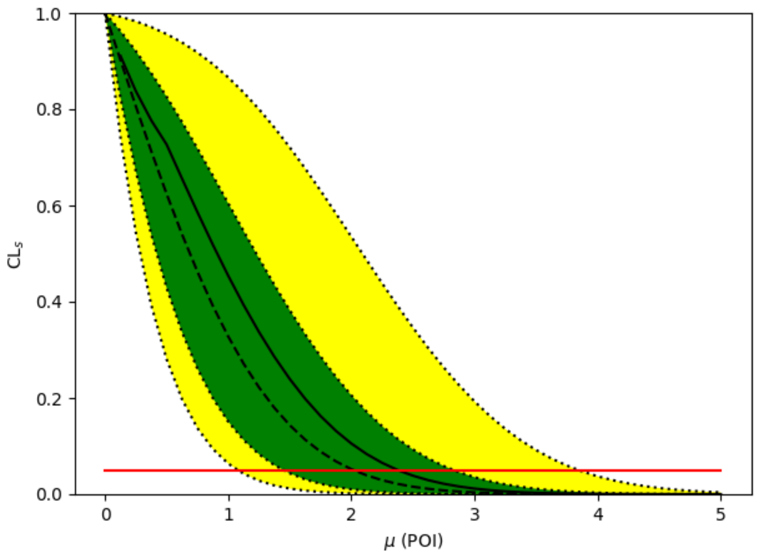


Figure 3.14: The calculated p-values (y-axis) for various upper-limit μ hypotheses, the parameter of interest (POI), for a toy model histogram with one bin. The computation is based on the $\tilde{q}_{\mu}$ test statistic. The dashed black line indicates the expected limit, computed with toy signal simulation, while the black line indicates the observed limit, computed from "data." The green band is the 1σ uncertainty on the p-value, and the yellow band is the 2σ uncertainty. Given a desired 95% confidence level, the intersection of the horizontal red line at $\alpha = 0.05$ with the black lines indicates the expected (dashed) and observed (solid) upper limits on μ. The green and yellow bands give the 68% and 95% uncertainty bands on the estimate of the upper limit.

whose distribution is also known analytically in the asymptotic limit, is

$$\tilde{q}_{\mu} = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{3.10}$$

The computation of this statistic, the probability distribution, and the p-value computation is implemented in the pyhf package used in this thesis [38]. An example plot of the p-value (and the uncertainty on the p-value) as a function of μ is for a toy analysis is shown in Figure 3.14.

3.3.3 Systematics

The above discussion concerns fitting the signal strength μ. However, in any fit, we also find the best-fit values for the nuisance parameters, which also need to be well-controlled.

Consider a single nuisance parameter θ. Often, before the analysis fit, this parameter is already constrained by previous measurements. Before the final fit, the parameter usually has some central estimated value θ_0 constrained by other measurements. This may be the energy scale calibration constant, efficiency, etc. If the calibration or control measurement is good, the constraint should be rather strong compared to any constraint from an analysis fit. An analysis generally will not have a better calibration measurement than the dedicated calibration analysis. Thus we expect that the best fit value of θ, denoted $\hat{\theta}$, should not be significantly different from θ_0. This is quantified in the pull, defined

$$\text{pull} = \frac{\hat{\theta} - \theta_0}{\sigma_\theta} \tag{3.11}$$

where the denominator is the uncertainty from the final fit. If the systematics are not overconstrained, then the pull should be distributed as a standard normal Gaussian, with central value of 0 and standard deviation of 1. Pulls significantly different from this indicate a problem in the fit, most typically an overconstraint, which means some systematic error has failed to be included. In this case, there is not enough freedom in the fit.

Another useful measure of the systematic errors that is often presented with the pulls is the impact. The idea is to vary the post-fit θ parameter by $\pm 1\sigma_\theta$ and check how the fit value $\hat{\mu}$ varies. When there are many nuisance parameter, each one is varied one at a time. This can be used to rank systematic errors, since the nuisance parameters with the largest impact indicate that $\hat{\mu}$ is very sensitive to their values.

Lastly, it is important to check nuisance parameter correlations. Very high correlations may indicate that some nuisance parameters are potentially redundant or have a complicated relationship that may not be well-controlled. One generally wants few parameters that are highly correlated.

3.3.4 Machine Learning

A key component of ATLAS data analysis is the use of machine learning. In the most general possible definition, machine learning is the use of algorithms that can be automatically optimized. For the purpose of this thesis, we will always be concerned with supervised learning. In supervised learning, we are given n known samples, and we try to model a function $Y = f(X)$ from input variables $X = \{x_1, x_2, ..., x_n\}$ to an output $Y = \{y_1, y_2, ..., y_n\}$. Note that in general the x_i and y_i can be tensors (of different sizes). Once the model as been trained on the n samples with known outputs, it can be applied to new input data to make predictions about the output.

If X and Y are continuous (or approximately continuous), the problem is called regression. The simplest regression model is a linear model, which with one output variable takes the form $y_i = \alpha \cdot x_i + \beta$ for a vector of parameters α and scalar offset β. The optimal parameters are solvable in an exact closed-form solution, where the optimum is defined as minimizing the least-squared error, $\sum_{i=1}^{n}(\alpha \cdot x_i + \beta - y_i)^2$.

Often, the output variable is categorical (and the inputs are still real numbers), such as a binary answer to a question (i.e. “is this shower consistent with an electron?”). In this case, the problem is known as classification. The simplest classification model is logistic regression. If we want to model a binary output, then we need a function that ranges between 0 and 1, the most common choice of which is the logistic function,

$$f(x_i) = \frac{1}{1 + e^{-(\alpha \cdot x_i + \beta)}} \tag{3.12}$$

where the parameters α and β are to be estimated from data, typically via a maximum-likelihood method in which the most likely values are found iteratively. The output of the model is viewed as a probability that a given an input x_i falls into one of the categories. Note that the logistic model is still based on a linear function in the exponential and is closely tied to linear regression. In particular, decision boundaries will always be flat hyperplanes in input space. The logistic model also generalizes well to multiple classification, such as in the closely related linear discriminant analysis.

Note that while both of these algorithms are nominally linear, basic non-linearities can easily be incorporated by providing, for example, the square of each variable or all second-order cross-terms. As long as this does not dramatically increase the dimensionality of the problem, this is sufficient in many cases.

In 90% of problems, linear and logistic regression will work perfectly well, and in the last 10% of cases, linear and logistic regression will achieve 90% of the best-case performance. However, when the absolute best performance is required and when there are strong non-linear dependencies among the inputs and outputs, more sophisticated algorithms are required that can model more complex functions, such as boosted decision trees and neural networks.

3.3.5 Boosted Decision Trees

The Boosted Decision Tree (BDT) has been a workhorse of ATLAS data analyses for many years. While it is beginning to be phased out in many cases by slightly more performant and more popular neural networks, BDTs are still a critical component of many analyses.

The basic element of a BDT is the decision tree, many of which will be combined in a particular way that significantly improves performance. A decision tree is a rather simple object, a series of one variable binary splits of the data, categorizing each data point into exactly one category. The structure of a decision tree is shown in Figure 3.15. The net effect is that the input parameter space is split into rectangular regions.

In each region, the output of the decision tree is selected to be the best fit of the known training outputs. For a regression problem, the output is the mean of all samples in that region. For a classification problem, the output is the most likely category in that region.

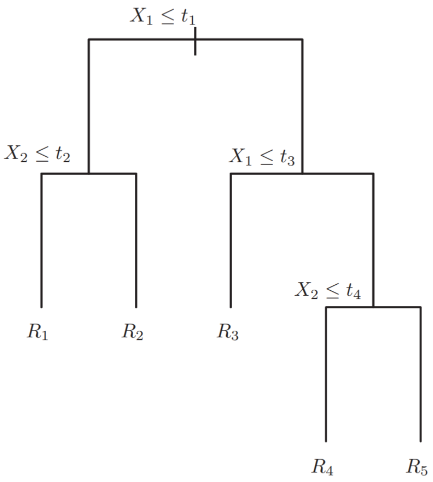


Figure 3.15: A simple decision tree on two variables, X_1 and X_2, resulting in five output regions, the output of which is selected to be optimal for that region [9].

In the case of regression, the tree is optimized by minimizing the residual sum of squares, $\sum(y_i - f(x_i))^2$ across all the regions. In the classification case, the error is often taken to be the total entropy, which for K classes and M output regions, is $-\sum_{m=1}^{M}\sum_{k=1}^{K} p_{mk} \log p_{mk}$, where p_{mk} is the fraction of samples in region m that fall in class k.

Typically, decision trees are grown using a greedy algorithm, such that at each stage, the variable and threshold are selected such that the error is reduced as much as possible. A stopping rule such as requiring a minimum number of samples in a region determines when to stop adding branches. Often, the tree is then pruned using a regularization algorithm that penalizes the total error for very large trees, such as by adding in a term to the error proportional to the number of regions.

Decision trees on their own, while easy to interpret, often have limited accuracy. Boosting is a general method of combining several copies of relatively weak learning algorithms to create a much better fit. In boosted decision trees, first a decision tree is fit to the data, resulting in a fit function $f_1(x_i)$ for the data set $\{x_i, y_i\}$. Then, for each data point, the fit residuals are computed with f scaled by a learning rate scalar λ, $r_i^1 = y_i - \lambda f_1(x_i)$. Then, another tree, $f_2(x_i)$ is fit to the residuals $\{x_i, r_i^1\}$, and the next iteration of residuals are computed, $r_i^2 = r_i^1 - \lambda f^2(x_i)$. This process is then continued iteratively until a good fit is acheived (for example, by checking a validation sample of data). For B boosting iterations, final output is then simply $f(x_i) = \sum_{n=1}^{B} \lambda^n f_n(x_i)$.

3.3.6 Neural Networks

Over the past 10 years or so, neural networks have exploded in popularity in industry and in academia, largely for their ability to create excellent predictions over large data sets with relatively little human input. In particle physics, one of the main attractive features of neural networks is their ability to capture non-linear correlations between variables. This is particularly useful in particle physics because of the nonlinear correlations between four-vectors, induced by underlying physics such as the conservation of momentum; complex phase-space probability distributions from cross-sections; or simply the fact that energy, momentum, and mass are related quadratically, not linearly.

Feed-Forward Neural Networks

The simplest and prototypical example of a neural network is the feed-forward neural network, which is also sometimes called somewhat obfuscatingly a multilayer perceptron. The architecture of a feed-forward neural network is meant to abstractly and very approximately represent the interconnection of neurons in the brain. Each neuron in the network receives inputs from n other neurons, with each input weighted by some learned value. The sum of the weighted inputs is then passed into an activation function, which roughly acts as a switch. In the simplest case of a step function activation function, if the inputs are above some threshold, the activation function outputs 0, and if they are below the threshold, it outputs 1. The threshold is a learned parameter for each neuron and often called the bias. The step function tends to behave non-optimally because of its discontinuity at the threshold. The most common activation function is probably the ReLU (Rectified Linear Unit) function, which outputs 0 below the threshold and outputs the input ($y = x$) above the threshold. Other common functions include the logistic function (sigmoids) or the hyperbolic tangent.

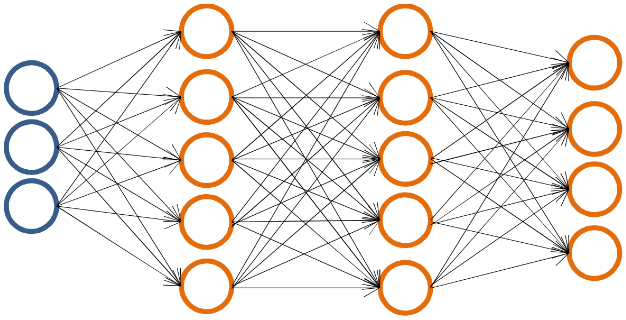


Figure 3.16: A fully-connected feed-forward neural network, with 3 inputs, two hidden layers with five neurons each, and four outputs [10].

In the simple feed-forward network, the neurons are arranged in layers, with neurons in one layer receiving inputs only from the previous layer. When every neuron in a layer is connected to every neuron in the previous layer, then the network is said to be fully-connected. The layers of the neural network that are not inputs or outputs are called the hidden layers, and networks with many hidden layers are called deep neural networks. Figure 3.16 is an example of the architecture of a fully-connected feed-forward neural network.

A typical neural network has a large number of weights that need to be learned. The general approach is the method of gradient descent. First, the performance of the network over a training data set is scored with a specified function, call the loss function. The output of the loss function is a scalar, the loss, that should be decreased to improve the fit of the network. A common loss function for regression problems, for example, is just the sum of squared residuals between the known training outputs and the network predictions, $\sum(y_i - f(x_i))^2$. In gradient descent, the gradient of the loss function is computed with respect to all of the weights in the network, and the weights are updated in the direction of steepest descent. The gradient is computed through back-propagation, in which the gradients of each layer are iteratively computed using the chain rule. In many cases, the gradient function can be computed analytically, which reduces the computational load. In each update of the weights, the input data are fed through the neural network again to produce a new loss and a new gradient, and the weights are then updated again. Each cycle is called a training epoch. In moderate to large networks, training can often be a bottleneck of neural network performance, especially in the absence of hardware acceleration such as GPUs.

A network that has not approached the loss function minimum and that therefore badly models the data is said to be underfit. On the other hand, neural networks are sufficiently flexible that overfitting is a major concern as well, where the neural network can predict each training data point very well, but is incorrect when given new data. Overtraining is often monitored by checking the loss function over a validation data set. As the network is trained, the validation loss will decrease as the fit improves. If the network begins to be overtrained, the validation loss will rise, creating the classic U-shaped overtraining curve. To minimize overtraining, one can employ a regularization scheme. One class of regularization methods adds a term to the loss function that penalizes large weights. Another common approach is dropout, in which some fraction of randomly-selected connections between neurons is removed during each training epoch.

A feed-forward network is relatively simple but rather effective in most cases. However, in some situations, it is insufficient, either because the input data are simply best represented in a different form (such as an image or text), or because different architectures can leverage properties of the problem. A convolutional neural network (CNN) is designed to process image data by grouping neighboring pixels via a kernel function to produce a single output. A recurrent neural network (RNN) is similar to a feed forward neural network, except that the outputs can be fed back as an input, which allows the network to have memory while processing a sequence of data.

Attention and Transformers

A popular concept in contemporary machine learning is the idea of attention, which was popularized by the transformer neural network proposed in 2017 [11].

Suppose each data sample we have consists of a set of input vectors. We will be concerned with the case where each of the input vectors contain information about a jet in the event, such as the four-momentum. We will also assume for simplicity that we have a fixed number of N input jets, which I denote x^i for $i = 1, ..., N$. The self-attention mechanism is a network architecture that can easily learn the relationships between the different x^i input vectors.

In self-attention, each input vector x^i is first multiplied by learned weight matrices, Q, K, and V to produce a query vector $q^i = Qx^i$, a key vector $k^i = Kx^i$, and a value vector $v^i = Vx^i$. The Q,K, and V matrices are to be learned and may be constrained (such as by setting $K = V$). Jet i is then compared to jet j by comparing the query vector for jet i with the key vector for jet j. This comparison is done via a dot product, $d^{ij} = q^i \cdot k^j$, corresponding informally to information asked *by* jet i *about* jet j. We say jet i attends to jet j. Each of these products is then run through a (scaled) softmax function in order to normalize the outputs for each query,

$$a^{ij} = \frac{\exp\left(d^{ij}/C\right)}{\sum_{k=1}^{N} \exp\left(d^{ik}/C\right)} \tag{3.13}$$

where C is a scaling factor, typically taken to be the square-root of the dimension of the key/query vectors, which ensures the dot-products aren't large for high-dimensional problems. Finally, for each jet i, we compute an output vector h^i by taking the weighted sum of the value vectors for every jet, weighted by the dot-product attention, $h^i = \sum_{j=1}^{N} a^{ij} v^j$.

This process can then be iterated with the output vectors h^i as inputs to another self-attention layer. In the first iteration, the network can learn pair-wise information. In the second iteration, the network can learn information including up to four of the input vectors, and this sharing of information repeats on each layer. The final output of the network will depend on the application. In this analysis, we will use the (normalized) self-attention scores directly to select which pairs of jets should be grouped together. With this output, it is then straightforward to derive a loss function based on whether the grouping is correct,and use the standard auto-differentiation with gradient descent to train the network.

In multi-head attention, we allow the network to learn multiple sets of query, key, and value weight matrices, one set for each "head". This allows the network to learn different

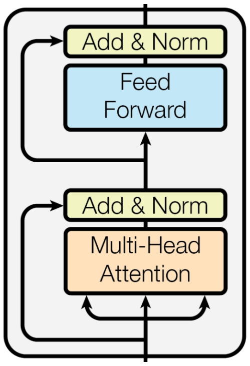


Figure 3.17: The transformer encoder layer from [11]. The inputs to the attention layer are directly added to the outputs in order to preserve the input information, and the normalization ensures the outputs stay in a small range, making training easier. The feed-forward component is a shallow network that allows additional processing of the attention information, with the residual input connection for the same reason.

definitions of attention. In the language processing setting, one definition of relevance may be just the previous word in the sentence. Another definition may learn something more complex, such as the nearest noun, for example.

A transformer is just a particular arrangement of multi-head self-attention layers. The classic transformer features an encoder and decoder. The encoder is a set of self-attention layers that transform the inputs into some abstract space containing information about the relationships between the vectors. The decoder does the reverse, taking the abstract vectors, feeding them through self-attention layers, and producing the desired output. The original use of the transformer had translation in mind, in which an input sentence in English is encoded into a high dimensional sentence space before being transformed in the decoder into the representation of that information in French, for example. An example encoder layer is shown in Figure 3.17.

In our data analysis, a particularly appealing feature of self-attention is that it can be made to be order-independent. This means we can, for example, feed the network a set of particle four-vectors which in general do not have any natural ordering, and the network can learn the relationships between them in an order-independent way.

CHAPTER 4

OBJECT DEFINITION AND CALIBRATION

4.1 From Bare Quarks to B-Jets

Since we are searching for the $hh \to 4b$ process, we are looking for a final state of four b quarks. However, identifying the presence of a b-quark is challenging, since b-quarks are not directly detectable, instead appearing in the detector as complex b-jet objects.

4.1.1 Jets

The first challenge with having a quark in the final state is that bare quarks (or gluons) are not measureable. In QCD, the strength of the strong coupling constant increases with distance so that the strength of the strong interaction for quarks separated by about a femtometer (10^{-15} m) is at an energy scale comparable to the mass of the light quarks. Increasing the separation beyond this means that it becomes energetically favorable for a gluon from the interaction to split into a quark-antiquark pair to neutralize the color charge to reduce the interaction strength between the two original quarks. The end result is that we have two separated hadrons (particles composed of at least two quarks) instead of just two separated bare quarks.

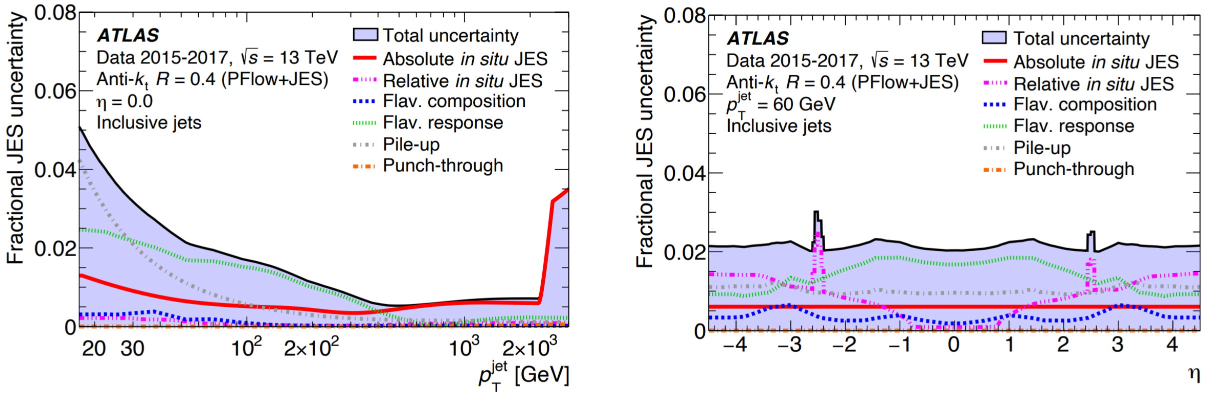
In particle collisions, this process is known as hadronization. Any quark or gluon that is produced in the hard scatter will fly away from the interaction point at effectively the speed of light. The large energy associated with the bare quark at distance scales of femtometers means that many quark-antiquark pairs will be created, resulting in a collection of many hadrons all traveling in approximately the same direction. Many of these hadrons are stable with respect to detector length scales, such as pions and kaons, or will almost immediately decay into stable hadrons. Thus for a bare quark, what is actually observed in the detector is a collection of collimated particles, mostly charged pions, with some contribution from photons (from $\pi^0 \to \gamma\gamma$ decays), charged kaons, and neutral hadrons. To measure information about the initial quark, these particles will be grouped together algorithmically in some way, and such a grouping of particles is called a jet.

There are many different jet reconstruction algorithms that have many different theoretical and experimental motivations. Ultimately, any jet definition will have some degree of arbitrariness to it, since exactly which particles come from the bare quark hadronization rather than other interaction-point QCD is not theoretically well-defined. Such issues, however, are usually corner cases and the jet concept works well.

The current state-of-the-art jet reconstruction algorithm in ATLAS is called "particle flow" (or "p-flow" for short). Traditional jet algorithms cluster together energy deposits in the calorimeters with some distance and momentum cutoff. These methods have several disadvantages, the most prominent of which at the LHC is complications due to pileup. The other soft proton-proton collisions in the event create a background of soft particles, and since the calorimeter alone cannot distinguish between different interactions, it is a challenging problem to filter out this background.

Particle flow algorithms, however, use tracking information in addition to the calorimeter and can reject pileup much more effectively. Additionally, while the calorimeter provides a more accurate measure at high energies, soft particles are more accurately measured in the tracker. Thus an ideal algorithm combines these two. However, to avoid double-counting, one needs to match particle tracks to clusters, measuring the "flow" of the particles. In practice, this is quite challenging due to the dense environment. A frequent occurrence, for example, is two particles sharing a calorimeter cluster, or the reverse – a single particle creating two calorimeter clusters. The details of the ATLAS algorithm that account for this can be found in [39].

The uncertainties in the measurements of jets are categorized into two primary sources, the jet energy scale (JES) and jet energy resolution (JER). The jet energy scale is effectively a calibration constant from the detector output voltages to jet energy in GeV. The current



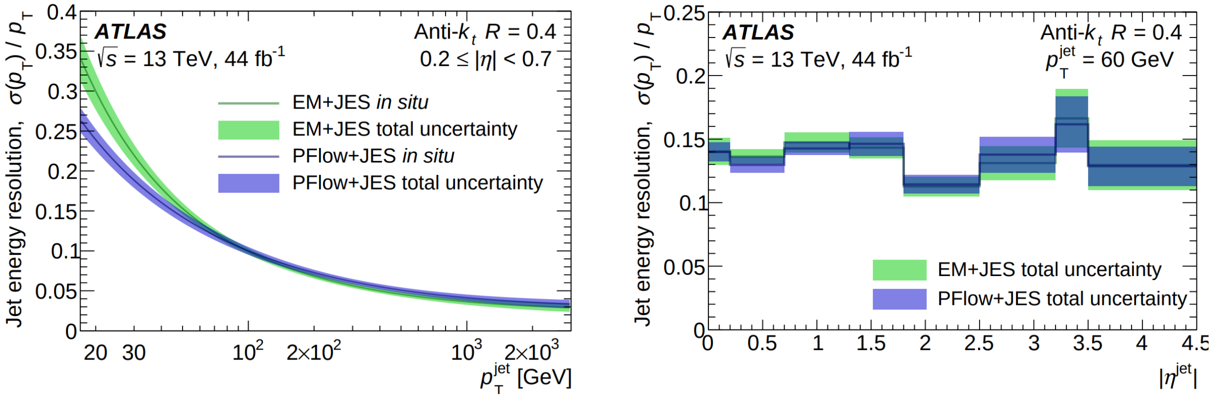
(a)

(b)

Figure 4.1: The Jet Energy Scale uncertainty as a function of p_T (a) and η (b), with the most important contributions shown [12].

method of measuring the JES is a two-stage procedure. First, Monte Carlo simulations of jets are used to scale the simulated detector response to the known four-momenta of the truth particles. Then, an additional MC-to-data correction is derived *in situ* using momentum balance methods, where jets are balanced against well-measured objects such as leptons and Z bosons, or at high energies, other low energy jets. Note that JES calibration does *not* usually calibrate jets so that their momentum matches the "true" quark momentum, since this is not measureable. The goal is to match the simulated jet response to the measured jet response in data. Variations due to uncertainty on the jet energy scale will tend to shift the energy of all the jets in the data sample in the same direction. The JES uncertainty as a function of p_T and η is shown in Figure 4.1.

The jet energy resolution is the uncertainty on a single-jet energy measurement due to random fluctuations in the jet composition, detector measurements such as hadronic shower shape, and jet algorithms. Variations due to the jet energy resolution uncertainty will tend to spread jet energies in both directions, broadening distributions. The jet energy resolution is shown in Figure 4.2. Quantifying and reducing the JES and JER uncertainties is a challenging task with a large number of nuisance parameters, and these tend to be the dominant uncertainties for many analyses with hadronic final states.



(a) (b)

Figure 4.2: The Jet Energy Resolution as a function of p_T (a) and η (b) for EMTopo (old) and PFlow (current) jet reconstruction algorithms [12].

4.1.2 B-Jets

So far, the above discussion applies to jets initiated from any flavor of quark or a gluon (except for the top quark, which decays in less than a femtometer, before hadronization occurs). QCD, however, is flavor-conserving, and this can be leveraged to distinguish some flavors of quarks. For light quarks, up and down quarks, flavor conservation is not useful. There will be many, many up and down quarks created via hadronization. However, for heavier quarks, such as b-quarks, c-quarks, and to some extent, strange quarks, these will not be created in hadronization. Thus if we can identify hadrons containing these quarks, we can distinguish the initial quark flavor, a procedure known as flavor tagging.

In practice, the easiest (and generally most useful) flavor to tag is that of b-quarks, known as b-tagging. A b-tagged jet is called a b-jet. In a b-jet, the initial b-quark forms a B hadron which has a lifetime of order $c\tau = 0.5$ mm before decaying into several other hadrons. At LHC energies, this lifetime is Lorentz boosted to several millimeters. Thus the characteristic feature of a b-jet is the presence of an additional vertex from the B hadron decay products that is displaced several millimeters away from the primary interaction point. Figure 4.3 is a cartoon drawing of a displaced vertex. Thus particle tracking to find the vertex is a critical component of b-tagging.

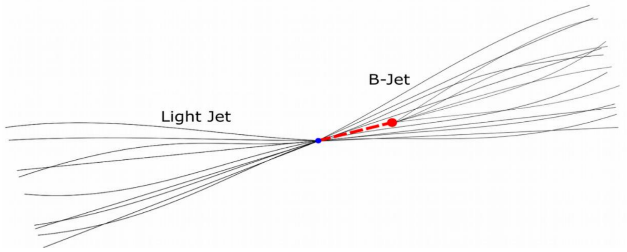


Figure 4.3: A cartoon drawing of a light jet and a b-jet originating from one primary vertex.

The challenge of b-tagging is to discriminate b-jets from c-jets, light jets, and τ-jets. Jets that contain a charm quark are called c-jets. During hadronization, the charm quark will form a D meson, which has a lifetime comparable to but slightly shorter than a B meson. Thus c-jets also tend to have secondary vertices and distinguishing b-jets and c-jets is challenging. In addition, some fraction of B-mesons decay into a D meson, so b-jets will often include D mesons and even an associated tertiary vertex. Light jets, while easier to distinguish, are much more numerous, and so a much better rejection is necessary. τ jets are primarily jets from the $\tau \to 3\pi + \nu$ decay that may also have other hadronic particles present from the primary vertex or pileup. These are usually less of a concern because τ's are rarer than light jets and easier to distinguish than c-jets, but they still must be considered.

Currently, ATLAS uses a two-stage b-tagging algorithm. First several low-level taggers are run that generally use a single b-tagging strategy. Then the outputs and key information from the low-level taggers are combined into a single high level tagger, currently a deep neural network called DL1 that has much better performance than any individual low-level tagger. The three main low-level taggers are RNNIP, the Soft Muon Tagger (SMT), and JetFitter. In the RNNIP, track impact parameters (the distance of closest approach to the interaction point, extending the track infinitely back) are fed into a recurrent neural network. A recurrent neural network naturally handles correlations between tracks better than previous algorithms (i.e. given one track with a large impact parameter, it is more likely that another track in the same jet also has a large impact parameter). The Soft Muon Tagger takes advantage of the fact that around 21% of B meson decays produce a muon.

It outputs simple variables such as the ΔR between the muon and the b-jet, the muon p_T, and some muon track quality variables. Lastly, the JetFitter algorithm reconstructs the topology of the jet tracks along the jet, finding any additional secondary or tertiary vertices. Finally, outputs or intermediate variables from each of these algorithms are combined as inputs into another machine learning algorithm. The modern state-of-the-art high-level b-tagger is based on a feed-forward neural network and is called DL1. The addition of the soft muon tagger can complicate the calibration, so DL1 without the SMT is called DL1r, and with the SMT, it is DL1rmu. DL1r is the current baseline version. An older algorithm based on a boosted decision tree and that was used in the trigger in Run 2 is called MV2.

The output of a b-tagger is ultimately a score between 0 and 1, with scores near 1 likely to be a b-jet and scores near 0 likely not a b-jet. Then a cut value can be chosen above which an object is called a b-jet and below which it is rejected. A given cut value will have some b-tagging efficiency and light/charm jet rejection based on the performance of the algorithm. The possible values of efficiency and rejection as the cut value is scanned are summarized in ROC curves, shown in Figure 4.4.

Several different cut values are selected as standard, and full calibrations are completed for these values, where the efficiency as a function of p_T and η is measured in data and in simulation, typically in $t\bar{t} \rightarrow \ell\ell\nu\nu bb$ events, so that the differences are understood. These events are chosen because $t\bar{t}$ pairs are copious at the LHC and this decay can be efficiently reconstructed without relying on b-tagging. Since top quarks decay to b-quarks 99.8% of the time, we know the two jets are true b-jets in data, and we can measure how frequently they pass the tagging.

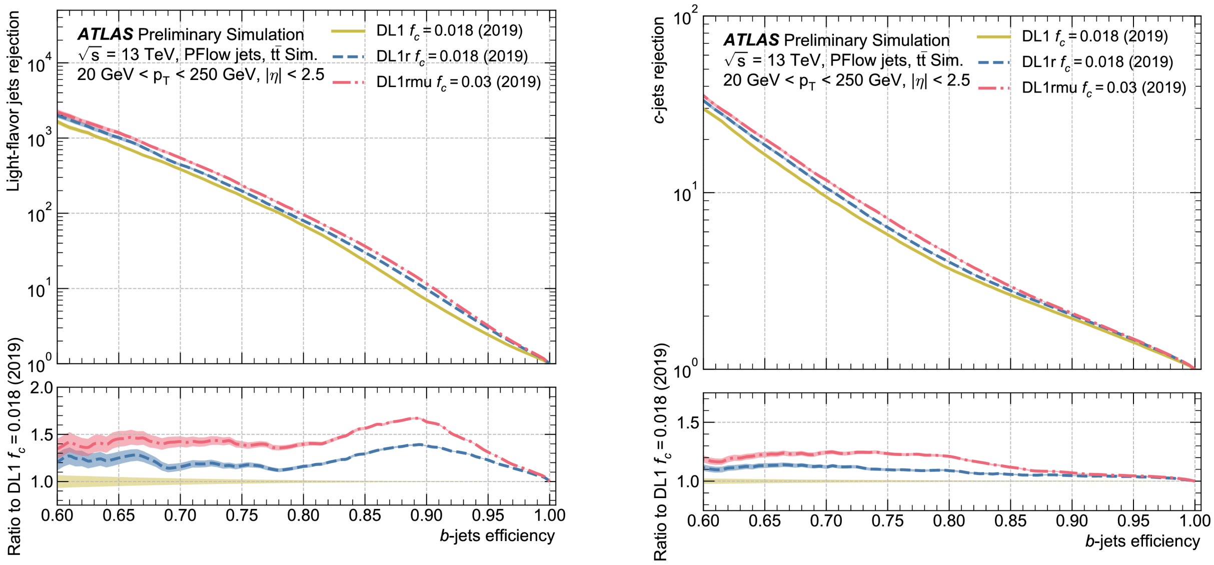
Additionally, the efficiency of the b-tagging algorithms need to be measured on light jets, charm jets, and τ jets in order to correctly model the fake rate. Measuring the light jet efficiency in data is relatively easy, given the very large number of light jets present. Similarly, because of hadronic τ-tagging, measuring the τ efficiency is relatively straightforward. The hardest quantity to measure is the c-jet efficiency, since there is no pure source of c-jets.

O.P.	b-jet	c-jet	l-jet	τ-jet
60%	61.1%	3.4%	0.096%	0.38%
70%	70.8%	1.0%	0.26%	1.4%
77%	77.6%	20%	0.61%	5.3%
85%	85.3%	50%	2.5%	25%

Table 4.1: The average efficiency of the DL1r b-tagger at different operating points (O.P.) for different objects, averaged over $t\bar{t}$ events.

There are two main approaches. The first and older approach is to use $c + W$ events, where the W decays leptonically. To distinguish the c jet, one either checks for a long-lived D meson or a muon from the c decay, either of which will need to be opposite the sign of the lepton from the W decay. The major downside of this approach is that specific decay chains are required, which increases the extrapolation uncertainty to inclusive c jets. The other approach uses $t\bar{t}$ events and is inclusive. Approximately 50% of W boson decays include at least one charm quark. Since this depends on the well-measured CKM matrix, the fraction of each flavor is well-predicted. One can then run b-tagging over a sample of W-boson decays, and given well-measured b-jet and light-jet efficiencies, the charm-jet efficiency can be inferred. In reality, this is a tricky procedure since the $t\bar{t}$ reconstruction biases the flavor content, and it requires a dedicated analysis [40].

In this round of the $hh \to 4b$ analysis, we use the 77%-efficient working point as baseline, after a scan of the options in the early stages of the analysis showed this to be optimal, though only slightly better than the 70% working point. Note that the "77%" indicates that the algorithm was, on average, 77% efficient on tagging b-jets over a particular $t\bar{t}$ sample. The efficiency as a function of jet p_T is shown in Figure 4.5, and the efficiency of each working point for various objects is shown in Table 4.1.



(a)

(b)

Figure 4.4: B-tagging ROC curves for (a) light-jet rejection and (b) charm-jet rejection. "DL1" uses an older impact parameter low-level algorithm based on likelihoods instead of the RNNIP algorithm based on a neural network. DL1r uses RNNIP, and DL1rmu uses RNNIP and the SMT.

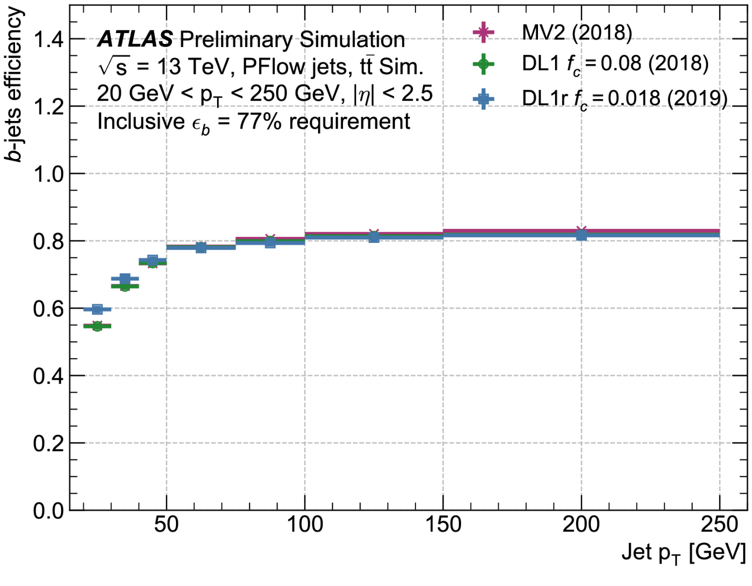
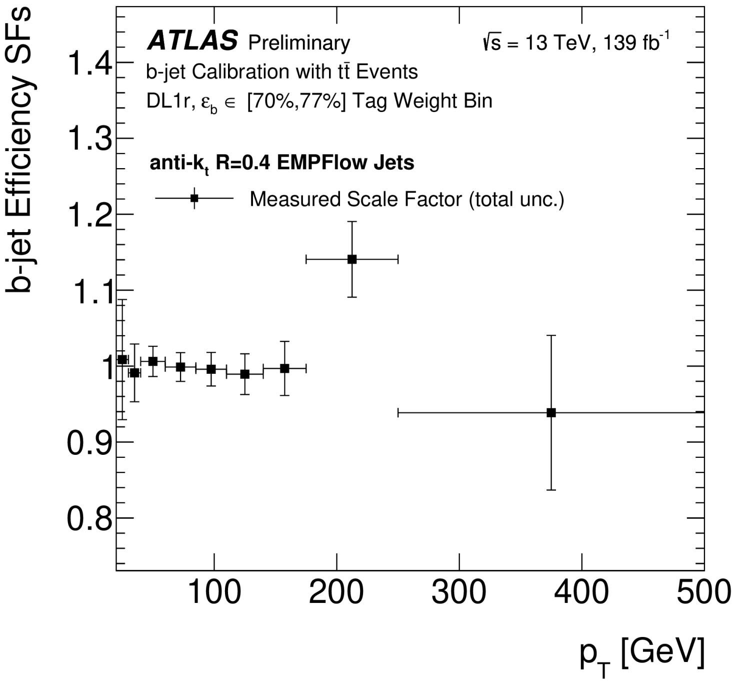


Figure 4.5: The b-tagging efficiency at the 77%-efficient working point as a function of p_T. MV2 is an older high-level algorithm that uses a BDT instead of a neural network. "DL1" uses an older impact parameter low-level algorithm based on likelihoods instead of the RNNIP algorithm based on a neural network.

4.2 B-Jet Calibration and Scale Factors

4.2.1 Introduction to Scale Factors

In ATLAS data analysis, we frequently analyze objects that are identified with some limited efficiency. When these objects are used to derive conclusions about physics, it is important that the efficiencies are well-understood. In particular, the efficiencies are usually slightly different when measured in data compared to when measured with Monte Carlo Simulation, and correcting for this is referred to as object calibration. Typically, calibrations are provided for single objects, and analyzers extrapolate from a single-object calibration to collisions that may contain several objects. This section demonstrates general features in this problem of extrapolating from one calibrated object to multiple calibrated objects. Throughout this section, I will use the example of the b-tagging, but this example generalizes easily to other objects.



(a)

Figure 4.6: The data/MC scale factor as a function of p_T for the baseline DL1r tagger. Note that the calibration at high p_T is challenging due to limited statistics.

B-jet calibration is provided via scale factors, which are defined to be the ratios of efficiency in data to efficiency in Monte Carlo for a given b-jet. Because efficiency depends on the parameters of the jet, such as the transverse momentum (p_T) and pseudorapidity (η), the scale factors will as well. Figure 4.6 shows an example of b-tagging scale factors as a function of jet p_T. Note that in many cases, scale factors are close to 1 and are therefore the most important when either precision is required or there are many b-jets.

4.2.2 One-Jet Events

To understand the basic use of scale factors, consider an ensemble of simulated events with exactly one (true) b-jet and no other objects. The goal is to perform b-tagging on this ensemble to estimate the expected number of events in data with one b-tag.

Assume for simplicity that all the jets are similar enough kinematically to have the same efficiency. In addition, an important, nontrivial assumption is that the total number of simulated events before b-tagging is equal to the total number of data events before b-tagging. This is a statement of luminosity normalization, which depends both on accurate simulation and accurate measurement of the total amount of collected data.

This problem is usually approached at the event-level by considering scale factors as modifying the weight of Monte Carlo events. If the scale factor for a given event is 0.95, for example, then the event weight is reduced by a factor of 0.95 so that this event is not falsely over-represented in the final histograms.

I claim instead that because ATLAS statistical analysis is based on event counts, the only important quantity is the estimate of the number of events in data. Therefore, I will approach the problem instead from the perspective of ensembles of events, which will be extremely useful for more complicated problems. Define N to be the total number of MC (or data) events before b-tagging. Let ϵ^{MC} be the b-tagging efficiency in simulation, and ϵ^{data} be the efficiency in data. Then, an estimate for the number of b-tagged events in MC and data is

$$
\begin{aligned}
N_{MC}^{tagged,est} &= \epsilon^{MC} N \\
N_{data}^{tagged,est} &= \epsilon^{data} N
\end{aligned}
\tag{4.1}
$$

However, after we run b-tagging on our simulation, we know N_{MC}^{tagged} and don't need to rely on an estimate. Instead, we can invert the problem and use N_{MC}^{tagged} to estimate N.

$$
N^{est} = \frac{N_{MC}^{tagged}}{\epsilon^{MC}} \tag{4.2}
$$

Plugging N^{est} for N in the second line of equation 4.1,

$$
N_{data}^{est,tagged} = \frac{\epsilon^{data}}{\epsilon^{MC}} N_{MC}^{tagged} \tag{4.3}
$$

The scale factor is defined to be the ratio of efficiencies that allows us to scale from MC to data.

$$
SF \equiv \frac{\epsilon^{data}}{\epsilon^{MC}} \tag{4.4}
$$

Note that equation 4.3 implicitly assumes a large enough N that the estimates are accurate. In reality, there will be binomial fluctuations, which we will examine more generally in section 4.2.6.

Another assumption that we will make throughout is that there is no uncertainty in the scale factor and efficiency measurements. This is merely because these are "standard" sources of uncertainty that are handled like other nuisance parameters in the analysis, i.e. through manual variations or through a full likelihood fit.

4.2.3 B-tagging Two Jets

Suppose we now have simulated events with exactly two truth b-jets, and we require both b-jets to be b-tagged. We'd like to derive a scale factor for this slightly more complicated process.

These two jets generally do not have the same kinematics. For simplicity, we will assume for now that the scale factor is entirely determined by the transverse momentum p_T (but the discussion generalizes easily when, for example, η is included). For one jet, we can derive a scale factor for every p_T bin, but for two jets, we'll need a scale factor for each possible combination of bins. This means that the scale factors are defined over the tensor product of the p_T bins. Concretely, suppose there are P bins of p_T, which can be indexed by $\alpha = 1, 2, ..., P$. In the one-jet case, our scale factor has one index for each p_T bin, SF_α In the two-jet case, the scale factor now has an index for both of the jet p_T's, $SF_{\alpha\beta}$. This means we will have P^2 different possible scale factors for two-jet events.

Let $N_{\alpha\beta}$ denote the number of events with jet 1's p_T in the α bin and jet 2's p_T in the β bin before any b-tagging. Let M represent the number of events with 2 b-tags in Monte-Carlo, and D represent the number of events passing the tagging requirement in data.

How do we compute the scale factor for this example? It's relatively intuitive. Assuming no correlations [1], the chance of jet 1 being tagged and jet 2 being tagged is just the product of the efficiencies. Thus

$$\begin{aligned} M_{\alpha\beta} &= \epsilon^{MC}(\alpha)\epsilon^{MC}(\beta)N_{\alpha\beta} \\ D_{\alpha\beta} &= \epsilon^{data}(\alpha)\epsilon^{data}(\beta)N_{\alpha\beta} \end{aligned} \tag{4.5}$$

where I have defined $\epsilon(\alpha)$ as the efficiency of tagging jet 1, given its p_T is in bin α. Similarly, the efficiency of tagging jet 2 is $\epsilon(\beta)$.

[1] This is a good assumption because b-tagging relies only on the internal structure of a jet. Jet formation occurs at a late time in the collision, and so will tend to be spatially and temporally separated from the formation of other jets. Examining the small higher-order corrections to this universal assumption in jet physics is an interesting and challenging line of research.

We can then write the scale factor as

$$SF_{\alpha\beta} \equiv \frac{\epsilon^{data}(\alpha)\epsilon^{data}(\beta)}{\epsilon^{MC}(\alpha)\epsilon^{MC}(\beta)} \tag{4.6}$$

such that

$$D^{est}_{\alpha\beta} = \frac{\epsilon^{data}(\alpha)\epsilon^{data}(\beta)}{\epsilon^{MC}(\alpha)\epsilon^{MC}(\beta)} M_{\alpha\beta} \tag{4.7}$$

This scale factor has the nice property of being just the product of the scale factors of each individual jet.

$$SF_{\alpha\beta} = SF_{\alpha} SF_{\beta} \tag{4.8}$$

Note that this discussion will generalize quite naturally to the case of J true b-jets when we require all of them to be tagged.

4.2.4 Two Jets with at Least One B-tag

Suppose we now have a sample of Monte Carlo events with two truth b-jets, and we request that at least one of the jets is b-tagged. As an example, this situation may occur when there are multiple rounds of b-tagging, particularly when using b-jet triggers. In this case, the first round of b-tagging is correlated but not identical to the second round, so to maintain maximal efficiency, the first round of b-tagging may require fewer b-tags.

In the above simple examples, it is rather intuitive how to get from jet-level to event-level scale factors. There is one well-defined scale factor to use for all the events falling into a given set of p_T bins. In this example, however, there are several different approaches that one might consider taking.

4.2.4.1 Approach 1: Check every jet

One way we could derive scale factors for the “two jets, at least one tagged” case is to derive a scale factor based on what jets are and aren’t tagged. That is, we check both jets, and if the jet is tagged, we include a factor of ϵ; if the jet failed to pass tagging, we include a factor of $1-\epsilon$. We do this to account for the fact that the chance of failure will be different in Monte Carlo versus in data.

There are three ways to obtain at least one b-tag among two jets, so we will consider each possibility as a separate category. Therefore, the number of Monte Carlo events in each category will be

$$\begin{aligned}
M^1_{\alpha\beta} &= \epsilon^{MC}(\alpha)\epsilon^{MC}(\beta)N_{\alpha\beta} \\
M^2_{\alpha\beta} &= \epsilon^{MC}(\alpha)\epsilon^{MC}(\bar{\beta})N_{\alpha\beta} \\
M^3_{\alpha\beta} &= \epsilon^{MC}(\bar{\alpha})\epsilon^{MC}(\beta)N_{\alpha\beta}
\end{aligned} \tag{4.9}$$

where a bar indicates the probability that a jet in that p_T bin is not b-tagged. That is, $\epsilon(\bar{\alpha}) = 1 - \epsilon(\alpha)$. The superscript on M indicates the category we are considering. Thus in this example, in category 1, both jets are tagged; in category 2, jet 1 is tagged and jet 2 is not; and in category 3, jet 1 is not tagged and jet 2 is tagged. There is an analogous expression for the data event counts.

Writing down the scale factors for each of the categories is straightforward.

$$\begin{aligned}
SF^1_{\alpha\beta} &= \frac{\epsilon^{data}(\alpha)\epsilon^{data}(\beta)}{\epsilon^{MC}(\alpha)\epsilon^{MC}(\beta)} \\
SF^2_{\alpha\beta} &= \frac{\epsilon^{data}(\alpha)\epsilon^{data}(\bar{\beta})}{\epsilon^{MC}(\alpha)\epsilon^{MC}(\bar{\beta})} \\
SF^3_{\alpha\beta} &= \frac{\epsilon^{data}(\bar{\alpha})\epsilon^{data}(\beta)}{\epsilon^{MC}(\bar{\alpha})\epsilon^{MC}(\beta)}
\end{aligned} \tag{4.10}$$

Therefore, for each p_T bin, we have three possible scale factors, depending on how the event is tagged, and one just needs to apply the appropriate scale factor based on which jets pass b-tagging.

4.2.4.2 Approach 2: Check until satisfied

This above approach might seem natural to some but arbitrary to others. A common way that scale factors are applied in ATLAS is to check one jet at a time until the tagging requirement is satisfied. In our example case, we will end up with two categories. If the first jet is tagged, we're done, and our scale factor is derived from the first-jet efficiency. If the first jet is not tagged, then we check the second jet. If the second jet is tagged, then we derive our scale factor from the chance of the first jet failing and the second jet being tagged.

The motivation for this method is the idea that not all of the tagging information for every jet should matter. Once we know that jet 1 is tagged, it doesn't matter whether jet 2 is tagged or not – the event will be accepted either way.

Explicitly, there are two possibilities with the following expected number of Monte Carlo events in each category

$$\begin{aligned} M^1_{\alpha\beta} &= \epsilon^{MC}(\alpha) N_{\alpha\beta} \\ M^2_{\alpha\beta} &= \epsilon^{MC}(\bar{\alpha})\epsilon^{MC}(\beta) N_{\alpha\beta} \end{aligned} \tag{4.11}$$

Therefore,

$$\begin{aligned} SF^1_{\alpha\beta} &= \frac{\epsilon^{data}(\alpha)}{\epsilon^{MC}(\alpha)} \\ SF^2_{\alpha\beta} &= \frac{\epsilon^{data}(\bar{\alpha})\epsilon^{data}(\beta)}{\epsilon^{MC}(\bar{\alpha})\epsilon^{MC}(\beta)} \end{aligned} \tag{4.12}$$

Note that the event-level scale factors we derive in this approach are distinct from the scale factors from approach 1. If jet 1 is tagged, then our scale factor with this method is given by line 1 in equation 4.12. In contrast, in approach 1, the scale factor includes an additional factor depending on the tagging information of jet 2. Thus we have two different well motivated approaches that give different answers.

4.2.4.3 Approach 3: Check all possibilities

There is yet another approach that seems well motivated. Perhaps we shouldn't be concerned with the actual realization of which jets were tagged, but rather we ought to account for every possible way in which a given event could have satisfied the b-tagging.

In this case, there is only one efficiency we consider for every event

$$M^1_{\alpha\beta} = \big(\epsilon(\alpha)\epsilon(\beta) + \epsilon(\alpha)\epsilon(\bar{\beta}) + \epsilon(\bar{\alpha})\epsilon(\beta)\big)\, N_{\alpha\beta} \tag{4.13}$$

which will lead to exactly one scale factor per p_T bin

$$SF^1_{\alpha\beta} = \frac{\epsilon^{data}(\alpha)\epsilon^{data}(\beta) + \epsilon^{data}(\alpha)\epsilon^{data}(\bar{\beta}) + \epsilon^{data}(\bar{\alpha})\epsilon^{data}(\beta)}{\epsilon^{MC}(\alpha)\epsilon^{MC}(\beta) + \epsilon^{MC}(\alpha)\epsilon^{MC}(\bar{\beta}) + \epsilon^{MC}(\bar{\alpha})\epsilon^{MC}(\beta)} \tag{4.14}$$

Yet again, we get a distinct scale factor for every event that is different from the previous two approaches.

4.2.5 General scale factor categorization

The problem now is that all of these approaches seem to be well motivated, and different people find different approaches to be "obviously correct." So which is actually correct?

The surprising answer is that all three are equally valid approaches (in the limit of large statistics), even though they produce different scale factors. Let's prove it.

In general, we will have J jets whose p_T bins can be indexed by $\{\alpha_1, \alpha_2, ..., \alpha_J\} \equiv \boldsymbol{\alpha}$. The number of events with jets falling into a given set of bins is $N_{\boldsymbol{\alpha}}$. These events can generically be divided into K categories, each of which has its own efficiency and scale factor, $\epsilon^k_{\boldsymbol{\alpha}}$ and $SF^k_{\boldsymbol{\alpha}}$. In each category k, we will have $M^k_{\boldsymbol{\alpha}} = (\epsilon^k_{\boldsymbol{\alpha}})^{MC} N_{\boldsymbol{\alpha}}$ MC events, and $D^k_{\boldsymbol{\alpha}} = (\epsilon^k_{\boldsymbol{\alpha}})^{data} N_{\boldsymbol{\alpha}}$ data events. We assume each event unambiguously falls into exactly one category. Equations 4.10, 4.12, and 4.14 are explicit examples of different categorizations.

Now, the total number of events in each category is not usually important. For example, in approach 2, we will almost never need to know how many events had "jet 1 failing tagging and jet 2 passing tagging." What we do want to know is the total count of the events among all categories. Therefore, we are interested in the sum of the categories

$$\begin{aligned} M_{\boldsymbol{\alpha}} &= \sum_{k=1}^{K} M_{\boldsymbol{\alpha}}^k = \sum_{k=1}^{K} (\epsilon_{\boldsymbol{\alpha}}^k)^{MC} N_{\boldsymbol{\alpha}} \\ D_{\boldsymbol{\alpha}} &= \sum_{k=1}^{K} D_{\boldsymbol{\alpha}}^k = \sum_{k=1}^{K} (\epsilon_{\boldsymbol{\alpha}}^k)^{data} N_{\boldsymbol{\alpha}} \end{aligned} \tag{4.15}$$

Our scale factors are computed per category, since they are computed at event level:

$$SF_{\boldsymbol{\alpha}}^k = \frac{(\epsilon_{\boldsymbol{\alpha}}^k)^{data}}{(\epsilon_{\boldsymbol{\alpha}}^k)^{MC}} \tag{4.16}$$

Using this set of scale factors, we can estimate the total number of data events by

$$\begin{aligned} D_{\boldsymbol{\alpha}}^{est} &= \sum_{k=1}^{K} SF_{\boldsymbol{\alpha}}^k M_{\boldsymbol{\alpha}}^k \\ &= \sum_{k=1}^{K} \left(\frac{(\epsilon_{\boldsymbol{\alpha}}^k)^{data}}{(\epsilon_{\boldsymbol{\alpha}}^k)^{MC}} \right) \left((\epsilon_{\boldsymbol{\alpha}}^k)^{MC} N_{\boldsymbol{\alpha}} \right) \\ &= \sum_{k=1}^{K} (\epsilon_{\boldsymbol{\alpha}}^k)^{data} N_{\boldsymbol{\alpha}} \\ &= D_{\boldsymbol{\alpha}} \end{aligned} \tag{4.17}$$

where we've implicitly used a large-N limit. Thus, we see that applying scale factors in each category will give the correct estimate for the total number of data events passing the tagging requirements.

This argument makes no assumptions about how the categories are chosen! This proves that all of the approaches above will give consistent results, when all of the categories are summed together.

Intuitively, this works because different scale factors from different categorizations occur at different rates in the ensemble of MC events. If we switch categorizations and end up with a scale factor for some events that is much larger than what we started with, then it will always be true that this scale factor occurs less frequently in the whole ensemble of MC events because a large SF corresponds to a small MC efficiency. The fact that the event-level scale factors are different is just an artifact of applying scale factors at the event level, as opposed to the event-collection level.

It is worth pointing out that in the proof, I did not assume any particular composition of the ensemble of events. There may be light jets intermixed with the b-jets, and the same argument holds. Categories in which a light jet is incorrectly tagged will merely have low efficiencies, but the estimate of data events will still be equally valid, so long as the composition of the Monte Carlo and the data are the same to begin with.

The result of this discussion is that there is a large amount of freedom in computing scale factors. In the large-N limit, when only the final counts are of concern, and when the efficiencies are all perfectly measured, then there is no *a priori* advantage of one method over another. This may not be the case if those assumptions are relaxed, as we will observe.

4.2.6 Finite Statistics: Multinomial Fluctuations

In all of the analysis thus far, we have been assuming large statistics. In reality, there will be fluctuations among the different $M^k_{\boldsymbol{\alpha}}$, which follow a multinomial distribution. These fluctuations will change the sum in the first line of equation 4.17, which will induce a variance in $D^{est}_{\boldsymbol{\alpha}}$.

From reference [41], the multinomial variance and covariance in the $M^k_{\boldsymbol{\alpha}}$ categories is given by

$$\begin{aligned}\mathrm{Var}(M^k_{\boldsymbol{\alpha}}) &= N_{\boldsymbol{\alpha}}(\epsilon^k_{\boldsymbol{\alpha}})^{MC}\left(1-(\epsilon^k_{\boldsymbol{\alpha}})^{MC}\right)\\ \mathrm{Cov}(M^k_{\boldsymbol{\alpha}}, M^j_{\boldsymbol{\alpha}}) &= -N_{\boldsymbol{\alpha}}(\epsilon^k_{\boldsymbol{\alpha}})^{MC}(\epsilon^j_{\boldsymbol{\alpha}})^{MC}\end{aligned} \tag{4.18}$$

The negative covariance just indicates that for a fixed $N_{\boldsymbol{\alpha}}$, when events are added to one category, they must have been removed from another category.

The variance of a sum of correlated random variables is

$$S = \sum_i a_i X_i \implies \mathrm{Var}(S) = \sum_i a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j) \tag{4.19}$$

Therefore, the variance in the estimated number of data events is

$$\begin{aligned}
\mathrm{Var}(D^{est}_{\boldsymbol{\alpha}}) &= \sum_{k=1}^{K} (SF^k_{\boldsymbol{\alpha}})^2 \mathrm{Var}(M^k_{\boldsymbol{\alpha}}) + 2 \sum_{i>j} SF^i_{\boldsymbol{\alpha}} SF^j_{\boldsymbol{\alpha}} \mathrm{Cov}(M^k_{\boldsymbol{\alpha}}, M^j_{\boldsymbol{\alpha}}) \\
&= N_{\boldsymbol{\alpha}} \left[\sum_{k=1}^{K} (SF^k_{\boldsymbol{\alpha}})^2 (\epsilon^k_{\boldsymbol{\alpha}})^{MC} \left(1 - (\epsilon^k_{\boldsymbol{\alpha}})^{MC}\right) - 2 \sum_{i>j} SF^i_{\boldsymbol{\alpha}} SF^j_{\boldsymbol{\alpha}} (\epsilon^i_{\boldsymbol{\alpha}})^{MC} (\epsilon^j_{\boldsymbol{\alpha}})^{MC} \right] \\
&= N_{\boldsymbol{\alpha}} \left[\sum_{k=1}^{K} (SF^k_{\boldsymbol{\alpha}})^2 (\epsilon^k_{\boldsymbol{\alpha}})^{MC} - \left(\sum_{k=1}^{K} SF^k_{\boldsymbol{\alpha}} (\epsilon^k_{\boldsymbol{\alpha}})^{MC} \right)^2 \right] \\
&= N_{\boldsymbol{\alpha}} \left[\sum_{k=1}^{K} SF^k_{\boldsymbol{\alpha}} (\epsilon^k_{\boldsymbol{\alpha}})^{data} - \left(\sum_{k=1}^{K} (\epsilon^k_{\boldsymbol{\alpha}})^{data} \right)^2 \right] \\
&= N_{\boldsymbol{\alpha}} \left[\sum_{k=1}^{K} (SF^k_{\boldsymbol{\alpha}} - 1) (\epsilon^k_{\boldsymbol{\alpha}})^{data} + \sum_{k=1}^{K} (\epsilon^k_{\boldsymbol{\alpha}})^{data} - \left(\sum_{k=1}^{K} (\epsilon^k_{\boldsymbol{\alpha}})^{data} \right)^2 \right]
\end{aligned} \tag{4.20}$$

and therefore, finally,

$$\mathrm{Var}(D^{est}_{\boldsymbol{\alpha}}) = N_{\boldsymbol{\alpha}} \epsilon^{tot,data}_{\boldsymbol{\alpha}} \left(1 - \epsilon^{tot,data}_{\boldsymbol{\alpha}}\right) + N_{\boldsymbol{\alpha}} \sum_{k=1}^{K} f^k_{\boldsymbol{\alpha}} \epsilon^{k,data}_{\boldsymbol{\alpha}} \tag{4.21}$$

where $\epsilon^{tot}_{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \epsilon^k_{\boldsymbol{\alpha}}$ is the total efficiency of passing the b-tagging requirement in any category, and $f^k_{\boldsymbol{\alpha}} \equiv SF^k_{\boldsymbol{\alpha}} - 1$.

The variance in $D^{est}_{\boldsymbol{\alpha}}$ is therefore made of two primary components. The first term in equation 4.21 is just the binomial variance for acceptance of data events given some total efficiency in data. This does not depend on scale factors but only on the total efficiency.

The second term is the additional variance added by the scale factor procedure. There are a few interesting things to note about this term. First, if the scale factors in several categories are less than one, then the total variance in $D^{est}_{\boldsymbol{\alpha}}$ can be less than the binomial uncertainty. Qualitatively, this happens when the Monte Carlo efficiency is larger than the data efficiency, resulting in more MC events passing than data events. Since we assume the scale factors are known perfectly, the larger number of events means that we have a higher-N estimate of $D^{est}_{\boldsymbol{\alpha}}$ than would be expected from the data efficiency alone. Figure 4.7 demonstrates the behavior of this term in a relevant toy model.

Second, it is not too hard to show via induction that if the total number of categories K is increased while the total efficiency is held constant, then the second term is non-decreasing. In fact, this term is always increasing unless categories are added with the same scale factors as already-existing categories, which is a trivial case. Thus the variance in $D^{est}_{\boldsymbol{\alpha}}$ can always be minimized by using the fewest possible categories. This makes intuitive sense, since more categories simply provide more opportunities for multinomial fluctuations.

4.2.7 Jet p_T ordering

Until this point, we have been concerned with one particular set of jet parameters $\boldsymbol{\alpha}$, which we have taken to be the set of jet p_T's. From equation 4.21, the fractional error in the estimate of $D^{est}_{\boldsymbol{\alpha}}$ scales like $\sqrt{\mathrm{Var}(D^{est}_{\boldsymbol{\alpha}})}/N_{\boldsymbol{\alpha}} \propto 1/\sqrt{N_{\boldsymbol{\alpha}}}$. Thus we want to maximize $N_{\boldsymbol{\alpha}}$, the number of events in each kinematic bin.

In many cases, the scale factor is independent from the order of the jets. That is, permutations of the α_i in $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_J\}$ all yield the same scale factor. Approaches 1 and 3 in sections 4.2.4.1 and 4.2.4.3 satisfy this property.

However, this is not guaranteed, and in fact, approach 2 in section 4.2.4.2 violates this permutation invariance. Naively, this means that approach 2 introduces additional variation due to fluctuations in the permutation of the jets, which reduces $N_{\boldsymbol{\alpha}}$ in each of the bins

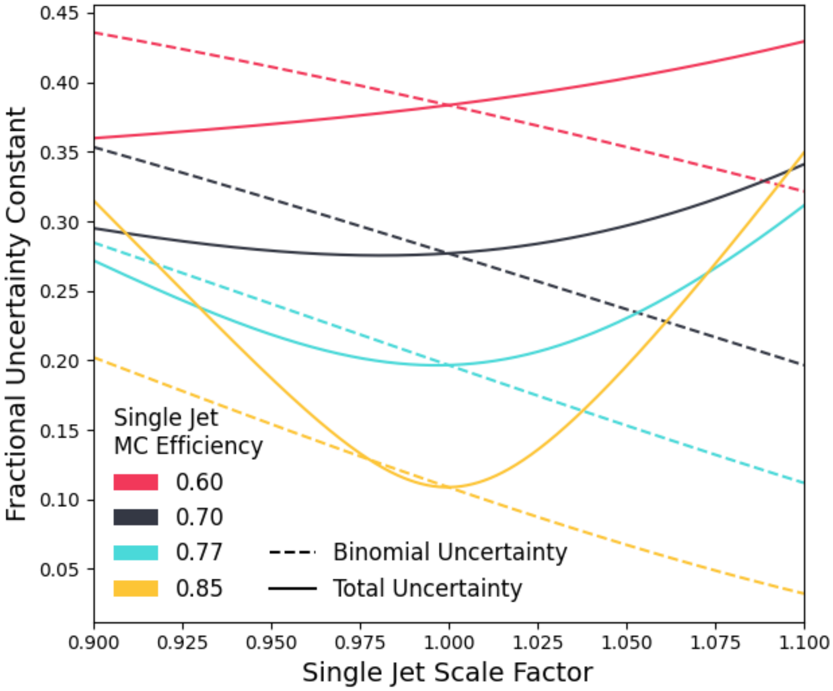


Figure 4.7: The size of the uncertainty with and without the additional contributions to the variance from the scale factor procedure. This is computed in an example of four true b-jets with at least two b-tags by using approach 2, the "check until satisfied" method discussed in section 4.2.4.2. The y-axis is the constant c in $\sqrt{\text{Var}(D^{est}_{\boldsymbol{\alpha}})}/N_{\boldsymbol{\alpha}} = c/\sqrt{N_{\boldsymbol{\alpha}}}$. Note that the uncertainty from categorical multinomial fluctuations can be a significant fraction of the total uncertainty. Notice that for scale factors slightly less than one, the total uncertainty is less than the binomial uncertainty. In addition, if the single-jet scale factor is much less than one, then this trend reverses, which is particularly evident for the 85% single-jet MC efficiency. In this case, categories with several failing jets will be more frequent in data, so the scale factor for that category will be greater than one, driving the total uncertainty back up.

and increases the variance in $D^{est}_{\boldsymbol{\alpha}}$. This effect can in fact be quite dramatic. For J jets in P possible p_T bins, there are J^P order-dependent permutations, and $\frac{(J+P-1)!}{J!(P-1)!}$ possible order-independent combinations. For the example scale factors shown in Figure 4.6, there are 9 bins in p_T. If we want to tag 4 b-jets (as is the case in the hh→4b analysis), then there would be $4^9 = 262144$ different permutations but only 495 combinations. Thus we would like to modify approach 2 to avoid violating permutation invariance.

The easiest way to achieve this is to simply sort the jets by p_T before checking any b-tagging information. Thus, all possible permutations of a given set of jets will be sorted into 1 p_T-ordered permutation. A very similar proof to that in section 4.2.5 shows that this p_T-sorting will not cause problems, so long as the counts in each permutation are summed in the final analysis. For example, p_T-sorting is valid if events with jet p_T's 50 GeV and 40 GeV are not later counted separately from events with jet p_T's 40 GeV and 50 GeV. This is true in almost all cases.

4.2.8 Categorical analysis

While a larger number of categories can lead to larger overall variance, one advantage of using more categories is that there is the flexibility to later treat each category separately. Because we apply a scale factor for each category, we will correctly estimate the number of data events in each of those categories.

In the example of 2 true b-jets with at least 1 jet b-tagged, we might later be interested in treating the 1 b-tag and 2-btag events separately. In this case, if we define our scale factor algorithm to pick categories such that 1-tag and 2-tag events are always in distinct categories, then we are free to examine each category independently. Approach 1 (section 4.2.4.1) happens to do this. In that example, all 2-tag events are in category 1. For the 1-tag events we can sum categories 2 and 3, and since each category is independently scaled correctly, the sum will be as well.

Approaches 2 and 3 (sections 4.2.4.2 and 4.2.4.3), however, do not satisfy this requirement. For example, if we were to derive scale factors with the "tag until satisfied" method, then look at only 2-tag events, we will have an incorrect normalization because the category where the first jet is tagged will include both 1- and 2-tag events.

Therefore, with careful categorization, it is possible to simultaneously calibrate multiple different interesting subsets of the data. Conversely, one needs to be careful to avoid using a subset of the data that is not calibrated.

4.3 Trigger Calibration and Trigger Scale Factors

One challenging aspect of this analysis is the calibration of the multiple b-jet triggers, the complete list of which is included in section 5.3. To correctly estimate the signal sensitivity, we need the trigger efficiency of the signal simulation to match the trigger efficiency for real signal in data. Note that the trigger calibration does not apply to data in this analysis, since we do not use simulated backgrounds in the final analysis.

There are several challenging aspects of the calibration. The first is deciding how to correct events that pass multiple triggers, where the different triggers may have different simulation-to-data corrections. The next challenge is to adjust for the jet p_T threshold turn-on curves, which are not identical between data and simulation. The last is to correct for the b-tagging efficiency that is different between simulation and data. In particular, the online and offline b-tagging efficiencies are correlated but not equal, creating a correlation in the corrections between the online and offline b-tagging.

4.3.1 Trigger Buckets

One complication with using multiple triggers is that a single event may pass any subset of them. The efficiency of each trigger can be calibrated and scale factors derived between data and MC. However, it is not clear which scale factor should be used if multiple triggers are passed. Additionally, if we check one trigger and the event fails, but it passes the next-checked trigger, the proper scale factor would include a factor for the differences between data and MC in the probability of failing the first trigger, a so-called “anti-scale factor.” Additionally, a rigorous treatment would also then need to use the efficiency of passing the second trigger *given* that the first trigger failed, and this is not measured or calibrated.

To avoid this complication, we use a strategy that ensures that each event is compared to exactly one trigger. Using offline-only information, each event is assigned to a trigger that it is very likely to pass, and this is done in a way to ensure that no trigger is starved of statistics. For example, suppose we have an event with a 350 GeV b-tagged jet offline. This event would very likely pass the high-p_T single b-jet trigger. It may also have a high chance of passing other triggers, such as the 4 jet trigger with 2 b-tags, but since the 1 b-jet trigger tends to have fewer total signal events, this category takes priority. Only after the trigger is chosen is the trigger decision for an event actually checked. If the event passes, then it is assigned to that trigger bucket and receives scale factors for that trigger. If the event fails, then it is discarded. An example flow chart of this process is shown in Figure 4.8.

The offline categories must be carefully designed in order to minimize "leakage," which is the loss of an event that would have passed another trigger. This is accomplished with two strategies. First, the triggers with the lowest signal efficiency are generally given the highest priority. By keeping the most efficient trigger as the "last resort" category, events that fail the offline selections will tend to pass this trigger. This strategy also maximizes the number of events for the inefficient triggers, reducing statistical variations. Secondly, the offline thresholds are chosen somewhat aggressively to ensure a very high chance that if the trigger is checked, the event will pass. This means that the majority of event leakage will be events that pass the final trigger, but by design, since the last trigger is the most efficient, the chance of the following is minimized: failing all offline selections AND failing the lowest priority trigger AND passing one of the other triggers. Thus by these strategies, leakage is kept to a minimum. Events that pass at least one of the triggers are only rarely discarded.

In practice, the offline bucket thresholds are chosen via a brute-force grid search that optimizes the final signal sensitivity.

4.3.2 Trigger p_T Scale Factors

The next challenge with the triggers is correcting for the different efficiencies as a function of jet p_T in data and simulation. In particular, we are often concerned with low-mass signals with jets near the trigger thresholds, meaning proper calibration of the trigger turn-on curves is important.

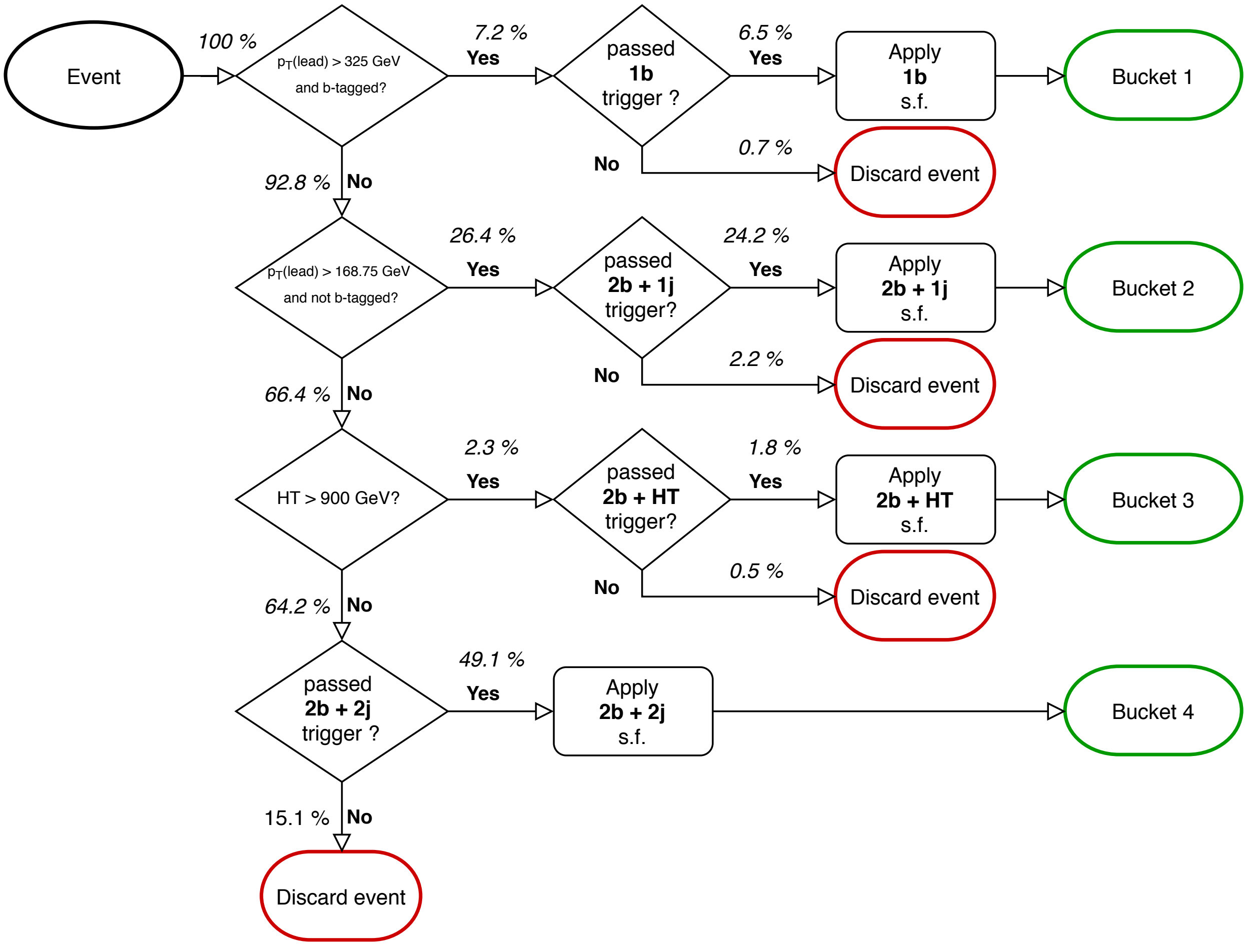


Figure 4.8: The trigger buckets strategy flowchart, with percentages shown from a G^*_{KK} graviton resonance signal sample.

The nontrivial dependence of efficiency on jet p_T arises from the differences between online and offline jets. The jets are reconstructed and calibrated differently, and the net effect is that a given offline jet p_T^{off} will map to a distribution of online jet p_T^{on}'s, generally at a slightly lower energy scale due to imperfect online calibrations. Thus, because the trigger cuts on p_T^{on}, a given offline p_T^{off} has a random probability of passing or failing the trigger, depending on whether it is randomly mapped above or below the online threshold.

Because this efficiency behavior depends only on the relationship between online and offline p_T, the probabilities for each offline jet to pass the threshold are independent. Thus for each offline jet, there is some p_T dependent probability of the jet passing a cut, $\epsilon(p_T)$, and the situation is very similar to that discussed in Section 4.2. In particular, we can match the online and offline jets by requiring they be within some ΔR of each other. In multijet events, we can just multiply the single-jet efficiencies to get the full-event trigger efficiency. This also means that the online and offline efficiencies can be compared at the single-jet level in order to derive jet-level scale factors that can be multiplied together to get event-level scale factors. Note that this story can be somewhat complicated by extra combinatorics if jet order can change between online and offline, but in practice this is rare enough to be negligible in most cases.

This logic applies identically to the case of the ATLAS two-stage trigger where we would then be concerned with the probability of a jet of a given offline p_T passing both the HLT and L1 thresholds.

In this analysis, the trigger p_T scale factors are calibrated using muon-triggered $t\bar{t}$ events, $t\bar{t} \to \mu\nu bbqq$, in data and simulation. The $t\bar{t}$ event selection requires exactly one muon that fired a muon trigger and with $p_T > 25$ GeV, at least four jets with $p_T > 35$ GeV, at least two offline b-tagged jets, missing transverse momentum, $E_T^{\text{miss}} > 20$ GeV, and $E_T^{\text{miss}} + m_T^W > 60$ GeV, where $m_T^W = \sqrt{2p_T^{\ell} E_T^{\text{miss}}(1 - \cos\Delta\phi_{\ell\nu})}$. This selection achieves a very high semi-leptonic $t\bar{t}$ purity without making kinematic cuts on the jets that could bias the efficiencies.

The single-jet efficiency measured in data or MC is then just measured by the number

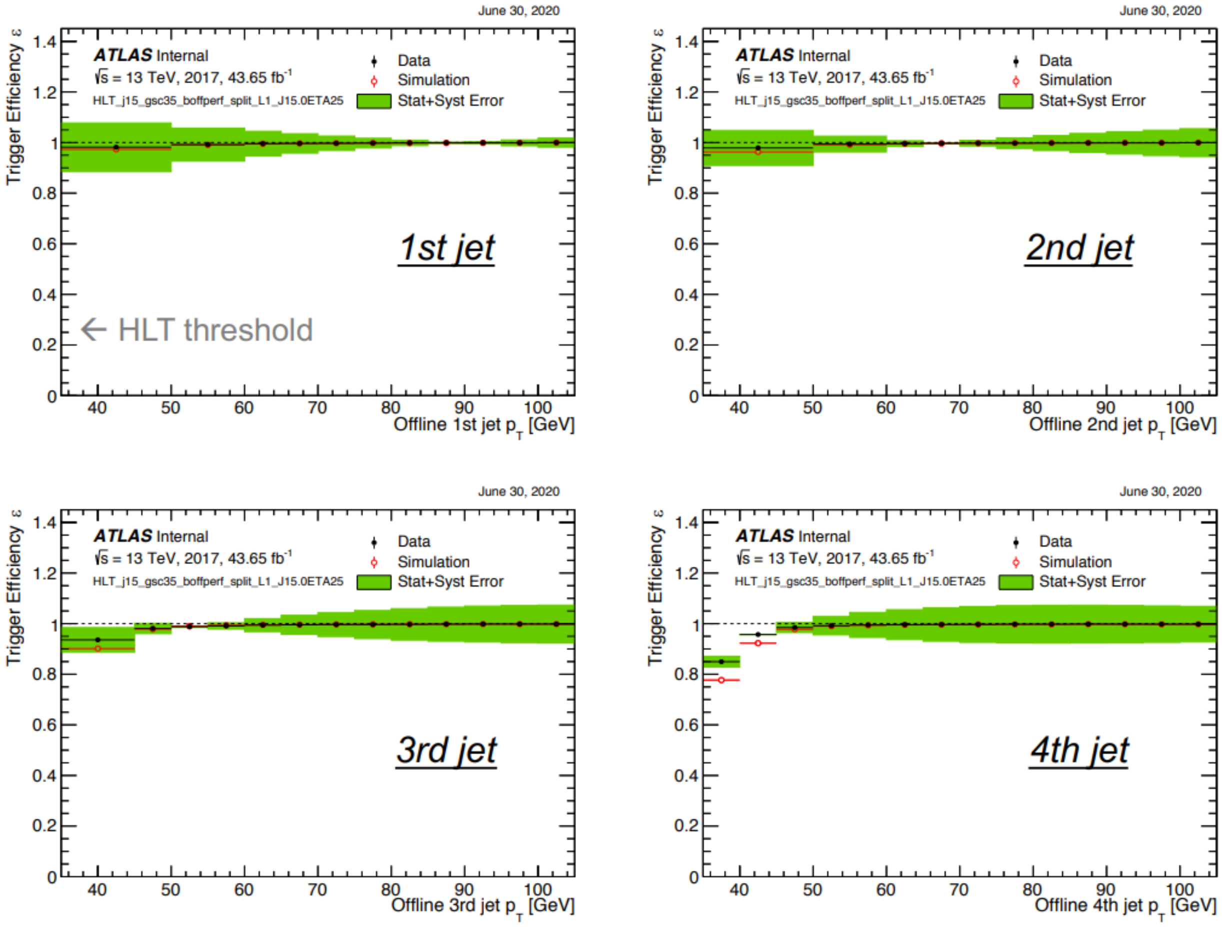


Figure 4.9: The single-jet HLT trigger efficiencies for jets to pass a 35 GeV trigger threshold. The trigger efficiencies shown concern a trigger that requires four jets above the 35 GeV. Notice the difference between data and simulation, particularly on the fourth-leading jet.

of events with a jet in a given p_T range passing the HLT threshold and the event selection, normalized by all events passing the event selection with a jet in that p_T range. Example derived single-jet efficiencies are shown in Figure 4.9.

4.3.3 Trigger b-tagging Scale Factors

The issue of using b-tagging online and offline is somewhat subtle, as the correlations between the two rounds of b-tagging are very high. Suppose we have the simple case of one (true) b-jet where we run online b-tagging and offline b-tagging. We are interested in the efficiency for the jet passing online AND offline b-tagging, $\epsilon(\text{on} \cap \text{off})$. By the definition of conditional probability, $\epsilon(\text{on} \cap \text{off}) = \epsilon(\text{on})\epsilon(\text{off}|\text{on}) = \epsilon(\text{off})\epsilon(\text{on}|\text{off})$. Thus in the Bayesian view, one can view the probability of passing both as the probability of passing the online b-tagging times the probability of passing the offline b-tagging given the prior that the online b-tagging passed.

If there are multiple b-jets and fewer b-tags are required online than offline (as is the case in this analysis), it will often happen that we are interested in the probability that a jet failed online b-tagging but passed offline b-tagging, $\epsilon(\overline{\text{on}} \cap \text{off})$, leading to conditional efficiencies of passing offline given a failure online (and also vice-versa). This means that the event-level online and offline b-tagging scale factor computations will generally need to be done together in one step.

This story is even further complicated by the fact that the online and offline b-tagging are not only using slightly different algorithms and jet definitions, but even different working points, where the efficiencies can be quite different and also still correlated. In the real case of requiring for example "at least two b-tags online at the 60% working point and at least four b-tags offline at the 77% working point," the scale factor computation can get quite complex.

Fortunately, the results of section 4.2 and still apply, especially the general proof in section 4.2.5 that there is a great deal of flexibility in computing scale factors. When computing joint online and offline scale factors, we can still divide the events into K categories, where now the categories depend on both the online and offline efficiencies. For example, suppose we have an event with two true b-jets and we require at least one online b-tag and two offline b-tags. Suppose also we are applying the "check until satisfied" method, and we are checking the online tagging first. I will always assume we are able to match the online and offline jets, which is generally true. In this case, there are two categories whose efficiencies are

$$\begin{aligned} \epsilon^1 &= \epsilon(1^{\text{on}})\epsilon(1^{\text{off}}|1^{\text{on}})\epsilon(2^{\text{off}}) \\ \epsilon^2 &= \epsilon(\overline{1^{\text{on}}})\epsilon(2^{\text{on}})\epsilon(1^{\text{off}}|\overline{1^{\text{on}}})\epsilon(2^{\text{off}}|2^{\text{on}}) \end{aligned} \tag{4.22}$$

The first category corresponds to the case where the first online jet is tagged so that we have satisfied the “at least one online b-tag” condition. Then offline we need to use a *conditional* probability for jet 1, since we checked it, but not for jet 2, since it was never checked online. In the second category, the first online jet fails b-tagging, but the second passes. Since both jets were checked online, the conditional probabilities have to be used for both jets.

The only assumptions made in the proof of section 4.2.5 are that the different categories are mutually exclusive, that the efficiencies are properly normalized, and that the categories will be considered inclusively in any final histogram. All of these assumptions still hold. The only differences are that the efficiencies of each category are somewhat more complicated to compute, and there will be more total categories given the number of different possible online/offline tagging combinations.

In practice, we follow the ATLAS standard of “check until satisfied” for p_T-sorted jets. We check the online tagging first, until we have enough tagged jets to satisfy the trigger requirement. Then we check the offline jets, applying conditional scale factors whenever the online jet information was considered. The online efficiencies, offline efficiencies, and online given offline conditional efficiencies are all provided by the b-jet trigger group. These efficiencies are measured by selecting a sample of nearly pure b-jets from $t\bar{t}$ events that are triggered with electron or muon triggers. These lepton triggers are dedicated for online b-tagging calibration and so the online b-tagging algorithms are run but unused in the decision. Then it is a relatively straightforward matter to later measure what fraction of the true b-jets from the top quark decays were successfully tagged.

Because the conditional efficiencies sometimes require several terms to express in terms of these three quantities, such as $\epsilon(\overline{on} \cap \overline{off}) = 1 - \epsilon(off) - \epsilon(on) + \epsilon(on|off)\epsilon(off)$, the scale factor for a given event is generally not factorizable into a product of single-jet scale factors and must be computed from the raw efficiencies. For this reason, we do the online and offline b-tagging scale factor computation in functionally one step. This method is not unique, but it does tend to yield relatively few categories, minimizing the multinomial fluctuations.

CHAPTER 5

EVENT SELECTION AND RECONSTRUCTION

5.1 Data Samples

The data used in this analysis consists of 126.7 fb^{-1} of data collected at 13 TeV from 2016 to 2018 using the ATLAS detector. The luminosity per year is shown in Table 5.1. Note that a few femtobarns of data were collected in 2015, but this data is not used due to a lack of a recent b-jet trigger calibration for the unique 2015 multi-b-jet triggers. Additionally, around 9 fb^{-1} of data collected in 2016 is not useable due to an inefficiency in the b-jet trigger related to online vertex reconstruction.

5.2 Monte Carlo Samples

The only Monte Carlo samples used in the final analysis are the non-resonant $hh \to 4b$ signal samples because the background is fully data-driven.

5.2.1 Standard Model Signal

Non-resonant $hh \to 4b$ signal simulation is used to optimize the analysis and determine the expected sensitivity. There are two samples that are used in this analysis, an older sample that approximates that top mass as infinite in loops, and a new sample that uses a fully correct, more difficult finite-top mass calculation. Both samples are generated at NLO. The older signal samples use the MC@NLO generator interfaced with Herwig7 for parton

Year	Luminosity
2016	24.6 fb^{-1}
2017	43.65 fb^{-1}
2018	58.45 fb^{-1}

Table 5.1: The total integrated luminosity used in this analysis per year.

showering and hadronization. The newer signal samples use Powheg for event generation and are again interfaced with Herwig for showering.

5.2.2 κ_λ Reweighting

We would like to be able to have signal samples at many different κ_λ values, but generating enough statistics at different values would require substantial CPU time. Instead, we can leverage the functional form of the cross-section to perform a reweighting to any κ_λ value given a few basis samples.

From the two leading order di-Higgs Feynman diagrams in Figure 2.4, we can see that the triangle diagram will be proportional to $\kappa_t \kappa_\lambda$, and the box diagram will be proportional to κ_t^2, where κ_t is the Higgs top quark Yukawa coupling normalized to the Standard Model. Thus the total di-Higgs amplitude will, at leading order, have the form $\mathcal{M} = K_1 \kappa_t^2 + K_2 \kappa_t \kappa_\lambda$, where K_1 and K_2 depend only on kinematics. If we're making some signal histogram, then the number of events in a given bin will be proportional to the differential cross section, which is in turn proportional to $|\mathcal{M}|^2$. Since we set limits using the shape of m_{hh}, and because m_{hh} signifcantly impacts the signal event kinematics, this is the most important variable to estimate. Thus we are interested in $\frac{d\sigma}{dm_{hh}}$, or more precisely, for a bin width of Δm_{hh}, $\frac{\Delta\sigma}{\Delta m_{hh}}$.

Squaring the amplitude, we find

$$\begin{aligned} \frac{d\sigma}{dm_{hh}} &\propto K_1^2 \kappa_t^4 + 2K_1 K_2 \kappa_t^3 \kappa_\lambda + K_2^2 \kappa_t^2 \kappa_\lambda^2 \\ &= A + B\kappa_\lambda + C\kappa_\lambda^2, \end{aligned} \tag{5.1}$$

so to leading order, the cross section depends quadratically on κ_λ with three constants. Then we use truth samples generated for $\kappa_\lambda = 0, 1$, and 10 to solve for A, B, and C in each bin of m_{hh}. If we want to find the m_{hh} distribution for a new $\kappa_\lambda^{\text{target}}$ value that wasn't generated, we can reweight each standard model event by first computing m_{hh}, then scaling that event weight by

$$\frac{A(m_{hh}) + B(m_{hh})\kappa_\lambda^{\text{target}} + C(m_{hh})(\kappa_\lambda^{\text{target}})^2}{\left(\frac{d\sigma}{dm_{hh}}\right)_{\kappa_\lambda=1}} \tag{5.2}$$

where the denominator is in practice just the number of events in that m_{hh} bin for the standard model.

One subtlety to ensure accuracy is that first, the truth m_{hh} values must be used without selections to derive the reweighting constants. Selection and reconstruction will distort the quadratic dependence. We also want enough statistics in each m_{hh} bin such that the constants are well determined, so roughly speaking, the basis κ_λ values chosen should cover well all possible m_{hh} values.

5.3 Triggers

The first event selection that happens is the online, live selection of which events to store to tape, the trigger. Triggering on the $hh \to 4b$ signal is quite difficult because of the all-hadronic final state, since most events at the LHC produce many jets. B-tagging must be run online in order to filter light jets, but this is quite computationally expensive due to running tracking for the complex b-tagging algorithms. The multi-b-jet triggers consumed a not-insignificant fraction of the online CPU resources in Run 2.

Because of the difficulties of triggering on our signal, no single trigger is optimal. Instead, we use a logical *OR* of several different b-jet triggers that target different regions of phase space. All of the triggers used are listed in Table 5.2, with a more human-readable explanation in Table 5.3 This list was determined via a brute-force optimization using the Standard Model signal, with additional triggers beyond the four primaries giving only small additional signal efficiency at the cost of increased complexity. The efficiency of each of the triggers in 2017 is shown as a function of m_{hh} in Figure 5.1, demonstrating that different triggers are optimal for different kinematic regions.

Note that we do not use the 2015 data, a loss of 3.22 fb^{-1}, because of difficulties matching

Short Name	Full Name
2b2j	HLT_2j35_bmv2c2060_split_2j35_L14J15.0ETA25
2b1j	HLT_j100_2j55_bmv2c2060_split
1b	HLT_j225_bmv2c2060_split

(a) 2016

Short Name	Full Name
2b2j	HLT_2j15_gsc35_bmv2c1040_split_2j15_gsc35_boffperf_split_L14J15.0ETA25
2b1j	HLT_j110_gsc150_boffperf_split_2j35_gsc55_bmv2c1070_split_L1J85_3J30
2bHT	HLT_2j35_gsc55_bmv2c1050_split_ht300_L1HT190-J15s5.ETA21
1b	HLT_j225_gsc300_bmv2c1070_split

(b) 2017

Short Name	Full Name
2b2j	HLT_2j35_bmv2c1060_split_2j35_L14J15.0ETA25
2b1j	HLT_j110_gsc150_boffperf_split_2j45_gsc55_bmv2c1070_split_L1J85_3J30
2bHT	HLT_2j45_gsc55_bmv2c1050_split_ht300_L1HT190-J15s5.ETA21
1b	HLT_j225_gsc300_bmv2c1070_split

(c) 2018

Table 5.2: Triggers used in each year with the ATLAS-standard trigger names.

online to offline jets and calibrating the triggers.

5.4 Offline Selection

After the trigger selection and the selection for good-quality data runs (from the standard ATLAS Good Runs List), we apply the offline event selections. We first require events to

Trigger Name	Description
2b2j	Four jets $p_T > 35$ GeV; At least 2 b-tags
2b1j	One jet $p_T > 150$ GeV; Two other jets $p_T > 55$ GeV; At least any two b-tagged
2bHT	Two jets $p_T > 55$ GeV and b-tagged; Scalar sum of top 5 jet p_T's > 300 GeV
1b	One jet $p_T > 300$ GeV, b-tagged

Table 5.3: Explanation of the primary triggers. The exact definitions vary slightly from year-to-year depending on the best available jet calibrations, and the b-tagging working points are as efficient as possible given that year's rate constraints. Note that the 2bHT trigger didn't yet exist in 2016.

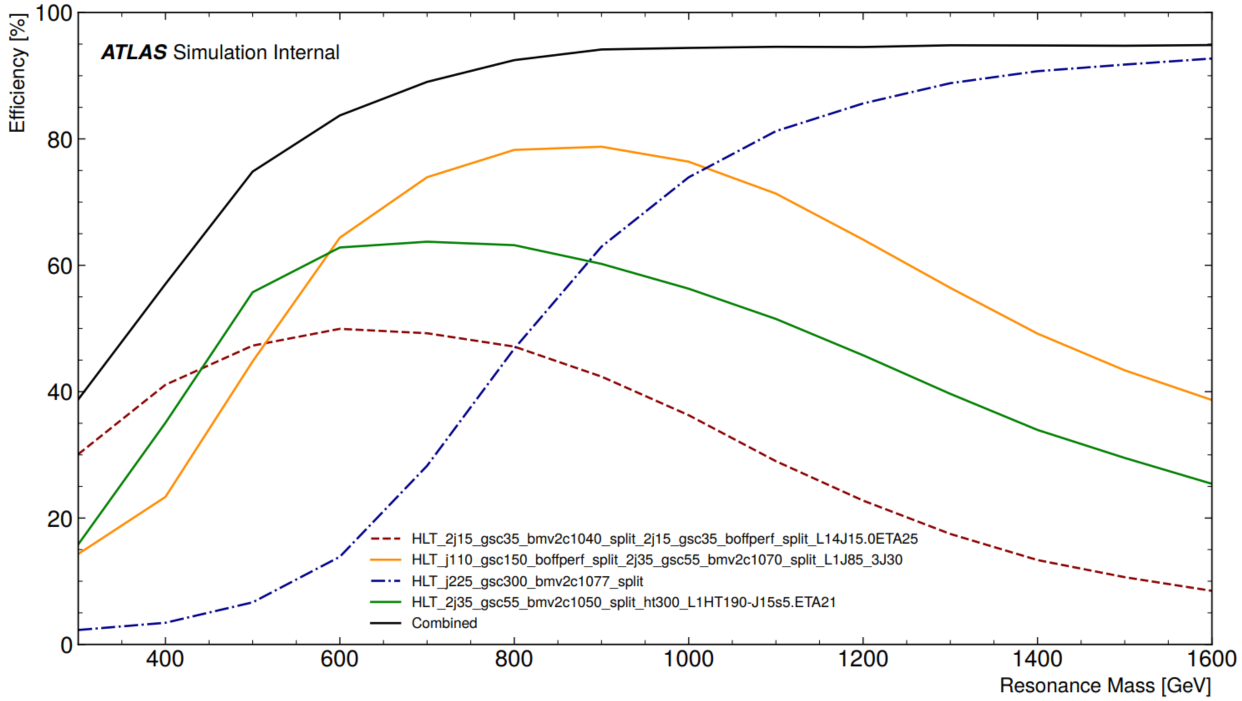


Figure 5.1: Simulated trigger efficiency for 2017 triggers for massive G^*_{KK} graviton resonances decaying to two Higgs's to four b-jets, prior to any analysis selections. The Standard Model nonresonant signal peaks around 400 GeV, though with a broad distribution.

have at least four jets with $p_T > 40$ GeV and $|\eta| < 2.5$. The p_T threshold is selected to align with the p_T thresholds of the lowest-p_T trigger, and the η cut ensures that the jets are within the tracker so that b-tagging can be run on all four jets. A study is currently underway to add a category that relaxes the p_T cut by leveraging the fact that three of our four triggers require fewer than four jets, albeit with higher thresholds on the remaining jets.

We then select for at least two b-tags among the jets with $p_T > 40$ GeV. The events with exactly two b-tags are used to derive the background estiamte. The events with four or more b-tags form the primary signal region. In events with exactly three b-tags, the fourth jet is recovered and used to form an alternative signal region, as discussed in section 5.6.

Next, we apply a cut to reject the $t\bar{t}$ components of the background. Since virtually every top quark decays to Wb, $t\bar{t}$ events always contain 2 true b-jets. We reject the leptonic component of the $t\bar{t}$ from leptonic W decays by ensuring that we have no isolated electrons or muons. The other 2 b-tags in 4b $t\bar{t}$ come from two different sources with comparable frequency. First, the W's may decay to cs or to light jets, with the charm jets or light jets being mis-tagged as b-jets. Secondly, $t\bar{t}$ may be produced in associated with a $b\bar{b}$ pair not from a top decay. Note that additional true b-jets appearing directly in the top decays is negligible because it is suppressed by the off-diagonal elements of the CKM matrix, with $\text{BR}(t \to bW \to bbq) \approx 0.06\%$.

To reject $t\bar{t}$, we define the variable X_{Wt},

$$X_{Wt} = \sqrt{\left(\frac{m_W - 80.4 \text{ GeV}}{0.1 m_W}\right)^2 + \left(\frac{m_t - 172.5 \text{ GeV}}{0.1 m_t}\right)^2} \tag{5.3}$$

where m_W and m_t are formed from every possible triplet of jets, and the combination with the minimum X_{Wt} is chosen. The 0.1 in the denominator is a rough approximation for the mass resolution. Then we cut and keep events with $X_{Wt} \geq 1.5$, for which there is no combination of jets that lead to invariant masses consistent with a top quark decay to a W boson. This cut ensures that $t\bar{t}$ contributes less than 10% of the background.

Next, every event is assigned to a trigger bucket as described in section 4.3.1. Any event that fails the trigger in that bucket is rejected. This leads to only a small loss of signal and background while significantly simplifying the trigger calibration. The trigger bucket cuts must also be run on data to ensure no kinematic biasing, even though it is only used to calibrate the signal simulation.

At this point, we pair the b-jets into "Higgs candidates," and compute the kinematics of the hypothesized parent Higgs. Of course, the pairs will only be a Higgs boson for signal, hence the name Higgs candidate. The pairing problem is discussed in detail in section 5.5. For events with only 3 b-tags, there is an additional step of choosing the fourth jet that is described in section 5.6.

In order to minimize the QCD backgrounds, we select events in which the two Higgs candidates are nearby in pseudorapidity, $\Delta\eta_{hh} < 1.5$. The di-Higgs signal is very central, so the Higgs candidates tend to have a small pseudorapidity gap. On the other hand, QCD backgrounds tend to increase closer to the beamline, so the events with a larger pseudorapidity gap tend to be predominately background.

Lastly, we define a signal region using the invariant masses of the Higgs candidates $m_{bb}^2 = -(p_1^\mu + p_2^\mu)^2 = (E_1 + E_2)^2 - |\vec{p_1} + \vec{p_2}|^2$. The Higgs candidates are sorted by p_T, with the higher-momentum Higgs candidate called the leading Higgs and the other the subleading Higgs. The masses of the leading and subleading Higgs candidates are plotted in a 2D histogram that we call the mass plane. With perfect jet energy scale and resolution, the signal would peak at (125 GeV, 125 GeV). However, the jet calibration and resolution is such that the peak is somewhat to the lower left, closer to (120 GeV, 110 GeV). The asymmetry is due to the fact that the energy scale and resolution is different for the higher-momentum jets from the leading Higgs candidate. The signal region (SR) is then defined as a region around the signal peak. The SR is defined by $X_{hh} < 1.6$, where

$$X_{hh} = \sqrt{\left(\frac{m_{h1} - 120\text{ GeV}}{0.1 m_{h1}}\right)^2 + \left(\frac{m_{h2} - 110\text{ GeV}}{0.1 m_{h2}}\right)^2} < 1.6 \tag{5.4}$$

This shape is not quite an ellipse because the independent variables (m_{h1}, m_{h2}) appear in the denominator. However, it approaches an ellipse in the limit as the 0.1 resolution factor in the denominator goes to 0, so it is approximately elliptical. The final limits will be set using only the events that fall in the SR.

Around the signal region, we define the Validation Region (VR), which is circular, and around the VR, we define the Control Region (CR). Specifically, the VR is defined by events outside the SR with

$$\sqrt{(m_{h1} - (120 \times 1.03)\text{ GeV})^2 + (m_{h2} - (110 \times 1.03)\text{ GeV})^2} < 30\text{ GeV} \tag{5.5}$$

and the CR is defined by events outside the VR and SR with

$$\sqrt{(m_{h1} - (120 \times 1.05)\text{ GeV})^2 + (m_{h2} - (110 \times 1.05)\text{ GeV})^2} < 45\text{ GeV} \tag{5.6}$$

The factors of 1.03 and 1.05 shift the center of the circles to better center on the shape of the signal region, which is not circular. The CR and VR are used to derive and to validate the background predictions, as described in Chapter 6.

5.5 Higgs Pairing

One challenging aspect of this analysis is that with four b-jets in the final state, there are three ways to pair them into two Higgs bosons, as diagrammed in Figure 5.2. Because the pairing algorithm will also be run on the background events, the choice of pairing algorithm is extremely important for determining the shape of the background in the massplane. For example, selecting only events where there are two pairs of jets with an invariant mass close to the Higgs mass will badly sculpt the background. Because there are many pairing combinations in every event, this would bias the background to select jet pairings that just happen to result in masses near the Higgs mass, causing the background to peak at (125, 125). This is an example of a bad pairing algorithm, even if the approach is highly efficient on signal and naïvely seems like a good approach. There are many pairing algorithms that have been considered for this analysis, several of which are discussed below.

Additionally, for the 2b data, the selection of the other jets to use for the Higgs candidate is usually part of the pairing algorithm. This selection can impact how similar 2b and 4b are, which can impact the uncertainties on the background estimate.

5.5.1 Pairing Strategies

Min-D_{hh} Pairing The last round of this analysis used a method that effectively picks a pairing where the two invariant masses are the most equal. In fact, because the signal peak is slightly off center, the pairing is chosen that minimizes the distance in the massplane to the line from (0,0) to (120,110), called D_{hh}. This is shown daigramatically in Figure 5.3. While this avoids biasing the background entirely into the signal region, it instead biases the

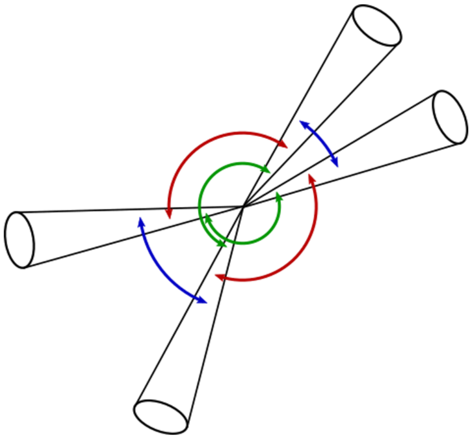


Figure 5.2: The three ways that four jets can be paired into two Higgs each decaying to a pair of jets.

background to lie along the diagonal line, as shown in Figure 5.3. By always picking pairings in the background that are closest to the line, the background *must* inevitably be sculpted, at least to some extent. In 2b data, the two untagged jets with the highest b-tag score were selected, a procedure that is not well-calibrated.

BDT Pairing Improving upon the D_{hh} pairing, the full-Run 2 resonance search used a Boosted Decision Tree (BDT) algorithm to pair the jets. By using other kinematic variables instead of the Higgs masses, the BDT somewhat relieves the background sculpting. In 2b data, the two additional jets are selected randomly among all jets with $p_T > 40$ GeV and $|\eta| < 2.5$.

For a given pairing hypothesis, the BDT is fed various kinematic information for that pairing, and outputs a "correct pairing" score. All three possible pairings are run through the BDT, and the pairing with the highest score is selected.

The inputs to the BDT for a given pairing are m_{hh} and the opening angles between the b-jet pairs, ΔR_1, ΔR_2, $\Delta\eta_1$, $\Delta\eta_2$, $\Delta\phi_1$, and $\Delta\phi_2$. The variable $m_{hh} = m_{4b} = (\Sigma E_i)^2 - |\Sigma\vec{p}_i|^2$ does not depend on the pairing and just serves to parameterize the BDT, since the angular dependence of the decay products depends strongly on the total mass of the system (i.e., at high m_{hh}, the Higgs bosons have a large kinetic energy and the decay products are Lorentz-

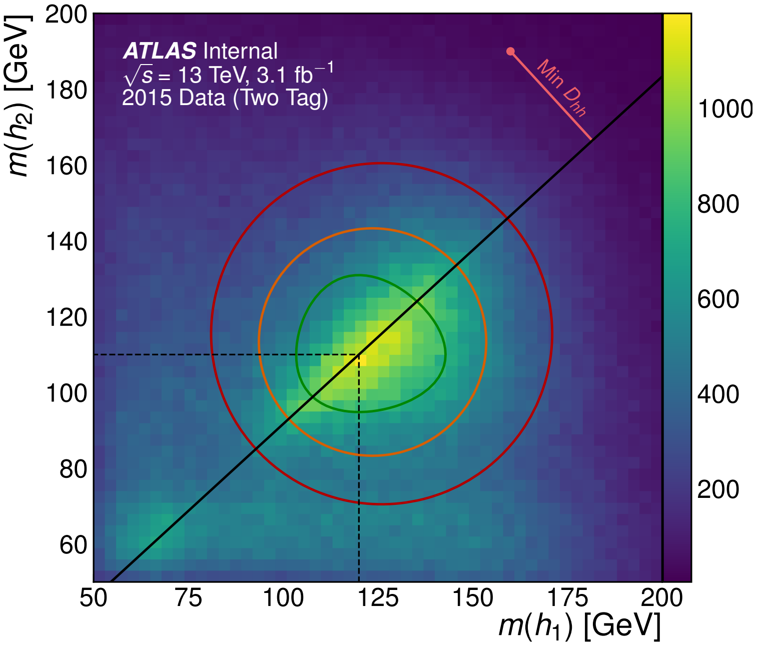


Figure 5.3: The Higgs massplane for min-D_{hh} pairing in 2015 2b data, a similar distribution to the 4b data. The signal region is drawn in green, the validation region in orange, and the control region in red.

boosted to smaller opening angles).

The efficiency of the BDT pairing as a function of m_{hh} is shown in Figure 5.4, in addition the the min-D_{hh} (discussed above) and min-ΔR (discussed below) pairing algorithms. The 2b and (blinded) 4b massplanes for the BDT pairing are shown in Figure 5.5. Notice the peak in the background.

Pair-A-Graph Pairing The Pair-A-Graph algorithm uses a neural network to predict which jet pairing is the most likely. The key feature of this algorithm is the use of self-attention, which allows relationships between jets to be learned in an permutation-invariant manner, as described in section 3.3.6.

The architecture of the Pair-A-Graph network is diagrammed in Figure 5.6. The inputs to the algorithm are p_T, η, ϕ, E, and the b-tagging quantile for the five leading jets. The first step of the algorithm is to embed these five-dimensional jet vectors into a 10-dimensional latent space via a learned linear embedding. The jet vectors are then run through one or two

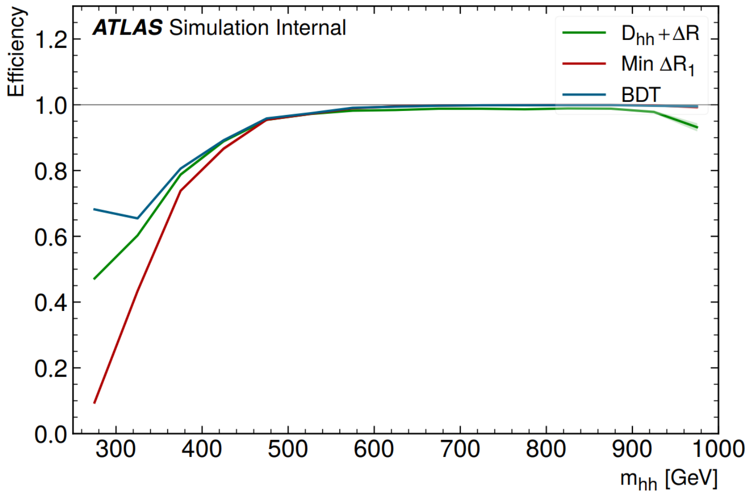
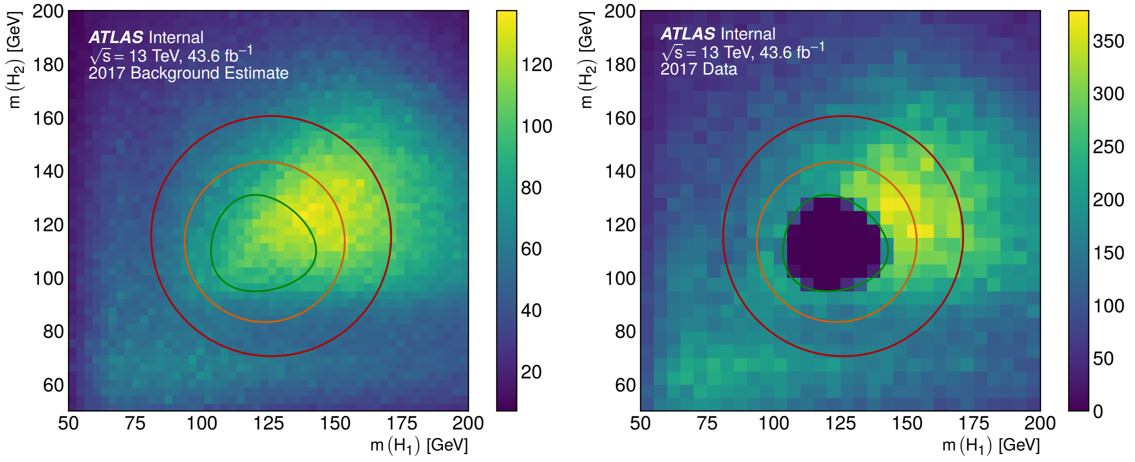


Figure 5.4: The efficiency vs m_{hh} for various pairing algorithms.



(a)

(b)

Figure 5.5: The backgrounds in the massplane for the BDT pairing. (a) The 2b-derived background estimate, described in section 6.1. (b) The blinded 4b data.

transformer encoder layers, and then a final multi-head self-attention layer. The network structure was tested without the transformer encoder layer, but the performance was found to be slightly worse. The output of the final mutli-head self-attention layer is the raw dot products of they query and key vectors, which will be one scalar for every possible pair of jets, called the edge score. Then, with five jets, there are 15 ways to create two pairs such that no jet is shared between the pairs (no one jet can be in two Higgs candidates). For each of these 15 valid pairings, the sum of pair scores is computed, and the selected pairing is that with the highest score. The loss function is the standard negative log likelihood (or equivalently, cross-entropy) for classification problems.

The algorithm can be easily visualized as a graph, where the nodes are the jets and the edges are the pairings between the jets. The network takes the input kinematics from all of the jets, and outputs a pairing score for every possible edge. The two non-adjacent edges with the highest sum of edge weights are taken as the pairing. This is shown diagrammatically in final stages of Figure 5.6.

Note that PairAGraph naturally handles differing numbers of b-tags, so does not need a special algorithm for selecting the additional jets in the 2b category. PairAGraph will tend to select the 2b jets that are naturally similar to 4b. Similarly, Pair-A-Graph can run out-of-the-box on 3b data, where it will automatically select the fourth jet.

The massplane for the Pair-A-Graph pairing algorithm is shown in Figure 5.7. One can immediately observe that there is no longer any peaking in the signal region and the background smoothly falls off from the lower left corner.

Min-ΔR Pairing The lack of a sculpted background peak is not a unique feature of the Pair-A-Graph algorithm, and in fact, a much simpler pairing algorithm also achieves good efficiency without sculpting the background. One simply pairs into one Higgs candidate the pair with the minimum ΔR between the jets. If there are exactly four b-jets, the other Higgs candidate is just the other pair of b-jets, and if there are more than four b-jets, the pair with

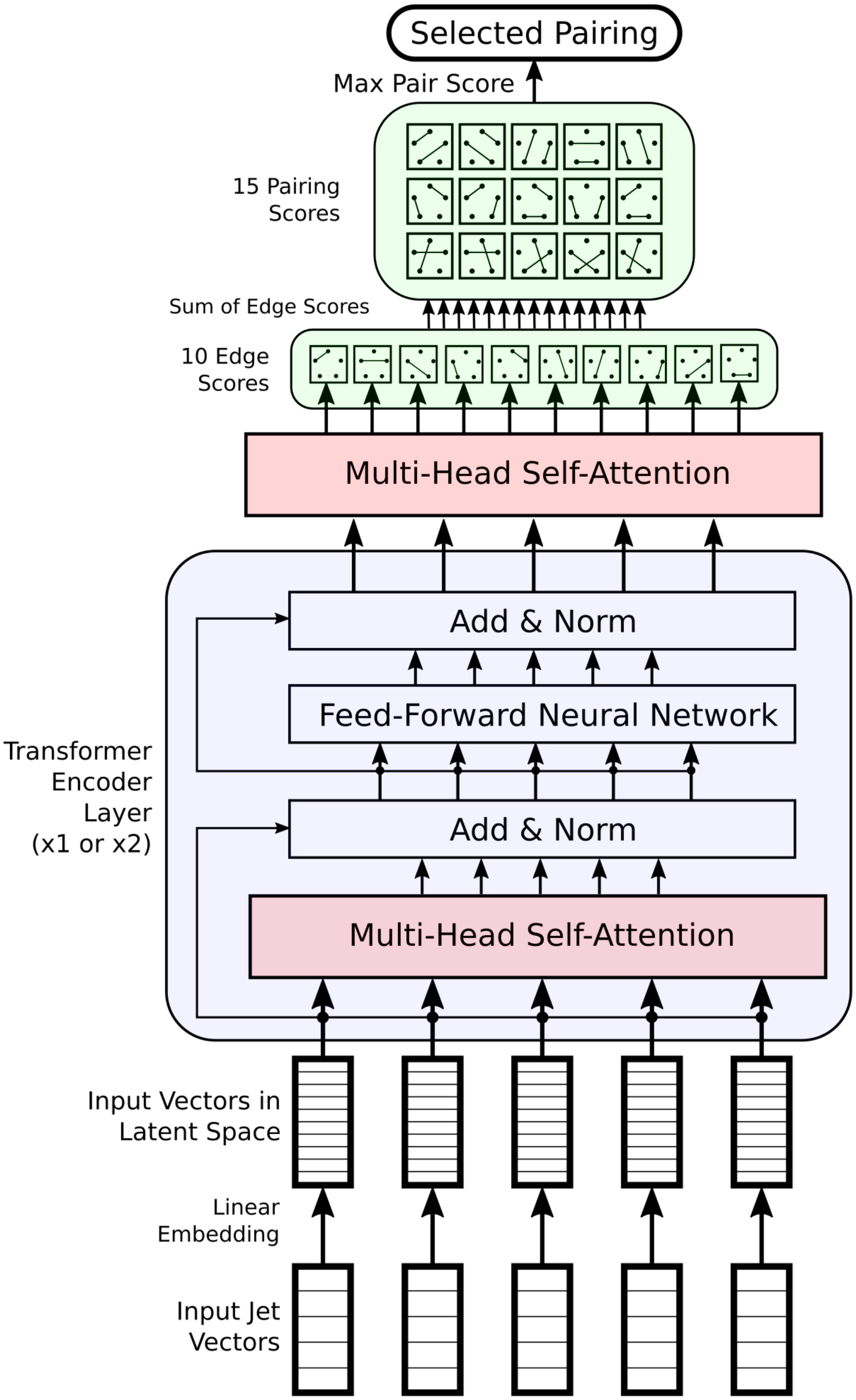

Selected Pairing
Max Pair Score
15 Pairing Scores
Sum of Edge Scores
10 Edge Scores
Multi-Head Self-Attention
Add & Norm
Feed-Forward Neural Network
Add & Norm
Multi-Head Self-Attention
Transformer Encoder Layer (x1 or x2)
Input Vectors in Latent Space
Linear Embedding
Input Jet Vectors

Figure 5.6: The Pair-A-Graph architecture.

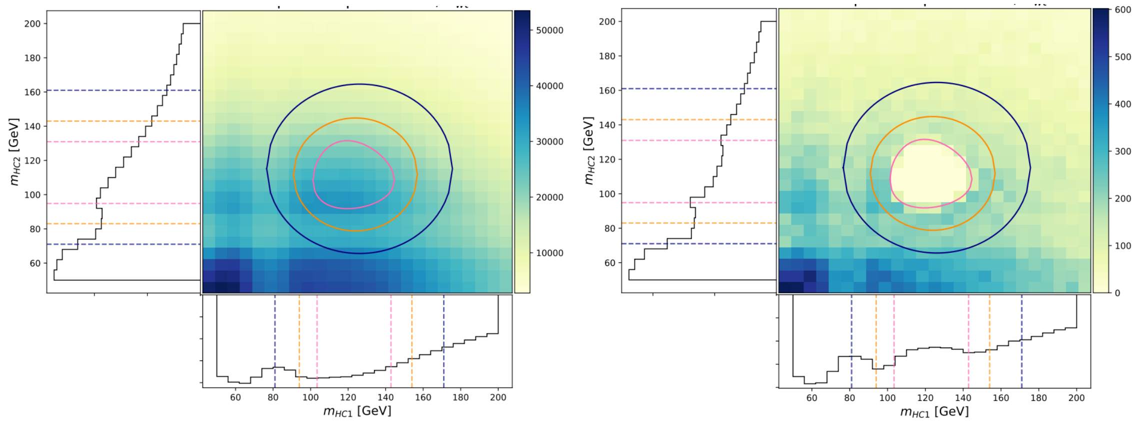


Figure 5.7: The massplane after the Pair-A-Graph pairing, with Pair-A-Graph trained on the Standard Model signal. The sharp band near 80 GeV is due to the X_{Wt} cut that removes jet pairs consistent with the W boson mass, 80 GeV. The band is present though less visible in the other massplanes due the large background peak.

the minimum ΔR can again be selected. The resultant massplane from the min-ΔR pairing strategy is shown in Figure 5.8. Because of the simplicity and effectiveness of this strategy, this is the baseline reconstruction algorithm used in this analysis.

5.6 The 3b Category

5.6.1 Motivation

This round of the analysis includes for the first time a category of events with only three b-tagged jets. The motivation is simple – at a roughly 77% average b-tagging efficiency for each b-jet, the probability that all four of our signal b-jets are tagged is only around $(0.77)^4 = 35\%$. However, the probability of having exactly three b-tags is $4(0.77)^3(1-0.77) = 0.42\%$. Thus naively, including a 3b category should more than double our signal acceptance and make a significant improvement to the final limit.

However, a challenge of the 3b category is that it suffers from higher backgrounds. Not only will the real multi-b backgrounds be higher for the same reason that there is more signal, but also there is a much higher probability of 2 real b-jets plus one mis-tagged jet. For light jets, where the mis-tag probability is on the order of 0.1%, this means we should

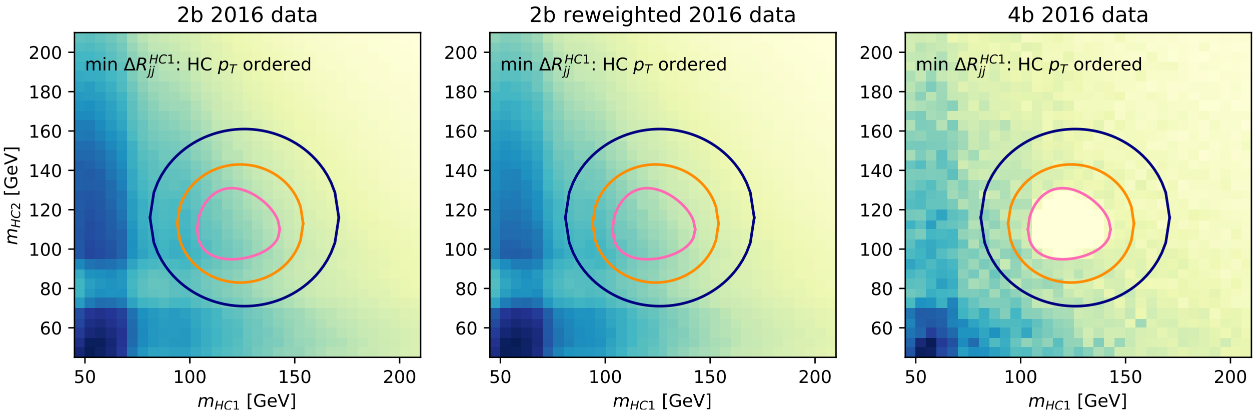


Figure 5.8: The massplane with min-ΔR pairing for 2b data (left), the nominal background estimate (center), and the blinded 4b data (right). The sharp band near 80 GeV is due to the X_{Wt} cut that removes jet pairs consistent with the W boson mass, 80 GeV. The band is present though less visible in the other massplanes due the large background peak.

expect this background to increase by a factor of order 10^3. Thus the size of the total QCD background may be much larger and must be checked. This also means that during the 3b reconstruction, it may be advantageous to make tighter selections than for 4b.

Another important consideration is that there are two ways for our signal events with four real b-jets to have only three b-tagged jets. First, as discussed above, one of the four b-jets may randomly fail the b-tagging selection. Second, one of the four b-jets may simply fall outside the jet selections ($p_T > 40$ GeV and $|\eta| < 1.5$). By checking the truth information and by performaing ΔR matching between the truth and reconstructed jets, I found that of all of our signal events with 3 reconstructed b-tags, around 50% had a fourth true b-jet within the acceptance that failed the b-tagging, and the other 50% had a b-jet outside the acceptance.

5.6.2 Fourth Jet Selection

This analysis fundamentally requires reconstructing the masses of two different Higgs bosons for background rejection and for the background modeling strategy. In the 3b category, only one of the Higgs bosons is immediately reconstructable, and we need to find a new method to reconstruct the other Higgs. The simplest approach is to devise an algorithm to select the fourth jet, since this will not only immediately provide the Higgs mass, but also the kinematics. One could alternatively imagine fitting a neural network, to predict the other Higgs kinematics given all the jets in the event, but we find the other approach more straightforward.

The question then is how to select the fourth jet, since the majority of signal events have more than four jets. In particular, we need to think carefully how to evaluate the performance of this algorithm. First, we obviously want to know in what fraction of events the algorithm picks the correct jet, which we can find by comparing the selected jet to the truth information. However, we know from above that in 50% of cases, there is no correct choice within in the jet selections. It is therefore ideal that the algorithm could also reject these events. Any signal event where the wrong jet is selected will become a background event, so the impact is two-fold: a signal event is lost, and a background event is gained. Even considering this, the total signal size is small, so the size of this effect should be small. However, if we reject these events, this means we are also rejecting additional background events that could look like signal. Given the higher 3b backgrounds, it is generally better to be aggressive on this front. Because of these considerations, we want to also optimize the number of rejected events in the case of a signal with a true b-jet outside the acceptance.

In total, there are five possible outcomes. First, if the truth b-jet is within the jet acceptance, then the algorithm can either

- Select the correct jet;
- Select the incorrect jet; or
- Reject the event;

and if the truth b-jet is outside the jet cuts, then the algorithm can either

- Select a jet (which is always incorrect); or
- Reject the event.

		4th Jet Inside Acceptance	4th Jet Outside Acceptance
Pick Jet	Correct Jet	63.4%	–
	Incorrect Jet	35.7%	100%
Reject Event		0%	0%

Table 5.4: The performance of the "highest p_T" fourth jet reconstruction algorithm.

All of these possibilities can be summarized in a table as in Table 5.4. The columns correspond to whether the fourth truth jet is matched to a reconstructed jet inside the jet acceptance cuts. The rows indicate whether the reconstruction algorithm picked the correct jet, picked the wrong jet, or rejected the event. Because the share of events between the columns is determined by the analysis acceptance cuts, it is not impacted by the reconstruction algorithm, so each column is separately normalized. The green cells indicate the correct behavior of the algorithm – either successfully choosing the correct jet, or correctly rejecting an event where the last jet is outside the acceptance. The other three cells indicate the possible error modes.

Highest-p_T The perhaps simplest strategy one might attempt is to pick the highest p_T non-b-tagged jet, since QCD background jets tend to be low energy. This strategy, however, picks the wrong jet around 1/3 of the time, as shown in Table 5.4. Furthermore, this algorithm can never reject events, indicating that background will likely be high. Adding a p_T cut on this jet may help to suppress the background, but it tends to also cause a much lower acceptance, as shown in Table 5.5.

Other similar simple selections tend to also either have a poor efficiency, or would lead to very strong sculpting of the background. For example, one might select the jet that leads to an invariant mass closest to the Higgs mass. This however, is a poor strategy for pairing b-jets when all four b-jets are known, as discussed above, and would only be worse when more jets are allowed to be checked. The background scultpting is therefore expected to be even worse than that of Figure 5.3.

		4^{th} Jet Inside Acceptance	4^{th} Jet Outside Acceptance
Pick Jet	Correct Jet	42.6%	–
	Incorrect Jet	27.7%	55.5%
Reject Event		29.8%	44.5%

Table 5.5: The performance of the "highest p_T" fourth jet reconstruction algorithm where the fourth jet is required to have $p_T > 80$ GeV.

1b-Loose and Pseudo-continuous b-tagging The raw output of the b-tagging algorithms is a b-tag score, with 0 indicating the jet is unlikely to be a b-jet, and 1 indicating that it is likely to be a jet. A threshold is chosen above which jets are considered b-jets for a given targeted efficiency, 77% in this analysis, and the efficiency of this threshold cut is carefully calibrated, providing scale factors that ensure agreement between simulation and data. One technique we might like to use for the fourth jet reconstruction is to look at the raw b-tag score, and pick jets that only just barely were below the threshold.

This technique has newly become possible in ATLAS through the use of pseudo-continuous b-tagging, which has only recently been calibrated. Typically, in analyses, one picks a threshold from the handful that are calibrated (currently 60%, 70%, 77%, and 85% efficiency working points), and uses only this threshold. Mixing can lead to calibration issues due to the fact that each threshold is calibrated on the same data-set, leading to very strong correlations. In pseudo-continuous b-tagging, one defines five different b-tagging bins, called quantiles, depending on which thresholds the b-tag score falls between. A b-jet in quantile 2, for example, would have a b-tag score that falls between the threshold for the 77% working point and the threshold for the 85% working point. Then, the efficiencies for b-jets falling into each bin are all calibrated together so that correlations are controlled. This method is called "pseudo-continuous" because it uses effectively a discretized version of the continuous b-tag score.

In the language of pseudo-continuous b-tagging, our 77% working point is equivalent to accepting jets with b-tag scores with quantile less than 3. For the fourth jet, we can then simply select the fourth jet to be one with b-tag score in quantile 2, equivalent to it passing

		4^{th} Jet Inside Acceptance	4^{th} Jet Outside Acceptance
Pick Jet	Correct Jet	46.3%	–
	Incorrect Jet	6.8%	12.7%
Reject Event		47.2%	87.3

Table 5.6: The performance of the “1b Loose” fourth jet reconstruction algorithm.

only the 85% efficient working point. In the case of multiple jets in this quantile, we can just pick the highest-p_T jet. Because this technique amounts to loosening the b-tagging requirement, it is sometimes called a “1b Loose” category.

One major challenge with this approach is that while pseudo-continuous b-tagging is calibrated in the offline reconstruction, it is not at the moment calibrated against the b-jet triggers, a complication discussed in section 4.3.3. We do not have the conditional scale factors for a jet falling into the pseudo-continuous bins *given* that it passed an online b-tagger. These could be estimated by taking the difference in efficiencies of the flat cut working points, such as $\epsilon(\text{Quantile 4}|\text{Online Pass}) = \epsilon(85\%|\text{Online Pass}) - \epsilon(77\%|\text{Online Pass})$, but this runs into the same correlations in the calibration that pseudo-continuous b-tagging was developed to circumvent in the purely offline case. That being said, this is likely a relatively small effect, since it impacts the scale factors somewhat indirectly, and the scale factors are applied only to the signal simulation.

Despite these complications, the baseline performance of the 1b Loose method is excellent, as shown in Table 5.6. When this method selects a jet, it tends to pick the correct jet the large majority of the time. Additionally, this method rejects the majority of cases where the fourth jet is outside the acceptance. Because this method is somewhat aggressive on the event rejections, it should have a low background, even if the signal efficiency is somewhat lower.

Kinematic Neural Network Because of the difficulties foreseen in calibrating a 1b-Loose strategy, we have also developed a jet selection algorithm that uses only the kinematic

		4^{th} Jet Inside Acceptance	4^{th} Jet Outside Acceptance
Pick Jet	Correct Jet	88.1%	–
	Incorrect Jet	4.7%	11.1%
Reject Event		7.2%	88.9%

Table 5.7: The performance of the "Kinematic Neural Network" fourth jet reconstruction algorithm.

information of the jets, without using any b-tagging information (aside from the fact that three of the jets are b-tagged). Rather than rely on manual kinematic cuts that may sculpt the background, we try using a neural network.

The architecture is a standard feed-forward network with 30 inputs and 8 outputs. The inputs to the neural network are the p_T, η, and ϕ of the 7 leading non-b-tagged jets, sorted by jet p_T, and the three b-tagged jets, separately sorted by jet p_T. The first 7 outputs correspond to selecting one of the 7 input untagged jets, and the last output corresponds to rejecting the event. A handful of different numbers of hidden layers and hidden layer widths are tried, with minimal differences in results. The loss function is the standard categorical cross-entropy for classification problems.

The efficiency results are summarized in Table 5.7. One major concern with this method is that the neural network could learn to reconstruct the Higgs mass, running into background sculpting issues. The plot of the background in the massplane is shown in Figure 5.9, which shows that this does not appear to be the case. While the performance on signal is impressive, this method turns out to have a much worse background rejection than the 1b Loose method, and despite higher signal acceptance, leads to an overall worse performance, as discussed in section 6.3. For this reason, this method is not optimized further.

Pair-A-Graph A relative newcomer to the 3b category is Pair-A-Graph, which as described in section 5.5.1, automatically includes 3b events by the architecture of the network. In principle, it should learn to reconstruct 3b events that are similar to the easier 4b events, leveraging the similarity in nature between the jet pairing problem and the fourth jet se-

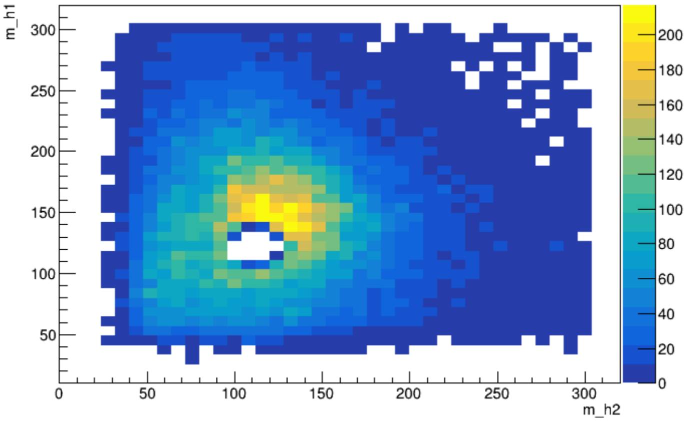


Figure 5.9: The Higgs candidate massplane for the 3b category with the kinematic neural network fourth jet reconstruction strategy.

lection problem. Because Pair-A-Graph uses both the jet four-vectors and the b-tagging quantile, it should hopefully perform better than either the 1b Loose or kinematic neural network methods alone.

There are two main disadvantages with directly using Pair-A-Graph for the fourth jet selection. First, Pair-A-Graph lacks a way to reject events. It always selects four jets to use in a pairing. Second, it makes use of the pseudo-continuous b-tagging quantiles as one of the inputs, therefore running into the same calibration issues as the 1b Loose category.

Selected Method The 1b-Loose method is selected for the final reconstruction due to the balance of performance and simplicity. However, there is substantial room for improvement in the future by including kinematic information as well.

5.7 Signal Region Optimization

The signal region in the Higgs candidate mass plane is defined by equation 5.4. This is the same definition as the previous iteration of the analysis. For the current analysis, this shape is determined to be nearly optimal in a brute-force optimization over signal region definitions of the form

$$X_{hh} = \sqrt{\left(\frac{m_{h1} - m_1^0}{r_1 m_{h1}}\right)^2 + \left(\frac{m_{h2} - m_2^0}{r_2 m_{h2}}\right)^2} < 1.6 \tag{5.7}$$

The optimization is a four-dimensional search over the parameters m_1^0 and m_2^0, which define the center point of the signal region, and r_1 and r_2, which define the width of the signal region. Because of the redundancy in degrees of freedom between X_{hh} and the r_i, X_{hh} is fixed to 1.6 for this study.

The quantity $S/\sqrt{B}$ is optimized using 2017 data and MC16d signal samples. S is computed from the integral over the signal region of a cubic spline interpolation of the mass plane histogram. The smoothing avoids flatness problems where the optimizer changes the region definition but no new events are gained or lost. Non-resonant signal MC is used as a good approximation of the resonant signals as well, since the mass plane is defined only by the Higgs masses and resolutions. A similar technique is used to compute B using the reweighted 2-tag data.

The same method is also used to optimize a purely elliptical signal region defined by

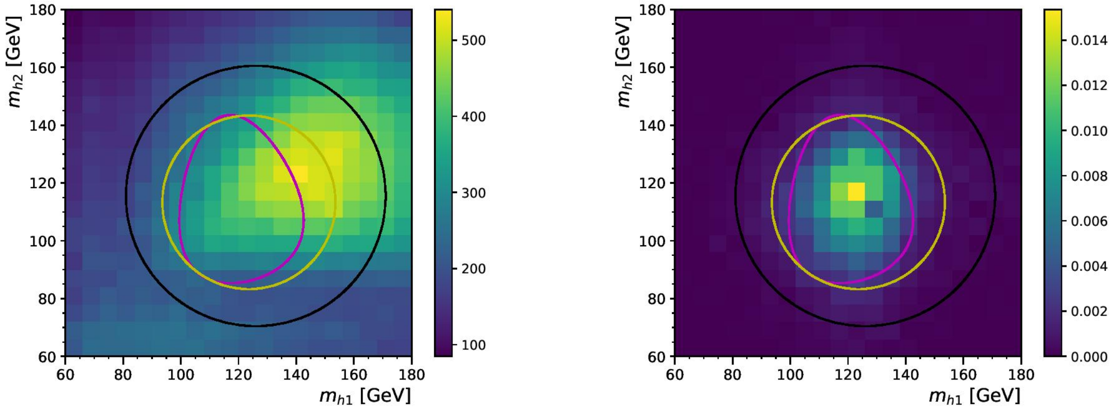
$$E_{hh} = \sqrt{\left(\frac{m_{h1} - m_1^0}{r_1 m_1^0}\right)^2 + \left(\frac{m_{h2} - m_2^0}{r_2 m_2^0}\right)^2} \tag{5.8}$$

The results of the optimization are shown in table 5.8, and the optimal signal regions are shown in figures 5.10 and 5.11. A 5-6% improvement in $S/\sqrt{B}$ is observed for the standard model non-resonant signal. The optimal definitions are also checked for resonant scalar signal samples. The largest improvement among all signal samples is for the 280 GeV scalar, which shows a 13% increase for both region shapes. For signal masses greater than 400 GeV, the improvements are comparable to or worse than the standard model non-resonant signal, dropping to a 1-2% increase for the 900 GeV scalar.

With the optimized definitions, the statistics in the validation region are greatly reduced, and the signal region extends past the validation region into the control region. A widening of the validation and control regions would thus be needed, and this is expected to increase the systematic error on the background shape due to training on events further away from the signal region. Given the size of the improvements, there is therefore no strong motivation for redefining the signal region, so the baseline definition is used.

Parameter	Baseline	Optimized X_{hh}	Optimized Ellipse
m_1^0	120 GeV	117.2 GeV	122.4 GeV
m_2^0	110 GeV	107.0 GeV	114.6 GeV
r_1	0.1	0.11	0.15
r_2	0.1	0.16	0.19
$\frac{(S/\sqrt{B})_{\text{optimized}}}{(S/\sqrt{B})_{\text{baseline}}}$	1.0	1.053	1.061

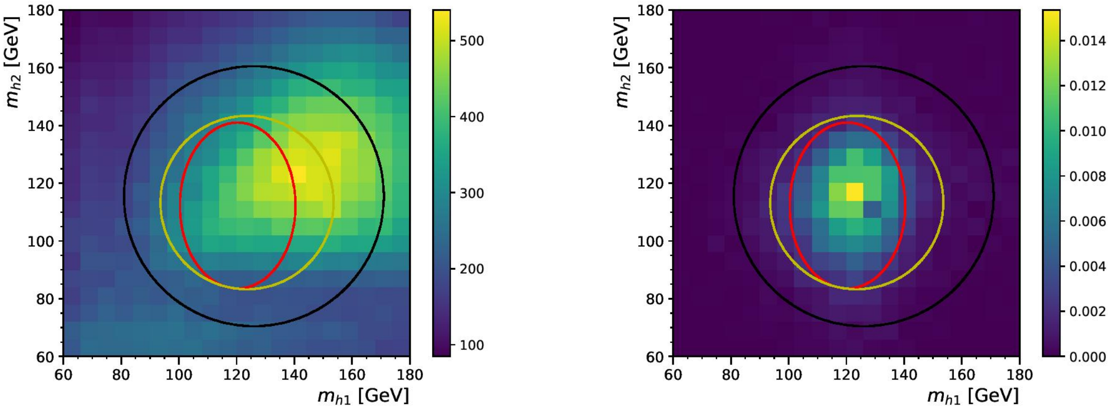
Table 5.8: The result of optimization of the signal region definition. Note that in all cases, the cut value (X_{hh} or E_{hh}) is held at 1.6, since there is a redundancy in degrees-of-freedom between the cut value and the r_i.



(a) Reweighted 2b

(b) Standard model non-resonant MC

Figure 5.10: Optimized X_{hh} signal region definition is shown in magenta. The baseline control and validation regions are outlined in black and yellow, respectively.



(a) Reweighted 2b

(b) Standard model non-resonant MC

Figure 5.11: Optimized elliptical signal region definition is shown in red. The baseline control and validation regions are outlined in black and yellow, respectively.

CHAPTER 6

BACKGROUND ESTIMATION

6.1 Nominal Strategy

One of the most important challenges in this analysis is the accurate prediction of the background. The Monte Carlo simulation for real 4b QCD is relatively poor, so we cannot rely on simulation for an accurate background estimate. Instead, it must be derived in the data.

As a small aside, the ultimate version of this analysis in the future would likely need to resort to QCD Monte Carlo simulation. Currently, we are able to average over the microphysics that leads to the QCD backgrounds. However, the best option would be to use a physically-motivated model, where much of the underlying physics could be accounted for, at least phenomenologically. This would require a dedicated measurement of multi-b QCD so that the Monte Carlo generators could be tuned to correct for the current differences. Such a measurement would require its own dedicated paper with enough turn-around time from the phenomenology community for the tuning to be implemented and used in the 4b analysis.

For the current analysis, the idea behind the data-driven nominal method is that differences between the distributions of events with 4 b-jets and events with 2 b-jets and 2 additional untagged jets should be relatively small. The motivation behind this assumption is that in QCD, b-jets are always pair produced at a gluon vertex. This could be a gluon from the hard scatter directly, in which case there are two high-p_T back-to-back b-jets, or it could be a gluon radiated in the final state that splits into two b-quarks, which will lead to two nearby b-jets with a small difference in relative momentum. The gluon-quark-quark vertex, however, has no preference for flavor, so these processes where the b-quark is replaced by any one of the other four light quarks (up, down, strange, or charm) should be equally likely, at least to first order.

There are three main effects that cause this equality to be inexact. First, the fact that the proton is composed primary of up and down quarks biases light jets to be more common. Second, gluons often do not split into separated well-defined jets, but will produce a single jet via the formation of hadrons or the radiation of additional gluons due to the color charge of the gluon. This leads to gluon jets that are not associated with a particular quark flavor. Lastly, the significant differences in the quark masses mean that the different quarks are not all kinematically equally likely, with the b quarks suppressed the most due to their higher masses.

Thus, we use the 2b2j data as a baseline background estimate, but then we correct it to match 4b more closely. We use the control region, which is an annulus around the signal region in the Higgs candidate massplane, to learn how to extrapolate from 2b to 4b. Then this extrapolation is applied in the 2b signal region to provide an estimate for the 4b signal region.

In practice, the differences are accounted for via event reweighting. Suppose we are considering some set of kinematic variables, $\boldsymbol{x}$, defined over the four jets. Then we can consider the probability that an event with these kinematics occurs in the 4b data, $P^{4b}(\boldsymbol{x})$, and the probability that this event occurs in the 2b data, $P^{2b}(\boldsymbol{x})$. We can define the ratio of probabilities,

$$w(\boldsymbol{x}) = \frac{P^{4b}(\boldsymbol{x})}{P^{2b}(\boldsymbol{x})} \tag{6.1}$$

If we know $w(\boldsymbol{x})$, then given a 2b event, we can use it to estimate 4b by assigning it a weight $w(\boldsymbol{x})$. When all of the weights are summed in histograms, then we should have distributions that estimate the 4b distribution. Thus the goal is to learn $w(\boldsymbol{x})$ as accurately as possible.

The fact that the kinematics are different but similar is just the statement that $w(\boldsymbol{x})$ should be relatively close to 1 after normalization. Generally speaking, weights far from 1 will tend to lead to larger uncertainties in the estimate because small variations in the input lead to large variations in the output. So while this method in principle could learn to reweight a flat distribution into 4b, the weights would be far from 1 and the uncertainties large.

In the last round of the analysis, the weights were learned via an iterative procedure, where weights would be derived from one kinematic variable at a time, applied to all of the events, then derived in the second variable, and so on. While this method worked well and was relatively easy to interpret, it fails to account for correlations between the chosen variables, which are non-negligible and in particular, difficult to estimate with this method.

In the current nominal reweighting procedure, we use a neural network to directly learn $w(\boldsymbol{x})$. The inputs to the neural network are 11 kinematic variables, and the output is a single scalar, the predicted weight for that event. The input variables are

- $\log(p_T)$ of the softest (4^{th} leading) of the four Higgs candidate jets
- $\log(p_T)$ of the second hardest (2^{nd} leading) of the four Higgs candidate jets
- $\log(\Delta R)$ between the closest two Higgs candidate jets
- $\log(\Delta R)$ between the other two Higgs candidate jets
- The average of the absolute values of the Higgs candidate η's
- $\log(p_T)$ of the di-Higgs system
- ΔR between the two Higgs candidate
- $\Delta\phi$ between the jets in the leading Higgs candidate
- $\Delta\phi$ between the jets in the subleading Higgs candidate
- $\log(X_{Wt})$, the log of the top veto variables
- The total number of jets in the event.

To learn $w(\boldsymbol{x})$, we need a loss function whose minimum is the ratio of the probability distributions. Let $R(\boldsymbol{x})$ denote the output of the neural network given input kinematic variables $\boldsymbol{x}$. Then one loss function that works is

$$L(R(\boldsymbol{x})) = E_{2b}\left[\sqrt{R(\boldsymbol{x})}\right] + E_{4b}\left[\frac{1}{\sqrt{R(\boldsymbol{x})}}\right] \tag{6.2}$$

where E_{2b} denotes the expectation value over the 2b data, and E_{4b} denotes the expectation value over the 4b data.

That this loss function yields the reweighting function is not hard to prove. First, from the definition of expectation value,

$$L(R(\boldsymbol{x})) = \int \left(\sqrt{R(\boldsymbol{x})}P_{2b}(\boldsymbol{x}) + \frac{1}{\sqrt{R(\boldsymbol{x})}}P_{4b}(\boldsymbol{x})\right) d\boldsymbol{x} \tag{6.3}$$

where $P_{2b}(\boldsymbol{x})$ is the true probability distribution of 2b events, and similarly for $P_{4b}(\boldsymbol{x})$. Minimizing the loss is then a standard functional minimization problem, whose minimum is given by the Euler-Lagrange equations. Denoting the integrand by $I(\boldsymbol{x})$,

$$\begin{aligned} 0 &= \frac{\partial I}{\partial R} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \frac{\partial I}{\partial\left(\frac{\partial R}{\partial x_i}\right)} \\ &= \frac{\partial I}{\partial R} \\ &= \frac{1}{2}\frac{1}{\sqrt{R(\boldsymbol{x})}}P_{2b}(\boldsymbol{x}) - \frac{1}{2}\frac{1}{[R(\boldsymbol{x})]^{3/2}}P_{4b}(\boldsymbol{x}) \\ &= \frac{1}{2}\frac{1}{\sqrt{R(\boldsymbol{x})}}\left[P_{2b}(\boldsymbol{x}) - \frac{1}{R(\boldsymbol{x})}P_{4b}(\boldsymbol{x})\right] \end{aligned} \tag{6.4}$$

which is clearly solved by $R(\boldsymbol{x}) = \frac{P_{4b}(\boldsymbol{x})}{P_{2b}(\boldsymbol{x})}$, the desired weight function.

In practice, a general neural network is not guaranteed to learn a positive weight function. To enforce this constraint, we actually train the network to output the function $Q(\boldsymbol{x}) = \log R(\boldsymbol{x})$. This ensures the weight function, which will be the exponential of the neural network output, is always positive. Substituting Q for R in the loss function, the transformed loss function is just

$$L(Q(\boldsymbol{x})) = E_{2b}\left[e^{\frac{1}{2}Q(\boldsymbol{x})}\right] + E_{4b}\left[e^{-\frac{1}{2}Q(\boldsymbol{x})}\right] \tag{6.5}$$

which will have the same minimum due to the monotonicity of the logarithm.

This method does have a handful of drawbacks. First, variables that are not directly used in the reweighting are not necessarily reweighted correctly. We have to be careful about the selection of reweighting variables and which variables the reweighted data model well. In practice, the only variable that is important to estimate correctly is m_{hh}, the invariant mass of the di-Higgs system, since we set our final limits based on the shape of the histogram of m_{hh}.

Another disadvantage of this method is that the differences between the control region and signal region are not explicitly handled. For example, a given set of kinematic variables may be relatively more likely in 4b compared to 2b in the signal region than in the control region where the network is trained. We account for control to signal region extrapolation uncertainty as described in below, in section 6.2.2.

6.2 Background Estimate Uncertainties

Quantifying the uncertainty of the background estimate is a tricky task. We can break it down into a few different components with the rather general bias-variance decomposition. Suppose we have some data with inputs $\boldsymbol{x}$ and unknown outputs y. The relationship is governed by some true but unknown function $f(\boldsymbol{x})$ and a noise term ϵ, where the noise term has a mean of zero and a finite variance. Thus, $y = f(\boldsymbol{x}) + \epsilon$. Our goal is to model $f(\boldsymbol{x})$ with some modeling function $\hat{f}(\boldsymbol{x})$ that may, for example, be trained in some manner on data where the output is known. Given the model, we would like to know the variance of the prediction compared to the true data, $E[(y - \hat{f}(\boldsymbol{x}))^2]$. One can show through some relatively straightforward algebraic manipulations that

$$E\left[(y - \hat{f}(\boldsymbol{x}))^2\right] = E[\epsilon^2] + E\left[\left(E[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})\right)^2\right] + \left(f(\boldsymbol{x}) - E[\hat{f}(\boldsymbol{x})]\right)^2 \tag{6.6}$$

The first term is the statistical noise. In the case that we are modeling a histogram, this will be just the Poisson error, the square root of the event counts in each bin, $\sqrt{N}$. The second term is the model variance. If we had many different independent and identically distributed data samples over each of which we learned a model function $\hat{f}(\boldsymbol{x})$, there would be a distribution of model functions. This is the uncertainty associated with the given training sample and training procedure. The last term is the model bias. This is the uncertainty associated with the fact that the model function is not the same function as the true unknown function.

In this analysis, the noise term is handled with the general likelihood-based framework described in Section 3.3.2. The other two terms must be estimated by hand and included in the final statistical analysis as nuisance parameters.

6.2.1 Model Variance and Bootstrapping

Because we have only one data sample (the LHC Run 2 dataset), we cannot directly measure the model variance. Instead, we rely on a common procedure called bootstrapping. In bootstrapping, the training data is resampled many times, with the model trained on each resampling. If the data sample contains N events, then we randomly draw events from the data sample N times, with replacement. In the limit of large N, this is equivalent to randomly assigning a weight to each event where the weight is drawn from a Poisson distribution with mean one. In this analysis, we do 100 different bootstrap resamplings with which we train 100 different neural networks, each initialized with different random weights.

There is a subtlety in how to convert the outputs of these 100 different bootstrapped neural networks to uncertainties in a histogram. The problem is that the output of the bootstrap procedure is 100 different weights for each event, providing *event-level* information. The ideal approach would be to make 100 different histograms of the variable of interest, one for each of the bootstrap resamplings, and the uncertainty would be the standard deviation of the counts in each bin. In practice, this runs into two problems. First is the practical problem of storing 100 different weights for every event, and second is the fact that there is no average weight for each event, which we would like to have. This means to plot the average histogram for any variable, we would have to plot all 100 variations and take the mean each time. In contrast, if we could define an average event weight, then we could use those weights to plot the nominal estimate for any distribution of interest.

To overcome this problem, we break the bootstrap uncertainty into two components. First, for each bootstrap sample, we compute a normalization factor,

$$\alpha = \frac{N_{4b}}{\sum_{i=1}^{N_{2b}} w_i} \tag{6.7}$$

which will be close to but not equal to one, due to the fact that the neural network does not necessarily preserve the normalization. Thus, we have 100 different α's, from which we can compute the mean and standard deviation, $\overline{\alpha}$ and σ_{α}. The size of the uncertainty on α we assign as a type of normalization uncertainty on the predicted histogram, where all bins fluctuate up or down by the same fractional amount.

In addition, we need to account for correlations in uncertainties between the bins. To do this, for each event we compute the mean and standard deviation of the 100 bootstrap weights. Then we assume the worst-case scenario where all events vary in the same direction at the same time. So we look at this histogram where every event weight is replaced by a fluctuation one standard deviation above the mean, and a histogram where every weight is replaced by a fluctuation one standard deviation below the mean. This will significantly overestimate the size of the variation in the histogram, but it should approximate the envelope of shape variations. Thus we re-normalize these variations to preserve the total integral of the histogram, treating them as norm-conserving shape fluctuations. The size of the fluctuations are then governed by the normalization uncertainty, as explained above. Together, these provide a shape and normalization uncertainty.

In practice, we have actually found that rather than the mean and standard deviation, our results are more stable when using the medians and interquartile ranges. Thus these are used in the baseline background uncertainty estimate. Additionally, because the bootstrap error is a bin-wise statistical uncertainty, it is added in quadrature to the Poisson uncertainty in each bin, rather than being treated as a separate nuisance parameter.

6.2.2 Model Bias and the Validation Region

The next challenge is to estimate the model bias, which is in general a more difficult task, since the underlying true distribution is unknowable. In this analysis, we believe that we learn the reweighting function very well in the control region. However, due to the slightly different kinematics in going from the control region to the signal region, the reweighting function no longer matches the true distribution. To estimate the size of this effect, we define a validation region in the massplane that is an annulus between the control region and the signal region. In the validation region, we train another reweighting neural network, and we compare the control-region-derived and validation-region-derived predictions of the signal region. The difference in these predicitons is assigned as an uncertainty.

In practice, in the validation region, we actually use the full bootstrapping procedure to derive 100 different validation region neural networks, the median of which is compared to the median control-region-derived prediction.

Additionally, rather than treating the difference as a single nuisance parameter, we split it into two nuisance parameters using the variable H_T, which is the scalar sum of the p_T of every jet in the event. Thus we have a low-H_T nuisance parameter, which is the difference between the control and validation derived predictions for events with $H_T < 300$ GeV, and a high-p_T shape nuisance parameter, defined similarly for $H_T > 300$ GeV.

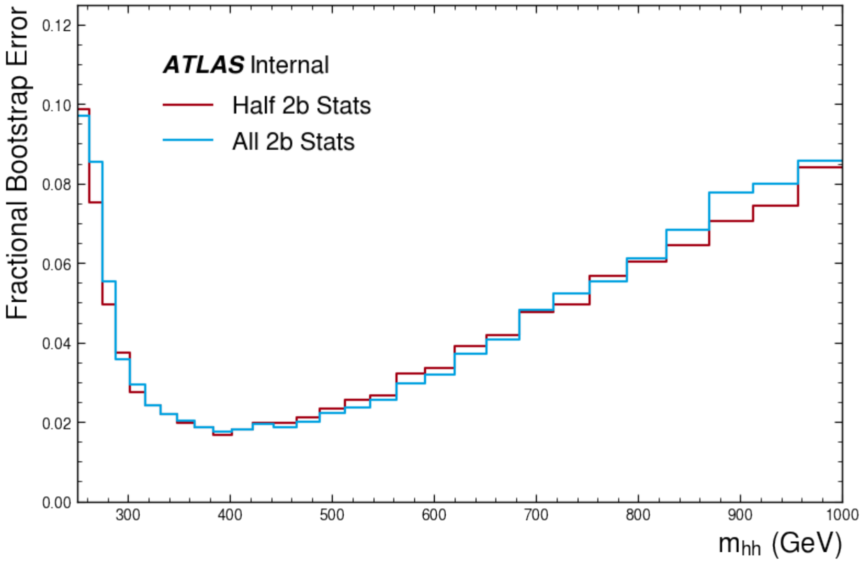
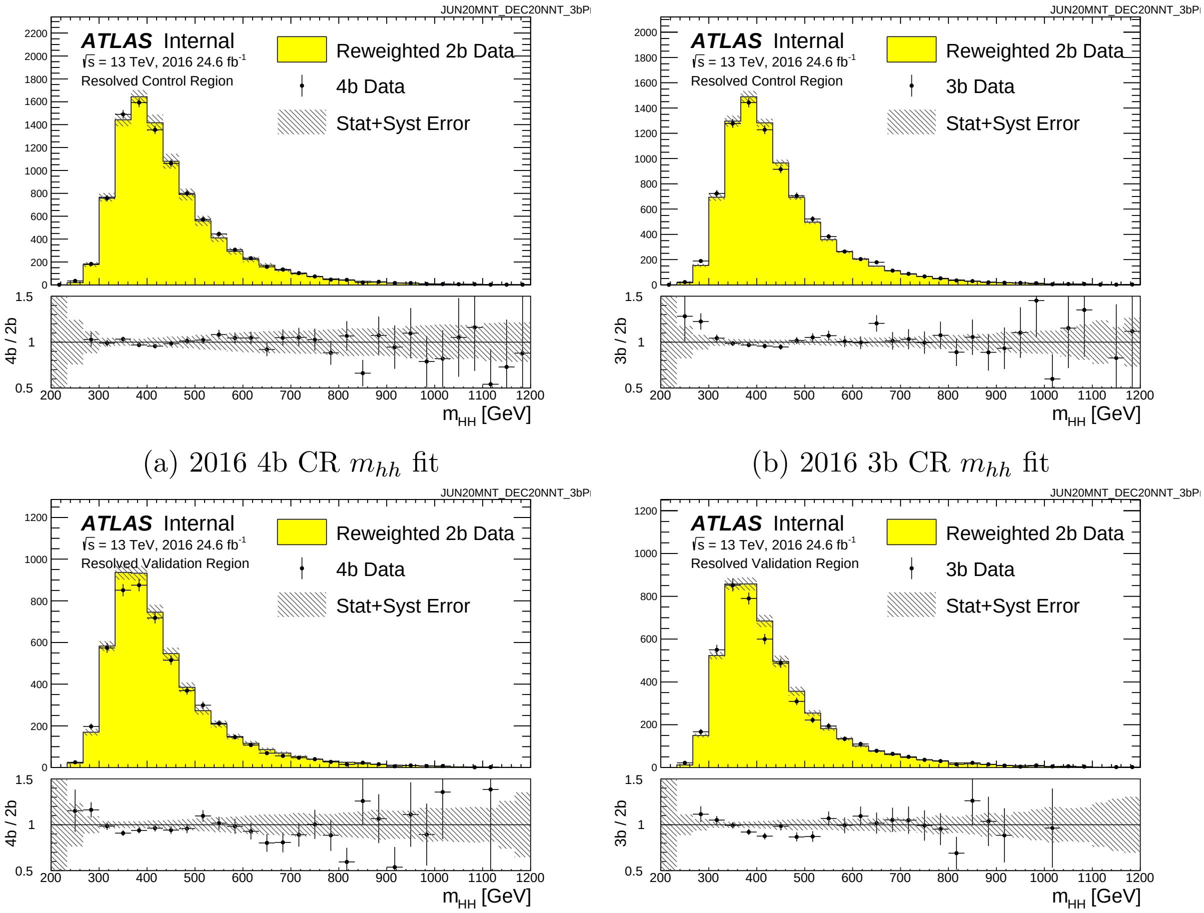


Figure 6.1: The relative bootstrap uncertainty for 2017 4b data, derived using both all of the 2b data and half of the 2b data.

6.3 3b Background Estimation

There are a few different options to derive a background for the 3b category. One could add it inclusively to 4b, for example, and derive one background estimate for the sum of the data. This is not preferred, however, because we know that due to the different flavor tag rejections, the backgrounds will be different between 3b and 4b. If 3b has a higher relative background, then we would be reducing the 4b S/B unnecessarily.

For this reason, we choose to derive a separate background estimate for 3b and 4b. However, if we do derive the background estimate using the same 2b data, then the estimate would be correlated in a complicated manner. For this reason, we need to split the 2b data, using a subset to derive the estimate for 4b and a subset for 3b. Reducing the 2b statistics, however, could increase the variance of the trained models, as measured by the bootstrapping procedure. Fortunately, even when halving the 2b statistics, the size of the bootstrap error is unchanged, as shown in Figure 6.1, indicating that it is the 4b and 3b statistics limiting the precision of the estimate.



(a) 2016 4b CR m_{hh} fit

(b) 2016 3b CR m_{hh} fit

(c) 2016 4b VR m_{hh} fit

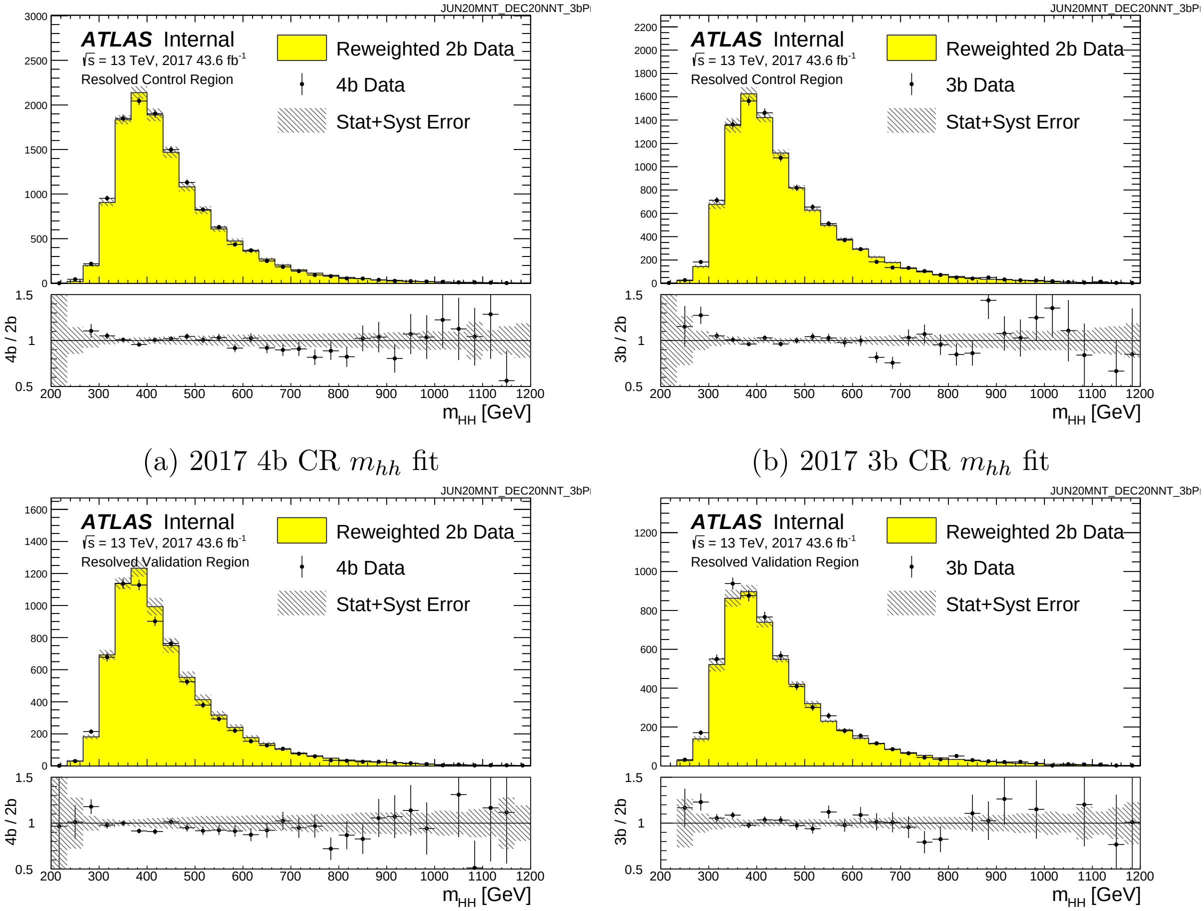
(d) 2016 3b VR m_{hh} fit

Figure 6.2: The 2016 2b reweighted predictions of the m_{hh} distribution in the control region (top) and validation region (bottom).

6.4 Background Fit

Since the variable m_{hh} is used to set the final limit, this is the most important variable to estimate correctly. Figures 6.2-6.4 show the m_{hh} prediction from the reweighted 2b data in the control and validation regions for each year, for both the 4b and 3b categories.

The histograms of all of the variables used to derive the background estimate are included in the Appendix.



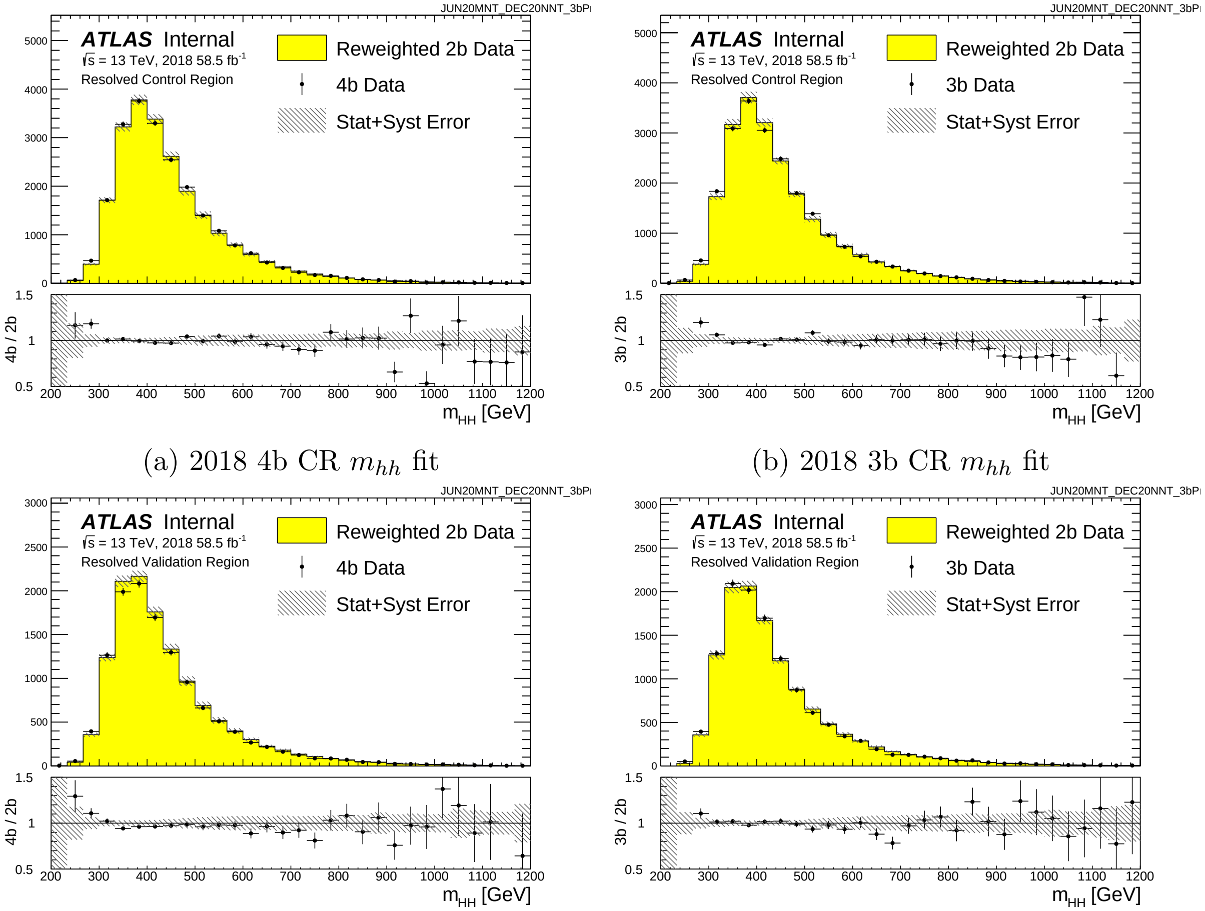
(a) 2017 4b CR m_{hh} fit

(b) 2017 3b CR m_{hh} fit

(c) 2017 4b VR m_{hh} fit

(d) 2017 3b VR m_{hh} fit

Figure 6.3: The 2017 2b reweighted predictions of the m_{hh} distribution in the control region (top) and validation region (bottom).



(a) 2018 4b CR m_{hh} fit

(b) 2018 3b CR m_{hh} fit

(c) 2018 4b VR m_{hh} fit

(d) 2018 3b VR m_{hh} fit

Figure 6.4: The 2018 2b reweighted predictions of the m_{hh} distribution in the control region (top) and validation region (bottom).

CHAPTER 7

RESULTS

7.1 Unblinded Results

The unblinded m_{hh} histograms for each year are shown in Figures 7.1, 7.2, and 7.3. The agreement between the predicted background and the observed data is reasonable, though there appears to be a slight systematic overestimate of the background at the peak. The last round of the analysis with only the 2015 and 2016 data (27.5 fb^{-1}) also observed a similar level of mis-modeling despite a smaller data set and a different background estimation method, which is evidence of the difficulty of modeling the background. Since we see no significant excesses, we use the observed data to set limits.

For the standard model signal, we can use these histograms to calculate a 95% expected and observed limit on μ, the Standard Model signal strength. The limit on μ corresponds to the largest μ at which, had the Standard Model signal had a cross section of $\mu \times \sigma_{\text{SM}}$, we would have been 95% certain to observe it. The limits are shown in Table 7.1. The limit on μ corresponds to an observed limit on the cross-section of $\sigma(pp \to HH) < 284$ fb, compared to the Standard Model expectation of $\sigma(pp \to HH) = 31.1$ fb, assuming the Standard Model Higgs branching ratio to $b\bar{b}$.

We also can set a limit on the Higgs self-coupling κ_λ. Because the m_{hh} distribution changes significantly as a function of κ_λ in addition to the cross-section, we treat each κ_λ value as a separate signal and compute the 95% confidence limit on μ for each point. The κ_λ values for which the observed μ is 1 correspond to the 95% confidence interval on κ_λ. The limit on the cross-section as a function of κ_λ is shown in Figure 7.4, corresponding to the 95% confidence interval $\kappa_\lambda \in [-5.5, 12.7]$. Given that the Standard Model Higgs self-coupling is $\lambda_{\text{SM}} = 0.13$, this corresponds to a constraint on the Higgs self-coupling of $\lambda_{hhh} \in [-0.71, 1.6]$.

Obs.	-2σ	-1σ	Exp.	+1σ	+2σ
9.12	7.7	10.0	13.1	19.3	24.5

Table 7.1: The observed and expected limits on the Standard Model signal strength μ.

7.2 Interpretation

The observed (expected) limit of 9.12 (13.1) represents a 41% (58%) improvement over the last round of the analysis, which achieved an observed (expected) limit of 12.9 (20.7). Given that the data set is 4.6 times larger, we see that the improvement is worse than the $\sqrt{N}$ scaling based on Poisson statistics, which would predict an expected limit around $20.7/\sqrt{4.6} = 9.7$. This is a statement of the fact that the analysis is systematics-limited, particularly by the size of the uncertainties on the background estimate.

Despite this, we believe that the estimates of the systematic errors are better controlled in the current analysis. There are several potential future improvements that are able to significantly reduce the total systematic error on the background estimate. One of the largest improvements comes from a new definition of the X_{Wt} variable that rejects the $t\bar{t}$ background. Currently, all permutations of jets are checked to see if any are consistent with the top mass, regardless of the b-tagging information. This definition causes the impact of this cut to be different between the 2b and 4b data, increasing the difficulty of the reweighting. A new definition requires also that the b-tagging information is consistent with a top decay chain. This does not impact the 4b data, but does improve the agreement between the 2b and 4b data, significantly reducing the systematic uncertainties on the background estimate. This change should allow this channel to reach an expected limit comfortably below $\mu = 10$ and stay competitive with the other primary di-Higgs decay channels.

7.3 Improvements from a 3b Category

One of the key analysis changes to be tested in this thesis is the extent to which including a 3b category improves the expected results. We can estimate this by using the background

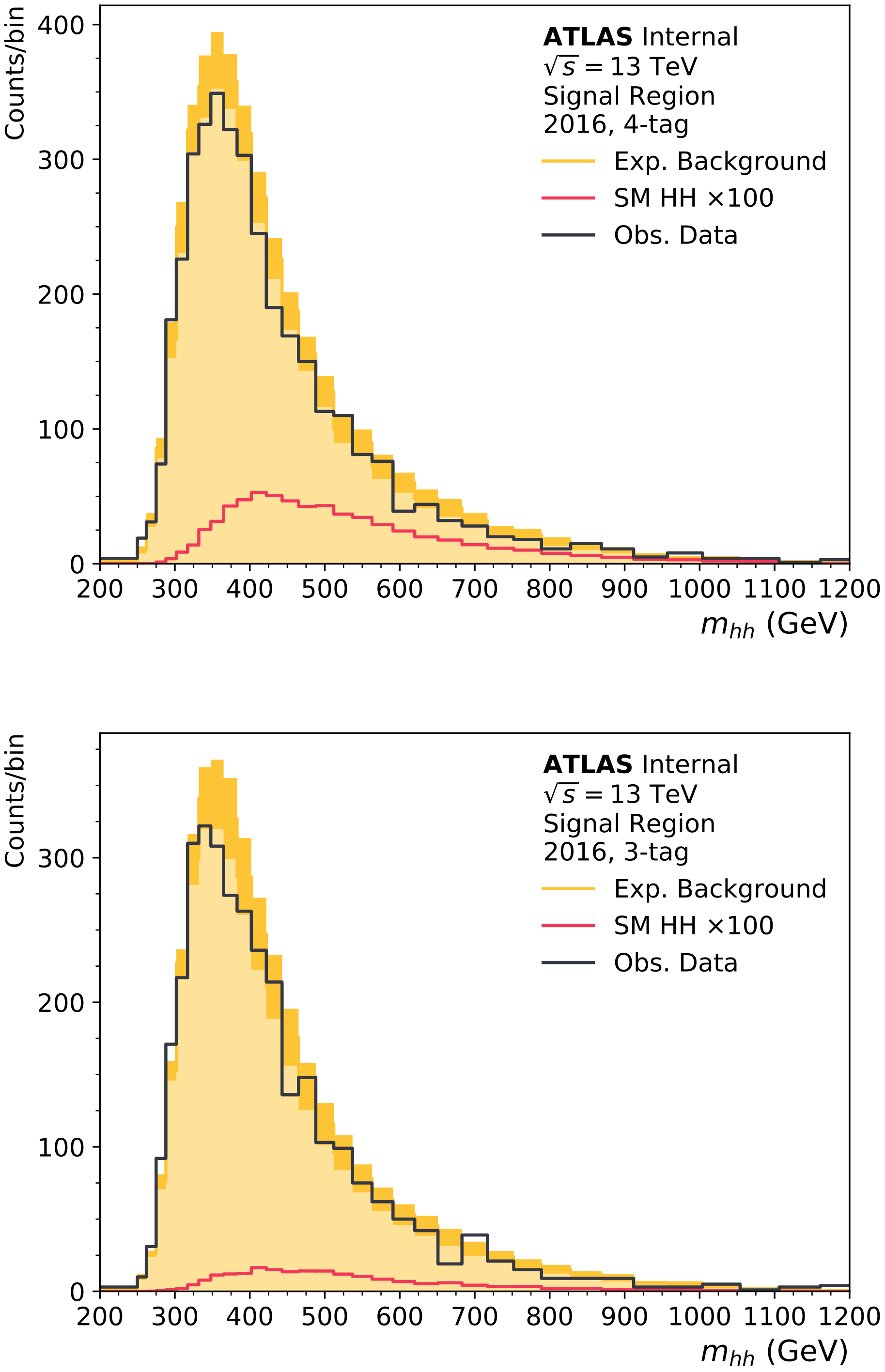


Figure 7.1: The observed and expected m_{hh} distributions for the 2016 data in the 4-tag (top) and 3-tag (bottom) categories. The total error in the background prediction is shown in dark yellow.

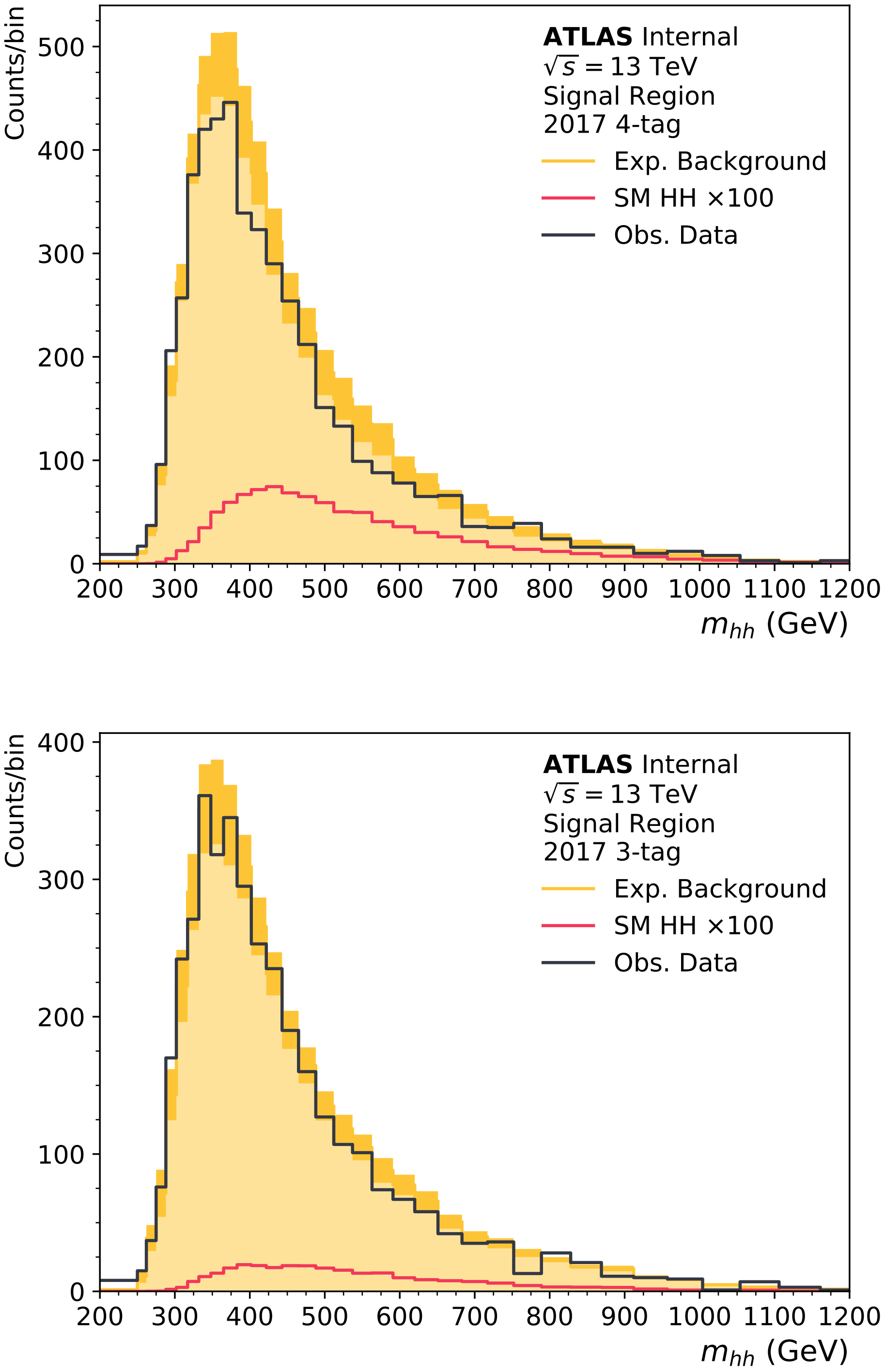


Figure 7.2: The observed and expected m_{hh} distributions for the 2017 data in the 4-tag (top) and 3-tag (bottom) categories. The total error in the background prediction is shown in dark yellow.

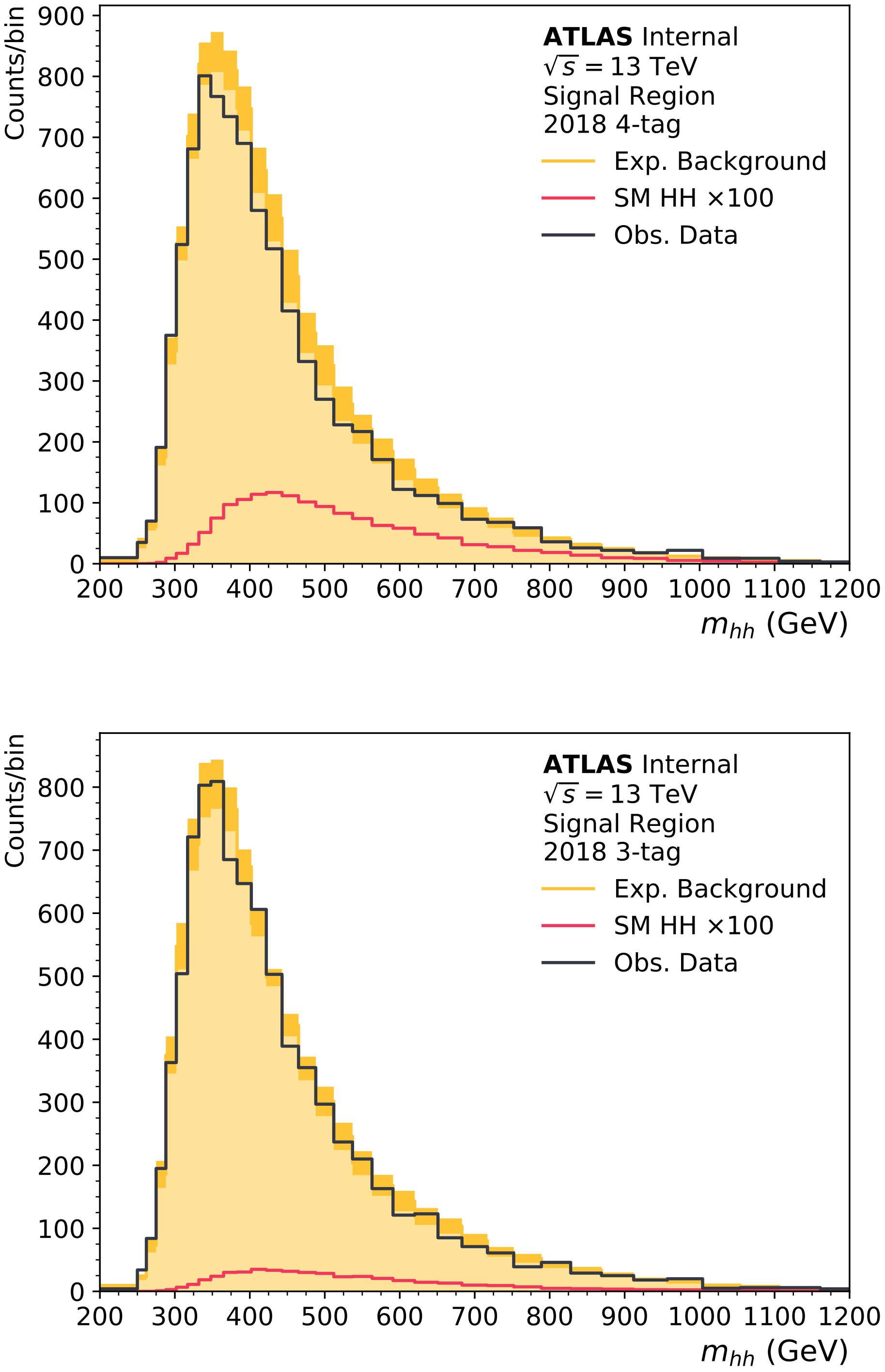


Figure 7.3: The observed and expected m_{hh} distributions for the 2018 data in the 4-tag (top) and 3-tag (bottom) categories. The total error in the background prediction is shown in dark yellow.

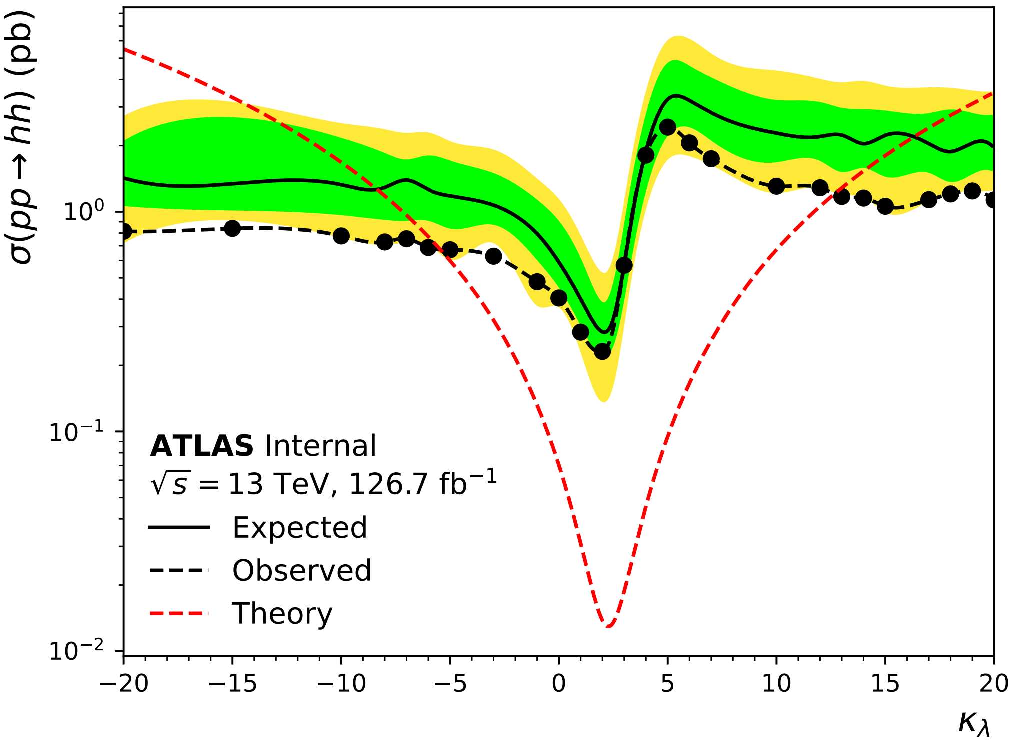


Figure 7.4: The 95% confidence upper limit on the cross-section as a function of κ_λ. The green and yellow bands are the 1σ and 2σ confidence interval on the expected limit.

prediction as the “observed” data in the calculation of limits. Figure 7.5 compares the expected limits as a function of κ_λ with and without a 3b category, including the full systematic errors. We can see that at the Standard Model, there is a 15% improvement in the expected limit, with a 10% improvement for other values of κ_λ.

However, improvements to the size of the systematic errors may change this conclusion. To test the extreme case of no systematic error, we can compute the limits assuming only statistical errors. The results are shown in Figure 7.6. Unfortunately, in this limit, the addition of the 3b category does not significantly impact the final expected limit. This is likely due to the smaller signal size, which is particularly evident in Figures 7.1-7.3. Thus for the 3b category to remain useful after the expected reductions in the systematic errors, the selection efficiency may need to be improved. Given that these results use the simplest fourth jet reconstruction, there is room for improvement.

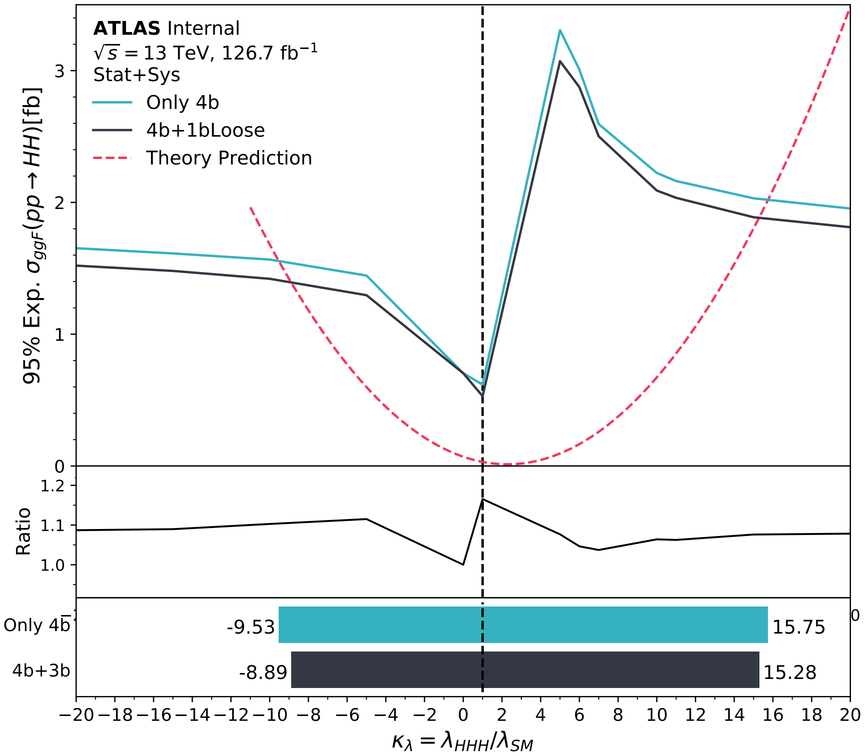


Figure 7.5: The expected limit as a function of κ_λ with and without a 3b category, including all systematic errors.

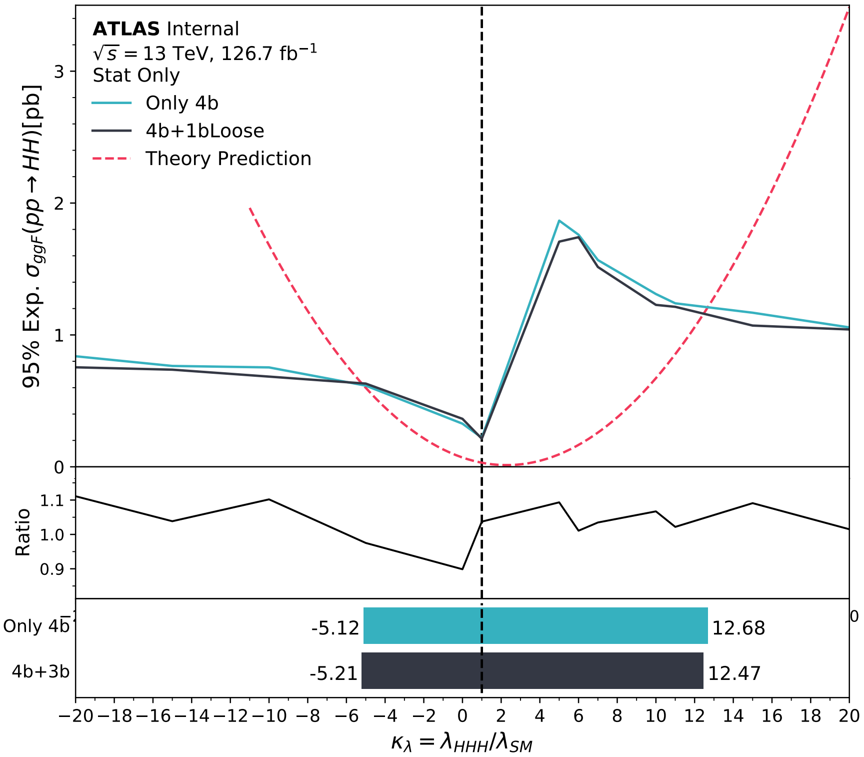


Figure 7.6: The expected limit as a function of κ_λ with and without a 3b category, including only statistical errors.

CHAPTER 8

FUTURE TRIGGERS

8.1 Run 2 Trigger Limitations

One of the major challenges of the nonresonant $hh \to 4b$ analysis is the trigger. Figure 2.6 shows that sensitivity to the Higgs self-coupling is highest at low m_{hh}, which means that the b-jet momenta will tend to be small. Thus we are tasked with designing a trigger for a low-momentum all-hadronic final state, for which the QCD backgrounds are enormous. Partly for this reason, $hh \to 4b$ has become a benchmark for future trigger design.

8.2 The Asymmetric Trigger Idea

One idea that is being implemented for Run 3 (and that is a contender for the HL-LHC) is the idea of using asymmetric jet trigger p_T thresholds. For our four-jet final state, we can use four different momentum thresholds. This technique can significantly increase the signal acceptance while maintaining or even reducing the total background rate. In fact, the asymmetric triggers were proposed in the context of FTK where full-scan tracking would be available, making b-tagging four jets much less time consuming. However, the idea has been sufficiently promising to be implemented even without FTK.

While the $hh \to 4b$ analysis in Run 2 used an "OR" of several different triggers, the most important was the 4j35 trigger that required four jets above 35 GeV, with at least two of them b-tagged (at various working points depending on the rate). For the Standard Model signal, this is decently efficient. However, it is least efficient at the lowest m_{hh} values, where the effects of the Higgs self-coupling are most important. When κ_λ is varied to be large, the Higgs self-coupling dominates over the triangle diagram, the signal has a soft m_{hh} spectrum, and the Run 2 triggers are inefficient.

In order to improve this, we would like to lower the jet thresholds. However, this is already a high rate and CPU-intense trigger, so we cannot flatly lower the threshold. Instead, we can try lowering just the softest jet threshold. It turns out that the soft signal events that fail the 4j35 trigger tend to have only one or two jets below the threshold. Thus we lower the threshold on the softest jet as much as possible. This alone would lead to a large increase in rate. However, it can be balanced by increasing the jet threshold on the leading jet. The leading jet for the signal, even at low m_{hh}, is rarely below 35 GeV, partly due to the combinatorics. There are four chances to get a high p_T jet (and similarly, four chances to get a low p_T jet). There is no need to maintain a 35 GeV trigger threshold when, as shown in Figure 8.1, the leading jet is almost never this soft. This logic does not just apply to the first and fourth jets; we can also lower the third jet threshold somewhat as long as we raise the second jet threshold to balance the increased rate.

In particular, one can use these tiered thresholds both at Level 1 and at the HLT. At Level 1, no b-tagging can be run due to the time it takes to reconstruct tracks, but one can still use asymmetric thresholds. In fact, the Run 3 baseline L1 trigger is actually only a three jet trigger because the fourth jet is so soft that it is often missed at L1.

The single-jet efficiencies as a function of p_T threshold (i.e. the number of events with the n^{th} jet above the given threshold) are shown in Figure 8.2.

Some preliminary threshold tuning has been done for Run 3 and for the HL-LHC. The final thresholds to be used are not yet fixed due to ongoing hardware upgrades to the trigger that improve jet reconstruction. The baseline thresholds are summarized in Table 8.1.

8.2.1 b-tagging and the Delayed Stream

One additional challenge with a four b-jet final state is that we must do some b-tagging in order to reject non-b-jet backgrounds. However, b-tagging online and offline are slightly different (since the jet reconstruction is different), so in order to maintain maximum offline acceptance, we want the loosest possible b-tagging online that still leaves the total trigger rate acceptable. Additionally, the 2b region has been critical for our background estimates

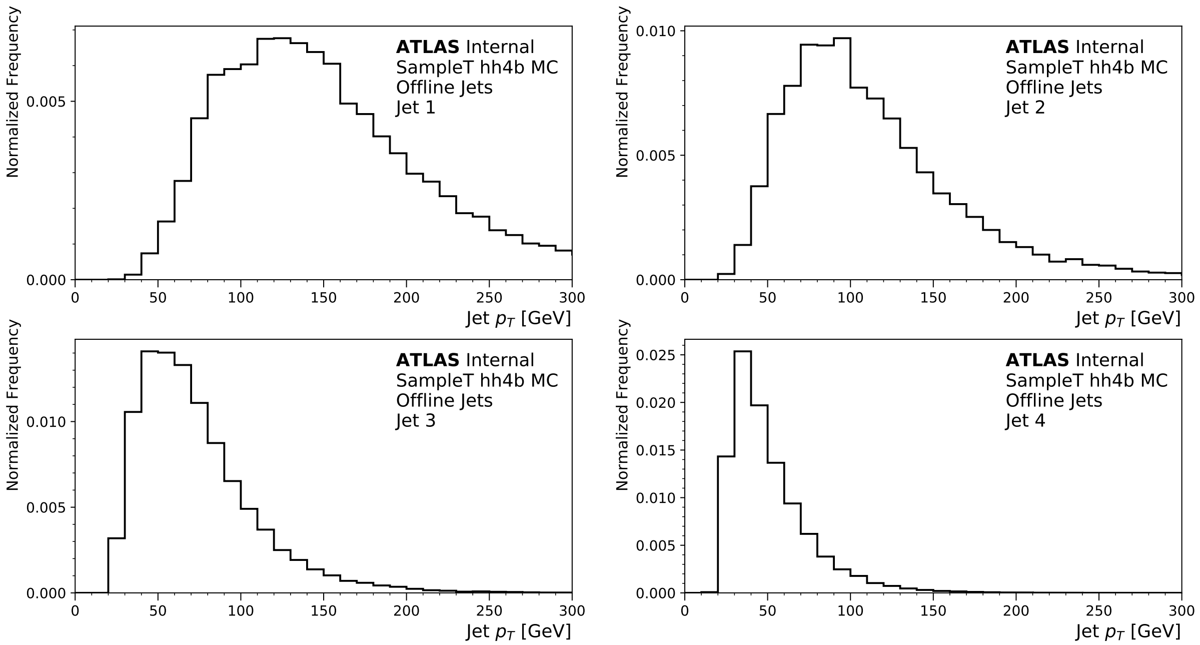
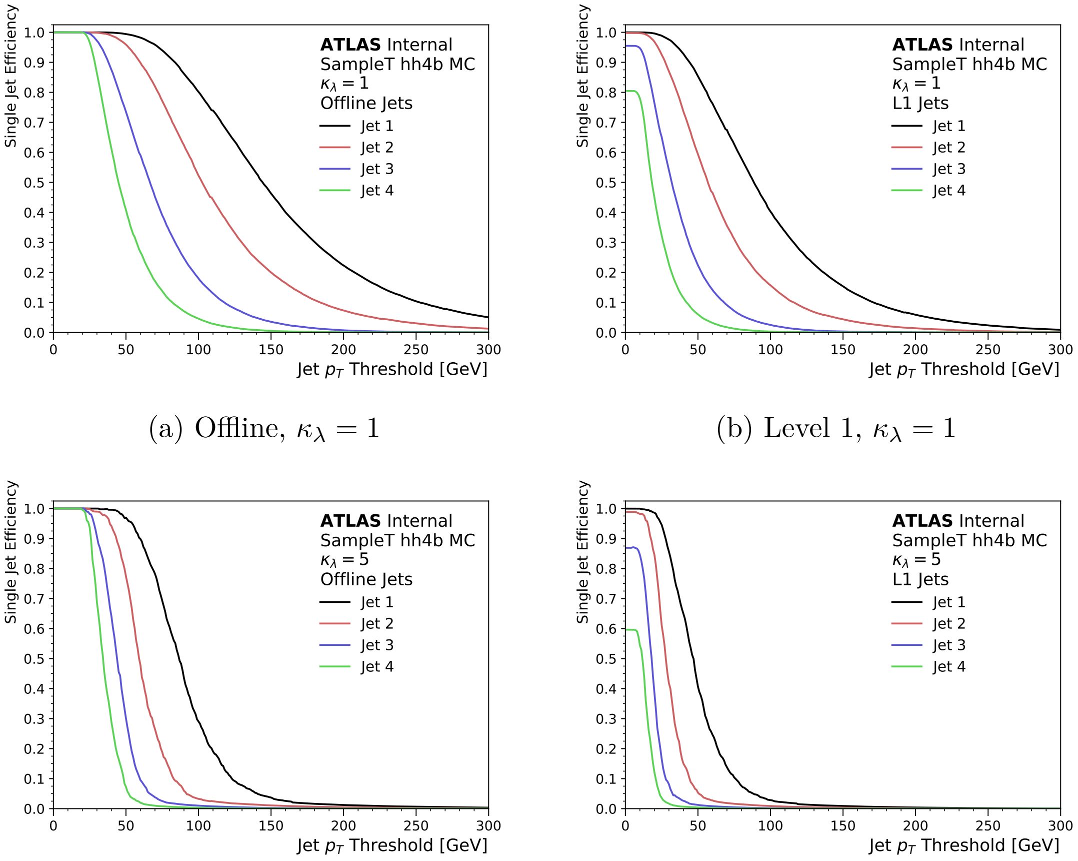


Figure 8.1: The Standard Model ordered offline jet p_T distributions from simulation. The cutoff at 20 GeV is due to the fact that jets below 20 GeV are not nominally reconstructed.

because it provides events that are very similar to 4b, so we would like to maintain a purely 2b trigger if possible.

There are currently two b-tagging options under consideration for Run 3, a 2b or a 3b trigger, where we request any two or three of the jets passing the p_T thresholds to be b-tagged. The 3b rate for the baseline is around 30 Hz, which is large but manageable. The 2b trigger, however, would have a rate near 160 Hz, which is too large for the main trigger stream. However, in Run 3, there is an option to write the triggered events to the "delayed stream," which stores the output data on a large bank of tapes in an un-reconstructed form. Then, when the CPUs are free, the data can be later reconstructed, usually once data-taking is over. This technique is feasible because the limiting factor for the output data rate is not the bandwidth of the readout, but the ability to do prompt reconstruction before saving the data. The total available bandwidth for the delayed stream is large, and the number of triggers using it appears to be small, so there is a lot of available bandwidth for this trigger. It is likely the final configuration will include a pre-scaled 2b trigger in the main trigger stream with the bulk of the data being recorded via the delayed stream.



(a) Offline, $\kappa_\lambda = 1$

(b) Level 1, $\kappa_\lambda = 1$

(c) Offline, $\kappa_\lambda = 5$

(d) Level 1, $\kappa_\lambda = 5$

Figure 8.2: The single jet efficiencies offline (close to HLT thresholds) and at Level 1 (with Run 2 reconstruction) for the standard model and $\kappa_\lambda = 5$. The plateaus at 20 GeV are due to the fact that jets below 20 GeV are not nominally reconstructed.

Trigger	Jet 1	Jet 2	Jet 3	Jet 4	Rate
Run 3 L1	45	15	15	–	7.4 kHz
Run 3 HLT	80	55	28	20	30/160 Hz (3b/2b)
HL-LHC L1	60 (110)	45 (90)	30 (60)	15 (30)	100 kHz

Table 8.1: Baseline asymmetric trigger p_T thresholds, in GeV, for the Run 3 L1 and HLT triggers, and an asymmetric L1 trigger at HL-LHC. The parentheses in the HL-LHC trigger indicate the calibrated offline-equivalent p_T thresholds that can be compared to Figure 8.1. The Run 3 HLT thresholds are close to offline values, but the Run 3 L1 thresholds are taken from the Run 2 L1 system and are far from offline. The L1 jet p_T calibration is however expected to be improved in Run 3 with the upgraded L1Calo trigger hardware.

Since we have now sorted the jets in the trigger, one might be tempted to request a trigger where the leading two jets are b-tagged. This, however, is a disaster for the efficiency. If the tagging efficiency is 77%, then the probability of a four b-jet event passing the trigger is the probability that the leading and subleading jets are both b-tagged, $(0.77)^2 = 59\%$. Instead, a better approach is to leverage combinatorics and ask that at least *any* two of the four jets are b-tagged. The total trigger efficiency in that case is $6(0.77)^2(1-0.77)^2 + 4(0.77)^3(1-0.77) + (0.77)^4 = 96\%$. The efficiency for real 2b will still be somewhat limited, but this is acceptable since 2b events are used only to derive the background estimate and are not statistically limited. Thus it is critical that the trigger algorithm run b-tagging down to the p_T of the lowest jet threshold. Because a goal of the asymmetric trigger is to lower the fourth jet threshold as much as possible, this means running b-tagging on low-p_T jets, which will always have a sizeable rate. The asymmetric trigger is thus computationally somewhat expensive.

8.3 Impact of the Asymmetric Triggers in Run 3

The asymmetric trigger will almost certainly be used throughout Run 3, starting in 2022. While the total integrated luminosity will only double by the end of Run 3, with the new triggers, our total number of signal events will increase by significantly more than a factor of 2, gaining the equivalent of substantial LHC running time.

Figure 8.3 shows the rates vs efficiency plot for various asymmetric trigger configurations, as well as for some primary Run 2 triggers. The rates are estimated from the central trigger simulation tuned to an instantaneous luminosity of 2×10^{34} cm^{-2} s^{-1}. The efficiency is computed with respect to signal with four offline b-tags, which at the time of this trigger simulation, used a 70% efficiency b-tagging working point. The conclusions are similar for different working points and for a 3b region. The cyan circle in the lower right corner is the baseline 3 b-tag Run 3 trigger, while the red star is the best 2018 trigger, based on 4j35.

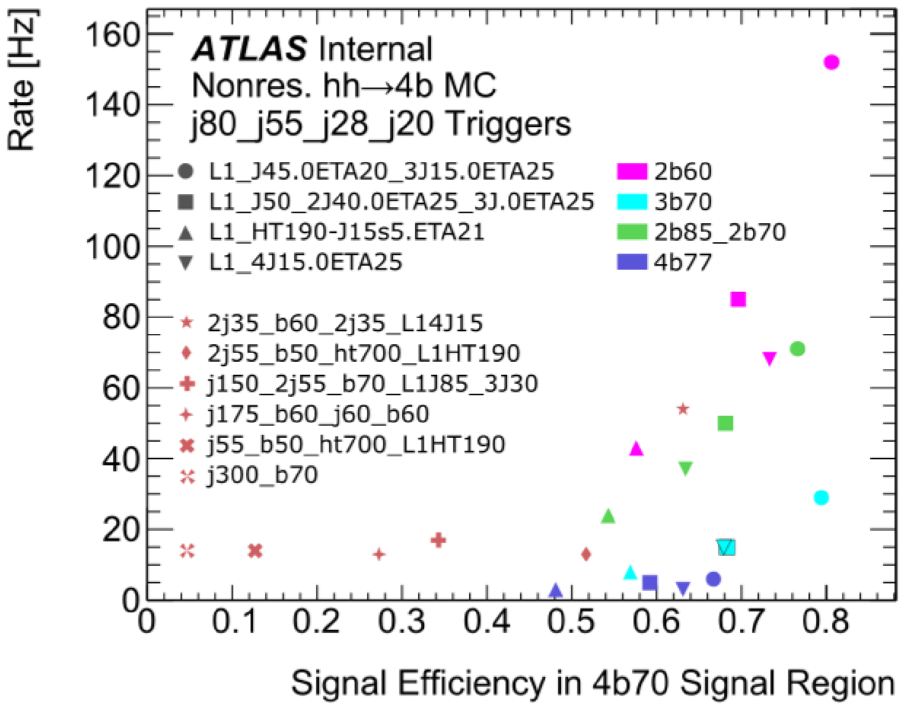


Figure 8.3: The rate vs efficiency for various asymmetric triggers and L1 seeds under consideration. The baseline new L1 trigger is represented by circles. The colors indicate different b-tagging working points in the HLT with the notation NbXX corresponding to N b-tags at the XX% efficient working point. The various red shapes correspond to the 6 single most efficient triggers in 2018.

One can see that the signal efficiency is substantially higher but with less than half the rate. The magenta cirlce in the upper right corner is the baseline 2b trigger for Run 3 that would be written to the delayed stream.

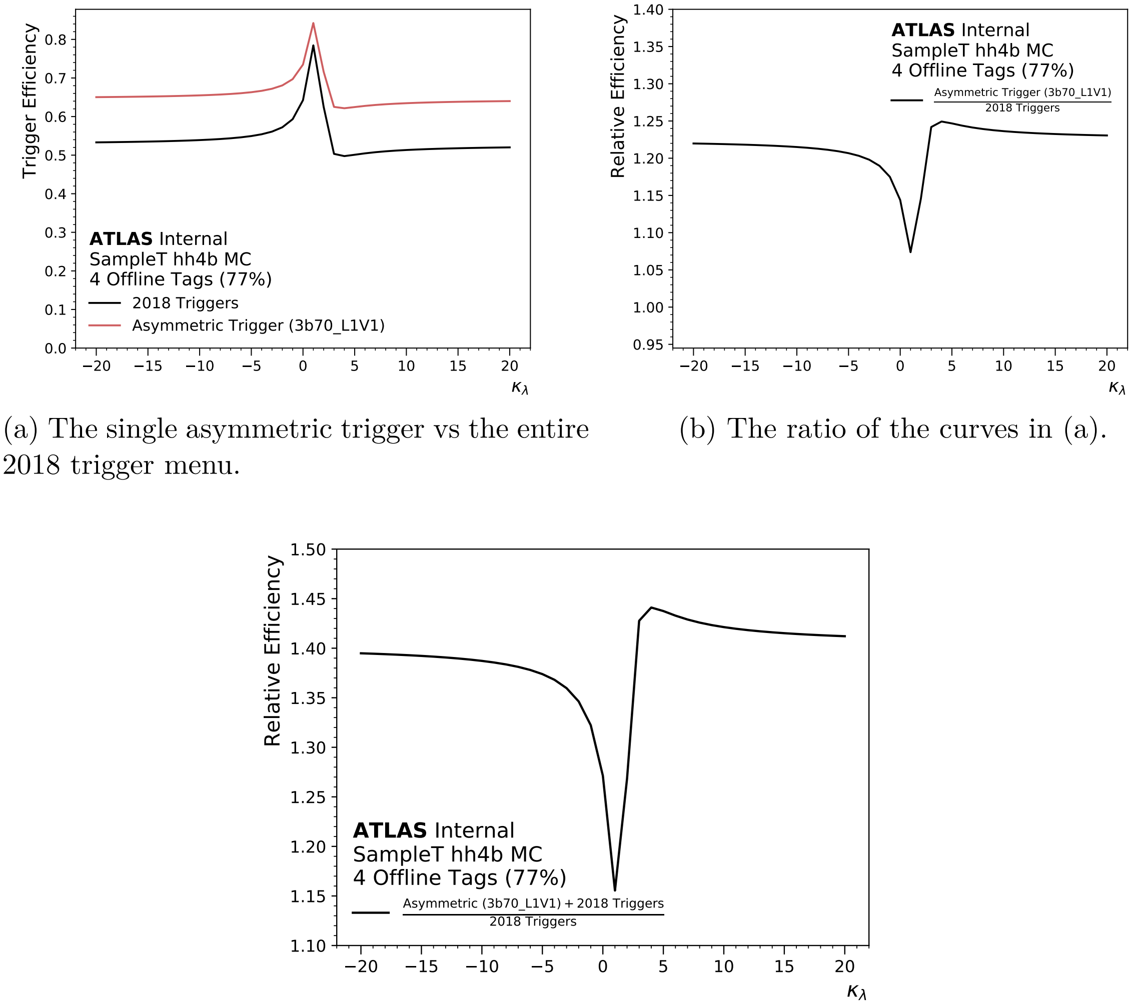
Figure 8.3 does not account for the fact that the different triggers in 2018 are all combined in the actual analysis. Perhaps the asymmetric trigger would fare poorly in comparison when the total "OR" is considered. Figures 8.4 and 8.5 show the efficiencies of a 3b and 2b trigger, respectively, when compared to the entire 2018 trigger menu. First, notice the spiking behavior near $\kappa_\lambda = 1$. This is due to the fact that the standard model is actually near the hardest m_{hh} spectrum. Even small deviation in κ_λ tend to significantly increase the proportion of low m_{hh} events.

From Figures 8.4(a,b) and 8.5(a,b), we can see that the asymmetric triggers are up to 20-25% more efficient than the entire 2018 menu. Thus even the “OR” of all of the triggers considered in Figure 8.3 is still significantly less efficient than the baseline trigger. None of the 2018 triggers can capture the low m_{hh} events. Furthermore, the 2018 triggers are still somewhat orthogonal to the asymmetric trigger. There is no strong motivation to remove the 2018 triggers in Run 3, and the combined efficiencies can reach over a 40% increase. This is because the 2018 triggers are better at catching events where the p_T is more democratically shared among the four jets, such as the 4j35 trigger or in H_T triggers that use the scalar sum of p_T's. There may also be some gains from the 1b triggers that require fewer total online jets.

Comparing Figures 8.4 and 8.5, we can see that when we require 4 offline b-tags, there is only a small efficiency gain in switching from 3b (70% working point) trigger to 2b (60% working point) trigger. This is because requiring 3 b-tags is already efficient for tagging 4 real offline b-jets. However, there is a dramatic difference for a 3b signal region, an efficiency increase over 20%, as shown in Figure 8.6. Thus for an optimal 3b region, the 2b trigger is strongly preferred. This efficiency bonus for the 3b region is in addition to the preserved background estimate that is the main motivator for using a 2b trigger.

The above plots show pretty dramatic increases in the signal efficiency. However, we know that the signal that is recovered is at low m_{hh}, where backgrounds are higher. An important question is how much this signal improves the final limits and whether any potential improvement is lost to larger backgrounds.

This is a difficult question to answer for several reasons. First, it is hard to predict future analysis improvements. Perhaps in the future, QCD will be measured precisely enough to reduce the background systematics to a level where the signal is immediately useful. There could also be better signal vs background discrimination to reduce the larger backgrounds at low m_{hh}. Regardless, it is not possible for a future “smarter” analysis to use this signal if it is not triggered.

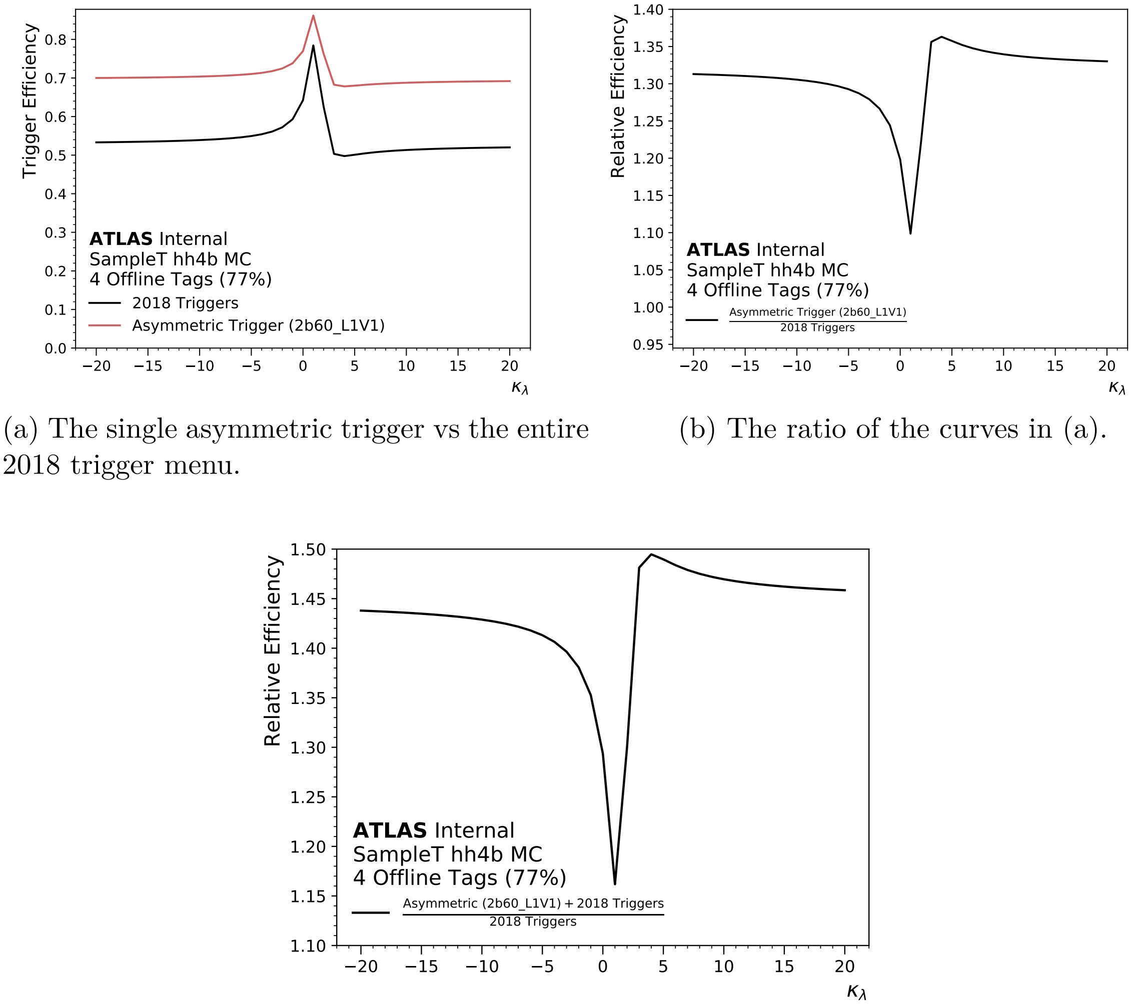


(a) The single asymmetric trigger vs the entire 2018 trigger menu.

(b) The ratio of the curves in (a).

(c) The total efficiency gain when the asymmetric trigger is combined with the 2018 triggers.

Figure 8.4: Trigger efficiencies as a function of κ_λ for the Run 3 baseline asymmetric trigger with 3 online b-tags at the 70% working point and an offline signal region requiring 4 offline b-tags at the 77% working point.

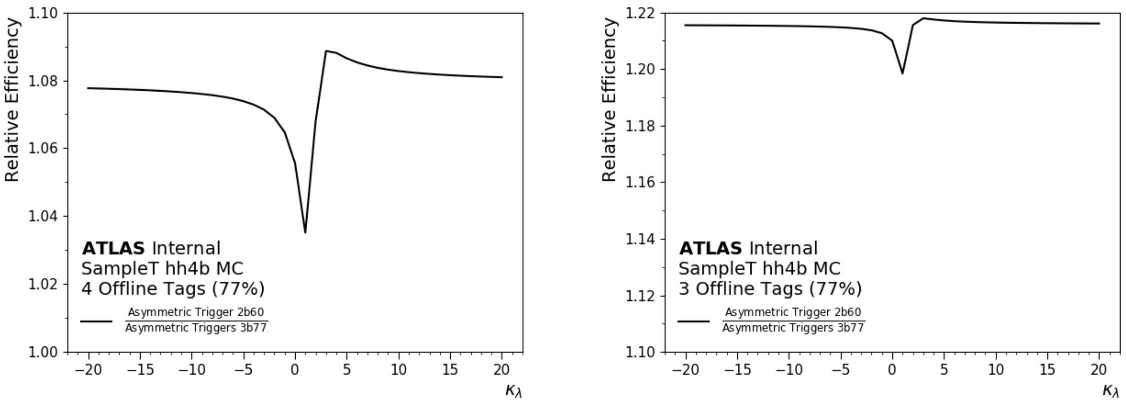


(a) The single asymmetric trigger vs the entire 2018 trigger menu.

(b) The ratio of the curves in (a).

(c) The total efficiency gain when the asymmetric trigger is combined with the 2018 triggers.

Figure 8.5: Trigger efficiencies as a function of κ_λ for the Run 3 baseline asymmetric trigger with 2 online b-tags at the 60% working point and an offline signal region requiring 4 offline b-tags at the 77% working point.



(a) 4 offline b-tags at the 77% working point (b) 3 offline b-tags at the 77% working point

Figure 8.6: Trigger efficiencies as a function of κ_λ for the Run 3 baseline asymmetric trigger for an offline signal region with 3 b-tags at the 77% working point.

The other major difficulty is that this trigger has never run in the menu, and we don't have any QCD data with four b-tags in this range of p_T. This presents a problem for our data-driven background estimate. For these reasons, we do not have a good estimate for the sensitivity increase expected in Run 3, and any estimate we could derive would be sufficiently inaccurate as to not be useful. One expects that with a 40% gain in efficiency, however, that the sensitivity will increase. Given the delayed stream bandwidth is not being filled, there is no strong reason not to implement this trigger. This will also be an excellent learning opportunity to see how well the trigger works in data and how it impacts the sensitivity, providing useful information for $hh \to 4b$ trigger design at the HL-LHC.

8.4 Asymmetric Triggers at the HL-LHC

The situation for the HL-LHC is even more complicated than for Run 3. Due to the higher pileup (reaching around 200 simultaneous collisions per bunch crossing), the QCD backgrounds are much higher per event and jet thresholds have to be increased. In the baseline trigger upgrade plan, the 4j trigger with two b-tags would have a single p_T threshold at 65 GeV offline-equivalent at Level 1, up from 40 GeV in Run 2 (35 GeV in the HLT). This,

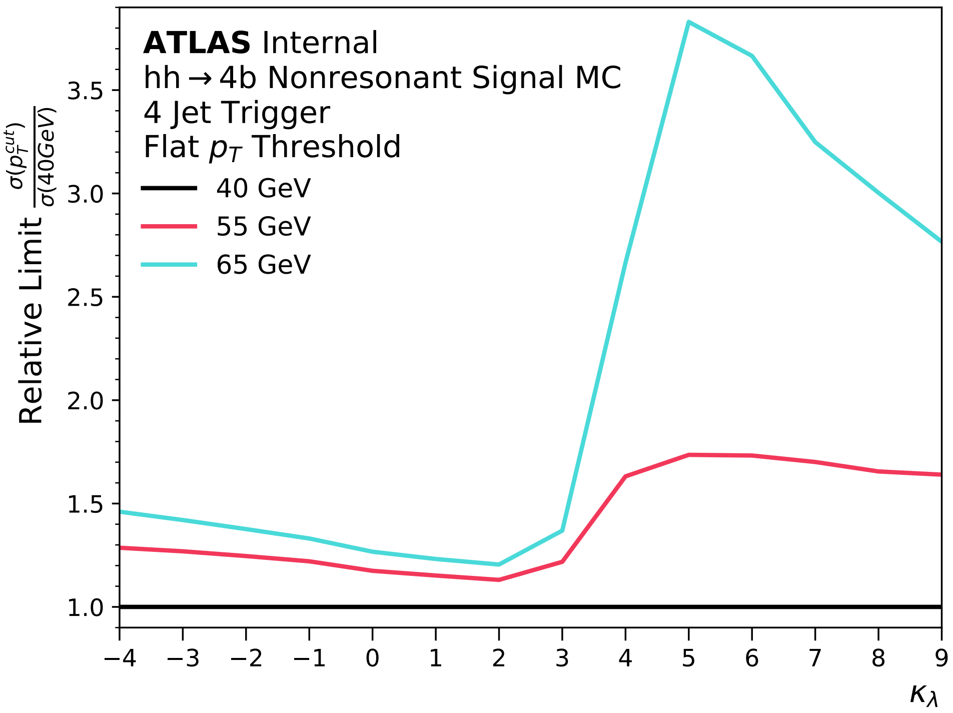


Figure 8.7: The change in analysis sensitivity for increased thresholds, assuming the baseline Run 2 analysis and systematics. The sensitivity loss is driven by signal acceptance loss.

however, would be a disaster for the $hh \to 4b$ analysis. The sensitivity of the Run 2 analysis with a 65 GeV threshold is shown in Figure 8.7. One can see that for the Standard Model, there is a 30-40% loss of sensitivity, while for $\kappa_\lambda = 5$, the sensitivity decreases by a factor of over 3.5.

One option to improve this situation was the proposed trigger upgrade L1Track. L1Track would use custom hardware to do fast regional tracking at Level 1. The additional information of the tracks would allow substantial improved background rejection and reduction of various trigger rates. The extra available bandwidth would then allow the 4 jet, 2 b-tag trigger threshold to be lowered to 55 GeV. The 55 GeV threshold is also shown in Figure 8.7, and one can see that we would still lose sensitivity at the level of 20-50%.

The 55/65 GeV thresholds, however, were determined before the asymmetric trigger idea became likely for Run 3. So an important question is how the asymmetric triggers perform at the HL-LHC. In particular, we are interested in designing two different asymmetric trigger

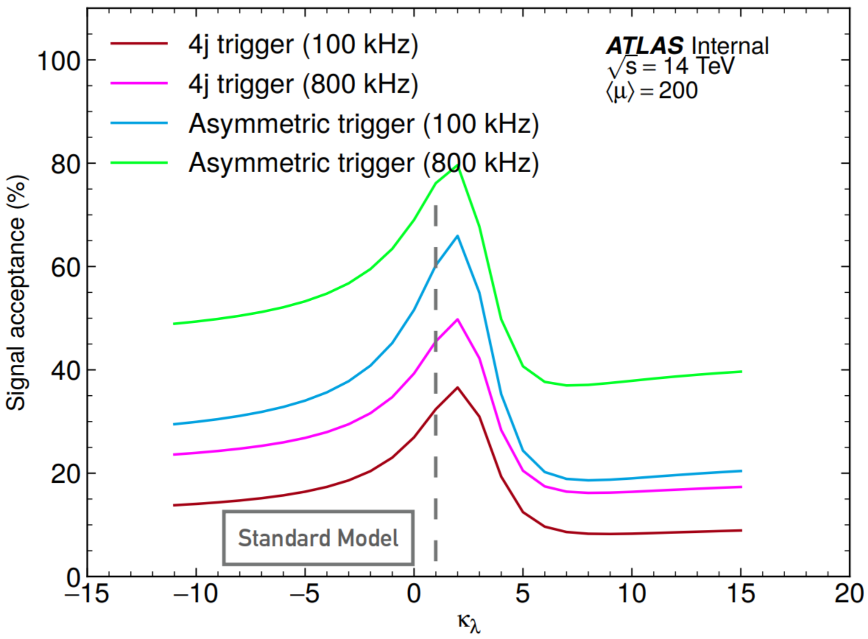


Figure 8.8: The efficiency of various proposed triggers for the HL-LHC.

thresholds, based on whether or not L1Track would be included to provide extra bandwidth. With L1Track, the target L1 trigger rate was 800 kHz, and without, the target was 100 kHz. In particular, it was important for the decision-making process to have an estimate of the final sensitivity changes, despite the difficulties of making such an estimate.

We first found a set of thresholds that matched the specified rates. The efficiencies are shown in Figure 8.8. We can see that switching from the flat 4 jet trigger at 100 kHz to an asymmetric trigger at 100 kHz doubles the signal efficiency. Then switching to the 800 kHz asymmetric trigger doubles the signal efficiency yet again.

The next step was to try to estimate the sensitivity. Since we do not have data at these low p_T's, we used a QCD simulation to estimate the background. While this will not be a precise sensitivity estimate, it should provide at least a reasonable qualitative estimate of the relative improvements. The rest of the analysis is the same as the Run 2 resonant analysis. The limits including only statistical uncertainties are shown in Figure 8.9. Our largest uncertainty is the shape uncertainty on the background, so limits with a one nuisance parameter shape uncertainty estimate are shown in Figure 8.10. In particular, we checked

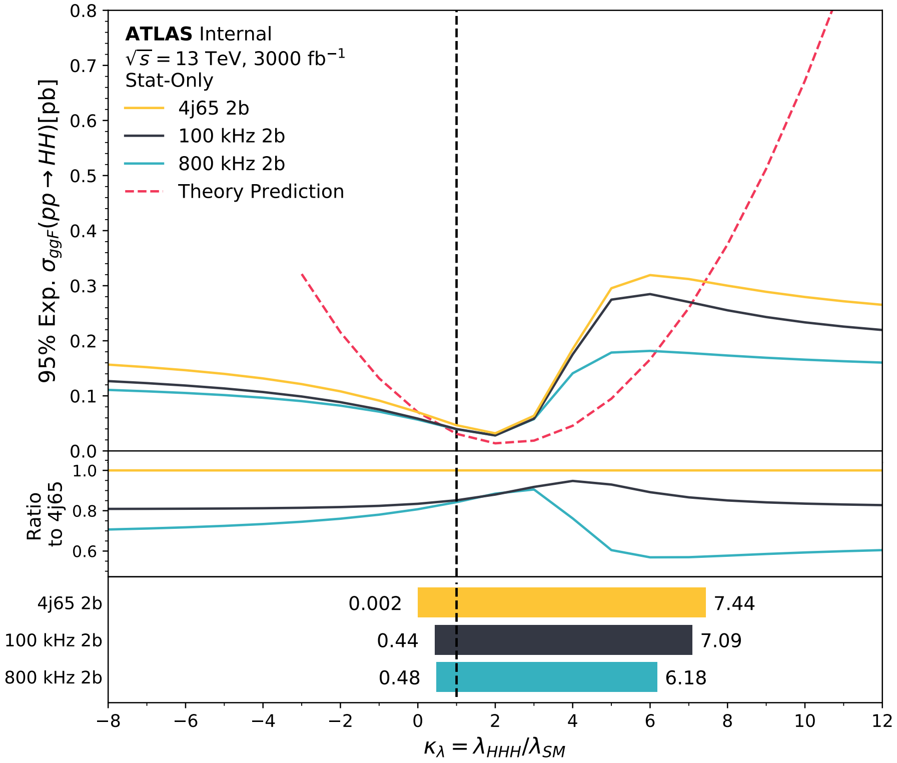
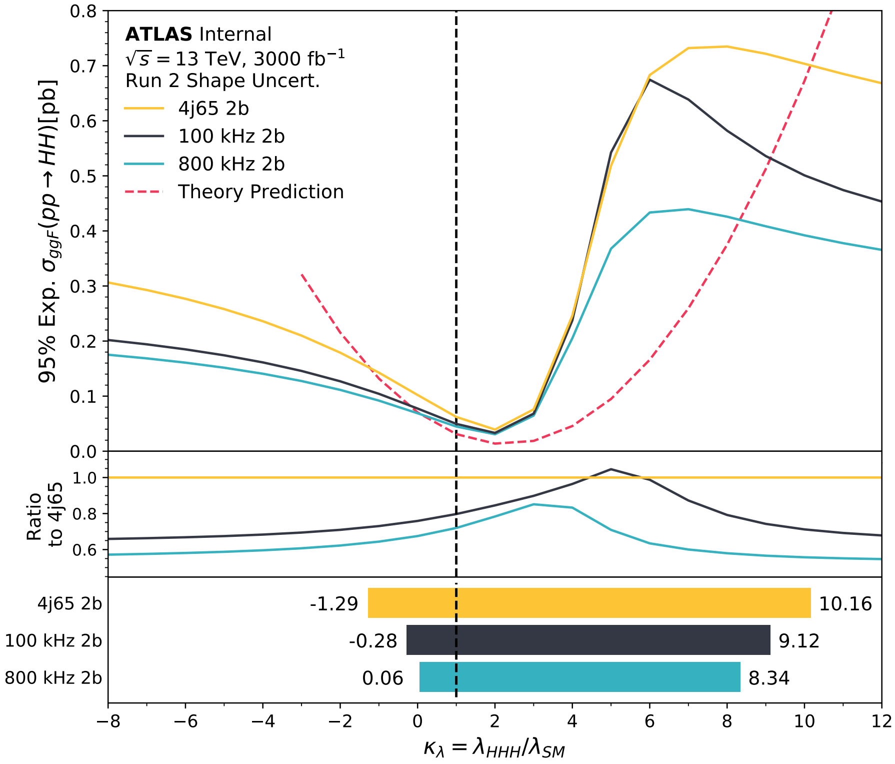


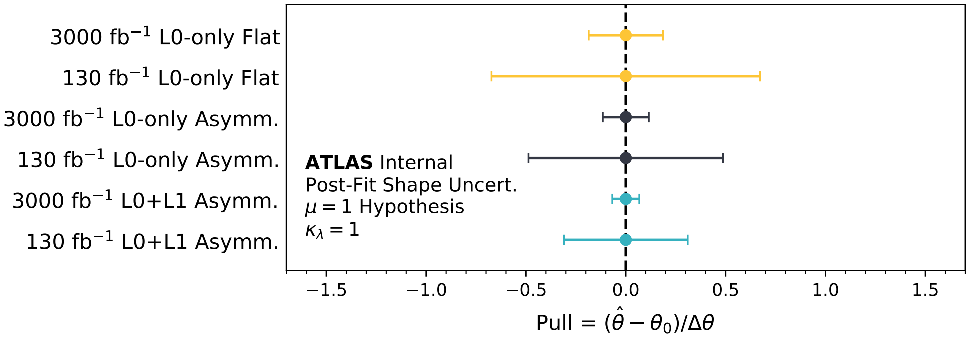
Figure 8.9: Stat-only limits for various trigger configurations at the HL-LHC. The yellow is the baseline no L1Track, flat four jet trigger. The black is the asymmetric trigger without L1Track, and the cyan is the asymmetric trigger with L1Track. The bottom panel shows the 95% exclusion range on κ_λ.

the nuisance parameter constraint at 3000 fb^{-1} for all trigger configurations, and indeed the nuisance parameter is heavily constrained. This means the actual limits will be somewhat worse than shown and that the estimate with systematics is not reliable, an expected behavior for such a difficult projection.

From both limits, we can see that the asymmetric triggers do indeed lead to significant improvements in the limits relative to the flat trigger, with the largest improvements being at varied κ_λ. Unfortunately, given these results in addition to other comparable analysis-specific studies, the trigger improvements from L1Track were deemed insufficient for the technical difficulties in implementation, particularly in the pixel readout electronics, and



(a)



(b)

Figure 8.10: (a) Limits on the cross section for various triggers, analogous to Figure 8.9, except with a single nuisance parameter for m_{HH} background shape uncertainty. (b) The post-fit nuisance parameter pulls when extrapolating to different quantities of data. The colors match the above triggers.

ATLAS has decided not to pursue the L1Track project. Thus the trigger bandwidth for the $hh \to 4b$ signal will be limited. Regardless, it does appear that the asymmetric trigger idea will be promising for the HL-LHC.

CHAPTER 9

FUTURE NEW IDEAS

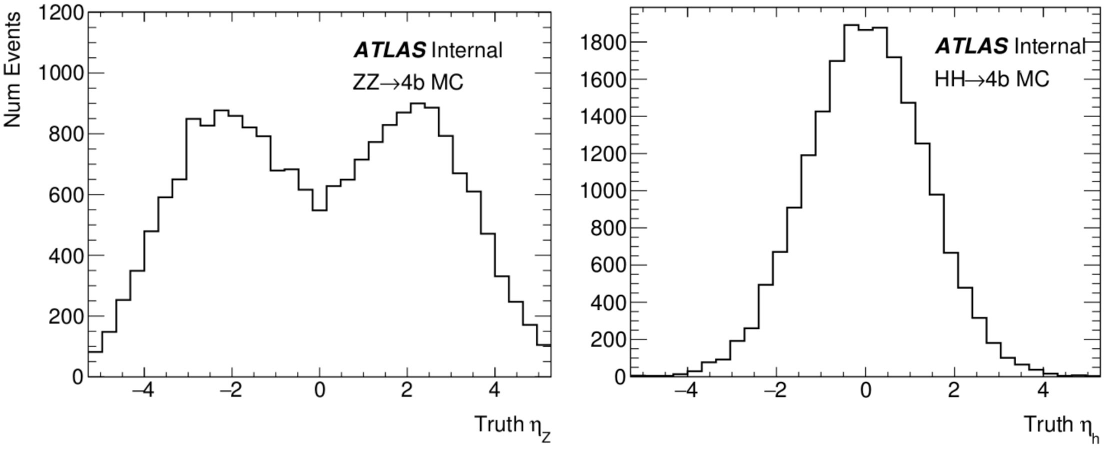
9.1 Alternative Signal Validation Regions

9.1.1 $ZZ \to 4b$

Given the challenges of the background estimate, it would be useful to have some way to validate the estimate with a real signal in data that we could unblind and observe before the Standard Model signal. One signal that in principle is very similar to $hh \to 4b$ is $ZZ \to 4b$. The cross-section of ZZ production is 350 times higher than that for standard model hh production, so while the $Z \to bb$ branching ratio is much smaller, the total cross section times branching ratio for $ZZ \to 4b$ is around 20 times that of $hh \to 4b$. Given that the last round of the analysis had an expected limit around $\mu = 20$, this seems like a promising validation region. One would simply define new signal, validation, and control regions around (m_Z, m_Z) in the massplane instead of around (m_h, m_h).

However, the ZZ signal is actually not much easier to observe than hh in the $4b$ final state because the ZZ production is extremely forward. Famously in diboson production, there is a small momentum transfer and the bosons tend to travel in the direction of the beam, with the probability peaking in the forward region. The distribution of pseudorapidity of the Z bosons is shown in Figure 9.1a. Compare this to pseudorapidity of the Higgs bosons in di-Higgs production in Figure 9.1b.

The problem with the forward Z bosons is twofold. First, the jets will tend to be outside the acceptance of the tracker, which reaches $|\eta| < 2.5$. Without the tracker, running b-tagging is impossible, so it is not possible to b-tag the jets. All ZZ events with a single b-jet outside the tracker will be very difficult to reconstruct. However, at the HL-LHC, the tracker will extend much further in pseudorapdity, and this problem could potentially be mitigated somewhat. The second issue is that the transverse momentum of the jets will be



(a) ZZ

(b) hh

Figure 9.1: The distribution of the truth boson pseudorapidity, η, for $pp \to ZZ$ and $pp \to hh$.

low. Already in the standard model hh production, accepting the lowest p_T jet in the trigger is challenging. The ZZ process is significantly softer in p_T, as shown by the histogram in Figure 9.2. The baseline four b-jet trigger in Run 2 has a threshold of 40 GeV, which will clearly cut out a large amount of signal. In total, requiring four b-jets with $|\eta| < 2.5$ and $p_T > 40$ GeV for all four jets has a selection efficiency around 0.8%, compared to a selection efficiency of 13% for the same cuts on $hh \to 4b$. Thus the total acceptance times branching ratio times cross section is similar between $ZZ \to 4b$ and $hh \to 4b$.

Not only is the signal selection efficiency low for ZZ, the backgrounds will also be higher. Because the jets tend to be forward, there will be higher QCD backgrounds, since QCD peaks in the forward region. We would have to adjust the selection $|\Delta\eta_{ZZ}| < 1.5$ that we are able to make with the di-Higgs process, which removes a substantial amount of QCD background. Additionally, the reconstructed masses will tend to be lower, which will also be associated with higher backgrounds, since the backgrounds in the massplane peak in the lower left corner, at the lowest masses. Thus while it may be possible to measure $ZZ \to 4b$, it would likely have to diverge from the hh analysis, limiting its use as a validation region in the first place.

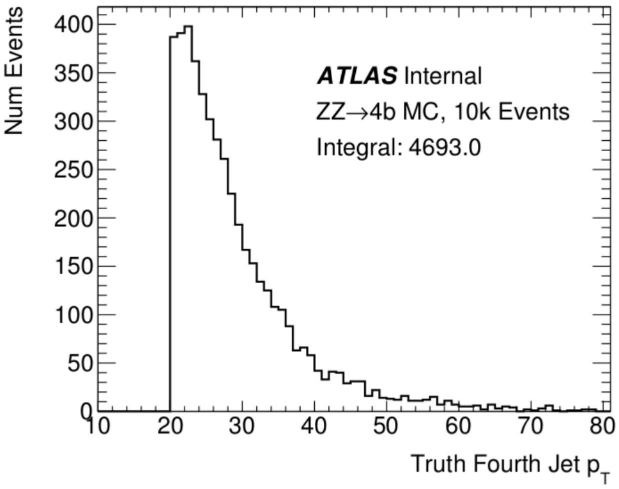


Figure 9.2: The p_T of the softest of the four b-jets from the $ZZ \rightarrow 4b$ decay, with a cut for $p_T > 20$ GeV.

9.1.2 $ZH \rightarrow 4b$

On the other hand, $ZH \rightarrow 4b$ may be a useful validation region soon. While the cross-section is lower than ZZ, it benefits from a higher branching ratio to $4b$ and because it is an s-channel process, will tend to be central. The total cross-section times branching ratio for $ZH \rightarrow 4b$ is around 7 times that of $hh \rightarrow 4b$, and since the kinematics should be similar, it should be observable around when the standard model search is sensitive to $\mu \approx 7$. The backgrounds will be somewhat higher because of the lower Z mass, but it should be possible to define off-diagonal signal, validation, and control regions and run the analysis in that region. Given the projected sensitivities, this validation region could even be useful now, though it has not been pursued. It should likely be pursued however for the Run 3 analysis to provide a critical check of the analysis.

9.2 Untested Background Estimation Ideas

9.2.1 Neural Network Perturbation

Let $\mathbf{x} = \{m_{h1}, m_{h2}, m_{hh}\}$ and let $\alpha = \{\alpha_1, \alpha_2, ...\}$ be some generic set of parameters. Then our 2b data can be described by a generic parameterized probability distribution $P_2 = f(\mathbf{x}; \alpha_2)$ for a particular parameter set α_2 of a model f. Similarly, our 4b data can be described by $P_4 = f(\mathbf{x}; \alpha_4)$.

A fundamental hypothesis in our analysis is that these distributions are similar to each other, so that 2b can be used to model 4b. In this language, that could be stated as $\Delta\alpha \equiv \alpha_4 - \alpha_2 \ll 1$. We can in principle therefore Taylor expand

$$P_4 = f(\mathbf{x}; \alpha_2 + \Delta\alpha) = f(\mathbf{x}; \alpha_2) + \Delta\alpha \cdot \nabla_\alpha f(\mathbf{x}; \alpha)|_{\alpha=\alpha_2} \tag{9.1}$$

Suppose now that f is some neural network regression model with inputs $\mathbf{x}$, parameters α, and output P. Because of the universal approximation theorem, this neural network can approximate the underlying distribution arbitrarily well. We can train our neural network on 2b data to get the distribution $P_2^{fit} = f(\mathbf{x}; \alpha_2)$. If the 2b and 4b data are not too different, then we can apply the Taylor expansion to get

$$P_4 - P_2^{fit} = \Delta\alpha \cdot \nabla_\alpha f(\mathbf{x}; \alpha)|_{\alpha=\alpha_2} \tag{9.2}$$

This is a linear system of equations with len($\Delta\alpha$)=#params unknowns. The number of equations is equal to the number of points $\mathbf{x}$ in the probability distributions that we sample. In principle, by using small enough bins in the histogram, or by using a smoothing algorithm like kernel density estimation, this number can be made arbitrarily large. We therefore in general have an overconstrained system of equations.

Let ΔP be the column vector of $P_4 - P_2^{fit}$ where the row index is the point $\mathbf{x}$, and let $D = \nabla_\alpha f(\mathbf{x}; \alpha)|_{\alpha=\alpha_2}$ be a matrix where the rows are the different points $\mathbf{x}$ and the columns are the components of the gradient.

The least-squares best-fit solution for $\Delta\alpha$ will therefore be the standard form

$$\Delta\alpha = (D^T D)^{-1} D^T \cdot \Delta P \tag{9.3}$$

which can be computed relatively easily with standard neural network libraries.

We therefore have an analytical expression for the best fit for our 4b data based on a small parametric perturbation from our 2b data.

9.2.2 Mass Plane Transfer Functions

The central hypothesis of the nominal background estimation is that the 2b and 4b data are similar. Traditionally, we interpret this to mean that $P_{2b}(m_{h1}, m_{h2}, m_{hh}) \approx P_{4b}(m_{h1}, m_{h2}, m_{hh})$. However, we can actually loosen this requirement somewhat. Really what we need is that the transfer function from one region of the massplane to another is similar between 2b and 4b. It is acceptable for 2b and 4b to be different if they scale the same way between the control and signal regions.

This motivates the following definition of the transfer function. Suppose regions of the massplane are index by $\vec{x}$, which may for example be histogram bins. The number of events in region $\vec{x}$ is $N(\vec{x})$. Then given the number of events in region $\vec{x}$, we would like to estimate the number of events in a different region $\vec{y}$. We therefore define the transfer function

$$t(\vec{x}, \vec{y}) = \frac{N(\vec{y})}{N(\vec{x})} \tag{9.4}$$

Thus to predict the number of events in region $\vec{y}$ given the events in region $\vec{x}$, one can just multiply by the transfer function.

Now, one way we can phrase the similarity between 2b and 4b is that the transfer functions should be similar. We can define the difference function $s(\vec{x}, \vec{y})$ to be the difference in transfer functions,

$$s(\vec{x}, \vec{y}) = t^{4b}(\vec{x}, \vec{y}) - t^{2b}(\vec{x}, \vec{y}) \tag{9.5}$$

One could instead define the ratio of transfer functions, though the following logic will be similar.

Thus, to predict the unknown number of 4b events in region $\vec{y}$, given the number of events in a known region $\vec{x}$

$$N^{4b}(\vec{y}) = [t^{2b}(\vec{x}, \vec{y}) + s(\vec{x}, \vec{y})]N^{4b}(\vec{x}) \tag{9.6}$$

Note that this estimate depends on the 2b transfer function, which is known, and the difference function, which is unknown when $\vec{y}$ is in the signal region.

However, if 2b and 4b are similar, then the difference function should be small, and in particular, it should be slowly varying. This means it should be easy to interpolate over the signal region, such as via Gaussian processes or a simple spline fit. One might imagine defining polar coordinates (r, θ) with respect to the center of the signal region, and using a Fourier series to fit to θ, where the high frequency components can be neglected.

Note that it is not too hard to show that $s(\vec{x}, \vec{y}) = 0$ if and only if $N^{2b} = N^{4b}$, as we would hope. The method of transfer functions merely provides a rephrasing of the problem to a function that may be easier to fit. This fit will incorporate quantities that we currently ignore, such as that adjacent bins will tend to have similar counts and that there are correlations between different bins in the control region, correlations which may extrapolate into the signal region and may improve the fit.

9.3 Optimal Variable Construction

Generically, the distribution of our four jets is given by probability density function over a 16-dimensional space, $P(\boldsymbol{X})$, where $\boldsymbol{X}$ represents a 16-dimensional coordinate. One important basis of this function is the four-momenta of all four of the b-jets,

$$\begin{aligned} P(\boldsymbol{X}) &= P(p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu) \\ &= P(E_1, p_1^x, p_1^y, p_1^z, E_2, p_2^x, p_2^y, p_2^z, E_3, p_3^x, p_3^y, p_3^z, E_4, p_4^x, p_4^y, p_4^z) \end{aligned} \tag{9.7}$$

Working directly with this distribution is unwieldy, and we would like to reduce the dimensionality. Fortunately, we can physically motivate the construction of several independent variables that should not depend on the particular underlying physics process being probed.

First, the system should be rotation-invariant for rotations in ϕ around the beam axis. Thus we can pick any single azimuthal variable, and factor it out as a uniform distribution. I will arbitrarily pick ϕ_1, one of the jet azimuthal variables, $\tan\phi_1 = \frac{p_1^y}{p_1^x}$. Which ϕ is selected does not matter, and there are many other options. Thus our full 16-dimensional probability distribution becomes the product of a 15-dimensional distribution and a 1 dimensional uniform distribution.

Next, we know that the mass of jets is determined by the physics of hadronization and not by any hard inelastic interactions that we want to probe, assuming that the majority of events with four b-tags are composed of four real b-jets. Thus the mass of each jet should follow a distribution governed entirely by soft QCD, which I denote $\mathrm{JetMass}(m)$, and the jet masses should be independent of each other and the overall event kinematics. Thus we can factor out of the joint probability a factor of the form $\prod_{i=1}^{4} \mathrm{JetMass}(m_i)$. This takes leaves us with an 11-dimensional distribution.

Next, the total z-momentum of the collision should be given by the proton parton density function and should be process-independent, assuming that throughout the massplane, the same relative fraction of different process is constant (which may be an aggressive assumption). Thus we can factor out $\mathrm{PDF}(p_{tot}^z)$, leaving us with 10 variables. Note that technically the PDF should be convolved with the detector resolution, but we will ignore this for now, defining it to be part of the $\mathrm{PDF}(p_{tot}^z)$ function.

One might be tempted to construct the x and y transverse momenta as well, motivated by the conservation of momentum. However, because we are considering only the four b-tagged jets, additional jets in the event can give a transverse boost. In particular, the presence of the extra jets is process-dependent and could potentially be used as a useful discriminator. However, we know that orthogonal to the direction of the transverse boost, there will always be a component of momentum that is conserved. The sum of the momenta orthogonal to the net transverse boost I will call $p_T^{\text{orth-boost}}$, and this variable should be distributed as a Gaussian with mean 0 and width determined by the detector resolution. This reduces the problem to 9 remaining variables.

Thus, so far, we have

$$\begin{aligned} P(\boldsymbol{X}) = f(x_1, ..., x_9) \times & \prod_{i=1}^{4} \text{JetMass}(m_i) \times \text{Uniform}(\phi_1; [0, 2\pi]) \\ & \times \text{PDF}(p_{\text{tot}}^z) \times \text{Normal}(p_T^{\text{orth-boost}}; 0, \sigma_{\text{det}}) \end{aligned} \tag{9.8}$$

for some unknown function f and some variables $\{x_i\}$ that will depend on the goal.

Now, how do we decide the remaining variables $\{x_i\}$, which will generally be strongly correlated and physics dependent? First, the $\{x_i\}$ must generally be a function of the four-momenta, $(x_1, ..., x_9) = h(p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu)$ for some multi-output function h. If we want to preserve information and avoid losing a dimension, we want to be able to compute all four four-momenta from a combination of the x_i and the already factored variables. That is, we want there to be some function g such that

$$\begin{aligned} (p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu) &= g(x_1, ..., x_9, m_1, ..., m_4, \phi_1, p_{\text{tot}}^z, p_T^{\text{orth-boost}}) \\ &= g(h(p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu), m_1, ..., m_4, \phi_1, p_{\text{tot}}^z, p_T^{\text{orth-boost}}) \end{aligned} \tag{9.9}$$

This requirement is a generalization of an invertability requirement and significantly constrains the form of h.

There are other constraints as well. The most restrictive constraint is that $h(p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu)$ should be invariant under permutations of the four-momenta, an invariance under the permutation group S_4. This leverages the fact that there is no natural ordering to the jets. In statistical language, the inputs to $P(p_1^\mu, p_2^\mu, p_3^\mu, p_4^\mu)$ are *exchangable*, which is similar to though looser than than independence. Functions that are invariant under the order of the inputs are, for example, the invariant mass of the four jets, $m_{4b} = -(\sum_{i=1}^4 p_i^\mu)^2$, and the maximum transverse momentum, $\max(p_T^1, p_T^2, p_T^3, p_T^4)$. Note that algorithms that first sort the jets, then execute some order-dependent function are valid because any individual sorting algorithm will have the same output regardless of the input order of the jets.

There is also a discrete $\mathbb{Z}_2$ symmetry. Both QCD and the SM processes are parity-invariant. That is, $P(\boldsymbol{X})$ should be invariant under the negation of all of the three-momenta, $(E_i, \vec{p}_i) \mapsto (E_i, -\vec{p}_i)$, as long as the weak interactions are negligible. Note that the collision system is invariant under a reflection of $\hat{z} \mapsto -\hat{z}$, but this is the same as a parity switch combined with a rotation about ϕ by π. The function h should ideally preserve parity and ϕ symmetry, though it may be possible to have violations cancel via correlations in the joint probability. It is not clear to me whether this is possible.

Some of the $\{x_i\}$ we can choose either through the desire for simplicity or for analysis precedent. For example, we expect that the reconstructed Higgs masses m_{h1}, m_{h2} and the total four-jet invariant mass, $m_{4b} = m_{hh}$ are powerful variables, both for background rejection and for background modeling. We also might want to explicitly compute the total transverse momentum boost. Unlike the other variables discussed, however, we expect correlations between these variables themselves, and between these variables and the remaining $\{x_i\}$. Thus while we can factor these variables, the remaining distribution must be conditioned on them. For example, we can write

$$f(x_1, ..., x_9) = f^*(u_1, ..., u_6 | m_{h1}, m_{h2}, m_{hh}) \times P(m_{h1}, m_{h2}, m_{hh}) \tag{9.10}$$

for the unconditional (marginal) distribution $P(m_{h1}, m_{h2}, m_{hh})$ and some new function f^* over potentially new variables $\{u_i\}$ that must be conditioned on m_{h1}, m_{h2}, and m_{hh}.

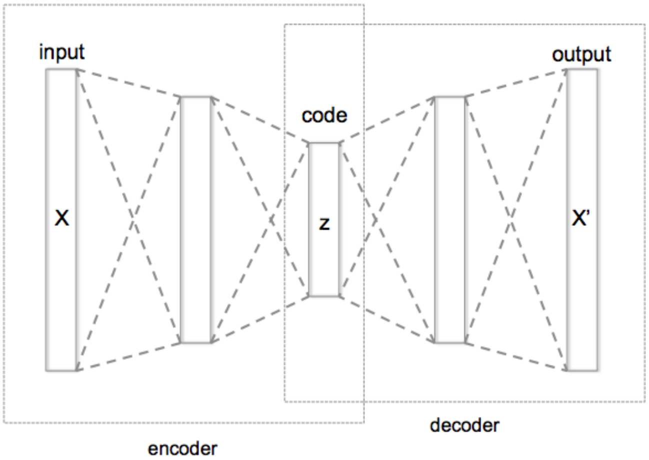


Figure 9.3: The schematic structure of an autoencoder, from Wikipedia [13]. The learned representation is z, with the goal to learn $X' = X$.

Whether we factorize these variables or not, there are some useful things we can do with this framing of the problem. For example, suppose we'd like to find the variables $\{x_i\}$ such that the information of the four-momenta are encoded as accurately as possible. We ought to be able to do this using an auto-encoder structure, for example. The basic idea of an auto-encoder is that it takes inputs, the p_i^μ in this case, passes them through a deep neural network, and attempts to predict exactly the same outputs, p_i^μ. It tries to learn the identity function. The key is that at some point in the network, we can create a bottleneck, a layer of limited width. This forces the network to encode all of the initial information in the bottleneck. This is shown schematically in Figure 9.3

In our case, we could build an auto-encoder that is basically two neural networks. They could be feed-forward networks, for example. The first takes the 16 p_i^μ inputs and has 9 outputs x_i. The second network would then take those 9 x_i in addition to the variables that we have already factored out, $m_1, ..., m_4, \phi_1, p^z_{\text{tot}}, p_T^{\text{orth-boost}}$, and attempt to output the original 16 p_i^μ. In the language of Equation 9.9, the first network learns the function h and the second network learns the function g. In this architecture, the first network, the encoder, must learn information independent from the already factored variables, and given the dimension of the subspace, it should need at least 9 outputs to do this. Despite this, there may also be some nontrivial dependencies among the x_i such that the neural network compresses the information into even fewer outputs, as is the case with more standard usage of auto-encoders. This would be interesting to check.

One problem with a feed-forward based architecture for the auto-encoder is that it is not inherently symmetric under interchange of the input momenta. We can circumvent this by first sorting the jets by p_T or other variables, but it would be nice if the permutation invariance were built into the network. This may be possible to do using an architecture based on the idea of Deep Sets, though I am inexperienced on this front [42]. The network architecture also will not respect the parity and rotation symmetries. Perhaps it is possible to design the network such that these are still respected, and this would be an interesting challenge.

A feature of the auto-encoder approach is that we know the learned $\{x_i\}$ will not be unique. Because we use nonlinear transformations, there will be infinitely many possible new sets of basis variables. The minimum of the loss function will be a wide valley, leaving a lot of flexibility. However, we can leverage this freedom to optimize over additional quantities. Suppose we want to do signal to background discrimination, where we ultimately want some output score between 0 and 1 such that events near 1 are likely signal and events near 0 are likely background. Because the 7 variables we have already factored out should be the same between signal and data, one can imagine doing background discrimination with only the $\{x_i\}$. We could feed the $\{x_i\}$ into some sort of binary classification algorithm, such as a simple feed forward network or something more complex. Then we could train the classifier such that the loss function of the auto-encoder is still minimized. For example, we could use a standard entropy loss function for classification, but with an additional term that penalizes straying too far from the autoencoder loss function minimum. Such a constraint term would have a parameter λ controlling the relative importance of the autoencoder that would need to be tuned empirically. This proposed architecture is summarized in Figure 9.4.

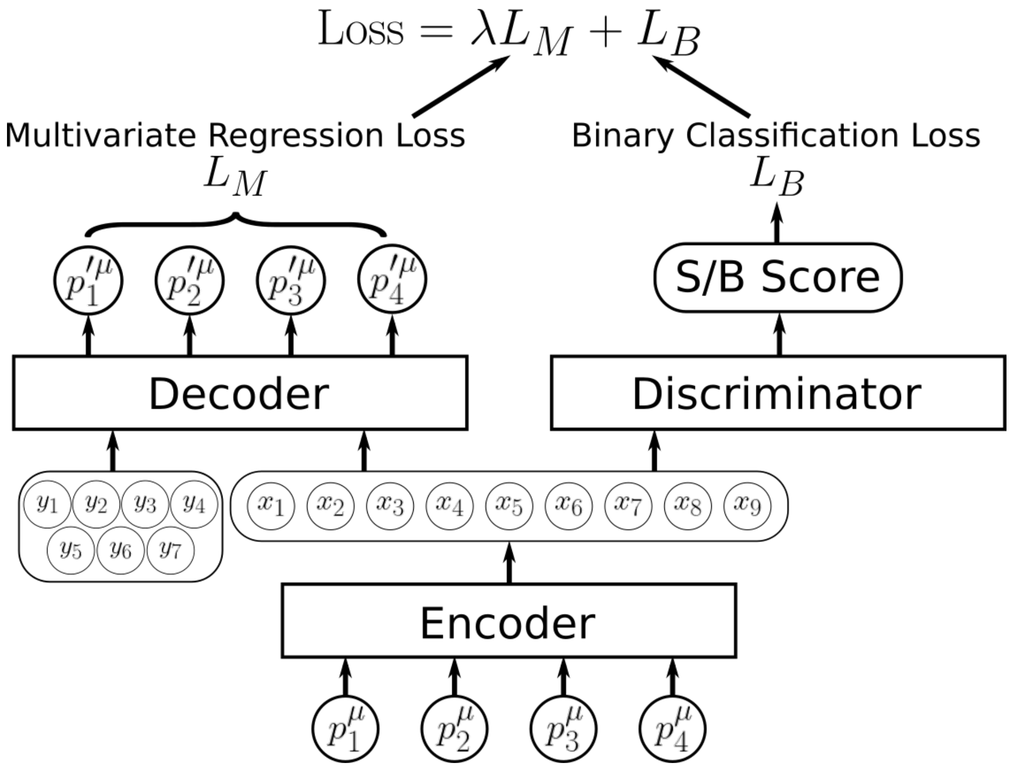

$\mathrm{Loss} = \lambda L_M + L_B$
Multivariate Regression Loss
L_M
Binary Classification Loss
L_B
$p_1'^\mu$
$p_2'^\mu$
$p_3'^\mu$
$p_4'^\mu$
S/B Score
Decoder
Discriminator
y_1 y_2 y_3 y_4 y_5 y_6 y_7
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9
Encoder
p_1^μ
p_2^μ
p_3^μ
p_4^μ

Figure 9.4: The architecture of the proposed network to find optimal variables $\{x_i\}$ given the goals to preserve all information and also use them for signal to background discrimination. There may be other more sophisticated loss functions that enable training both of these goals. The $\{y_i\}$ are the 7 variables that should be independent of the physics we are trying to probe, $\{y_i\} = (m_1, m_2, m_3, m_4, \phi_1, p^z_{\mathrm{tot}}, p_T^{\text{orth-boost}})$. The exact choice of architecture for the encoder, decoder, and discriminator is somewhat flexible.

While having a discriminator is nice, this network also outputs variables $\{x_i\}$ that should encode physically meaningful information. In particular, the learned $\{x_i\}$ should be a set that is a sensitive signal to background discriminant. It would be interesting to see if the network learned to compute m_{hh}, for example, since this variable has a large impact on the event kinematics. Note that one could connect the autoencoder to other networks targeting $\{x_i\}$ with different nice features, such as those that are best for background modeling. Our current background estimate uses an ultimately arbitrary set of variables, but it may be possible to instead train it with the auto-encoder to learn the best possible set of variables for the background estimate.

CHAPTER 10

POSTSCRIPT

This analysis is hard. The backgrounds are large and difficult to model. The signal is small and challenging to reconstruct. Before the LHC powered on, few expected the $b\bar{b}b\bar{b}$ to make a significant contribution to the studies of the Standard Model Higgs self-coupling.

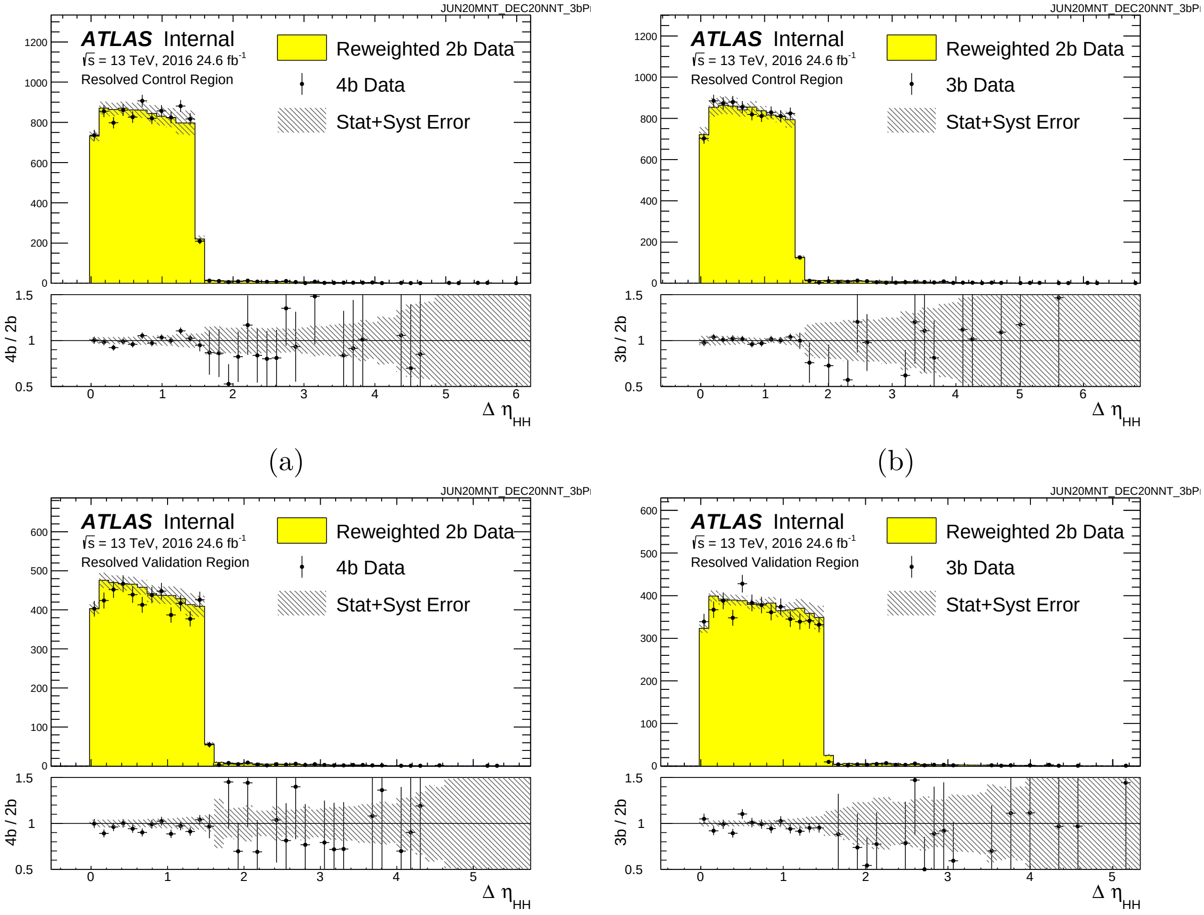
Yet, these challenges are exactly what make this analysis attractive to so many, with the size of the analysis group swelling in recent years. There has been a constant stream of new collaborators with creative new ideas, often utilizing state-of-the-art concepts in data analysis. It is for this reason that the $b\bar{b}b\bar{b}$ channel has managed to stay competitive with the relatively much more straightforward $bb\gamma\gamma$ and $bb\tau\tau$ analyses.

The expectation is that eventually, this analysis will gradually begin to lag significantly in sensitivity behind these other analyses. However, there is still an enormous amount of room for optimization. It seems to me that each round of the analysis uses the cutting-edge ideas from the previous round. The 36 fb^{-1} analysis did not use a neural network for the background prediction, though the idea had been considered and was suspected to be powerful. Right now, we have many such ideas that remain unused but that may lead to potential improvements going forward. It remains to be seen which are the best.

Even though I am leaving the field of physics, I am looking forward to watching how far this analysis will be pushed over the next 10 or 15 years. I also look forward to the day that the LHC experiments announce a 5σ observation of Standard Model HH production (or an observation of non-Standard Model HH production!) Current projections place a 5σ observation as only marginally possible, even with combined ATLAS and CMS data. However, seeing all of the improvements that have been made and that are planned across the di-Higgs group, I have faith that we will comfortably reach a 5σ observation by the end of the HL-LHC and make a good measurement of the Higgs self-coupling.

APPENDIX

Background Estimate Plots



(a)

(b)

(c)

(d)

Figure A.1: Distributions of $\Delta\eta_{HH}$ in 2016 4b and 3b data for the control and validation regions, comapred to the 2b-derived background estimate.

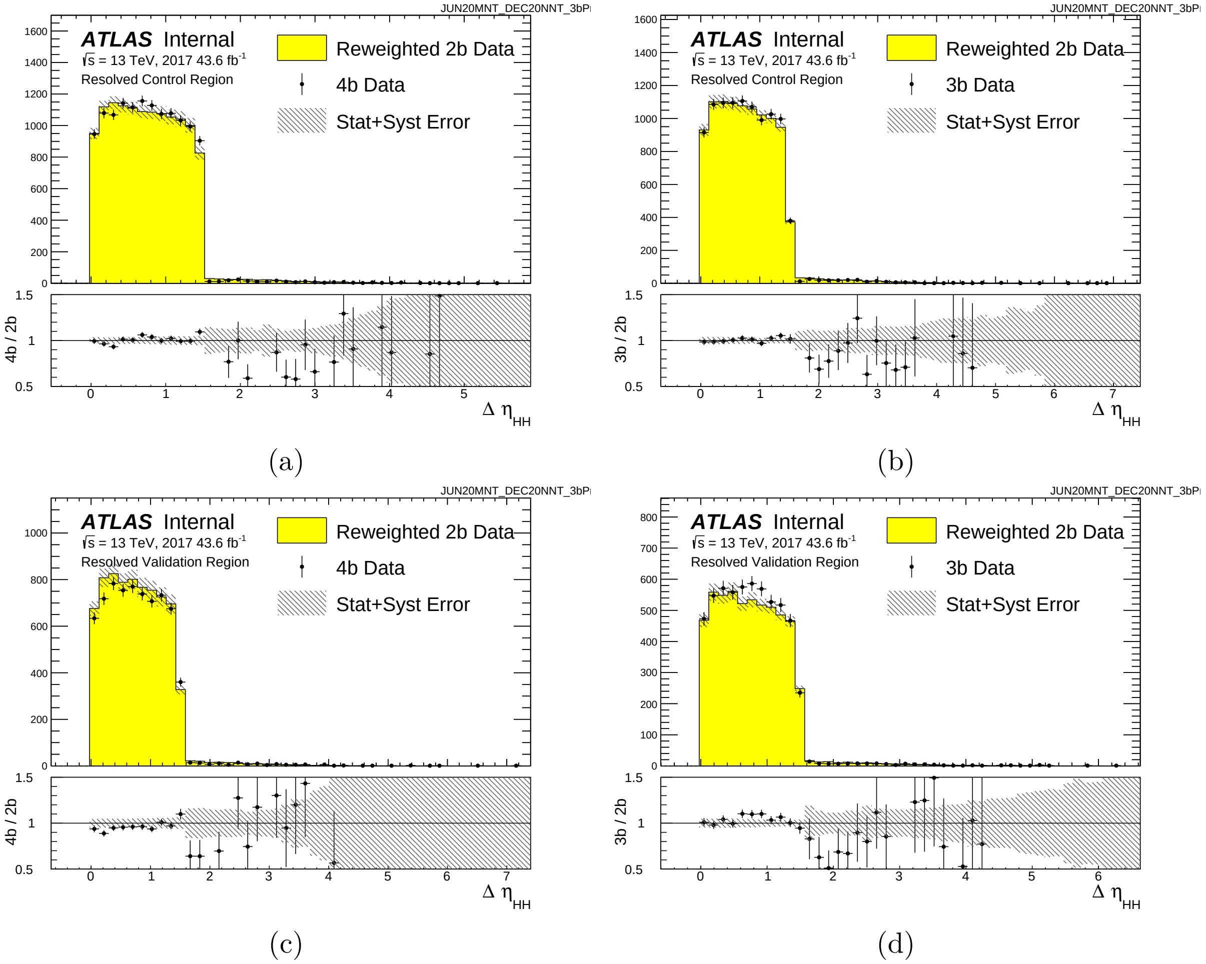


Figure A.2: Distributions of $\Delta\eta_{HH}$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

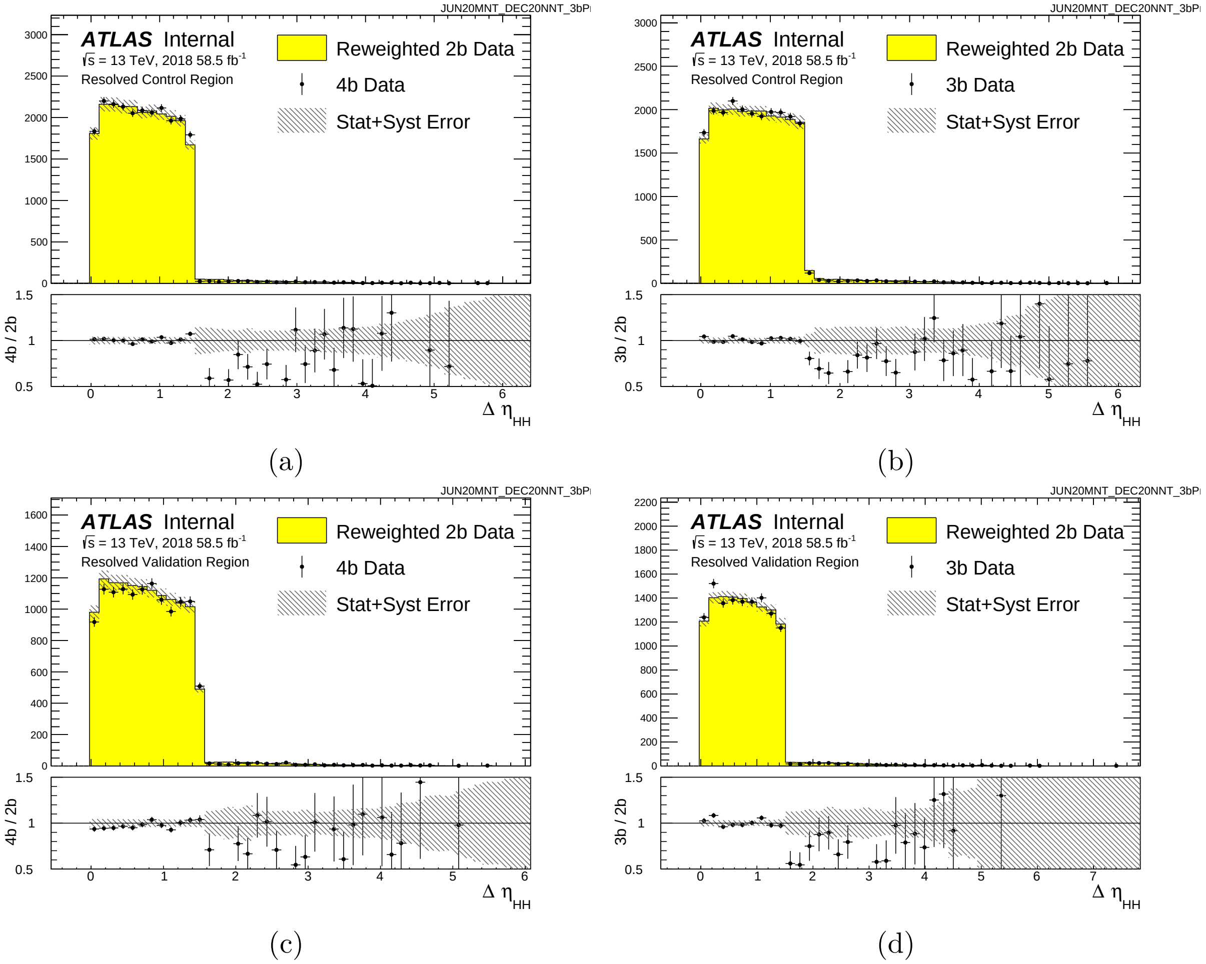


Figure A.3: Distributions of $\Delta\eta_{HH}$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

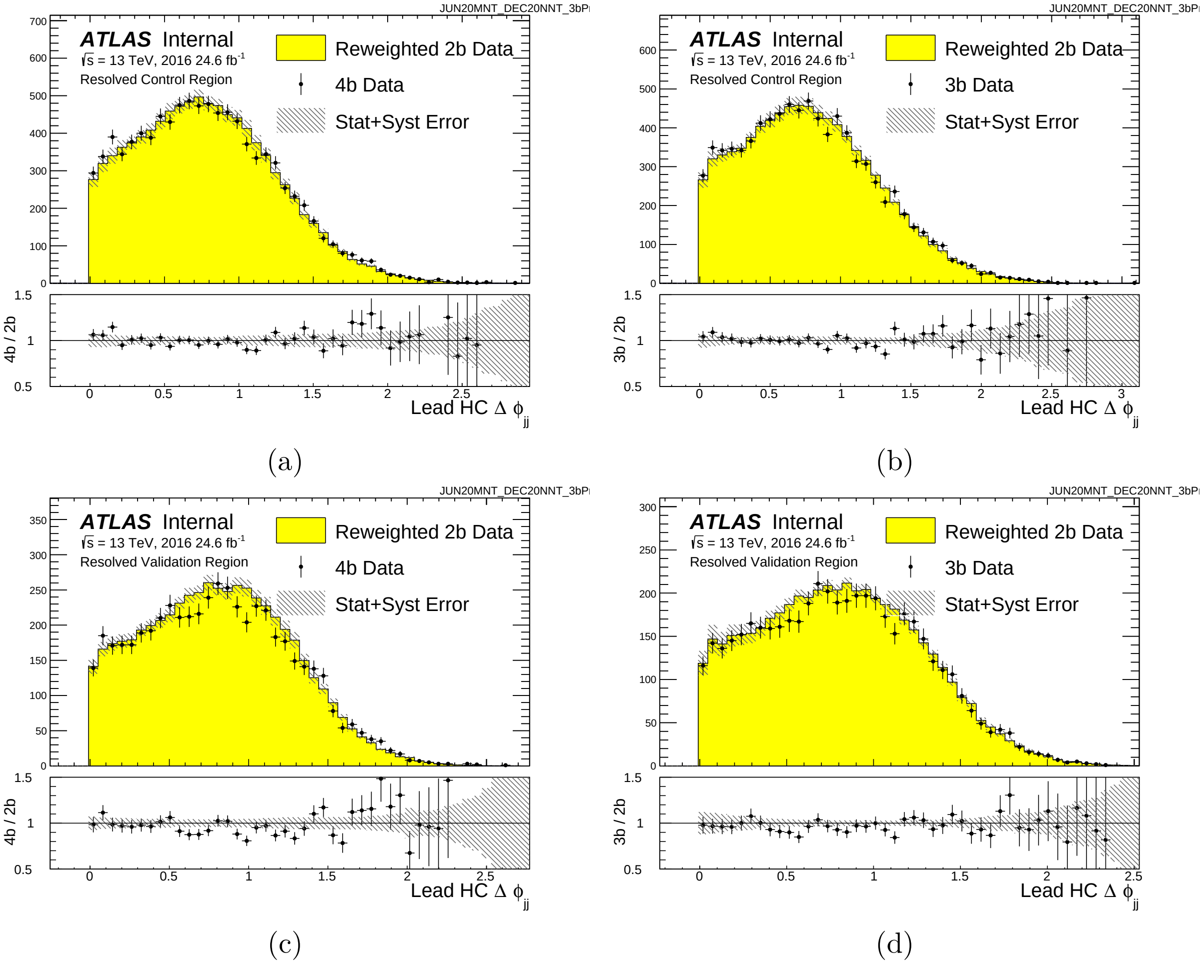
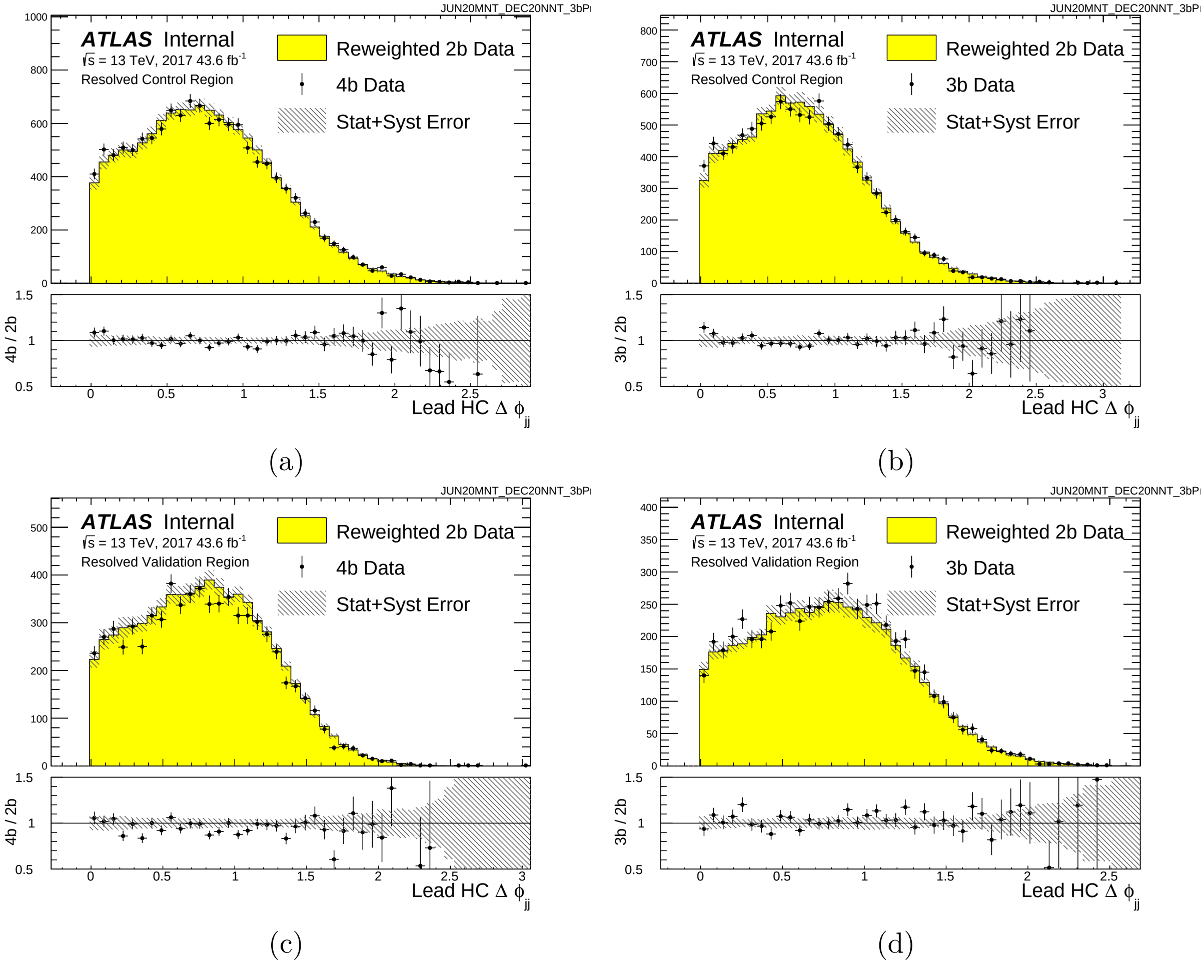


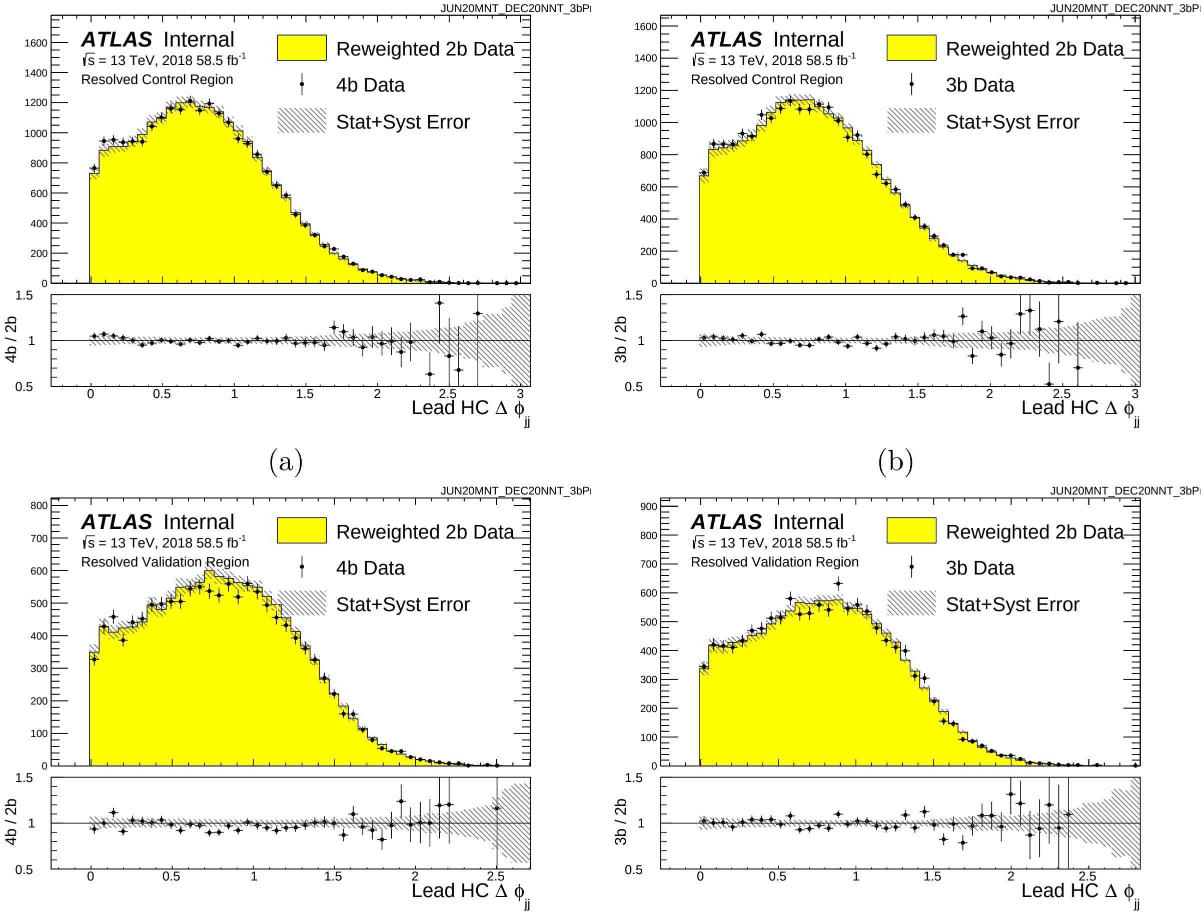
Figure A.4: Distributions of the leading Higgs candidate $\Delta\phi_{jj}$ in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b)

(c) (d)

Figure A.5: Distributions of the leading Higgs candidate $\Delta\phi_{jj}$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



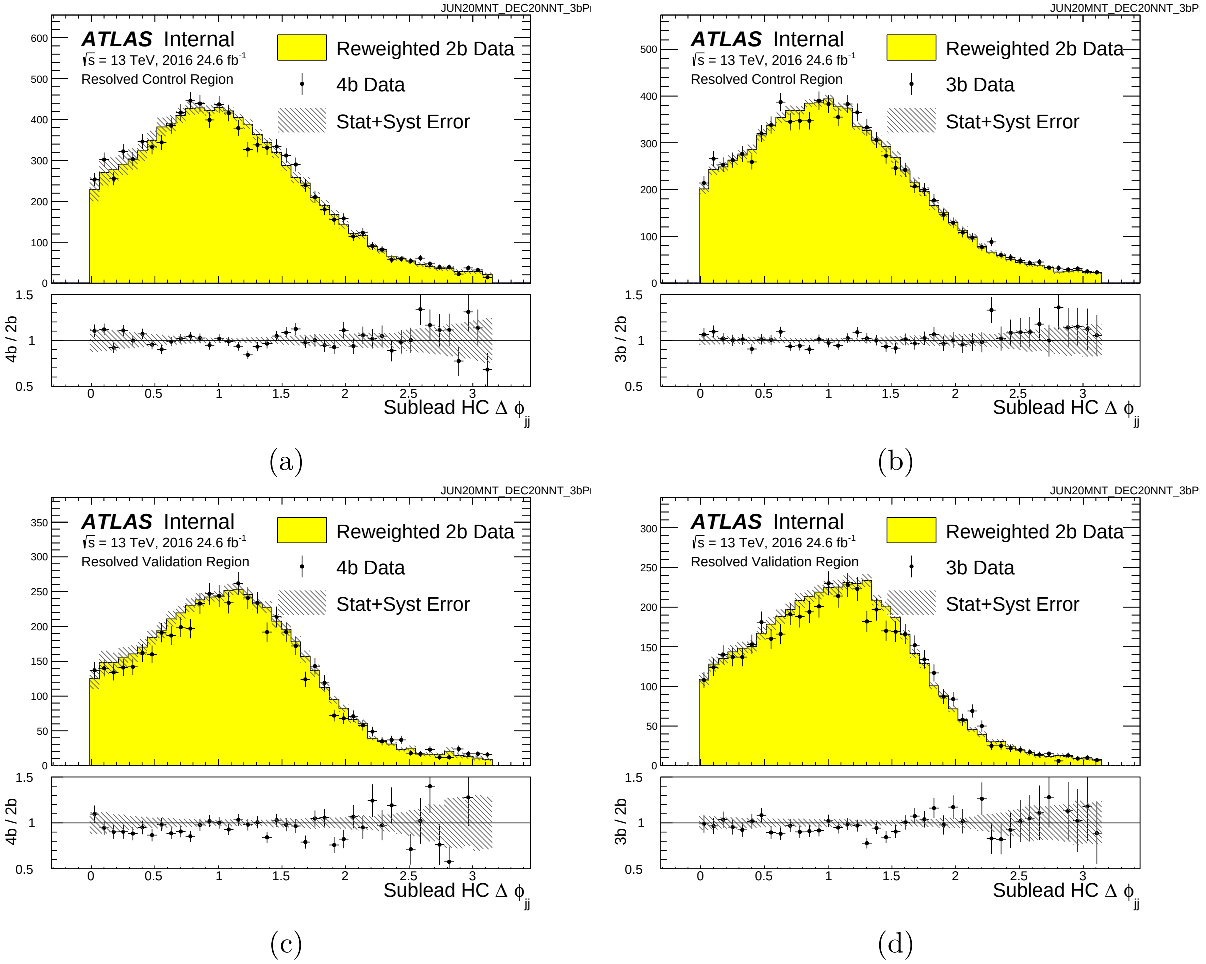
(a)

(b)

(c)

(d)

Figure A.6: Distributions of the leading Higgs candidate $\Delta\phi_{jj}$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



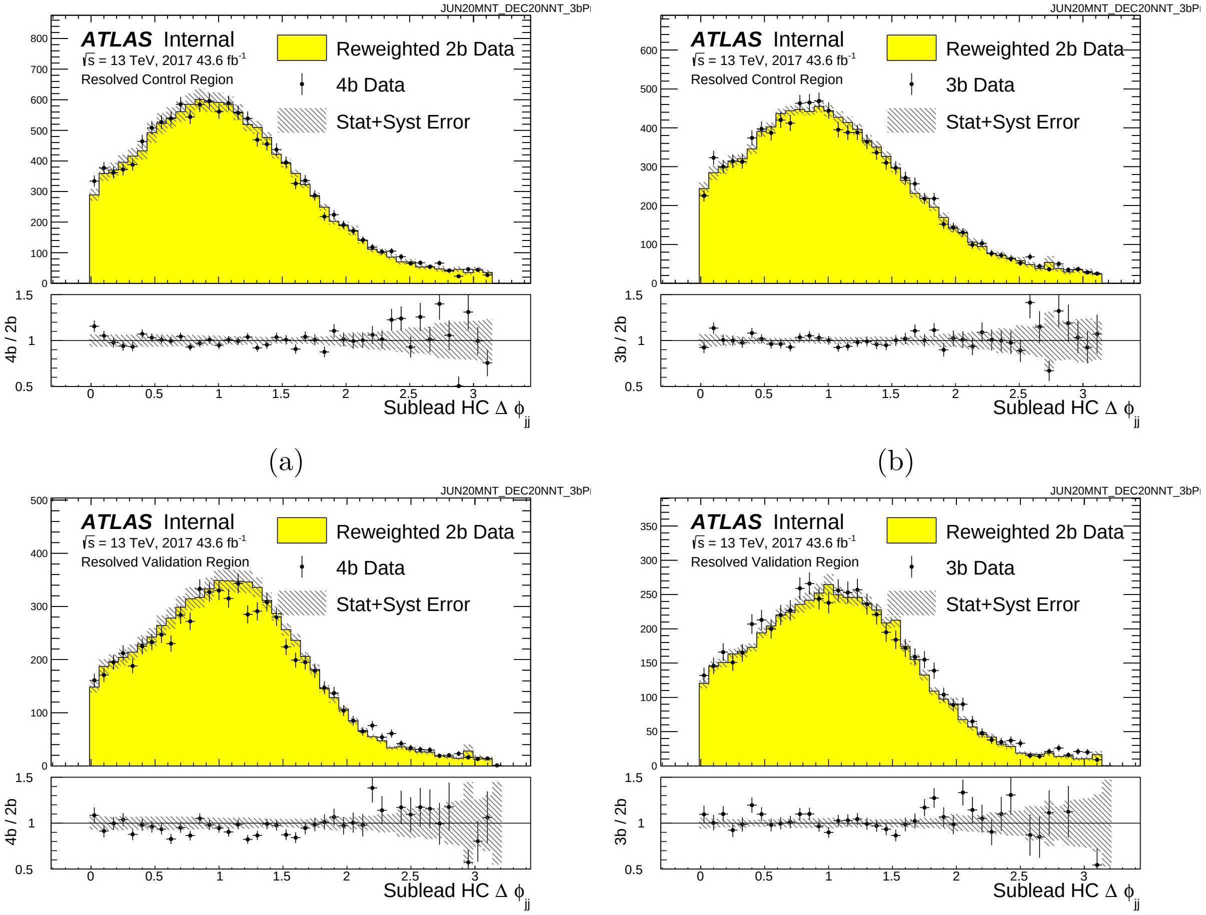
(a)

(b)

(c)

(d)

Figure A.7: Distributions of the subleading Higgs candidate $\Delta\phi_{jj}$ in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



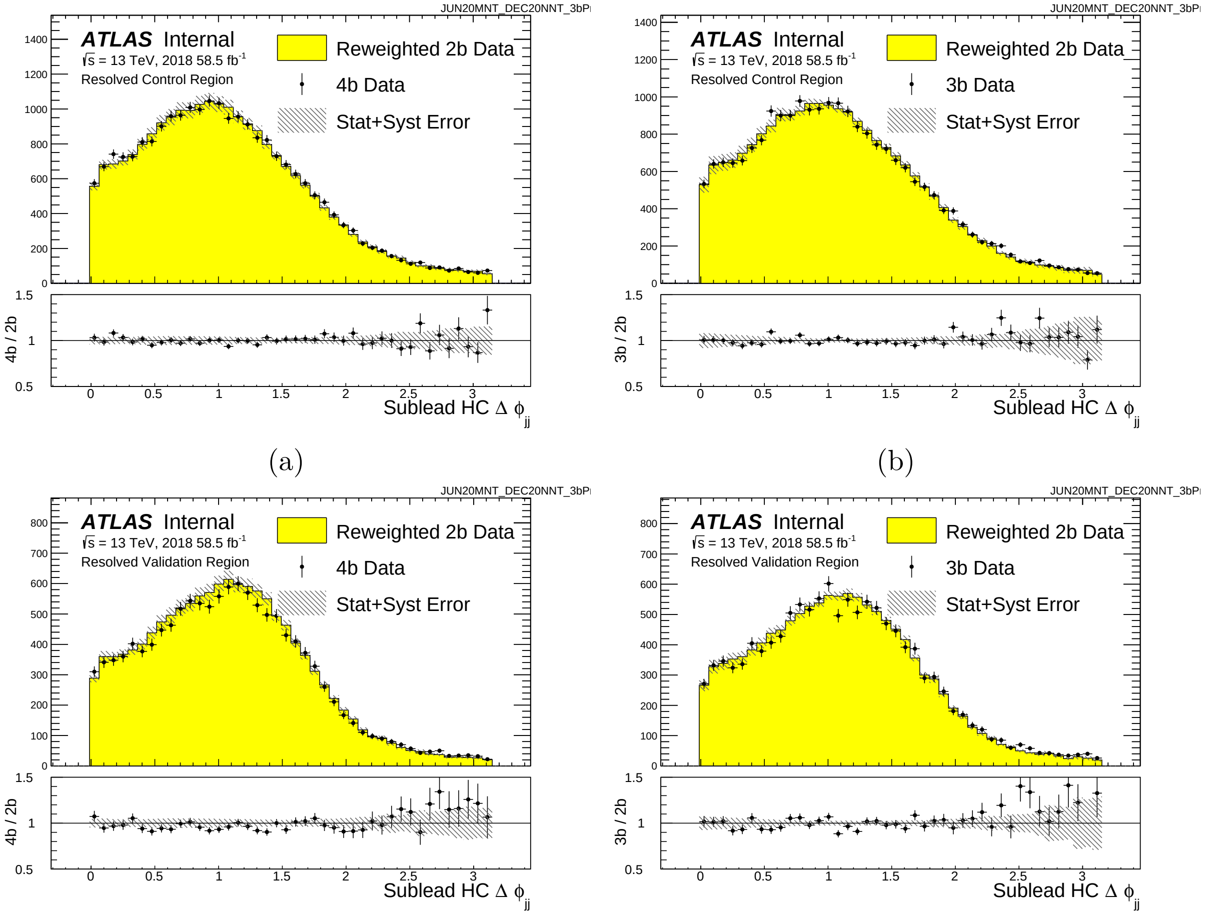
(a)

(b)

(c)

(d)

Figure A.8: Distributions of the subleading Higgs candidate $\Delta\phi_{jj}$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



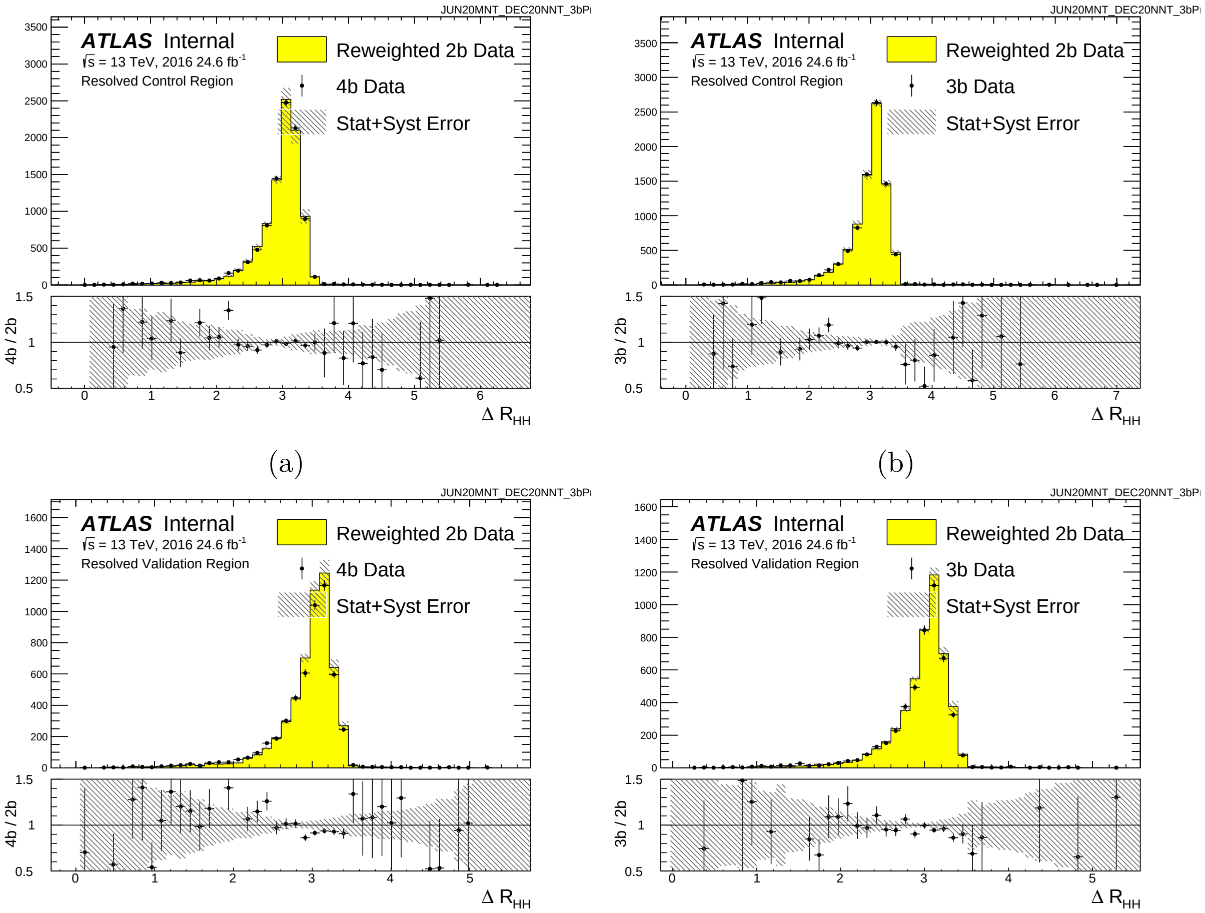
(a)

(b)

(c)

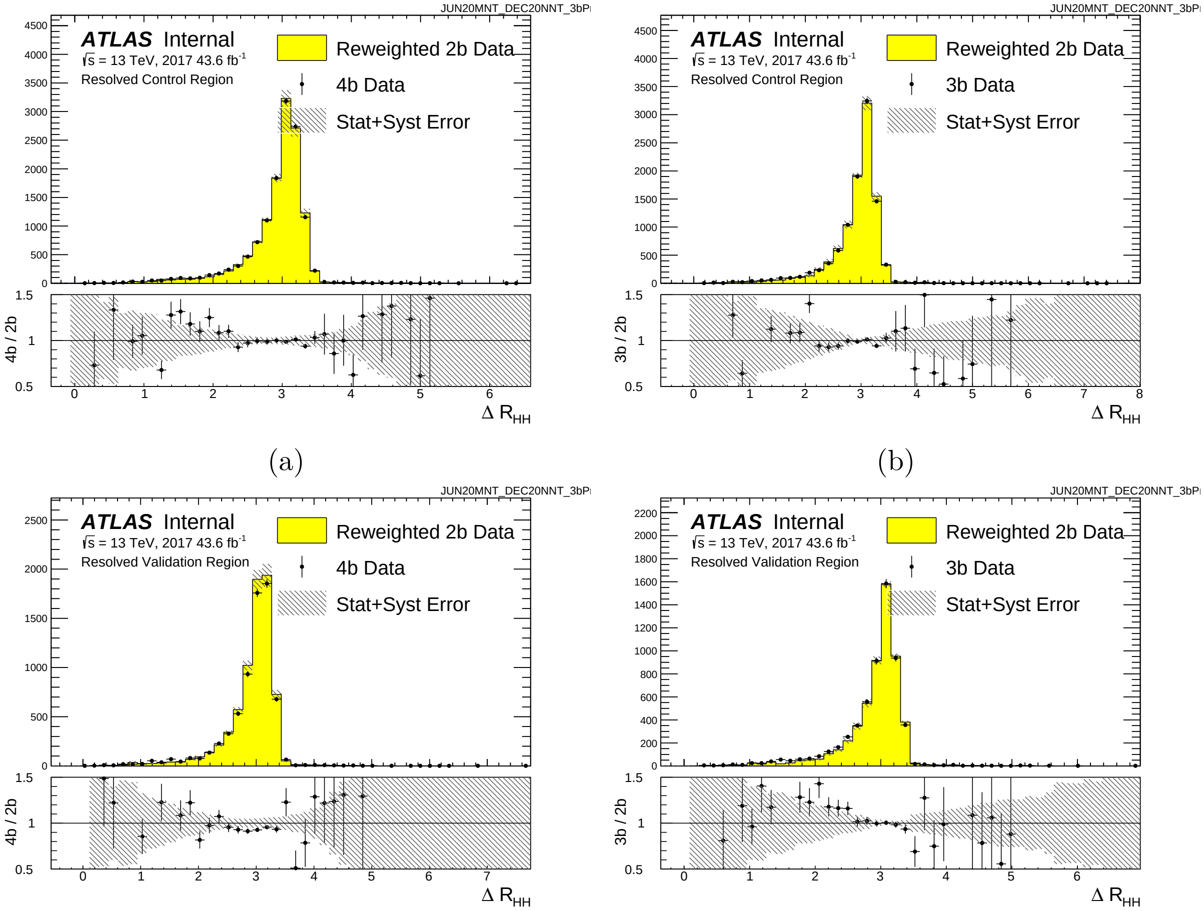
(d)

Figure A.9: Distributions of the subleading Higgs candidate $\Delta\phi_{jj}$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b) (c) (d)

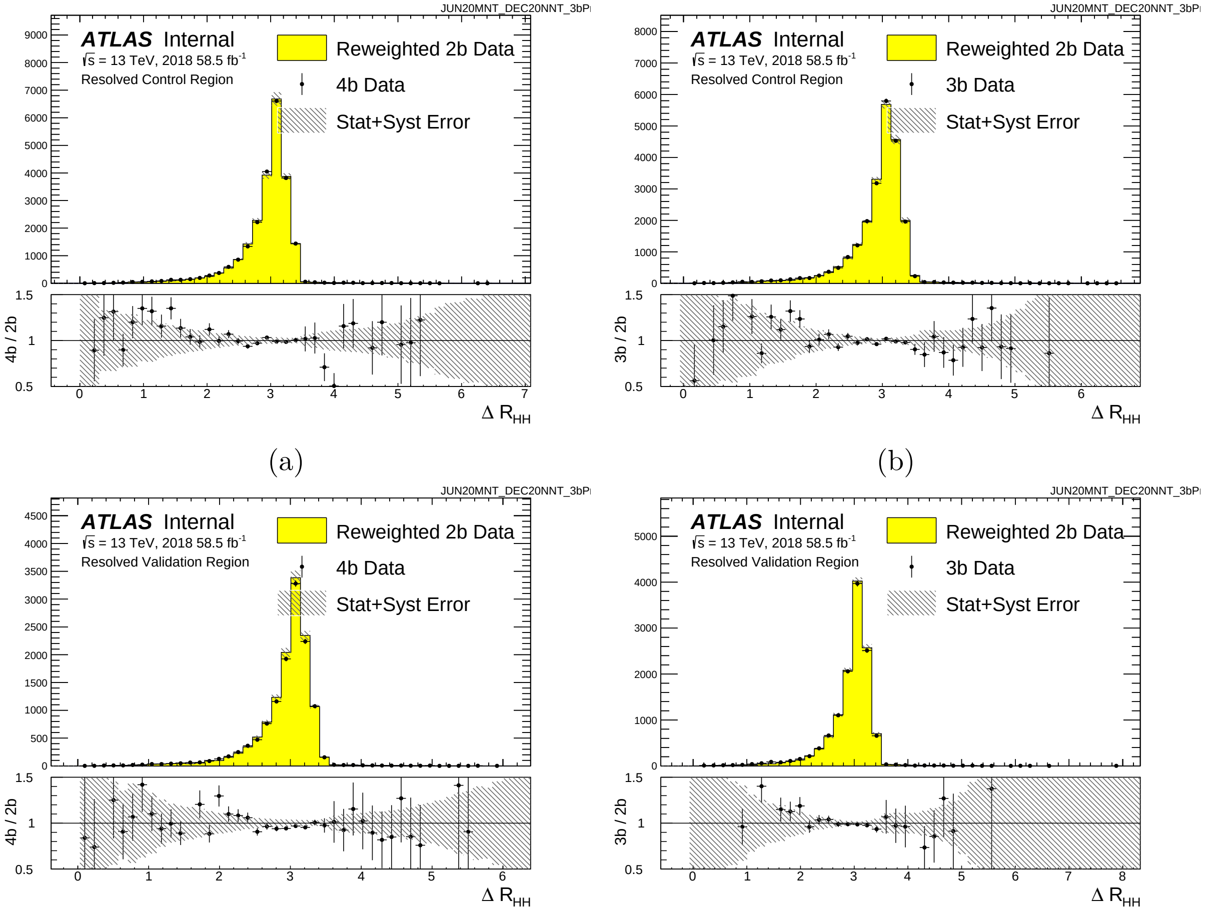
Figure A.10: Distributions of ΔR_{HH} in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b)

(c) (d)

Figure A.11: Distributions of ΔR_{HH} in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



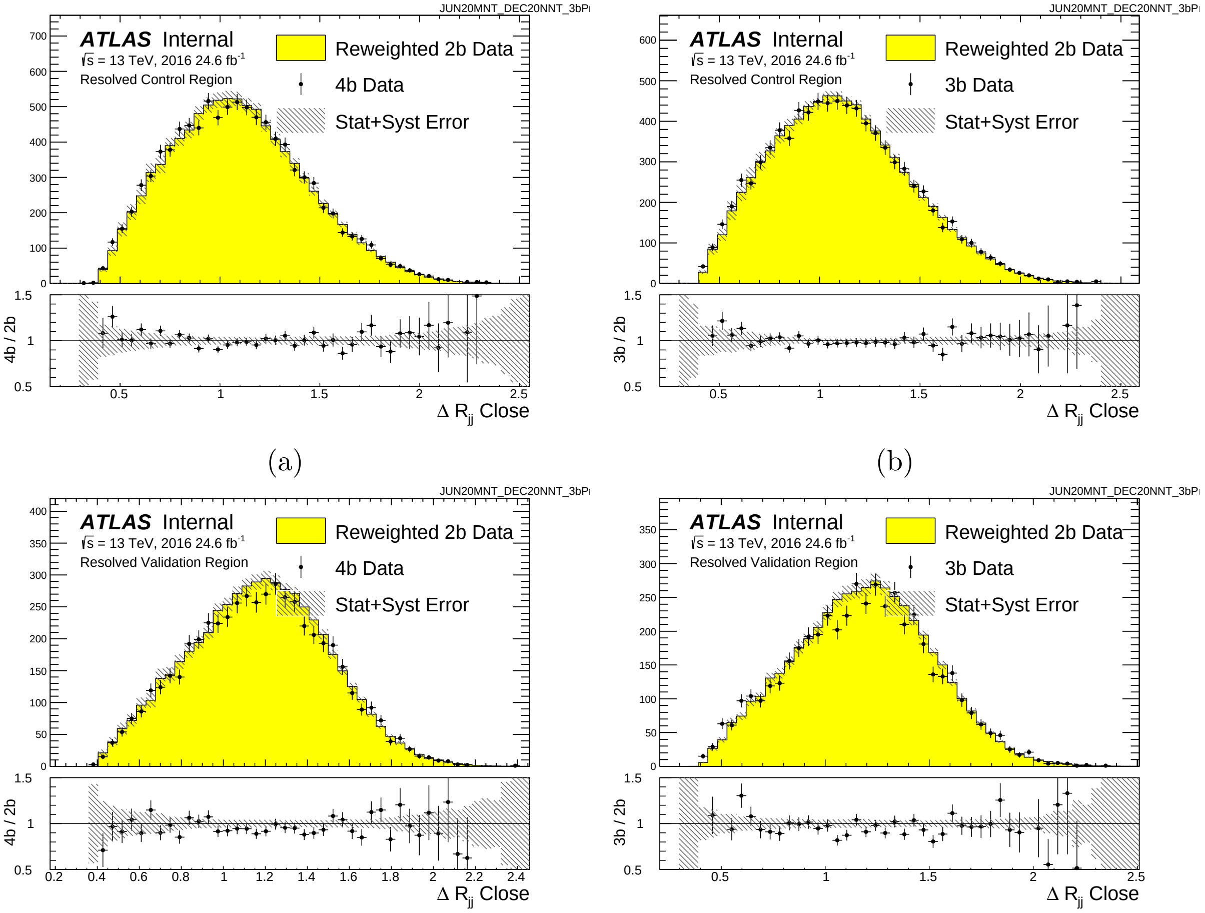
(a)

(b)

(c)

(d)

Figure A.12: Distributions of ΔR_{HH} in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



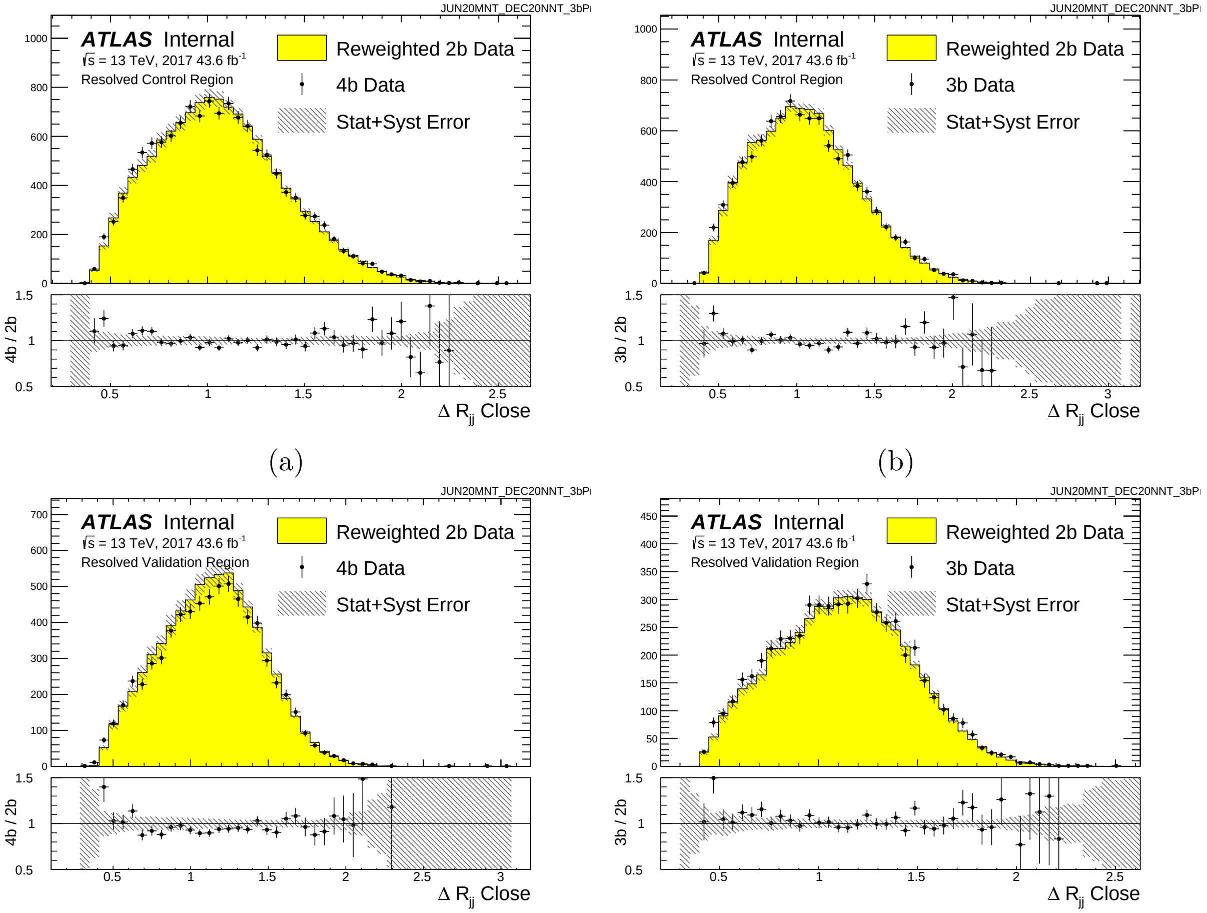
(a)

(b)

(c)

(d)

Figure A.13: Distributions of the leading Higgs candidate ΔR_{jj} in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



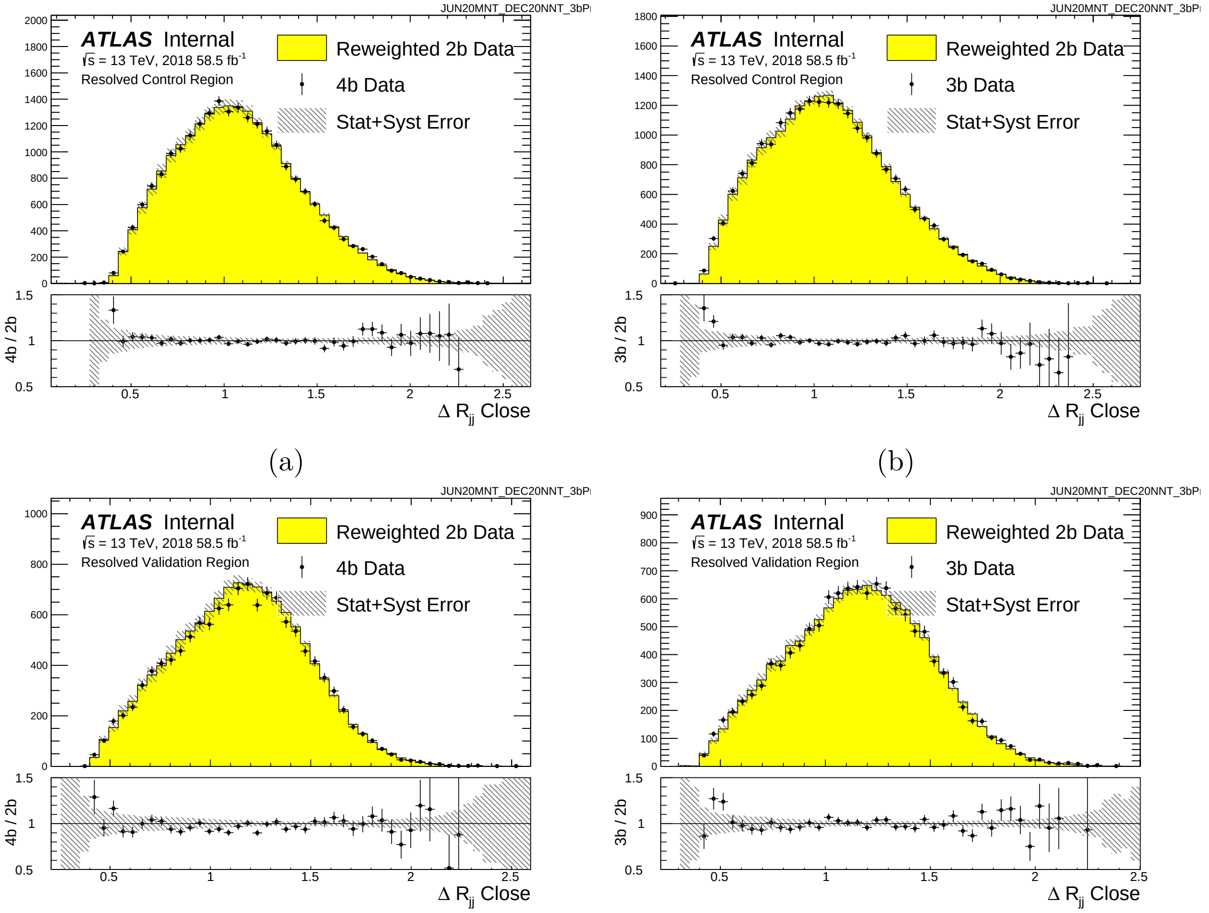
(a)

(b)

(c)

(d)

Figure A.14: Distributions of the leading Higgs candidate ΔR_{jj} in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



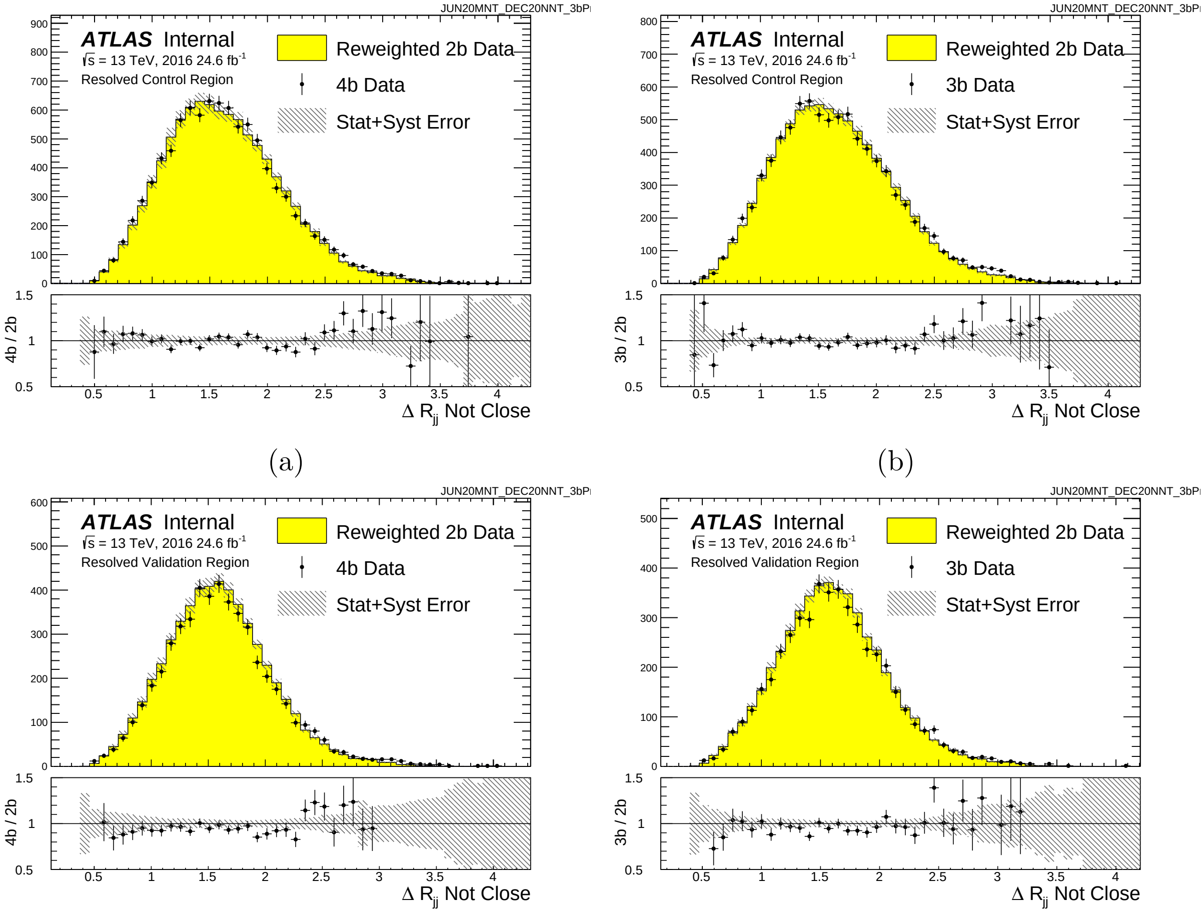
(a)

(b)

(c)

(d)

Figure A.15: Distributions of the leading Higgs candidate ΔR_{jj} in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



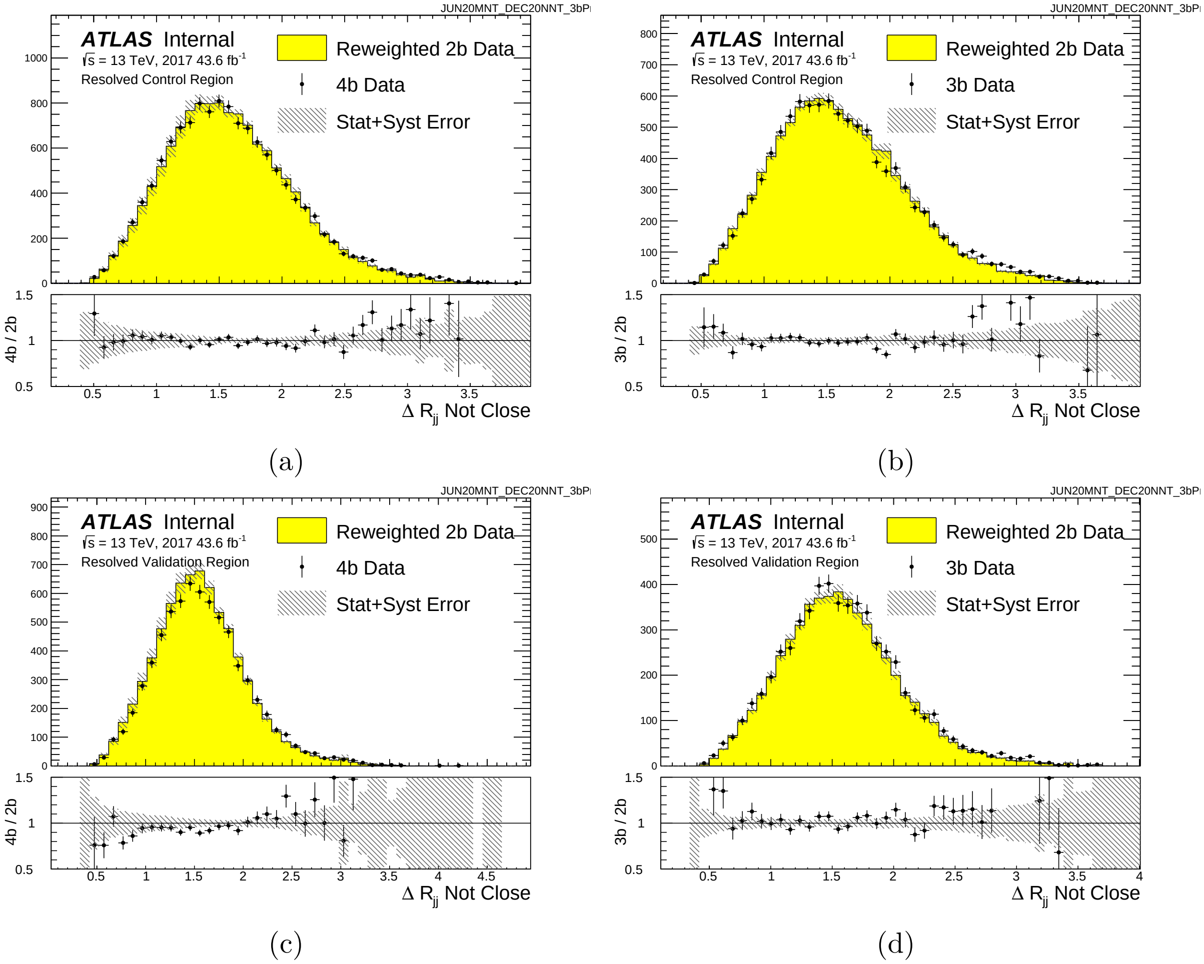
(a)

(b)

(c)

(d)

Figure A.16: Distributions of the subleading Higgs candidate ΔR_{jj} in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a)

(b)

(c)

(d)

Figure A.17: Distributions of the subleading Higgs candidate ΔR_{jj} in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

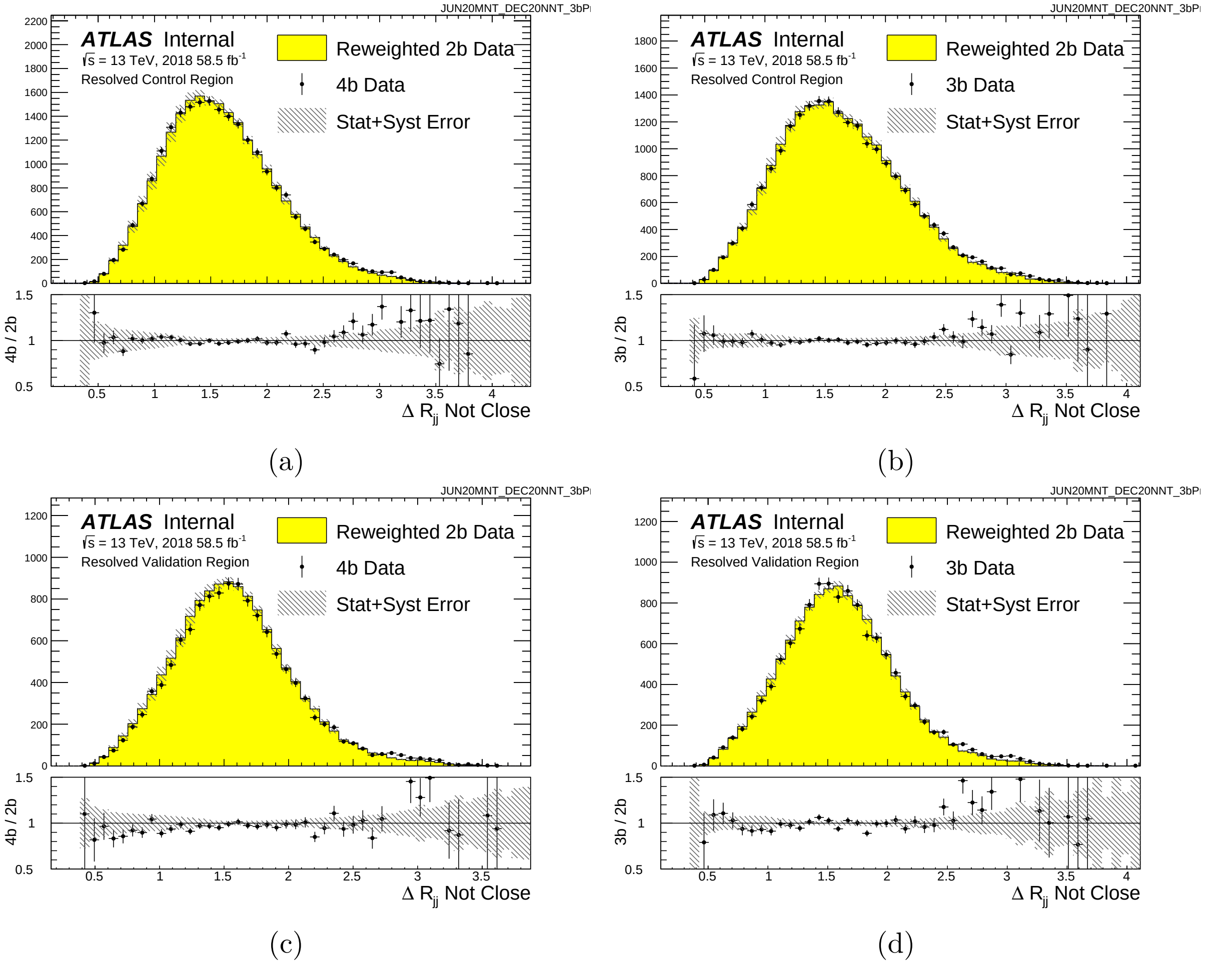
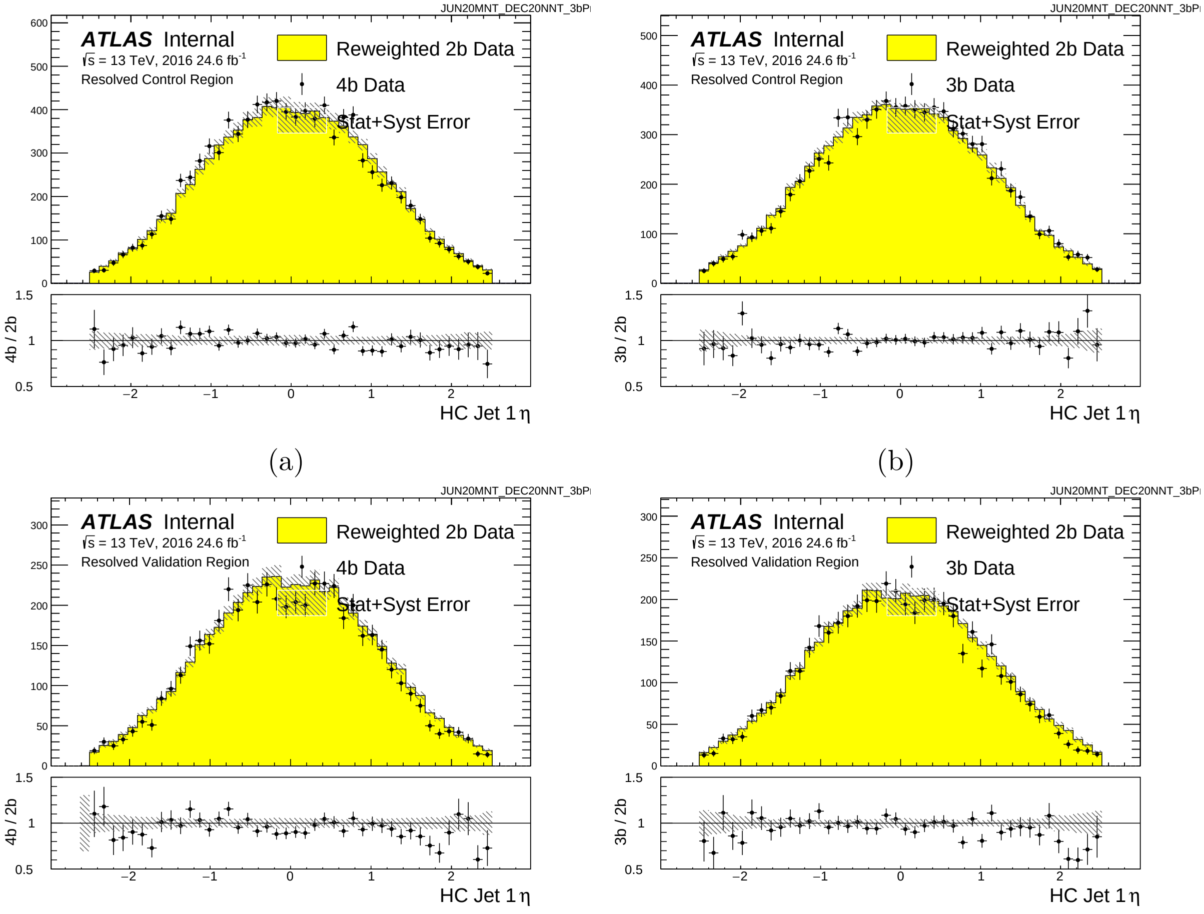


Figure A.18: Distributions of the subleading Higgs candidate ΔR_{jj} in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



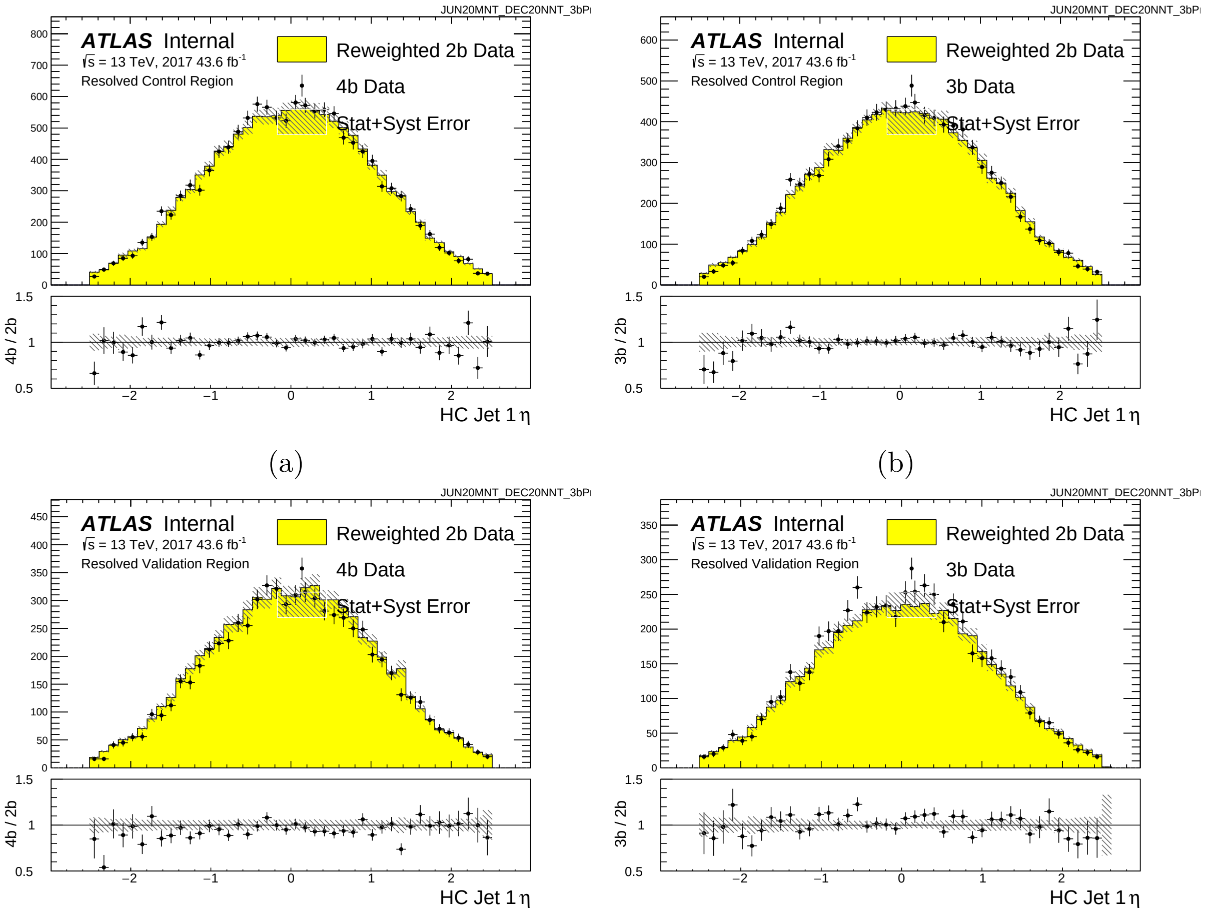
(a)

(b)

(c)

(d)

Figure A.19: Distributions of the first jet η in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



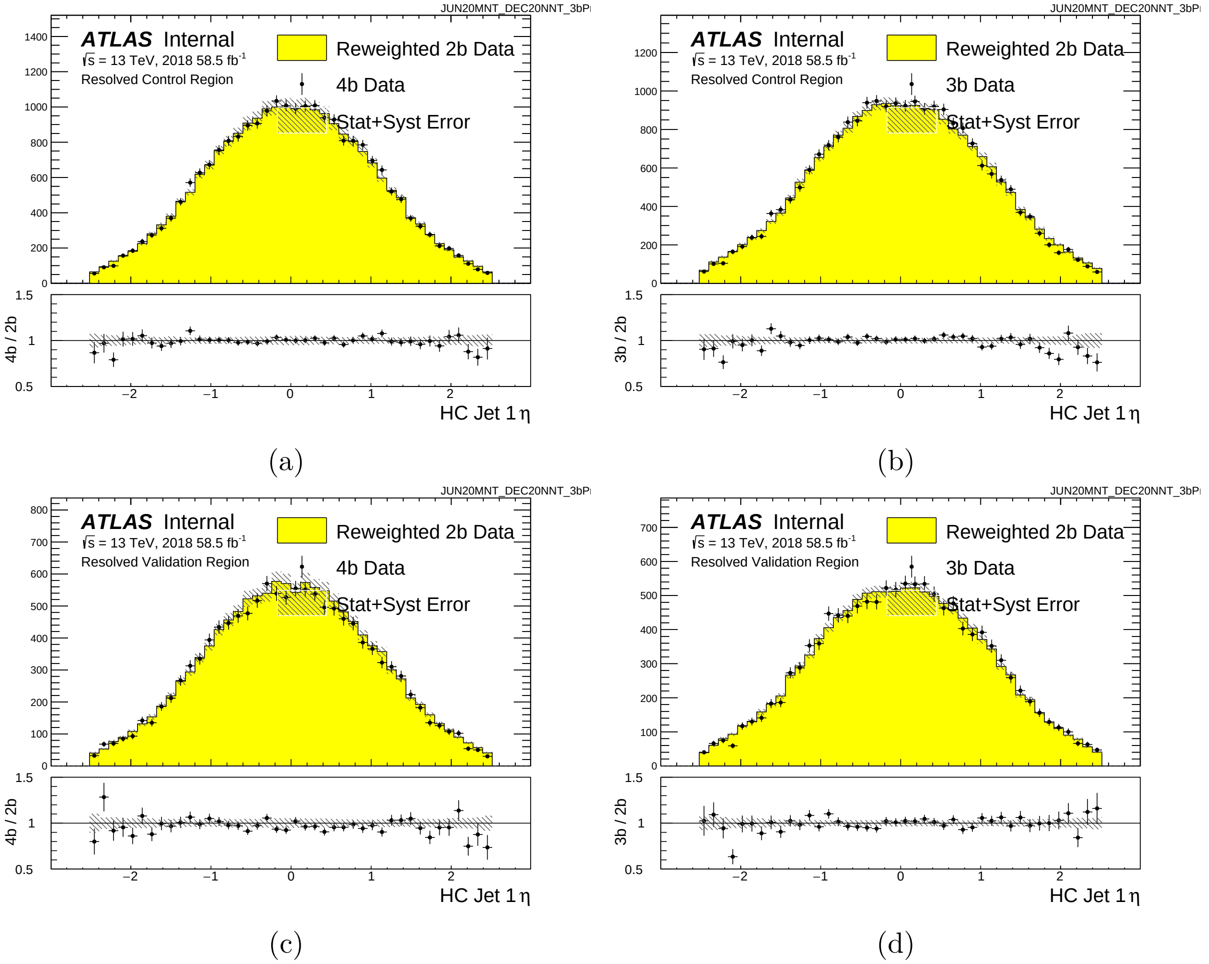
(a)

(b)

(c)

(d)

Figure A.20: Distributions of the first jet η in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



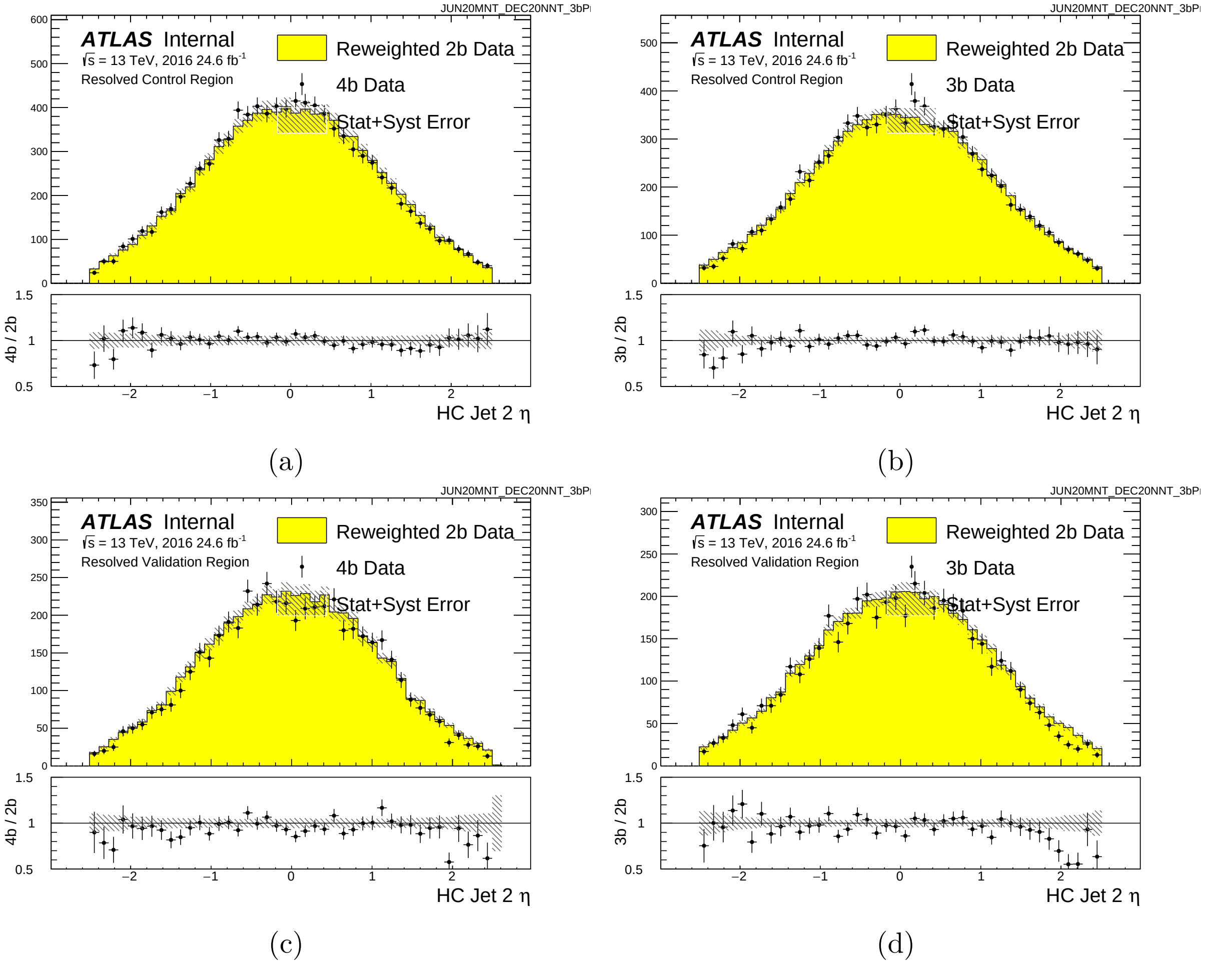
(a)

(b)

(c)

(d)

Figure A.21: Distributions of the first jet η in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



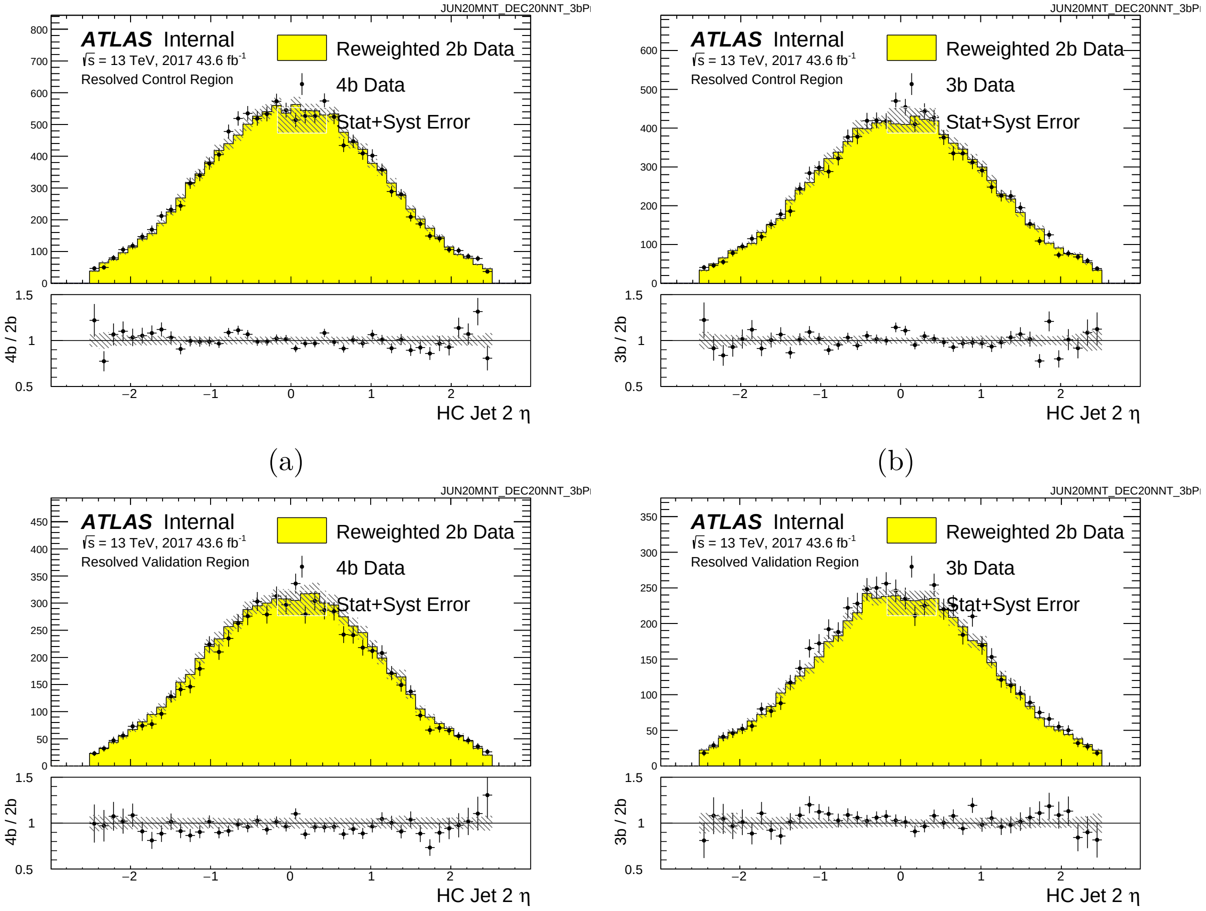
(a)

(b)

(c)

(d)

Figure A.22: Distributions of the second jet η in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a)

(b)

(c)

(d)

Figure A.23: Distributions of the second jet η in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

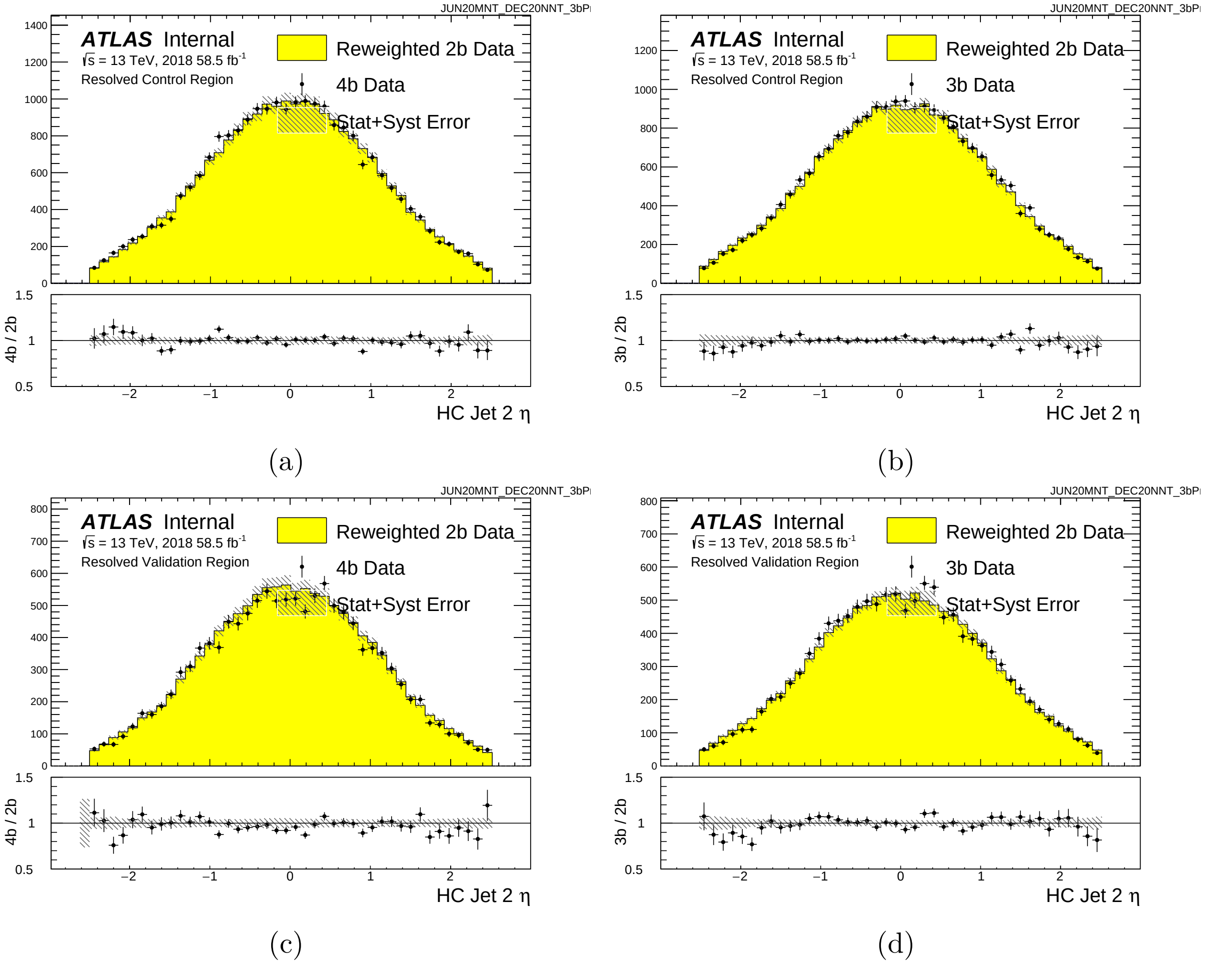
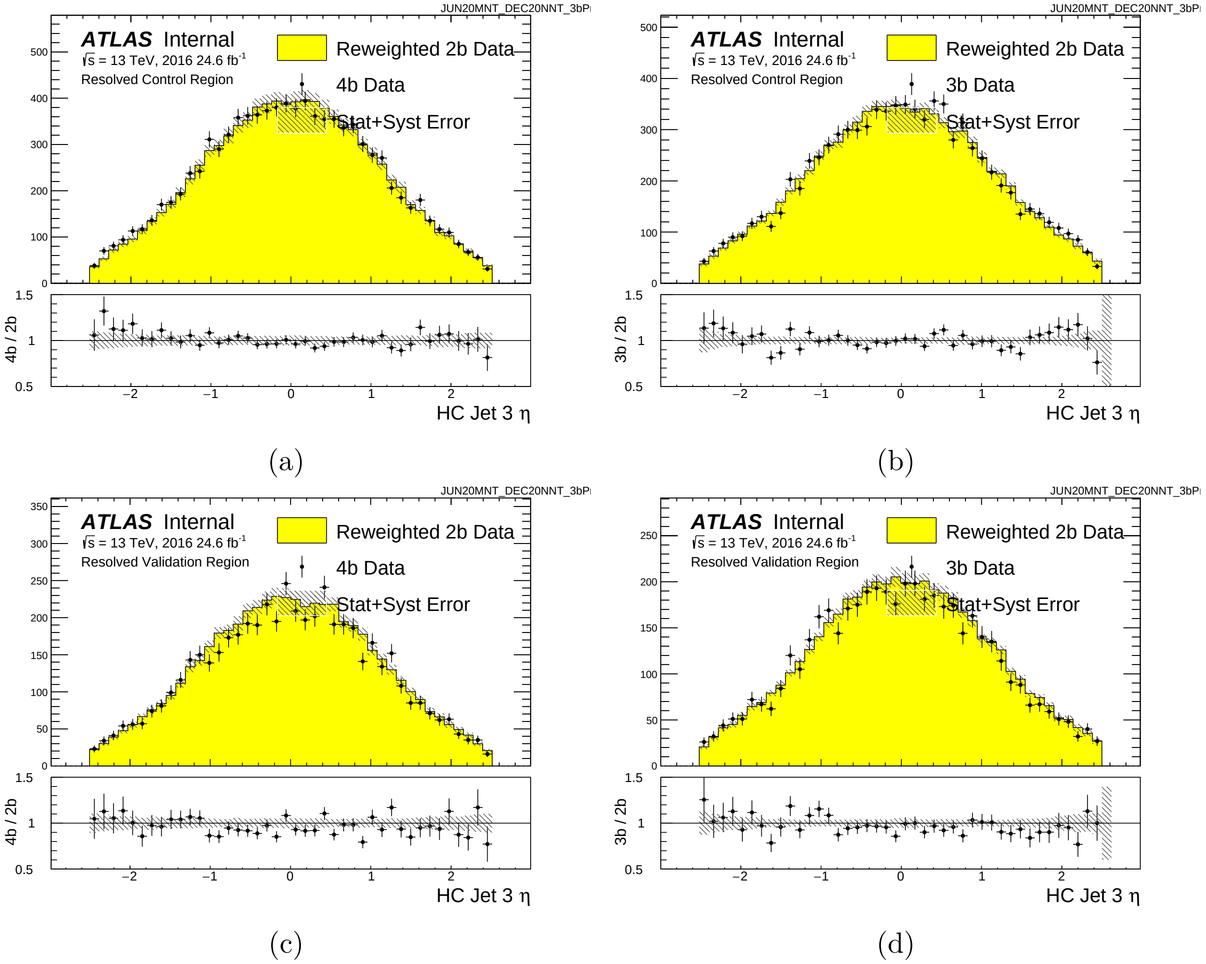


Figure A.24: Distributions of the second jet η in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



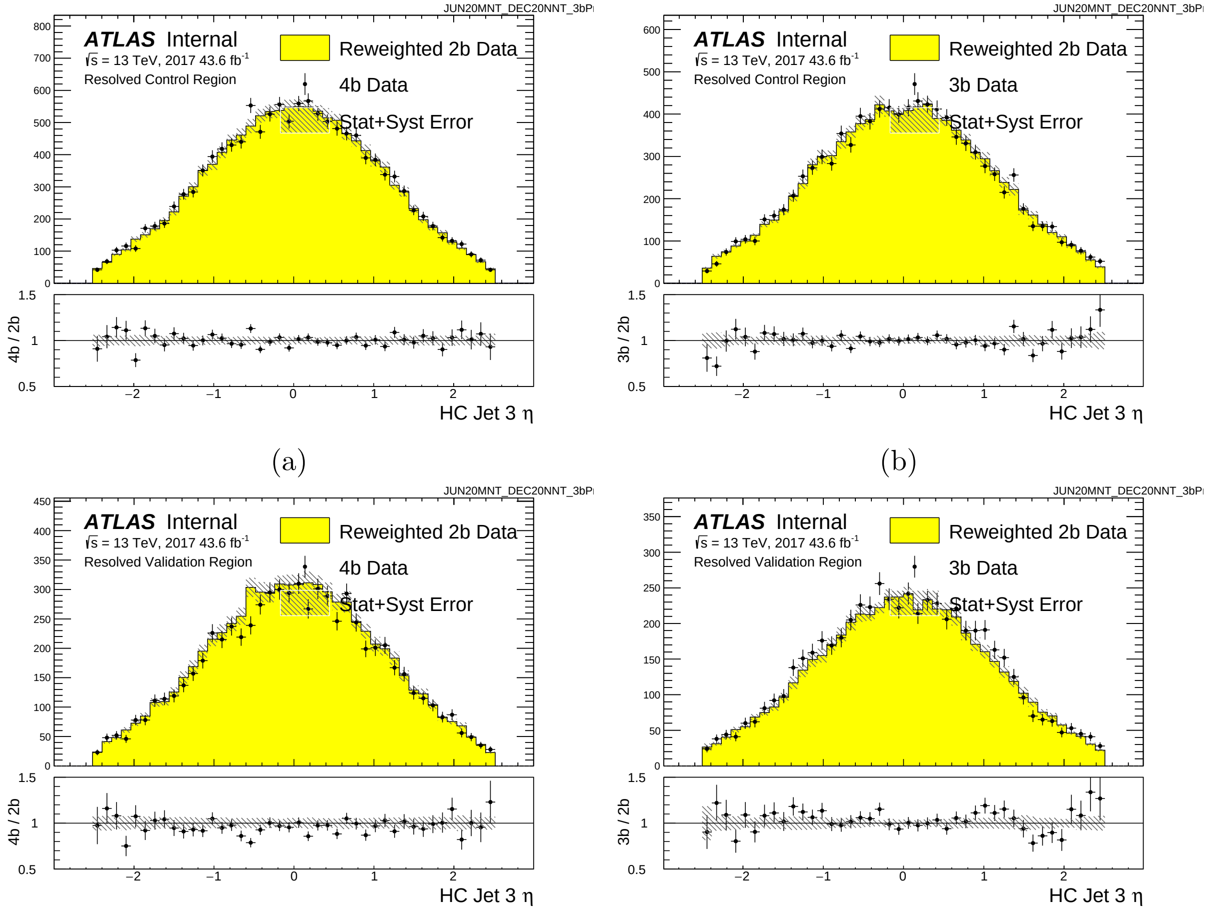
(a)

(b)

(c)

(d)

Figure A.25: Distributions of the third jet η in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



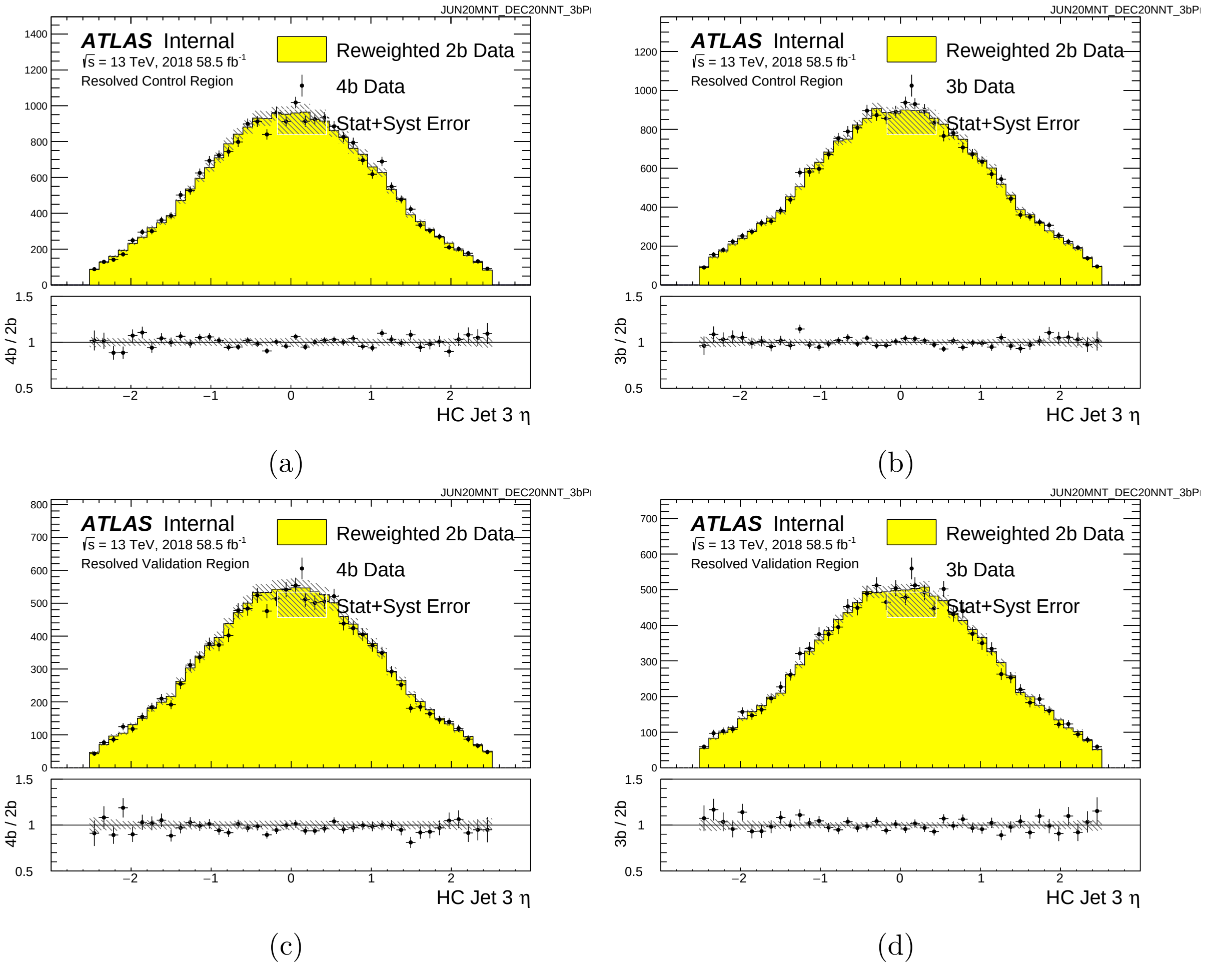
(a)

(b)

(c)

(d)

Figure A.26: Distributions of the third jet η in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



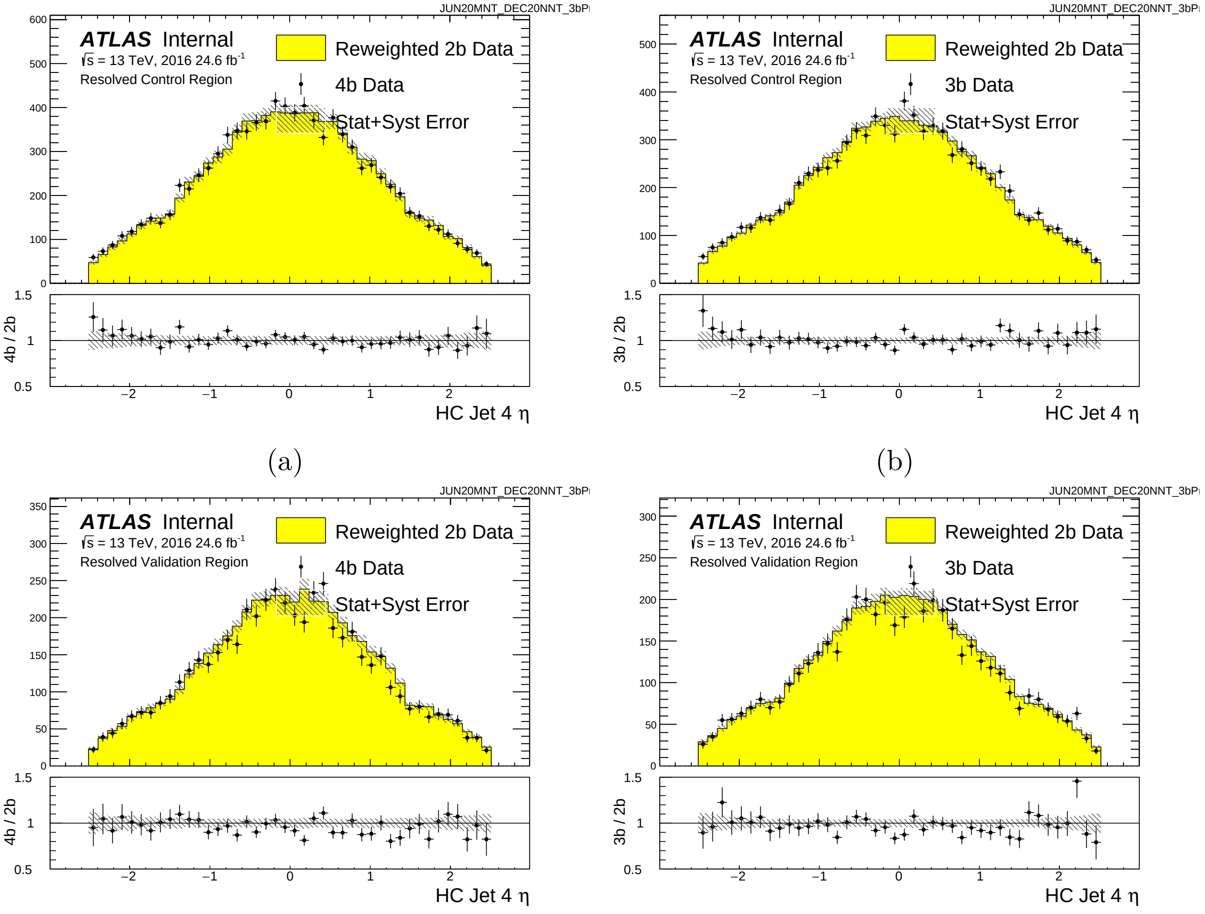
(a)

(b)

(c)

(d)

Figure A.27: Distributions of the third jet η in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



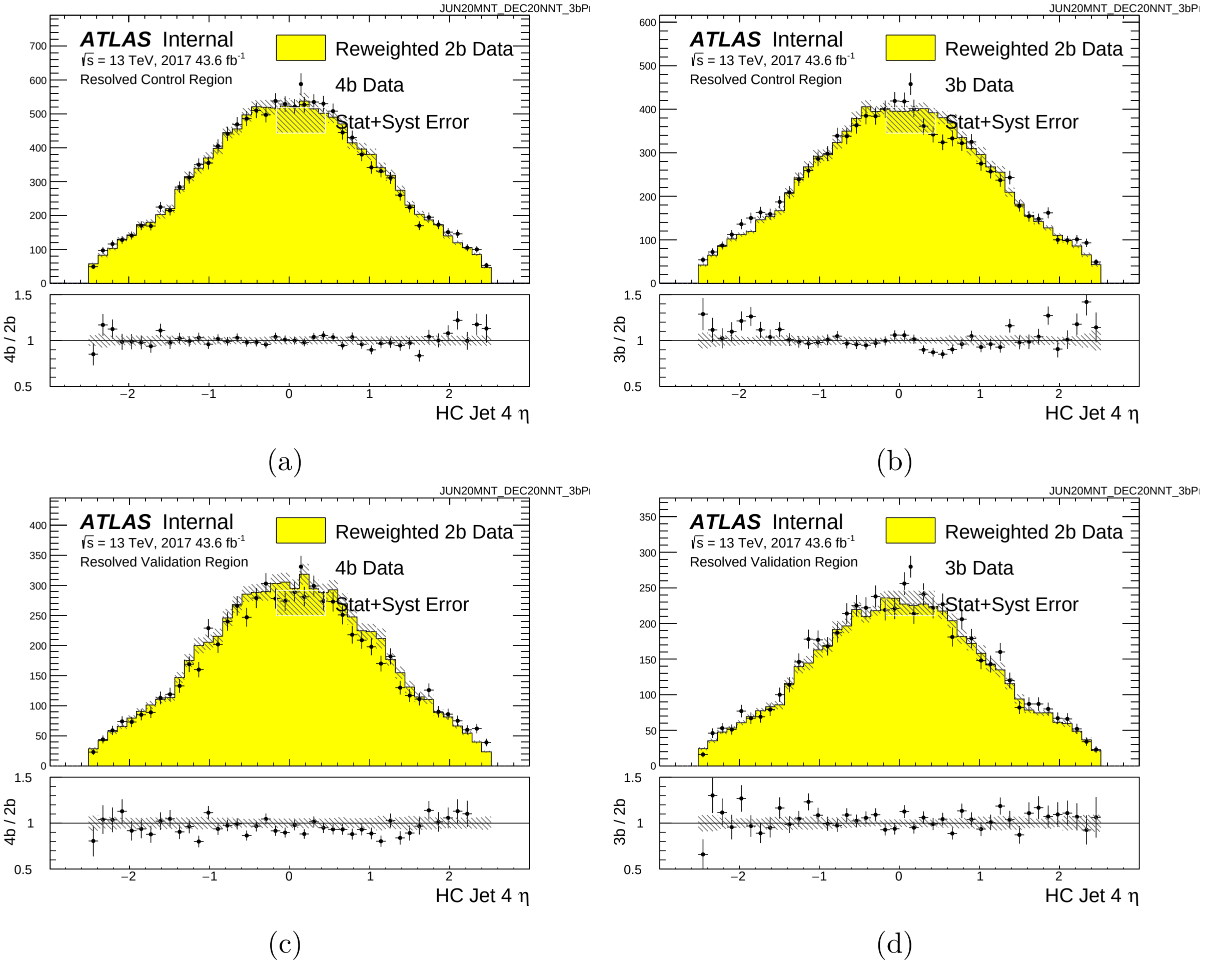
(a)

(b)

(c)

(d)

Figure A.28: Distributions of the fourth jet η in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



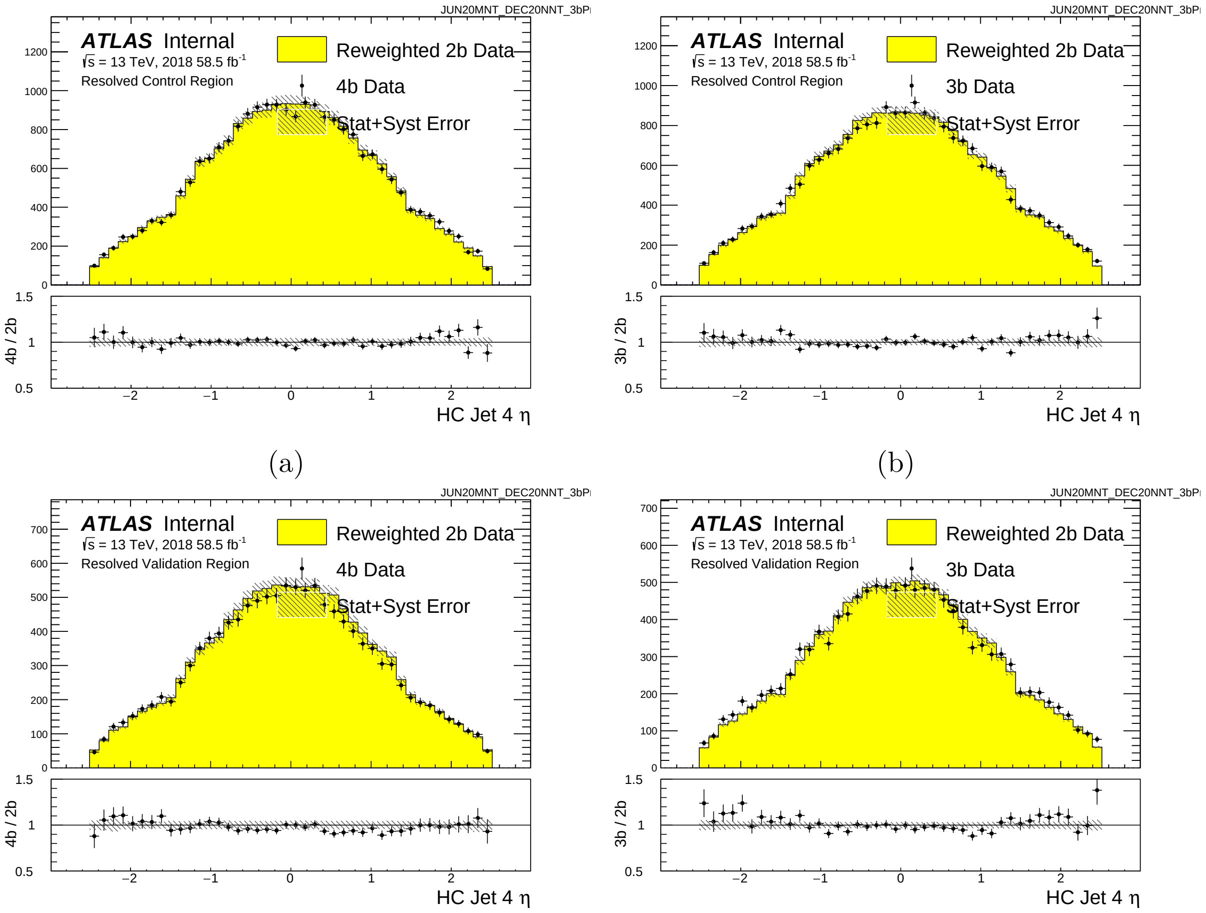
(a)

(b)

(c)

(d)

Figure A.29: Distributions of the fourth jet η in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



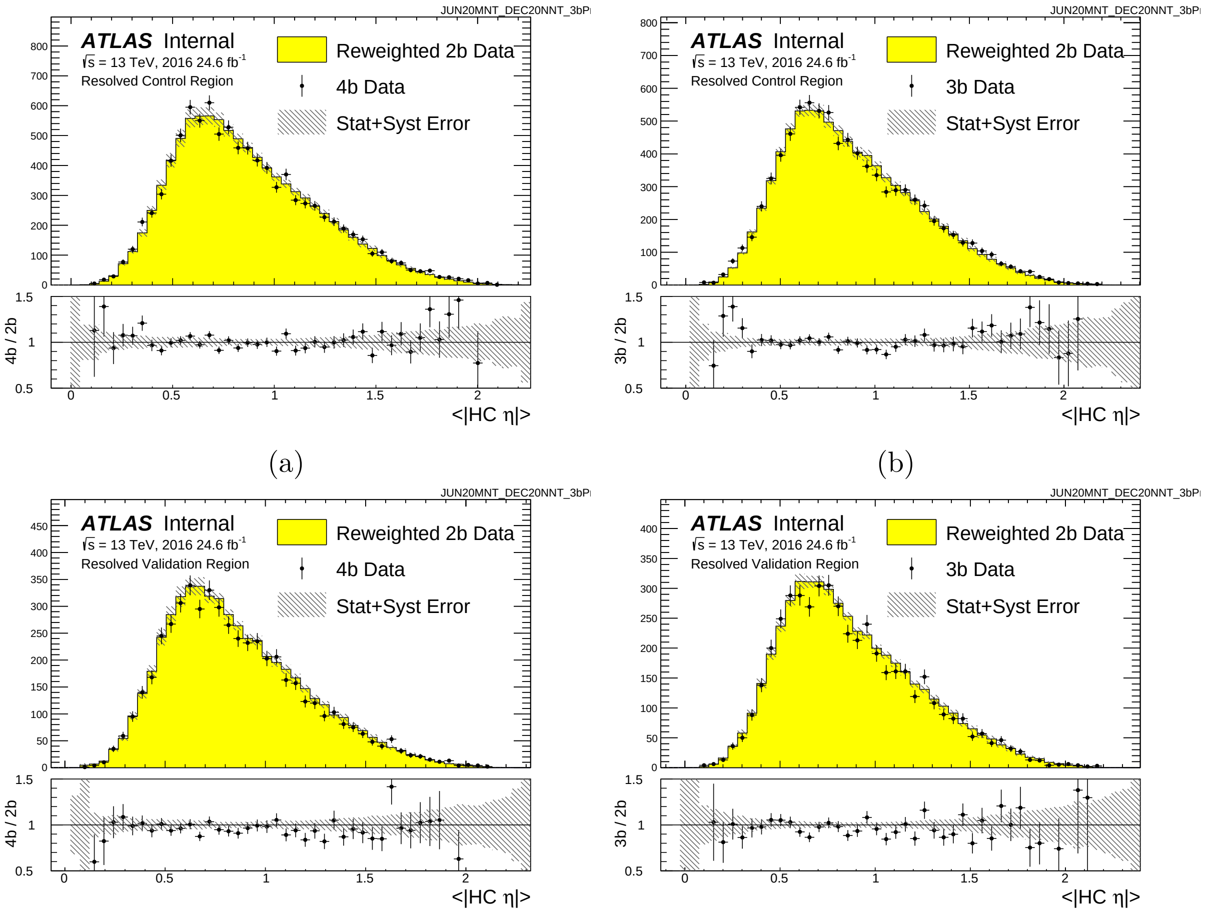
(a)

(b)

(c)

(d)

Figure A.30: Distributions of the fourth jet η in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



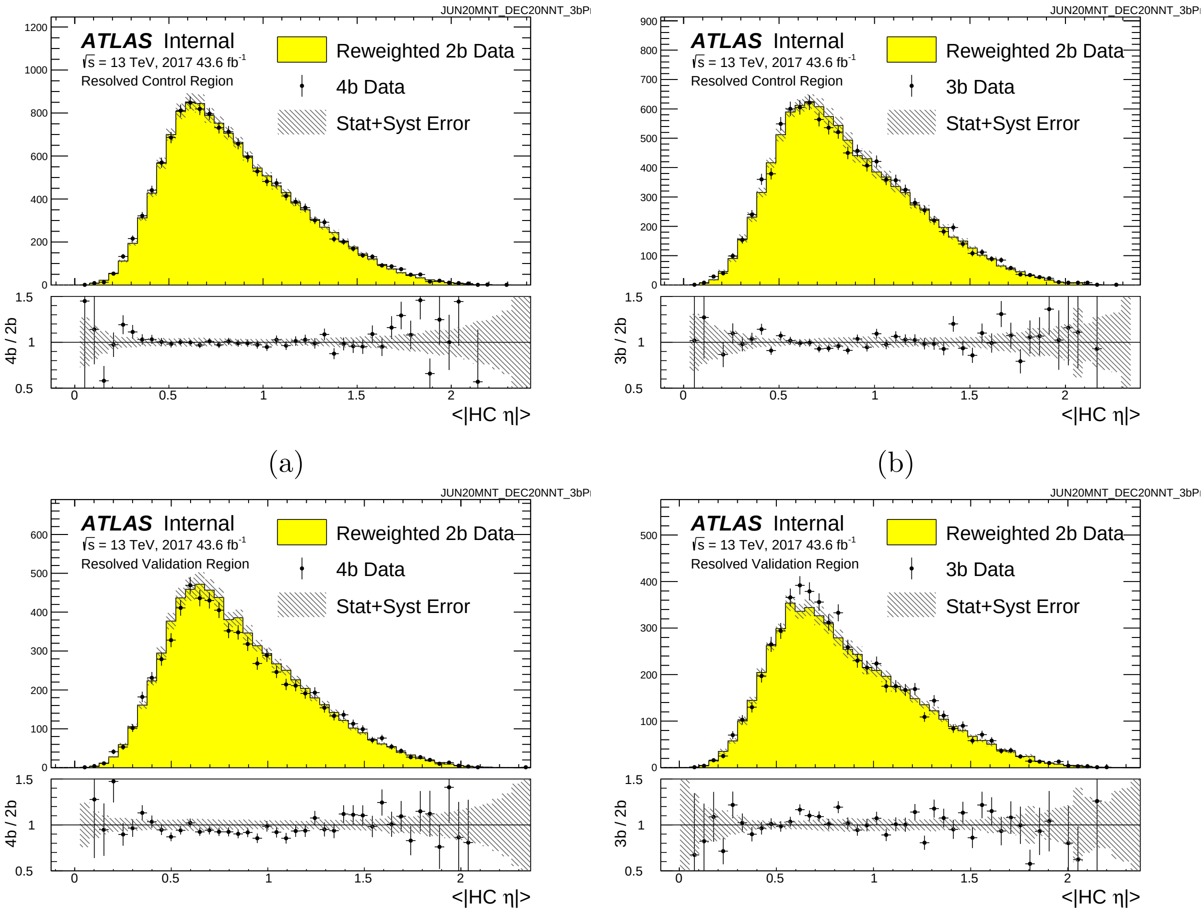
(a)

(b)

(c)

(d)

Figure A.31: Distributions of the average of the absolue value of the four jet η's in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



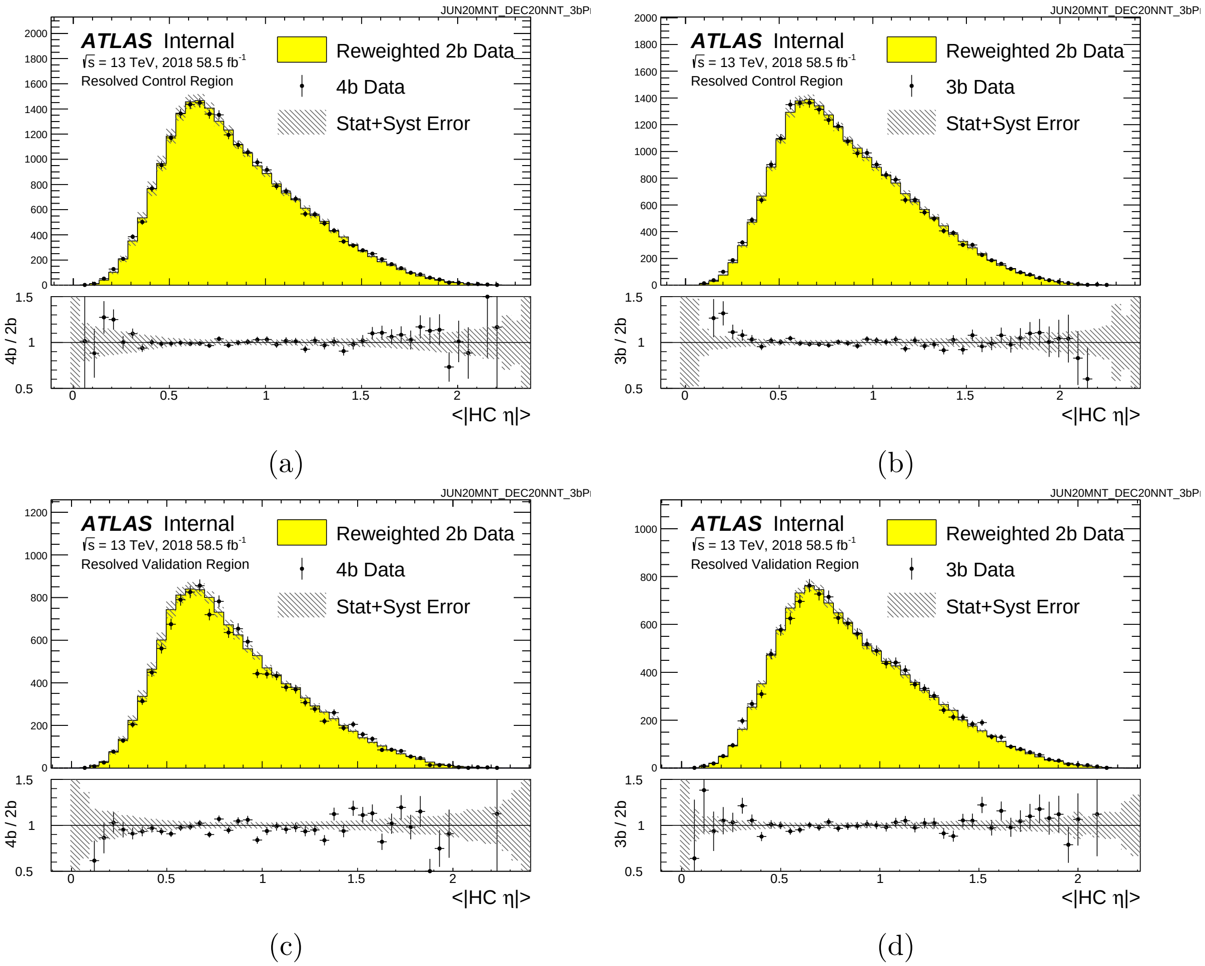
(a)

(b)

(c)

(d)

Figure A.32: Distributions of the average of the absolue value of the four jet η's in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



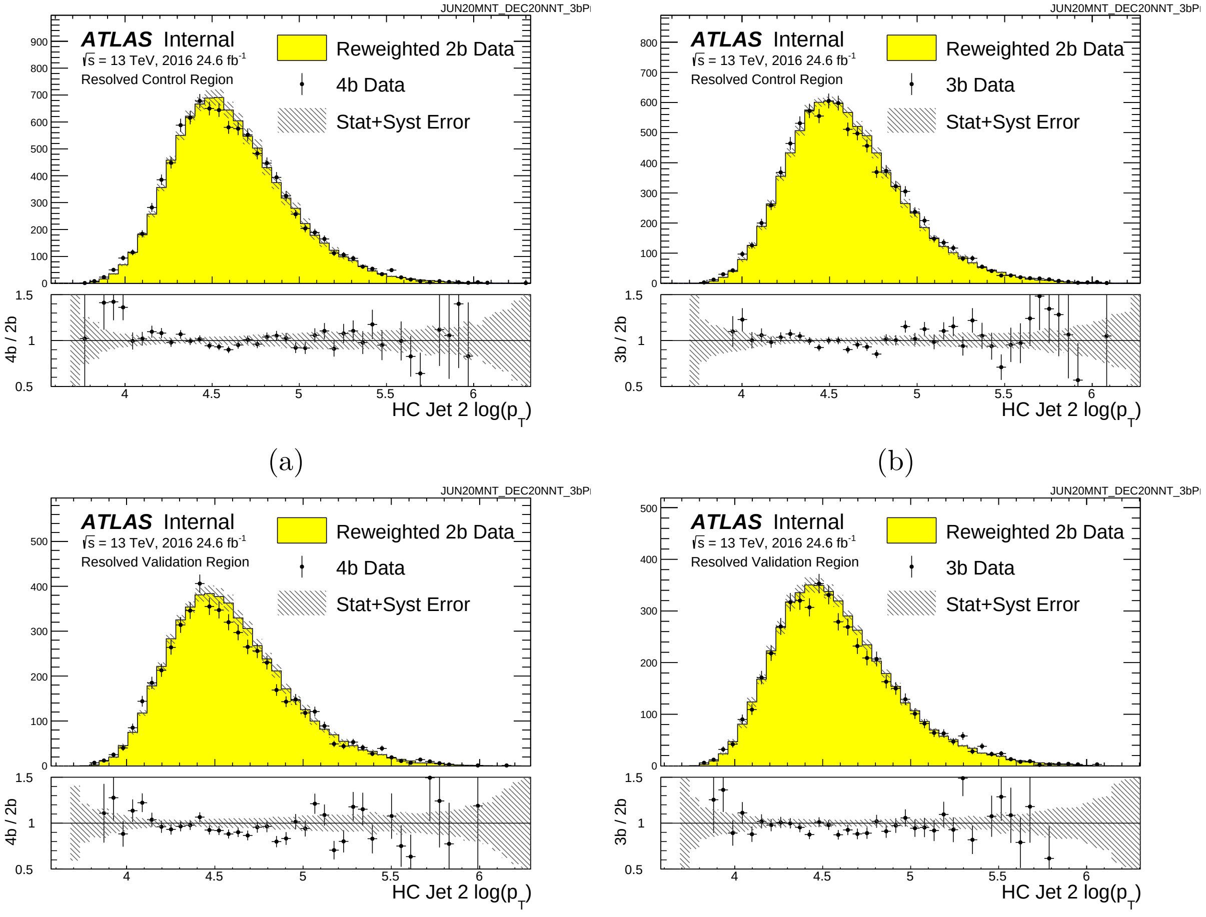
(a)

(b)

(c)

(d)

Figure A.33: Distributions of the average of the absolue value of the four jet η's in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



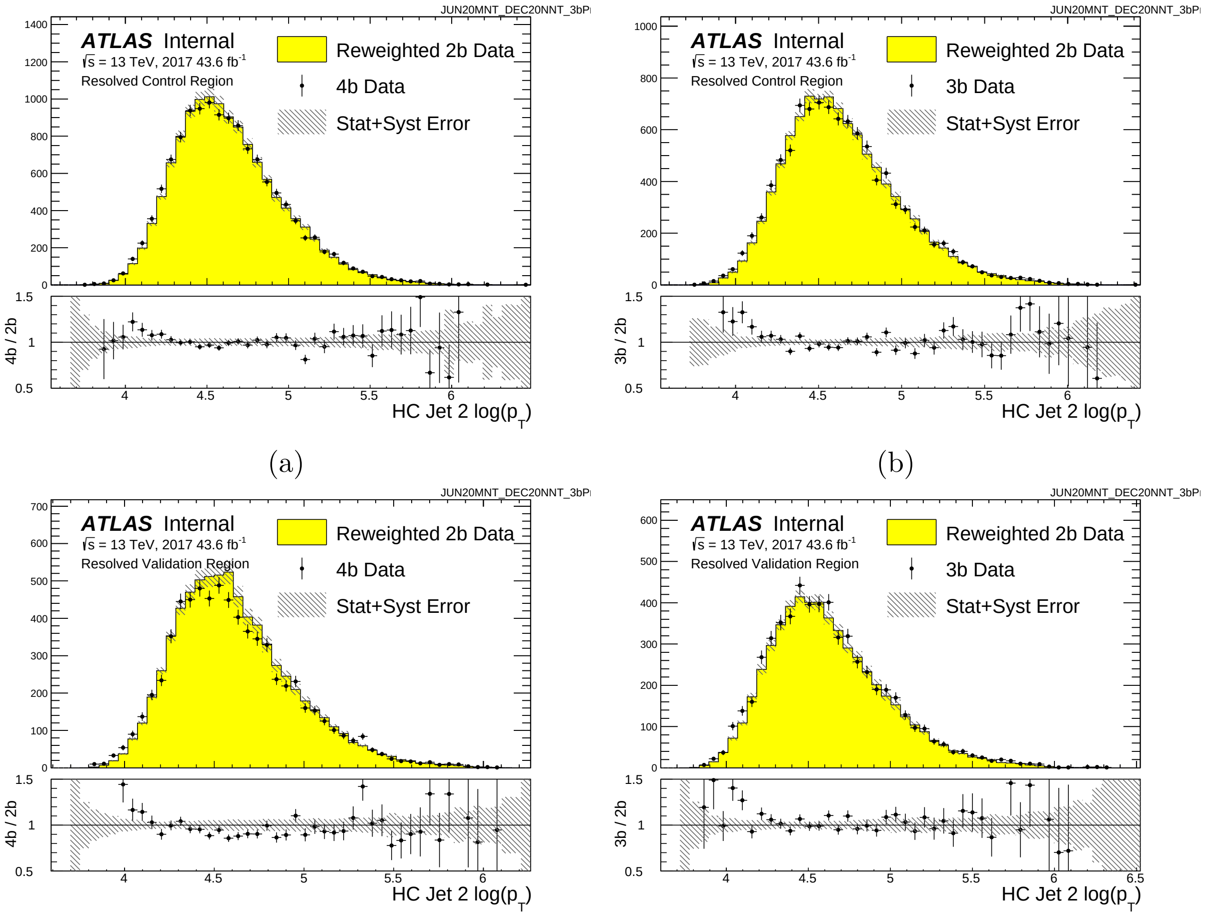
(a)

(b)

(c)

(d)

Figure A.34: Distributions of the second jet $\log p_T$ in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



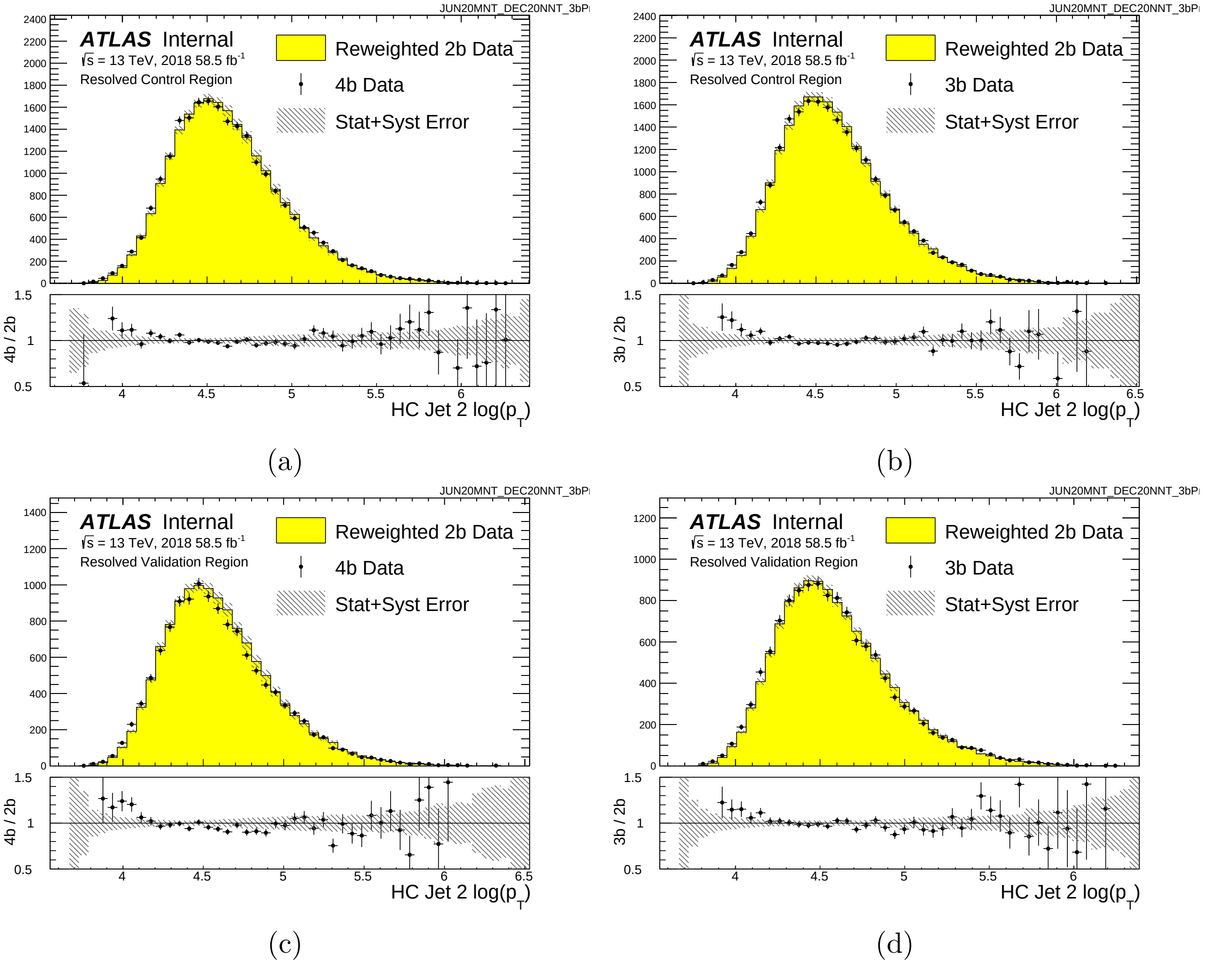
(a)

(b)

(c)

(d)

Figure A.35: Distributions of the second jet $\log p_T$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



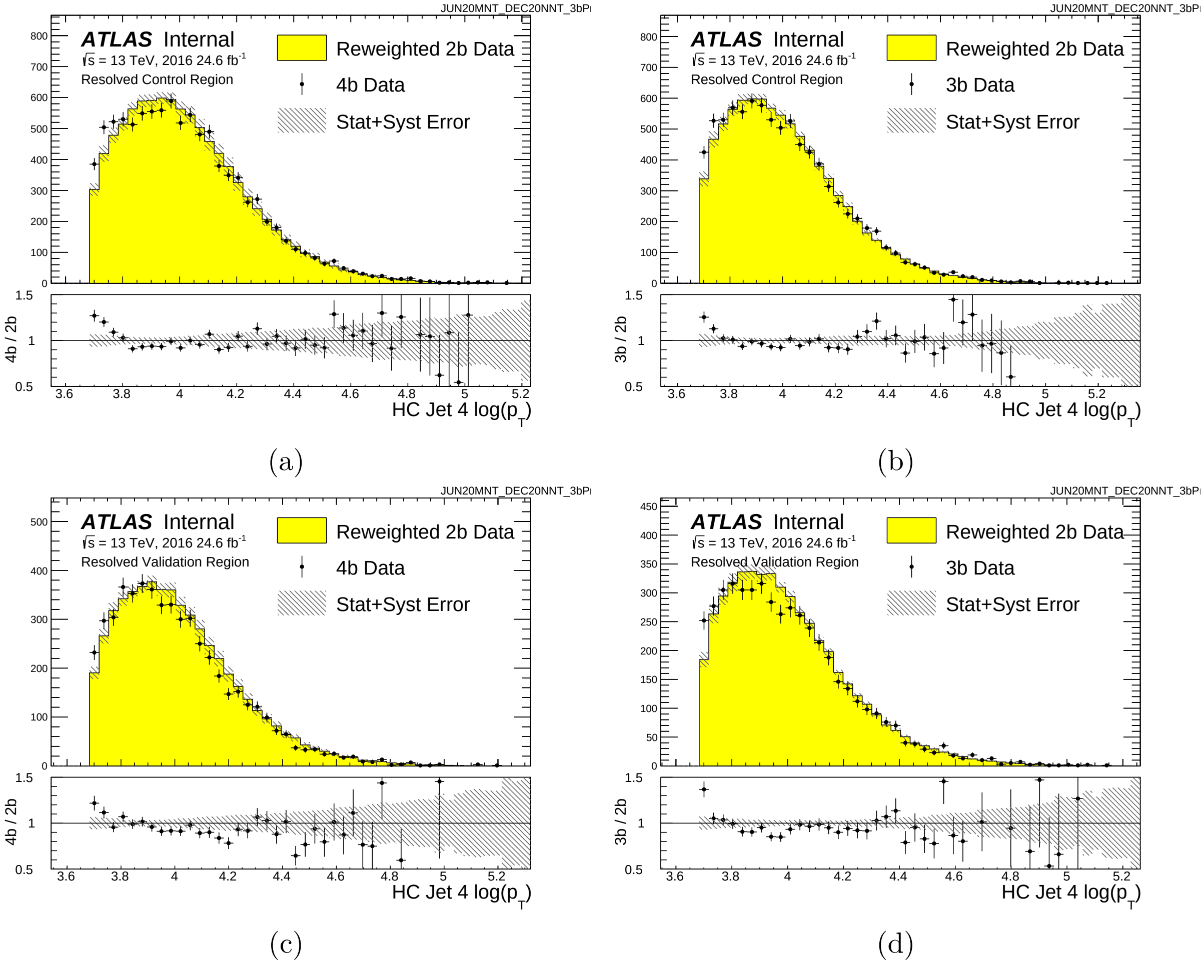
(a)

(b)

(c)

(d)

Figure A.36: Distributions of the second jet $\log p_T$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



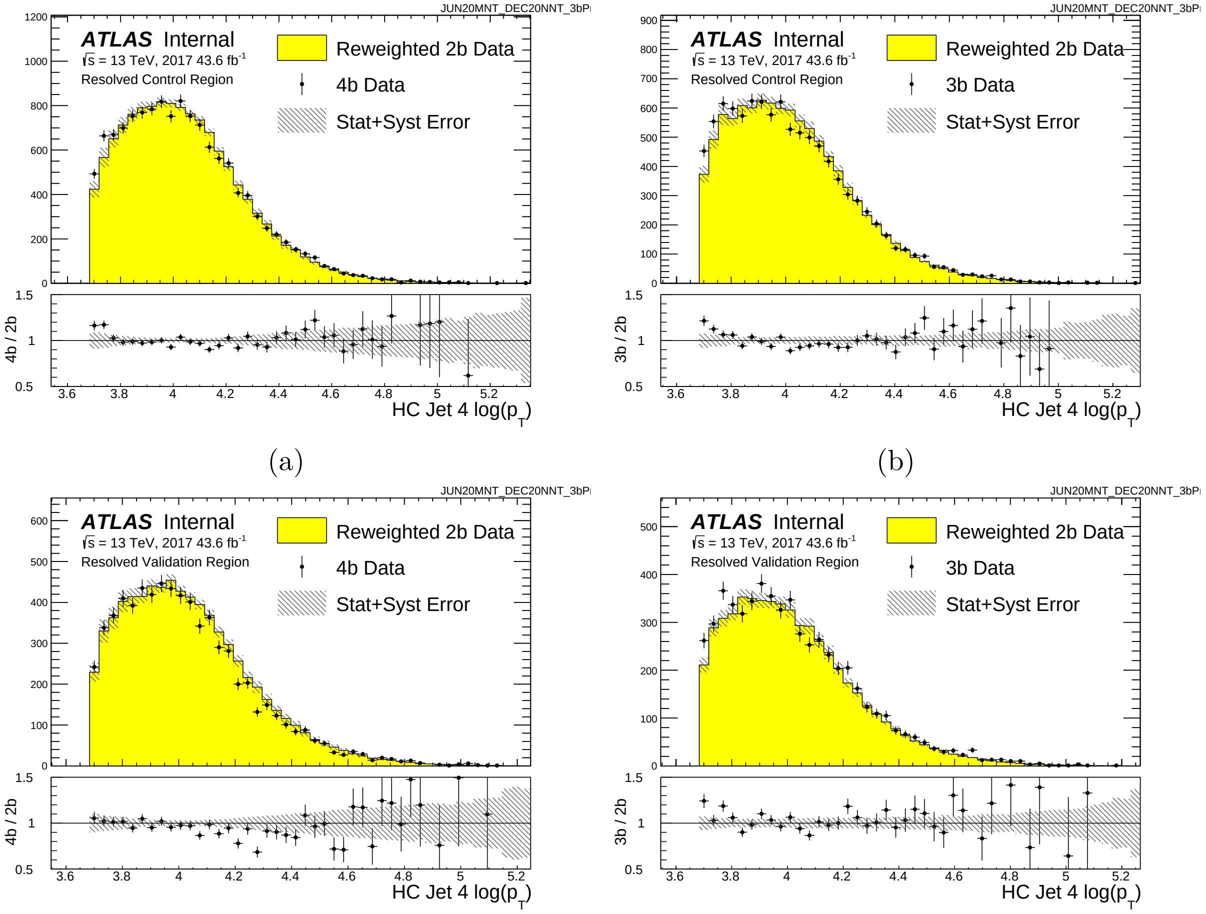
(a)

(b)

(c)

(d)

Figure A.37: Distributions of the fourth jet $\log p_T$ in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



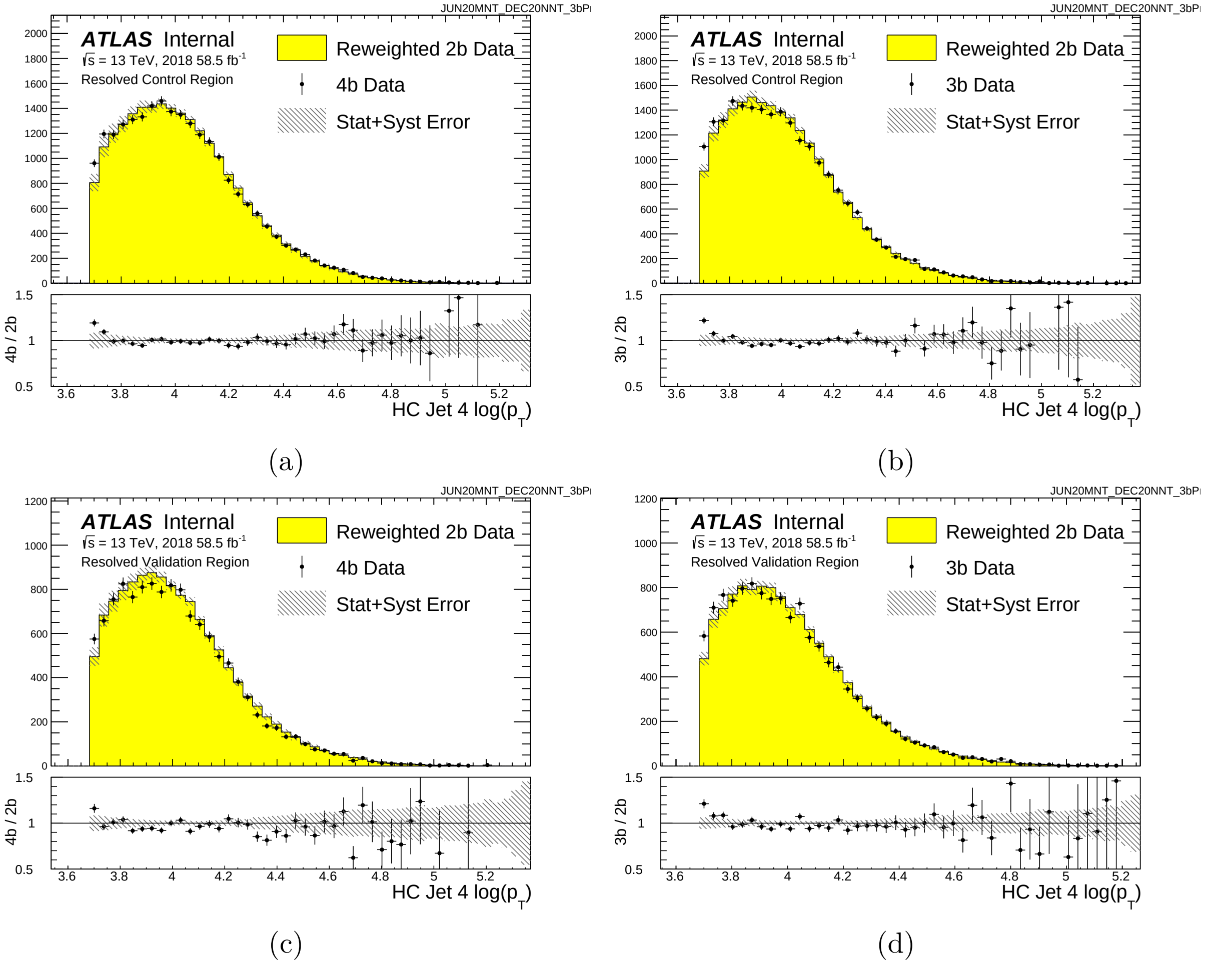
(a)

(b)

(c)

(d)

Figure A.38: Distributions of the fourth jet $\log p_T$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



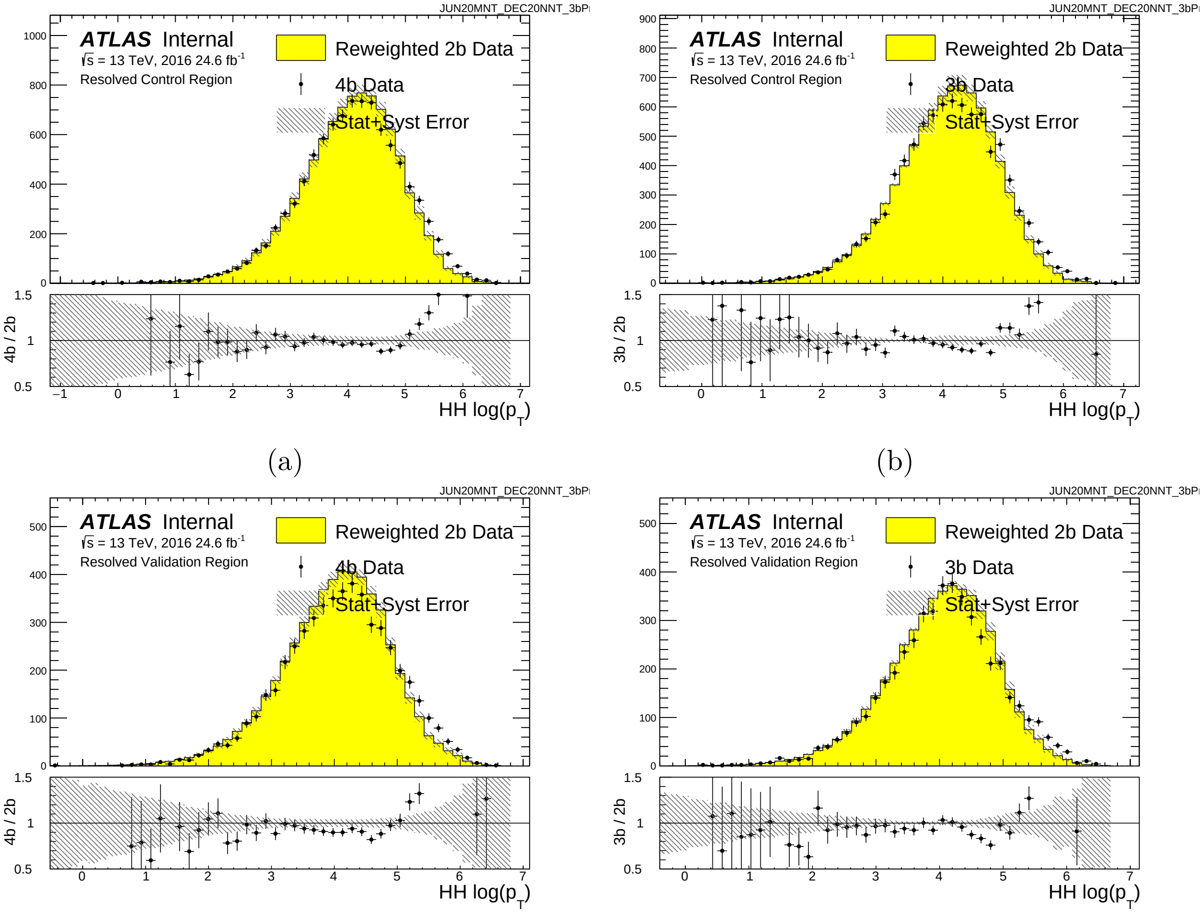
(a)

(b)

(c)

(d)

Figure A.39: Distributions of the fourth jet $\log p_T$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



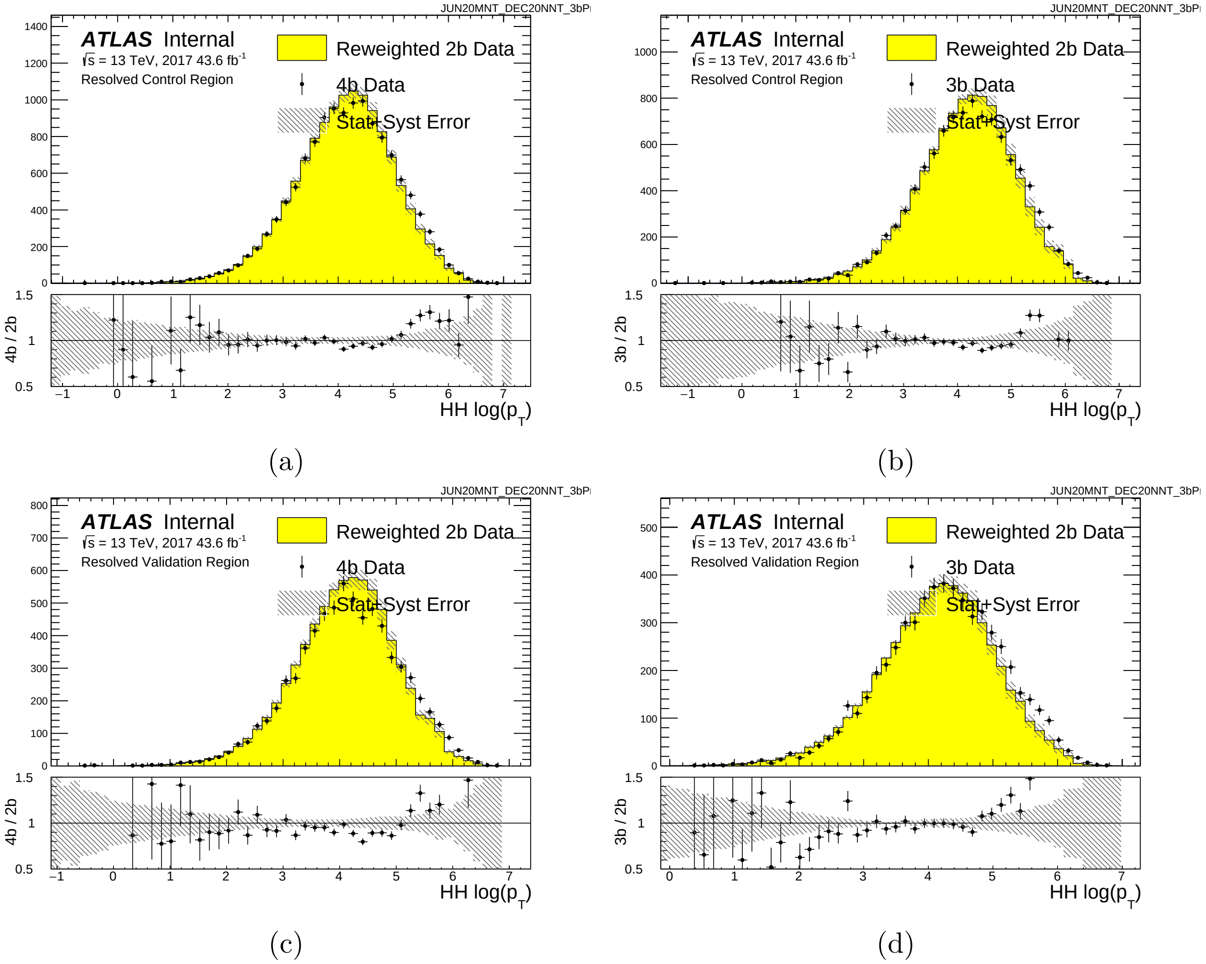
(a)

(b)

(c)

(d)

Figure A.40: Distributions of $\log p_T^{hh}$ in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



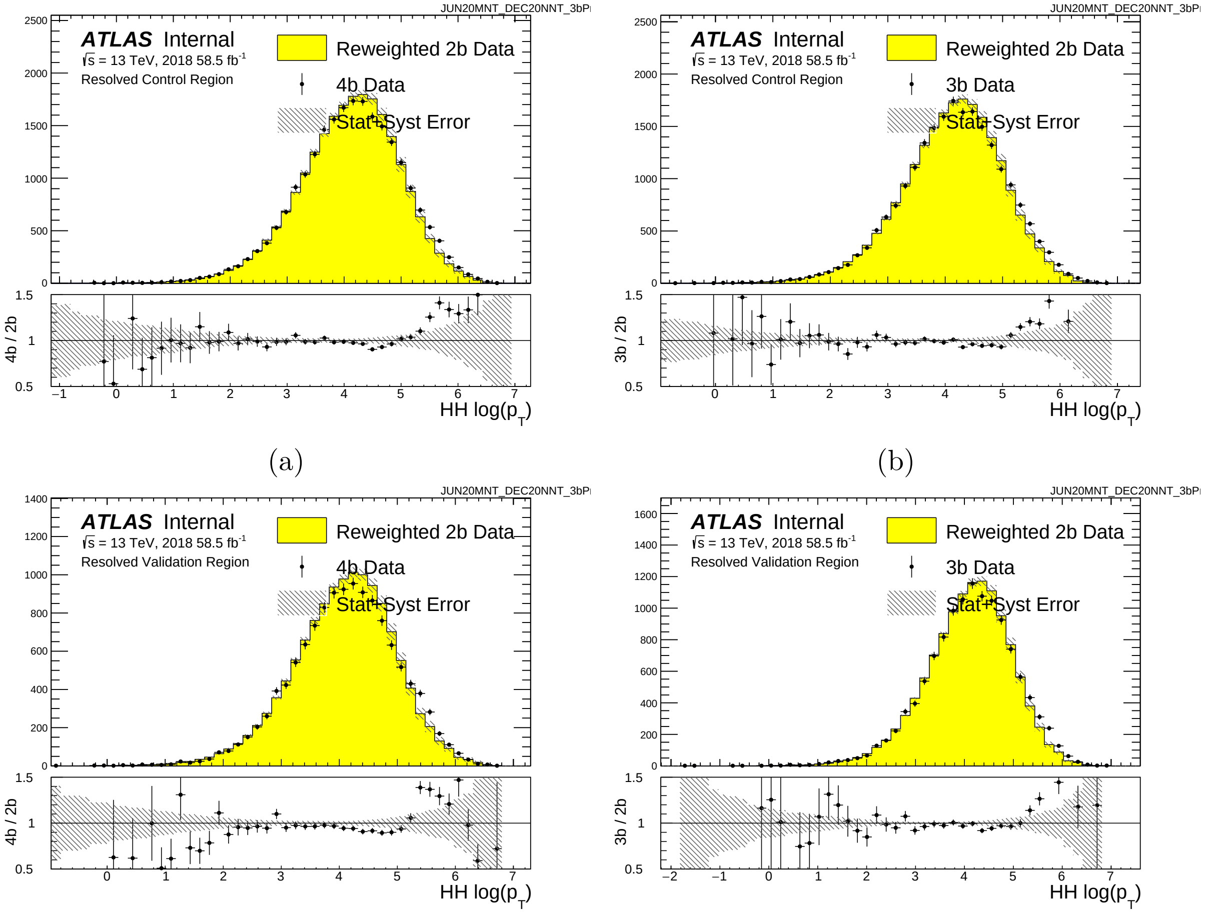
(a)

(b)

(c)

(d)

Figure A.41: Distributions of $\log p_T^{hh}$ in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



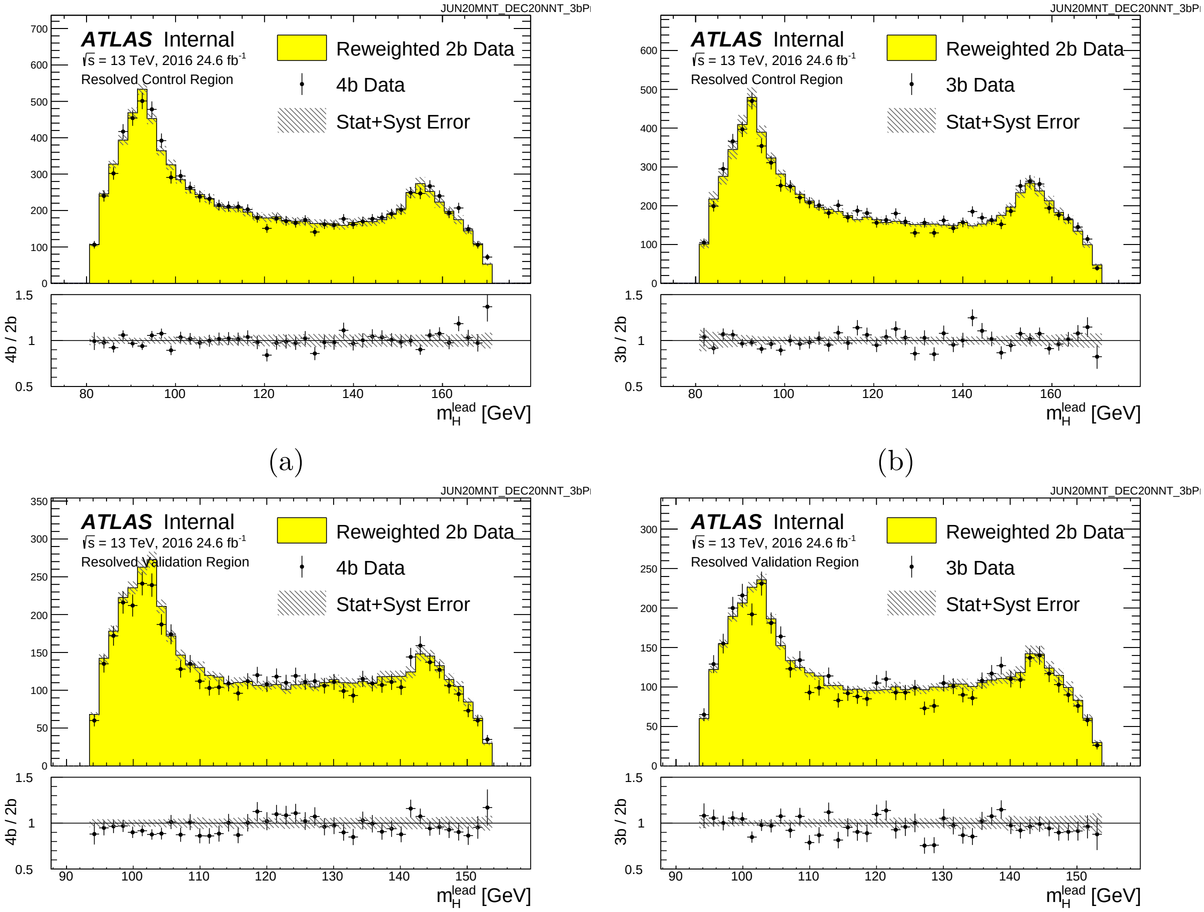
(a)

(b)

(c)

(d)

Figure A.42: Distributions of $\log p_T^{hh}$ in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



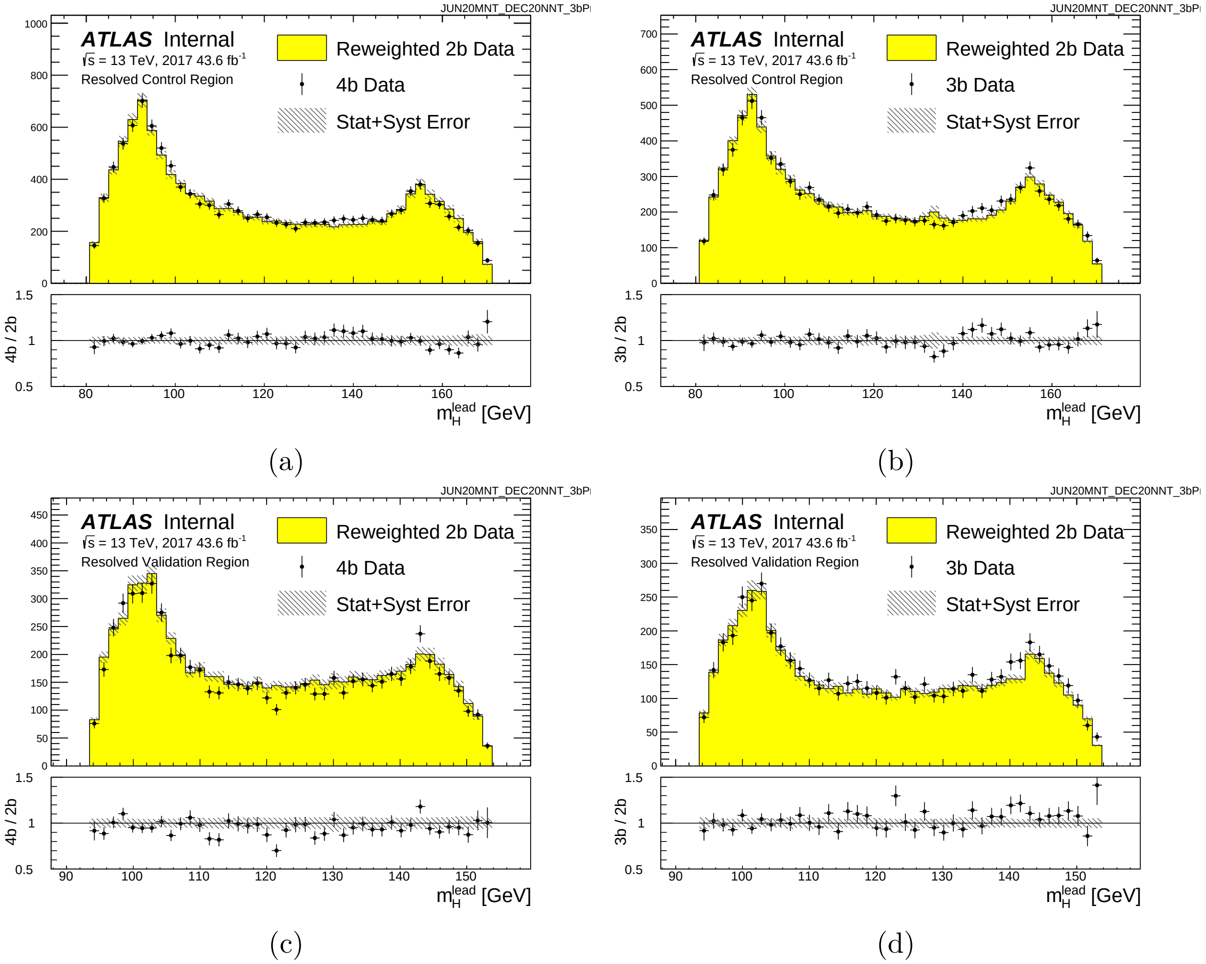
(a)

(b)

(c)

(d)

Figure A.43: Distributions of leading Higgs candidate mass in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



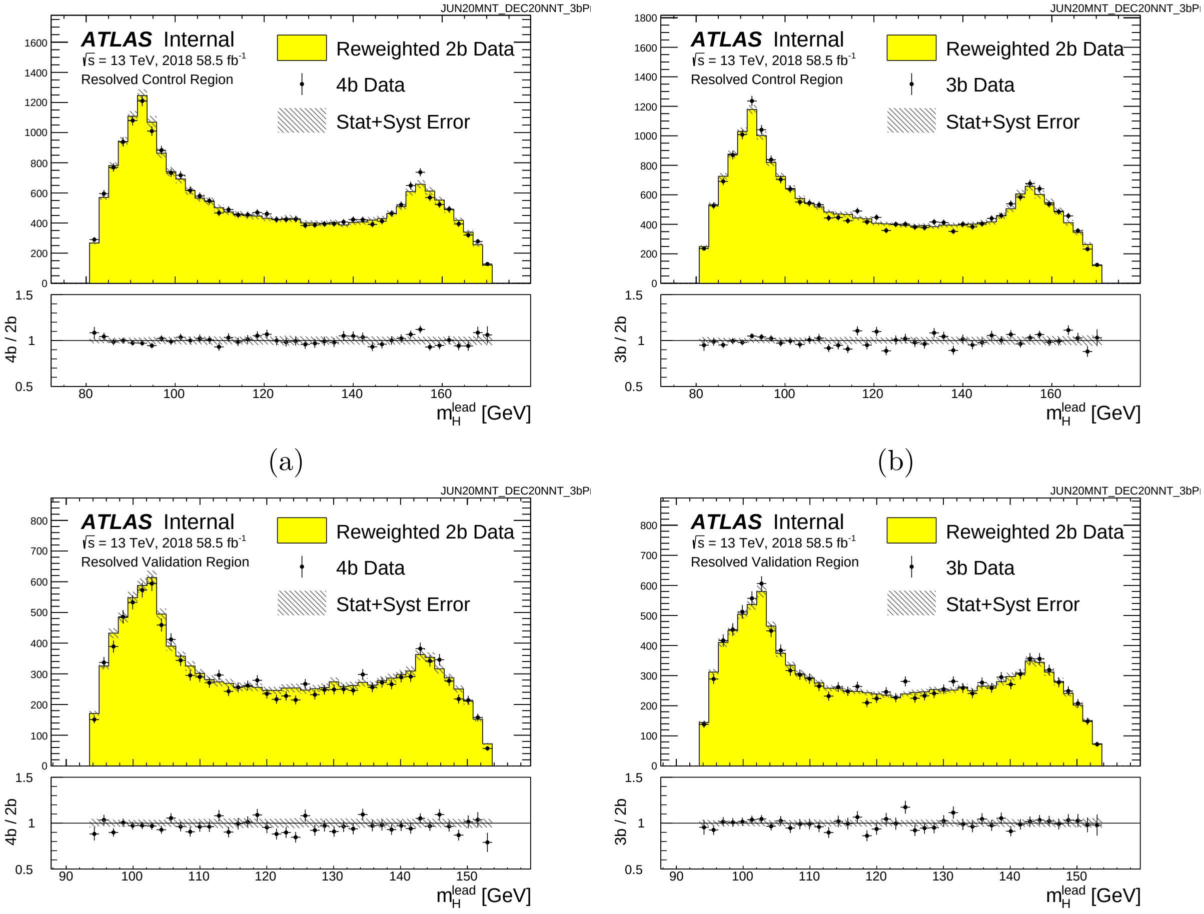
(a)

(b)

(c)

(d)

Figure A.44: Distributions of leading Higgs candidate mass in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



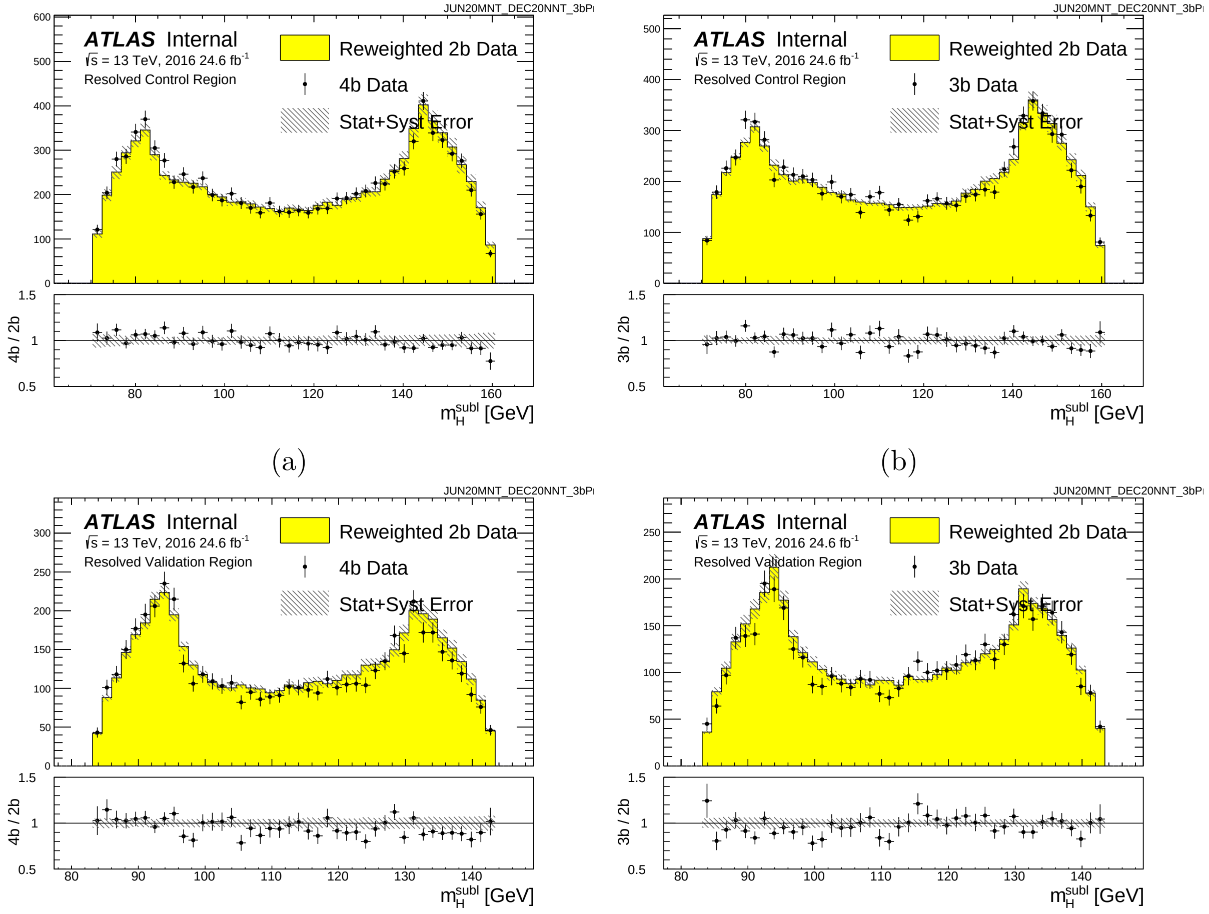
(a)

(b)

(c)

(d)

Figure A.45: Distributions of leading Higgs candidate mass in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



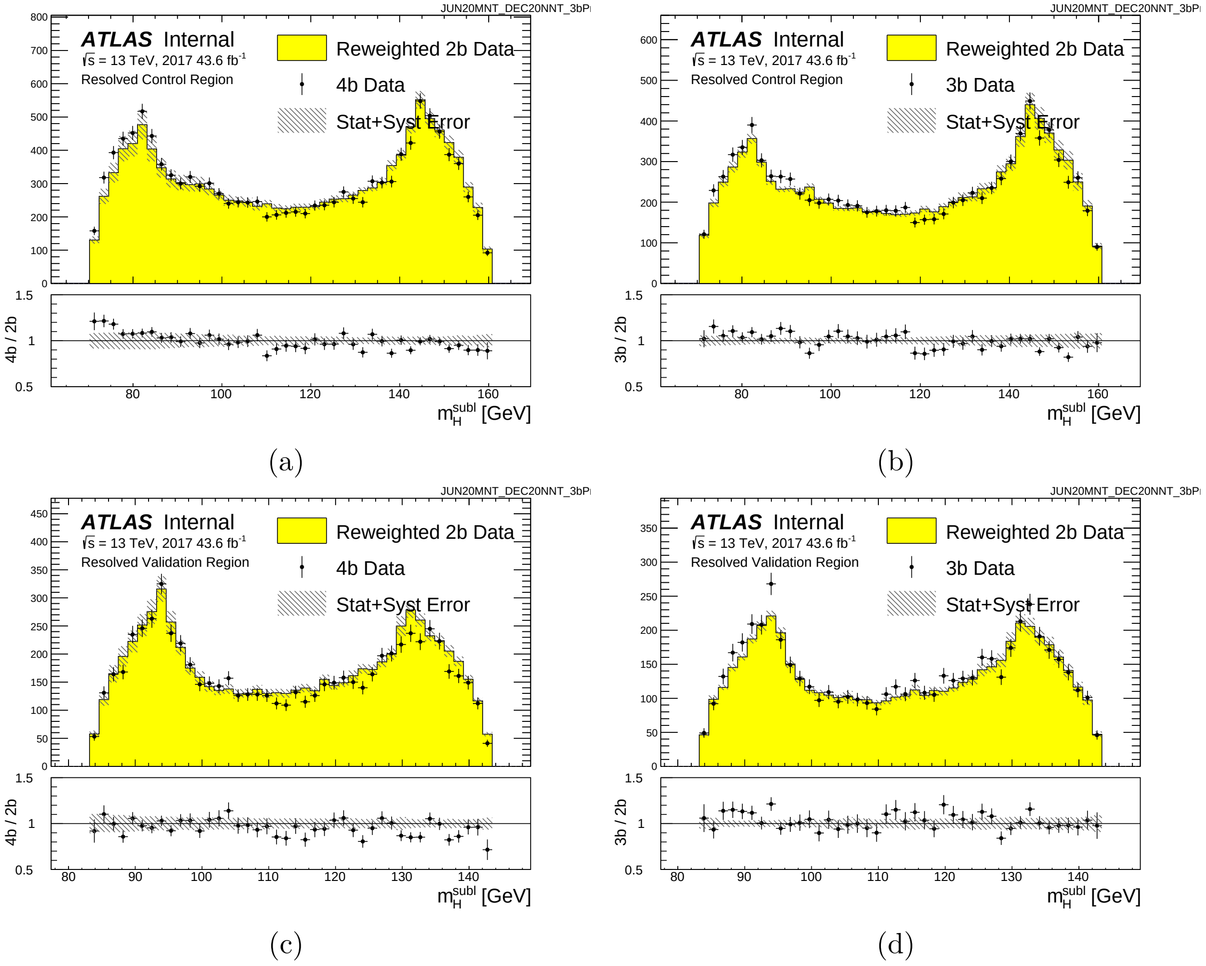
(a)

(b)

(c)

(d)

Figure A.46: Distributions of subleading Higgs candidate massin 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



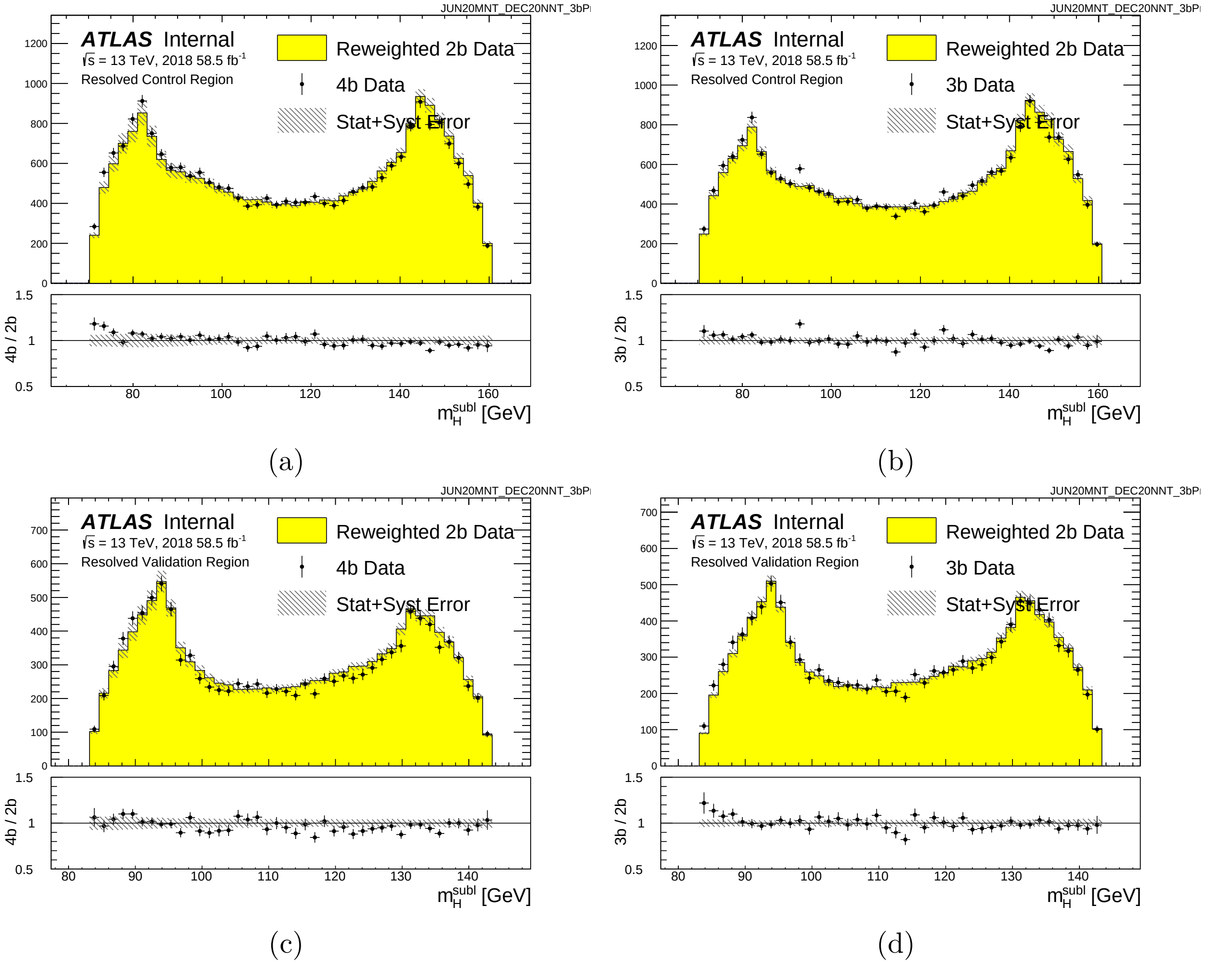
(a)

(b)

(c)

(d)

Figure A.47: Distributions of subleading Higgs candidate massin 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



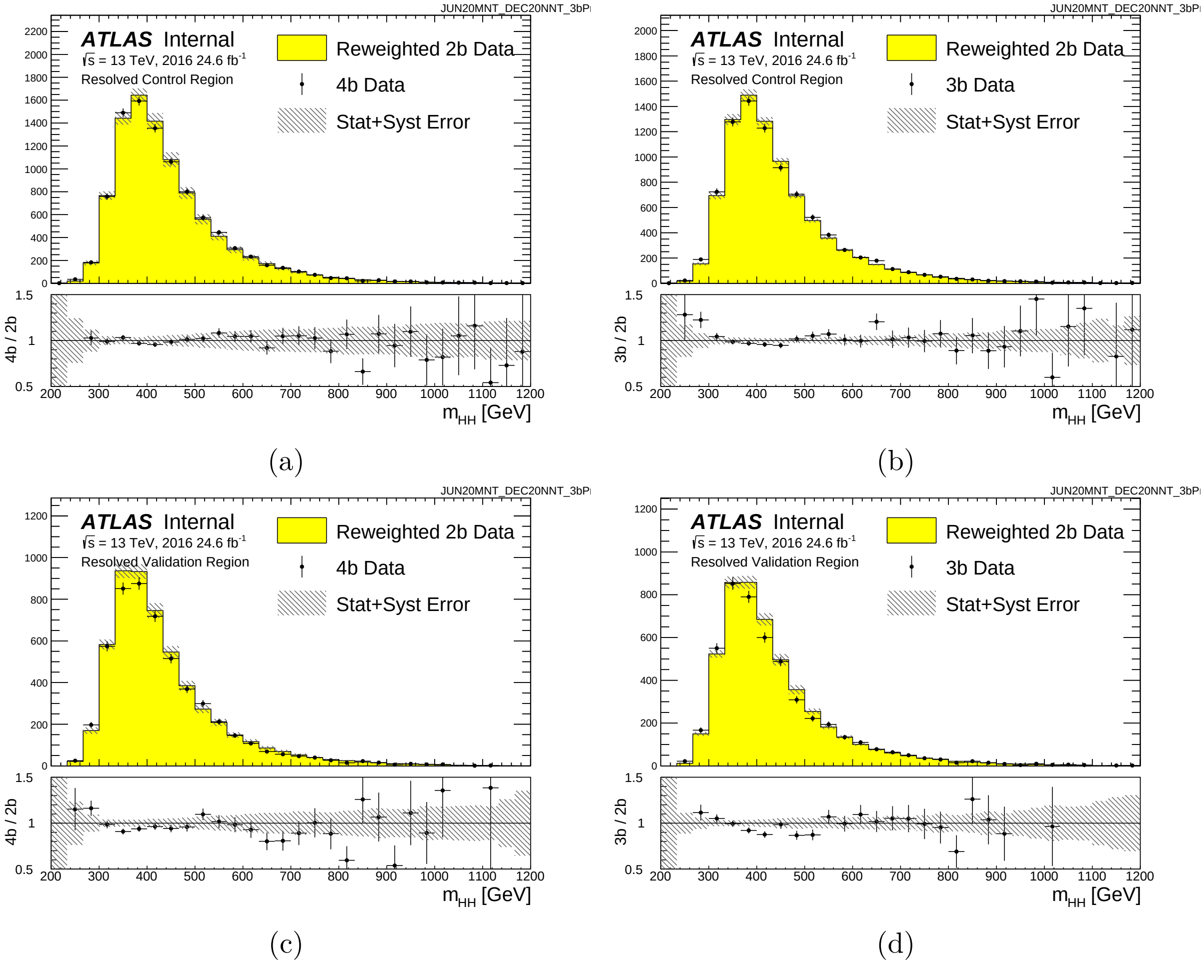
(a)

(b)

(c)

(d)

Figure A.48: Distributions of subleading Higgs candidate massin 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



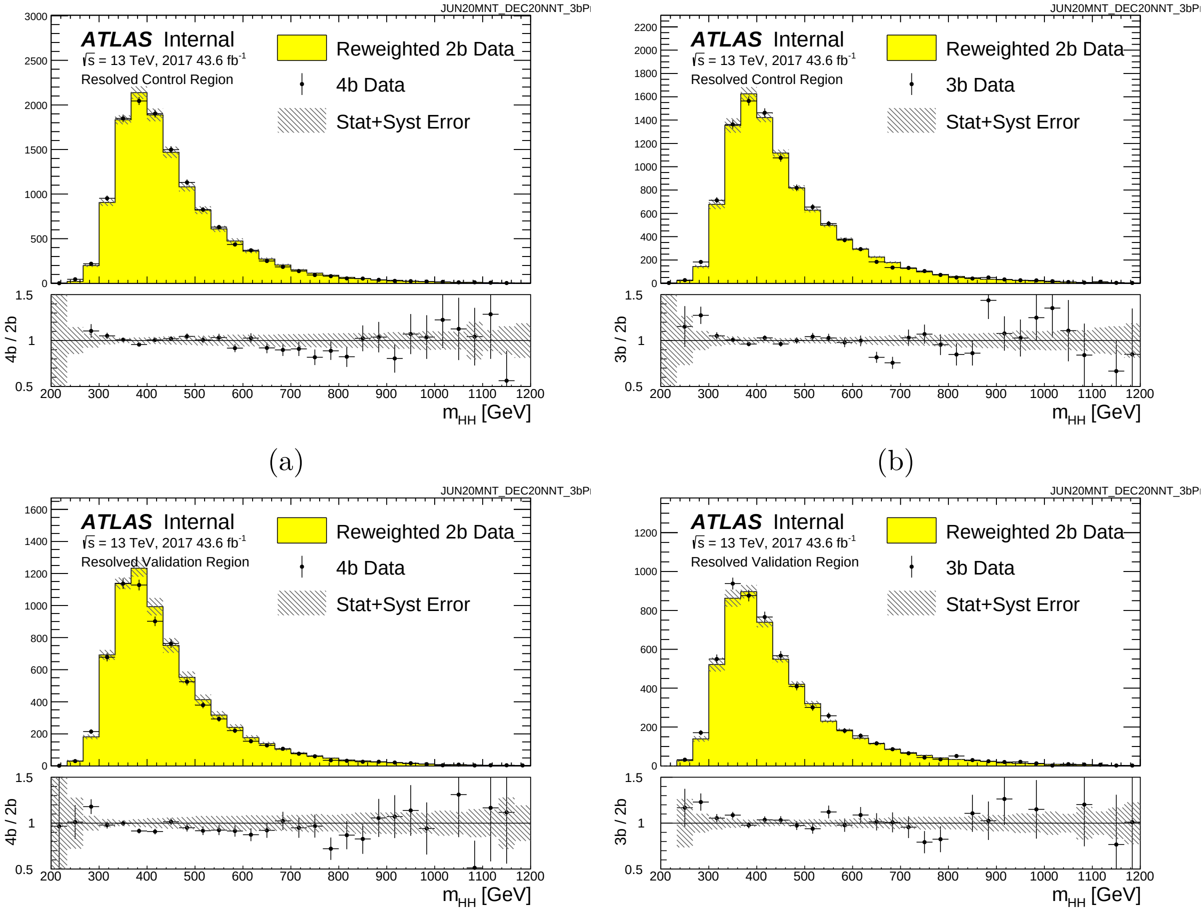
(a)

(b)

(c)

(d)

Figure A.49: Distributions of four-jet invariant mass in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



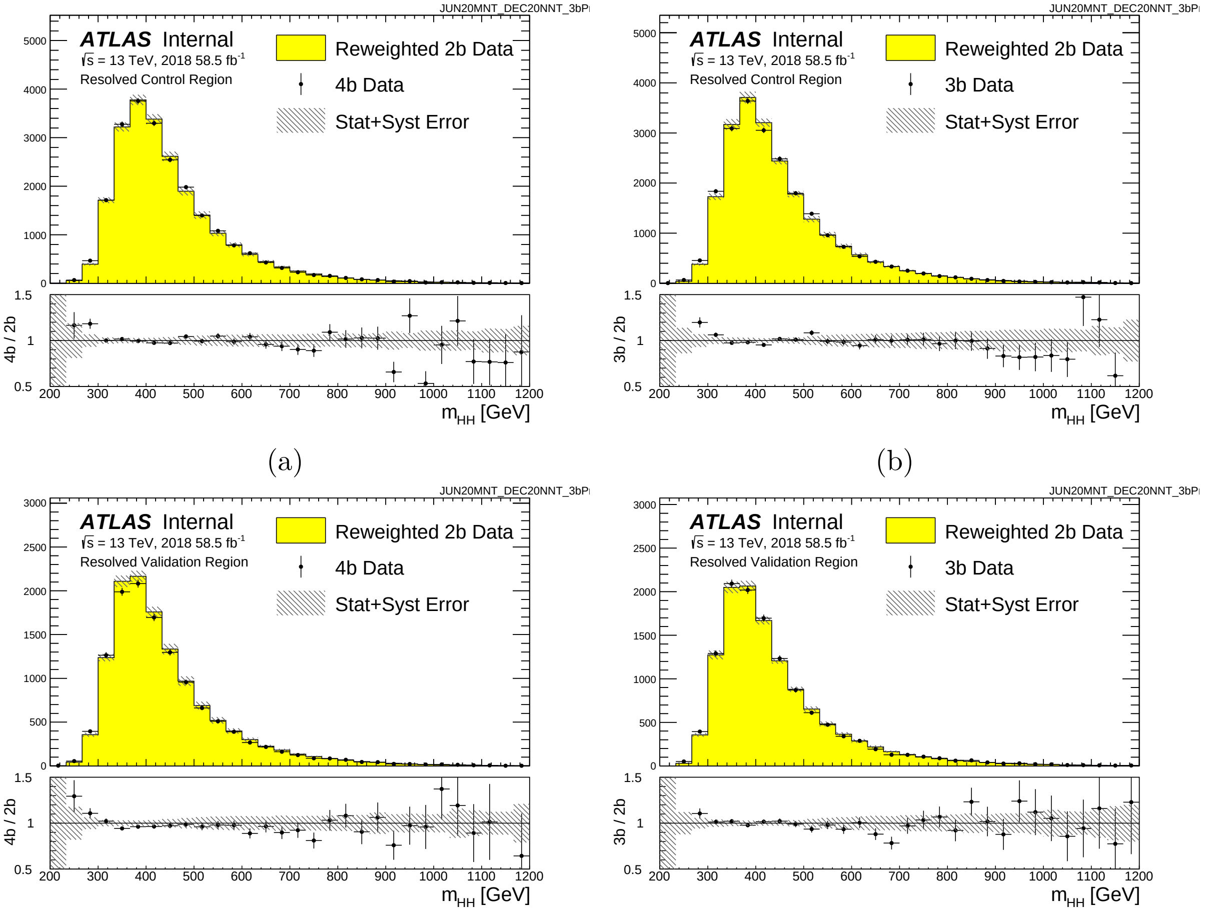
(a)

(b)

(c)

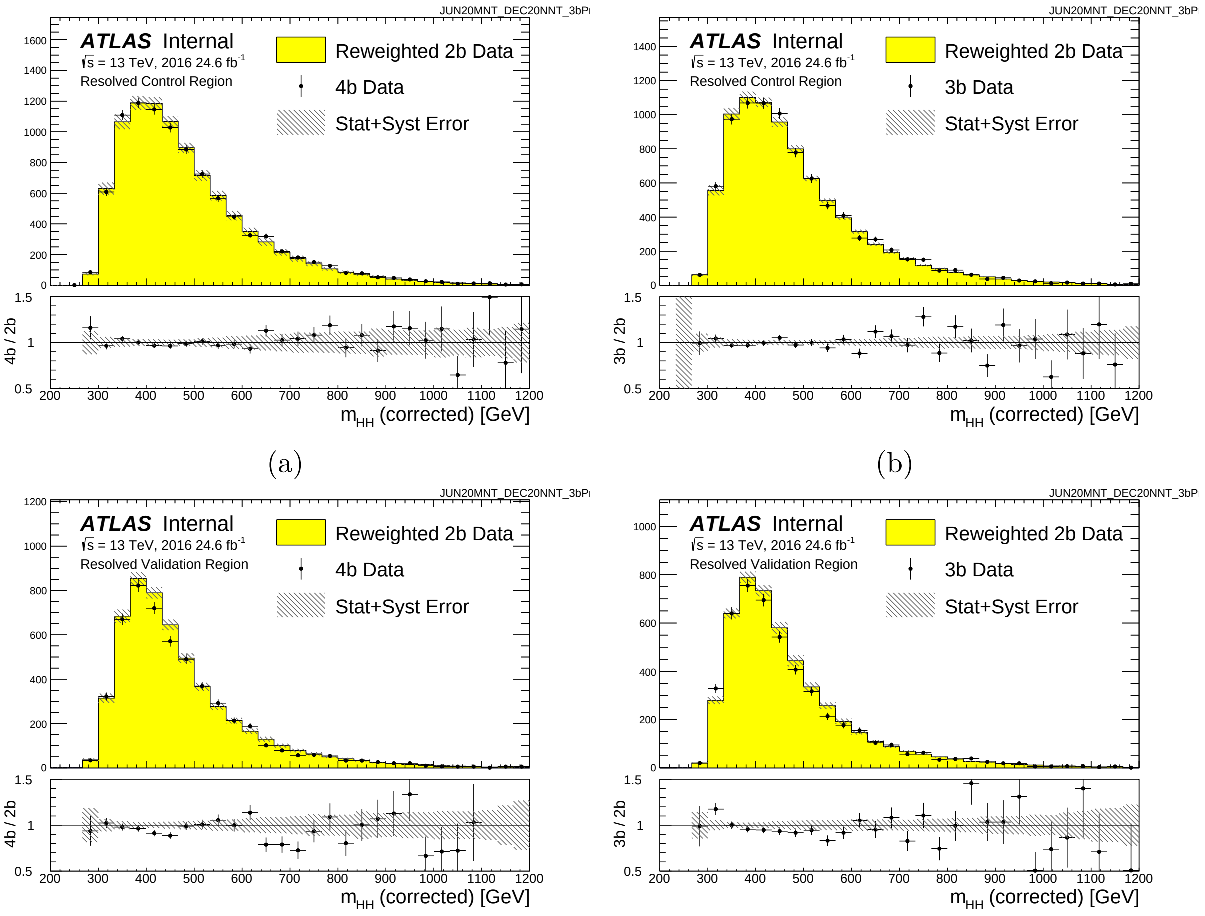
(d)

Figure A.50: Distributions of four-jet invariant mass in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b) (c) (d)

Figure A.51: Distributions of four-jet invariant mass in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



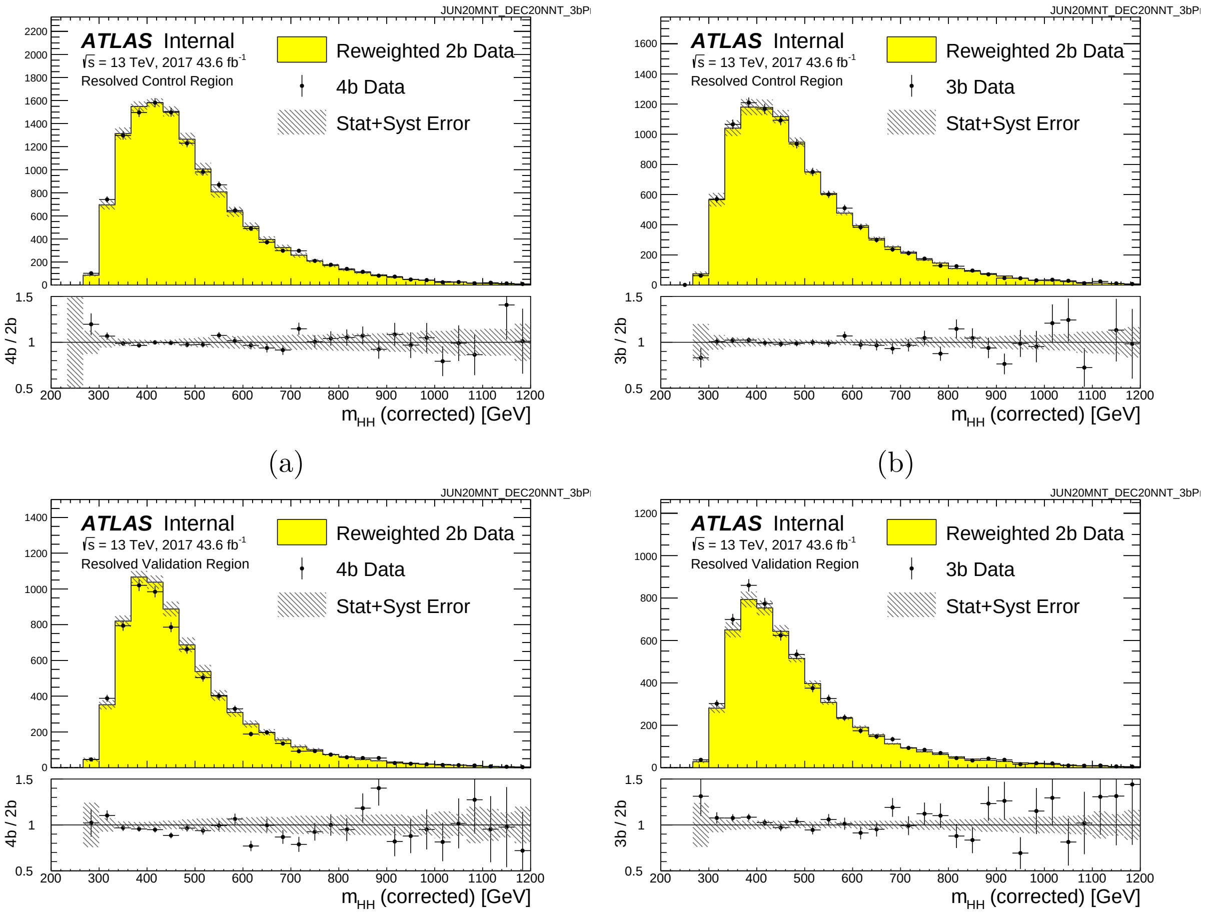
(a)

(b)

(c)

(d)

Figure A.52: Distributions of four-jet invariant mass with the Higgs candidate masses scaled to 125 GeV in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



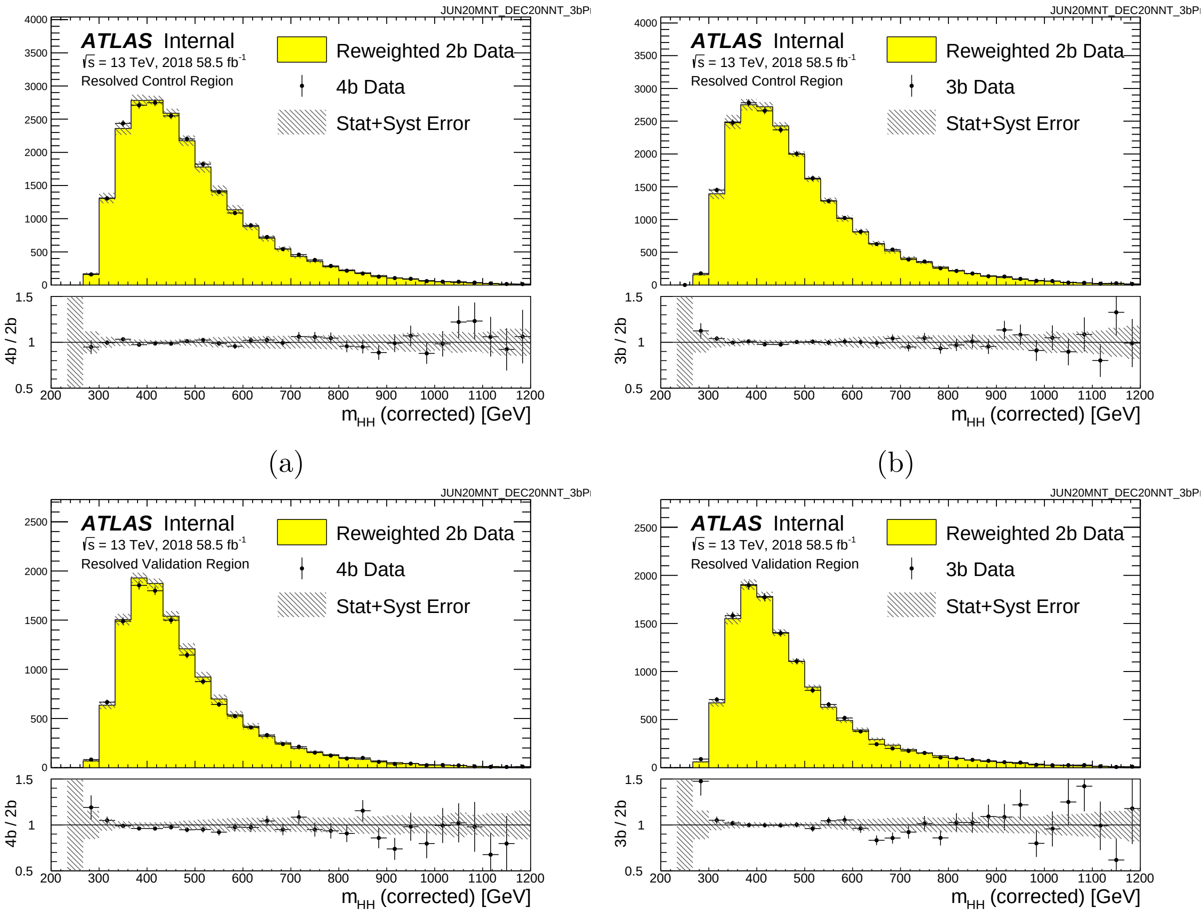
(a)

(b)

(c)

(d)

Figure A.53: Distributions of four-jet invariant mass with the Higgs candidate masses scaled to 125 GeV in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



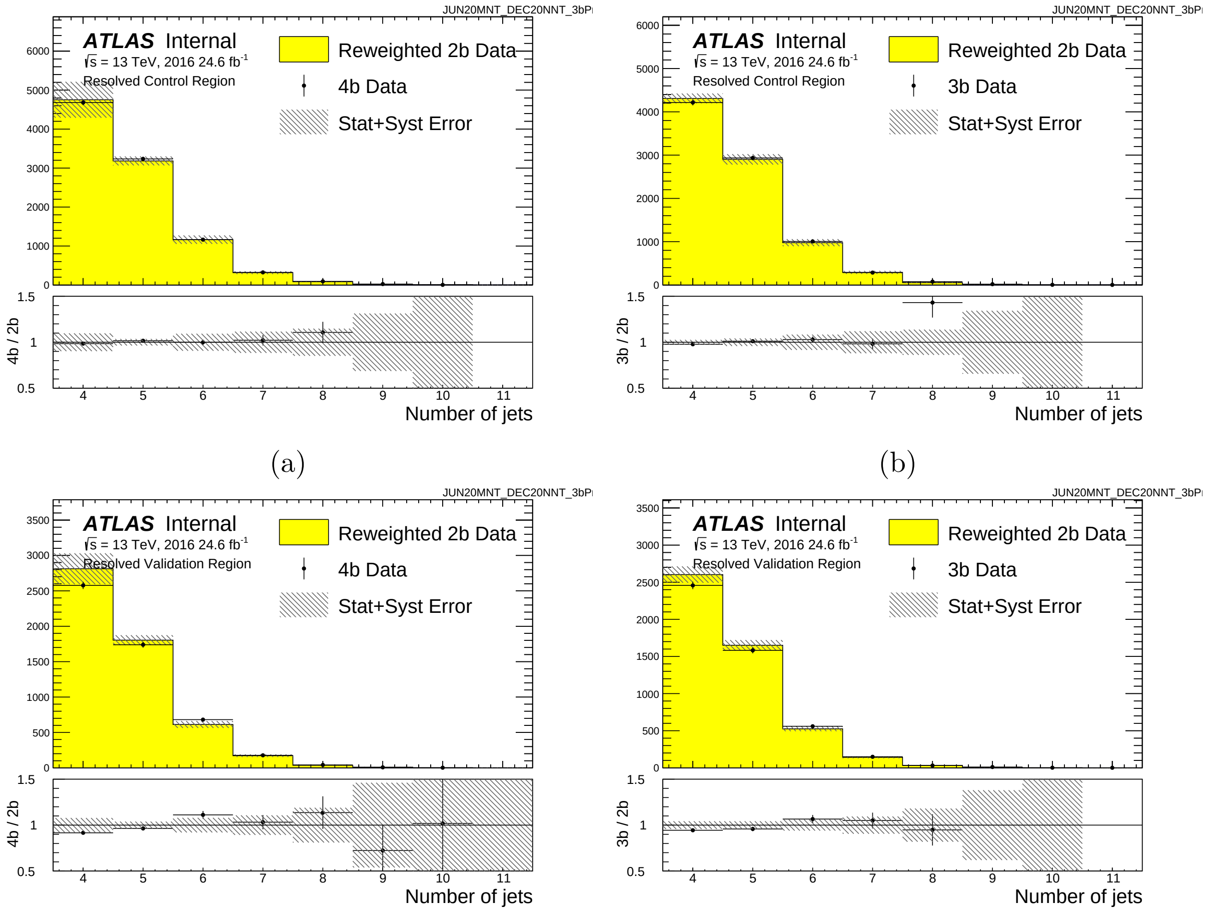
(a)

(b)

(c)

(d)

Figure A.54: Distributions of four-jet invariant mass with the Higgs candidate masses scaled to 125 GeV in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



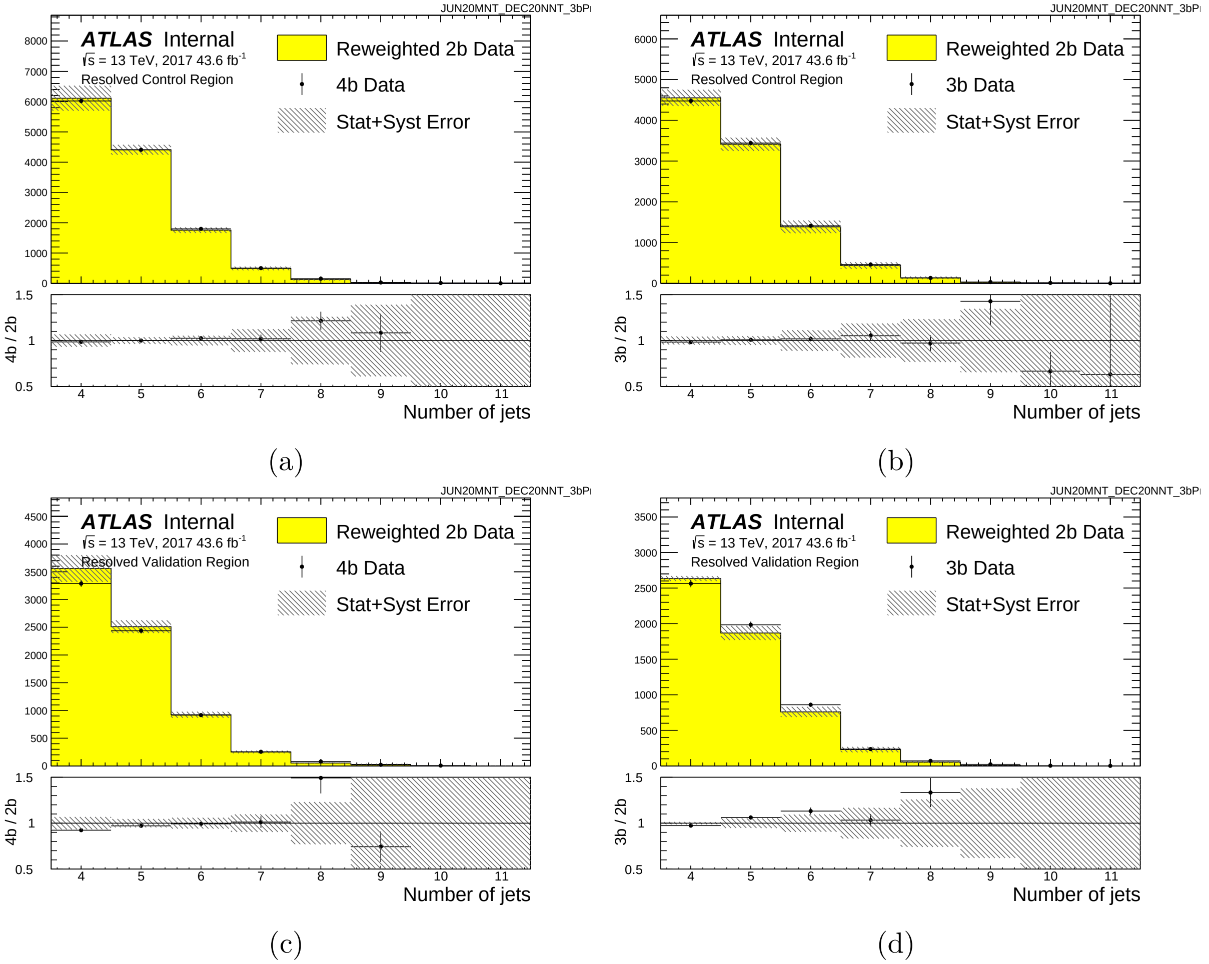
(a)

(b)

(c)

(d)

Figure A.55: Distributions of the number of jets in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



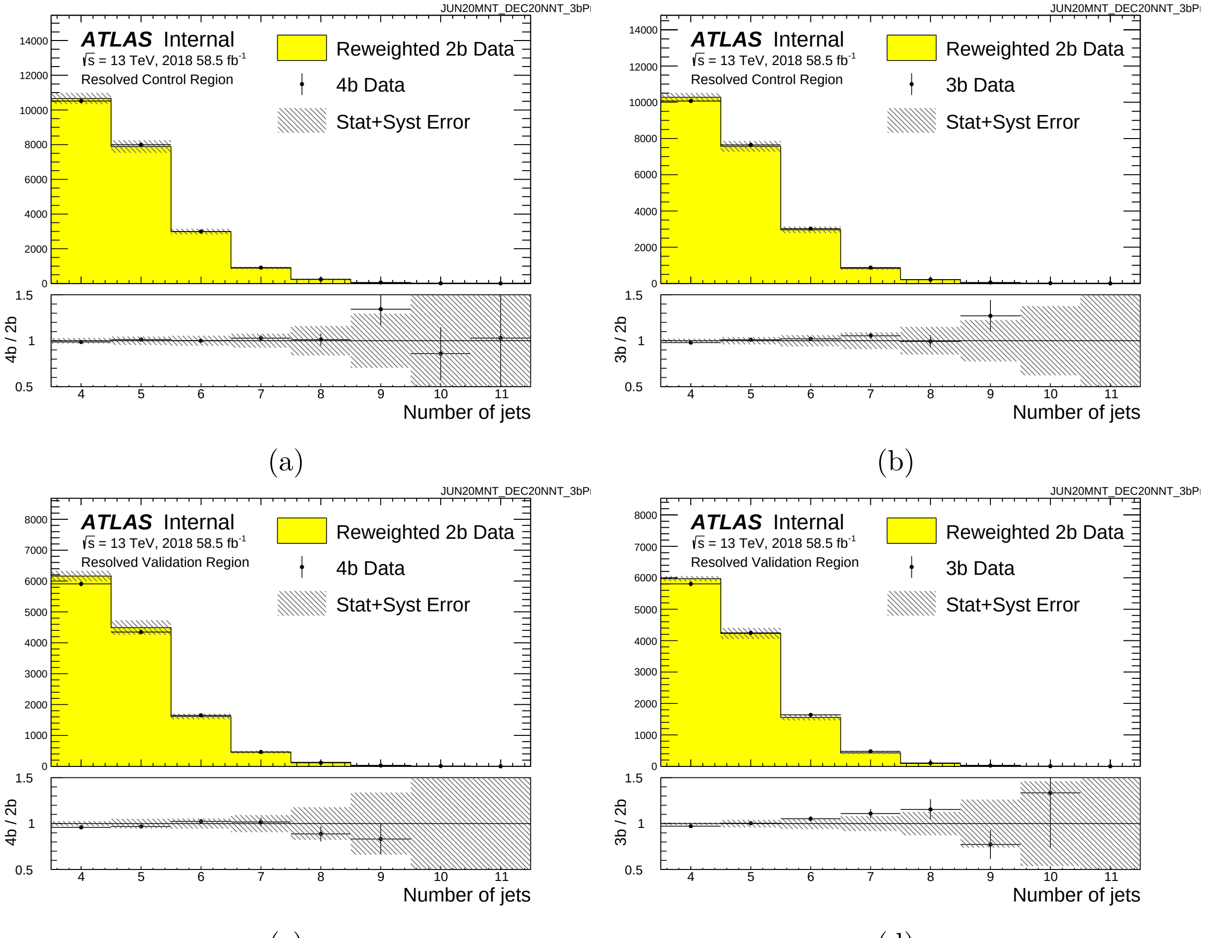
(a)

(b)

(c)

(d)

Figure A.56: Distributions of the number of jets in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



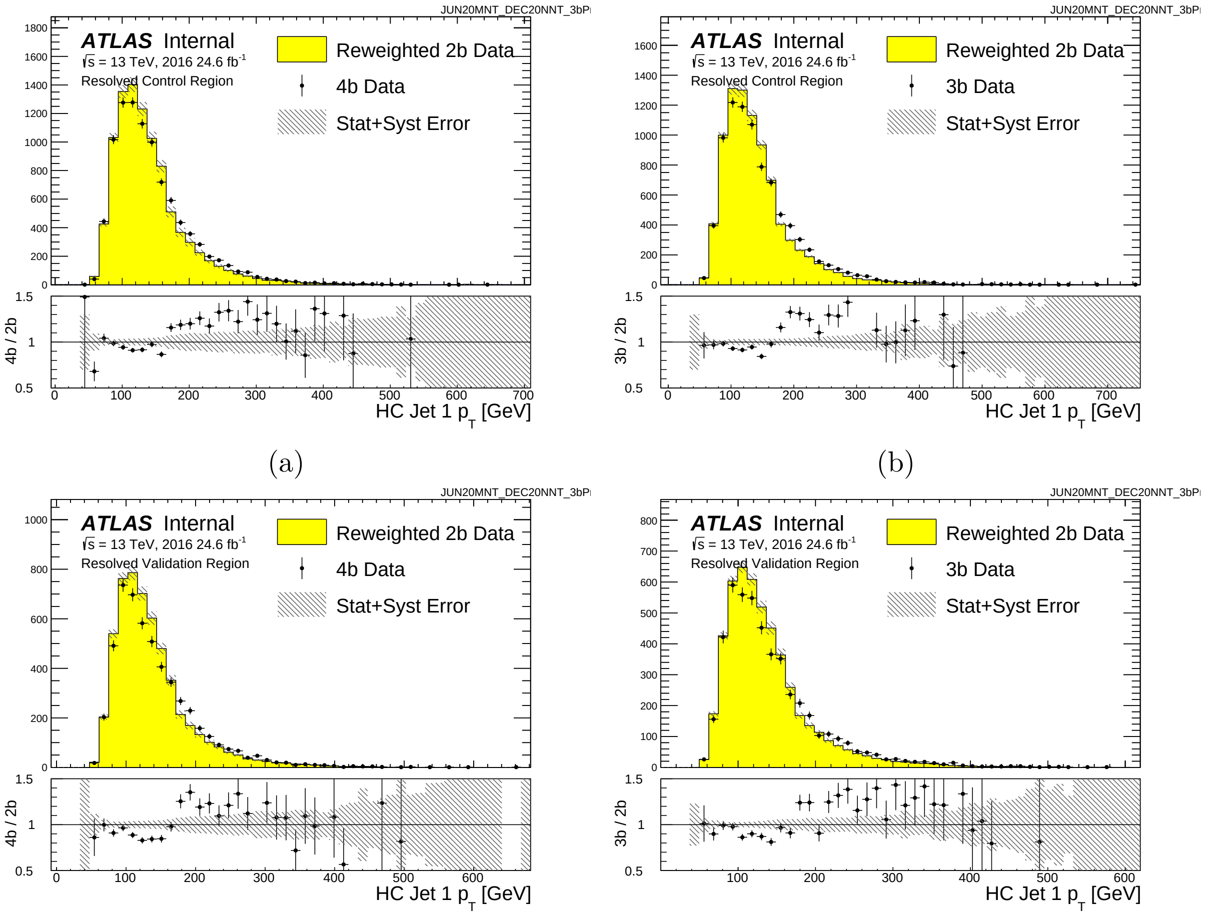
(a)

(b)

(c)

(d)

Figure A.57: Distributions of the number of jets in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



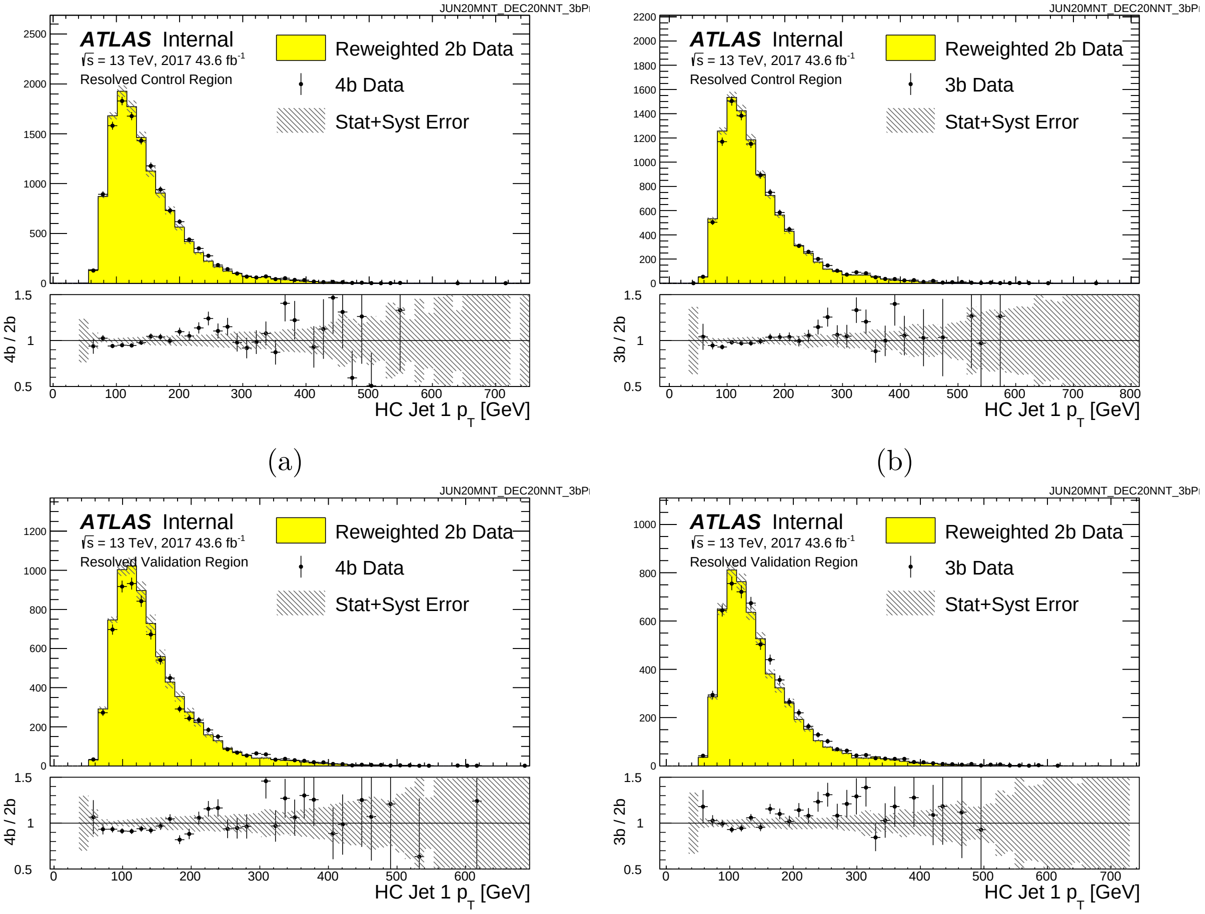
(a)

(b)

(c)

(d)

Figure A.58: Distributions of the first jet p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a)

(b)

(c)

(d)

Figure A.59: Distributions of the first jet p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

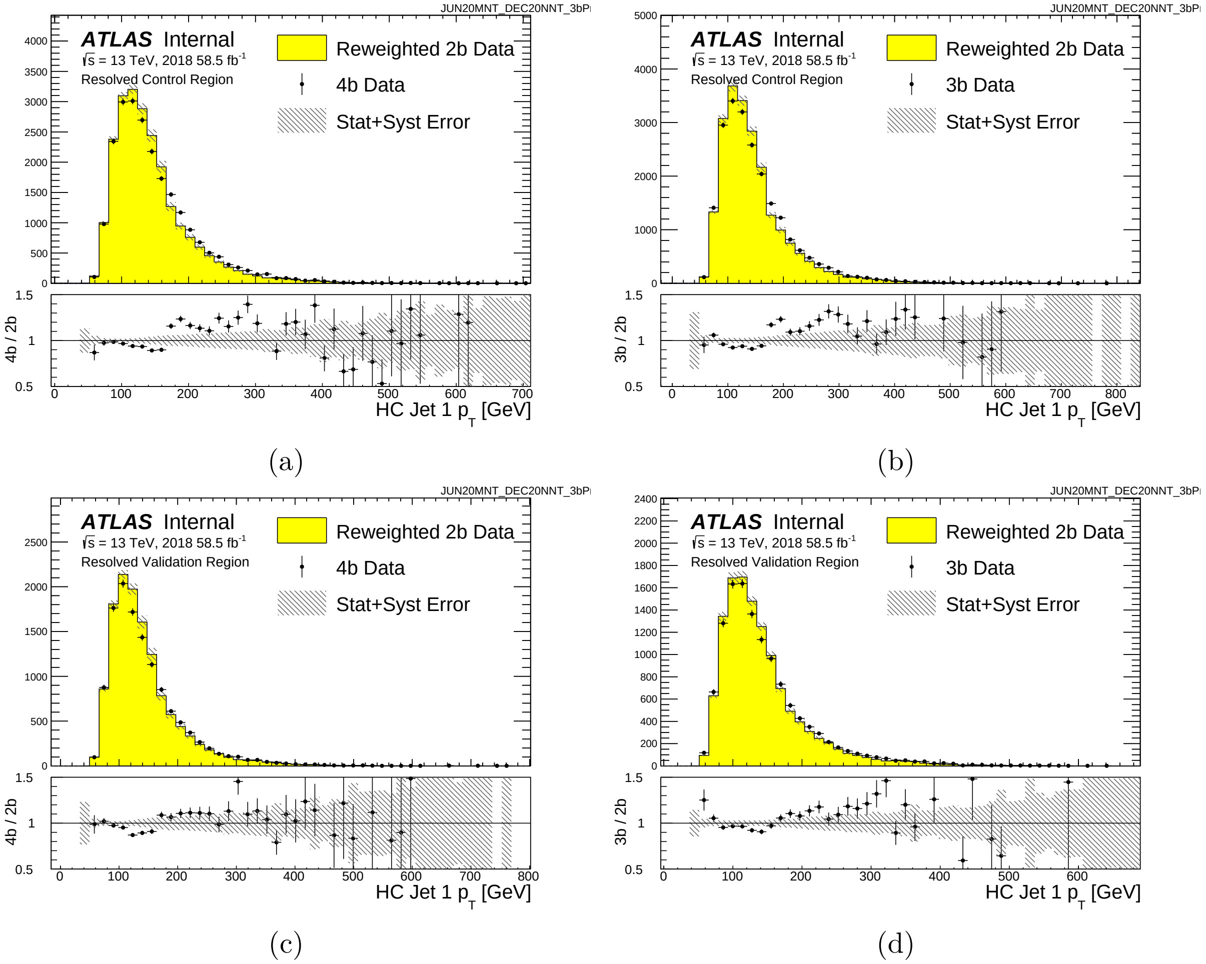
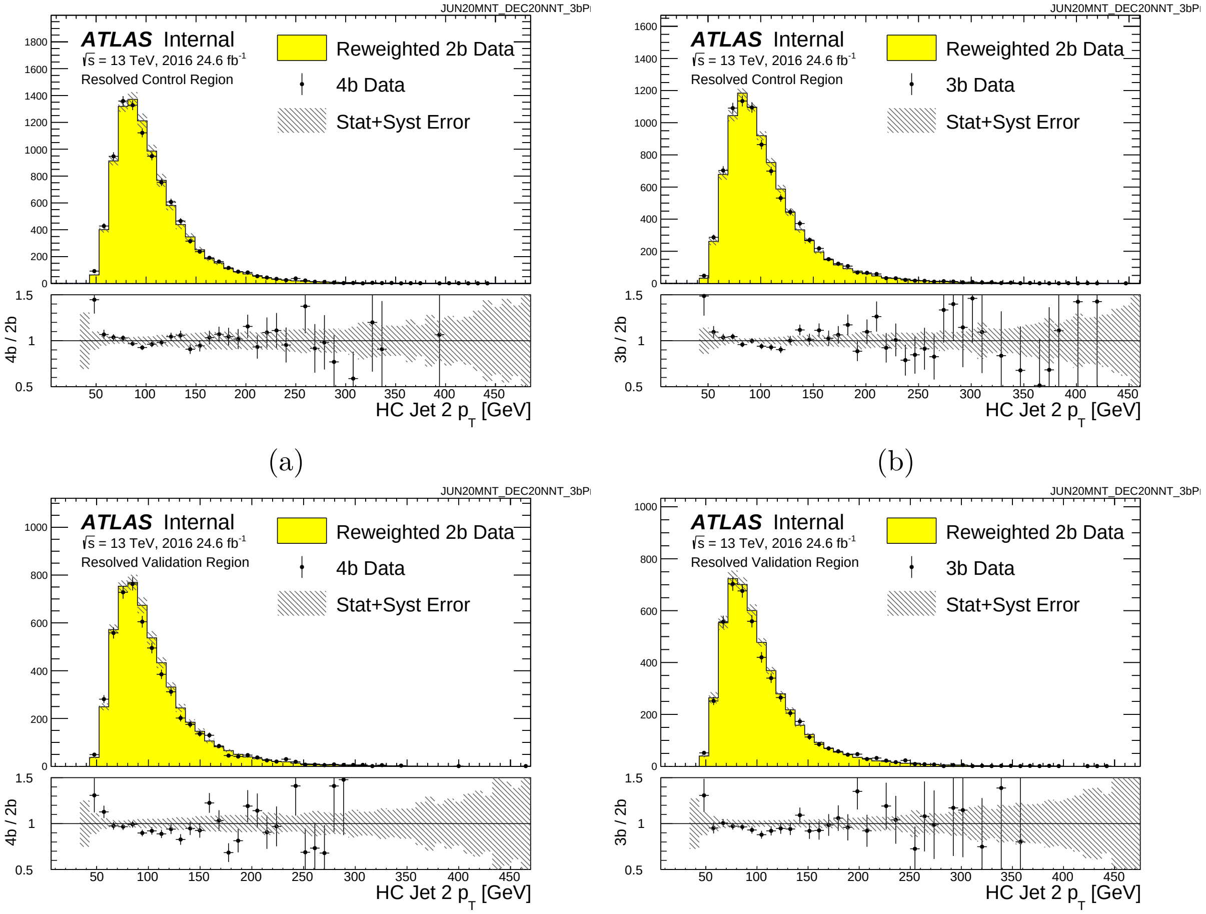


Figure A.60: Distributions of the first jet p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



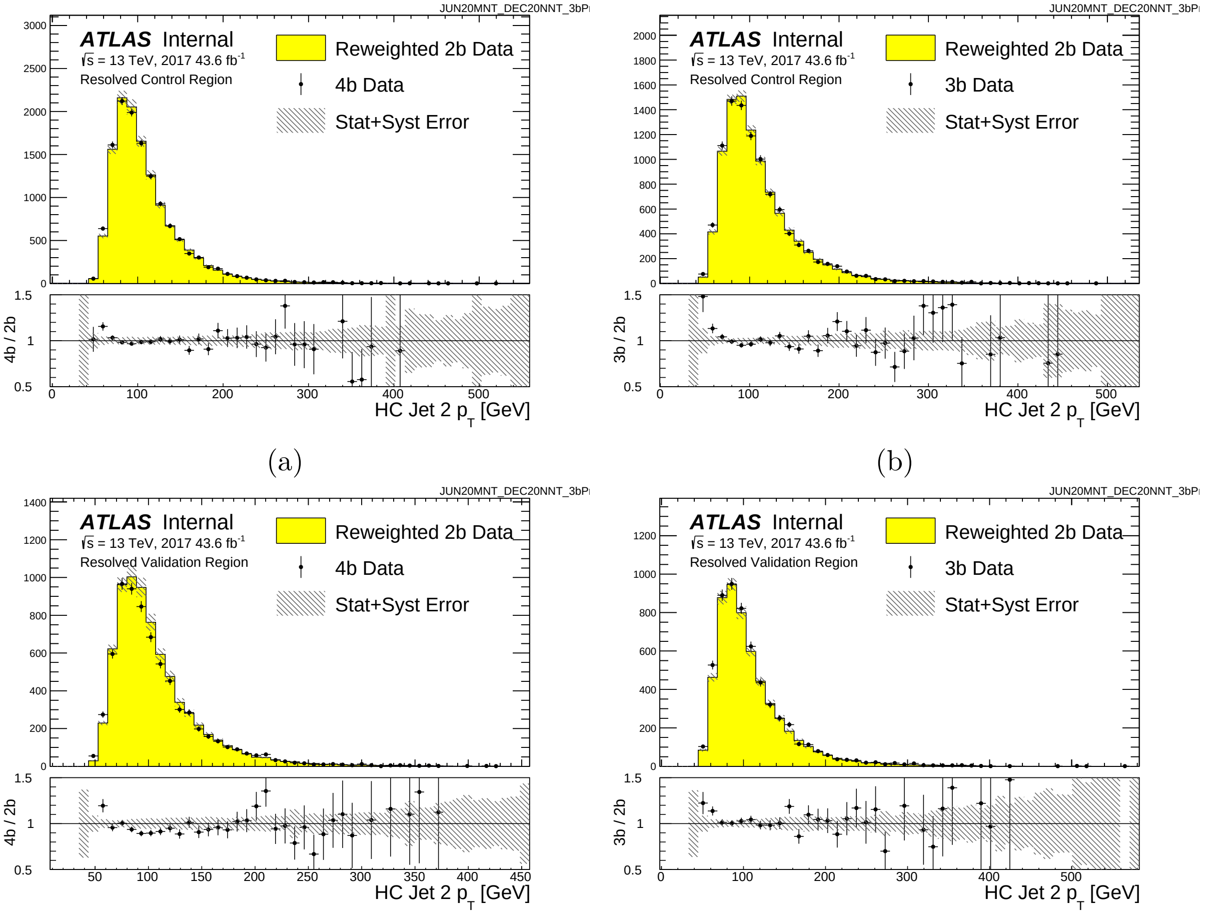
(a)

(b)

(c)

(d)

Figure A.61: Distributions of the second jet p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



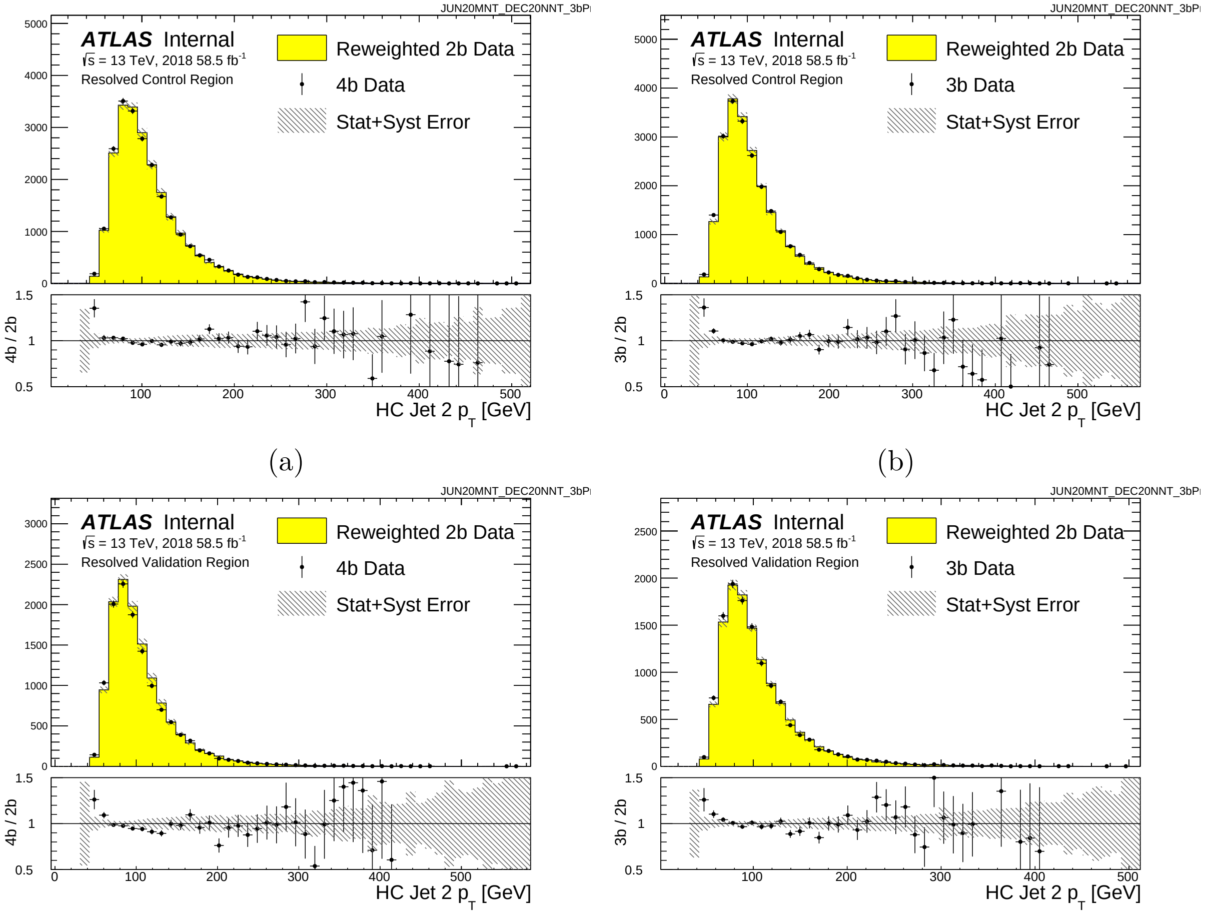
(a)

(b)

(c)

(d)

Figure A.62: Distributions of the second jet p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



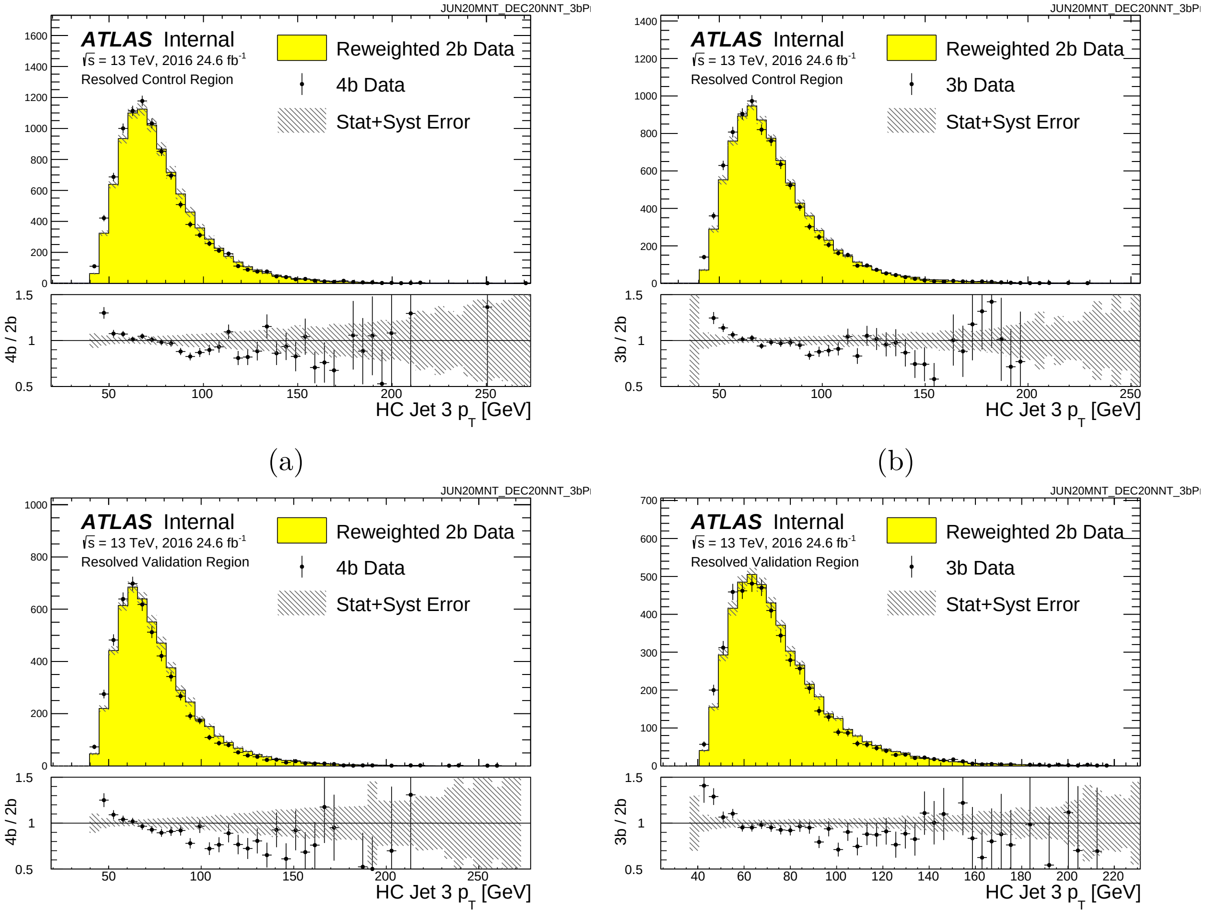
(a)

(b)

(c)

(d)

Figure A.63: Distributions of the second jet p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



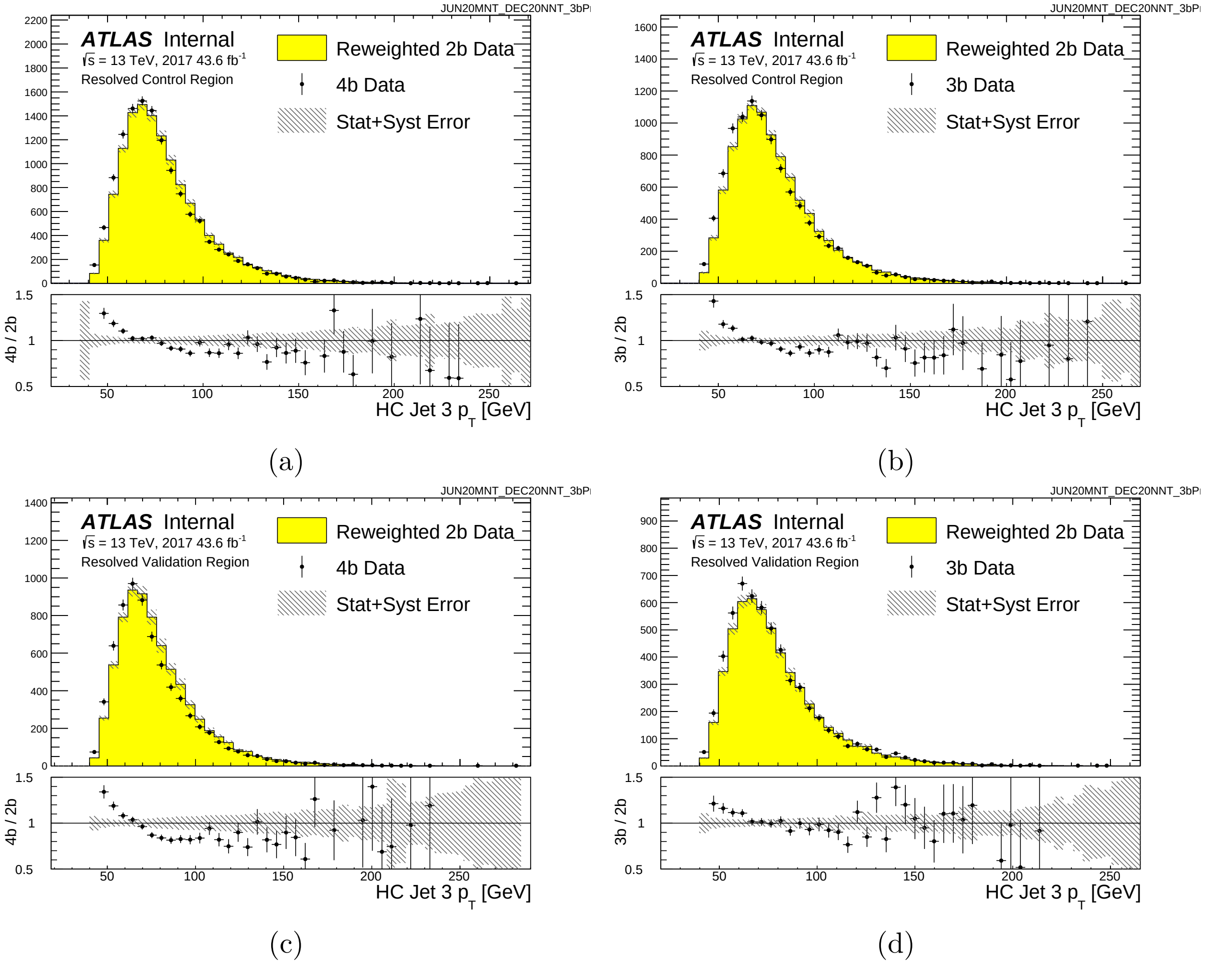
(a)

(b)

(c)

(d)

Figure A.64: Distributions of the third jet p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



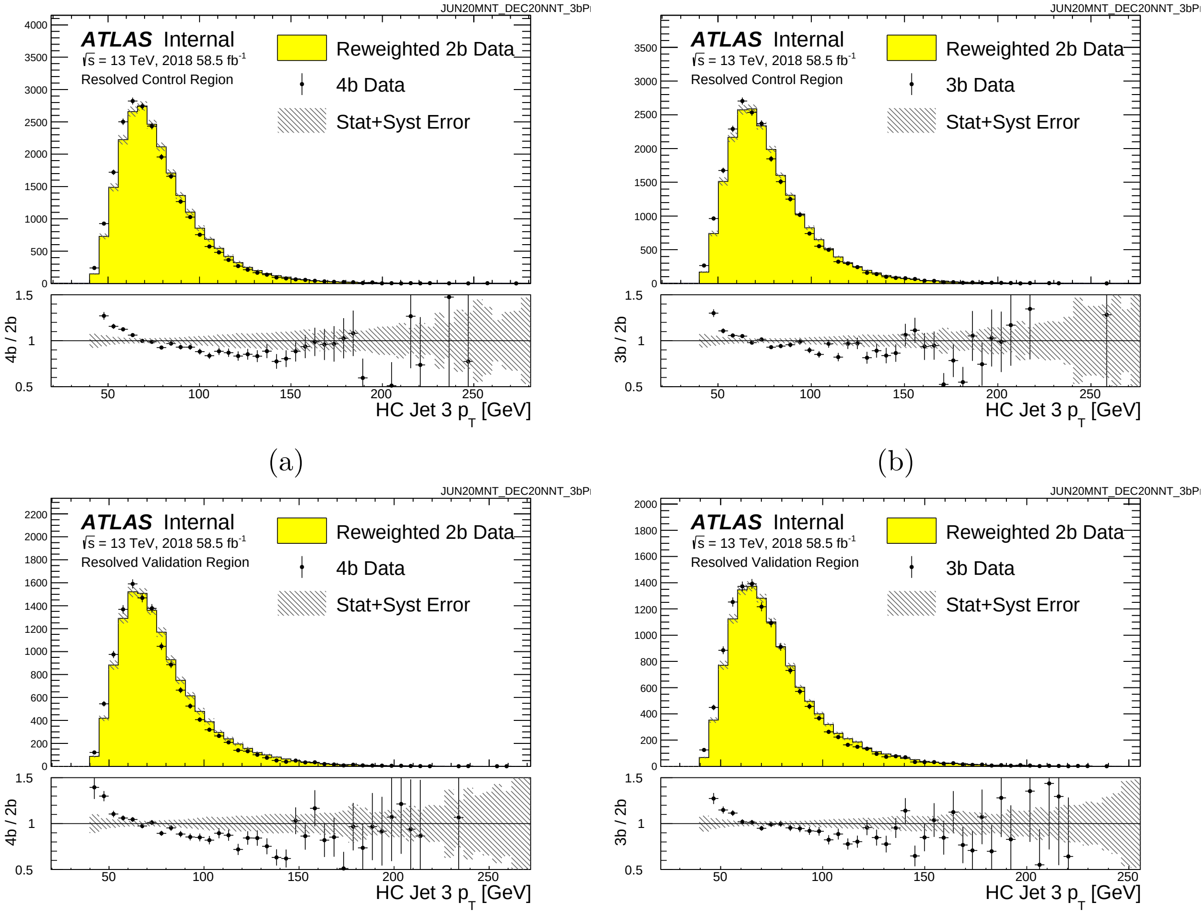
(a)

(b)

(c)

(d)

Figure A.65: Distributions of the third jet p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



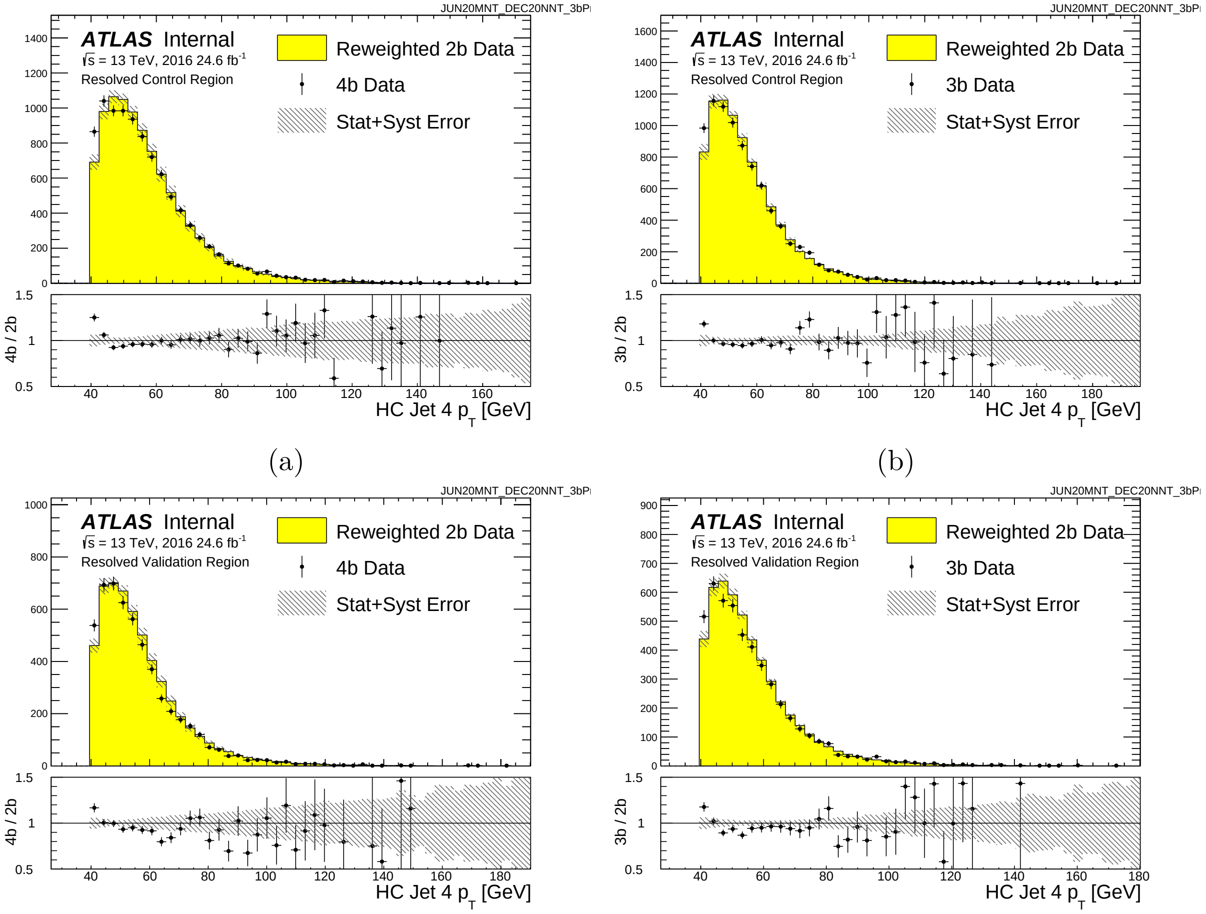
(a)

(b)

(c)

(d)

Figure A.66: Distributions of the third jet p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



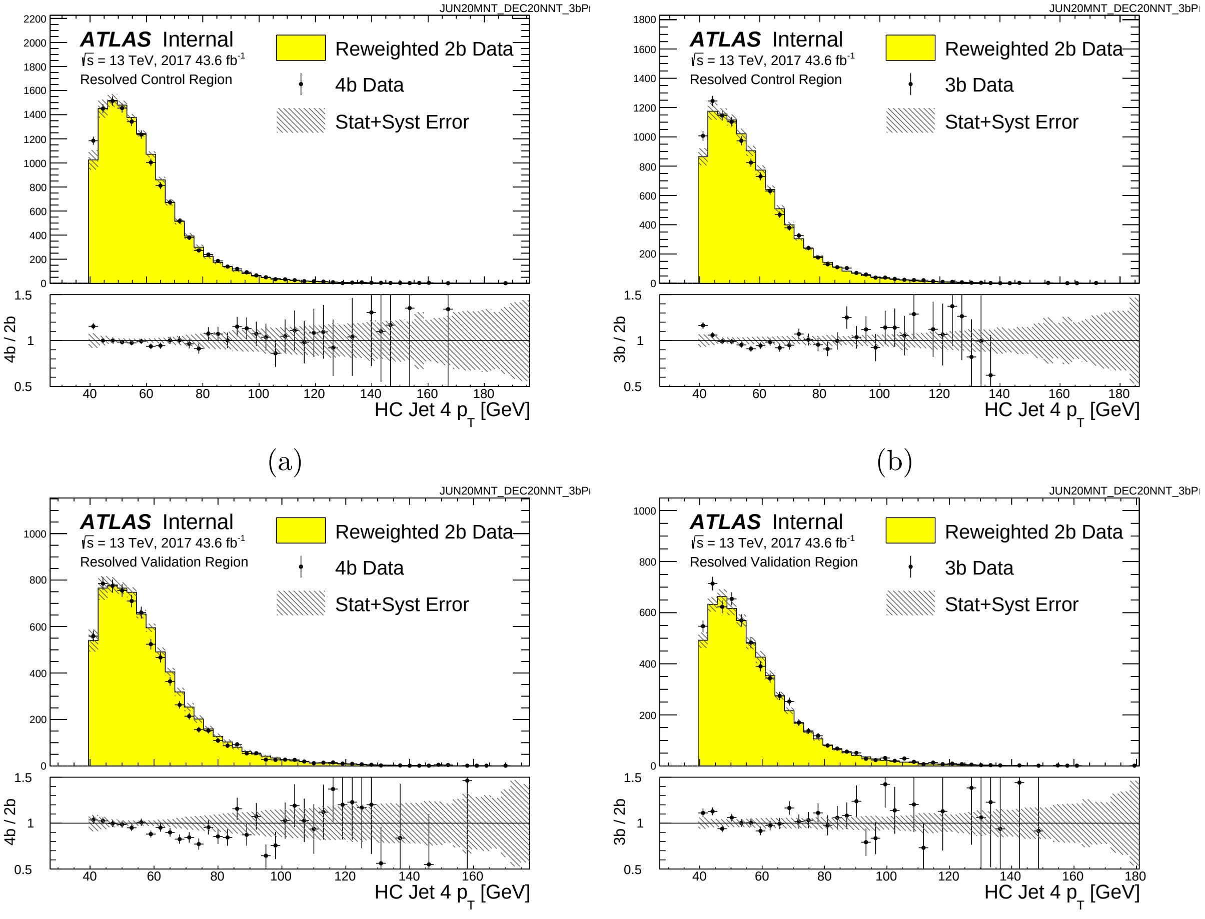
(a)

(b)

(c)

(d)

Figure A.67: Distributions of the fourth jet p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



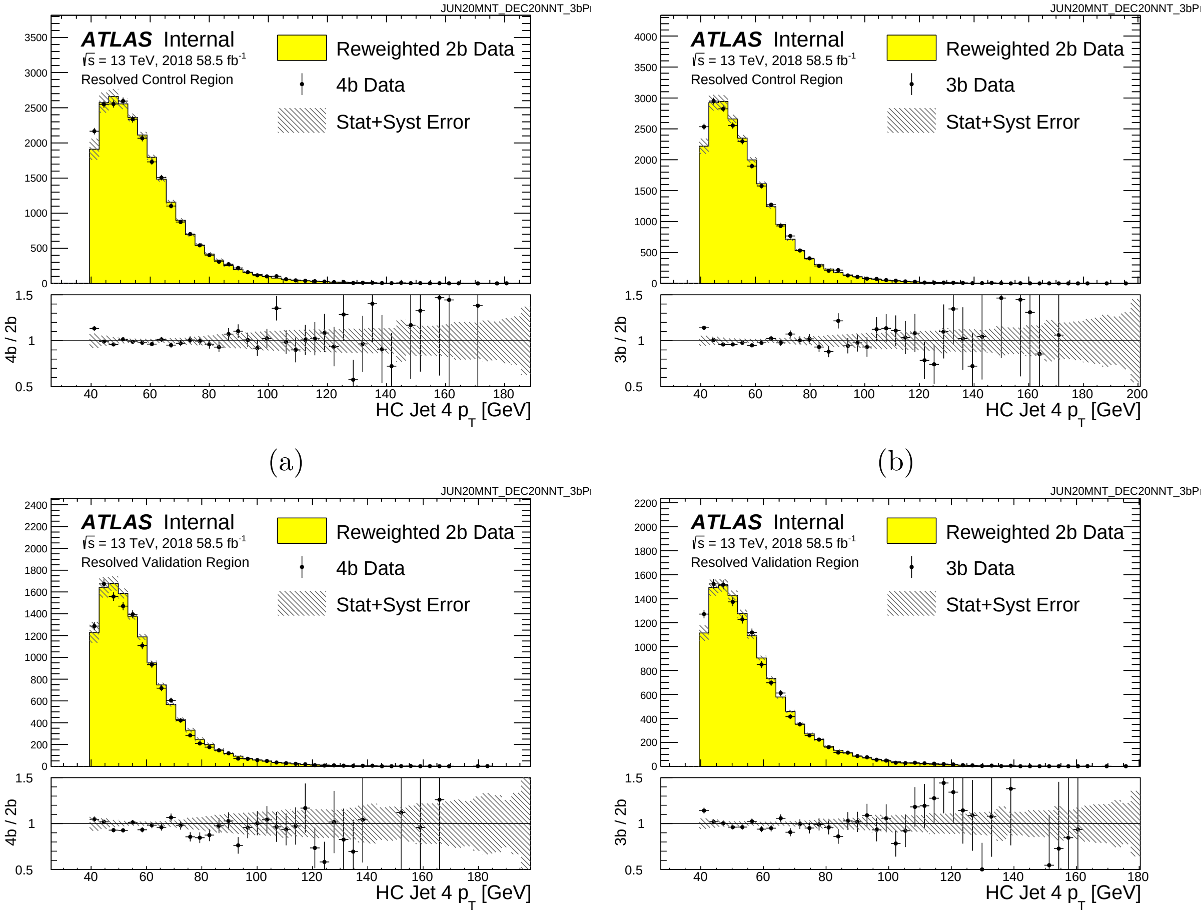
(a)

(b)

(c)

(d)

Figure A.68: Distributions of the fourth jet p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



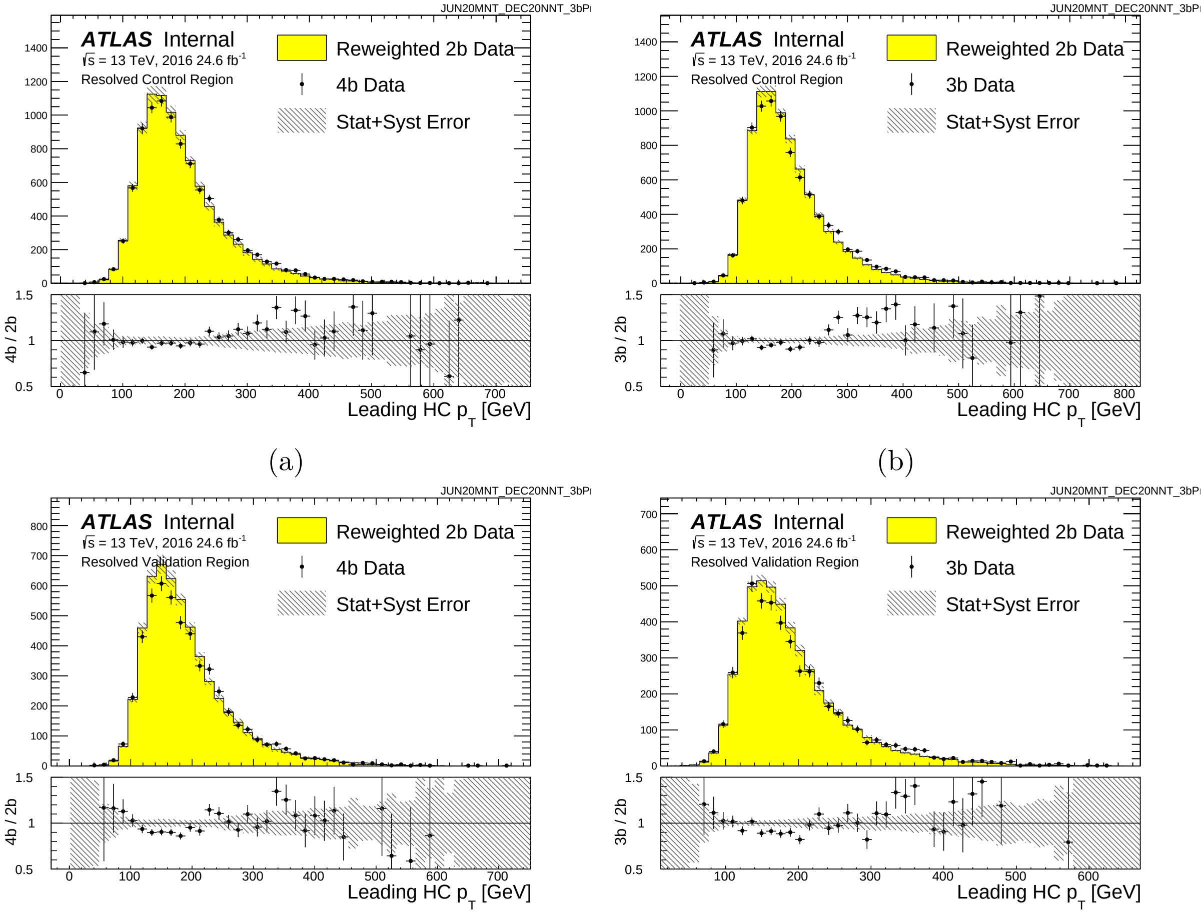
(a)

(b)

(c)

(d)

Figure A.69: Distributions of the fourth jet p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



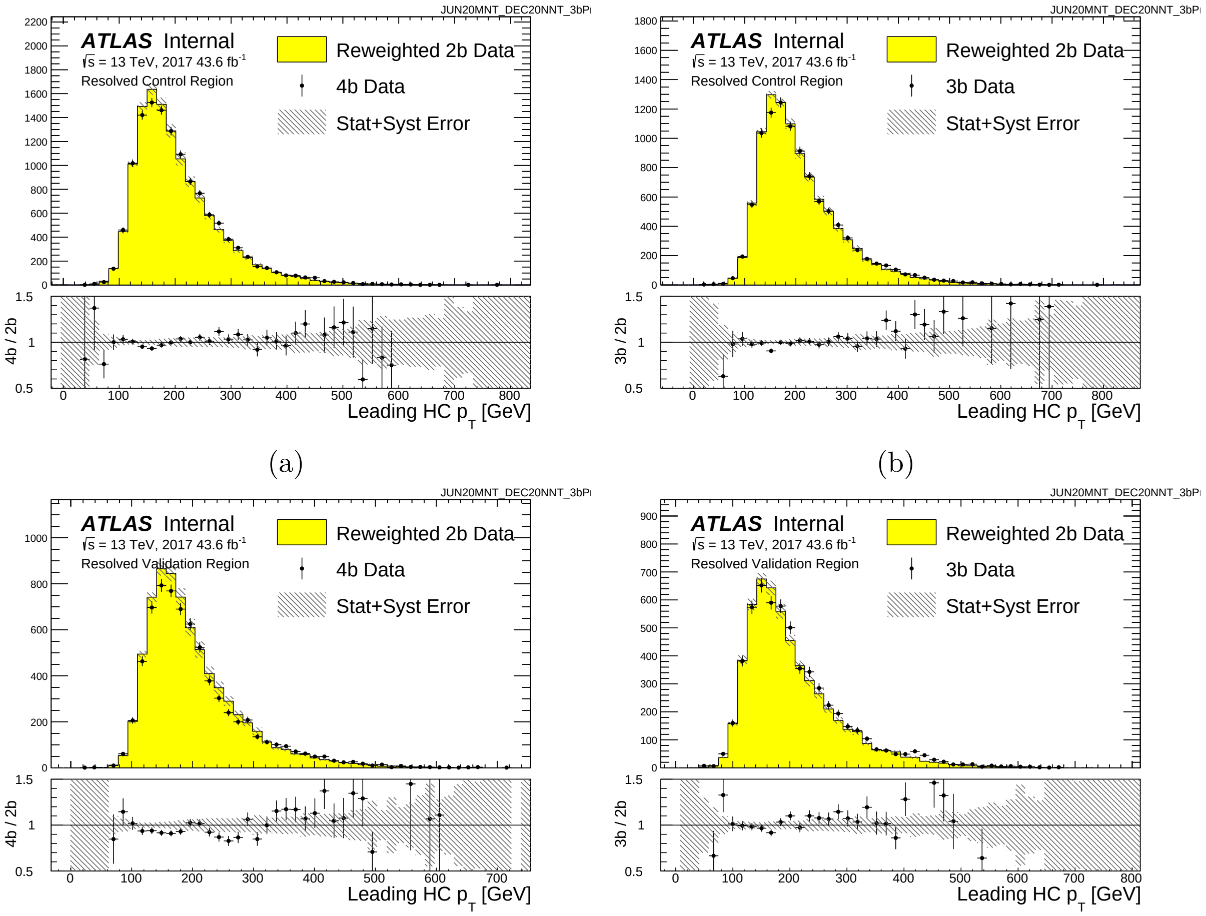
(a)

(b)

(c)

(d)

Figure A.70: Distributions of the leading Higgs canddiate p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



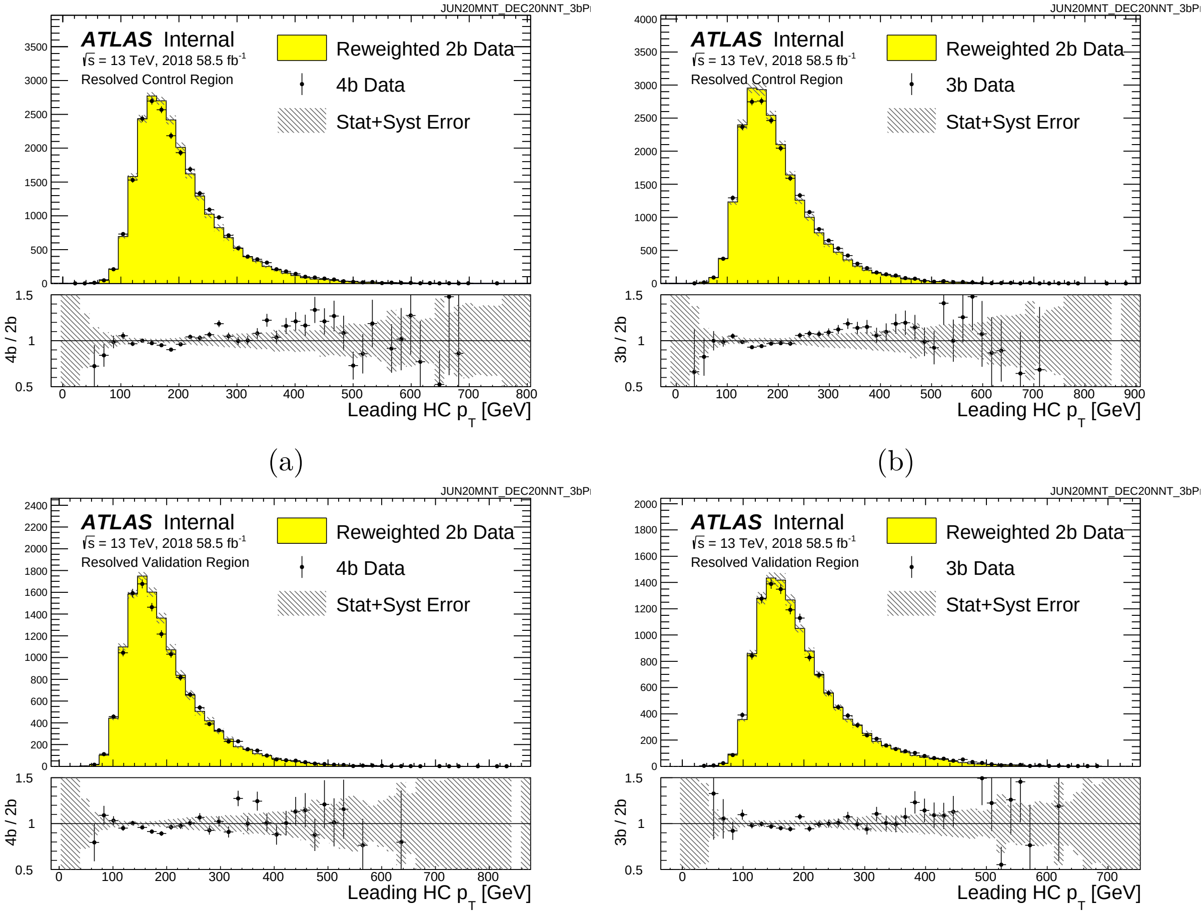
(a)

(b)

(c)

(d)

Figure A.71: Distributions of the leading Higgs canddiate p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



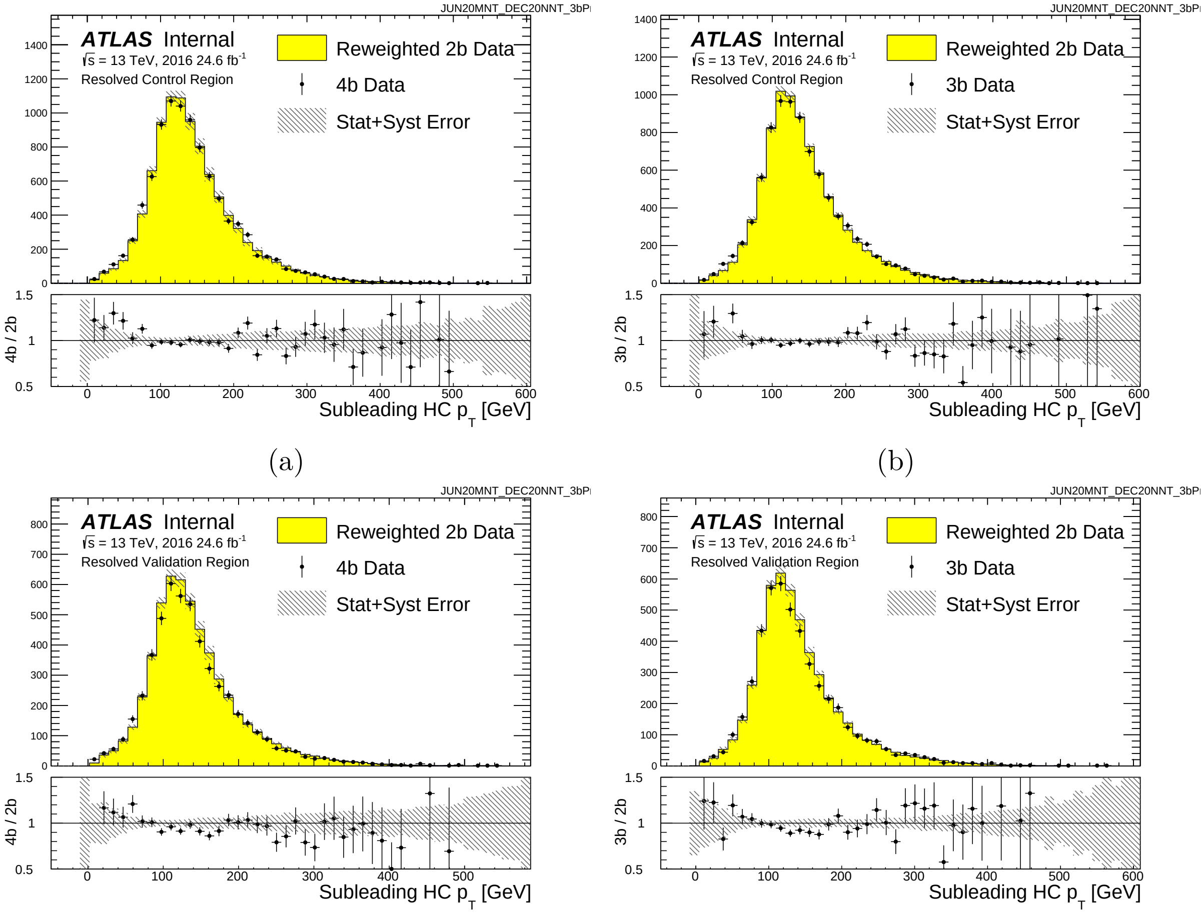
(a)

(b)

(c)

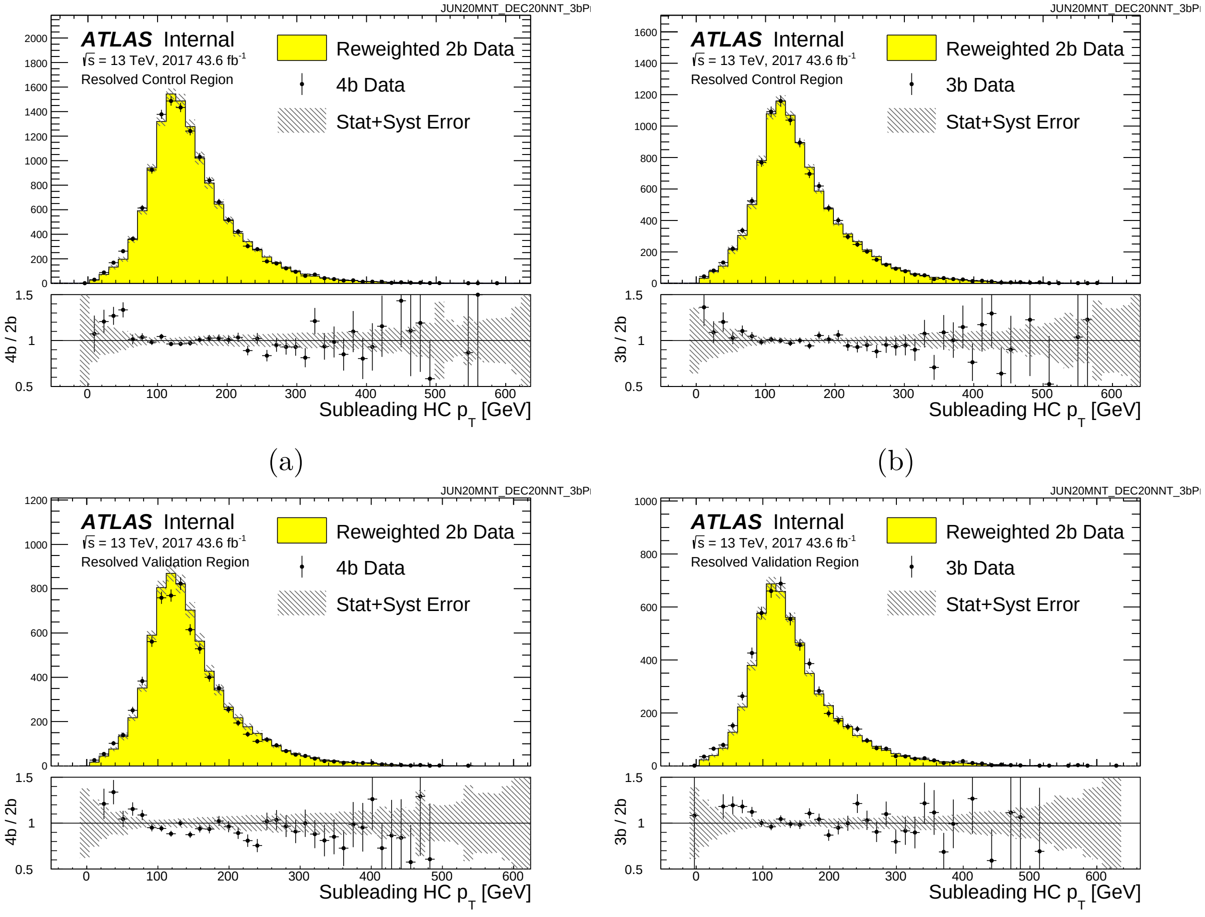
(d)

Figure A.72: Distributions of the leading Higgs canddiate p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b) (c) (d)

Figure A.73: Distributions of the subleading Higgs candidate p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



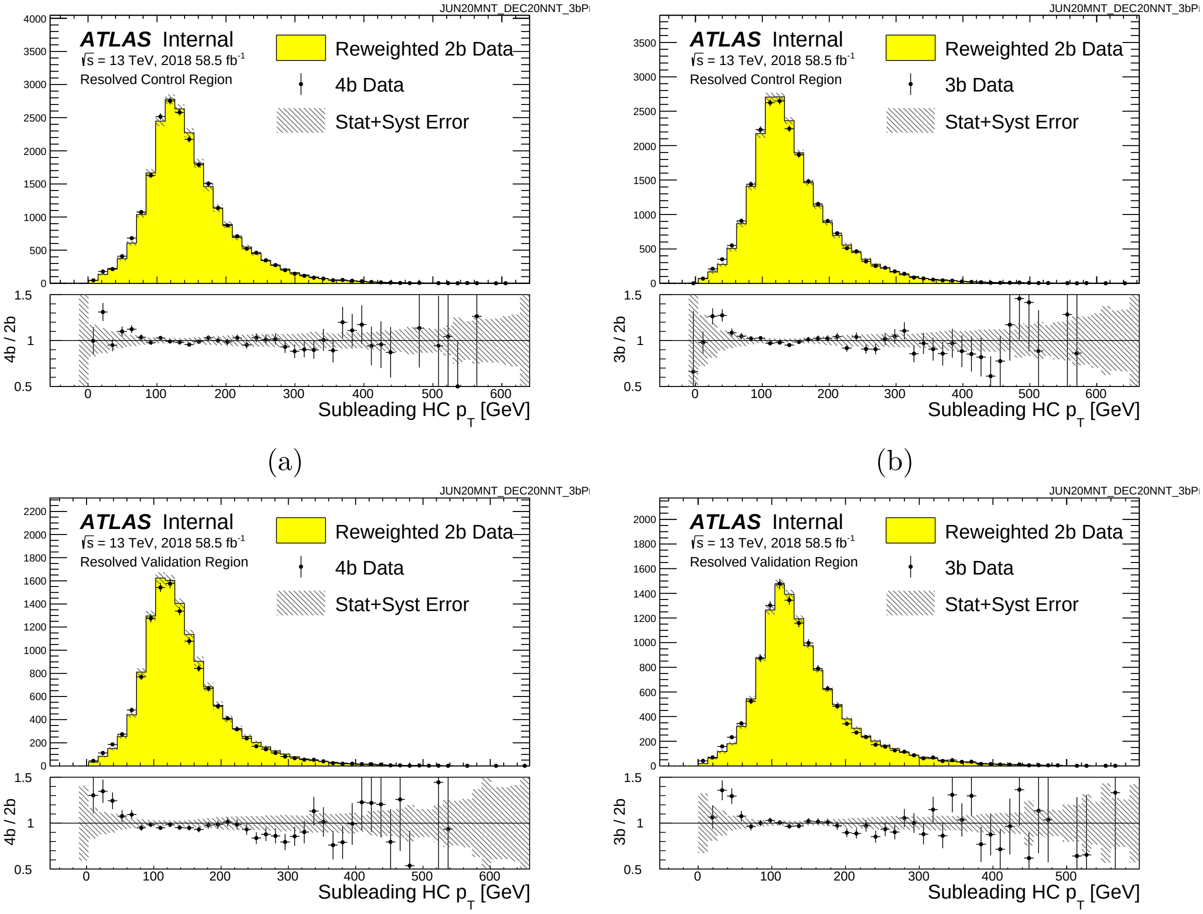
(a)

(b)

(c)

(d)

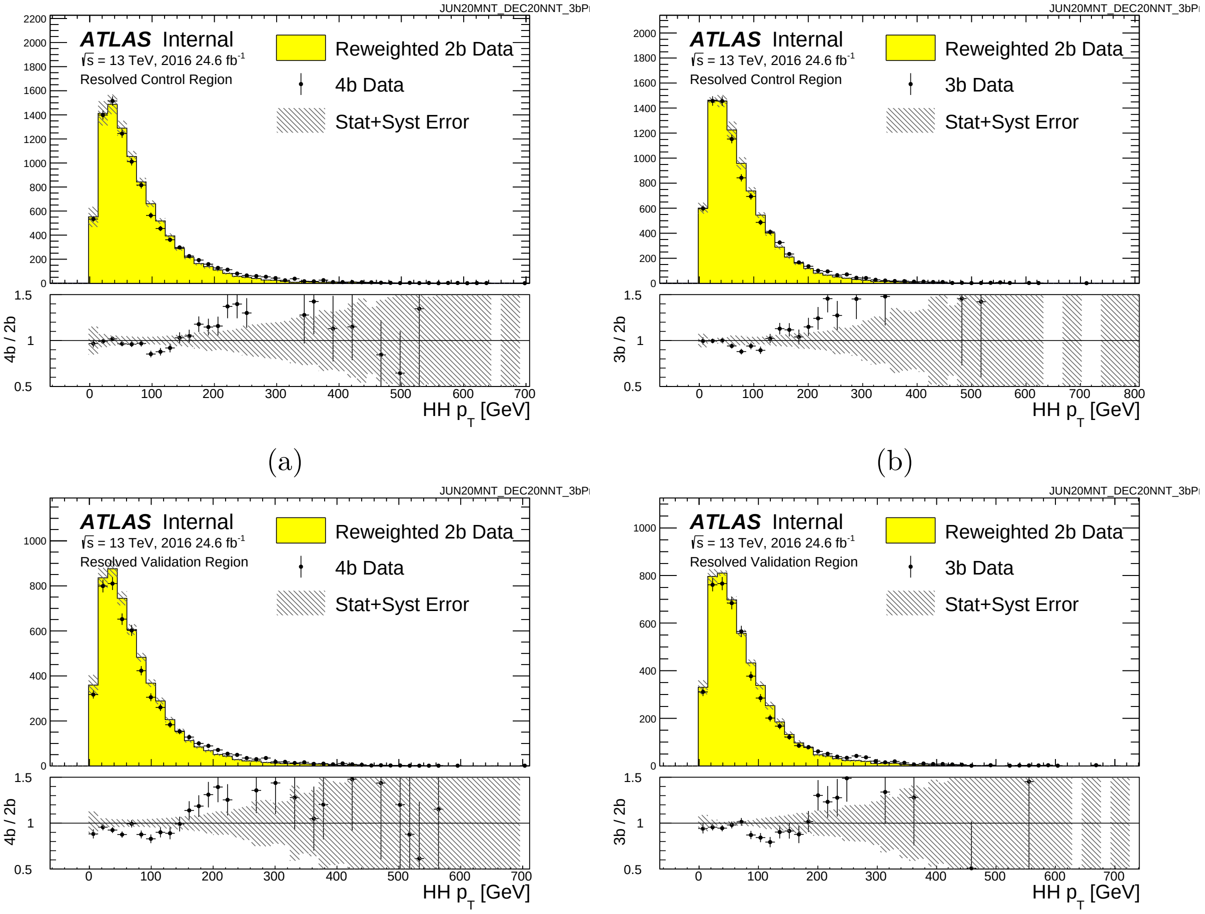
Figure A.74: Distributions of the subleading Higgs candidate p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a) (b)

(c) (d)

Figure A.75: Distributions of the subleading Higgs candidate p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



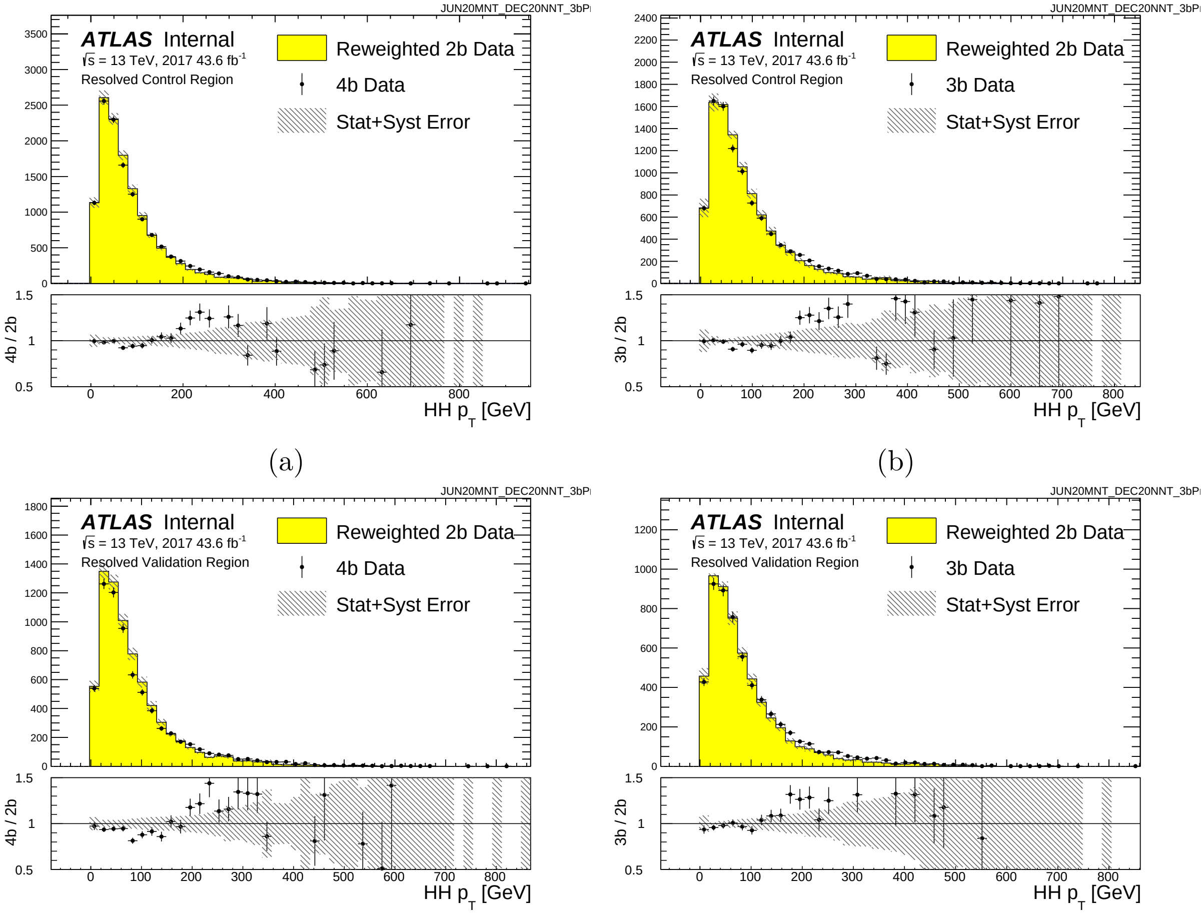
(a)

(b)

(c)

(d)

Figure A.76: Distributions of the di-Higgs system total p_T in 2016 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



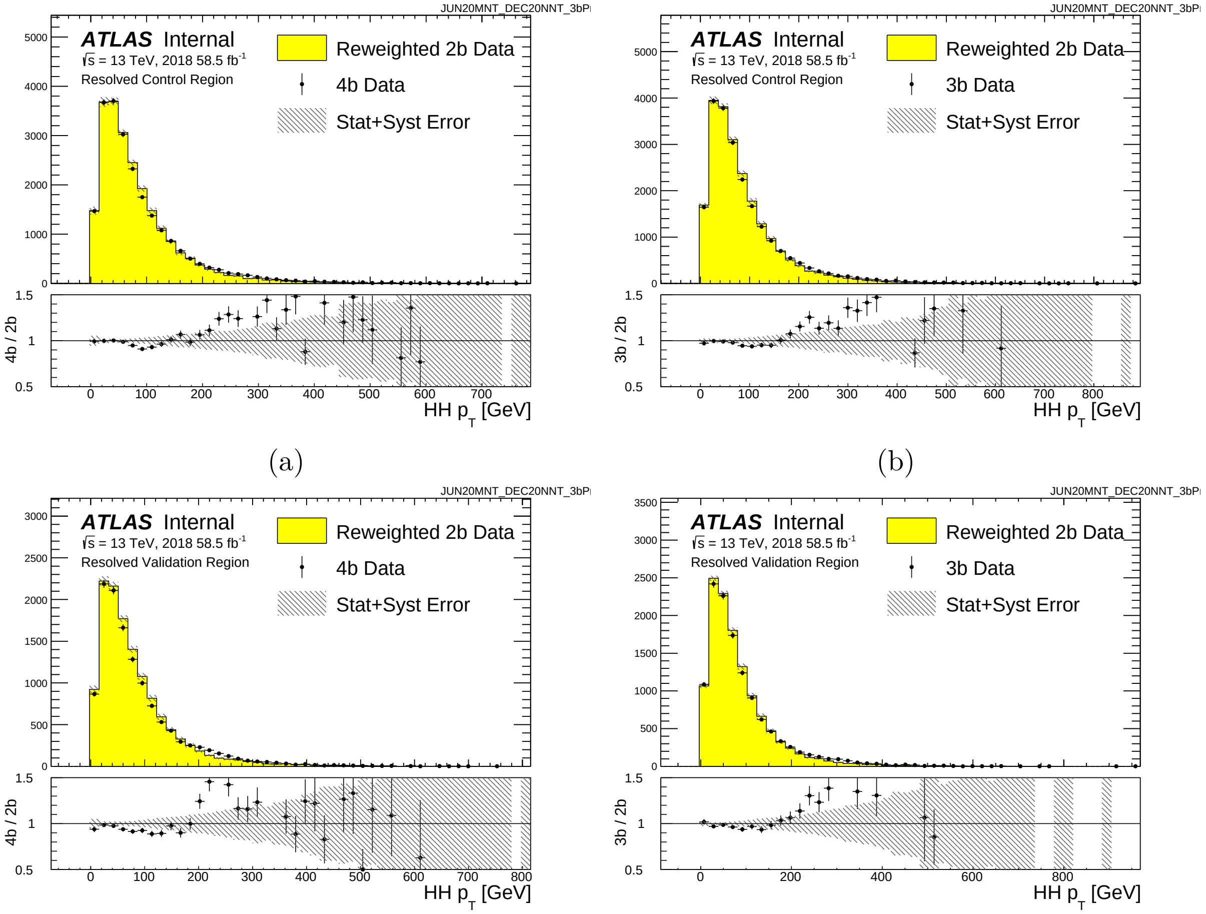
(a)

(b)

(c)

(d)

Figure A.77: Distributions of the di-Higgs system total p_T in 2017 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.



(a)

(b)

(c)

(d)

Figure A.78: Distributions of the di-Higgs system total p_T in 2018 4b and 3b data for the control and validation regions, compared to the 2b-derived background estimate.

REFERENCES

[1] Biagio Di Micco, Maxime Gouzevitch, Javier Mazzitelli, and Caterina Vernieri. Higgs boson potential at colliders: Status and perspectives. *Reviews in Physics*, 5:100045, Nov 2020.

[2] Fermilab applied physics and superconducting technology division. `https://td.fnal.gov/srf-department`. Accessed: 2020-12-12.

[3] M. J. Syphers D. A. Edwards. *An Introduction to the Physics of High Energy Accelerators.* Wiley, 2004.

[4] Laurent Guiraud. Model of an LHC superconducting quadrupole magnet. Aimant quadripôle supraconducteur. Jan 2000.

[5] Andre Holzner. Quadrupole magnet — Wikipedia, the free encyclopedia, 2016. [Online; accessed 22-Dec-2020].

[6] W. Badgett. A Final Review of the Performance of the CDF Run II Data Acquisition System. *J. Phys. Conf. Ser.*, 396:012004, 2012.

[7] M (CERN) Aleksa, W (Pittsburgh) Cleland, Y (Tokyo) Enari, M (Victoria) Fincke-Keeler, L (CERN) Hervas, F (BNL) Lanni, S (Oregon) Majewski, C (Victoria) Marino, and I (LAPP) Wingerter-Seez. ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report. Technical Report CERN-LHCC-2013-017. ATLAS-TDR-022, Sep 2013. Final version presented to December 2013 LHCC.

[8] *ATLAS tile calorimeter: Technical Design Report.* Technical Design Report ATLAS. CERN, Geneva, 1996.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[10] Andrew Ng. Machine learning. Coursera Online Course.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[12] ATLAS Collaboration. Jet energy scale and resolution measured in proton-proton collisions at $\sqrt{s} = 13$ tev with the atlas detector, 2020.

[13] Chervinskii. Autoencoder — Wikipedia, the free encyclopedia, 2015. [Online; accessed 28-Jan-2021].

[14] David Griffiths. *Introduction to Elementary Particles*. Wiley, 2008.

[15] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. 083C01.

[16] Xabier Cid Vidal and Ramon Cid Manzano. Lhc parameters: Taking a closer look at the lhc. Online; accessed 22-Dec-2020.

[17] Eugene P. Wigner. On Unitary Representations of the Inhomogeneous Lorentz Group. *Annals Math.*, 40:149–204, 1939.

[18] Steven Weinberg. *The Quantum Theory of Fields*. Cambridge University Press, 1995.

[19] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2014.

[20] K. Abe, Y. Haga, Y. Hayato, M. Ikeda, K. Iyogi, J. Kameda, Y. Kishimoto, M. Miura, S. Moriyama, M. Nakahata, and et al. Search for proton decay via p→e+ 0 and p→ + 0 in 0.31 megaton · years exposure of the super-kamiokande water cherenkov detector. *Physical Review D*, 95(1), Jan 2017.

[21] Georges Aad et al. Search for new non-resonant phenomena in high-mass dilepton final states with the ATLAS detector. *JHEP*, 11:005, 2020.

[22] A.D. Sakharov. Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe. *Sov. Phys. Usp.*, 34(5):392–393, 1991.

[23] Gianfranco Bertone and Dan Hooper. History of dark matter. *Reviews of Modern Physics*, 90(4), Oct 2018.

[24] T. Aoyama, N. Asmussen, M. Benayoun, J. Bijnens, T. Blum, M. Bruno, I. Caprini, C.M. Carloni Calame, M. Cè, G. Colangelo, and et al. The anomalous magnetic moment of the muon in the standard model. *Physics Reports*, 887:1–166, Dec 2020.

[25] G. W. Bennett, B. Bousquet, H. N. Brown, G. Bunce, R. M. Carey, P. Cushman, G. T. Danby, P. T. Debevec, M. Deile, H. Deng, and et al. Final report of the e821 muon anomalous magnetic moment measurement at bnl. *Physical Review D*, 73(7), Apr 2006.

[26] Antonio Pich. Flavour anomalies, 2019.

[27] John Rumble, editor. *CRC Handbook of Chemistry and Physics.* CRC Press, 2020.

[28] Peter Linstrom. Nist chemistry webbook, nist standard reference database 69, 1997.

[29] S.I. Sukhoruchkin and Z.N. Soroko. Atomic mass and nuclear binding energy for w-172 (tungsten): Datasheet from landolt-börnstein - group i elementary particles, nuclei and atoms · volume 22b: "nuclei with z = 55 - 100" in springermaterials (https://doi.org/10.1007/978-3-540-70609-0_2124). Copyright 2009 Springer-Verlag Berlin Heidelberg.

[30] Marek Gaździcki and Mark I. Gorenstein. *Hagedorn's Hadron Mass Spectrum and the Onset of Deconfinement*, pages 87–92. Springer International Publishing, Cham, 2016.

[31] Manuel Meyer. *Electroweak Phase Transition in the Standard Model and its Inert Higgs Doublet Extension.* PhD thesis, Institute for Theoretical Physics, Universität Bern, 2017.

[32] Lisamission.org.

[33] Nan Phinney. SLC final performance and lessons. *eConf*, C00082:MO102, 2000.

[34] Kek reclaims luminosity record. *CERN Courier*, 2020. Online; accessed 22-Dec-2020.

[35] G. Antchev, P. Aspell, I. Atanassov, et al. First measurement of elastic, inelastic, and total cross-sections at $\sqrt{s} = 13$ TeV by TOTEM and overview of cross-section data at LHC energies. *Eur. Phys. J. C*, 79, 2019.

[36] ATLAS Outreach. ATLAS Fact Sheet : To raise awareness of the ATLAS detector and collaboration on the LHC. 2010.

[37] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011.

[38] Heinrich, Lukas and Feickert, Matthew and Stark, Giordon. pyhf: v0.5.4.

[39] M. Aaboud, G. Aad, B. Abbott, J. Abdallah, O. Abdinov, B. Abeloos, S. H. Abidi, O. S. AbouZeid, N. L. Abraham, and et al. Jet reconstruction and performance using particle flow with the atlas detector. *The European Physical Journal C*, 77(7), Jul 2017.

[40] Measurement of b-tagging Efficiency of c-jets in $t\bar{t}$ Events Using a Likelihood Approach with the ATLAS Detector. Technical Report ATLAS-CONF-2018-001, CERN, Geneva, Mar 2018.

[41] Glen Cowan. *Statistical Data Analysis.* Oxford University Press, 1998.

[42] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2018.