A SEARCH FOR LIGHT PAIR-PRODUCED PARTICLES THAT EACH DECAY INTO AT LEAST FOUR QUARKS

By

ELLIOT JOHN HUGHES

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

written under the direction of

John Paul Chou

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2018

ABSTRACT OF THE DISSERTATION

A search for light pair-produced particles that each decay into at least four quarks

By ELLIOT JOHN HUGHES

Dissertation Director:

John Paul Chou

A search for pair-produced particles with masses above 100 GeV that each decay into at least four quarks is presented. The search focuses on a signal model consisting of a squark that decays into four quarks through an intermediate higgsino with a hadronic R-parity-violating coupling. Data from proton–proton collisions with 13 TeV center-of-mass energy, collected by the Compact Muon Solenoid experiment at the Large Hadron Collider during 2015 and 2016, are analyzed for resonances in the average mass spectrum of pairs of large jets with high transverse momentum, similar mass, and substructure consistent with four partons. A novel data-driven technique is used to estimate the dominant quantum-chromodynamic multijet background events. The significant top-quark decay background events are accounted for by simulation. No statistically significant excess of observed events over expected events in the data is found. The pair production of squarks with masses between 0.10 and 0.72 TeV and gluinos with masses between 0.10 and 1.41 TeV are excluded at 95% confidence level.

Acknowledgments

Thank you,

John Paul Chou, for advising, suggesting, helping, caring, financing, objecting, and waiting;

Sunil Somalwar, for conversations and machinations;

Peter Thomassen, for rational questions and enthusiastic anecdotes;

Alejandro Gomez Espinosa, for telling me to do things a better way;

Savvas Kyriacou, for nonsense and creativity, often in tandem.

Monica and Mark Hughes, for maternal and paternal presence, respectively, and most other contributions besides;

and Clyde Pesh, for aesthetic inspiration and, perhaps more importantly, for pulling me from the path of an approaching car when I was looking the other way and asserting, asserting, asserting.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

If you take matter[1] and partition it into sets of identical elements, then these elements are called *particles* and the sets *species*. In many cases, a particle may be further divided into other different particles—a proton, for example, consists of quarks and gluons—so it's useful to distinguish between *elementary particles*, particles that aren't known to be divisible, and *composite particles*, particles consisting of elementary particles bound together.

While particles of the same species are identical, they can still have individual degrees of freedom. The probability distribution describing the likelihood that a particle has a given set of values for these degrees of freedom is called a *state*. Two particles of the same species can have very different behaviors if they occupy different states, but you can't conclude that one particular particle is in one particular state, only that there are two particles in two states. In the context of the mathematical formalism of quantum mechanics [1], this indistinguishability characteristic of species directly results in two possible *species statistics*: species in which the particles can occupy the same state and those in which they can't [2]; the former are called *bosons*, and the latter *fermions*. Quantum mechanics also introduces the particle property *spin*, an inherent quantized angular momentum. After integrating special relativity [3] into quantum mechanics [4–6], the spin–statistics theorem concludes that bosons must have integer spin and fermions must have half-integer spin [7, 8].

[1] You may choose to conceive of "matter" tautologically—matter is that which is constituted of particles—or axiomatically—matter consists of the generally understood actors in physical reality. In any case, I'm using a more general definition for "matter" than the usual definition that only refers to things with mass, which needlessly emphasizes a confusing physical property.

1.1 The Standard Model

The Standard Model (SM) is the most complete description of the empirically understood dynamics of 30 elementary species that constitute our known universe. It consists of a relativistic quantum field theory (QFT) that describes to the best of our ability the observable particles and their interactions with one another. The six elementary boson species and 24 elementary fermion species of the SM are listed in Table 1.1.

The elementary boson species are mediators of four fundamental particle interactions: the electromagnetic force, the weak force, the strong force, and the Higgs mechanism and coupling[2]. The gravitational force is conspicuously absent from this list; it has yet to find a place in the SM.

The elementary fermion species are everything else. It would be nice to have a simple description of these fermions that mirrors the cleaner conception of the bosons, but there doesn't appear to be one. The elementary fermions compose all of the physical things we're familiar with, the chemical elements formed by protons, neutrons, and electrons, but these elements also contain gluons and photons, both elementary bosons. And the fermions can't be classified as the particles that are experiencing the fundamental interactions because the elementary bosons directly interact with one another: $W^{\pm}$ bosons and photons interact through electromagnetism, gluons interact with other gluons through the strong force, $W^{\pm}$ bosons interact with Z bosons and themselves through the weak force, and Z and $W^{\pm}$ bosons interact with the Higgs boson.

But perhaps this lack of simplicity is itself the important message to take away from my attempts to further summarize elementary fermions. The SM is just one part of the description of the universe that the community of scientifically inclined humans has derived so far, one part of what seems to be an inherently complicated description.

[2] I'm tempted to call this coupling a "force", too, but it would obscure the nuanced nature of the Higgs interaction.

Table 1.1: A list of all 30 elementary particle species in the SM and the values of their mass and electric charge (q).

Name	Symbol	Mass[3]	q	Discovery (order)	
Bosons					
Photon	γ	0	0	1923[4] [9] (2/30)	
W^+ boson[5]	W^+	80 GeV	+1	1983 [10] (23/30)	
W^- boson	W^-	"	−1	"	
Z boson[6]	Z	90 GeV	0	1983 [11] (25/30)	
Gluon	g	0	0	1979 [12] (22/30)	
Higgs boson[7]	H	100 GeV	0	2012 [13, 14] (30/30)	
Fermions					
Electron	e^-	0.5 MeV	−1	1897 [15] (1/30)	Leptons
Positron	e^+	"	+1	1932 [16, 17] (3/30)	
Muon	μ^-	0.1 GeV	−1	1947 [18, 19] (4/30)	
Antimuon	μ^+	"	+1	"	
Tauon	τ^-	2 GeV	−1	1975 [20] (18/30)	
Antitauon	τ^+	"	+1	"	
Electron neutrino	ν_e	-[8]	0	1953 [21] (6/30)	
Electron antineutrino[9]	$\overline{\nu}_e$	"	"	"	
Muon neutrino	ν_μ	-	0	1962 [22] (8/30)	
Muon antineutrino	$\overline{\nu}_\mu$	"	"	"	
Tauon neutrino	ν_τ	-	0	2001 [23] (28/30)	
Tauon antineutrino	$\overline{\nu}_\tau$	"	"	"	
Up quark	u	2 MeV	$+\frac{2}{3}$	1969 [24, 25] (10/30)	Quarks
Up antiquark	$\overline{u}$	"	$-\frac{2}{3}$	"	
Down quark	d	5 MeV	$-\frac{1}{3}$	"	
Down antiquark	$\overline{d}$	"	$+\frac{1}{3}$	"	
Charm quark	c	1 GeV	$+\frac{2}{3}$	1974[10] [26] (16/30)	
Charm antiquark	$\overline{c}$	"	$-\frac{2}{3}$	"	
Strange quark	s	0.1 GeV	$-\frac{1}{3}$	1969 (see "up quark")	
Strange antiquark	$\overline{s}$	"	$+\frac{1}{3}$	"	
Top quark	t	172 GeV[11]	$+\frac{2}{3}$	1995 [27, 28] (26/30)	
Top antiquark	$\overline{t}$	"	$-\frac{2}{3}$	"	
Bottom quark	b	4 GeV	$-\frac{1}{3}$	1977 [29] (20/30)	
Bottom antiquark	$\overline{b}$	"	$+\frac{1}{3}$	"	

The elementary fermions are evenly divided into two groups: *leptons* and *quarks*. Each

[3] The species' masses have been measured to a much higher precision than what I've presented here; I just show one significant figure for intuitive understanding.

[4] Photons have an extremely complicated history, beginning long before the seeds of the SM framework were planted. Therefore, I've chosen to list the "discovery" year as the year in which the first incontrovertable evidence for the particle behavior of light was published.

[5] The "W" stands for "weak", referring to the $W^\pm$ bosons' participation in the weak force. The word "boson" is often appended to the names (as I've done) to make them more readable, an annoyance that would be fixed by giving them less weak names.

[6] The "Z" stands for "zero", referring to its neutral electric charge.

[7] The Higgs boson is the only elementary particle species with a proper name.

[8] In the SM, what we call a neutrino is not a mass eigenstate but, rather, a flavor eigenstate; these states are superpositions of each other. In the strictest definition of the SM, the neutrinos are massless and so this disinction is meaningless, but we empirically know that neutrinos have very small masses.

[9] We still don't know if an antineutrino is distinct from its neutrino partner or if they're just different states of the same species. In any case, they're considered distinct in the SM formalism.

[10] This date corresponds to the discovery of the J/ψ meson. This discovery was quickly followed by the idea of a charm quark, which was mostly confirmed by the 1976 discovery of the first D mesons [30].

[11] I cite the mass of the t with a little more precision than the other species because it's relevant to the analysis I present.

of these groups are evenly divided into three *generations* and, at the same time, into two *types* (see Table 1.2). Fermions of the same type are almost identical to those in the same generation. The fact that the leptons and quarks have the same organizational structure is not completely explained.[12]

After all of this categorization, the last remaining elementary fermion structure is particle–antiparticle pairing.[13] The only observable difference between a normal particle[14] and its associated antiparticle is in their electric charges: an antiparticle has the opposite value of the normal particle's charge.

Table 1.2: The organizational structure of the elementary fermions.

Leptons	e^-, e^+	μ^-, μ^+	τ^-, τ^+	Charged leptons
	ν_e, $\overline{\nu}_e$	ν_μ, $\overline{\nu}_\mu$	ν_τ, $\overline{\nu}_\tau$	Neutrinos
Quarks	u, $\overline{u}$	c, $\overline{c}$	t, $\overline{t}$	Up-type quarks
	d, $\overline{d}$	s, $\overline{s}$	b, $\overline{b}$	Down-type quarks
		Generations		

The two lepton types are the *charged leptons*[15] and the neutrinos. Each lepton has a property called *lepton number* (L), which has a value of 1 for normal leptons and -1 for antileptons. All interactions involving leptons conserve the summed L of all incoming leptons.[16] The three lepton generations are called *lepton flavors* and are named after the charged leptons that occupy them. There's an additional lepton number for each lepton flavor, but while these quantities are conserved in close-range interactions, they're violated

[12] It's probably just a curious emergent phenomenon that might better be called a coincidence; you can find beautiful patterns in twelve random numbers. The fact that there are three generations of each fermion group can be motivated by chiral anomaly cancellation, but maybe some day we'll discover fourth-generation fermions, though they'd have to be pretty different given observed constraints [31].

[13] You can think of $W^\pm$ bosons as a particle–antiparticle pair, but the mathematical description of the $W^\pm$ bosons is functionally different from that of the fermions.

[14] We desperately need a nominclature revolution in physics! I almost had to write "anti-antiparticles".

[15] There's no single word for the electron-like leptons. The SM is ugly.

[16] This property is one of the reasons that we keep the idea of antineutrinos around. If neutrinos are their own antiparticles, then we will have to tweak L conservation.

by neutrino oscillation.

Quarks are the basic building blocks of most composite particles. Quark-composed particles are called *hadrons* and are grouped according to the number of valence quarks they contain: *mesons* for two, *baryons* for three, *tetraquarks* for four, *pentaquarks* for five, and so on.[17] There are six quark–antiquark groups, called *quark flavors*, organized into two types: up type and down type. Much like the leptons' L, the quarks have a conserved[18] property called *baryon number* (B), which is $\frac{1}{3}$ for normal quarks and $-\frac{1}{3}$ for antiquarks.

In the SM QFT, particles of each elementary species are modeled by quanta of tensor or spinor fields on a Lorentzian manifold that represents spacetime. The particle fields are related to one another in a single Lagrangian density ($\mathcal{L}$). This sounds nice, but the SM $\mathcal{L}$ is far too complicated for there to be an analytical solution to its field equations. But, by applying perturbation theory to the vacuum state, it's still possible to calculate important observable quantities from the SM $\mathcal{L}$, namely the scattering matrix (S-matrix). The S-matrix relates the initial and final states of a particle interaction, and therefore defines cross sections and decay rates for physical processes. A convenient approach to calculating an interaction's S-matrix is through the use of *Feynman diagrams*, elegant symbolic representations of the mathematical terms in the perturbative expansion. Feynman diagrams for the lowest order terms, called *tree-level* diagrams, are also important visualizations of basic interactions.

The SM QFT formalism describes the electromagnetic and weak interactions as components of a unified *electroweak* interaction mediated by massless gauge bosons that also couple to the Higgs field. By imposing local gauge invariance and spontaneous symmetry breaking on the combined Higgs–electroweak component of the SM $\mathcal{L}$, and by asserting that the Higgs field has a nonzero vacuum expectation value (v), the distinct electromagnetic and weak interactions are resolved as well as the masses of the $\mathrm{W}^{\pm}$ and Z bosons. At the

[17] So far, no one has observed hadrons with more valence quarks than pentaquarks, which were observed for the first time in 2015 [32].

[18] Technically, B can be violated by sphalerons [33], which have never been observed but are still considered to be in the SM.

same time, the SM $\mathcal{L}$ contains terms detailing interactions between the Higgs field and the elementary fermions. The effect of spontaneous symmetry breaking on these terms result in this Higgs interaction resolving into two parts: masses for the fermions and a coupling between the fermions and the Higgs boson[19].

Oddly, the strong force is more self-contained than the other interactions. While the quarks interact with every elementary boson, gluons only interact with quarks and themselves. In the SM QFT, this gluon-mediated interaction is described by quantum chromodynamics (QCD). The basis of QCD, is a conserved quantity called *color charge* that has three types, called *colors*; each color can have a value of ± 1. A quark occupies a superposition of single-color states with a value of 1 for normal quarks and -1 for antiquarks. A gluon occupies a superposition of two-color states that sum to 0. Free particles must be in a color state in which all three colors are expressed in equal magnitudes. This color-charge model was proposed [34, 35] to account for the fact that quarks and gluons are not observed outside of hadrons, a principle termed *confinement*. Therefore, products of strong interactions will immediately form hadrons with other quarks and gluons, which are created spontaneously from the vacuum if necessary, in a process called *hadronization*.

1.2 Beyond the Standard Model

Although the SM is the most comprehensive and empirically tested physical model to date,[20] it isn't sufficient. Despite significant effort, the SM hasn't incorporated any description of gravity, a startling gap within the context of astrophysical observations that indicate the existence of enormous quantities of mass and energy unattributed to SM particles. Additionally, neutrino oscillation experiments reveal that the neutrinos are superpositions of three

[19] This Yukawa coupling is the Higgs "force" I was hesitant to call by that name.

[20] This phrasing is disingenuous, since the SM was generally defined to agree with what was observed. In any case, it's an unprecedented descriptive success.

mass eigenstates,[21] the details of which are not yet understood enough to fully be incorporated in the SM. The SM also doesn't account for the observed particle–antiparticle-quantity asymmetry; it could be described by CP-violating terms in the SM $\mathcal{L}$, but observations constrain them well below the necessary values.

Perhaps equally concerning are the puzzling quirks of the SM. For example, while the elementary fermions have a nice organizational structure, there doesn't appear to be any nice pattern to their masses.[22] But this oddness pales in comparison to a deeply disturbing aspect of the SM QFT: renormalization. Some higher-order terms in the perturbative expansion of an S-matrix element have divergent contributions. These higher-order corrections can be parametrized in terms of an upper limit, Λ, on the integral over the interaction's momentum transfer. By adjusting (*renormalizing*) the fundamental constants in the SM QFT to values that incorporate the corrections up to Λ and then ignoring these divergent higher-order terms, the S-matrix can be evaluated accurately, and Λ becomes a physical representation of the energy scale above which the SM is an inadequate description. For the strong and electroweak interactions, the divergent terms are proportional to $\log(\Lambda)$. Therefore, setting Λ orders of magnitude higher than the interaction constants only results in a minor renormalization. This is called *natural* behavior. But the Higgs field is problematic: its divergent terms are proportional to Λ^2, meaning that the value used for the Higgs boson mass in S-matrix calculations is highly dependent on the energy scale; this is called the *hierarchy problem*.

An elegant theoretical solution to the hierarchy problem is *supersymmetry* (SUSY) [36–39]. According to SUSY, SM elementary particle species are partnered with new species with the opposite statistics. Each SM elementary fermion species is partnered with a new boson species, which is given the same name as the fermion with an "s" before it; therefore, each quark is assigned a *squark*. The SUSY partners of the elementary SM bosons are less

[21] This is pretty annoying, since every other massive particle in the SM is defined as a single mass eigenstate.
[22] In fact, even the ordering is wonky: in the first quark generation, the down-type quark has a greater mass than the up-type quark while the opposite is true for the other two generations.

straightforward because the partners are assigned before imposing spontaneous symmetry breaking, and an additional Higgs boson species has to be introduced in order to correctly reproduce the Higgs interaction with respect to quarks. The SUSY partners of SM elementary bosons are named by taking the associated boson names and appending "ino" to them; therefore, the SUSY partners of gluons and Higgs bosons are called *gluinos* and *higgsinos*, respectively. The additional terms in the Higgs mass correction introduced by these new SUSY species exactly cancel the corresponding SM ones, effectively erasing the hierarchy problem. But a consequence of imposing SUSY is the introduction of couplings between SUSY and SM particles that violate L and B conservation. In order to counteract this unobserved feature of the theory, a species property called R *parity* is defined:

$$R \equiv (-1)^{3(B-L)+2S}, \tag{1.1}$$

where S is the species's spin. Under the assumption that L and B are defined to be the same for the SUSY partners of SM species, R has a value of ± 1 for SM and SUSY species, respectively. By asserting that R parity is conserved (RPC) [40], all of the SUSY interactions that violate B and L are disallowed. Another attractive feature of RPC SUSY is that the lightest SUSY particle would necessarily be stable, and therefore would provide a particle solution to the dark matter conundrum.

Unfortunately, there hasn't been any direct experimental evidence for RPC SUSY, despite the fact that natural RPC SUSY particles should be produced in measurable quantities at collider experiments [41–44].

The resolution that I pursue discards RPC, allowing R parity to be violated (RPV) [45]. As long as the RPV interactions are small, SUSY can still agree with the L and B conservation measurements. And while the theory no long answers the dark matter question, it's able to preserve naturalness by still asserting that the SUSY species have masses relatively close to their SM partners. Certain RPV models are still largely unconstrained by

observations [46].

The particular RPV model that I probe hypothesizes a small RPV hadronic coupling that allows a squark to decay to two quarks. By assuming that a higgsino mass is smaller than the squark mass, a relatively light squark can decay to a quark and a higgsino, where the higgsino decays to three quarks via an off-shell squark decaying through the RPV interaction (see Figure 1.1, left). Therefore, this process ultimately describes a light squark decaying to four quarks. Because the RPV coupling is assumed to be small, the initial squarks are expected to be produced in pairs. I also consider a case in which a gluino decays to five quarks in a similar manner, which demonstrates the inclusive nature of the search I perform (see Figure 1.1, right).

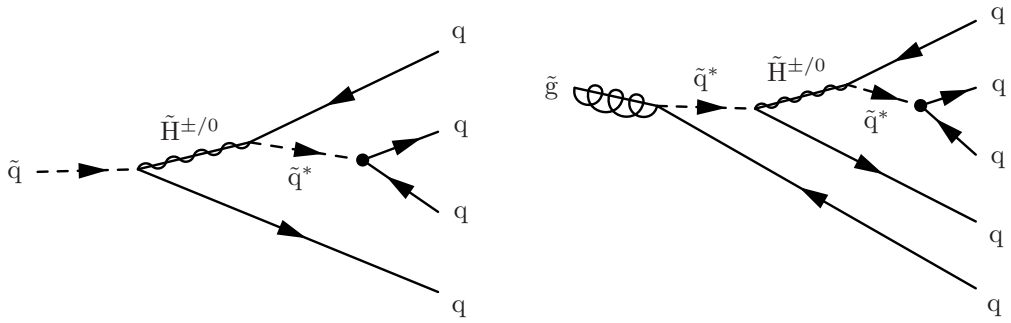


Figure 1.1: Feynman diagrams for a squark decaying into four quarks via an intermediate higgsino and a three-body decay involving an off-shell squark with a hadronic RPV coupling (left) and for a gluino decaying into five quarks via a three-body decay into two quarks and a higgsino, which decays into three quarks as in the squark case (right). I expect the initial squarks or gluinos to be produced in pairs, meaning that I look for two of the same decay in each event.

1.3 Search

I'm looking for new particles with masses on the order of hundreds of GeV that are produced in pairs by proton–proton collisions at the Large Hadron Collider and that each decay into at least four final-state quarks. While quarks are certainly fascinating, the motivation for this search goes far beyond fandom. Proton–proton collisions produce an extreme quantity of hadrons through expected stong-interaction processes, but the current understanding of

these processes is still imprecise. Therefore, fighting through this smokescreen to discover a new physical process is an engaging task; one of the reasons for pursuing this search has been to develop and employ an elegant procedure for estimating the majority of these expected processes, using the data itself. Although a number of other searches for similar decays exist, none constrain new particles with masses below about 400 GeV that decay entirely into quarks, nor do any concentrate on the particular collisions that I do. I focus on a hypothesis that these undiscovered particles are squarks that decay to four final-state non-top quarks (see Figure 1.1), but the analysis is general enough that it's sensitive to other models.[23]

Following a procedure detailed in Chapter 4, I analyze proton–proton collision data recorded by the Compact Muon Solenoid detector (see Chapters 2 and 5) and cluster reconstructed particles into large jets (see Chapter 3), selecting events that contain two similar jets that each likely encompass four final-state quarks (see Chapter 6). Using the average mass of these two jets as my variable of interest, I perform a fit of the expected signal and estimated background (see Chapter 7) distributions to the data. I don't observe an excess of data events over the background prediction, meaning I don't discover the hypothesized particles; I set upper limits on pair-production cross sections (see Chapter 8).

[23] Alongside the four-quark final-state hypothesis, I also show results for a five-quark case arising from pair-produced gluinos. For purely selfless reasons, I leave other models and topologies to future graduate students.

Chapter 2

Experiment

I analyze proton–proton collision data collected by the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC) near Geneva, Switzerland.

2.1 The Large Hadron Collider

The LHC [47] is a 26.7 km circumference particle collider on the border between Switzerland and France at the European Organization for Nuclear Research (CERN). Built in the tunnel that once housed the Large Electron–Positron collider (LEP, 1989–2000) [48], the LHC consists of two approximately circular beam pipes containing oppositely circulating particle beams that intersect at four different interaction points (IPs), each monitored by particle detector experiments. Although an LHC beam can contain either protons or lead nuclei (depending on the time of year), I only consider proton-beam collisions in this analysis.

The LHC's proton beams consist of groups of about 10^{11} protons,[24] called *bunches*. Each bunch has an energy of 6.5 TeV, and so the center-of-mass energy of LHC collisions 13 TeV. The proton bunches are separated in time by 25 ns, resulting in proton-bunch collisions, called *bunch crossings*, at a rate of 40 MHz.

During a single bunch crossing, there are roughly 30 proton–proton interactions, the record of which is referred to as an *event*. The number of proton–proton interactions per unit area per unit time, called *instantaneous luminosity* (L), is the standard measure of the data collection rate. Of the many proton interactions in a collision, I'm interested in studying

[24] This is about the same as the number of neurons in a human brain.

those that involve a large transfer of momenta to their products. The softer interactions in an event, called *pileup*, which make up the significant bulk of the proton interactions, and the associated initial- and final-state radiation and lingering collision products from the previous event are collectively called the *underlying event*. A host of techniques are implemented to sift through and account for the underlying event in order to most accurately identify the physical processes of interest.

2.2 The Compact Muon Solenoid detector

The CMS [49, 50] detector is a roughly cylindrical particle detector located at Point 5 of the LHC. The detector is a combination of four particle-detecting *subdetectors* located within a magnetic field created by a large superconducting solenoid. The first subdetector—the one closest to the IP—is the tracker, which measures the paths of charged particles. Surrounding the tracker is the electromagnetic calorimeter (ECAL), which measures the energy of charged particles and photons, followed by the hadron calorimeter (HCAL), which measures the energy of neutral hadrons. These two calorimeters work by absorbing particles that hit them and measuring the resulting particle showers using scintillation light. The only SM particles that typically pass through the calorimeters are muons and neutrinos. Muons are detected by the fourth subdetector called the muon system, which sits beyond the coils of the solenoid.

The CMS detector uses a right-handed Cartesian coordinate system. The origin is defined to be the expected IP, and the x-axis points radially inward toward the center of the LHC ring, while the y-axis points radially upward; therefore the z-axis points tangentially along the beam pipe in the counterclockwise direction. Because the particles produced in collisions radiate outward from the IP, it's often more useful to use a cylindrical or spherical coordinate system. In these cases, the longitudinal axis is taken to be the z-axis, the azimuthal angle $\phi \in (-\pi, \pi]$ is defined from the positive x-axis toward the positive y-axis, and the polar angle $\theta \in [0, \pi]$ is defined from the positive z-axis. Instead of θ, it's

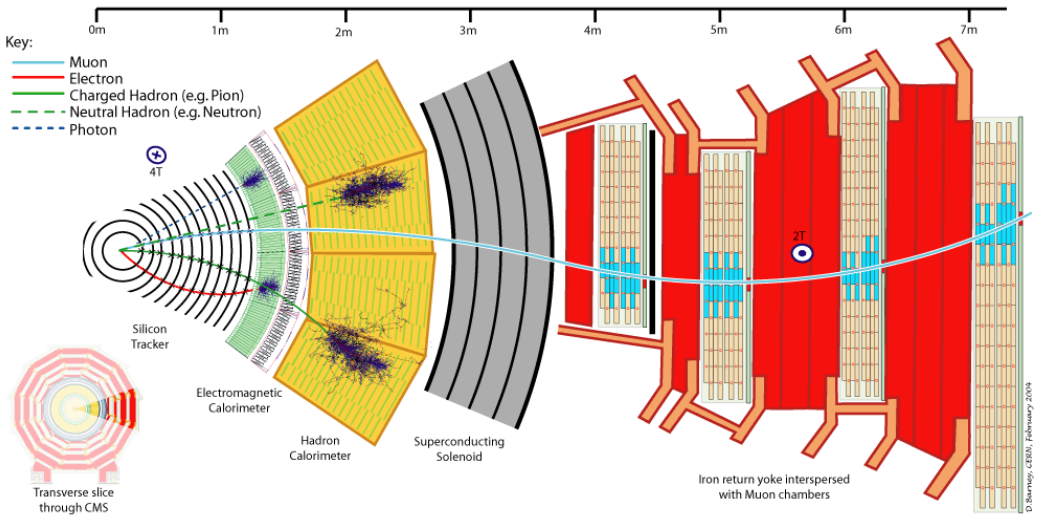


Figure 2.1: A portion of a transverse slice of the CMS detector, illustrating the subdetectors and examples of different particle signatures. [51]

more practical to consider the pseudorapidity

$$\eta \equiv -\log\left[\tan\left(\frac{\theta}{2}\right)\right]$$
$$\in [-\infty, \infty]$$

because η differences are Lorentz invariant in the massless limit. The spherical radial coordinate is denoted r, and the cylindrical coordinate is denoted ρ.

In order to detect as many particles from an event as possible, each CMS subdetector is divided into a *barrel*, a longitudinally aligned cylindrical segment occupying a region of ρ, and two *endcaps*, circular segments occupying a region of the transverse plane at either end of the barrel. Thus, the two proton beams pass through the centers of the layered endcaps at either end of the detector and collide at the IP within the detector's concentric barrels.

The tracker consists of an inner pixel detector, in the region $4.4\,\text{cm} < \rho < 10\,\text{cm}$, and an outer strip detector, in the region $10\,\text{cm} < \rho < 1\,\text{m}$. Both components are composed of silicon semiconductor junctions which exhibit small electric currents when passing charged particles excite valence electrons into the conduction band, an observation termed a *hit*.

The ECAL, occupying the region $1\,\mathrm{m} < \rho < 1.8\,\mathrm{m}$, consists of lead-tungstate ($PbWO_4$) crystals backed by photodetectors. $PbWO_4$ is optically transparent, allowing it to absorb and scintillate simultaneously. In fact, it emits 80% of its scintillation light in 25 ns, which is the LHC bunch crossing period. $PbWO_4$ is also particularly well-suited to a compact calorimeter because of its short radiation length of 0.89 cm and small Molière radius of 2.2 cm.

The HCAL, occupying the region $1.8\,\mathrm{m} < \rho < 3\,\mathrm{m}$, is composed of passive brass absorbers interleaved with plastic scintillating tiles. The optical output from regions of the HCAL detector in η–ϕ–ρ space are summed together and forwarded to photodetectors. The analog signal output by each photodetector is converted to a digital signal by a charge-integrating and encoding integrated circuit, which outputs a single value for each bunch crossing. In addition to the barrel and endcap components, the HCAL includes two additional components: the hadron outer system, which sits just beyond the solenoid and increases the accuracy of energy measurements by observing any showers that extend past the barrel, and the hadron forward system, which is located at either end of the CMS detector and measures hadronic energy that passes too close to the beamline to be detected by an HCAL endcap.

The muon system consists of 250 drift tubes in the barrel and 468 cathode strip chambers in the endcaps, both of which use gas ionization to record passing muons. These trackers are accompanied by resistive plate chambers for associated timing and momentum measurements.

It's impractical to store data from every collision at the LHC, therefore CMS has a two-step trigger system to store the most potentially interesting fraction of observed events. The first step is the level-1 trigger that stores each incoming event in a buffer and spends about 1 µs per event to determine if it should be accepted. The accepted events are forwarded to the high-level trigger (HLT), which uses simple techniques to decide if an event is pertinent to a set of analyses. Events that pass the HLT are sent to the data storage

management system that coordinates storage and access of the events around the globe. Observed events are analyzed in the context of CMS software (CMSSW) [50] and the ROOT analysis framework [52].

Chapter 3

Objects

The determination of physical objects from the basic subdetector measurements of an event is called *reconstruction*, a process that starts with the interpretation of the digital signals read out from the detector's front-end electronics and concludes with the identification of individual particles via the particle flow (PF) algorithm [53]. The collection of these reconstructed particles describes the basic physical processes underlying their production. In particular, I define clusters of reconstructed particles called *jets* in order find events consistent with two initial particles of the same mass that each decayed into many quarks.

3.1 Basic objects

A *track* is a charged particle trajectory identified by hits observed in the tracker. From the collection of hits, subgroups of particular pixel hits consistent with the beginning of a charged particle trajectory[25] are extracted as *track seeds*. Each track seed is then extrapolated, and hits near to this expected trajectory are fitted to define a track as a continuous path. In order to reduce the combinatorial complexity of this procedure, the algorithm is performed iteratively on the hit collection, each time with an emphasis on a slightly different trajectory type (e.g., high momentum, displaced, etc.), until a final track collection is defined.

Tracks are grouped into *vertices* by an iterative fitting procedure that associates tracks with a common starting z-position along the beamline. Vertices are sorted in descending

[25] The charge of a particle depositing a hit is assumed to be ± 1. Without this assumption, a particle momentum couldn't be infered from the geometry of a track.

order based on the amount of p_{T} associated with the tracks they contain;[26] the first vertex is called the *primary vertex*, the rest are called *pileup vertices*, referring to the fact that they're likely derived from the event's pileup [50].

Energy deposits in the calorimeters are grouped to form *calorimeter clusters* in an analogous way to track reconstruction. Calorimeter cells that made particularly large energy measurements are selected to be *cluster seeds*, each defining a calorimeter cluster following a procedure that associates the cluster seed with all nearby cells. Since non-seed cells may be associated with more than one cluster, the energies of the clusters are assigned by a simultaneous fit that assumes a Gaussian energy deposition centered at each cluster seed.

The PF algorithm derives reconstructed particles from the tracks, vertices, and calorimeter clusters (*PF elements*) by linking them together. The linking algorithm associates PF elements into *PF blocks* based on their proximity in the η–ϕ space. Individual particles are then derived from each PF block in the following order: muons, electrons and isolated photons, and finally charged and neutral hadrons created by hadronization and any remaining photons.

In general, reconstructed objects are assigned four-momenta. Typically, four-momenta are expressed by a position in η–ϕ space, an energy, and a *transverse momentum* defined as $p_{\mathrm{T}} \equiv \sqrt{p_x^2 + p_y^2}$, where p_x and p_y are the x- and y-components of the momentum, respectively. This parametrization is particularly useful because the p_{T} of the particles produced in the collision must sum to zero, since the initial protons are only traveling in the z-direction.

3.2 Jet clustering

The process of deciding how many jets there are in an event, and to which jet a given object belongs, is called *jet clustering* [54]. Beginning from a collection of objects,[27] jet

[26] To understand vertex ordering, you need to know about jet clustering, which I describe in Section 3.2. For each vertex: 1) cluster its tracks into AK4 jets, 2) assign the vertex a sorting key corresponding to the quadratic sum of the p_{T} of each of its track jets, 3) add to this key the square of the vertex's missing transverse energy, calculated as the vectorial sum of the vertex's track jets' p_{T} values.

[27] In general, jets need not be formed from reconstructed particles, as they are in the presented analysis; for example, a jet could be constructed from lower-level detector objects. In fact, the ordering of vertices

clustering algorithms group these objects into a new collection of jets, where each jet is usually represented by the vectorial sum of its constituent objects' four-momenta.

Today, the most prolific jet clustering algorithms are the k_{T}, anti-k_{T} (AK), and Cambridge–Aachen (CA) algorithms. All of these algorithms operate recursively using the same general principle: for each object i in the object collection, calculate

$$d_{\mathrm{b}i} \equiv {p_{\mathrm{T}}}_i^{2p} \tag{3.1}$$

where $p \in \{-1, 0, 1\}$ for the k_{T}, CA, and AK algorithms, respectively; and for each other object j, calculate

$$\Delta R_{ij}^2 \equiv (y_i - y_j)^2 + (\phi_i - \phi_j)^2 \,, \tag{3.2}$$

$$d_{ij} \equiv \min\left({p_{\mathrm{T}}}_i^{2p}, {p_{\mathrm{T}}}_j^{2p}\right) \frac{\Delta R_{ij}^2}{R^2} \tag{3.3}$$

where y is rapidity, and the *jet size* R is a free parameter loosely describing a cone size for the jets; then, if the smallest d_{ij} is less than $d_{\mathrm{b}i}$ combine the objects i and j into a new object by summing their four-momenta, otherwise put object i into the output jet collection; finally, repeat the process until every object is in the jet collection. Jet collections are sorted by the jets' p_{T}, in descending order, where the first n jets are called the n *leading jets*.

Jet collections are denoted by the jet clustering algorithm used, followed by ten times the particular value of R. For example, I mainly use a jet collection clustered with CA with $R = 1.2$ (CA12).

A general measure of the amount of jet activity in an event is

$$H_{\mathrm{T}} \equiv \sum_{i=1}^{n_{\mathrm{jets}}} p_{\mathrm{T}i} \tag{3.4}$$

where n_{jets} is the number of jets from a particular jet collection that pass a basic selection.

involves clustering tracks into jets (see Footnote 26).

For my purposes, the H_{T} definition I use is derived from one of the triggers I employ to select data events (see Appendix A): the H_{T} is constructed from jets in the AK8 jet collection that each have $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $|\eta| < 2.5$.

3.3 Jet corrections

I'm looking for the pair-production of a new particle, a physical process that would result in a single vertex. Therefore, the objects associated with the pileup vertices are nuisances that need to be accounted for. This process of selecting and correcting against pileup is called *pileup mitigation*. The first pileup mitigation technique I employ is charged-hadron subtraction (CHS), an algorithm that loops over the reconstructed particle collection before jet clustering is performed and filters out all charged hadrons that originate from a pileup vertex consisting of at least four tracks [55]. The CHS algorithm selection is fairly conservative, since it would be unhelpful to accidentally remove important particles; it removes only about 50% of charged particle pileup and, of course, no neutral particle pileup.

After jets are clustered, they need to be further corrected for the pileup they contain that wasn't removed by CHS, for idiosyncrasies of the CMS calorimeters' responses, and for differences between simulation and data (if the jets under consideration are from simulated events). These corrections are applied to the jets' four-momenta, and so they're called *jet-energy corrections* (JECs) [55].

It's desirable for the mass of a jet that consists of a single particle's decay products to reflect that particle's mass. Unfortunately, pileup and detector noise can significantly muddle the mass scale and resolution of a jet. Additionally, non-resonant processes, such as light-quark and gluon production, can result in jets with large masses when components of the associated showers are distributed throughout many jets, an effect that can significantly overwhelm other processes in the mass range of interest. *Jet grooming* algorithms remove unwanted contributions from these sources; particularly, I use the *pruning* jet grooming algorithm [56]. To prune a jet, recluster the constituent objects of the jet, and at each step

in which objects i and j would be combined to form a new object k, require that

$$\frac{\min\left(p_{\mathrm{T}i},\, p_{\mathrm{T}j}\right)}{p_{\mathrm{T}k}} > z_{\mathrm{cut}}, \tag{3.5}$$

$$\Delta R_{ij} < D_{\mathrm{cut}}, \tag{3.6}$$

otherwise discard the softest of objects i and j and continue reclustering, where $z_{\mathrm{cut}} = 0.1$ and $D_{\mathrm{cut}} = 0.5$ are chosen to be the standard CMS pruning parameters and ΔR_{ij} is the positive square root of Equation 3.2.

3.4 Jet substructure

An event's jet collection is a general description of its hadronic activity, but jets are simple objects that describe complex phenomena. Classifying particular jets based on their constituent particles is an important step in a refined analysis. I identify jets that likely contain at least four hadronized particles; I also find jets that likely contain at least one b quark, in order to select events containing top quarks.

If a jet is sufficiently large, it can contain a number of smaller jets from different quarks and gluons. In the context of the large jet, these internal jets are termed *subjets*. For my purposes, I don't need to know very much about a jet's subjets other than the general number of them. I therefore use the N-subjettiness variables, denoted τ_N [57]. These quantities are determined for a particular jet by reclustering its constituents with the k_{T} algorithm until N subjets remain and then calculating

$$\tau_N \equiv \frac{1}{d_0} \sum_i p_{\mathrm{T}i} \min\left(\Delta R_{1i},\, \ldots,\, \Delta R_{Ni}\right), \tag{3.7}$$

where i iterates over the jet constituents and $d_0 \equiv \sum_i p_{\mathrm{T}i} R$, where R is the jet size. The relative N-subjettiness $\tau_{kl} \equiv \tau_k / \tau_l$, for some integers k and l, is a useful characterization of a jet's substructure.

The procedure for identifying jets that were likely formed by the decay of a bottom quark is called *b tagging* [58]. Hadrons that contain bottom quarks have relatively long lifetimes; a B meson has a mean lifetime of $\sim$1 ps, and therefore the jet produced by its decay will be displaced from the primary vertex by some tenths of a mm, a distance resolvable by the CMS tracker pixels. This displaced jet vertex is called a *secondary vertex*. I use the combined secondary vertex version 2 (CSVv2) b-tagging algorithm, which incorporates a jet's secondary vertex and constituent track information in a neural network to return a discriminator variable d_{CSVv2}. The larger the value of d_{CSVv2}, the more likely the jet is to be identified as the product of a bottom quark. I employ the loose working point for d_{CSVv2}, which is defined by the value of d_{CSVv2} above which jets are b tagged with a 10% misidentification probability. This working point corresponds to about an 80% b-tagging efficiency.

Chapter 4

Procedure

I'm looking for pair-produced RPV squarks or gluinos that each decay into at least four quarks. I call these physical processes *signals*.[28] Each signal is defined by a particular squark or gluino mass (*signal mass*); I assume the intermediate higgsino mass is 75% or 60% of the squark or gluino mass, respectively, a configuration that distributes the initial mass energy equally between the initial decay quarks.[29] I consider 11 squark masses between 0.1 and 0.8 TeV and 17 gluino masses between 0.1 and 1.5 TeV, values I chose to adequately cover the range of signal masses to which the analysis is sensitive.

I consider observed events in which the products of a signal would be detected relatively near to the initial signal particles' trajectories. I do this by clustering the reconstructed particles in an event into CA12 jets, and then I select events that contain two CA12 jets (one for each of the pair-produced signals) that each have large transverse momentum and substructure consistent with at least four subjets. I call such CA12 jets *fatjets*.

Since the signal mass is a free parameter in the search, I use the average mass ($\overline{m}$) of the two leading fatjets as the variable of interest, defined as $\overline{m} \equiv (m_1 + m_2)/2$, where m_1 and m_2 are the pruned masses of the fatjets—from this point, all jet masses discussed are assumed to be pruned. I filter the events containing two fatjets by requiring that the two

[28] Technically, this usage is incorrect. A "signal" should actually describe a sign that one of these processes exists, such as a process's contribution to a particular distribution; it shouldn't be used to describe the process itself. But this language comes from a particular worldview that I dispute—I don't think things exist in any way beyond their "signals"—so I have no problem adulterating it. Actually, I'd prefer to drop "signal" in favor of a term like "desideratum", from which I could derive "desideraton" to refer to a particular particle I want to find, but I don't want to add too much unnecessary confusion.

[29] This is simply an educated choice driven by convenience, since I didn't have the resources to simulate a wide variety of signals.

fatjets have relatively similar masses, since the two pair-produced signals are assumed to be identical.

Not every selected event will actually result from a signal; they're more likely to come from known SM processes that look like a signal, called *backgrounds*. I consider three background processes: generic QCD processes that result in jets formed by the hadronization of non-top quarks and gluons, called *QCD multijet*; pair production of top quarks, called $t\bar{t}$; and $W^{\pm}$ and Z boson decays. The $W^{\pm}$ and Z boson background events can be disregarded because I don't expect very many of them to pass the selection, and those that I do expect look very similar to QCD multijet events (see Section 7.1); they're implicitly included when I account for the significant QCD multijet background (see Section 7.2).

I extract a particular signal's contribution to the observed events by performing a maximum-likelihood fit of the expected signal and estimated background components to the observed $\overline{m}$ distribution. Each component in the fit is an $\overline{m}$ probability density function (PDF): the signal and $t\bar{t}$ background $\overline{m}$ PDFs are taken from simulation, while the QCD multijet background $\overline{m}$ PDF is constructed using a data-driven method. Each $\overline{m}$ PDF is assigned three fit parameters: a normalization factor, a shift of the median of the PDF from the nominal value, and a width factor about the median of the PDF. These fit parameters are called the *normalization*, *shift*, and *stretch*, respectively. Fitting the $\overline{m}$ PDFs to the data allows me to quantify the differences in shape between simulation and data.

Chapter 5

Datasets

The analysis I conduct involves observed and simulated events, each organized into separate datasets. The observed events were measured by the CMS detector and are referred to as *data* (see Section 5.1). The simulated events are specific to individual signal and background processes (see Section 5.2).

5.1 Data

I analyze a total integrated luminosity of 38.2 fb^{-1} of data collected by CMS in 2015 and 2016. From the entire collection of recorded data, I consider events related to the analysis by selecting only those that passed at least one of the relevant triggers that depend on H_{T} and jet p_{T} (see Appendix A). These events are included in the *JetHT* CMS datasets, documented in Table 5.1. In order to prevent events which might contain significant experimental issues from contaminating my results, I ignore problematic luminosity sections identified by the CMS data quality monitoring system by applying the *golden JSON* luminosity mask files shown in Table 5.2. I make sure that the distributions of relevant observable quantities recorded in the 2015 datasets are similar to those recorded in 2016, otherwise I might introduce systematic effects derived from combining the datasets from the different years (see Appendix B).

Table 5.1: The JetHT CMS datasets and corresponding integrated luminosities.

Dataset name	$\int L\,dt$ [**fb^{-1}**]
/JetHT/Run2015C_25ns-16Dec2015-v1/MINIAOD	0.017
/JetHT/Run2015D-16Dec2015-v1/MINIAOD	2.241
2015 total:	**2.258**
/JetHT/Run2016B-03Feb2017_ver2-v2/MINIAOD	5.751
/JetHT/Run2016C-03Feb2017-v1/MINIAOD	2.573
/JetHT/Run2016D-03Feb2017-v1/MINIAOD	4.242
/JetHT/Run2016E-03Feb2017-v1/MINIAOD	4.025
/JetHT/Run2016F-03Feb2017-v1/MINIAOD	3.105
/JetHT/Run2016G-03Feb2017-v1/MINIAOD	7.576
/JetHT/Run2016H-03Feb2017_ver2-v1/MINIAOD	8.435
/JetHT/Run2016H-03Feb2017_ver3-v1/MINIAOD	0.216
2016 total:	**35.923**

Table 5.2: The luminosity masks used for each analyzed run period.

Run period	**Luminosity mask filename**
2015	Cert_13TeV_16Dec2015ReReco_Collisions15_25ns_JSON.txt
2016	Cert_271036-284044_13TeV_23Sep2016ReReco_Collisions16_JSON.txt

5.2 Simulated events

Simulated events are created in three steps: generation, showering, and reconstruction. Generation software calculates S-matrix elements from tree-level and higher-order Feynman diagrams that correspond to a particular physical process resulting from a proton–proton collision. Randomized events are then derived from these S-matrix elements, where each event is represented by a record of the basic kinematic variables of each particle involved in the interaction. Showering software then decays the final-state partons in each generated event using Monte Carlo (MC) techniques, resulting in a collection of simulated particles that could potentially be observed by a detector. Therefore, the final step in simulating events is simulating the interaction of these showered particles with a detector model in order to form a reconstructed event as similar to data as possible.

I simulate pair-produced squark and gluino signal events using MadGraph5_aMC@NLO [59] for the generation step, PYTHIA 8.2 [60] for the subsequent parton showering, and

CMSSW 8 for the event reconstruction. I normalize squark and gluino events to the theoretical cross sections calculated to next-to-leading-order plus next-to-leading-logarithm precision [61]. For the squark production cross sections I use those computed for top squarks. The datasets that correspond to the simulated signal events for each signal mass and the associated production cross sections are shown in Table 5.3.

Simulated events for background processes relevant to the analysis are produced centrally by CMS. I analyze two different QCD multijet datasets that differ slightly in their simulation procedure and binning: $\mathrm{QCD}_{2\to2}$ and $\mathrm{QCD}_{2\to4}$. The $\mathrm{QCD}_{2\to4}$ dataset is simulated using MadGraph5_aMC@NLO to generate up to four partons which are then showered by PYTHIA 8.2; the sample is binned in H_T. The $\mathrm{QCD}_{2\to2}$ dataset is simulated using PYTHIA 8.2 to generate two partons which are subsequently showered using the same software; the sample is binned in $\hat{p}_\mathrm{T}$, the value of both initial partons' p_T. These QCD multijet datasets are listed in Table 5.4. Simulated $\mathrm{t\bar{t}}$ events are generated with POWHEG v2 [62] and showered with PYTHIA 8.2. I also analyze simulated $\mathrm{W}^\pm$ + jets and Z + jets events where the $\mathrm{W}^\pm$ or Z boson decays hadronically, and other relevant diboson processes. The $\mathrm{t\bar{t}}$ and non-QCD datasets are listed in Table 5.5.

Table 5.3: The simulated squark and gluino datasets and corresponding production cross sections.

Dataset name	Entries	Cross section [pb] [61]
Squark mass points		
/Sq100To4J /elhughes-miniaod_sq100to4j_moriond17cutht700_v2-c25660a1db7b69b092e620682f3b66ea/USER	757100	1521.11
/Sq150To4J /elhughes-miniaod_sq150to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	249.409
/Sq175To4J /elhughes-miniaod_sq175to4j_moriond17cutht700_v1-c25660a1db7b69b092e620682f3b66ea/USER	100000	121.416
/Sq200To4J /elhughes-miniaod_sq200to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	64.5085
/Sq250To4J /elhughes-miniaod_sq250to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	21.5949
/Sq300To4J /elhughes-miniaod_sq300to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	8.51615
/Sq400To4J /elhughes-miniaod_sq400to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	1.83537
/Sq500To4J /elhughes-miniaod_sq500to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	100000	0.51848
/Sq600To4J /elhughes-miniaod_sq600to4j_moriond17cutht700_v1-c25660a1db7b69b092e620682f3b66ea/USER	100000	0.174599
/Sq700To4J /elhughes-miniaod_sq700to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	99900	0.0670476
/Sq800To4J /elhughes-miniaod_sq800to4j_moriond17cutht700-c25660a1db7b69b092e620682f3b66ea/USER	99900	0.0283338
Gluino mass points		
/Sg100To5J /agarabag-miniaod_sg100to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	907600	82500.0
/Sg150To5J /agarabag-miniaod_sg150to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	13400.0
/Sg175To5J /agarabag-miniaod_sg175to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	6500.0
/Sg200To5J /agarabag-miniaod_sg200to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	3440.0
/Sg250To5J /agarabag-miniaod_sg250to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	1140.0
/Sg300To5J /agarabag-miniaod_sg300to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	445.0
/Sg350To5J /agarabag-miniaod_sg350to5j_moriond17_cutht500_v2-c25660a1db7b69b092e620682f3b66ea/USER	99700	195.0
/Sg400To5J /agarabag-miniaod_sg400to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	99700	93.7
/Sg450To5J /agarabag-miniaod_sg450to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	99400	48.1
/Sg500To5J /agarabag-miniaod_sg500to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	970800	26.1
/Sg550To5J /agarabag-miniaod_sg550to5j_moriond17_cutht500_v2-c25660a1db7b69b092e620682f3b66ea/USER	99800	14.8
/Sg600To5J /agarabag-miniaod_sg600to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	100000	8.68
/Sg650To5J /agarabag-miniaod_sg650to5j_moriond17_cutht500_v2-c25660a1db7b69b092e620682f3b66ea/USER	99900	5.26
/Sg800To5J /agarabag-miniaod_sg800to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	186000	1.4891
/Sg1000To5J /agarabag-miniaod_sg1000to5j_moriond17_cutht500_v2-c25660a1db7b69b092e620682f3b66ea/USER	85500	0.325388
/Sg1250To5J /agarabag-miniaod_sg1250to5j_moriond17_cutht500-c25660a1db7b69b092e620682f3b66ea/USER	154200	0.0627027
/Sg1500To5J /agarabag-miniaod_sg1500to5j_moriond17_cutht0-c25660a1db7b69b092e620682f3b66ea/USER	100000	0.0141903

Table 5.4: The simulated QCD multijet datasets and corresponding cross sections. For both $QCD_{2\to4}$ and $QCD_{2\to2}$, I combined the nominal and extended datasets to get the most statistics available.

Dataset name	Entries	Cross section [pb]
$QCD_{2\to4}$		
/QCD_HT500to700_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	18929951	31630.0
/QCD_HT700to1000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	15629253	6802.0
/QCD_HT1000to1500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	4767100	1206.0
/QCD_HT1500to2000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	3970819	120.4
/QCD_HT2000toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	1991645	25.25
/QCD_HT500to700_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v2/MINIAODSIM	43341392	31630.0
/QCD_HT700to1000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	29783527	6802.0
/QCD_HT1000to1500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	10360193	1206.0
/QCD_HT1500to2000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	7855883	120.4
/QCD_HT2000toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	4047360	25.25
$QCD_{2\to2}$		
/QCD_Pt_300to470_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	4150588	7823.0
/QCD_Pt_470to600_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	3959986	648.2
/QCD_Pt_600to800_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	3896412	186.9
/QCD_Pt_800to1000_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	3992112	32.293
/QCD_Pt_1000to1400_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	2999069	9.4183
/QCD_Pt_1400to1800_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	396409	0.84265
/QCD_Pt_1800to2400_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	397660	0.114943
/QCD_Pt_2400to3200_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	399226	0.00682981
/QCD_Pt_3200toInf_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v3/MINIAODSIM	391735	0.000165445
/QCD_Pt_300to470_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	18253032	7823.0
/QCD_Pt_600to800_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	9622896	186.9
/QCD_Pt_800to1000_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	15704980	32.293
/QCD_Pt_1000to1400_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	6982586	9.4183
/QCD_Pt_1400to1800_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	2477018	0.84265
/QCD_Pt_1800to2400_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	1552064	0.114943
/QCD_Pt_2400to3200_TuneCUETP8M1_13TeV_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM	596904	0.00682981

Table 5.5: The simulated $t\bar{t}$ and $W^{\pm}$ and Z boson datasets and corresponding cross sections.

Dataset name	Entries	Cross section [pb] [63]
/TT_TuneCUETP8M2T4_13TeV-powheg-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	77081156	730.0
/WJetsToQQ_HT-600ToInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	1026587	95.14
/ZJetsToQQ_HT600toInf_13TeV-madgraph /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	996000	5.67
/ZZTo4Q_13TeV_amcatnloFXFX_madspin_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	30454227	6.842
/ZZTo2L2Q_13TeV_powheg_pythia8 /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	496436	1.999
/WWTo4Q_13TeV-powheg /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	1998400	52.4
/WWToLNuQQ_13TeV-powheg /RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM	1999200	34.9

Chapter 6

Selection

I sift through the data and simulation to find events consistent with the signal. The first step is to identify a baseline set of events that satisfy loose criteria for relevance, a process called *preselection* (see Section 6.1). I then define subsets of these preselected events, called *analysis regions*: a *signal region* (SR), containing the events that look the most like the signal, and *control regions* (CRs), containing events that look similar to the background events I expect in the SR. To do this, I cluster the reconstructed particles in each event into CA12 jets, applying the corrections outlined in Section 3.3. Then, I form each analysis region from events that contain two leading CA12 jets that pass particular requirements on substructure, called the region's *fatjet tag*, and together satisfy relational criteria based on their m and η, called the region's *fatjet pairing*. The definitions of the fatjet tag and pairing for the SR and CRs are provided in Table 6.1.

6.1 Preselection

The data I analyze have already been filtered by the triggers I use (see Appendix A), but the same trigger selection can't easily be applied to simulated events. I avoid this difficulty by applying an $H_{\mathrm{T}} > 900\,\mathrm{GeV}$ requirement to all events, which is equivalent to the trigger selection within 98% for CA12 jet masses greater than 100 GeV.

Signal events contain two particles of the same mass that decay hadronically. All of the decay products of an initial particle need to be contained in a single jet in order for that jet's mass to correctly reflect the initial particle's mass. The more p_{T} the initial particle

has, the more likely the decay products will be close to one another, and the higher the p_{T} of the resulting jet. Therefore, I select events in which the two leading CA12 jets each have $p_{\mathrm{T}} > 400\,\mathrm{GeV}$. I further require that these two CA12 jets each have $|\eta| < 2.0$, which ensures that the jets' constituents are dominated by those in the barrel and endcap regions of the detector. Finally, I make sure that these jets contain at least some hadronic component and aren't derived from calorimeter noise by imposing a restriction on the neutral and charged jet constituents with a loose jet identification [64].

The fraction of events that pass the preselection is called the *selection acceptance*. The signal selection acceptance increases with the squark mass, as shown in Figure 6.3 (top left). This is because a signal CA12 jet's contribution to the event H_{T} is greater for higher squark masses, since the energy stored in the squark mass is converted to momenta for its decay products.

Table 6.1: The analysis regions, each defined by a fatjet tag and pairing. For each CR, there's an additional b-tagged version that includes $d_{\mathrm{CSVv2}} > 0.5426$ in the fatjet tag. The first group corresponds to the nominal analysis regions; the second group are alternate analysis regions that aren't used in the results but are defined for various checks.

Analysis region	**Fatjet tag**	**Fatjet pairing**
SR	$\tau_{21} < 0.75$, $\tau_{42} < 0.50$, and $\tau_{43} < 0.80$	$\Delta\eta < 1.0$ and $A_{\mathrm{m}} < 0.1$
Inclusive CR	$\tau_{21} < 0.75$, $\tau_{42} < 0.55$, $\tau_{43} < 0.90$, and $\tau_{42} > 0.50$ and/or $\tau_{43} > 0.80$	"
Loose SR	$\tau_{21} < 0.75$, $\tau_{42} < 0.50$, and $\tau_{43} < 0.90$	"
Loose CR	$\tau_{21} < 0.75$, $\tau_{42} < 0.60$, $\tau_{43} < 0.95$, and $\tau_{42} > 0.45$ and/or $\tau_{43} > 0.80$	"
Loose-τ_{42} CR	$\tau_{21} < 0.75$, $0.45 < \tau_{42} < 0.60$, and $\tau_{43} < 0.80$	"
Loose-τ_{43} CR	$\tau_{21} < 0.75$, $\tau_{42} < 0.45$, and $0.80 < \tau_{43} < 0.95$	"
Inverted-$\Delta\eta$ CR	$\tau_{21} < 0.75$, $\tau_{42} < 0.45$, and $\tau_{43} < 0.80$	$\Delta\eta > 1.0$ and $A_{\mathrm{m}} < 0.1$

6.2 Signal region

I define the SR by an event selection that results in the best sensitivity to the pair-produced squark signal relative to the total background, estimated in simulated events. Arriving at the

SR definition is a process called *optimization*. Five observables are particularly well-suited to discriminate between signal and background events, called *selection variables*. First, there are three substructure selection variables, applicable to each of the two leading CA12 jets:

- τ_{21}, which loosely differentiates between jets with and without substructure
- τ_{42}, which differentiates between the four-subjet substructure of squark signal and two-subjet substructure typical of QCD multijet events
- τ_{43}, which differentiates between the four-subjet substructure of squark signal and three-subjet substructure typical of $\mathrm{t\bar{t}}$ events

Second, there are two selection variables that relate the two leading CA12 jets, indexed 1 and 2:

- mass asymmetry, defined $A_{\mathrm{m}} \equiv |m_1 - m_2| / (m_1 + m_2)$, which is expected to be small for signal and $\mathrm{t\bar{t}}$ background events compared to QCD multijet events
- η difference, defined $\Delta\eta \equiv |\eta_1 - \eta_2|$, of which small values indicate that the jets are back-to-back

I optimize the selection variables by systematic trial and error. With five selection variables, it's not only a time-consuming computational task to find the the best cut values simultaneously, it's also an ill-defined problem. Different metrics for signal significance favor different optimal points; additionally, optimization is performed on simulated events, so the derived optimal point is likely just a good estimate, regardless of the metric. In the end, I pick memorable values near enough to the optimal values.

For each of τ_{21}, A_{m}, and $\Delta\eta$, I optimize the cut value by eye from the variable's distribution after cuts have been applied to the four other selection variables. Distributions of this kind are called $N-1$ distributions. The $N-1$ distributions for these three selection

variables are shown in Figure 6.1.[30] The cut values I choose are $\tau_{21} < 0.75$, $A_{\mathrm{m}} < 0.1$, and $\Delta\eta < 1.0$.

I optimize the τ_{42} and τ_{43} selection variables by calculating the Z_{bi} signal significance [65] for combinations of different values of these two variables (after the other selection variable cuts have been applied), using the combination that gives the best overall significance, considering each squark mass. The significance values for different τ_{42} and τ_{43} cuts, for six different squark masses, are shown in Figure 6.2. The different signal masses favor different cut values, but I choose to favor the lighter masses by using the values $\tau_{42} < 0.50$ and $\tau_{43} < 0.80$.

Although the optimization is based on the squark signal, I apply the same selection when analyzing the gluino signal. This is largely a choice of convenience.

The SR selection is tight, meaning that the fraction of preselected events that are selected, called the *selection efficiency*, is quite small. In general, it's more useful to consider the fraction of total events that passed a selection, which is the selection acceptance times efficiency. These values for the signals, plotted in Figure 6.3 (top left), range from $< 0.001\%$ for the 0.1 TeV squark mass signal to $\sim$1% for the 0.8 TeV squark mass signal. Nevertheless, the selection still reconstructs a narrow resonance for the low masses (see Figure 6.3, bottom) because the production cross sections are so large. The full width at half maximum degrades for higher signal masses because the squark decay products are less likely to be completely contained in a single CA12 jet. For each simulated signal and background dataset, the number and percentage of simulated events that pass each step in the SR selection are shown in Table 6.2. Because the simulated QCD multijet datasets have limited statistics after such a tight selection, I define a loose SR by slightly opening up the N-subjettiness cuts; I only use the loose SR for crosschecks.

The simulated $\overline{m}$ distributions for the 200 GeV squark signal and the QCD multijet and

[30] The process of constructing $N-1$ distributions is iterative, because it's inherently circular. I just show the final set of these distributions.

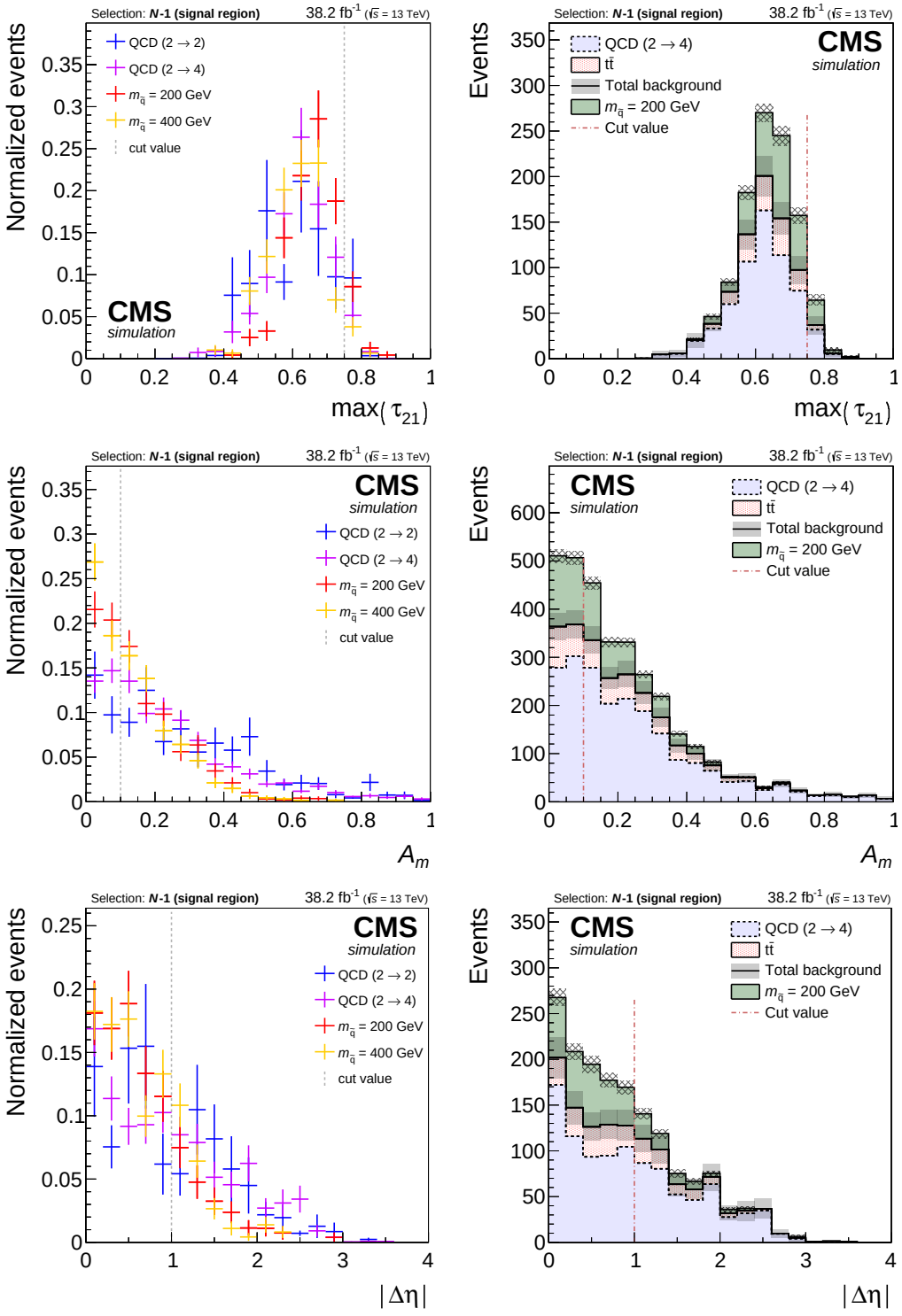


Figure 6.1: The $N-1$ distributions for τ_{21} (top row), A_{m} (middle row), and $\Delta\eta$ (bottom row) with the optimized cut value indicated by a dashed line. The left column compares two squark signals to the two different simulated QCD multijet datasets; the distributions are normalized, so in each case the cut value lies near to the crossover point between the signal and background. The right column shows the 200 GeV squark signal, $\mathrm{t\bar{t}}$ background, and $\mathrm{QCD}_{2\to4}$ background, illustrating the relative quantities of each process after a given cut.

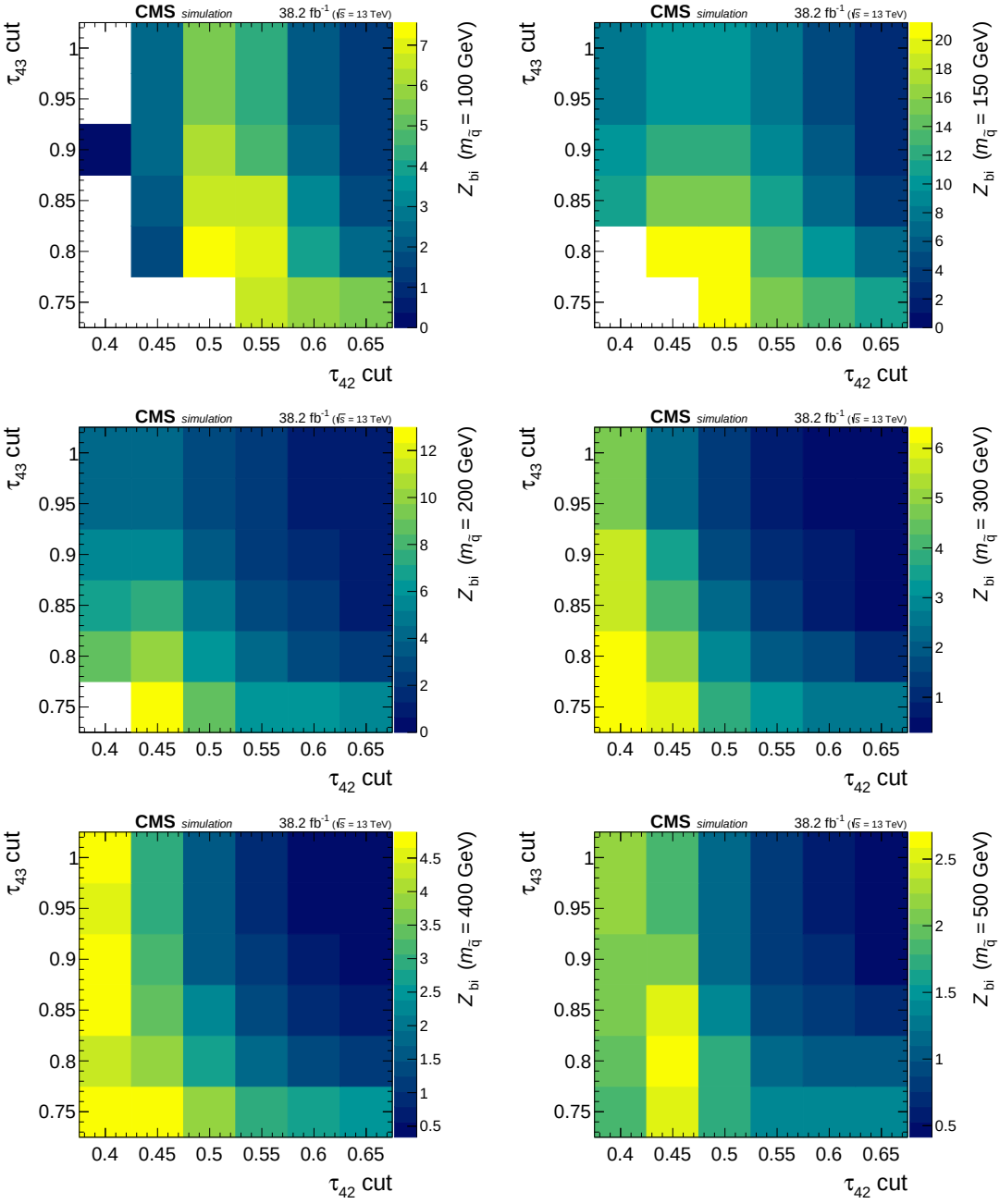


Figure 6.2: Signal significance for the 100, 150, 200, 300, 400, and 500 GeV quark masses (top-left to bottom-right) with respect to simulated $\mathrm{QCD}_{2\to4}$ and $\mathrm{t\bar{t}}$ background, where each bin represents a different pair of cut values for τ_{42} and τ_{43} and the Z_{bi} is calculated from signal and background events in a 50 GeV window around the nominal signal mass.

$t\bar{t}$ backgrounds are shown in Figure 6.3 (top right). The agreement between the simulated events and data is quite poor, but this isn't surprising. The QCD multijet dataset has low statistics in the SR, but producing more events wouldn't fix the disagreement. Simulating QCD is an area of active research; the methods are inexact and computationally challenging. It's unrealistic to expect the simulated QCD multijet $\overline{m}$ distribution to have the same shape and normalization as that of the data, since the SR consists of such rare events. It's precisely for this reason that I estimate the QCD multijet background with data.

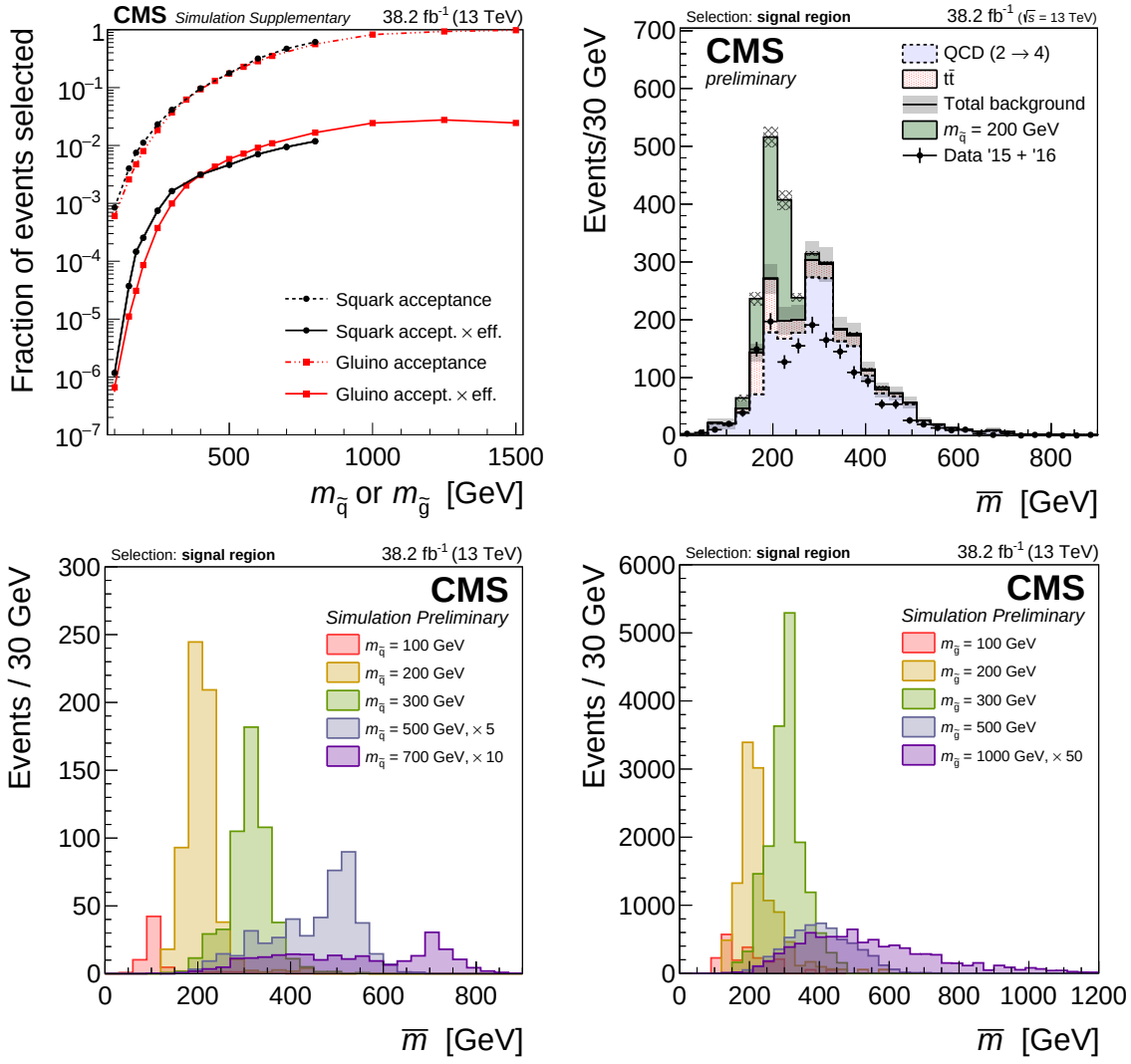


Figure 6.3: The selection acceptance and acceptance times efficiency for the squark and gluino signals (top left), the simulated 200 GeV squark signal and the QCD multijet and $t\bar{t}$ backgrounds (top right), and a subset of the simulated $\overline{m}$ distributions in the SR for the squark (bottom left) and gluino (bottom right) signals.

Table 6.2: The simulated dataset acceptances after each variable selection in the SR. The number of events (N) that pass each selection criterion are normalized to 38.2 fb^{-1}. The percentage values represent the fraction of the total expected number of events without any selection applied; this total isn't well-defined for the binned datasets, so percentages refer to the fraction of preselected events in these cases.

Cut →	**Gen-level $H_T > 700$ GeV**		**Pre-selection**		**$\|\Delta\eta\| < 1.0$**		**$A_m < 0.1$**		**τ_{ij} cuts**	
Sample ↓	N	%	N	%	N	%	N	%	N	%
Squark signals										
$m_{\tilde{q}} = 100$ GeV	180035.5	0.31 %	49242.6	0.085 %	36230.1	0.062 %	10237.9	0.018 %	68.1	< 0.001 %
$m_{\tilde{q}} = 150$ GeV	155025.2	1.63 %	38151.5	0.40 %	28232.9	0.30 %	8541.3	0.090 %	355.2	0.004 %
$m_{\tilde{q}} = 175$ GeV	136752.1	2.95 %	34839.2	0.75 %	25696.0	0.55 %	8008.7	0.17 %	676.7	0.015 %
$m_{\tilde{q}} = 200$ GeV	118885.9	4.83 %	27498.4	1.12 %	20280.7	0.82 %	6231.7	0.25 %	625.7	0.025 %
$m_{\tilde{q}} = 250$ GeV	89919.2	10.9 %	19123.6	2.32 %	14057.0	1.70 %	4365.2	0.53 %	615.2	0.075 %
$m_{\tilde{q}} = 300$ GeV	68882.3	21.2 %	13461.5	4.14 %	9908.9	3.05 %	2930.8	0.90 %	528.6	0.16 %
$m_{\tilde{q}} = 400$ GeV	47012.9	67.1 %	6801.7	9.71 %	4935.0	7.04 %	1259.6	1.80 %	219.7	0.31 %
$m_{\tilde{q}} = 500$ GeV	18916.6	95.6 %	3516.4	17.8 %	2569.1	13.0 %	623.8	3.15 %	91.8	0.46 %
$m_{\tilde{q}} = 600$ GeV	6631.5	99.5 %	2116.1	31.7 %	1536.1	23.0 %	342.2	5.13 %	47.2	0.71 %
$m_{\tilde{q}} = 700$ GeV	2558.2	99.9 %	1200.3	46.9 %	876.4	34.2 %	191.4	7.48 %	24.1	0.94 %
$m_{\tilde{q}} = 800$ GeV	1081.8	100.0 %	663.3	61.3 %	488.2	45.1 %	102.9	9.51 %	12.8	1.18 %
Gluino signals										
$m_{\tilde{g}} = 100$ GeV	7222606.0	0.23 %	1897783.6	0.060 %	1223758.9	0.039 %	234125.7	0.007 %	2083.5	< 0.001 %
$m_{\tilde{g}} = 150$ GeV	5637964.0	1.10 %	1323531.4	0.26 %	860051.8	0.17 %	197312.7	0.039 %	5675.4	0.001 %
$m_{\tilde{g}} = 175$ GeV	5293466.0	2.13 %	1171079.2	0.47 %	758280.2	0.31 %	190891.5	0.077 %	7638.7	0.003 %
$m_{\tilde{g}} = 200$ GeV	5082827.0	3.87 %	1049802.5	0.80 %	682156.1	0.52 %	178085.8	0.14 %	11267.9	0.009 %
$m_{\tilde{g}} = 250$ GeV	4508775.5	10.4 %	804533.7	1.85 %	521390.8	1.20 %	137612.3	0.32 %	16291.9	0.037 %
$m_{\tilde{g}} = 300$ GeV	3994372.5	23.5 %	631872.3	3.72 %	406965.8	2.40 %	100189.2	0.59 %	16902.0	0.099 %
$m_{\tilde{g}} = 350$ GeV	3326991.8	44.7 %	461400.6	6.20 %	300753.6	4.04 %	79689.4	1.07 %	15110.6	0.20 %
$m_{\tilde{g}} = 400$ GeV	2588439.8	72.4 %	334434.0	9.35 %	220740.1	6.17 %	56308.4	1.57 %	11091.1	0.31 %
$m_{\tilde{g}} = 450$ GeV	1666310.1	90.7 %	239671.9	13.1 %	159576.5	8.69 %	41121.2	2.24 %	7926.5	0.43 %
$m_{\tilde{g}} = 500$ GeV	970937.8	97.4 %	173736.2	17.4 %	118676.9	11.9 %	29742.0	2.98 %	5808.9	0.58 %
$m_{\tilde{g}} = 550$ GeV	561272.4	99.3 %	129250.5	22.9 %	89307.1	15.8 %	22098.1	3.91 %	4082.1	0.72 %
$m_{\tilde{g}} = 600$ GeV	330795.9	99.8 %	94586.8	28.5 %	66301.1	20.0 %	16535.2	4.99 %	3022.3	0.91 %
$m_{\tilde{g}} = 650$ GeV	200704.3	99.9 %	70910.5	35.3 %	50374.8	25.1 %	12488.7	6.22 %	2190.5	1.09 %
$m_{\tilde{g}} = 800$ GeV	56853.8	100.0 %	32057.4	56.4 %	23553.1	41.4 %	5604.2	9.86 %	948.8	1.67 %
$m_{\tilde{g}} = 1000$ GeV	12423.3	100.0 %	10231.9	82.4 %	7627.8	61.4 %	1581.6	12.7 %	302.9	2.43 %
$m_{\tilde{g}} = 1250$ GeV	2394.0	100.0 %	2236.9	93.4 %	1706.8	71.3 %	370.1	15.5 %	66.2	2.76 %
$m_{\tilde{g}} = 1500$ GeV	541.8	100.0 %	531.4	98.1 %	408.4	75.4 %	81.1	15.0 %	13.3	2.45 %
Backgrounds										
$t\bar{t}$	-	-	122562.1	0.44 %	70531.3	0.25 %	12159.6	0.044 %	353.3	0.001 %
$QCD_{2\to4}$	-	-	27046808.0	100.0 %	14115947.0	52.2 %	1558676.0	5.76 %	1929.1	0.007 %
$QCD_{2\to2}$	-	-	24538206.0	100.0 %	15214732.0	62.0 %	1368669.2	5.58 %	1357.3	0.006 %
$W^{\pm}$ + jets	-	-	262646.9	100.0 %	147052.0	56.0 %	14342.6	5.46 %	104.0	0.040 %
Z + jets	-	-	23309.1	100.0 %	13581.3	58.3 %	1392.8	5.98 %	5.9	0.025 %

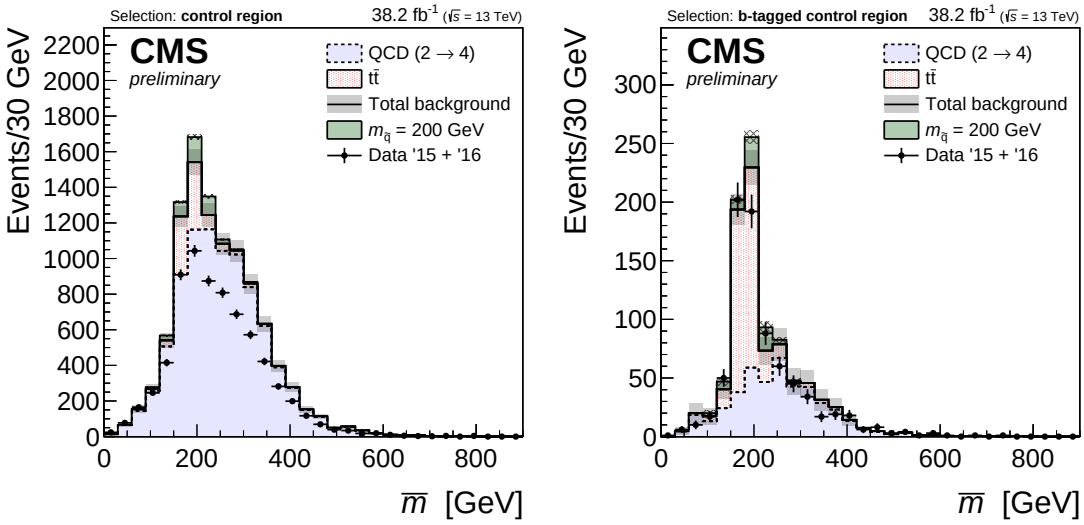


Figure 6.4: Distributions of $\overline{m}$ from simulated events and data in the inclusive (left) and b-tagged (right) CRs.

6.3 Control regions

I use the CRs to estimate the systematic uncertainties that I apply to the shape and normalization of the SR $\overline{m}$ PDFs in my final results. I define two nominal CRs that border the SR in τ_{42}–τ_{43} space: a QCD-multijet dominated region called the *inclusive CR*, and a $t\bar{t}$-dominated region called the *b-tagged CR*. The definition for the inclusive CR is shown in Table 6.1. The b-tagged CR is defined as a subregion of the inclusive CR by including a loose b-tag requirement of $d_{\text{CSVv2}} < 0.5426$ in the inclusive CR fatjet tag. There's no official method for applying the b-tagging algorithm to CA12 jets, and so I assign to each CA12 jet a d_{CSVv2} corresponding to the minimum d_{CSVv2} value of all AK4 jets that lie within a distance of 1.0 from its center, measured in η–ϕ space.

Although the CRs are defined to classify background events, they still could potentially contain signal events. The signal events present in a CR are referred to as *contamination*. The maximum contamination I expect from a squark signal in the inclusive CR is ∼4%, resulting from the 200 GeV squark; the expected b-tagged CR contamination is higher, at ∼10%, resulting from the 150 GeV squark (see Appendix C). Although the potential contribution of this contamination isn't large, I still include a procedure to prevent it affecting

the measurements of the systematic uncertainties in the CRs (see Section 7.4).

In addition to the nominal CRs, I define eight other CRs by defining four alternate CRs, each paired with a b-tagged version (see Table 6.1). These CRs are exclusively used to test the QCD multijet estimation procedure; they don't play a role in the final results.

Chapter 7

Backgrounds

I estimate the contribution of each background in the data differently. For a given analysis region, I take the $t\bar{t}$ $\overline{m}$ PDF to be the simulated $t\bar{t}$ fatjet-pair $\overline{m}$ distribution, I ignore the $W^{\pm}$ and Z boson contributions to the background (see Section 7.1), and I construct the $\overline{m}$ PDF for the QCD multijet background from the data (see Section 7.2).

I compare the estimated total background $\overline{m}$ PDF to the $\overline{m}$ distribution it's intended to describe in a procedure called a *closure* test. Using each simulated QCD multijet dataset, I perform a closure test in the SR and the inclusive CR, comparing the constructed QCD multijet $\overline{m}$ PDF and the simulated $\overline{m}$ distribution. These closure tests highlight modest differences between the estimated and observed distributions in simulation and demonstrate that the fitting procedure successfully brings the two into agreement. I also perform closure tests in the data, comparing the QCD multijet and $t\bar{t}$ $\overline{m}$ PDFs to the observed $\overline{m}$ distribution in the CRs. These closure tests further validate the fitting method and provide important estimations of the uncertainties associated with the rate and shape of the QCD multijet and $t\bar{t}$ backgrounds.

7.1 Contribution: $W^{\pm}$ and Z bosons

The $W^{\pm}$ and Z boson background contributions in the SR and CRs are negligible. The expected contributions of different $W^{\pm}$ and Z boson processes in each analysis region, based on simulation, are shown in Figure 7.1. In the SR, for example, I expect 104.0 ± 19.9 $W^{\pm}$ + jets events and 5.9 ± 1.2 Z + jets events, compared to the 2007.5 ± 136.0 events I

expect from QCD multijet events. The total contribution from $W^{\pm}$+jets isn't insignificant, but its $\overline{m}$ distribution is very similar to that of QCD multijet events. Since the QCD multijet $\overline{m}$ PDF normalization is a fit parameter that's left to float in the final result, I neglect these vector boson processes altogether, assuming that they're included in the QCD multijet background component prediction.

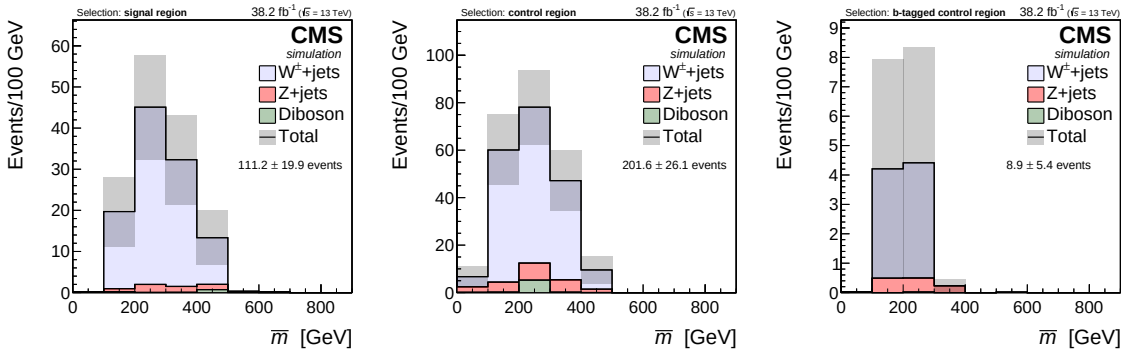


Figure 7.1: The simulated $W^{\pm}$ and Z boson $\overline{m}$ distributions in the SR (left), inclusive CR (middle), and b-tagged CR (right). The total number of events predicted in each analysis region is indicated on the respective plot, contributions that are negligible when compared to the expected QCD multijet and $t\bar{t}$ events.

7.2 Contribution: QCD multijet

For a given analysis region, I construct the QCD multijet background $\overline{m}$ PDF from the data in order to avoid the mismodeling of simulation. The basic procedure is to draw individual fatjets from a high-statistics collection of data events and to form artificial pairs between them. I identify events in which at least the leading CA12 jet satisfies the particular analysis region's fatjet tag; the number of such events is more than an order of magnitude greater than those required to have a fatjet pair. I form a collection of fatjets by taking the leading fatjet from each of these events. From this collection of leading fatjets from different events, I form all possible fatjet pairs—each pair must satisfy the fatjet pairing requirements on A_{m} and $\Delta\eta$. The $\overline{m}$ distribution of these artificial fatjet pairs is an $\overline{m}$ PDF for QCD multijet events.

This construction treats the m distribution of leading fatjets as an m PDF, denoted $P(m)$, from which two fatjet masses are sampled to form an $\overline{m}$ PDF, denoted $P_{\text{avg}}(\overline{m})$:

$$P_{\text{avg}}(\overline{m}) = \int_0^{2\overline{m}} P(x) \cdot P(2\overline{m} - x) \cdot \theta\left(0.1 - \left|\frac{x - \overline{m}}{\overline{m}}\right|\right) dx, \tag{7.1}$$

where θ is the Heaviside step function that imposes the A_{m} fatjet pair requirement; while not represented in Equation 7.1, I impose the $\Delta\eta$ fatjet pair requirement by hand when I implement this construction in practice.

Fatjets from QCD multijet events don't contain the products of partons with well-defined masses, and so the $P(m)$ derived from them has a significant p_{T} dependence that has to be taken into account. The correlation between the p_{T} and mass of leading QCD multijet fatjets is shown in Figure 7.2. I correct for this effect by constructing different $\overline{m}$ PDFs from the different groups of events that fall into consecutive H_{T} windows; I then combine these $\overline{m}$ PDFs with different weights according to the analysis region's observed H_{T} distribution. Figure 7.3 shows $P(m)$ and the derived $P_{\text{avg}}(\overline{m})$ with and without this H_{T} reweighting (left) and the observed H_{T} distribution (right), demonstrating the significant effect of the reweighting on the high-$\overline{m}$ tail of the distribution.

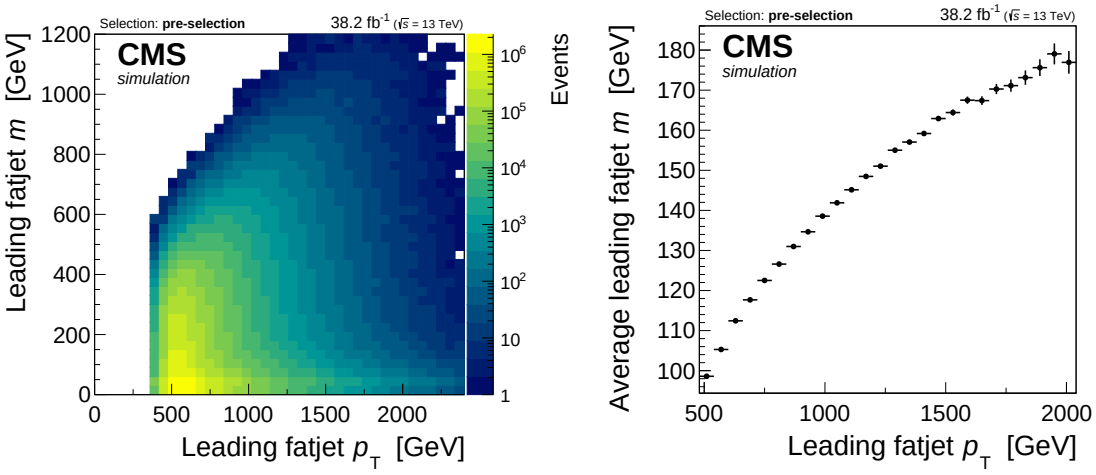


Figure 7.2: Distributions of p_{T}-leading fatjet p_{T} and mass (left) and the associated profile plot (right), which shows that the mass of a fatjet is on average greater for larger p_{T} values.

A final worry is the contamination of the QCD multijet $\overline{m}$ PDF from signal or other

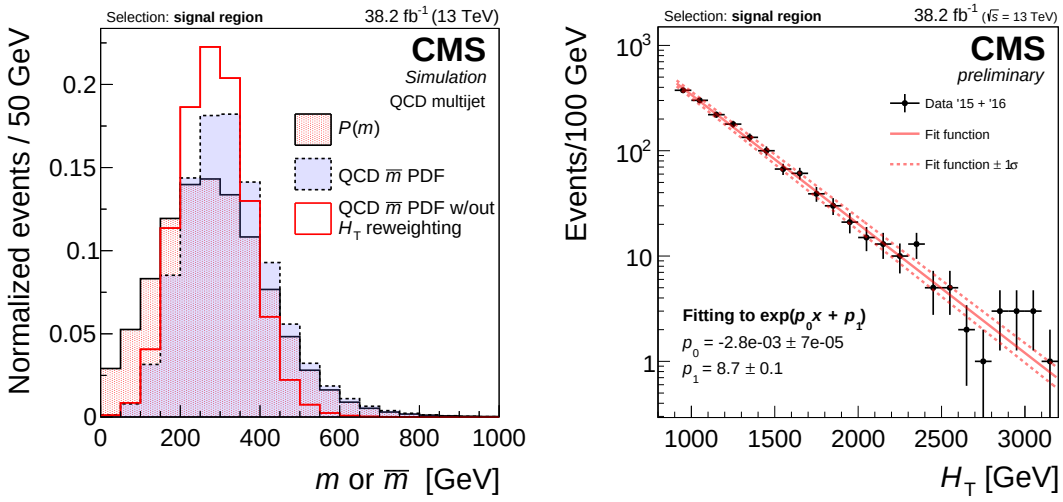


Figure 7.3: Left: the m distribution of the tagged p_{T}-leading fatjets, $P(m)$, in the SR and the $\overline{m}$ PDFs derived from these fatjets with and without H_{T} reweighting. Right: the H_{T} distribution in the SR of the data and the exponential fit from which I derive the H_{T}-reweighting factors.

backgrounds. There's nothing in the construction that prevents fatjets from $\mathrm{t\bar{t}}$ events, for example, from contributing to it. This worry is assuaged by noting the numerical dominance of QCD multijet events over all other processes in the events used to construct $P(m)$. In any case, a systematic effect from contaminating events will be incorporated in the results of the fit of the background $\overline{m}$ PDFs to the data.

In practice, I evaluate Equation 7.1 and the requisite H_{T} reweighting through a binned procedure. First I construct a three-dimensional histogram consisting of the leading fatjets' m and η and the event's H_{T} (see Figure 7.4). For each H_{T} bin in the histogram, I perform the integral in Equation 7.1 by calculating the $\overline{m}$ of all pairs of bins that pass the fatjet pair requirement, evaluated using the bin's central m and η values. The contribution of each bin pair to $P_{\mathrm{avg}}(\overline{m})$ is weighted by the product of the two bins' integrals, i.e. their respective number of fatjets. The resulting $P_{\mathrm{avg}}(\overline{m})$ collection is summed with weights derived from the fit to the sample's H_{T} distribution.

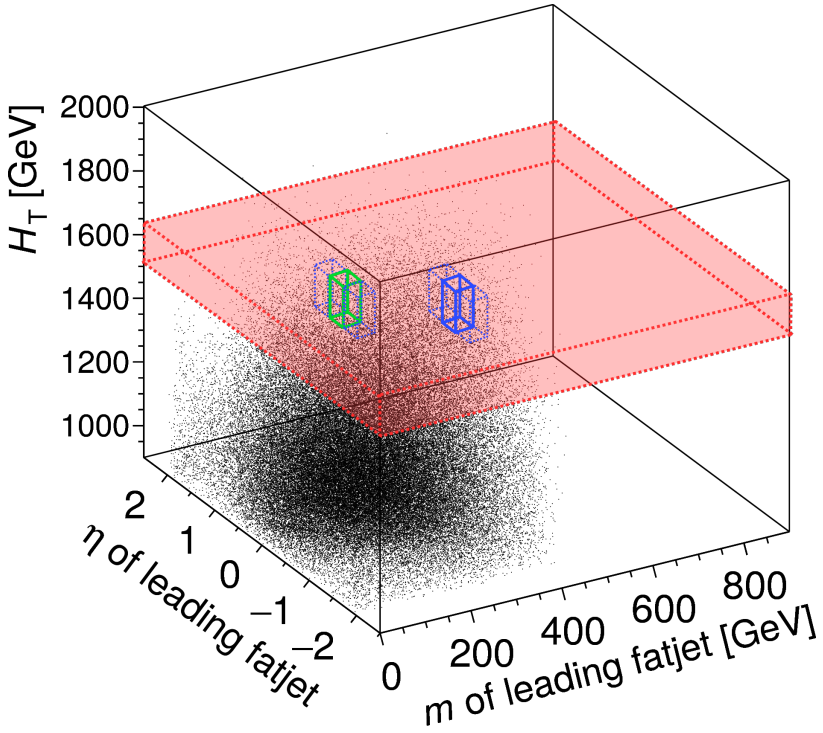


Figure 7.4: An illustration of the practical QCD multijet $\overline{m}$ PDF construction, using a histogram of leading fatjet variables and event H_{T}. A three-dimensional histogram is necessary: m in order to calculate a fatjet pair's $\overline{m}$ and A_{m}, η to calculate a fatjet pair's $\Delta\eta$, and H_{T} to apply the H_{T} reweighting. For each each H_{T} bin in the histogram (shaded red region), I calculate $P_{\mathrm{avg}}(\overline{m})$ by forming every possible pair of m–η bins (green and blue boxes are examples).

7.3 Closure tests in simulation

I conduct closure tests between the expected and observed $\overline{m}$ distributions using simulated QCD multijet events. In this case, the expected distribution is the QCD multijet $\overline{m}$ PDF, and the observed distribution is the fatjet pair $\overline{m}$ distribution. In the loose SR[31] and the inclusive CR, I fit the expected distribution to the observed distribution, quantifying the result with the post-fit values of the fit parameters and the *pull* distribution, defined in each $\overline{m}$ bin as

$$\text{pull} \equiv \frac{N_{\text{obs}} - N_{\text{exp}}}{\delta N_{\text{obs}}}, \tag{7.2}$$

where N_{obs} is the number of events in the observed distribution's bin, N_{exp} is the number of events expected distribution's bin, and δN_{obs} is the statistical uncertainty associated with N_{obs}. I don't assign any uncertainty to the expected distribution, and so the pull uncertainty is exactly 1.

The results of the simulated QCD multijet closure test are shown in Figure 7.5. For the loose SR in the $\text{QCD}_{2\to4}$ dataset, the $\overline{m}$ PDF needs to be shifted to the left by approximately 15 GeV and compressed approximately 5% to best match the observed distribution. The inclusive CR requires similar adjustments. The $\text{QCD}_{2\to2}$ dataset closure requires larger corrections than the $\text{QCD}_{2\to4}$, but these values are still consistent between the loose SR and inclusive CR. The better agreement in the high-$\overline{m}$ tail and the higher statistics of the $\text{QCD}_{2\to4}$ dataset are the main reasons that I used the $\text{QCD}_{2\to4}$ dataset for optimization. For each pull distribution, I calculate the χ^2 per number of degrees of freedom (χ^2/NDF), but I only calculate it up to 600 GeV because including the large uncertainties in the high-$\overline{m}$ tail would make these values artificially small. The χ^2/NDF of each closure test is reasonable, indicating what can be seen by eye: the fit parameters I choose can adequately describe all

[31] I show the results of this test in the loose SR because there are more statistics there than the SR. The SR results are good for $\text{QCD}_{2\to4}$, but the lower statistics of $\text{QCD}_{2\to2}$ events are much more limiting (see Appendix D).

of the significant differences between the expected and observed distributions.

7.4 Closure tests in data

I use closure tests in the CRs of the data to demonstrate the accuracy of the background estimation. These closure tests are similar to those for the QCD multijet simulation, but in the data it's necessary to include a $t\bar{t}$ $\overline{m}$ PDF as well as the QCD multijet $\overline{m}$ PDF. I assign independent fit parameters to each of these $\overline{m}$ PDFs and fit their sum to the observation. The results of the closure tests in the inclusive and b-tagged CRs of the data are shown in Figure 7.6. The shift parameters in the CRs are consistent with zero; the stretch parameters impose a broadening of the $\overline{m}$ PDFs, but there's a known $\sim$10% underestimation of the jet energy resolution in simulation [55], so this effect is expected. The pull distributions demonstrate that the background estimation method reproduces the data in the CRs. The success of the method in the inclusive CR, which has much greater statistical precision than the SR, is particularly comforting.

The post-fit values of the $\overline{m}$ PDF fit parameters are meaningful. The $t\bar{t}$ fit parameters have physical interpretations: the normalization parameter describes the difference in the fatjet-tagging acceptance between data and simulation, and the shift and stretch parameters are measures of the fatjet-mass scale and fatjet-mass resolution, respectively. The QCD multijet shift and stretch fit parameters reflect minor uncorrected correlations between the two leading fatjets in an event.

Although the data closure tests demonstrate agreement between the estimated backgrounds and observation, signal contamination in the CRs could contribute to the values of the fit parameters, thereby inflating the systematic uncertainties I assign to the $\overline{m}$ PDFs in the SR. I account for this possibility by repeating the closure tests for each squark signal under the assumption that the signal is present in the observed $\overline{m}$ distribution. In each case, I include the signal's simulated $\overline{m}$ distribution as a third $\overline{m}$ PDF in the fit, assigning it an independent normalization but the same shift and stretch parameters used by the $t\bar{t}$

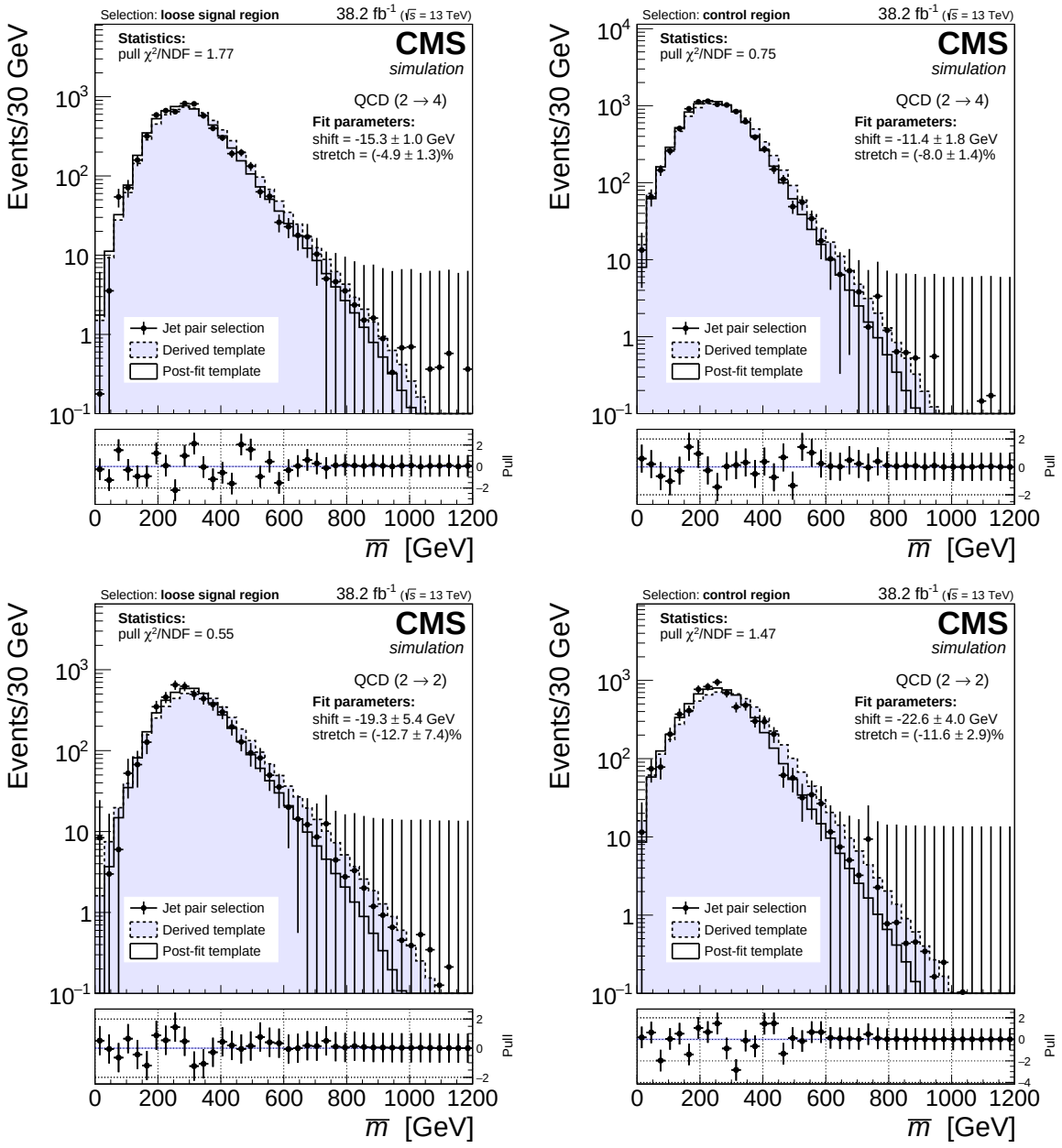


Figure 7.5: Comparisons between the QCD multijet $\overline{m}$ PDF, referred to as "template" in the legends, and the fatjet pair $\overline{m}$ distributions in simulated QCD multijet events from the $\text{QCD}_{2\to4}$ (top) and $\text{QCD}_{2\to2}$ (bottom) datasets in the loose SR (left) and inclusive CR (right). The same plots for the SR are shown in Appendix D (see Figure D.1).

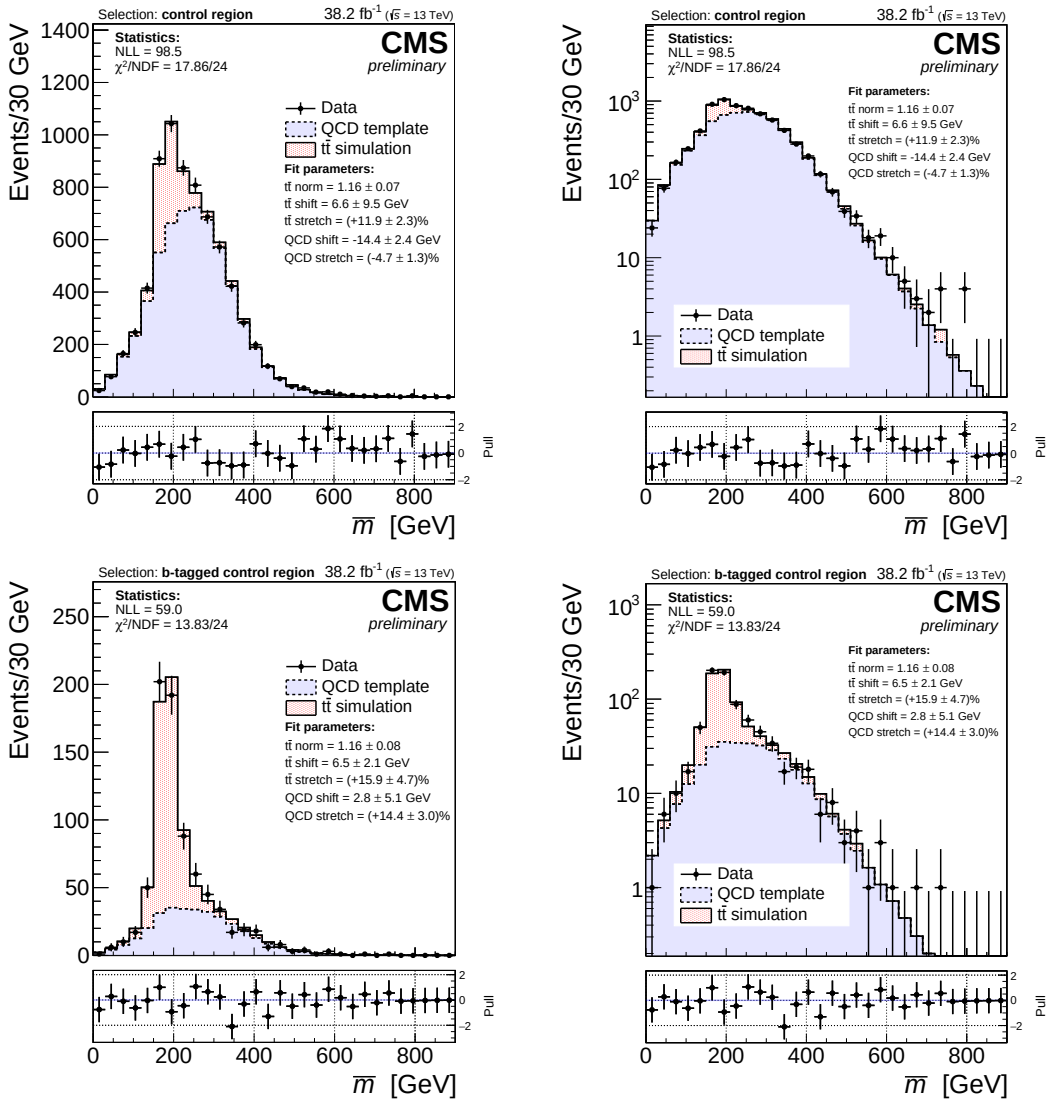


Figure 7.6: Closure tests in the inclusive (top) and b-tagged (bottom) CRs of the data. The post-fit values of the fit parameters are indicated on each plot, except for the arbitrary QCD multijet normalization.

$\overline{m}$ PDF; the results of this procedure for the 150 GeV squark signal are shown in Figure 7.7.

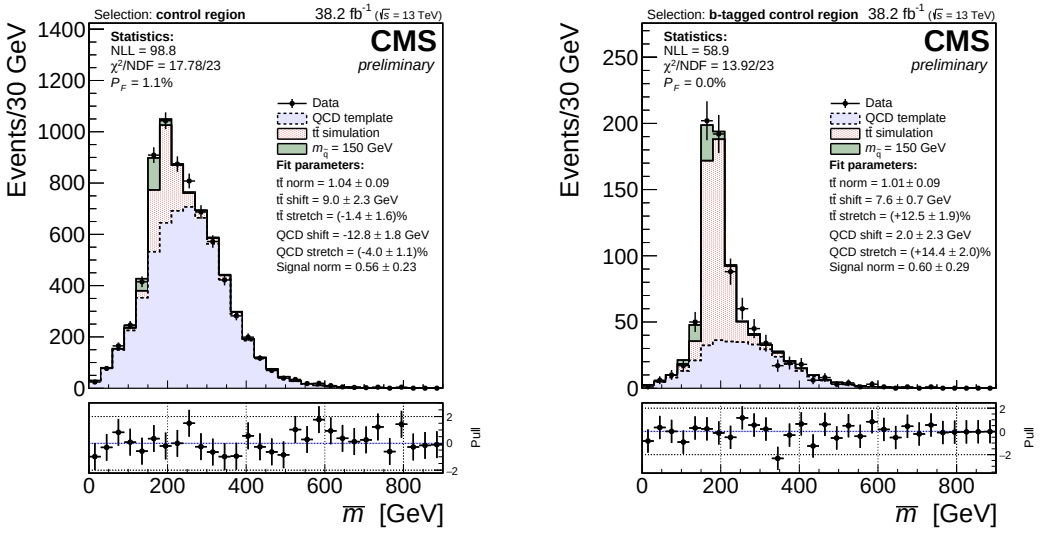


Figure 7.7: Closure tests including the 150 GeV squark signal in the inclusive (left) and b-tagged (right) CRs. The P_F of each F-test, relative to the null hypotheses shown in Figure 7.6, are noted on each plot; the P_F values are less than 5% in both cases, meaning that the null hypotheses are better descriptions of the data. I perform the same tests for each squark signal, leading to the same conclusion in each case.

In each CR, I compare the result of each signal-included closure test to the no-signal (*null hypothesis*) result presented in Figure 7.6. I quantify the comparison with an *F-test* [66]. For each squark signal, I calculate

$$F_{\text{sig}} \equiv \left(\frac{\chi^2_{\text{null}} - \chi^2_{\text{sig}}}{\chi^2_{\text{sig}}} \right) \left(\frac{n - p_{\text{sig}} - 1}{p_{\text{sig}} - p_{\text{null}}} \right), \tag{7.3}$$

where χ^2_{null} and χ^2_{sig} are the χ^2 values between the post-fit estimate and the data in the null hypothesis and signal hypothesis cases, respecively, p_{null} and p_{sig} are the number of fit parameters in the null hypothesis and signal hypothesis closure tests, respectively, and n is the number of $\overline{m}$ bins. The p_{null}, p_{sig}, and n parameters define a regularized incomplete beta function called the *F-distribution*, $F\left(F_{\text{sig}}, p_{\text{sig}} - p_{\text{null}}, n - p_{\text{sig}}\right)$, a cumulative distribution function describing the probability that the fit involving the signal hypothesis is a better description than the null hypothesis, for a given value of F_{sig}, called the *F-probability* (P_F).

If P_F is less than 5%, then the null hypothesis is taken to be the best description of the data; otherwise, I should use the signal hypothesis closure results when estimating systematic uncertainties. The F-tests demonstrate that the null hypothesis should be assumed in both CRs.

To crosscheck this procedure, I artificially add a signal $\overline{m}$ distribution to that of the data (*signal inject*) and repeat the closure tests. The 150 GeV signal-injected closure and F-test results for the null and signal hypotheses are shown in Figure 7.8. The signal-injected results consistently demonstrate that the signal hypothesis corresponding to the injected signal has the highest P_F and that P_F is greater than 5%, as expected. Therefore, I'm confident that there isn't significant signal contamination in the data CRs.

I study the limitations of the background estimation procedure by performing the null hypothesis closure test in each of the alternative CRs of data, defined in Table 6.1. In the inclusive loose CR (see Figure 7.9), the QCD multijet estimate disagrees with observation below $\overline{m} \lesssim 100\,\text{GeV}$. The loose CR therefore demonstrates a qualitative upper limit on the statistics and difference in N-subjettiness from the SR for alternate CR choices. The loose-τ_{42} and loose-τ_{43} CRs (see Figures 7.10–7.11) each emphasize a contribution to the nominal CRs from a particular N-subjettiness region. The closure in these two regions is very good, which demonstrates that the background estimation isn't very sensitive to the particular substructure of the fatjets. Finally, the closure tests for the inverted-$\Delta\eta$ CRs (see Figure 7.12) illustrate the reasonable success of the background estimation in a region where I don't expect very accurate fatjet-mass reconstruction. The successful closure tests for these alternative CRs lend confidence to the positive results of the nominal CRs and show that while the CR definition matters it doesn't have to be chosen particularly precisely.

I summarize the post-fit values of the fit parameters for each background contribution in each control region in Figures 7.13–7.14. There's reasonable agreement between the normalization and stretch parameters for each control region and background, but the shift

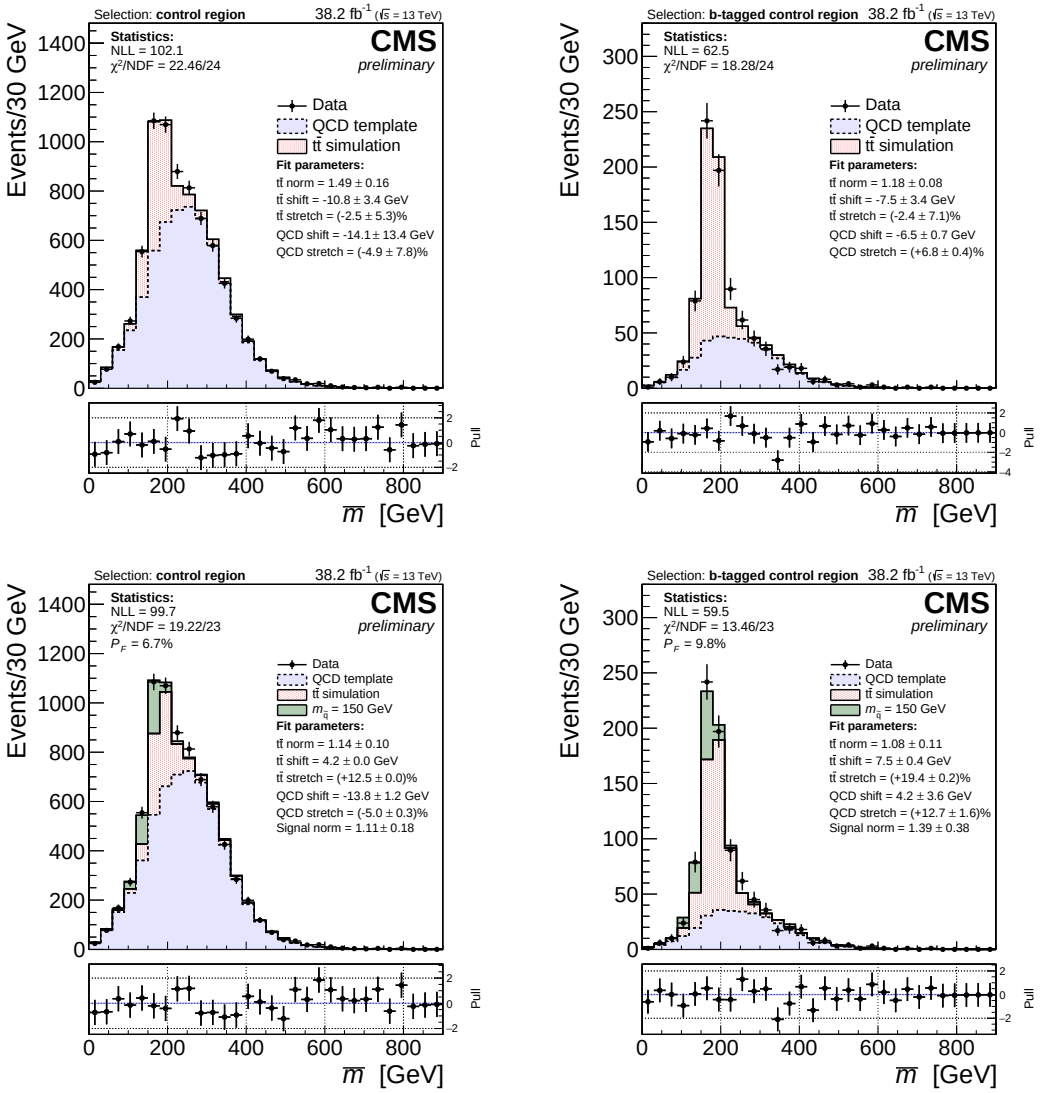


Figure 7.8: Closure tests in the CRs after the 150 GeV squark signal has been injected into the data, in order to crosscheck the signal contamination mitigation. The null-hypothesis closure tests (top) in the inclusive (left) and b-tagged (right) CRs demonstrate that despite signal contamination, the expected background can still agree with the data, but the inflation of the $t\bar{t}$ normalization parameter hints at a problem. The signal-hypothesis closure tests (bottom) show that P_F is greater than 5%, and so the procedure would be to extract the systematic uncertainties from these post-values, not those of the null hypothesis.

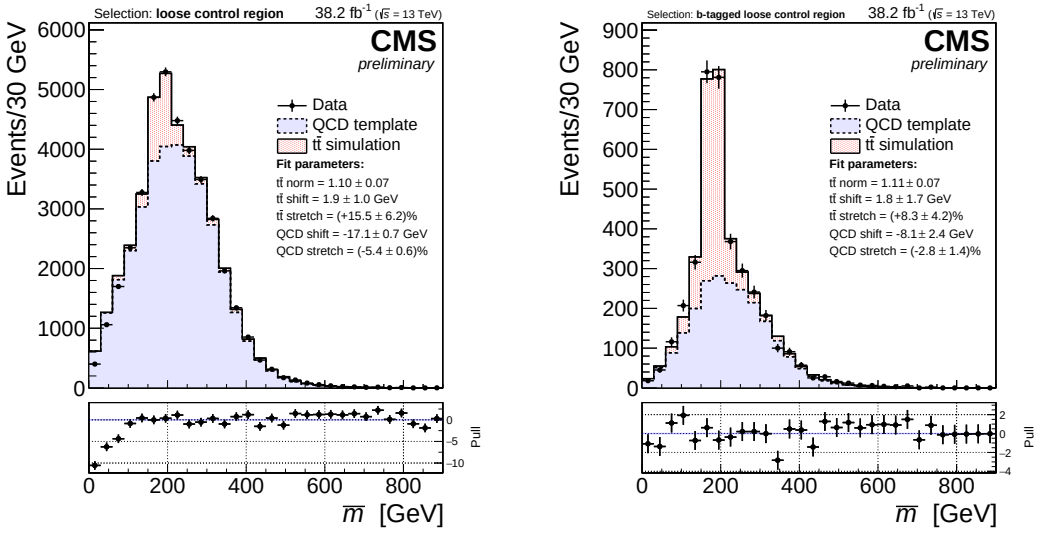


Figure 7.9: Background estimation in data loose control regions

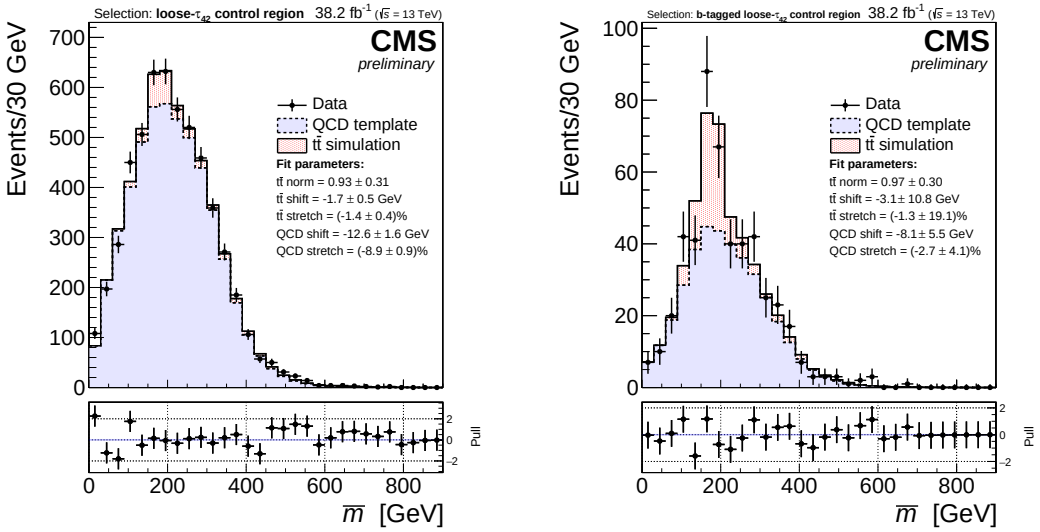


Figure 7.10: Background estimation in data loose-τ_{42} control regions

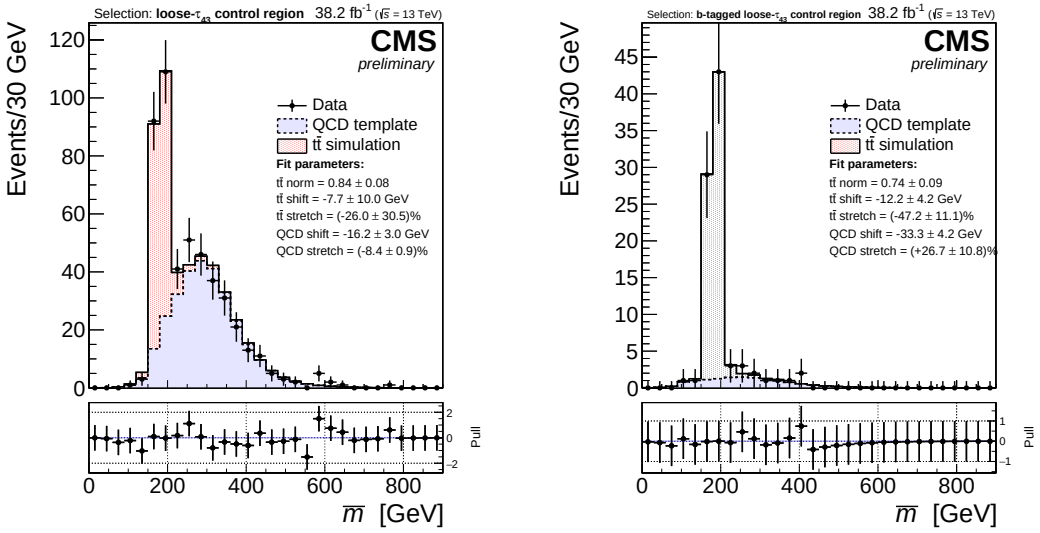


Figure 7.11: Background estimation in data loose-τ_{43} control regions

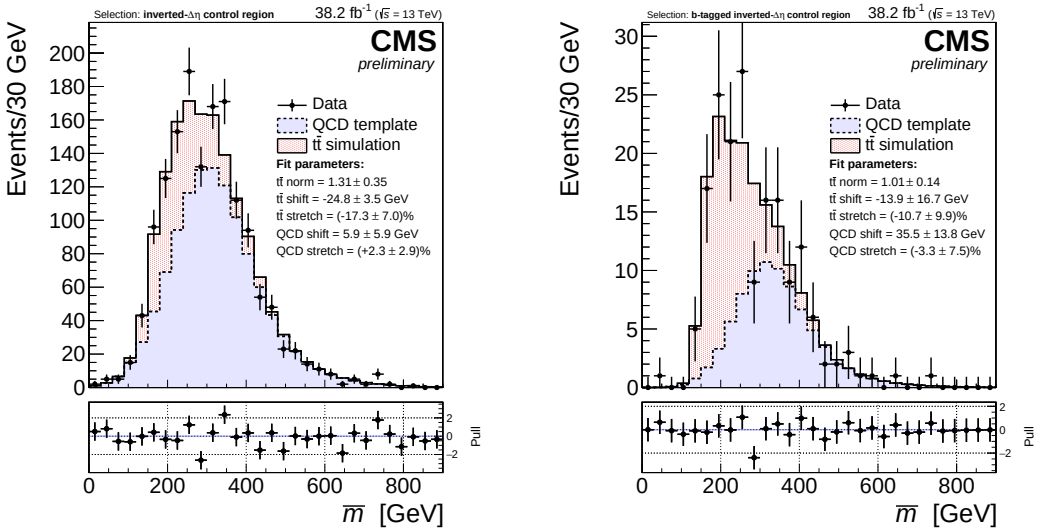


Figure 7.12: Background estimation in data inverted-$\Delta\eta$ control regions

parameters can differ significantly for the inverted-$\Delta\eta$ control region. This is not unexpected, since the cuts that define each region are responsible for shaping the $\overline{m}$ distribution differently.

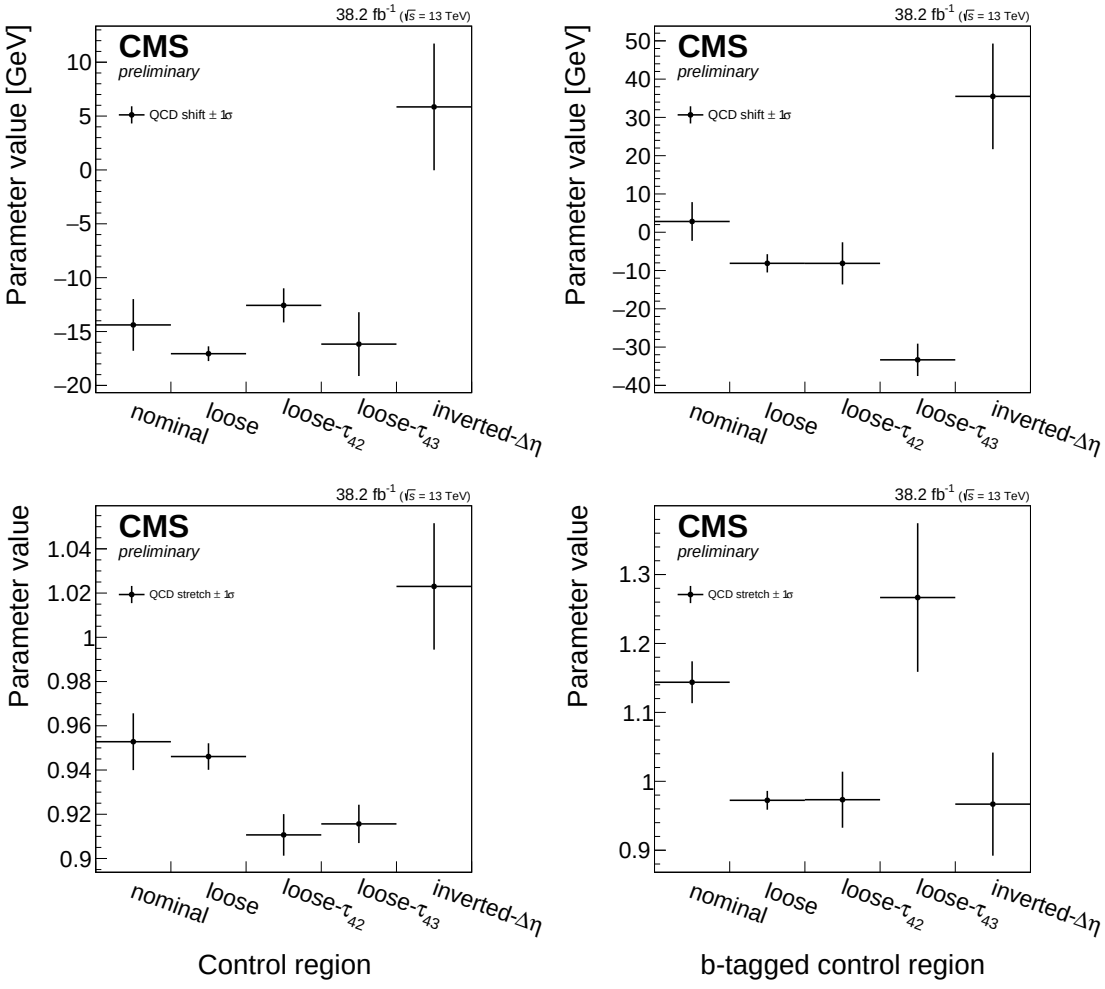


Figure 7.13: The QCD multijet fit parameters for the inclusive (left) and b-tagged (right) CRs.

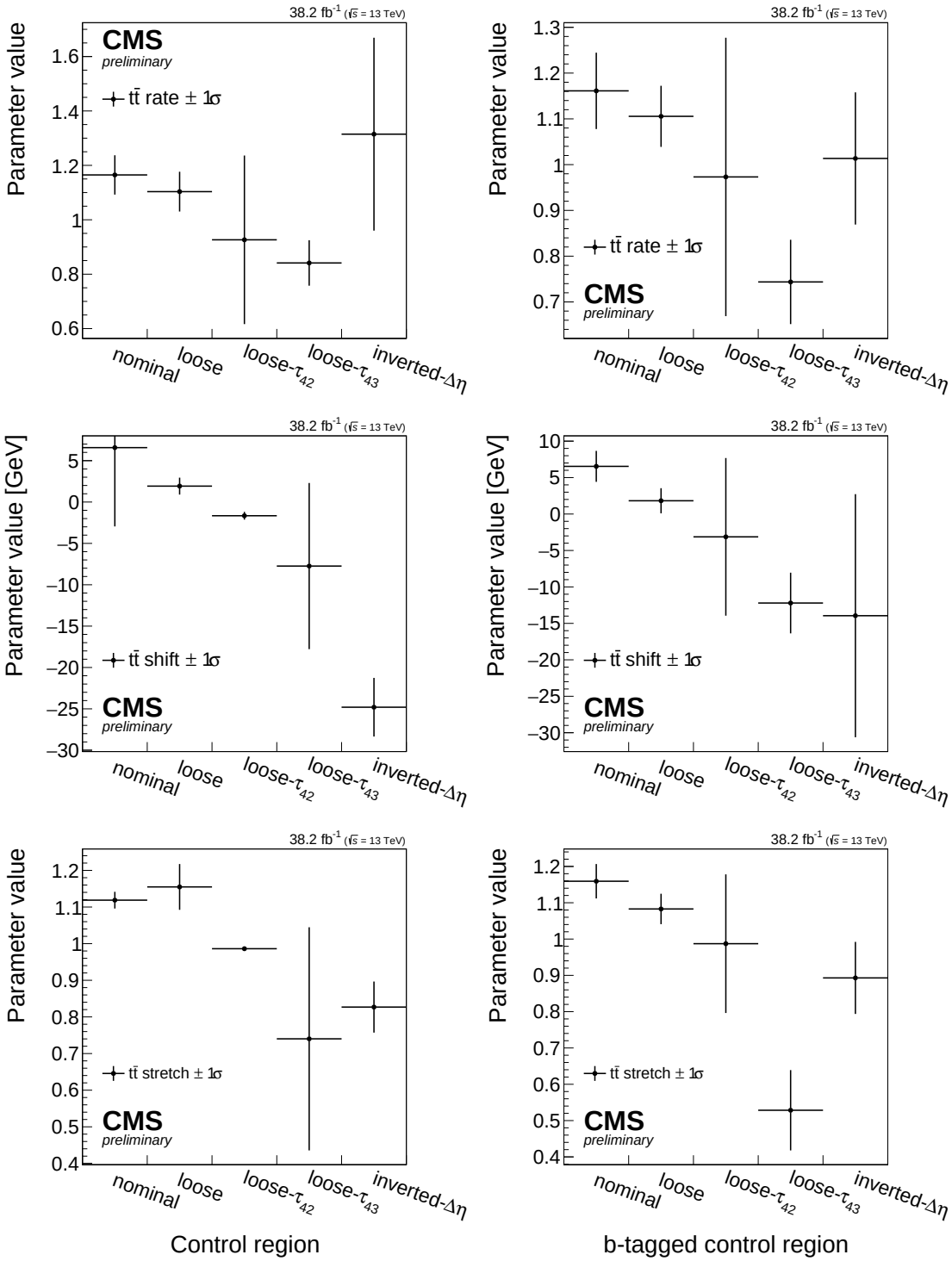


Figure 7.14: The $t\bar{t}$ fit parameters for the inclusive (left) and b-tagged (right) CRs.

Chapter 8

Results

The data don't exhibit an excess of events over the background prediction, meaning I didn't discover one of the signals. This conclusion is already fairly clear from Figure 6.3—the data in the plot in the upper right can't qualitatively accommodate most of the signals shown in the bottom row, even if the amplitude of the simulated QCD multijet $\overline{m}$ distribution is decreased. I demonstrate the lack of a signal discovery in a more robust way by fitting the background components to the data in the SR, checking that the fit closes within the uncertainty I assign to the components. I perform a maximum-likelihood fit of the $t\bar{t}$ and QCD multijet $\overline{m}$ PDFs to the data, allowing each PDF to change based on normalization, shift, and stretch parameters. The initial and final values of these parameters are shown in Table 8.1, and the post-fit $\overline{m}$ distributions are shown in Figure 8.1. The post-fit values of these parameters are consistent with their nominal values, which means that the background components successfully describe all of the data.

8.1 Limits on signal pair production

A small signal could still potentially fit in between the data and the background prediction in the SR. To systematically quantify this possibility, I calculate a statistical upper limit on the amplitude, called *signal strength* (A), of each signal $\overline{m}$ PDF in turn. Each upper limit is interpreted as the upper limit on the particular signal pair-production cross section. If this measured upper limit is lower than the theoretical value of the pair-production cross section, the particular signal hypothesis is ruled out by observation, and so the signal model is said to be *excluded*.

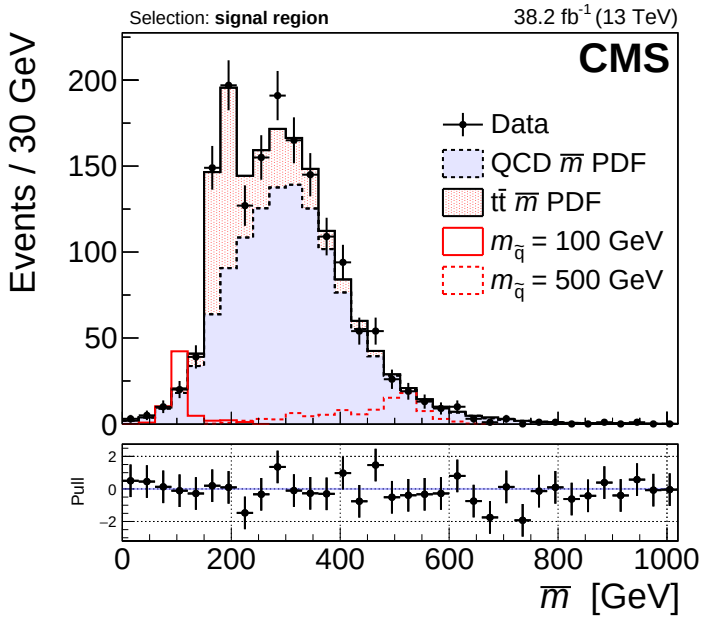


Figure 8.1: The $\overline{m}$ distribution in the data and the predicted (post-fit) background contributions. Additionally, the $\overline{m}$ PDFs for squarks with masses of 100 and 500 GeV are shown.

Using the data in the SR and the background and signal $\overline{m}$ PDFs, I apply a Bayesian method [67, 68] to compute 95% confidence level (CL) upper limits on squark or gluino pair production and decay to non-top quarks. The 95% CL upper limit is the value below which 95% of the PDF of A lies. Dictated by the principle of marginalization, the PDF of A, given the data distribution D, is

$$P(A \,|\, D) = \int_{N_1} \cdots \int_{N_k} P(A, \{N_i\} \,|\, D)\; dN_1 \ldots dN_k, \tag{8.1}$$

where N_i is the i^{th} nuisance parameter, and the *joint posterior distribution*, $P(A, \{N_i\} \,|\, D)$, is derived from Bayes's Theorem as

$$P(A, \{N_i\} \,|\, D) \propto \mathscr{L}(A, \{N_i\}, D)\, \mathcal{P}(A, \{N_i\}), \tag{8.2}$$

where $\mathscr{L}$ is the likelihood and $\mathcal{P}$ is the product of the prior distributions of A and the

nuisance parameters. I perform the computation of Equation 8.1 using the Bayesian Analysis Toolkit (BAT) [69], which performs the marginalization with Markov Chain MC [70] techniques, assuming a Poisson log-likelihood.

For each signal, I assign nuisance parameters that correspond to the fit parameters of the $t\bar{t}$, QCD multijet, and signal $\overline{m}$ PDFs. These nuisance parameters are assigned Gaussian priors except for the QCD multijet normalization, which is assigned a flat prior because I don't have a robust prediction for the total rate of QCD multijet events in the SR; this procedure is reasonable because the QCD multijet shape is much broader than that of the signals, allowing the data in the SR to directly constrain this background. The standard deviation of the QCD multijet $\overline{m}$ PDF shape nuisance parameter priors are derived from the CR data closure tests, but no initial shift or stretch are applied—the means of their prior distributions are set to the nominal values—since these values in the SR are likely different from those measured in the CRs. The $t\bar{t}$ rate and shape nuisance parameter priors are similarly derived from the CR measurements, and I also start the marginalization procedure with their initial values reflecting no shift or stretch, as that's consistent with measurement. Interpreting the $t\bar{t}$ shift and stretch as jet-mass scale and resolution uncertainties, respectively, I apply the same uncertainties to the signal shape by constraining the associated nuisance parameters to match those of the $t\bar{t}$ background. The signal normalization nuisance parameter, which represents the combined uncertainties associated with the signal acceptance and L, is assigned the same standard deviation as that of the $t\bar{t}$ events, since the signal and $t\bar{t}$ production and decay topologies are similar. The means and standard deviations of all of the nuisance parameters are shown in Table 8.1.

I calculate two limits for each signal mass: an *observed limit*, which is derived from the data, and an *expected limit*, which assumes a background-only hypothesis. A significant discrepancy between the two limit values would indicate either an important oversight, such as a mismodeled or absent background, or potentially new physics. To calculate expected limits, I run two hundred toys based on the predicted background distribution

Table 8.1: The nuisance parameters corresponding to each rate and shape parameter of the background and signal PDFs before and after the maximum-likelihood fit. Except for the QCD multijet $\overline{m}$ PDF normalization, which is floating—the value of which is simply the event yield with statistical uncertainty—each nuisance parameter has a Gaussian PDF and is reported as its mean ± its standard deviation.

Parameter	**Pre-fit value**	**Post-fit value**
QCD multijet $\overline{m}$ PDF normalization	floating	1222 ± 35 events
QCD multijet $\overline{m}$ PDF shift	0 ± 17 GeV	-8 ± 4 GeV
QCD multijet $\overline{m}$ PDF stretch	(0 ± 18) %	(-1 ± 3) %
$t\bar{t}$ $\overline{m}$ PDF normalization	1.00 ± 0.24	1.08 ± 0.14
$t\bar{t}$ $\overline{m}$ PDF and signal shift	0 ± 16 GeV	-10 ± 6 GeV
$t\bar{t}$ $\overline{m}$ PDF and signal stretch	(0 ± 20) %	(15 ± 9) %
Signal $\overline{m}$ PDF normalization	1.00 ± 0.24	-

(see Figure 8.1). I assign the expected limit a value and uncertainty corresponding to the median limit and standard deviation, respectively, derived from these toys.

The resulting limits on signal pair production are shown in Figure 8.2. Assuming the top-squark production cross section, squark masses between 0.10 and 0.72 TeV are excluded. Gluinos decaying to five quarks with masses between 0.10 and 1.41 TeV are also excluded. The post-fit total background estimation agrees with the data. The posterior distributions of the nuisance parameters confirm that the background component predictions are not significantly different from the estimates.

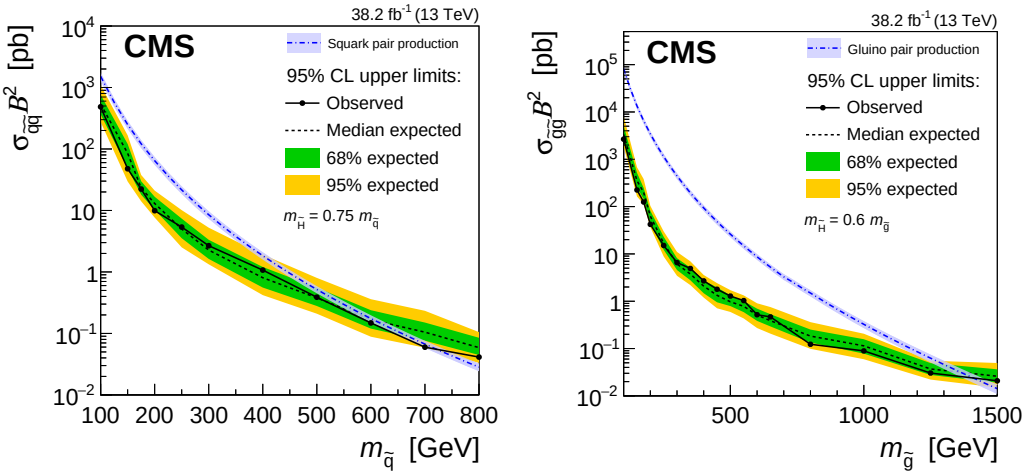


Figure 8.2: The expected and observed limits on the squark (left) and gluino (right) signal production.

Chapter 9

Conclusions

I look for new particles that decay to at least four quarks, but I don't find any. From the data, I derive limits on the pair production of squarks that decay to four quarks and gluinos that decay to five quarks through an RPV coupling. I exclude squarks with masses between 0.10 and 0.72 TeV and gluinos with masses between 0.10 and 1.41 TeV at 95% CL.

Appendix A

Trigger selection

I form a subset of data events that are relevant to my analysis by applying the logical "or" of the triggers shown in Table A.1 to the JetHT datasets. These triggers select events with significant jet activity. The cut values of the variables used by each trigger are chosen by a dedicated group in CMS with the goal of accepting as many events as possible given data collection rate limitations.

Table A.1: The definitions of the triggers I use in my analysis. The values in parentheses are applicable to the `Run2016H` dataset; they were changed to account for the increased L of the LHC near the end of the 2016 data-taking period. I analyze data events that pass at least one of the HLTs in the first group present in this table. I use the last HLT in the table to select orthogonal events for my trigger efficiency study.

Trigger name	**HLT definition**
`HLT_PFHT800(900)`	$H_\mathrm{T} > 800\,(900)$ GeV, where H_T is defined from AK4 jets with $p_\mathrm{T} > 40$ GeV and $\lvert\eta\rvert < 3.0$.
`HLT_AK8PFHT700(750)_TrimR0p1PT0p3Mass50`	$H_\mathrm{T} > 700\,(750)$ GeV, where H_T is defined from AK8 jets with $p_\mathrm{T} > 150$ GeV and $\lvert\eta\rvert < 2.5$ (the H_T definition I use in my analysis) and at least one AK8 jet with trimmed $m > 50$ GeV.
`HLT_PFJet450`	At least one AK4 jet with $p_\mathrm{T} > 450$ GeV.
`HLT_AK8PFJet360_TrimMass30`	At least one AK8 jet with $p_\mathrm{T} > 360$ GeV and trimmed $m > 30$ GeV.
`HLT_AK8DiPFJet300(280)_200_TrimMass30_BTagCSV_p087(20)`	At least one AK8 jet with $p_\mathrm{T} > 280\,(300)$ GeV, at least another with $p_\mathrm{T} > 200$ GeV, at least one AK8 jet with trimmed $m > 30$ GeV, and at least one b-tagged calojet (a jet formed from calorimeter deposit objects instead of PF objects).
`HLT_Mu50`	At least one muon with $p_\mathrm{T} > 50$ GeV and $\lvert\eta\rvert < 2.5$. (I used this HLT to measure the efficiencies of the HLTs above, not to select data events for my analysis.)

The HLT variables refer to special (online) objects calculated quickly and upstream of the PF (offline) reconstruction I used in my analysis. It's therefore important to check that

the offline H_{T} cut I used in my preselection is efficient with respect to these trigger cuts.

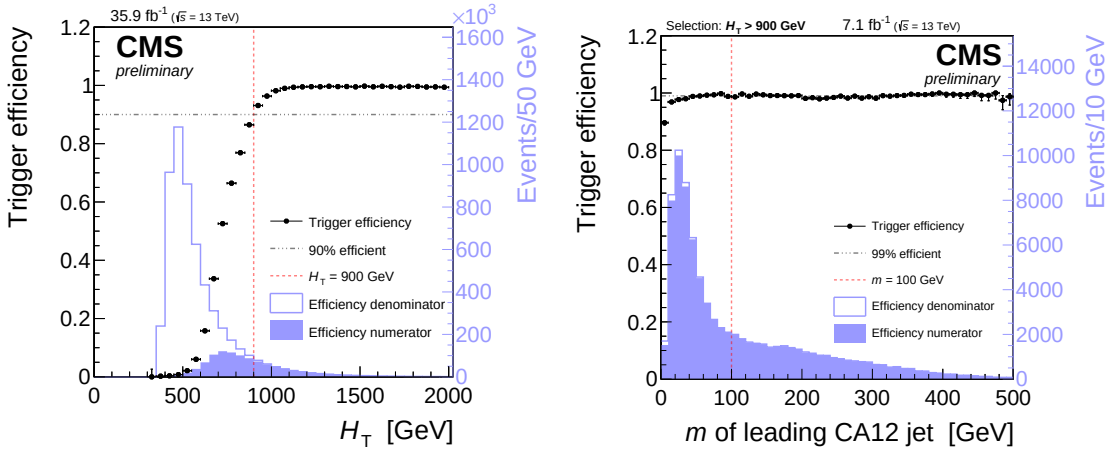


Figure A.1: The trigger efficiency

Appendix B

Data dataset comparisons

In the inclusive CR, I compare relevant analysis observables between the 2015 and 2016 JetHT datasets in Figures B.1–B.2. The only remarkable difference is a systematic shift in the τ_{21} distributions. The systematic uncertainties I assign will cover the difference; in any case, the overall disagreement is negligible, especially considering the much lower event yield in the 2015 datasets. Therefore, I don't have any problem combining the datasets from both years.

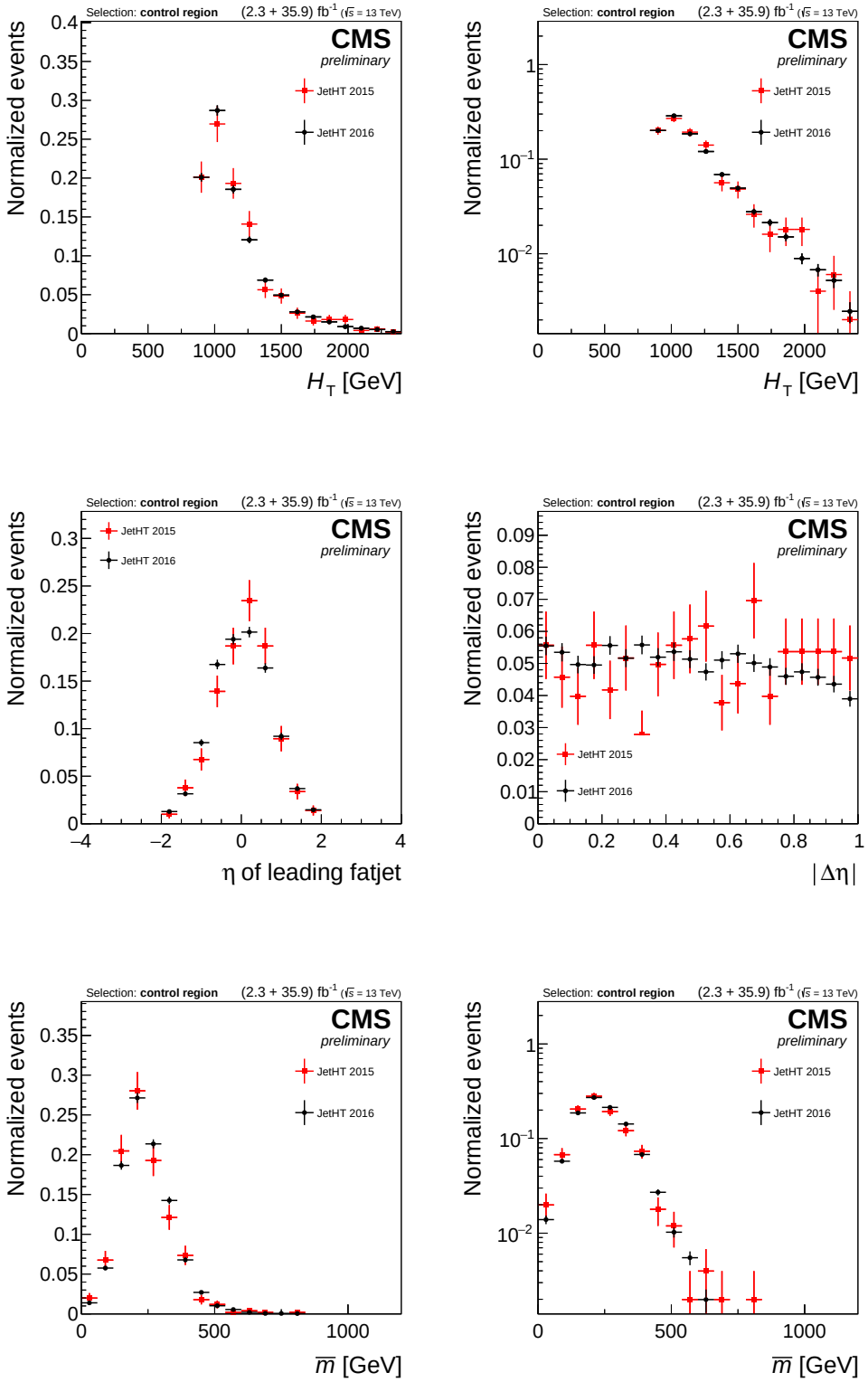


Figure B.1: Normalized distributions of kinematic variables for 2015 (red) and 2016 (black) JetHT datasets.

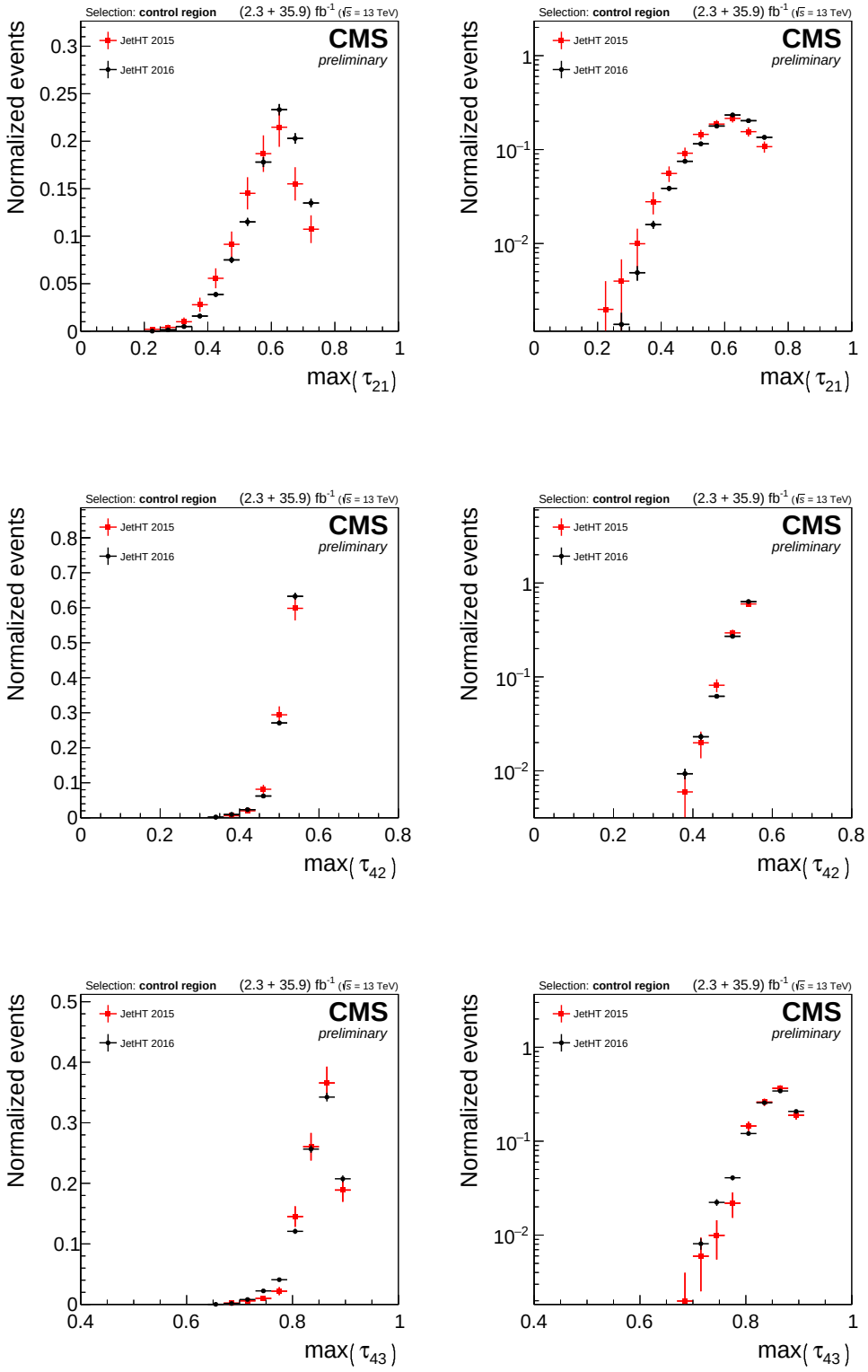
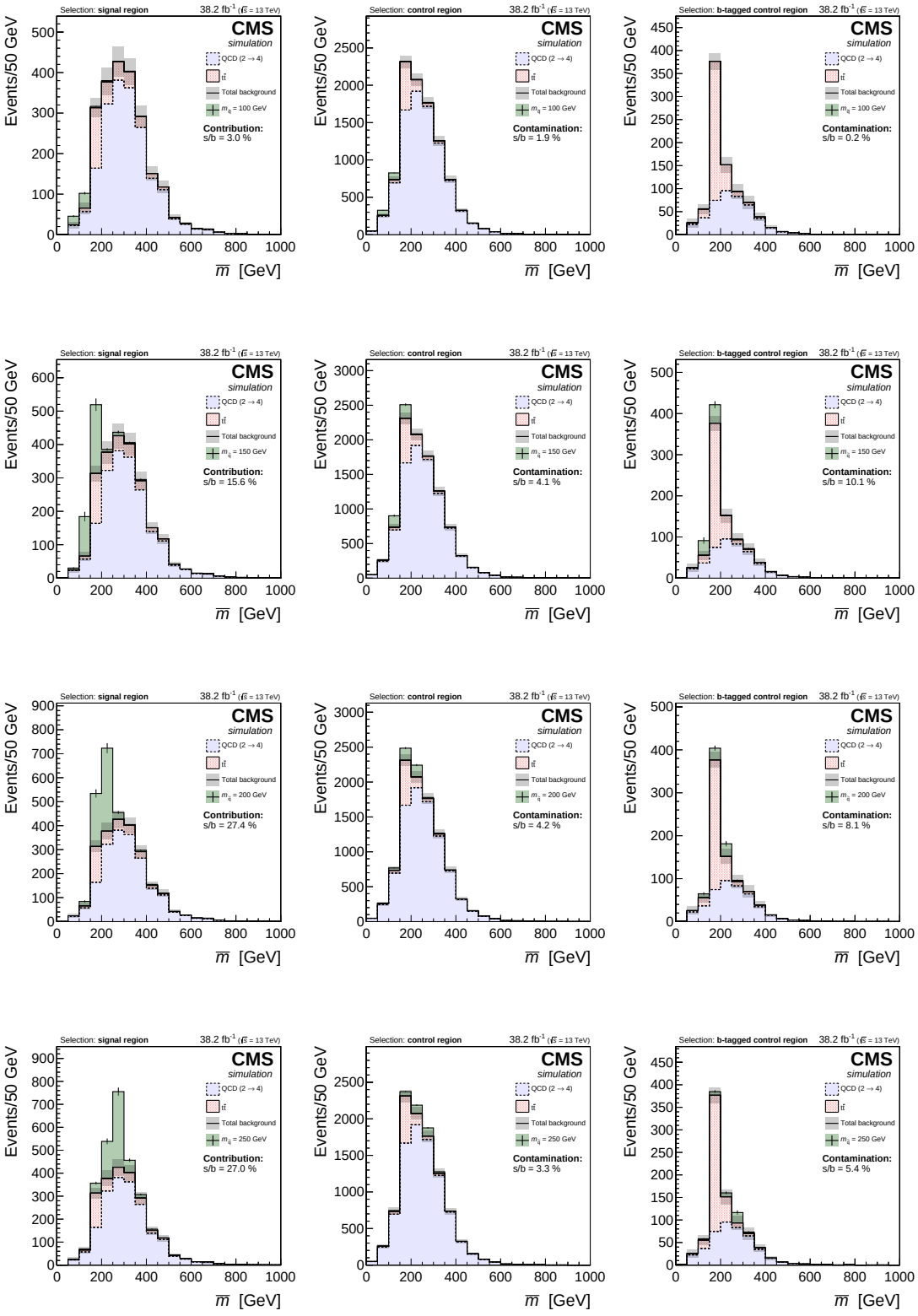


Figure B.2: Normalized distributions of N-subjettiness variables for 2015 (red) and 2016 (black) JetHT datasets.

Appendix C

Signal contamination

The squark signal contribution in SR and contamination in the CRs for seven squark masses are shown in Figure C.1.


Selection: signal region
Selection: control region
Selection: b-tagged control region
38.2 fb^{-1} ($\sqrt{s}$ = 13 TeV)
CMS
simulation
QCD (2 → 4)
$t\bar{t}$
Total background
$m_{\tilde{q}}$ = 100 GeV
$m_{\tilde{q}}$ = 150 GeV
$m_{\tilde{q}}$ = 200 GeV
$m_{\tilde{q}}$ = 250 GeV
Contribution:
Contamination:
s/b = 3.0 %
s/b = 1.9 %
s/b = 0.2 %
s/b = 15.6 %
s/b = 4.1 %
s/b = 10.1 %
s/b = 27.4 %
s/b = 4.2 %
s/b = 8.1 %
s/b = 27.0 %
s/b = 3.3 %
s/b = 5.4 %
Events/50 GeV
$\overline{m}$ [GeV]

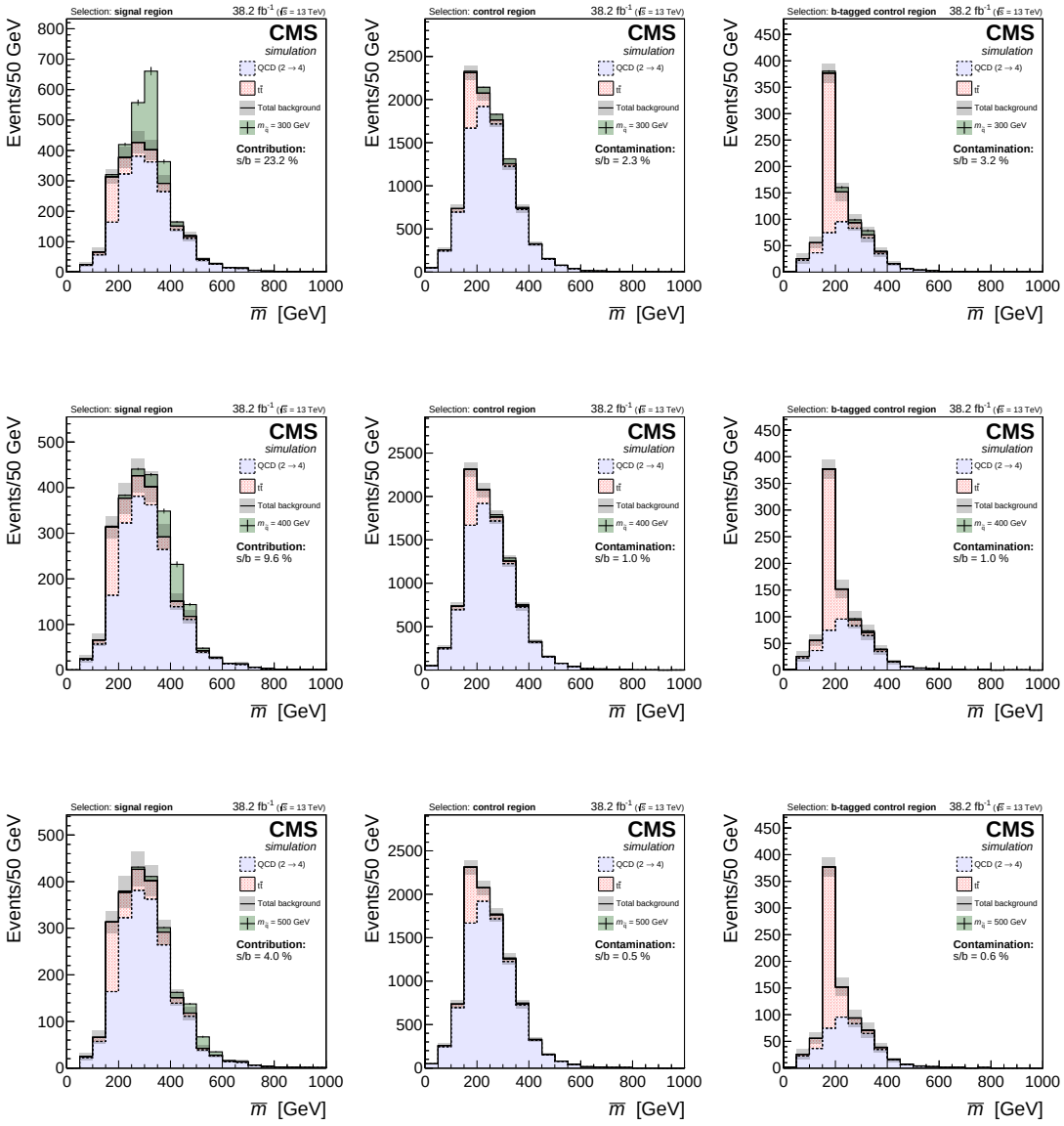


Figure C.1: Simulated distributions of $\overline{m}$ for QCD multijet, $\mathrm{t\bar{t}}$, and seven squark signal masses between 0.1 and 0.5 TeV (top to bottom) in the SR (left), inclusive CR (middle), and b-tagged CR (right). The ratio of signal events to background events in the respective analysis region is indicated on each plot.

Appendix D

Background estimation suppliment

I conduct a closure test in the SR of each simulated QCD multijet dataset. The poorer statistics in the $\mathrm{QCD}_{2\to2}$ dataset make this comparison less useful than that in the loose SR, but the closure test conclusions are the same: events from the $\mathrm{QCD}_{2\to4}$ dataset are estimated more accurately and need less adjustment than those from the $\mathrm{QCD}_{2\to2}$ dataset.

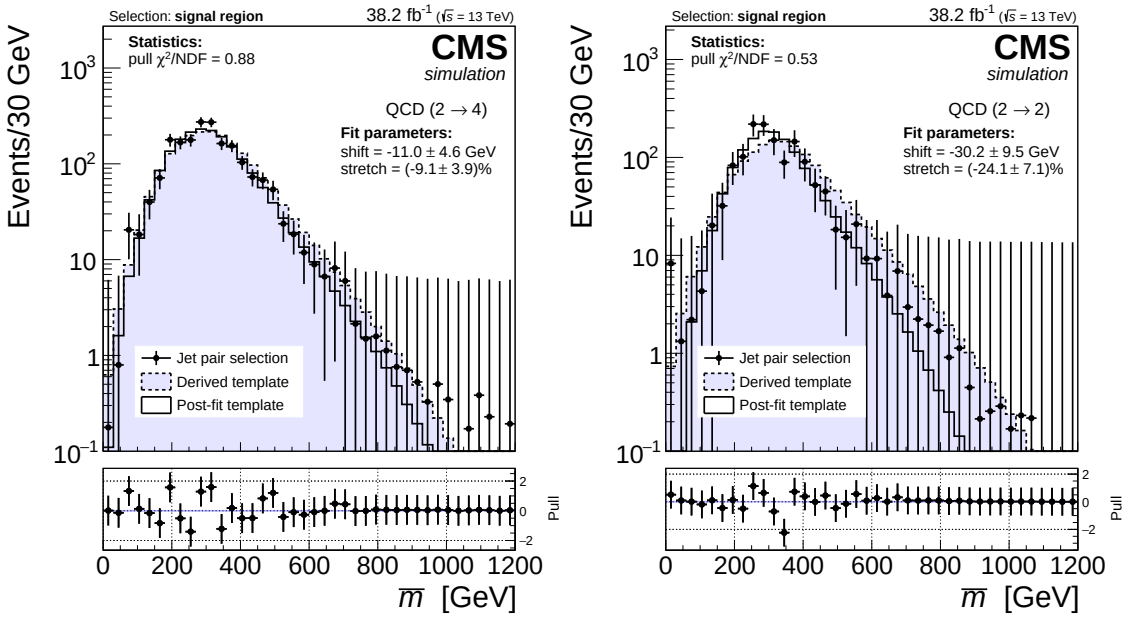


Figure D.1: Comparisons between the QCD multijet $\overline{m}$ PDF and the fatjet pair $\overline{m}$ distributions in simulated QCD multijet events from the $\mathrm{QCD}_{2\to4}$ (left) and $\mathrm{QCD}_{2\to2}$ (right) datasets in the SR. The statistics in the simulated datasets are poor, but the fit is still able to find agreement.

Bibliography

[1] W. K. Heisenberg, "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen.", *Z. Phys.* 33 (1925) 879, doi:10.1007/BF01328377.

[2] W. Pauli, "Über den zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren", *Z. Phys.* 31 (1925) 765, doi:10.1007/BF02980631.

[3] A. Einstein, "Zur Elektrodynamik bewegter Körper", *Ann. Phys. (Berl.)* 322 (1905) 891, doi:10.1002/andp.19053221004.

[4] P. A. M. Dirac, "The quantum theory of the electron", *Proc. Royal Soc. Lond. A* 117 (1928) 610, doi:10.1098/rspa.1928.0023.

[5] P. A. M. Dirac, "A theory of electrons and protons", *Proc. Royal Soc. Lond. A* 126 (1930) 360, doi:10.1098/rspa.1930.0013.

[6] P. A. M. Dirac, "The Principles of Quantum Mechanics". Oxford University Press, 1930.

[7] W. Pauli, "The connection between spin and statistics", *Phys. Rev.* 58 (1940) 716, doi:10.1103/PhysRev.58.716.

[8] J. Schwinger, "The theory of quantized fields. I", *Phys. Rev.* 82 (1951) 914, doi:10.1103/PhysRev.82.914.

[9] A. H. Compton, "A quantum theory of the scattering of X-rays by light elements", *Phys. Rev.* 21 (1923) 483, doi:10.1103/PhysRev.21.483.

[10] J. Aubert et al., "The ratio of the nucleon structure functions F_2^N for iron and deuterium", *Phys. Lett. B* 123 (1983) 275, doi:10.1016/0370-2693(83)90437-9.

[11] G. Arnison et al., "Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider", *Phys. Lett. B* 126 (1983) 398, doi:10.1016/0370-2693(83)90188-0.

[12] D. P. Barber et al., "Discovery of three-jet events and a test of quantum chromodynamics at PETRA", *Phys. Rev. Lett.* 43 (1979) 830, doi:10.1103/PhysRevLett.43.830.

[13] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B* 716 (2012) 30, doi:10.1016/j.physletb.2012.08.021.

[14] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* 716 (2012) 1, doi:10.1016/j.physletb.2012.08.020.

[15] J. J. Thompson, "Cathode rays", *Philos. Mag.* 44 (1897) 293, doi:10.1080/14786449708621070.

[16] C. D. Anderson, "The apparent existence of easily deflectable positives", *Science* 76 (1932) 238, doi:10.1126/science.76.1967.238.

[17] C. D. Anderson, "The positive electron", *Phys. Rev.* 43 (1933) 491, doi:10.1103/PhysRev.43.491.

[18] C. M. G. Lattes, H. Muirhead, G. P. S. Occhialini, and C. F. Powell, "Processes involving charged mesons", *Nature* 159 (1947) 694, doi:0.1038/159694a0.

[19] C. M. G. Lattes, G. P. S. Occhialini, and C. F. Powell, "Observations on the tracks of slow mesons in photographic emulsions", *Nature* 160 (1947) 453, doi:10.1038/160453a0.

[20] M. L. Perl et al., "Evidence for anomalous lepton production in e^+–e^- annihilation", *Phys. Rev. Lett.* 35 (1975) 1489, doi:10.1103/PhysRevLett.35.1489.

[21] F. Reines and C. L. Cowan, "Detection of the free neutrino", *Phys. Rev.* 92 (1953) 830, doi:10.1103/PhysRev.92.830.

[22] G. Danby et al., "Observation of high-energy neutrino reactions and the existence of two kinds of neutrinos", *Phys. Rev. Lett.* 9 (1962) 36, doi:10.1103/PhysRevLett.9.36.

[23] DONUT Collaboration, "Observation of tau neutrino interactions", *Phys. Lett. B* 504 (2001) 218, doi:10.1016/S0370-2693(01)00307-0.

[24] E. D. Bloom et al., "High-energy inelastic e–p scattering at 6° and 10°", *Phys. Rev. Lett.* 23 (1969) 930, doi:10.1103/PhysRevLett.23.930.

[25] M. Breidenbach et al., "Observed behavior of highly inelastic electron-proton scattering", *Phys. Rev. Lett.* 23 (1969) 935, doi:10.1103/PhysRevLett.23.935.

[26] J. J. Aubert et al., "Experimental observation of a heavy particle J", *Phys. Rev. Lett.* 33 (1974) 1404, doi:10.1103/PhysRevLett.33.1404.

[27] CDF Collaboration, "Observation of top quark production in $\overline{\mathrm{p}}\mathrm{p}$ collisions with the Collider Detector at Fermilab", *Phys. Rev. Lett.* 74 (1995) 2626, doi:10.1103/PhysRevLett.74.2626.

[28] D0 Collaboration, "Observation of the top quark", *Phys. Rev. Lett.* 74 (1995) 2632, doi:10.1103/PhysRevLett.74.2632.

[29] S. W. Herb et al., "Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton–nucleus collisions", *Phys. Rev. Lett.* 39 (1977) 252, doi:10.1103/PhysRevLett.39.252.

[30] I. Peruzzi et al., "Observation of a narrow charged state at 1876 MeV/c^2 decaying to an exotic combination of K$\pi\pi$", *Phys. Rev. Lett.* 37 (1976) 569, doi:10.1103/PhysRevLett.37.569.

[31] ALEPH Collaboration, "Determination of the number of light neutrino species", *Phys. Lett. B* 231 (1989) 519, doi:10.1016/0370-2693(89)90704-1.

[32] LHCb Collaboration, "Observation of $J/\psi p$ resonances consistent with pentaquark states in $\Lambda_b^0 \to J/\psi K^- p$ decays", *Phys. Rev. Lett.* 115 (2015) 072001, doi:10.1103/PhysRevLett.115.072001.

[33] F. R. Klinkhamer and N. S. Manton, "A saddle-point solution in the Weinberg–Salam theory", *Phys. Rev. D* 30 (1984) 2212, doi:10.1103/PhysRevD.30.2212.

[34] O. W. Greenberg, "Spin and Unitary-Spin Independence in a Paraquark Model of Baryons and Mesons", *Phys. Rev. Lett.* 13 (1964) 598, doi:10.1103/PhysRevLett.13.598.

[35] D. B. Lichtenberg, "Unitary Symmetry and Elementary Particles". Academic Press, 1970.

[36] Y. A. Gol'fand and E. P. Likhtman, "Extension of the algebra of Poincaré group generators and violation of P-invariance", *JETP Lett.* 13 (1971) 323, doi:10.1142/9789814542340_0001.

[37] J. Wess and B. Zumino, "Supergauge transformations in four dimensions", *Nucl. Phys. B* 70 (1974) 39, doi:10.1016/0550-3213(74)90355-1.

[38] R. Barbieri, S. Ferrara, and C. A. Savoy, "Gauge models with spontaneously broken local supersymmetry", *Phys. Lett. B* 119 (1982) 343, doi:10.1016/0370-2693(82)90685-2.

[39] H. P. Nilles, "Supersymmetry, supergravity and particle physics", *Phys. Rept.* 110 (1984) 1, doi:10.1016/0370-1573(84)90008-5.

[40] G. R. Farrar and P. Fayet, "Phenomenology of the production, decay, and detection of new hadronic states associated with supersymmetry", *Phys. Lett. B* 76 (1978) 575, doi:10.1016/0370-2693(78)90858-4.

[41] R. Barbieri and G. F. Giudice, "Upper bounds on supersymmetric particle masses", *Nuclear Physics B* 306 (1988) 63, doi:10.1016/0550-3213(88)90171-X.

[42] R. Kitano and Y. Nomura, "Supersymmetry, naturalness, and signatures at the LHC", *Phys. Rev. D* 73 (2006) 095004, doi:10.1103/PhysRevD.73.095004, arXiv:hep-ph/0602096.

[43] C. Brust, A. Katz, S. Lawrence, and R. Sundrum, "SUSY, the Third Generation and the LHC", *JHEP* 03 (2012) 103, doi:10.1007/JHEP03(2012)103.

[44] M. Papucci, J. T. Ruderman, and A. Weiler, "Natural SUSY endures", *JHEP* 09 (2012) 035, doi:10.1007/JHEP09(2012)035, arXiv:1110.6926.

[45] R. Barbier et al., "R-parity-violating supersymmetry", *Phys. Rep.* 420 (2005) 1, doi:10.1016/j.physrep.2005.08.006, arXiv:hep-ph/0406039.

[46] J. A. Evans and Y. Kats, "LHC coverage of RPV MSSM with light stops", *JHEP* 04 (2013) 028, doi:10.1007/JHEP04(2013)028, arXiv:1209.0764.

[47] L. Evans and P. Bryant, "LHC machine", *JINST* 3 (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[48] S. Myers, "The LEP collider, from design to approval and commissioning", 1991. doi:10.5170/CERN-1991-008.

[49] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[50] CMS Collaboration, "Technical proposal for the Phase-II upgrade of the Compact Muon Solenoid", CMS technical proposal CERN-LHCC-2015-010 and CMS-TDR-15-02, CERN, 2015.

[51] M. Lapka, "Interactive slice of the CMS detector", 2011. Accessed on September 14, 2017.

[52] R. Brun and F. Rademakers, "ROOT—An object oriented data analysis framework", *Nucl. Instr. Meth. Phys. Res. A* 389 (1997) 81, doi:10.1016/S0168-9002(97)00048-X.

[53] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* 12 (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:inst-det/1706.04965.

[54] G. P. Salam, "Towards jetography", *Eur. Phys. J. C* 67 (2010) 637, doi:10.1140/epjc/s10052-010-1314-6, arXiv:hep-ph/0906.1833.

[55] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* 12 (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[56] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Recombination algorithms and jet substructure: pruning as a tool for heavy particle searches", *Phys. Rev. D* 81 (2010) 094023, doi:10.1103/PhysRevD.81.094023, arXiv:0912.0033.

[57] J. Thaler and K. Van Tilburg, "Identifying boosted objects with N-subjettiness", *JHEP* 3 (2011) 15, doi:10.1007/JHEP03(2011)015.

[58] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* 13 (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[59] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* 7 (2014) 79, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[60] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", *Comput. Phys. Commun.* 191 (2015) 159, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[61] C. Borschensky et al., "Squark and gluino production cross sections in pp collisions at $\sqrt{s} = 13$, 14, 33 and 100 TeV", *Eur. Phys. J. D* 74 (2014) 3174, doi:10.1140/epjc/s10052-014-3174-y, arXiv:1407.5066.

[62] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method", *JHEP* 11 (2007) 70, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[63] ATLAS Collaboration, "Measurement of the W^+W^- production cross section in pp collisions at a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS experiment", *Phys. Lett. B* 773 (2017) 354, doi:10.1016/j.physletb.2017.08.047.

[64] CMS Collaboration, "Jet algorithms performance in 13 TeV data", CMS physics analysis summary CMS-PAS-JME-16-003, 2016.

[65] R. D. Cousins, J. T. Linneman, and J. Tucker, "Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process", *Nucl. Instr. Meth. Phys. Res. A* 595 (2008) 480, doi:10.1016/j.nima.2008.07.086, arXiv:physics/0702156.

[66] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of P", *J. Royal Stat. Soc.* 85 (1922) 87, doi:10.2307/2340521.

[67] D. Sivia and J. Skilling, "Data Analysis: A Bayesian Tutorial". Oxford University Press, 2nd edition, 2006. ISBN 9780191546709.

[68] Particle Data Group Collaboration, "Review of particle physics", *Chin. Phys. C* 40 (2016) 100001, doi:10.1088/1674-1137/40/10/100001.

[69] A. Caldwell, D. Kollár, and K. Kröninger, "BAT—The Bayesian analysis toolkit", *Comput. Phys. Commun.* 180 (2009) 2197, doi:10.1016/j.cpc.2009.06.026, arXiv:0808.2552.

[70] M. Metropolis et al., "Equation of state calculations by fast computing machines", *J. Chem. Phys.* 21 (1953) 1087, doi:10.1063/1.1699114.