THE 2D ALPHABET BACKGROUND MODELING METHOD AND ITS USE IN THE SEARCH FOR AN EXCITED BOTTOM QUARK

by

Lucas Corcodilos

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

October, 2021

Abstract

Given the recent prominence of jets in many LHC analyses, modeling multijet backgrounds produced via QCD processes has become a prominent issue. Most physics-based simulations of QCD processes produce so few events in the signal phase space that statistical uncertainties become dominant. Even when not considering the simulated event yields, the phase space considered is often not modeled well by the physics simulation of these events. Presented in this dissertation is a robust solution to model these backgrounds with a data-driven method that simultaneously fits for simulation-based models of other backgrounds as well as Beyond Standard Model signals. The search for a heavy resonance decaying to a top quark and a W boson in the fully hadronic final state is presented as a full example of an analysis using this novel background modeling method.

The analysis is performed using data from proton-proton collisions at a center-of-mass energy of 13 TeV, corresponding to an integrated luminosity of 137 fb^{-1} recorded by the CMS experiment at the LHC. The search is focused on heavy resonances, where the decay products of each top quark or W boson are expected to be reconstructed as a single, large-radius jet with a distinct substructure. The production of an excited bottom quark, b^{*}, is used as a

benchmark when setting limits on the cross section for a heavy resonance decaying to a top quark and a W boson. The hypotheses of b^* quarks with left-handed, right-handed, and vector-like chiralities are excluded at 95% confidence level for masses below 2.6, 2.8, and 3.1 TeV, respectively. These are the most stringent limits on the b^* quark mass to date, extending the previous best limits by almost a factor of two.

Thesis Committee

Primary Readers

Petar Maksimović (Primary Advisor)
Professor
Department of Physics and Astronomy
Johns Hopkins Krieger School of Arts and Sciences

Andrei Gritsan
Professor
Department of Physics and Astronomy
Johns Hopkins Krieger School of Arts and Sciences

Alternate Readers

Amitabh Basu
Associate Professor
Department of Applied Mathematics and Statistics
Johns Hopkins Whiting School of Engineering

Jacob Bernstein
Associate Professor
Department of Mathematics
Johns Hopkins Krieger School of Arts and Sciences

Robert Leheny
Professor
Department of Physics and Astronomy
Johns Hopkins Krieger School of Arts and Sciences

Acknowledgments

To my defense committee, thank you for taking the time to consider my research in sincerity.

To the office staff in the JHU Department of Physics and Astronomy (particularly Kelley Key), thank you for providing a welcoming environment that runs smoothly so students can focus on work instead of red tape.

To my parents, thank you for encouraging me in science when I was young, sending me to summer camps, motivating me to take more difficult classes, and forcing me to take SAT tutoring when I needed it but didn't want to take it. You inspired me to do more through self-motivation.

To my high school chemistry teacher, Lisa Yurgel, thank you for encouraging me to take AP Physics when I was planning on taking an easier class. Without your faith in me, I wouldn't have ever found physics.

To Yuri Gershtein, thank you for the opportunity to work on CMS research at Rutgers and for your professional guidance which lead me to Johns Hopkins.

To Kevin Nash, thank you for your patience in teaching me how to do CMS research. Having now had to do the same for younger students, I understand how much time and effort had to be spent to get me started in real research.

To Matt Goldman, Tolby Lew, Nick Chin, Leon Thanikal, and Joey Huang, thank you for surprising me by showing up to the defense of this thesis. Through our time at Rutgers and after (including with Omar Bakir and Faizan Munshi), you all have been great friends and our time spent online during the COVID-19 pandemic helped me keep my sanity.

To my friends from the JHU Physics and Astronomy department, thank you for the great community you've provided, where support is favored over competition. The graduate student population in the department is a great example of how academic environments should operate.

Specifically to Josh Kable, Oz Amram, and Cris Mantilla Suarez, thank you for your friendship over the past several years. You have helped me grow academically, mentally, and emotionally while you answered my physics questions, listened to me complain, and provided your unique perspectives. Particularly, Josh's regular board game nights inspired the formulation of the first chapter of this thesis and Oz regularly helped fill in gaps in my knowledge over the last few months as I prepared this dissertation.

To Samantha Kose, thank you for all of the support you've provided over the past $11\frac{1}{2}$ years as my partner. As you put it, you are my "hype team", supporting me when I'm right and telling me when I'm wrong. Your constant support makes it easy to forget but any time I think about specific milestones, you're always there. Over the past two years, you've spent an incredible amount of emotional energy to help me with my own worries - especially when you had to take the role of my friends in physics who I didn't see regularly because of the pandemic. Above all else, thank you for being my

best friend.

To my adviser, Petar Maksimovic, thank you for saying, "This is supposed to be fun." While not every moment in physics or life can be "fun", being your student has been a fun experience. You've taught me an incredible amount: enough about statistics to be a Statistics Contact in CMS and to have a dissertation about statistical modeling; enough about people to be able to navigate interactions in a large organization; enough about patience to get a result through CMS approval; enough about impatience to realize when "better is the enemy of the good". These are all skills and lessons I'll use for the rest of my life.

Finally, to my cat, Chuck, thank you for having a higher emotional capacity than many humans. Your time in my lap around 3:30pm every day I worked from home was always welcome.

Preface

One of the more frustrating parts of studying physics (at any level) is the reaction others have - "Oh, I always hated physics class," "That's above my pay grade," "I wouldn't understand any of that." For some reason, physics is seen as esoteric when in reality, its most basic principles influence every part of our lives. From electricity that powers our homes to the combination of red, blue, and green light that creates images on our screens, from the GPS in every phone that can locate our position on Earth to the fluid dynamics that lifts airplanes, from the greenhouse effects warming the planet to the gravity that keeps us all from floating away, fundamental physical laws are everywhere. Who wouldn't want to to understand them?

Of course, let's not sugar-coat studying the subject. Even the basics can become un-intuitive quickly and the math unwieldy. So I don't advocate that everyone should study physics and, in fact, I think a world of only physicists would be a terrible place. However, forcing someone to struggle with the math or to take tests on the material are what I believe normally kills interest in the subject among most people. The intricacies should only be there when someone has the taste for more but not as a barrier to understanding *any* of the magic of the subject.

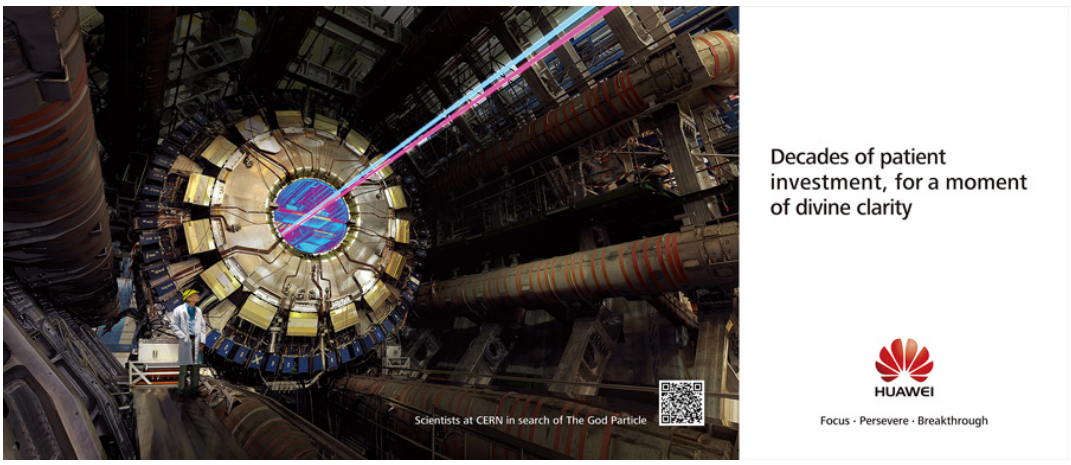

Figure 1: This advertisement was featured in an airplane magazine during a flight I took to attend a workshop. I laughed so hard that I tore out a copy to keep for myself. My favorite part is that the beams can only exist in a vacuum so the scientist should really be shown as a pink mist during their "moment of divine clarity."

On the flip side, science media often does the opposite. Core concepts of modern research have been so simplified that we get images like the one in Fig. 1 - who knew that protons could be visibly red-shifted or blue-shifted[1]?

With this belief, I'd like to make my own contributions to the field more accessible by structuring this dissertation in a way that is welcoming to a non-expert audience. After all, those who already know the intricacies of high energy particle physics don't need to read my explanation, those who *want* that explanation would be better served by a formal text (with which I do not plan to compete), and those who remain are exactly those readers that I don't want to leave behind.

After this document is submitted to the university library, no one may ever read it again. But if you *are* reading this from the archives of the university

[1] At least, this is what I guess the colors mean. I was originally tricked into thinking these were supposed to be beams of different particles which is exactly why this representation is so annoying.

and fall in that final group, I hope you find it an accessible explanation of one flavor of modern physics research. If you can put this down and a week later still remember a part of what you read, I'll consider this dissertation a success.

Table of Contents

List of Tables

List of Figures

Chapter 1

The Standard Model: The Game

The Standard Model (SM) of particle physics is, at its simplest, just like the setup to any board or card game. The "model" is just a big set of rules that explain the pieces in the game and how they interact with one another. It even comes with dice-like pieces such as the CKM matrix which determine at what probabilities certain interactions occur.

The universe is constantly "playing the game" and we humans can "play" as well by creating particle interactions that are otherwise rarely created by the universe. Of course, there is no way to "win". Instead, we keep playing the game (or watching the universe play) and check to see the rules as we understand them are being followed. Because, unlike your standard board game, this one does not come with the rule book.

We have instead tried to discern the rules ourselves, partly out of fascination and partly because the better you know the rules, the better you know how to use them to your advantage. At the level of elementary particles, we are still learning the rules and we know this because there are certain moves the universe has made or pieces that it has played that do not follow our

current version of the rules. The simplest example is gravity - at the moment, there is nothing in our rule book to explain what causes two objects with mass to be attracted to each other!

As another example, there are a class of "pieces" that we call "dark matter" that we see the universe using but that are also not in our rule book. We know the universe is using these pieces because we can look at distance galaxies and other astronomical objects and see them behaving as if they are more massive than we estimate based on the light that they emit and the way that they move. From these observations, we know there must be another type of matter collecting around these objects that is massive ("massive" here meaning "has mass") and is attracted by gravity to other massive objects. The key difference being that this mysterious matter does not emit *or absorb* light (i.e. photons) and it is thus "dark". The interactions of dark matter with SM particles are also extremely feeble.

There is additionally "dark energy", which we know even less about. We observe that the universe is expanding at an accelerating rate and, from the basic principle of energy conservation (a game rule), there must be an energy source to cause this acceleration and we dub that source "dark energy". The only other thing we know about dark energy is that it makes up about 68% of the universe while dark matter accounts for 27%. This means there is only 5% of the universe that we actually understand - and even that 5% can give us headaches.

There are a few other places where we know we have the rules wrong but for now, think of these as unexplained loopholes in the rules as we understand

them. Several players ("theorists") come up with some proposed expansions to the rules - these are called "Beyond Standard Model" (BSM) theories - and the play testers ("experimentalists") play the game to test them, reporting back if the new rules seem to do a better job explaining our observations than the current set of rules. If not, we move on to the next attempt.

While we have extensively verified the current set of rules, no new fundamental additions have been established since the discovery of the Higgs boson in 2012. Since then, various hypotheses have been tested and rejected, slowly reducing the possible set of new rules we can add. The unfortunate reality may be that the fundamental laws governing the roles of pieces like dark matter are beyond our current experimental capabilities. Maintaining the board game analogy, the universe may be playing the game at a level we simply are not ready to play at yet - or at a level we may never be capable of playing. Of course, that is a pessimistic view - the proverbial "rage quit" - and one that is impossible to prove without testing every possibility.

So while accepting that reality as a possibility, we consider the current version of the rule book and present just the pieces necessary to understand the original work presented in this dissertation. The remaining sections of this chapter will cover just those rules that are necessary before Ch. 2, which describes some of the techniques experimentalists use to "see" these particle interactions. Chapters 3 and 4 will then present the author's original work which includes a novel background modeling technique and an example analysis testing a possible modification of the SM, respectively.

1.1 Feynman Diagrams

A critical component to our "game" that is useful to introduce first is the Feynman diagram; a clever visual representation of particle interactions which also serves as a mathematical framework (though I leave that part to a more robust resource). A generic version is presented in Fig. 1.1 which can be used as a basis for specific particle interactions later.

First, in all of the Feynman diagrams we will consider, all vertices are defined by exactly three connecting lines called "propagators" which are our model's particles. Additionally, we adopt the convention that time flows from left to right and particles on the left are "incoming" and on the right are "outgoing". This means that in Fig 1.1, the interaction is an incoming particle decaying into two other particles. If we rotate the diagram 180 degrees, the interaction is instead two particles annihilating to create a single new particle. Critical to this example is that the possible propagators that can form a vertex do not depend on which direction time flows.

The remaining rules to the game, detailed in the next sections, then dictate which particles can interact with each other and how strongly the particles interact ("couple") with each other, which determines the probabilities to see one interaction over another. The Feynman Diagram is a way of representing these rules diagrammatically, as will be shown. Everything beyond a single vertex is just chaining together these fundamental interactions.

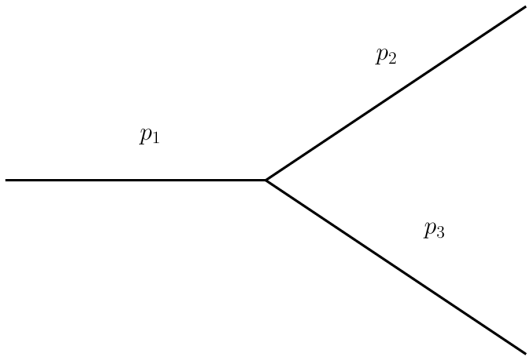

Figure 1.1: A very basic Feynman diagram showing the three point vertex formed by three unspecified propagators, p_1, p_2, and p_3.

1.2 The (relevant) game pieces

Fig. 1.2 shows a depiction of the Standard Model as one table of elements, complete with labeled categories, shaded regions, and associated particle quantities. These are the pieces to our game.

One brief note to make is on the units of mass in this diagram - eV/c^2. An eV is an "electron-volt" and is a convenient unit of energy when discussing energies on the particle scale. Recalling Einstein's famous $E = mc^2$ which relates energy and mass, one can quickly get $m = E/c^2$. Thus mass has units of eV per c^2, as Fig. 1.2 shows. However, we often define $c = 1$ and simply write the mass as eV, which will be used later on.

1.2.1 Bosons

Digging further into Fig. 1.2, the first subdivision of particles is into fermions (left three columns) and bosons (right two columns). While fermions and bosons can be differentiated by their spin (fermions having odd half-integer spin and bosons having integer spin), bosons also differentiate themselves

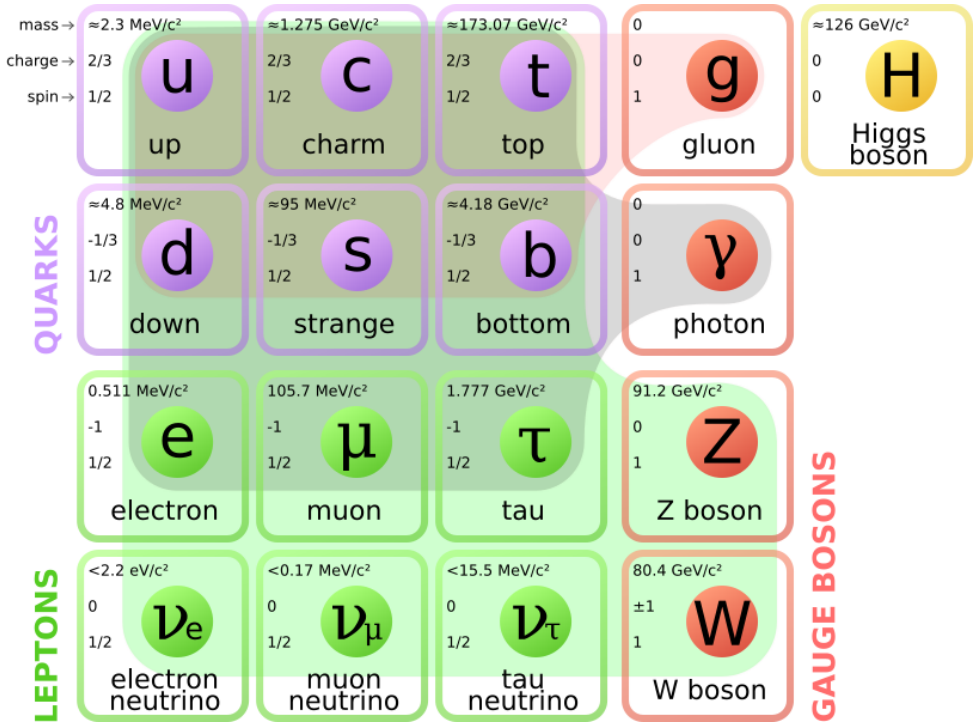

Figure 1.2: Table of the particles that make up the Standard Model of particle physics. Figure adapted from Ref. [1].

in that they are the so-called "force carriers" in the model. In other words, interactions can only happen when a force is involved which means at least one boson must be involved in each interaction to "carry" the force.

Consider two particles sitting at rest, one positively charged and the other negatively charged. As the saying goes, "opposites attract", so if the particles are released, they will start traveling towards each other. But how do the particles "know" to do that? How are they able to "communicate" so that the positive particle knows the negative particle is nearby?

The answer is the electromagnetic field in which they each exist. Imagine two bowling balls (our charged particles) on a stretchy spandex surface that is pulled taught (the field). With no balls on the spandex, the surface does

not stretch and remains "null" (also called the "vacuum state"). If the balls are added but sufficiently separated, they stay in place. If they are brought close enough, the warping of the spandex by one ball will affect the spandex near the other and they will roll towards each other. In a similar way, the charged particles exist in an electromagnetic field and because they have non-zero charge, they warp the field, "feel" each other's effect on the field, and move correspondingly.

This may seem confusing since there are no "fields" in Fig. 1.2, only particles. However, these are not particles like the lint on a shirt. Instead, their quantum nature gives them wave–particle duality which allows them to be described as both a particle and a wave. In this formulation, they are excitations, or "quanta", of their associated field (or fields) and there are no "particles" in the traditional sense, only fields and their excitations.

Then, the answer for how the particles "communicate" is via a force carrier - an excitation of the field which can propagate information and, in doing so, changes the value of the field from its vacuum state. For the electromagnetic force, this is the photon; for the strong force, the gluon; for the weak force, the W and Z bosons; for the Higgs force, the Higgs boson. These are real, detectable particles which themselves can have properties that allow for interactions between each other and even self interactions. However, they can also be "virtual" (sometimes also called "off shell" or "off mass shell") in that they exist for physically unallowable amounts of time (and potentially at physically unallowable masses or energies) but which is possible due to the

Heisenberg uncertainty principle which is defined mathematically as

$$\Delta x \Delta p \geq \frac{\hbar}{2} \quad \text{or} \quad \Delta E \Delta t \geq \frac{\hbar}{2} \tag{1.1}$$

where Δ represents "uncertainty in", x is position, p is momentum, E is energy, t is time, and $\bar{h}$ is Planck's constant divided by 2π (approximately $6.6 \times 10^{-16}\,\text{eV} \cdot s$). This means that a particle's momentum at a given position (or energy at a given time) cannot be known with infinite precision and to be more certain of the position (energy) is to be less certain of the momentum (time) and vice versa. This means that, for example, if a particle must exist at a very specific energy, the time at which it will exist at that energy is largely uncertain.

Every force also has an associated "charge" and if the charge for a given particle is non-zero, then it can "couple" (i.e. interact) with the force. Among the bosons, only the Higgs and gluon carry the same charge that they mediate - the Higgs couples to mass and the gluon to "color charge". The self-interactions of the gluon will have consequences in Sec. 1.2.4 and when discussing jets in Sec. 2.3.2. To study the forces of the remaining bosons, we should first introduce the fermions.

1.2.2 Fermions: Leptons

Fermions can be thought of as the particles that bind together to create the atoms and molecules we find in our everyday lives - protons, neutrons, and electrons are all made of (or are themselves) fermions. Fermions are also split into two types of particles called quarks and leptons.

Leptons are spin-$\frac{1}{2}$ particles with integer electric charge that can exist in space by themselves. Among the leptons, you will hopefully recognize at least one familiar face in the electron! However, they also come in two flavors - charged and neutral - where each charged lepton has a corresponding neutral partner called a "neutrino", which is (nearly) massless. Because neutrinos have no electric or color charge and are effectively massless, they are very hard to detect, as will be mentioned again in Sec 2.2.

The heavier partners of the electron are the muon and the tau, with the tau being heavier. Because of their masses and their ability to decay to lighter leptons, the muon and tau are unstable particles, making it energetically favorable for them to decay to electrons; a process which is only possible because leptons can couple to the weak force. Thus, the muon and tau are not nearly as common as the electron in everyday life. However, muons are regularly seen in the form of cosmic rays, high energy particles that come to Earth from astronomical objects such as the Sun. This is possible because muons have a relatively long life-time which means they can travel the distance between the Sun and Earth before decaying - provided they are traveling fast enough to experience the relativistic effect of time dilation. This attribute of muons will be relevant again in Sec 2.2 since muons pass largely unscathed through the CMS detector (part of the reason the muon is the namesake of CMS).

Finally, leptons are able to interact with the electromagnetic force, the weak force, and the Higgs force. The Higgs interaction strength is orders of magnitude smaller, creating a small coupling. The electromagnetic force is

the familiar force that, when moving charged leptons, produces what we call "electricity". It is also responsible for other common phenomenon such as the photoelectric effect and ionization. The weak force is the interaction that allows the decay of the heavier charged leptons to the electron and is also why the muon has such a long lifetime, an example which will be explained further in Sec. 1.2.7 once a few more game pieces have been introduced.

1.2.3 Fermions: Quarks

Quarks come in up-types and down-types (upper and lower purple rows in Fig. 1.2) with the primary difference being that up-type quarks have electric charge of -1/3 and down-type have electric charge of 2/3. The three generations of quarks (three purple columns in Fig. 1.2) differ primarily in mass, especially for the top quark which is the heaviest particle in the SM.

Unique to quarks is that they are the only other particles besides the gluon that are color charged and thus have the ability to interact with the strong force. Describing the charge with "color" is just a bookkeeping measure to describe a charge with three possible values that, when summed together, cancel each other. In this case, one can combine equal amounts red (r), green (g), and blue (b) and get back "white"/zero. Additionally, just as an electric charge can be positive or negative, the color charge can also be anti-red ($\bar{r}$), anti-green ($\bar{g}$), or anti-blue ($\bar{b}$), where color and anti-color also cancel each other. Thus, each quark is charged with either r, g, b, $\bar{r}$, $\bar{g}$, or $\bar{b}$.

Another attribute of the strong force is that it increases in strength as the

distance between particles increases - this is in contrast to the previous electromagnetic force example of "bowling balls on spandex" where the bowling balls will not affect each other if there is sufficient distance between them. Instead, pulling two bound quarks apart increases the energy of the bond at an increasing strength until it is more energetically favorable to form two new quarks, each of which form two new "color tubes" with the originally bound quarks. As a result, only "color neutral" bound states of quarks ("hadrons") have ever been observed, a phenomena called "confinement". Conversely, as the distance between quarks gets smaller (and the energy scale increases), the strong force actually becomes *weaker*, a property of the theory called "asymptotic freedom" - as in, particles charged under the strong force start to behave as if they are "free" once the interaction distances asymptotically approach zero.

Quarks are also able to interact with the electromagnetic and weak forces. The electromagnetic interaction of quarks is primarily relevant since the CMS detector can detect electrically charged particles. The weak interaction creates more significant consequences for quarks. In particular, exchange of a W boson has the ability to change the flavor of a quark from up-type to down-type (or vice versa). Because the W boson comes in two types - electrically charged either +1 or -1, electric charge is conserved. However, the interaction is possible for far deeper reasons which are perhaps best presented with historical context in Sec 1.2.5.

1.2.4 Quarks and the strong force

There are a few consequences that result from the strong force rules presented so far. The first is that quarks can exist in either color-anti-color pairs (called "mesons") or in triplets of rgb or $\bar{r}\bar{g}\bar{b}$ (called "baryons").[1] As examples of bound states, the proton and neutron are both baryons. The proton is made up of two up quarks and a down quark (uud) and the neutron is made of one up quark and two down quarks (udd). One can add up the individual electric charges of the quarks to see these mixtures gives +1 and 0, respectively.

The second consequence is that hadrons are not made of *just* the two or three quarks. Consider the mass difference between the uud quarks, 9.4 MeV total, and the proton they combine to create, 938 MeV- this is not a negligible difference and there must be something else going on! Recall that the binding of quarks is a result of the strong force which, as described in Sec. 1.2.1, means the quarks are constantly exchanging virtual gluons. As also stated, gluons are charged under the strong force so they can interact with themselves as well as quarks. As a result, the two or three quarks defining the hadron are actually "valence quarks" which are held together by a "sea" of quarks and gluons popping in and out of existence at unmeasurabley small scales of time and distance - and thus behaving "freely". The sea quarks and gluons thus make up the vast majority of the proton's mass! This will have consequences when discussing proton-proton collisions in Ch. 2 since, at high collision energies that probe small distances, the collisions are really between this soup of gluons

[1] Bound states made of mixtures of multiple pairs or triplets are not disallowed and have actually been observed but they are very rare and will be ignored for the purposes of this dissertation.

and quarks.

1.2.5 Quarks and the weak force

The existence of the weak force was first proposed by Enrico Fermi in 1933 as a way to describe beta decay, the process in which a neutron could decay into a proton, raising the atomic number by one (turning the atom into a new element) and emitting just the "beta" particle (an electron) - or so it seemed. The mass difference of the neutron and proton should leave a specific amount of energy to allocate to the mass of the electron and its kinetic energy. In other words, the electron should always be emitted at roughly the same momentum. Experimentalists instead observed a distribution of electron energies, implying a violation of the conservation of energy. They also knew of no fundamental mechanism which would allow a neutron to convert into a proton, emitting an electron in the process. Thus, Fermi's theory predicted a four-way interaction between the neutron, proton, electron, and the newly predicted neutrino.

The proposed "weak interaction" caused the decay and the neutrino served as the mechanism for carrying away an undetected amount of energy since it was assigned no electric charge and only participated in interactions involving this new force. As it turned out, this was the low-energy effective field theory for the full weak interaction theory. That is, Fermi got the rules correct when dealing with low energies. This is an amazing feat when one considers that the quark model, which is crucial to our current understanding, was proposed 30 years later! As a result, the full theory took nearly a century to develop - which is why we leave the historical re-telling here. What remains between 1933 and

the discovery of the Higgs in 2012 is lengthy, involves mathematical intricacies beyond the non-expert level targeted in the Preface, and will do no better to serve a non-expert. Instead, we can work from Fermi's initial observations and proposed theory to justify the modifications that take Fermi's theory to the current set of SM rules, albeit with one otherwise unexplained addition, the Cabibbo–Kobayashi–Maskawa (CKM) matrix.

Recalling from Sec. 1.1 the "rule" that vertices are defined with three intersecting propagators, the first modification to Fermi's theory is to have two different three-way interactions, connected by a shared particle, instead of one four way interaction. Additionally, we now know the interaction is not between the proton and neutron but between the constituent quarks. Therefore, the modified model looks like the diagram in Fig. 1.3 where the first interaction is between an up-type quark, down-type quark, and a W boson and the second between a charged lepton, a lepton neutrino, and a W boson, with the W being the connecting piece. This means that not only can the weak force change charged leptons to lepton neutrinos (and vice-versa) and up-type quarks to down-type quarks (and vice-versa) but it is also a mechanism to cross between fermion types. This is how beta decay is possible - a d quark in the neutron changes to a u (causing the neutron to change to the proton) via emission of a W boson which then decays into an electron and an electron neutrino.

Another artifact of the full theory that we have largely ignored so far is that there are actually three bosons associated with the weak interaction - W^+, W^-, and Z. The oppositely charged W^+ and W^- keep quark decays from being

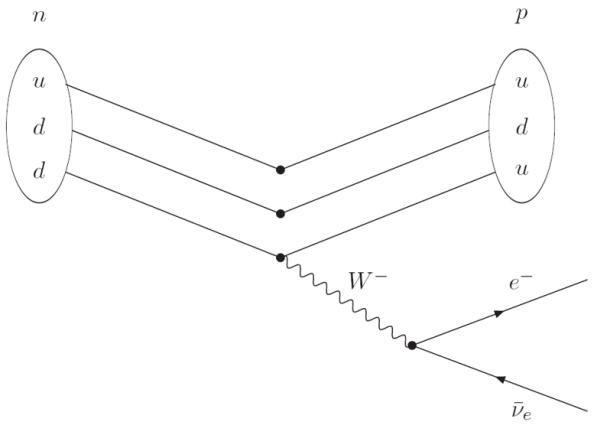

Figure 1.3: A Feyman diagram describing beta decay. Provided by Ref. [2]

one-directional. That is, they allow for both up-to-down and down-to-up interactions. They also allow for both electron and anti-electron (positron) interactions. However, there also exists the Z boson which has no electric charge. This allows for neutral weak currents where the two other particles involved in the three-way interaction will have electric charges that sum to zero. For example, an electron and positron or quark and anti-quark can annihilate to produce a Z boson (Fig. 1.4) or, in reverse, a Z boson can decay into oppositely charged leptons or a quark, anti-quark pair (Fig. 1.5). The Z can also interact with two neutrinos which is common in inelastic scattering of neutrinos against electrons bound in atoms - the neutrino "bounces off" the electron via exchange of a virtual Z boson (Fig. 1.6) just as two electrons "bounce" off each other by exchanging a virtual photon!

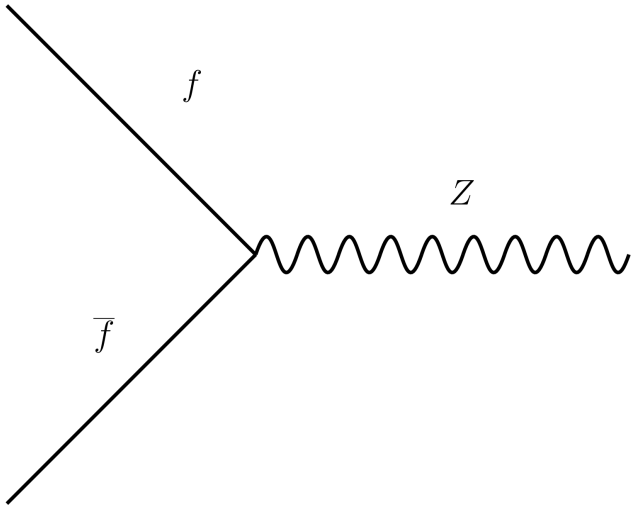

Figure 1.4: A Feyman diagram describing two fermions annihilating into a Z boson.

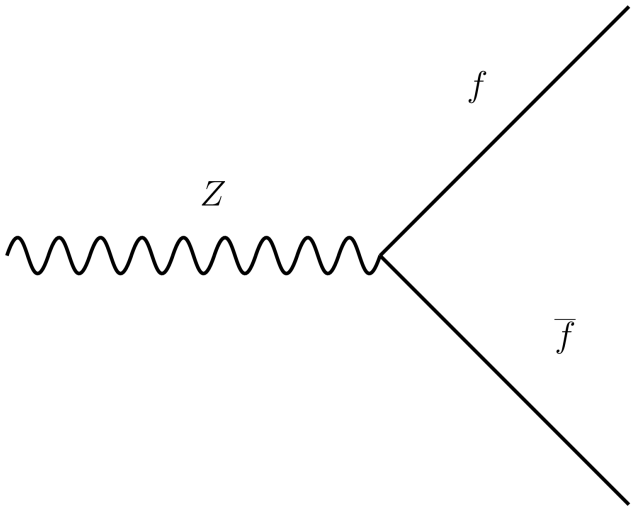

Figure 1.5: A Feyman diagram describing a Z boson decaying to two fermions.

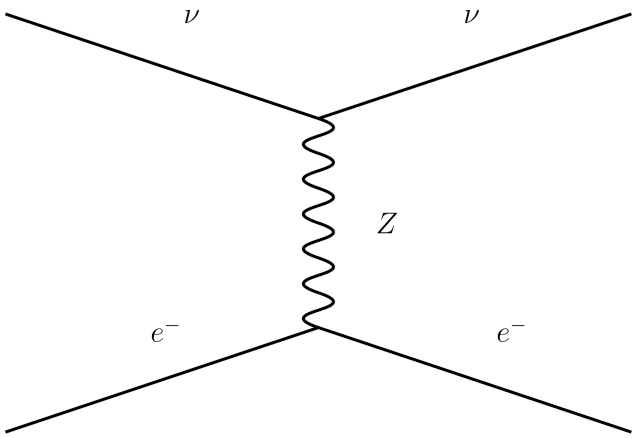

Figure 1.6: A Feyman diagram describing inelastic scattering of a neutrino against an atomic electron.

1.2.6 Quark mixing: or How to Make Cake

The final relevant detail in the rules of the weak force is the mechanism that allows for quarks to mix such that the d can turn into the u in the first place. This "mixing" of quark types is a result of two fundamental concepts from mathematics which themselves do not require math to explain: a "basis" and "superposition".

A "basis" is just a chosen representation of the "pieces" of a composite object. For example, an ice cream cake can be represented by some combination of cake, icing, and ice cream. Those three ingredients form a basis. However, I could also form a basis for ice cream cake with milk, sugar, and flour. That is, some mixture, or "superposition", of these ingredients can help me describe the cake in a different representation[2]. The different representations are useful for different scenarios. If assembling the cake, I want the cake, icing, and ice

[2] As you probably know, there are more ingredients than this but suspend your disbelief for the sake of simplicity

cream representation. If I need to make the cake from scratch, I will need to buy milk, sugar, and flour and pair that with a recipe that describes how to mix (superimpose) the ingredients so I can get to the basis of cake, icing, and ice cream. In other words, even if the primary components for assembly are cake, icing, and ice cream, the fundamental components are milk, sugar, and flour.

Similarly, the quarks and W and Z bosons in Fig. 1.2 are the "components for assembly" but they are not the fundamental representations of the components (milk, sugar, flour) of the weak force (the cake). The SM quarks are specifically a superposition of quarks in a different representation, denoted q'.

The cake analogy breaks when one considers that, while there is way to represent cake, icing, and ice cream with a mixture of milk, sugar, and flour, there is no way to do the reverse and represent milk, sugar, and flour as a mixture of cake, icing, and ice cream. This *is* possible with the quarks. Additionally, the superposition of quarks and their wave nature mean that a quark of one basis measured in another basis cannot be all three "ingredients" simultaneously - it becomes just one. For example, a d measured in the lab (the basis being the SM table in Fig. 1.2) is a superposition of u', c', and t'. However, when it interacts with the weak force, it must be represented as just *one* of u', c', and t'. The so-called "wave function" of q collapses to one q', which is chosen by quantum mechanical probabilities built into the wave function. And, the same process must happen to go *back* to the representation in the lab. That is, we do not observe d decay into q' and W; we observe d decay into q and W, where d collapsed to q' in the weak interaction's basis

$$\begin{bmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{bmatrix} = \begin{bmatrix} 0.97370 \pm 0.00014 & 0.2245 \pm 0.0008 & 0.00382 \pm 0.00024 \\ 0.221 \pm 0.004 & 0.987 \pm 0.011 & 0.0410 \pm 0.0014 \\ 0.0080 \pm 0.0003 & 0.0388 \pm 0.0011 & 1.013 \pm 0.030 \end{bmatrix}$$

Figure 1.7: The CKM matrix. Each element in the matrix represents the coupling of one up-type, down-type pair of quarks, with all combinations considered. Note that the diagonal elements are all close to 1, indicating that mixing within a generation is the strongest.

and the q' wave function, itself a superposition of u, c, and t, collapses into one of u, c, and t that we then observe.

These mixtures are summarized by the CKM matrix, shown in Fig 1.7, which condenses this back-and-forth into single values that represent the relative strengths of the couplings (or mixing) of each up-down quark pair. Fig. 1.8 additionally shows possible decays, including their relative likeliness, in a rotated version of the quark table in Fig. 1.2.

1.2.7 Example scenarios

As an exercise of how these concepts plug into the rest of the rules, we consider a few examples. The first is why neutrons and protons are the most abundant baryons in our world. There is nothing disallowing the second and third generation of quarks - c, s, b, and t - to form hadrons with other quarks. However, the weak force gives a mechanism for these quarks to decay causing them to be unstable. It is perfectly true that other hadrons exist and even have long enough lifetimes to be observed. For example, the K^+, made up of $u\bar{s}$, has a lifetime of order 10^{-8} which means it can travel 10 feet if traveling at the speed of light! But this is short given the scale of "universe time", making valence quark decay inevitable and eventually leaving the lightest possible

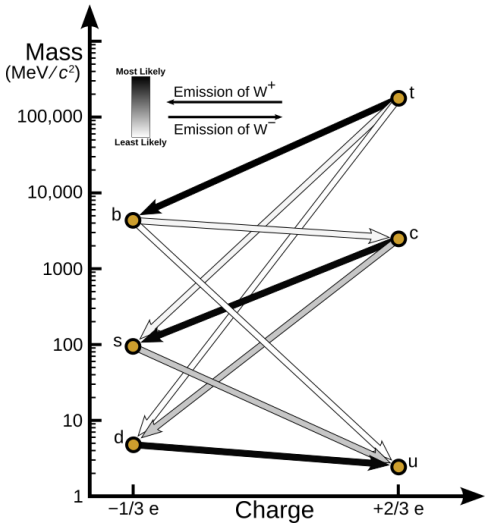

Figure 1.8: A graphical representation of the quark transitions rates due to the weak interaction. Provided by Ref. [3]. The x-axis distinguishes up-type and down-type by the charge while the y-axis shows the mass of each particle on a logarithmic scale. the color of the arrows represents the likelihood of a transition with black being most likely and white being least likely. The direction of the arrow indicates the direction of transition based on the mass differences (arrows always flow down in mass).

triplets - *uud* and *udd* - as our two building blocks for atomic nuclei. Though, note that not even the neutron is safe! It has a lifetime of almost 15 minutes but it too has a non-zero probability to decay via d to u. The proton is largely believed to be stable - though there are active efforts to measure the proton's decay which, having never been observed, place limits of its half-life at no less than 10^{34} years.

The next example is of muon decay, mentioned in Sec. 1.2.2. As stated, the muon has a long lifetime (2×10^{-6} seconds) and this, paired with relativistic effects, is how they can get from space to Earth without decaying. The tau has a much shorter lifetime of 10^{-13} seconds and as a result, we do not observe them as cosmic rays. Unlike the quarks, the weak force couples equally to all leptons - there is no CKM-like matrix mixing trick. However, recall that charged leptons decay to a lepton neutrino and a W boson (Fig. 1.9). That means the muon, a 106 MeV particle, must decay into a massless (but still energetic) neutrino and an 80,000 MeV W boson[3]! By energy conservation, this cannot happen and so the muon must sit and wait for a virtual W which, by the Heisenberg uncertainty principle, can have mass compatible with energy conservation provided it exists for a very short period of time. The 2 GeV tau has a much shorter lifetime (3×10^{-13} seconds) primarily because of its mass being closer to the mass of the real W. Specifically, the lifetime of a lepton is proportional to $\frac{1}{m_l^5}$. Taking the ratio of $\frac{1}{m_\mu^5}$ to $\frac{1}{m_\tau^5}$, we're left with a factor of about $\frac{1}{16^5}$ - or about 10^{-6}. This accounts for most of the difference but the tau additionally has a larger decay phase space since it has enough mass to decay

[3] Usually written as 80 GeV but implicit unit conversions do not convey the comparison well.

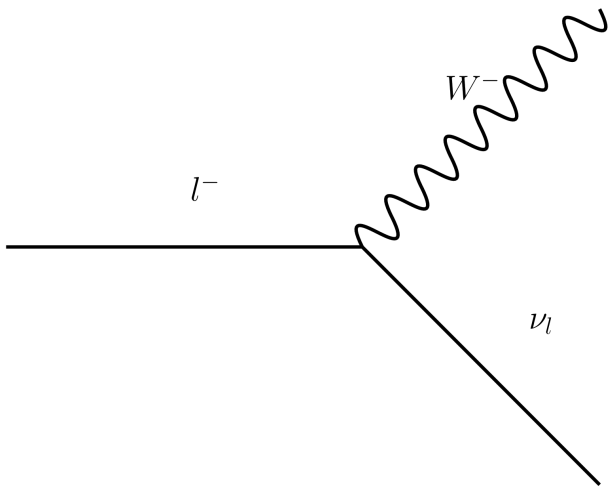

Figure 1.9: Generic decay of a charged lepton via the weak interaction.

to (via the W) uc and us as well as to a muon (in addition to the electron). Accounting for the CKM terms for uc and us, the phase space creates another factor of five. In combination, this gives 3.8×10^{-13} seconds for the lifetime of the tau with some precision lost relative to the true value since this is just an approximate example to show the dominant effects.

Finally, consider a similar scenario but for quarks. Of the six quarks, only the top quark is heavier than the W boson, making it the only quark which can decay to a real W. By the CKM matrix, the top quark decays to a W boson and bottom quark nearly 100% of the time and so the top quark's lifetime is very short. Conversely, the resulting bottom quark has a relatively long time - why? As we saw with the muon, the need for a virtual W boson is a factor since the bottom is only 4 GeV in mass. However, consider the other elements of the CKM matrix involving the bottom quark; they are 0.04 and 0.004 for the b mixing with the c and u, respectively. Thus the bottom quark eventually decays to a c or u but waits about 10^{-12} seconds to do so because

the mixing terms are small. This is not unique to the bottom quark; the lifetime of the strange quark is even longer at about 10^{-9} seconds. However, the long lifetime of the bottom quark will be relevant again in Ch. 2 when discussing bottom quark identification techniques.

1.3 The Excited Quark Expansion

To end this chapter, we consider an expansion to the SM rules which can be motivated with a short retelling of how humans have viewed particles through history.

Since ancient times, philosophers and physicists alike have discussed the prospect that all pieces of the observable universe are made up of one or more "indivisible" particles. The Greek philosopher Democritus proposed the theory of atomism (from the Greek *atomos*, meaning "uncuttable"). Atomism describes the world as being made of an infinite number of indivisible particles. After sorting materials based on experimental observations, scientists developed a theory that a common piece connects the elements and, borrowing from Democritus' approach, they called it the atom. J.J. Thompson's discovery of the first subatomic particle – the electron – and Hans Geiger and Ernest Marsden's observation of the nucleus showed that the atom was also composed of parts, leading to the development of the Bohr model. So much for "uncuttable"!

As we know today, two smaller parts make up the nucleus inside atoms - the proton and the neutron - and combinations of these two particles with electrons can create any atom. In the latter half of the 20th century, we learned that even the proton and neutron are divisible into quarks and gluons. Just as

different combinations of protons, neutrons, and electrons can create different atoms, different quark combinations can construct different hadrons.

The electron remains, as far as we know, indivisible and is considered a "fundamental" particle along with the quarks, bosons, and other leptons, that make up the Standard Model of particle physics. But if there is any lesson to be learned from the history above, it is that there is no reason to declare that we have reached the most fundamental of pieces to our universe! Thus, physicists continue to probe the existence of a substructure to the so-called "fundamental particles".

One such model is for an excited bottom quark, b^*. If a quark is made up of other constituent particles, the energy required to split them would be higher than the energies provided by the LHC. However, it may be possible to "excite" the bottom quark in a way similar to an electron being excited to a higher orbit in an atom by the absorption of a photon. A quark could instead be excited by the absorption of a gluon, as these are abundant in the proton–proton collisions made by the LHC. In such an interaction, the energy of the gluon would be converted into internal energy of a bottom quark which would then appear as a very heavy particle, called an excited bottom quark. The new particle would be so massive that any decay to two standard model particles will result in vast amounts of binding energy being transferred to the lighter decay products as kinetic energy. In that case, such a new particle would create a back-to-back signature with the decay particles in opposite hemispheres of the detector, as shown in Fig. 1.10.

The excited bottom quark could decay back into a gluon and a bottom

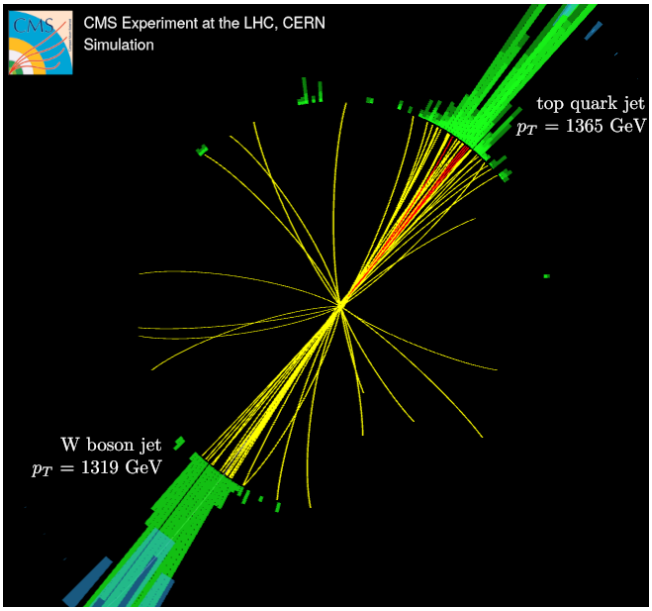

Figure 1.10: Event display of a simulated b^* event in the CMS detector with one jet tagged as a W boson and one as a top quark.

quark (just as an excited atom eventually emits a photon so that the electron decays back into a lower electron shell). However, the decay into a top quark and W boson is predicted to be two times more likely. The analysis described in Ch. 4 additionally only considers the case when the top quark and W boson both decay into large energetic hadronic showers since it is the most probable decay of both particles. The resulting Feynman diagram describing both the production and decay is presented in Fig. 1.11

1.3.1 Overview of analysis strategy

Unfortunately, there are many ways that hadronic showers are made in LHC collisions, so one needs to first discern excited bottom quark events from similar-looking background collisions. The analysis in Ch. 4 uses specialized algorithms to identify top quarks and W bosons. In particular, the algorithms

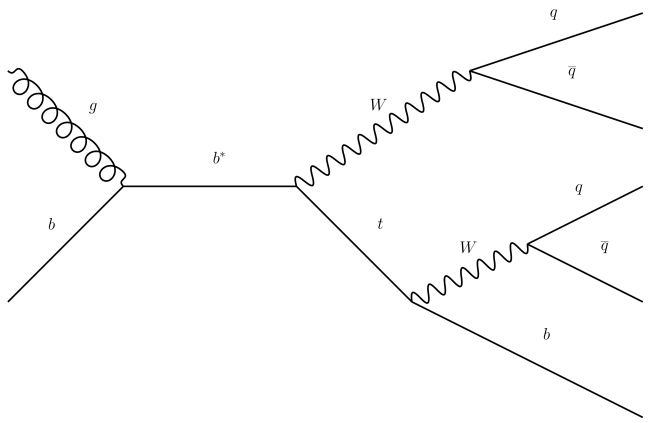

Figure 1.11: Feynman diagram describing the production and decay of the proposed excited bottom quark, b^*.

analyze the showers to determine how many "cores" are inside the more extensive showers. A W boson decays into two light quarks so will have two cores. The top quark will have three cores because it will decay into a bottom quark and a W boson. Finally, one can identify the bottom quark inside the top quark shower because a particle containing the bottom quark will travel a measurable distance in the CMS detector.

There are still many collisions from other, already known, particles that exist in the data even after using these dedicated algorithms to identify the top quark and W boson. These are accounted for by using a combination of simulation and measurements of the data. These estimates create a model of the background of Standard Model particles that have a similar signature in the detector. When analyzing the data, one can compare this background-only model to a model where simulated excited bottom quark events have been added to the background.

Because we do not know the mass of the excited bottom quark, scenarios

with excited bottom quarks of different masses and predictive models are compared against the background-only model. The statistical analysis of the data checks if one of these models describes the data better than the background-only model. If the model with a signal does not reproduce the data significantly better than the background alone, we instead set a limit on our sensitivity to detect excited bottom quarks below a certain mass.

More details on the analysis follow in Ch. 4.

Chapter 2

Experimental overview

As was stated in Ch. 1, the better we know the rules, the better we are able to use them to our advantage. With this in mind, this chapter covers how high energy particle experiments are constructed using some of the rules explained in Ch. 1.

The following sections will cover information about the Large Hadron Collider, the Compact Muon Solenoid detector, and some of the techniques used to reconstruct objects relevant to the analysis described in Ch. 4. Before that though, we first consider the much simpler Rutherford gold foil experiment which is instructive for understanding the bigger picture of a particle physical experiment.

Though commonly named after Rutherford, the gold foil experiment was actually performed by Rutherford's graduate students, Geiger and Marsden. The experiment was designed to test the so-called "plum pudding model" of the atom which proposed that the positive charge of the atom was uniformly distributed in the atom, with negatively charge "raisins" distributed evenly within the positive "pudding". Positively charged alpha particles made of

two neutrons and two protons were fired at a thin gold foil with the expectation that the heavy alpha particles would pass through the foil relatively undeflected, as shown on the left side of Fig. 2.1.

However, Geiger and Marsden instead observed a small but significant number of alpha particles being deflected at large angles relative to their incident direction, as shown on the right hand side of Fig. 2.1. The large angle scattering was eventually attributed to collisions of the alpha particles with dense atomic cores which we now call nuclei. Thus, Rutherford proposed his own model of the atom where the positive charge is concentrated at the center (along with the majority of its mass).

While this result played a crucial role in the advancement of nuclear and particle physics, the experimental setup and observation are the most relevant parts for our consideration. In particular, Rutherford, Geiger, and Marsden had a hypothesis about the angular distribution of the alpha particles after colliding with the gold foil and tested it by simply counting the number of events as a function of the deflection angle. The hypothesis was that the events would be distributes as a relatively narrow peak at zero degrees, with no events occurring outside a window of maybe a few degrees. The observed distribution was instead a slightly narrower peak with outliers far from a deflection angle of zero. This, at its simplest, is a binned shape analysis performed as a function of the angle of the detector. It is also, at its fundamentals, no different from how many CMS analyses are performed today.

We still wrap the interaction point with a cylinder - albeit much longer

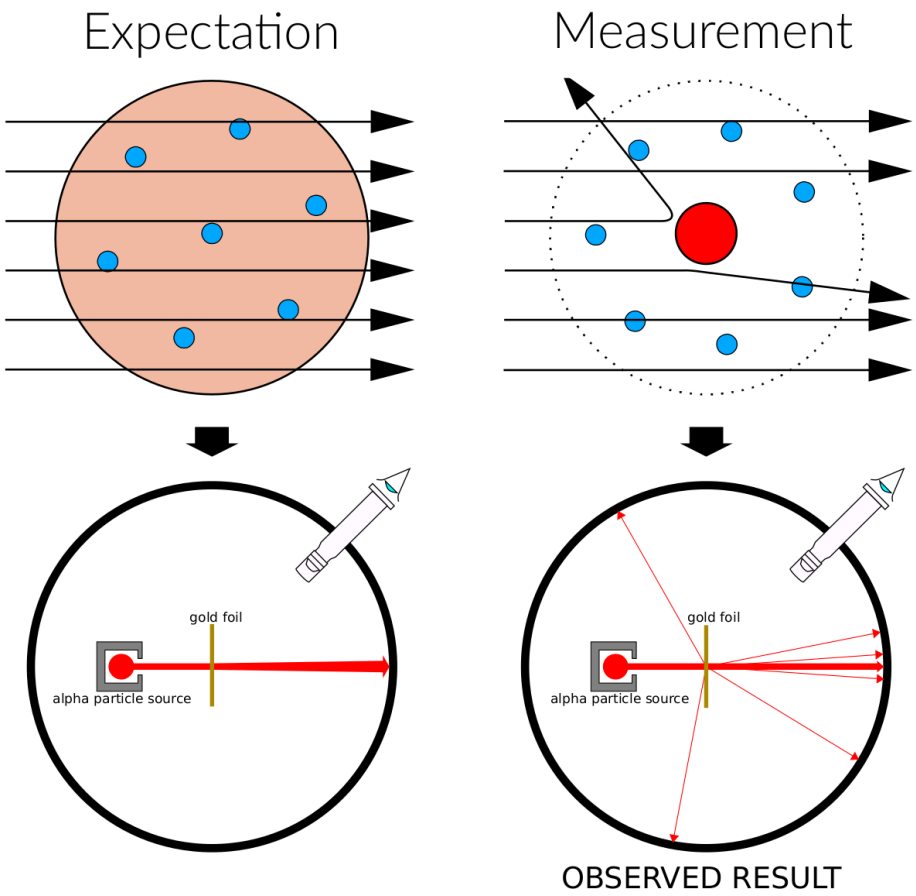

Figure 2.1: Cartoons of the expected and observed outcomes of the Rutherford gold foil experiment. The top row shows the two models of the atom while the bottom row shows the experimental outcomes from the corresponding atomic model. Adapted from Ref. [4].

and with end caps; we still use techniques to detect the positions of charged particles coming out from the interaction; we still examine the distribution of events as a function of geometric variables; and we still propose hypotheses about the distribution and look for discrepancies once the measurement is performed. The experiments have only become more complex as the possible interactions and the number of particles involved has grown. Therefore,

before covering the data analysis any further, let us first examine how to produce collisions and measure what comes out in modern high energy collider experiments.

2.1 The Large Hadron Collider (LHC)

The LHC is a particle accelerator and collider. The "L" in "LHC" refers to the fact that the apparatus itself is very large - a 27 km ring, called a "synchrotron", sitting 100 meters underground on the border of France and Switzerland. The "H" indicates that it manipulates hadrons, defined in Ch. 1. The "C" means the hadrons are "collided" together at high energies - the highest energies ever achieved, in fact.

While linear particle accelerators are also possible, a synchrotron has several advantages. First, there does not need to be more length to speed up a particle - just keep "kicking it" so the next go-around is always at a higher energy. Second, the two colliding beams can share the apparatus but rotate in opposite directions. Finally, for a given beam energy, the two beams collide at a higher energy than one beam colliding against a fixed target. In other words, the collision has a higher "center-of-mass energy," a standard measure of the collision energy no matter if there are two beams colliding or just one colliding with a stable target. The LHC center-of-mass energy (denoted with $\sqrt{s}$) is 13 TeV and is the highest energy at which any particles have been collided by humans.

One of the reasons the energy is so high is because the LHC uses protons which are relatively heavy. Of course, collision energies cannot be infinitely

increased just by using heavier particles. Heavier particles have higher inertia which means they are harder to accelerate. [1] However, lighter particles have their own disadvantages. One downside of the synchrotron is that charged particles lose energy via radiation when traveling fast and bent by a magnetic field. This effect is worse for lighter particles (like the electron) compared to heavier particles (like the proton), making an electron synchrotron a poor option. In other words, the proton has a Goldilocks mass that strikes a balance between these two effects.[2]

Another advantage of using hadrons is that their constituent quarks and gluons interact via the strong force. Once two protons collide head-on at such high energies, they stop interacting as composite protons and start behaving like jello-y blobs of quarks and gluons that mush together, the collisions then being between the protons' constituent quarks and gluons. The protons are just the chaperon to get the quarks and gluons to the party and are a result of the fact that quarks and gluons cannot exist by themselves. But if they cannot exist by themselves, how do they collide?

As discussed in Ch. 1, asymptotic freedom causes the strong force interactions between quarks and gluons to become weaker as the energy scale increases and the distance between particles decreases. This is why we can think of the protons as just the vehicle to carry the quarks and gluons. In this "hot soup" of quarks and gluons, particles can start hitting each other and with luck, two of them hit dead-on creating a "hard scatter" (a "soft scatter"

[1] Note that the LHC does also collide lead particles at certain times, albeit at lower center-of-mass energies.

[2] Plus, they are easily made by stripping hydrogen of their electrons.

would be the equivalent of a grazing deflection). Considering conservation of energy in this process, one may realize that this collision is *not* going to be at $\sqrt{s} = 13\,\text{TeV}$.

The energy in the "hot soup" is distributed according to Parton Distribution Functions (PDF) which describe the probability to find a constituent particle (parton) with a given fraction of the total momentum of the system and as a function of the total energy of the system. Depending on the particle and energy scale, the PDFs can have different shapes but will asymptotically approach zero as the momentum fraction increases, as can be seen in Fig. 2.2, which shows the product of the momentum fraction x with the distribution function f. Thus, two colliding particles will only carry some portion of the 13 TeV and the product of the four-momenta of two particles, $x_1 \cdot x_2$, then determines the energy scale of the collision and the resulting decay products. Since higher momentum fractions are less likely, lower energy collisions are more likely than higher energy collisions and even the highest energy collisions will not be at 13 TeV.

Additionally, despite the term "proton–proton collisions" implying that singular protons go around the ring, there are actually "proton bunches" which are spaced at 25 ns.[3] The advantage of these bunches is that they increase the probability of two protons colliding in a hard scatter in a given "bunch crossing." The corresponding disadvantage is that there many soft collisions and the products of these collisions can pollute the products of the interesting hard scatter. This is an effect called "pileup" which will be addressed in

[3]Note that everything is moving at very near the speed of light so this corresponds to about 7.5 meters.

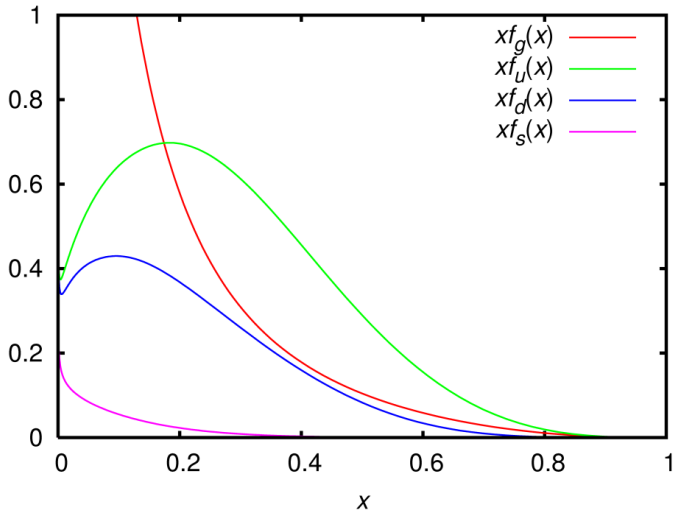

Figure 2.2: A plot showing the product of longitudinal momentum fraction, x, and PDF, f, as a function of x for the gluon (red), up (green), down (blue), and strange (magenta) and with a resolution scale of 2 GeV. As can be seen, all four become smoothly falling distributions above x of 0.4. Provided by Ref. [5].

Sec 2.3.2.

While the physics of accelerators and colliders is far more complex than what has been presented here, these are the basics needed to understand the next chapter in the story - the detector.

2.2 The Compact Muon Solenoid (CMS)

Out of the four large detectors around LHC ring, the one we focus on is the Compact Muon Solenoid. The detector being dubbed "compact" perhaps seems to be a misnomer - the entire apparatus is about four stories tall. However, the amount of material is very densely packed into layers that look like a cylindrical onion. The order of the layers is designed to optimally use the SM "rules" to identify particles coming from the primary collision vertex. Looking

at a "pizza slice" of the detector in Fig. 2.3 shows how each layer interacts with different types of particles.

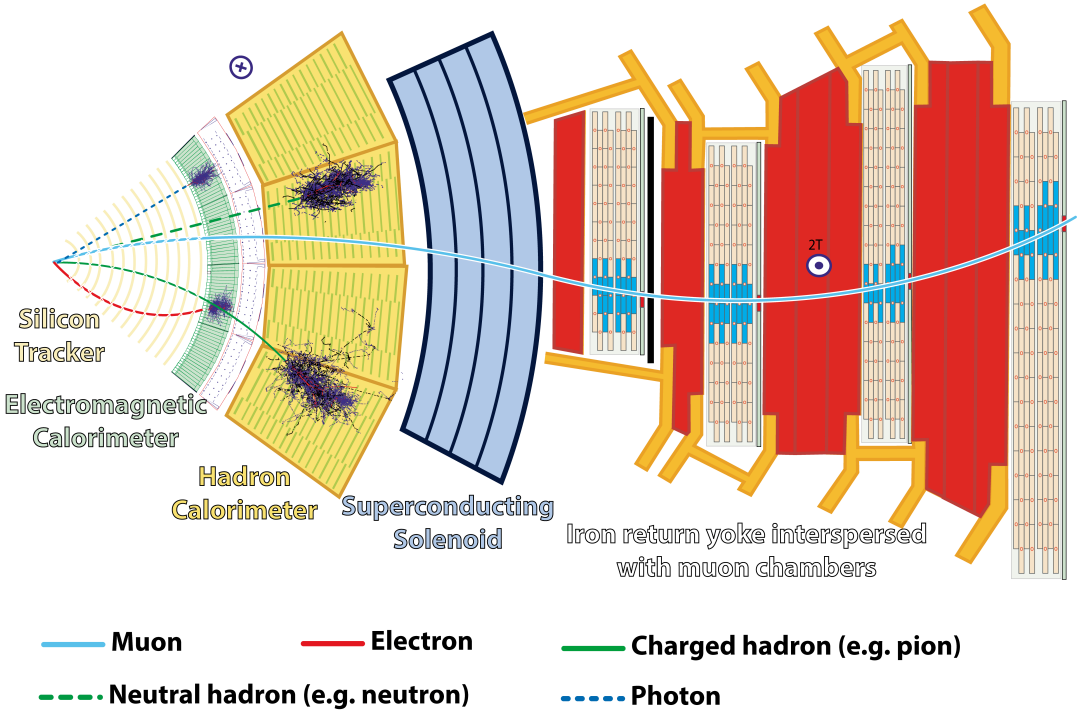

Figure 2.3: Diagram of the CMS detector layers, presented from the perspective off looking down the beamline and isolating a "pizza slice". The sub-detectors are labeled along with the direction of the magnetic field and several different particles whose paths are traced through each layer. Provided by Ref. [6].

The silicon tracker works similarly to a modern digital camera - particles that interact via the electromagnetic force (charged particles plus photons) interact with the silicon material, depositing charge which is then collected to register a positional "hit". The sub-detector is made up of tens of millions of pixels giving it very precise positional tracking which makes it possible to reconstruct the particle's track as it passes through each layer.

This is powerful for identifying bottom quarks, which, if you recall, have long lifetimes. If a bottom quark is produced from a hard scatter, it will have

enough momentum to travel several centimeters before decaying. The decay products will leave tracks in the pixel detector that converge back to a vertex that is "displaced" from the vertex of the hard scatter. Thus, by measuring this displacement, one can identify potential bottom quarks in the event. Of course, there are *many* charged particles passing through the silicon tracker for a given event and the detector does not see a "line" as a track - it just sees individual hits at each layer. This makes the track reconstruction a complicated and computationally intensive process.

In addition to being able to determine the positions of particle paths, the tracker makes use of the fact that it lies within the CMS "solenoid" which creates a constant magnetic field and causes all charged particles to have curved paths. While this further complicates the track reconstruction algorithm, it also allows the reconstruction algorithm to identify the momentum and sign of the charge of the particle passing through the layers based on how much the track curves and in which direction.

The next two layers of the detector are the electromagnetic (ECAL) and hadronic (HCAL) calorimeters, respectively. As can be seen in Fig. 2.3, these layers are the "brick walls" that stop most particles from leaving the detector. Like the silicon tracker, the ECAL interacts with electromagnetically interacting particles. However, the lead tungstate crystals used in its construction are much denser than than the silicon in the tracker and are designed to capture the energy of 98% of electrons and photons with energies up to 1 TeV. While, the ECAL's primary goal is not to interact with neutral particles or to stop hadrons, two thirds of hadrons coming from the primary collision will have

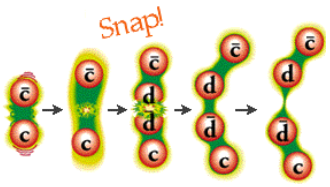

Figure 2.4: A cartoon of quark color tubes breaking and reforming as new quarks "snap" into existence. Provided by Ref. [7]

also interacted with this layer before making it to the HCAL. However, any charged hadron or muon will not be significantly slowed down.

This is why the HCAL comes *after* the ECAL. With the electrons and photons filtered out, what remains are just the hadrons and muons. As the name suggests, the HCAL is designed to stop hadrons and measure their energy. It does this with alternating layers of absorbing material (either brass or steel) and active scintillator, which absorb and measure energy respectively. Specifically, the dense absorbing material causes nuclear interactions via the strong force that slow the incoming hadrons. When colliding with the absorbing layer, the particles turn into cascading showers via a process called "hadronization" and the scintillating layer collects the energy of the shower's much lighter constituents.

Hadronization is a result of the strong force interactions described in Sec. 1.2.4. Recall that, because of confinement, two quarks traveling away from each other will eventually "snap" their color tube, forming two new quarks in the process, as shown in Fig. 2.4. As high energy quarks leave the collision, this process happens many times, forming a particle shower called a "jet", as shown in Fig. 2.5. The cascade eventually ends with the final particles being collected as energy deposits in the scintillator layers.

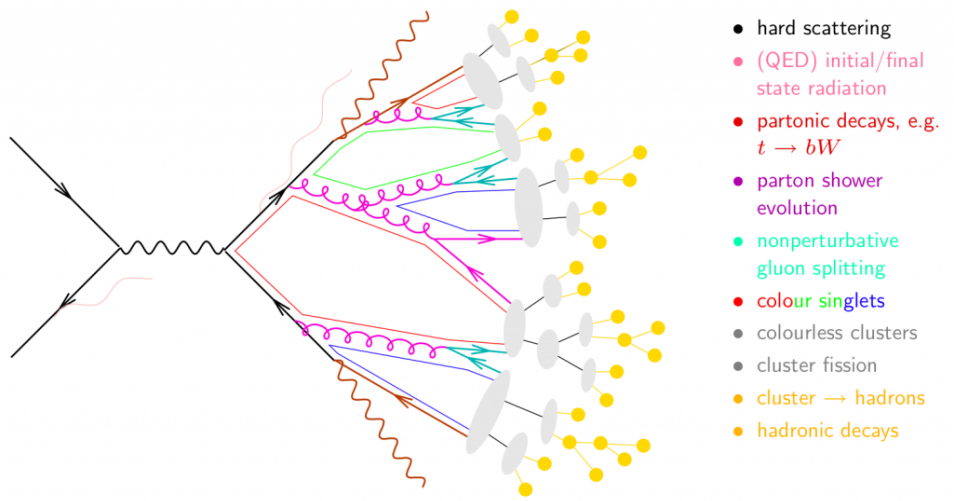

Figure 2.5: A modified Feynman diagram showing an example of initial decay particles showering via hadronization. Provided by Ref. [8]

This brings us to the last layer of the detector: the muon chambers. The unique muon detection system is what earned "muon" a place in the "CMS" acronym. The muon chambers live of the solenoid wrapping and have the job of identifying the remaining muons. Notice that I do not say "stopping the muons". The muon is 200 times heavier than an electron so if an electron is a marshmallow (about 7g), the muon is a World War II mortar shell (1.4 kg)! As I once heard at an APS talk, "They're muons so they're real porkers!"

The muon does not interact via the strong force so we are only left with electromagnetic interactions to slow them, which is not sufficient without a lot of dense material. However, because all of the other layers of the detector have been used to filter out other particles, we know only muons exist at the final layer and thus, we know the mass of all interacting particles at this layer.

We also know the magnetic field. As can be seen in Fig. 2.3, the muon path starts to curve in the opposite direction from the path inside the solenoid. This

is because the direction of the magnetic field outside the solenoid points in the opposite direction compared to the direction inside the solenoid. Thanks to the iron return yolk which alternates between blocks of the muon chambers, the magnetic field on the outside of the detector is a constant value due to the fact that iron saturates at 2 T. Thus, with the deduced particle mass and the constant magnetic field, one only needs to measure the curvature of the track to get the muon momentum.

As also can be seen in Fig. 2.3, the size of the detectors at this layer are quite large and so the positional precision is lower than for the ultra-precise silicon tracker (which also has the benefit of interacting with a muon before it loses energy to the other layers). However, the isolated tracks in the final layer inform the position and momentum that can be extrapolated in reverse to inform the reconstruction algorithms in the tracker. In fact, all of the detector information is used in cohesion in a "particle-flow" (PF) reconstruction and event description algorithm so that all of the distinguishing characteristics of the different layers can be used synchronously; another way our understanding of the SM rules inform how we "play the game".

Specifically, the algorithm links interactions in the various layers together based on their spatial proximity with stronger links created by closer detector elements. Blocks of elements are created to represent the possible set of links that could be combined to construct a PF particle. Muons are considered first since the muon chambers provide a unique signature and the tracker provides precise measurement of muon position and momentum. The elements used to reconstruct the muon are then removed from the algorithm before considering

electrons and energetic and isolated photons - both of which use links between the ECAL and tracker. Note that while the ECAL informs the link, the curved tracks of the electron are used to reconstruct its momentum. The electron and high energy photon elements are removed and the remaining hadrons and photons are reconstructed from what blocks remain. The hadrons specifically use links between the tracker, ECAL, and HCAL with the momentum of charged hadrons being also determined by information from the tracker. More details on the detector components and this algorithm can be found in Ref. [9].

2.2.1 Detector coordinates and common variables

The CMS detector also has a coordinate system and set of common kinematic variables which should be defined before proceeding. A cartoon of the cylindrical detector is provided in Fig. 2.6 which shows the standard Cartesian (x, y, z) parameterization as well as the two common angular coordinates, which take the origin as the center of the cylinder: the azimuthal angle, ϕ, and the "pseudo-rapidity" η. To understand the pseudo-rapidity, we should first define rapidity, y, as

$$y = \frac{1}{2} \ln \left(\frac{E + p_L}{E - p_L} \right) \tag{2.1}$$

where E is the energy of the particle and p_L is the longitudinal component of the three-momentum. When the magnitude of the three-momentum, $|\vec{p}|$ is much greater than the mass of the particle, the pseudo-rapidity converges to the rapidity. The rapidity is a preferred quantity because differences in rapidity, Δy, are Lorentz invariant to boosts along the longitudinal axis. Additionally, particles are distributed roughly evenly as a function of rapidity.

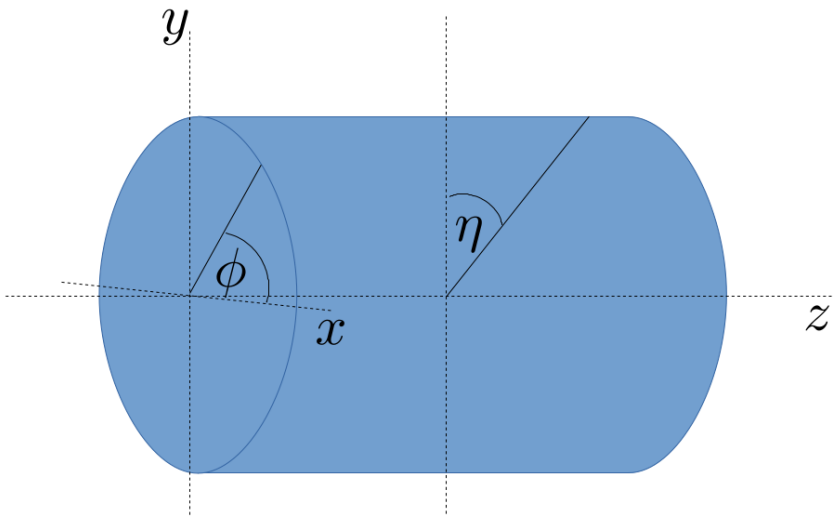

Figure 2.6: A rough sketch of the CMS detector coordinate system.

In the limit of ultra-relativistic particles, Eq. 2.1 simplifies to η which is defines as

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{2.2}$$

where θ is the traditional polar angle, defined as zero when parallel to z. It is the more common coordinate compared to y since it only relies on the polar angle, θ.

Note that the radial distance is not a terribly interesting quantity which is why it is not shown in Fig. 2.6. The particles coming from the collision are obviously traveling outwards but their distance from the interaction is not part of any conservation law we can use. Instead, we care about particle momentum and, in particular, the transverse momentum, p_{T}. This, combined with ϕ and η, allows us to reason about the kinematics of the decay products and to inform the particle-flow algorithm when it reconstructs the event.

The final quantity of interest for the sake of this dissertation is the H_T

which is the scalar sum of the p_T of the hadronic activity (i.e. jets) in the event. It serves as a simple tool to quickly identify events high energy hadronic activity (a sign of interesting physics). Analyses will sometimes simplify the definition as the scalar sum of the p_T of just those jets in the event that the analysis cares about - i.e. just the first few very energetic jets that capture most of the true H_T.

2.3 Data analysis and jet reconstruction

In addition to the far more sophisticated and complicated experimental setups that modern particle physics necessitates when compared to Rutherford's experiment, the analysis of the data is also more complicated. For example, Rutherford, Geiger, and Marsden were interested in all alpha particle collisions with the gold foil. In CMS data analysis, only a subset of the data will be interesting for a given analysis. If the goal is to look for evidence of a BSM particle decaying to exactly two high energy leptons, it would not make sense to look at events with anything other than exactly two high energy leptons.

Creating a subset of the data that matches the analysis criterion is called creating a "selection". The events after the selection are also sometimes referred to as a "region", referring to the fact that the selection is isolating a region of the data phase space. For gold foil experiment, the phase space was defined entirely by the angle of the alpha particle after colliding with the foil. They were interested in the entire phase space so there was no selection. With modern high energy physics data, there are many, many more variables defining the space, each being a different "axis" or "dimension" of the space.

By selecting on one, we segment the space along the axis of the variable being selected. Such a variable is thus treated as a "discriminant." One can also project the events onto one or more of the axes and use the corresponding distribution to make a measurement - this would be a "measurement" variable.

Another key difference between the gold foil experiment and CMS data analysis is that CMS uses computer-made simulations to model physics processes. The simulation can be used to try to describe the data or as just a source to validate a procedure. In any case, simulations are created using Monte Carlo (MC) methods which sample probability distributions to create simulated particle physics interactions. As one might imagine, this is incredibly complicated because the entirety of the rules described in Ch. 1 (including those pieces skipped over) have to be transcribed for the computer to understand. Based on these probabilities, the computer generates a random number for each vertex in the interaction chain to determine what the event looks like - which forces are involved, the lifetime of the resulting particles, the kinetic energy and position in space of the particles, etc. Then the outgoing particles must have their interactions with the detector also simulated! The simulation is thus split into two primary steps: one for the physics simulation and one for the simulation of the detector's response to "the physics".

At their simplest, the simulation steps are already very complicated. However, modeling the underlying fundamentals means that the final distributions in any given measurement variable are relatively trustworthy. Or at least, they are physically motivated which means they are good enough for us to make decisions with before we even look at data.

When MC simulation is sufficiently accurate, we can also use it to model known SM processes in the data. Even if there are residual differences between simulation and the data, they can be measured and the measurements of the differences can be used to apply corrections to the simulation. Uncertainties in the measurement can also be propagated as systematic uncertainties in the analysis. The case of simulation not being sufficiently accurate for a given physics process is covered in Ch. 3. In these cases, physical intuition is paired with model building and validation techniques to measure the contribution of the process in data as a function of the measurement variable(s).

Below, are several concepts needed to understand data-driven modeling techniques described in both Ch. 3 and Ch. 4

2.3.1 Data triggering

As stated in Sec. 2.1, proton bunches are collided every 25 ns; that is 40 million collisions per second! Even in a very hard scatter with very high energy particles coming out from the interaction vertex, the next collision will be happening before the previous event can even be detected, let alone recorded. Additionally, the majority of the events do not contain hard scatters with interesting physics that one would care to study.

Thus, CMS uses a hardware trigger system called the "Level 1" which acts extremely quickly to identify whether a given event contains evidence of an interesting hard scatter or not and whether the rest of the detector should read out information to the computers storing the raw data. The Level 1 triggering system will specifically choose about 100,000 events per second out of the

billions.

Note that this system requires very precise timing information so that the various electrical signal channels arrive to the trigger in sync. Assuming the electrical signal in a cable can travel at the speed of light, signals between events are separated by about 25 feet of cable. For a four story tall detector, the cables can have lengths that vary well beyond the scale of 25 feet and so the various signals must be synced to ensure the signals arrive together, giving the Level 1 hardware has as much time as possible to make a decision before the next group of signals arrive. This synchronization is provided by the data acquisition system (DAQ). An effect called "trigger prefiring" was discovered for the data collected in 2016 and 2017 that was a result of mis-timings of the trigger signals. This is described more in Sec 4.7.

If an event is considered interesting by the Level 1 trigger, the information is passed to the High-Level Trigger (HLT) which is a software-based system. It will perform rudimentary reconstruction of physics objects and make decisions based on these to select about 400 events per second to be read out to storage. The physics objects will be reconstructed more robustly in separate offline processing. The events recorded are guaranteed to have at least one trigger fired but, an event could pass multiple triggers. Therefore, CMS analyses can use the HLT Boolean values as simple filters to select a subset of data that has desirable attributes defined by the HLT. More can be read about the CMS trigger system in Ref. [10].

The fast reconstruction at the level of the trigger means that it is not fully efficient at selecting events near the thresholds defined for recording the event.

For example, if a trigger requires at least 1 TeV for the H_{T} of an event, the selection efficiency for events with true H_{T} just below 1 TeV will not be 0% (and just above 1 TeV will not be 100%). The trigger's rough reconstruction will be wrong by some amount which will create a "trigger turn-on" over which the trigger efficiency sharply rises and then slowly plateaus to 100%, as can be seen in Fig 4.1.

This is particularly relevant because the effect must be simulated in MC simulation of backgrounds and signal. The trigger selection *is* fully simulated but, just as with any part of the simulation, it is imperfect. In particular, the turn-on curve may be different from the one in data for the same HLT. One common solution to this discrepancy is to measure the efficiency in a subset of the data that is separate from the one considered in an analysis. This subset could be selected, for example, using the logical AND with another trigger that selects events one would not normally care about but that do not bias the efficiency one is attempting to measure. For example, if measuring the efficiency for an H_{T} based trigger in an all-hadronic analysis, one could use the logical AND with a muon based trigger to create a subset of events in which to measure the efficiency of the H_{T} trigger to select events.

Then one could measure the ratio of the efficiency in data to the one in simulation as a function of some variable relevant to the trigger's efficiency (H_{T} in the above example). This correction is then applied to the simulation in the signal region selection *with* the trigger bit selection included, thus weighting the simulated events to match the data. An alternative solution is to measure the efficiency of just the data in the orthogonal selection and then

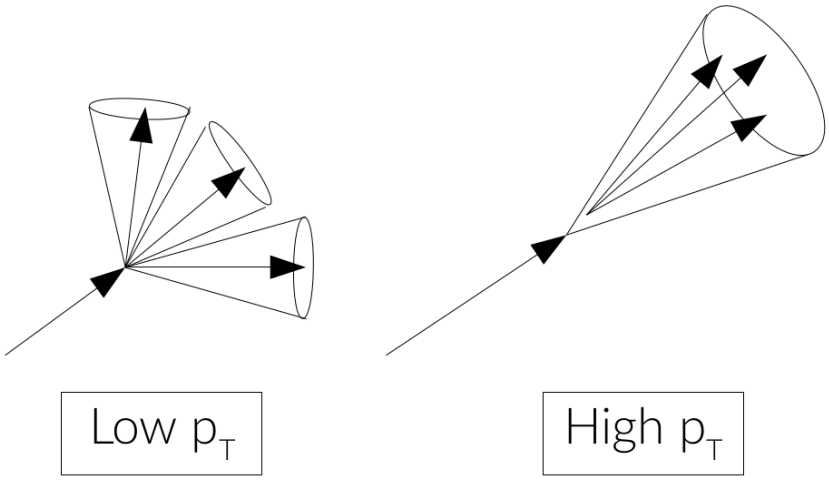

Figure 2.7: Cartoon of jet formation via collimating of the decay products.

apply this as a weight directly to the simulation *without* the simulated trigger bit selection applied. The latter is used in the analysis described in Ch. 4.

2.3.2 Jet reconstruction

Fig. 2.5 showed how particles undergoing hadronization produce showers of particles. If the original particle has a large amount of energy (either kinetic or inherently in its mass), the lighter decay products will be "boosted" (i.e. have high momenta). In the lab frame, this will cause the shower to "collimate", as shown in Fig. 2.7. As stated in Sec. 2.2, these collimated showers of hadrons are called "jets" and depending on the amount of boost and the reconstruction algorithm used, will be of varying size and shape.

Various algorithms exist for reconstructing a group of particles as a jet including the PF algorithm described in Sec 2.2. One of the more common jet reconstruction algorithms used by CMS analyses (including in Ch. 4), is

the anti-k_T[4] algorithm which defines d_{ij} as the distance between two entities, i and j, in the candidate jet and d_{iB} as the distance between entity i and the beam using

$$\begin{aligned} d_{ij} &= \min\left(\frac{1}{p_{T,i}}^2, \frac{1}{p_{T,j}}^2\right) \frac{(\Delta y_{ij})^2 + (\Delta\phi_{ij})^2}{R^2} \\ d_{iB} &= \frac{1}{p_{T,i}^2} \end{aligned} \tag{2.3}$$

where R is a chosen a parameter analogous to the jet "radius", $p_{T,i}$ is the transverse momentum of particle i, and Δy_{ij} and $\Delta\phi_{ij}$ are the difference in rapidity and azimuthal angle, respectively, between particles i and j.

For a given i, the algorithm will compare d_{ij} and d_{iB}. If d_{ij} is the smaller of the two, entities i and j will be combined. If d_{iB} is the smaller, then i is called a jet and its constituents are removed from consideration in the rest of the clustering. If i is a hard particle (high p_T) and j is a soft particle (low p_T), then the algorithm will cluster the soft particle to the hard particle, defining a new entity which is the combination of i and j. Thus, the anti-k_T algorithm will merge soft particles into the closest hard particle. This is in contrast to the k_T algorithm which merges soft particles together first. An example of the clustering is shown in Fig. 2.8 for $R = 1$ [11]. The most common R values used by CMS are $R = 0.4$ for "skinny" AK4 jets and $R = 0.8$ for "fat" AK8 jets (where "AK" denotes "anti-k_T "). Fat jets are used exclusively in the analyses described in Ch. 4.

[4]Note that k_T is used to denote the transverse momentum of a particle which was already defined in Sec. 2.2.1 as p_T. For consistency, p_T will continue to be used in this dissertation.

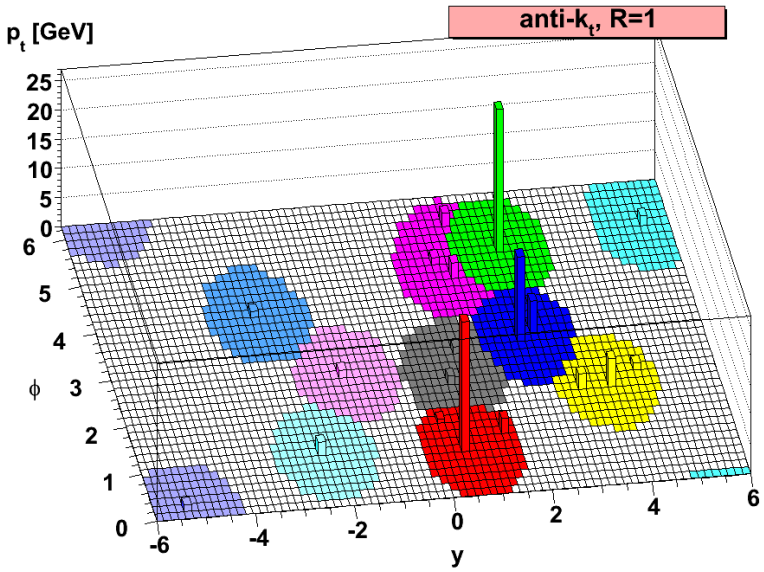

Figure 2.8: An example of jets clustered using the anti-k_{T} algorithm and $R = 1$. The z-axis is the p_{T} of the particle in a given (ϕ, y) bin while the colored regions represent the edges of the reconstructed jets based upon these particles. Provided by Ref [11].

Recall from Sec 2.1 when discussing the proton bunches that, even if a hard scatter occurs between two protons, many soft, glancing collisions still occur. In fact, there can be multiple primary vertices reconstructed, as can be seen in Fig. 2.9. As stated, the resulting noise is called "pileup" and it can easily contaminate jets.

To remove this contamination, each jet constituent is evaluated for its probability to have originated from pileup with an algorithm called Pileup Per Particle Identification, or PUPPI [12]. Using the precise positional tracking of the CMS pixel detector, charged particles originating from pileup can be traced back to origins along the beamline that are displaced from the primary hard collision. However, the tracker cannot account for electrically neutral particles. PUPPI accounts for these by calculating a probability that the neutral

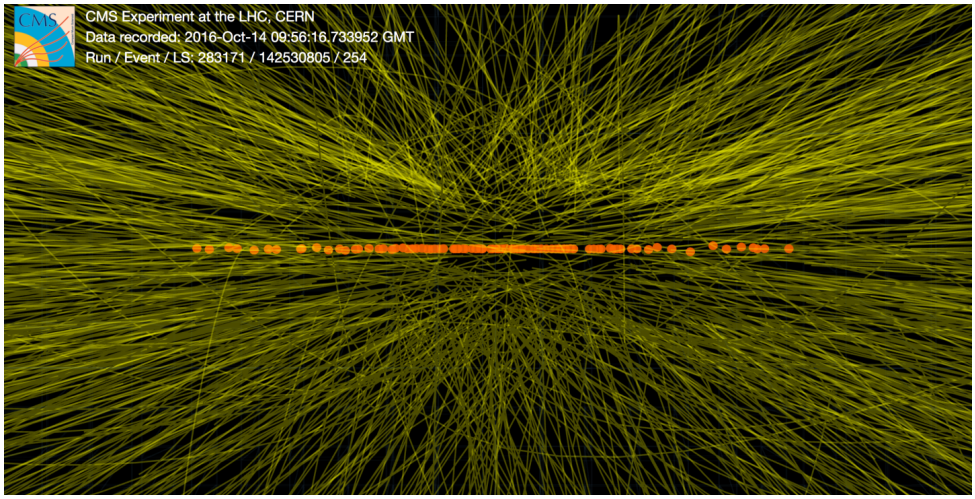

Figure 2.9: A CMS event display showing the contributions of many soft scatters that produce pileup. Provided by Ref [13].

particle originates from a pileup interaction based on how close the associated calorimeter deposits are to the deposits of the charged particles from the primary interaction. Neutral particles from the primary interaction will be close by while neutral particles from pileup will be evenly distributed in all directions. Neutral particles that are not close enough to be fully accepted but not far enough to be totally rejected are weighted by the calculated probability so they contribute less to the final jet.

In addition to pileup, other soft radiation originating from nearby objects will contaminate jets. This is accounted for with "jet grooming" techniques which remove this radiation. Commonly used is the "soft drop" [14] algorithm which iteratively declusters the jet into the last two components considered by the anti-k_{T} algorithm and checks if the component with lower p_{T} is characteristic of soft, wide-angle radiation (i.e. the relative p_{T} is less than the relative angular separation between the two components). If it is, drop it and repeat by

declustering the higher p_{T} component again. If it is not, call the combination of the two components a "soft drop jet".

The combination of the PUPPI and soft drop algorithms naturally reduces the energy (and masses) of the initial anti-k_{T} jets. However, they also improve the resolution on the jet mass when jets originate from heavy particles such as W bosons, Z bosons, H bosons, or top quarks while reducing the magnitude of the mass for jets originating from QCD processes. While this improves the jet mass as a discriminant between signal and jets originating from QCD, it also leaves behind a subset of QCD jets which are not well modeled by simulation, as will be discussed in Ch. 3.

While the substructure of jets can be used to more accurately reconstruct the jets themselves, it can also be used to identify jets that originate from heavier particles. Considering again the collimating phenomena shown in Fig. 2.7, if a heavy particle like a top quark (172 GeV) decays to a W boson (80 GeV) and a bottom quark (4 GeV), the W and bottom will merge into a single jet but their own decay products will *also* collimate. This creates distinguishable "cores" of energy inside the jet which can be identified as "subjets". This scenario is only considered for fat jets which have a radius large enough to encapsulate such substructure. On the other hand, a fat jet originating from a QCD process will not have a heavy particle among the first decay products and will thus, on average, not have multiple high energy cores.

There are various ways to leverage this substructure to discriminate QCD jets from heavy particle jets. One simple technique is to use the "N-subjetiness",

τ_N, of a jet which measures how consistent a jet is with having N cores and is defined as

$$\tau_N = \frac{1}{\sum_k p_{T,k} R_0} \sum_k p_{T,k} \min[\Delta R_{1,k}, \Delta R_{2,k}, ... \Delta R_{N,k}] \tag{2.4}$$

where $p_{T,k}$ is the transverse momentum of particle k and $\Delta R_{J,k} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is the distance between particle k and subjet J. This means that particles closer to the axis of the subjet will contribute less to the summation. Additionally, a higher momentum particle at the same distance from a subjet as a low momentum particle, will contribute more to the summation. This means that *lower* τ_N indicates the jet is consistent with having N cores.

By the nature of $\Delta R_{J,k}$, the value of τ_N will decrease as N increases. Therefore, jet substructure is quantified by comparing two N-subjetiness values against each other. For example, to check if a jet is more consistent with having three subjets than two subjets, one would use τ_3/τ_2. This ratio will be between 0 and 1, with values closer to 0 indicating that the jet is more consistent with having three cores.

Additionally, the long lifetime of bottom quarks means that they travel detectable distances inside the tracker. Even inside of a jet, secondary vertex reconstruction can identify a displaced vertex and associate it with the presence of a bottom quark. In this way, subjets can be b tagged, thus providing another discriminant which is particularly useful for identifying top quarks and Higgs bosons. When these substructure quantities (or a combination of them) are used to identify a jet as originating from a particular particle, the jet is considered "tagged".

Finally, while they will not be covered any further in this dissertation, there exist new taggers which use neural networks to identify jets as originating from non-QCD processes. These networks typically take the so-called "particle-flow candidates" as input to classify jets as "signal"-like or QCD-like, where "signal" could be one of W, Z, H, top, bottom, or some BSM particle.

These taggers are incredibly efficient but come with their own set of disadvantages. They are highly efficient at selecting signal and also pure since they remove almost all backgrounds not associated with the particle of the tagged jet. Any remaining contribution that may still have a large *relative* contribution to the total background amount is difficult to model since the simulation of that background will be mostly rejected by the tagger. Any data-driven method will also have difficulty since the search can be near background-free! Additionally, the neural networks indirectly learn the importance of jet mass which will bias the jet mass distributions of backgrounds passing the tagger to look more like signal, making it harder to differentiate the background and signal if the jet mass is a measurement variable for the analysis. There are methods which have successfully decorrelated the jet mass from the neural network classification but they do not work without a performance penalty. Finally, the performances in data and simulation tend to be very different, especially when selecting for the jets that are "most" signal like (i.e. very high classification scores). The corresponding corrections to simulation to make it more consistent with data can be significant, with large uncertainties that are dominated by the statistical uncertainties from the low event yields associated with this selection. So while they are undeniably better than their

predecessors, they are not infallible.

Of course, simpler identification tools like N-subjetiness also perform differently in simulation and data. Given that τ_N is dependent on the distances of each particle k to their N subjets, a simple mismodeling of the positions of particles in the jet hadronization can bias the τ_{N+1}/τ_N distribution. A similar argument can be made for the subjet b tagging. There are therefore corrections to improve the agreement in the tagging efficiencies between simulation in data. For N-subjetiness and b tagging, the ratio of efficiencies is close to 1 which makes the corresponding impact of the "scale factors" small - at least relative to some scale factors for the neural network taggers which can be as low as 0.5.

These scale factors are typically derived as a function of p_{T} and sometimes η. They are measured by dedicated groups within CMS which use data enriched in events with the type of jets being measured. Of course, one cannot select the enriched data by using the tagger itself. To overcome this, the measurement is typically performed using a "tag-and-probe" method (or similar) in a sample that is known to contain the jets of interest. One hemisphere of the event is designed to contain an object (or objects) that do not rely on the jet tagger but can nonetheless be identified to select on the data (the tag) and the other hemisphere contains the jet to evaluate the tagger on (the probe).

As an example for top jet tagging, the data is selected to be consistent with $\mathrm{t\bar{t}}$ events. Most SM events containing at least one top quark are $\mathrm{t\bar{t}}$

events[5] and so by identifying a top quark on one side of the event, the sample will already be enriched with top jets in the other hemisphere. The first top can be identified without using the top jet tagger by isolating leptonically decaying tops which will be composed of an AK4 jet consistent with a bottom quark (identified with the displaced secondary vertex) and a high energy lepton (the top goes to a bottom and a W and the W decays to a lepton and a neutrino)[6]. The leptonic decay of the top is a very "clean" channel since there's no hadronic activity beyond the bottom quark jet. Additionally, the high p_{T} lepton (electron or muon) can easily be identified by the tracker and/or muon chambers. Again, we use the rules to our advantage!

With a relatively pure sample of top quarks, the efficiency to tag them can be calculated in data. Of course, in simulation we *know* what created the jet and we can look at this generator "truth" to measure the tagging efficiency in the simulation.

As is always the case, the technique is imperfect and there are associated uncertainties in the scale factor measurement. Thus, when using these scale factors in an analysis, the scale factor uncertainties must be accounted for as systematic uncertainties. Because there is a p_{T} (and possibly η) dependence, simulation shape templates are used, as described next in Ch. 3.

[5] There are single top events with much lower probability to be produced but which exist nonetheless. These are handled by the analysis teams measuring the scale factors but I ignore them here for simplicity.

[6] See? All of that background from Ch. 1 is coming in handy!

Chapter 3

2D Alphabet background estimation

The most central aspect of any data analysis is the construction of the model used to fit the data. As with any hypothesis testing, the model must be able to describe a null hypothesis and at least one alternative hypothesis which, when individually fit to data, can be compared to determine if the results of the alternative model are significantly different from those of the null. In the case of a search for BSM physics, the null hypothesis is the Standard Mode, or the so-called "background only" or "b-only" hypothesis while the alternative hypothesis is the "signal plus background" or "s+b" hypothesis.

Technically, these are one and the same if one formulates the model as $r \cdot s + b$ where b is the estimate for the amount of background (typically as a function of some nuisance parameters), s is the theoretical signal that has been proposed (usually normalized to the cross-section proposed by the theory), and r is the "signal strength" which, when set to zero, returns the b-only model and when set to one, returns the theory proposal. When not frozen to a specific value, r is treated as an unconstrained free parameter in the model and is

measured in the fit to data. Thus, it could be measured as any value above zero.

The "fit" to data is done by constructing a binned negative log likelihood (NLL) which is then minimized as a function of the parameters of the full model. A Poisson likelihood is used since the data are counts. The generic form of the NLL is then

$$-\ln L(\vec{d};\vec{\theta}) = \sum_{i=1}^{N_{\text{bins}}} \left[n_i(\vec{\theta}) - d_i \ln n_i(\vec{\theta}) \right] \tag{3.1}$$

where d is the measurement of the data and $\vec{\theta}$ is the set of all parameters for model n. Since the values for d are all determined by observation, the responsibility of the analyzer is to just construct a model $n(\vec{\theta})$ which can describe the data well, given external measurements made by other analyses. This chapter presents one such model for heavy BSM resonances decaying to jets.

A "resonance" simply refers to a process that produces a peaked distribution, typically in mass. Searches for resonances caused by BSM physics typically involve three different types of processes which must be modeled: non-resonant backgrounds, resonant backgrounds, and the BSM signal. The non-resonant backgrounds are "smooth" and usually exponentially falling while the resonant background is a background that peaks at a specific value. For example, $t\bar{t}$ processes produce a resonance in the reconstructed jet mass at about 172 GeV.

Depending on the selection on data and which variable(s) is used to make

the measurement, these processes can have different shapes. The ability to constrain the contribution of signal in these analyses is directly dependent on making the signal shape sufficiently different from the background shapes. If the signal is a peak on a smooth background, then the search is referred to as a "bump hunt". A bump hunt is a powerful search method since the background and signal shapes are different. If the shapes of the signal and background are too similar, the parameters controlling each will be degenerate and thus correlated, making it difficult to measure the uncertainties in the individual parameters (specifically the parameter of interest, the signal strength). Similarly, if one hopes to constrain the relative contributions of different backgrounds, the background shapes themselves should be different from each other.

The mock distributions of Fig. 3.1 show examples of two unnamed variables which exhibit different shape characteristic; one where the non-resonant and resonant background shapes are very similar and another where they are very different. Since the signal resonates in both observables, these axes could be combined into one two-dimensional space such that the signal is still a peak in the two-dimensional space, the resonant background is a falling ridge, and the non-resonant is falling in both directions. The discriminating power of these distributions will be central to the methodology of this thesis.

3.1 Motivation

Given the progress that has been made over the past decade on hadronic jet identification, more high energy physics experiments are using jets in their

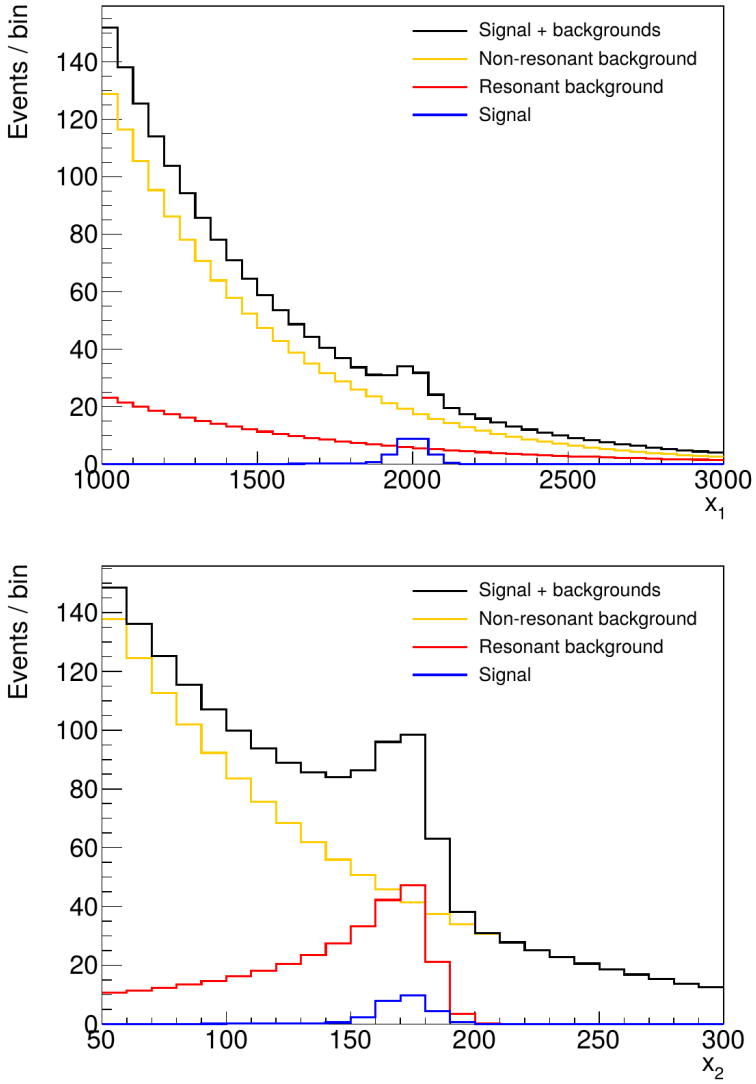

Figure 3.1: Distributions for an example bump hunt. The upper plot shows the distribution of non-resonant (yellow) and resonant (red) backgrounds and the signal (blue) along variable x_1, where both backgrounds are smoothly falling. The total of all three is shown in black. The lower plot shows the same but along a different variable x_2 where the resonant background peaks at the same location as the signal. In the upper plot, the two background contributions are not distinguishable when just examining the black line. In the lower plot, the signal and resonant background are not distinguishable from the shape of the black line. Combining the axes for a two-dimensional plot would make all three components distinct from one another.

analyses. The use of jets leads to a dramatic increases in the acceptance of multijet backgrounds from QCD processes. Removing these backgrounds requires identification of substructure that would be inconsistent with the soft radiation of most QCD jets. Jet mass grooming techniques are also used to remove soft radiation and pile up, thus removing further QCD contributions.

While most simulations of resonant backgrounds are considered acceptable for analysis use (within uncertainties which are well measured), simulations of QCD multijet processes are often not sufficient; especially when QCD background is the greater than about 25% of the total SM background in the distributions being measured, as is the case for any analysis using more than one jet with no obvious way to veto QCD events that are signal-like.

QCD events are simulated in bins of H_{T} to overcome the fundamental fact that higher energy events are less probable to produce. By isolating the simulation to ranges of energy, equal amounts of events can be generated in those ranges so that analyses concerned with high energy events are not left with only a small portion of relevant simulation. However, most jet-based searches require more than just a selection on event energy. They also typically select on jet attributes such as the groomed mass and jet substructure, including subjet bottom quark identification. While these advancements in jet reconstruction are effective at removing backgrounds originating from QCD, the combinatoric nature of QCD means that events still remain and those events exist in a subspace of physics that is not well modeled by generic multijet simulation. This simulation is thus inadequate for describing these remaining events, necessitating another method for modeling these background

contributions.

Conceivably, one could use the QCD simulation and just account for differences with the data using scale factors and for the statistical uncertainties in simulation using more model parameters. However, more often than not, one can correlate the shape of the multijet distribution in a region enriched in multijet background (called a "control region") to the distribution in the multijet-depleted region of interest (typically the one enriched in signal, the "signal region"). Even if the selection on the jet attributes changes the shape of the multijet distribution, it should remain "smooth" and non-resonant. Since the shape of the distribution is smooth by design (since it is enriched in multijet events), then the ratio of the two is also smooth and can be described with a parametric function, called the "transfer function". Multiplying the control region distribution by the transfer function "transfers" the information in the control region distribution in order to the signal region distribution.

3.2 A brief history of multijet background estimation

As one might expect, 2D Alphabet is a two dimensional version of the Alphabet method which itself is named after the ABCD method. All three are data-driven background estimation methods that use a transfer function to estimate the background contribution in a signal region based on the shape of the contribution in the control region. Working from relatively simple principles, we know that QCD multijet background will not resonate at a specific mass. The selection of QCD jets with substructure and a non-zero

groomed mass will naturally sculpt the jet mass distributions if unaccounted for but, even if sculpted, the resulting shape of the jet mass distribution will, in most cases, not contain a "peak" characteristic of a true resonance. On the other hand, the invariant mass of a system of jets will have relatively little mass sculpting. Additionally, since the probability for high energy collisions falls off as the momentum fraction of the parton's increases (from Sec 2.1), the invariant mass will also behave like a falling distribution.

Thus, the intricacies of simulating the exact physics are not necessary to estimate multijet background contributions in a given binned distribution. Instead, a binned shape analysis only requires a model that can accurately model these mass distributions. Of course, selecting jets with substructure and a heavy groomed mass increases the relative contribution of background processes such as $t\bar{t}$, W+jets, and Z+jets which can naturally contain jets with these characteristics. The simulation of these processes though is accurate enough that MC simulation can be used in conjunction with systematic uncertainties associated with known discrepancies between data and simulation, as will be discussed in Sec 3.4.

If the shapes of the background distributions in the control region and signal region are identical, then the transfer function is just a constant factor which only changes the normalization from one region to the other. However, having a "flat" transfer function is unusual as there is typically a shape dependence along the axis considered in the distribution. The ABCD method measures data distributions along an axis in selection regions A, B, and D (see Fig 3.2) which are enriched in background and depleted of signal. In

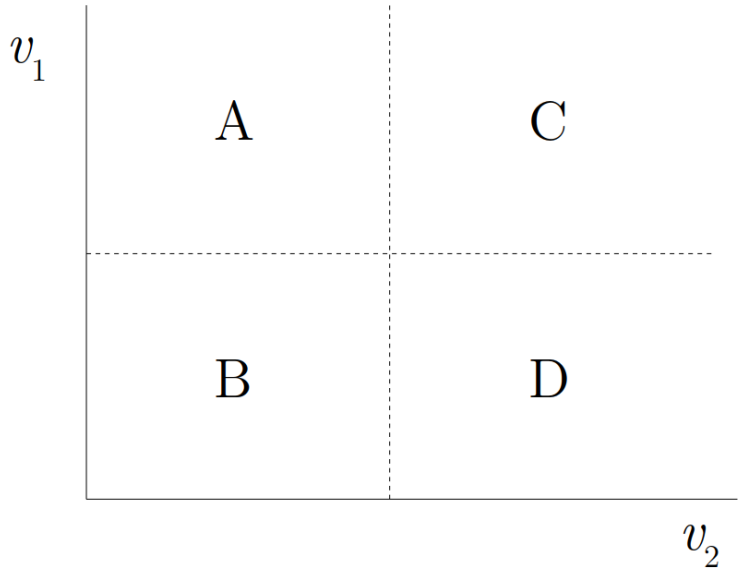

Figure 3.2: A graphical representation of the ABCD background estimation regions. The two axes, v_1 and v_2, are discriminating variables in the hypothetical analysis and C is considered the only signal enriched region.

the figure, v_1 and v_2 are the selection variables and the C region is the signal region of the analysis. Binned distributions are created along the measurement variable, x_1, for each of the four regions (not pictured). The ratio of the multijet components in regions A and B, A/B, and the same for C and D, C/D, are assumed to be equal and therefore, $A/B * D = C$ where A/B is the transfer function and C is the estimate of the multijet contribution in the signal region. Technically speaking, the binned events in region D are weighted by the binned (or possibly analytic) A/B as a function of x_1 to get the estimate of the multijet background along the axis x_1.

While this method has proven successful in the past, it has the disadvantage that it extrapolates the shape of the background to region C by assuming that A/B and C/D are equal. If a different v_2 can be chosen such that the

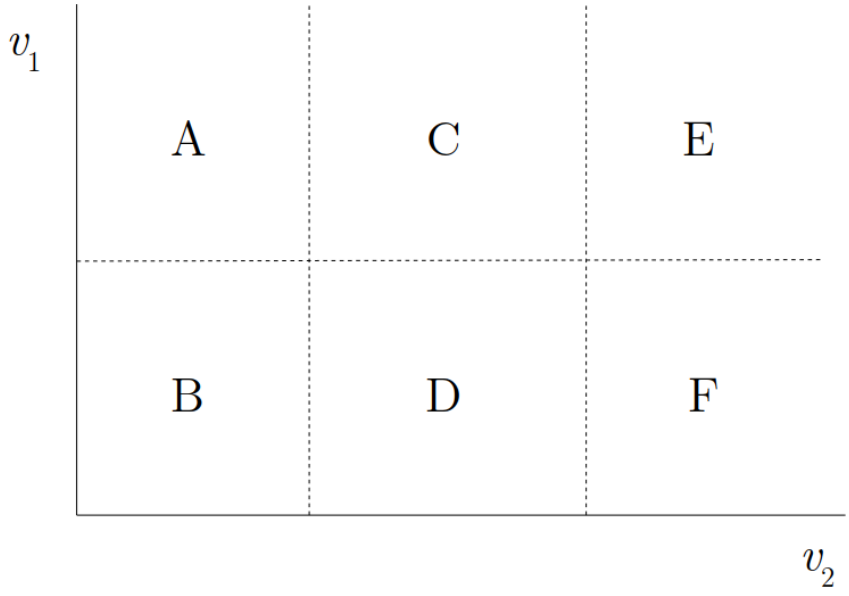

Figure 3.3: A graphical representation of the Alphabet background estimation regions. The two axes, v_1 and v_2, are discriminating variables in the hypothetical analysis and C is considered the only signal enriched region.

signal lies in the middle of the axis (still region *C* of Fig 3.3), one can instead interpolate the background using the "sidebands" of v_2, which is often more robust than extrapolation. It would be inconvenient to call this the ABCDEF method so it is instead referred to as "Alphabet" since it uses so many letters.

Of course, now one cannot use the simple $A/B * D = C$ equality. Instead, the v_2 axis is itself binned and two distributions along v_2 are created - one for each the upper and bottom segments of v_1, as shown in Fig. 3.3. The ratio of the upper and lower histograms is calculated to create a ratio-per-bin along v_2. Now these values can be fit with a function that will interpolate through the middle signal region, shown in Fig. 3.4 where the middle region is empty since the signal region, *C*, must remain blinded to avoid bias. The value of the transfer function, $f(v_2)$, in the middle of the v_2 axis can then be used to

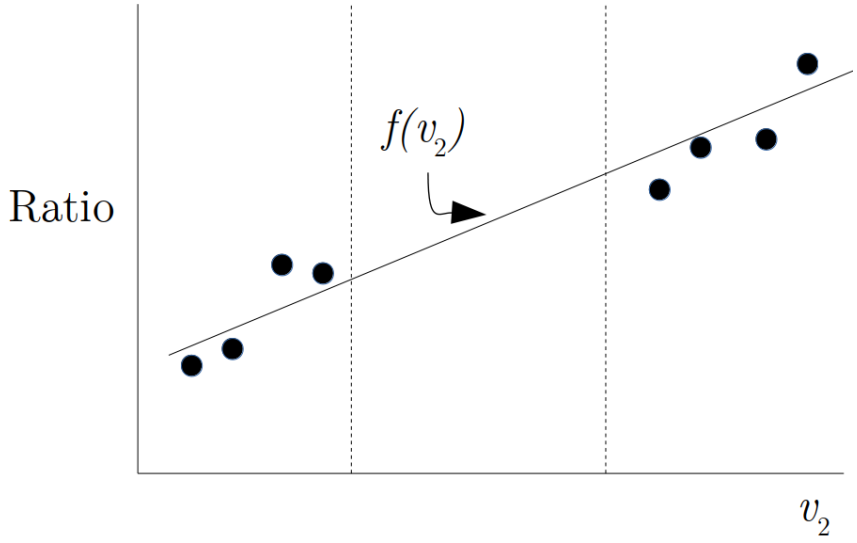

Figure 3.4: A cartoon of the ratio of the upper and lower segments of the v_1 axis from 3.3. The black dots represent the ratio calculated per-bin in v_2 which are fitted with the function $f(v_2)$ represented with the solid line. The vertical dashed lines denote the boundaries of the signal region along the v_2 axis.

calculate the distribution in C with $f(v_2) * D = C$.

Note that v_2 is not our measurement variable, x_1. It is simply the variable in which we measure the transfer function. The technical aspect of creating the QCD estimate in the C region is the same as was done for the ABCD method but A/B (which was a function of x_1) is replaced by $f(v_2)$, which is independent of x_1.

We make the jump to 2D Alphabet by asking the interpolation method of Alphabet to measure f in x_1 as well to get $f(v_2, x_1)$. Then we get the best of both worlds. In fact, v_2 can then be classified as a measurement variable and we can write more cleanly, $f(x_1, x_2, \vec{p})$ where $\vec{p}$ are the parameters of the parametric function which have been ignored to this point for convenience. They will be a part of the list of all likelihood parameters, $\vec{\theta}$.

Thus, while 2D Alphabet provides a framework to model the total background plus signal model, its name is derived from its data-driven background estimate of combinatorial backgrounds that are otherwise poorly modeled by simulation. In many cases, the background being modeled is QCD multijet production. However, depending on the selection, there may be other backgrounds accounted for as well such as W+jets or Z+jets (where the "jets" are still from QCD processes).

There are several advantages to the 2D Alphabet method. First, 2D Alphabet has the interpolation power of Alphabet that the ABCD method misses. Second, there is typically a shape dependence of the transfer function in x_1 that vanilla Alphabet ignores. For 2D Alphabet, the entire 2D space of x_1 vs x_2 can be used to constrain the transfer function. Third, if there are any backgrounds resonant in x_2 (similar to the distributions in Fig. 3.1), we can better constrain them by fitting the distinct peak in x_2. Finally, we get a complete model of the background in the dimensions we care about without needing to play games of weighting binned events along x_1 by the event's value in x_2.

The last point is perhaps the most powerful of the five because it allows us to build a model of the background without measuring $f(x_1, x_2, \vec{p})$ before fitting to data! Instead, it describes the binned PDF completely. We can then add it into the total background model which already consists of backgrounds based on MC simulation that can morph shape based on systematic uncertainties. Then $f(x_1, x_2, \vec{p})$ can be measured simultaneous to the parameters determining the other backgrounds in the likelihood fit to data (that is also performing the signal extraction). This has nice benefits like the fact that

all correlations between the parameters in $\vec{p}$ and the other nuisances will be accounted for by construction.

Another degree of freedom in this method is in the definition of the control region distribution. One could describe it using the distribution of simulated events in the control region. However, one can also assume that, since this control region is depleted of signal, the difference between the data and the MC simulated backgrounds is a robust estimate of the remaining background-only events that one would like to describe. In fact, it is guaranteed to be the most accurate representation of the non-resonant background because it is, by definition, whatever remains after the resonant components are removed. This leaves open the question of what happens when the resonant background simulation is wrong but that is addressed differently depending on the specifics of the full model being built (examples of which will be presented in in Sec. 3.4).

3.3 More complex transfer functions

The two-dimensional transfer function, $f(x_1, x_2, \vec{p})$, does not have to be just the ratio between the upper and lower regions of x_1 as described above. This method is commonly referred to as a "pass-fail ratio" and defines the QCD "pass" yield for a bin i as

$$n_P^{QCD}(i) = n_F^{QCD}(i) \cdot f(x_1(i), x_2(i), \vec{p}) \tag{3.2}$$

where $n_F^{QCD}(i)$ is the value of the QCD prediction in bin i of the "fail" control region and $x_1(i)$ and $x_2(i)$ are the values at the center of bin i along the

respective axes.

More terms can be added to the transfer function as long as they do not exist in a dimension outside of (x_1, x_2). One could, for example, have a purely one-dimensional transfer function that is independent of either x_1 or x_2. One could also develop a term that uses the QCD simulation which, while not perfect, can still help to describe the transfer function. As will be shown in Ch. 4, this can be done relatively simply by defining the transfer function as

$$f(x_1(i), x_2(i), \vec{p}) = g(x_1(i), x_2(i), \vec{p}) \cdot \frac{n_P^{MC}(i)}{n_F^{MC}(i)} \tag{3.3}$$

where g is some function of x_1 and x_2 with coefficients $\vec{p}$ (as defined previously for f) and $n_P^{MC}(i)$ and $n_F^{MC}(i)$ are the "pass" and "fail" distributions of the QCD MC in bin i, respectively. Since f is still the "pass-fail ratio" then g is the so-called "ratio of ratios".

However, actually calculating the ratio in binned MC distributions, $\frac{n_P^{MC}}{n_F^{MC}}$, is difficult. Recall that the number of simulated events in the "pass" is small enough that the relative statistical uncertainty is large. Regardless of the formulation, f is still required to be "smooth" and the natural fluctuations of $\frac{n_P^{MC}}{n_F^{MC}}$ are not conducive to this requirement.

To overcome this issue, one can "smooth" the MC distributions using a Kernel Density Estimation (KDE) method [15]. Recall that the relative Poisson uncertainty is given by $\frac{\sqrt{N}}{N}$ where N is the bin yield. The KDE algorithm smooths the distributions by treating each bin as a Gaussian kernel centered on the bin and with a width and amplitude defined as a function of N. When N is large, the kernel is narrow since the relative uncertainty is small. When

N is near zero, the kernel is wide, effectively distributing the yield across multiple neighboring bins. The final distribution is the sum of all kernels which is a smooth distribution.

With n_P^{MC} and n_F^{MC} smoothed by the KDE algorithm, $\frac{n_P^{MC}}{n_F^{MC}}$ is also smooth and then suitable to use in Eq. 3.3.

3.4 Simulation templates

As stated earlier, the full 2D alphabet model describes more than just the multijet background from QCD processes. In fact, if other backgrounds exist in an analysis' search, including them as part of the total model is beneficial for correlating the changes in all backgrounds. Then the transfer function can be measured simultaneous to the other background contributions as well as the signal.

These other backgrounds are typically not derived from data (though that does not have to be true) but are instead derived from simulation. As was discussed in Sec. 2.3, simulation does not agree exactly with data. This is true for jet reconstruction algorithms but it is also true for many other effects that must be simulated or corrected in the simulation. Examples that are unrelated to jets include trigger efficiencies, pile up simulation, and detector effects that were not understood at the time of simulation and therefore not simulated correctly.

Dedicated studies are performed to evaluate the impact of these systematic effects and they conclude whether a numeric correction must be made to the simulation to improve the agreement with data. Correction or not though, the

studies also evaluate an associated systematic uncertainty which indicates that the "next" measurement of the effect will be within one standard deviation of existing measurements 68% of the time.[1]

Variations of each of these systematic effects (related to jets or otherwise) will change the simulation distributions. Some may change the distributions uniformly such that every bin of the distribution is increased or decreased by some percentage amount. This is the case, for example, in the uncertainty of the total amount of collected data, called the "luminosity". However, the distributions could also change shape by being skewed, tilted, or morphed. In other words, the distributions in the bins still all change in a correlated manner but the changes per-bin are not always at the same relative magnitude as the other bins or in the same direction.

Accounting for these shape variations in our model requires a mathematical formulation such that just one parameter, α, in the set of model parameters, $\vec{\theta}$, can act as the knob that morphs the shape between the positive and negative variations of the effect, interpolating the nominal value and extrapolating beyond the 68% uncertainty interval [16].

For the sake of generalization, α should take the value of 0 at the nominal bin value, +1 for the "up" variation, and -1 for the "down" variation.[2] With the "up" and "down" variations of the binned distributions created ahead of time (called "templates"), the values can be used to define the three points that map the α values of -1, 0, and +1 to the bin's down, nominal, and up values,

[1] Note that this is different from saying that there is a 68% probability that the true value is within the uncertainties.

[2] Note that "up" and "down" here are just a naming convention - the bin's value does not have to increase or decrease, respectively.

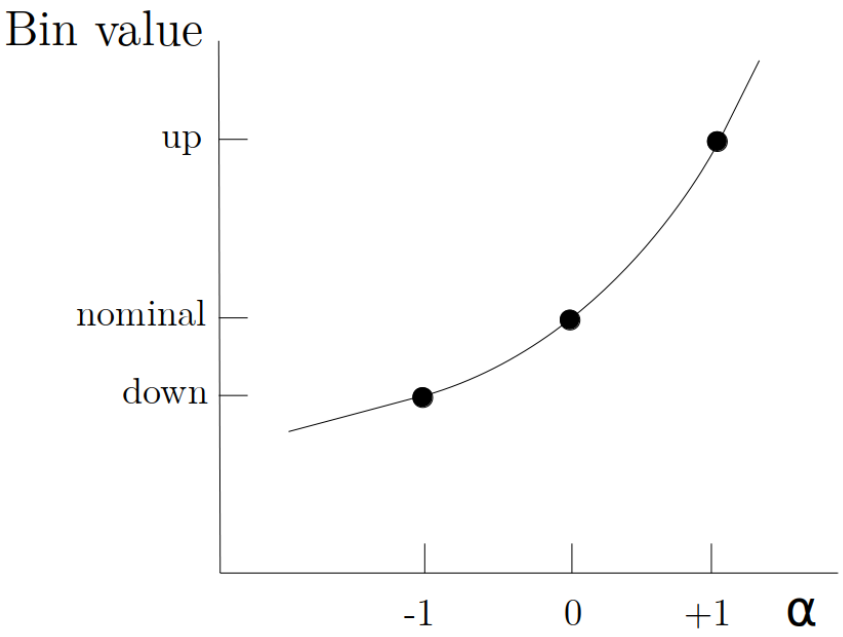

Figure 3.5: Cartoon of a piece-wise morphing function fitted to the three points in the "bin value" vs α space. The three points determine the coefficients of the mapping function completely, allowing for quadratic interpolation between α of -1 and +1 and linear extrapolation below -1 and above +1.

respectively. Figure 3.5 shows a representation of these three points in "bin value" vs α.

The functional interpolation between the up, nominal, and down values for a given bin is relatively easy to perform using a three parameter function such as a quadratic equation or an exponential. The extrapolation below -1 and above +1 could just follow the quadratic parameterization used for the interpolation. However, linear functions are the common alternative since they make fewer assumptions about how the bin value will change beyond the one standard deviation range defined by the shape templates. Conveniently, the parameters of the line are uniquely defined by the conditions that the full function be continuous and differentiable. Thus, the point-slope form of a line

can be used with the "point" being the +1 α or -1 α point and the "slope" being the slope of the quadratic at said point. Then our full piece-wise continuous mapping of α to a given bin's value is determined by just three parameters. An example function is plottes in Fig 3.5 with the points that have already been referenced. Note that other parameterizations can be used including summations of "pieces" along the α axis which are multiplied by sigmoid functions which turn "on" and "off" the pieces in a continuous way.[3]

We construct this function for each bin in our histogram and tie them all to the same α axis such that all bins are correlated via α. This is done for each of the systematic uncertainties to build the entire model for that single background. Additionally, if a systematic uncertainty is shared across multiple backgrounds or distributions, those distributions also all share the same parameter so that if the knob is turned, they all morph at the same time.

Finally, one needs to encode the knowledge of the 68% confidence into the model. This is done by defining a unit Gaussian as a function of α which will act as a penalty term in the negative log likelihood. The likelihood is multiplied by this additional term, which is maximized when α is zero. However, since the NLL is used for the minimization, the multiplicative term is effectively a penalty goes like α^2, being *minimized* at $\alpha = 0$. In other words, the farther α strays from zero, the faster the penalty increases. This gives the model a preference for keeping the shapes within their one standard deviation band.

[3]This is actually closer to what is actually used since these do not require quadratic equations to be solved for every bin. However, the details are beyond the scope of this description.

Despite the constraints, this morphing can have non-negligible effects, especially if the uncertainties are very large. This is specifically an issue if the multijet background in the control region is calculated as the data minus the nominal shapes of the MC simulated backgrounds. If the MC simulated background distribution needs to morph in the multijet control region, there must be some way to account for the corresponding change in the estimate of the multijet background, otherwise the agreement with data will be poor in that region.

To do this, the bins of the multijet background estimate can be constructed as free parameters which are able to take any value. This means that if a template-based background morphs to improve the model's agreement with data in the signal region (and thus hurting the agreement in the multijet control region), the numerical minimizer can improve the agreement in the control region by changing the value of the multijet component in the control region at no penalty to the negative log likelihood value (other than that it presumably improves the agreement with the data in the control region since the initial value is now wrong). Thus, by construction, the multijet component in the multijet control region will always change to provide good agreement with the data in that region. Keep in mind that this can have negative effects in the signal region though since the signal region multijet component is a function of the control region component (via the transfer function)!

The entire system is one complicated balancing act with all of the pieces tied together. However, one need only understand how the individual parameters tie into the model. All of the simultaneous movement is left to the

computer when it performs the numerical minimization of the NLL.

3.5 2D Alphabet as software

While "2D Alphabet" does mean the model building method described above, it is also the name of a generic software tool that has been built over the past several years which implements the model generically. It is designed generically to do any type of analysis that could use the methods or pieces of the methods described above.

The software is written and usable within Python and is designed to be the liaison between the initial distributions (from data and simulation) and the likelihood construction and minimization; all while also producing standardized plots and statistical studies that are of publication quality. The simulation templates and data can be specified via a JSON configuration file and the data-driven background model is defined by interacting with the Python API.

After constructing the input histograms and the definition for the data-driven components like the transfer function, the pieces are passed to the Higgs Analysis Combine Tool,[4] also just called "Combine", which builds the morphing infrastructure, the likelihood itself, and performs the minimization using MINUIT [17]. The output fit result is then interpreted by 2D Alphabet to produce the plots and statistical tests.

One significant addition is made to the central Combine release to make

[4]Named for its origin in the CMS Higgs analysis group and not because it is specific to Higgs-related analyses.

Combine 2D Alphabet-friendly - the `RooParametricHist2D` class. This is identical to the `RooParametricHist` class already provided by Combine but takes as input ROOT's two-dimensional histogram object, `TH2`, instead of a `TH1` [18]. The accompanying changes to accommodate this class are made in the Combine code. Additionally, code has been added and modified in Combine Harvester to plot the raw 2D distributions from the fit result. The `RooParametricHist` class is particularly important because it can store `RooRealVar` and `RooFormulaVar` objects which support the unconstrained, floating control region bins and Eq. 3.2, respectively.

The 2D Alphabet software also includes an interface to a standardized goodness of fit test. The test is performed by generating pseudo-data toys from the background-only model and evaluating the model's fit to the toys with the saturated chi-squared test statistic [19], which roughly follows a chi-squared distribution. The same saturated test statistic is evaluated in the fit to data. The value from the fit to data is then compared against the distribution in the toys. If the p-value of the data relative to the toys is greater than 0.05, the fit is considered healthy.

Signal injection tests are also included and performed by constructing the model with a frozen value of the signal strength, r_{inj}, the injected amount. Pseudo-data toys are generated from the $r_{inj} \cdot s + b$ model and fit with the model $r \cdot s + b$ (where r is freely floating again). The fit result for each toy is plotted as a function of $\frac{r - r_{inj}}{r_{Err}}$ where r_{Err} is the uncertainty on r from the fit of the toy. The fit is unbiased if the distribution of toys in this variable is a symmetric Gaussian centered at zero and with a width of one.

Finally, one can perform a Fisher F-test [20] to evaluate whether the transfer function parameterization over-fits the data. For a given number of parameters in $\vec{p}$, n_p, if a function parameterized with $n_p + 1$ parameters leads to a significantly better fit, the $n_p + 1$-th parameter is said to be justified. Pseudo-data toys are again generated but only from the simpler of the two models being compared. They are then fit separately using the n_p and $n_p + 1$ models. The saturated test statistic is calculated for both versions of the fit of the same toy and a new test statistic, proportional to the difference in the two saturated values, is calculated for each toy. The distribution of toys in this new test statistic will follow the F-distribution. The same test statistic is evaluated in data and if the p-value is greater than 0.05 relative to the F-distribution, the more complex model is justified. All three of these tests were used to validate the model used for the analysis described in Ch. 4.

In addition to the analysis in Ch. 4, two other CMS analyses on track for publication use the 2D Alphabet software to construct, fit, present, and test a 2D Alphabet model. The final results of both analyses use this modeling method and associated software. Because of the time taken to convert the ideas in the methodology to generic software, it is accessible for others to use, thus improving the physics program at CMS beyond the single analysis presented in this dissertation.

Chapter 4

Search for heavy resonance decaying tW production at $\sqrt{s}$= 13 TeV in the fully hadronic state

4.1 Introduction

The SM has been extensively verified by experiment, nonetheless there exists evidence that the SM is only an effective theory. Many possibilities for physics beyond the SM have been proposed, including the possibility that quarks are composite. Such quarks would have an internal structure that, excited, could produce a state with higher mass [21, 22]. Such a phenomenon is predicted by Randall–Sundrum models [23, 24] and models with a heavy gluon partner [25, 26, 27].

In this analysis, we search for a heavy resonance decaying to a top quark t and a W boson in the fully hadronic final state, using proton-proton (pp) collision data at a center-of-mass energy of 13 TeV. The search uses data corresponding to an integrated luminosity of 137 fb^{-1} recorded by the CMS experiment [28] at the CERN LHC during 2016–2018.

As a benchmark resonance, we consider an excited bottom quark, referred to as a b^* quark [22]. The strong interaction is the dominant production mechanism and can produce a single b^* quark at the LHC via the collision of a bottom quark and a gluon, $bg \to b^*$. The interaction is described by the Lagrangian

$$\mathcal{L}_1 = \frac{g_s}{2\Lambda} G_{\mu\nu} \bar{b} \sigma^{\mu\nu} (\kappa_L^b P_L + \kappa_R^b P_R) b^* + \text{h.c.}, \tag{4.1}$$

where g_s is the strong coupling, $G_{\mu\nu}$ is the gauge field tensor of the gluon, $\bar{b}$ is the bottom quark field, $\sigma^{\mu\nu}$ is the Pauli spin matrix, b^* is the excited bottom quark field, and Λ is the scale of compositeness [21], which is chosen to be the mass of the b^* quark. The chiral projection operators are represented as P_L and P_R, and κ_L^b and κ_R^b are the relative coupling strengths [29].

The $b^* \to tW$ decay is the dominant decay channel, with a branching fraction of approximately 40% for a b^* quark with $m_{b^*} > 1.2\,\text{TeV}$ [29]. The decay takes place through the weak interaction and is described by the Lagrangian

$$\mathcal{L}_2 = \frac{g_2}{\sqrt{2}} W_\mu^+ \bar{t} \gamma^\mu (g_L P_L + g_R P_R) b^* + \text{h.c.}, \tag{4.2}$$

where g_2 is the $SU(2)_L$ weak coupling and g_L and g_R are the relative couplings of the W boson to the left- and right-handed b^* quark, respectively [29]. The full interaction chain is then $bg \to b^* \to tW$. The b^* quark width is expected to be less than 10% of the b^* quark mass, which leads to a distinct resonant structure in the mass spectrum.

Three hypotheses for the left- and right-handed b^* quark couplings are

considered:

$$\text{left-handed (LH):} \quad \kappa_L^b = g_L = 1 \text{ and } \kappa_R^b = g_R = 0, \tag{4.3}$$

$$\text{right-handed (RH):} \quad \kappa_L^b = g_L = 0 \text{ and } \kappa_R^b = g_R = 1, \text{ and} \tag{4.4}$$

$$\text{vector-like (LH+RH):} \ \kappa_L^b = g_L = 1 \text{ and } \kappa_R^b = g_R = 1. \tag{4.5}$$

Searches for the b^* quark in the tW decay mode have been performed at the LHC by the ATLAS Collaboration at $\sqrt{s} = 7\,\text{TeV}$ [30] and by the CMS Collaboration at 8 TeV [31]. Additionally, searches for a b^* quark decaying to a bottom quark and a gluon were conducted by the CMS Collaboration at 8 TeV [32] and by the ATLAS Collaboration at 13 TeV [33]. The CMS tW decay mode search included a combination of fully hadronic, lepton+jets, and dilepton final states, and excluded b^* quark masses at 95% confidence level (CL) below 1.4, 1.4, and 1.5 TeV, for the left-handed, right-handed, and vector-like hypotheses, respectively.

Given the range of these exclusions, the present analysis considers a b^* quark with a mass greater than 1.2 TeV. For these mass values, the top quark and the W boson are commonly produced with a high Lorentz boost. Because of this, the hadronic decay products of the top quark and the W boson can each merge, resulting in two massive, large-radius jets, referred to as a "top jet" and a "W jet", respectively. These jets have a distinct substructure that is used to discriminate them from the background [34, 35]. The b^* quark mass is reconstructed as the invariant mass of the top jet and W jet system, m_{tW}. This variable, along with the reconstructed top jet mass, m_t, is used to search for

the b^* quark resonance.

The background is dominated by jets produced through the strong interaction, referred to as quantum chromodynamics (QCD) multijet production, and is estimated using multijet-enriched control regions based on inverting the top jet selection criteria. The SM W+jets and Z+jets production backgrounds are also accounted for with this technique. The $t\bar{t}$ background is estimated with simulation templates fit to data simultaneously in the signal region and a dedicated control region enhanced in $t\bar{t}$ production that constrains the simulation templates.

A binned maximum likelihood fit to data is performed in the two-dimensional m_{tW} versus m_t distribution, in a process where the signal and background models are fit simultaneously. From this fit, b^* quark mass limits are derived for the three b^* chirality hypotheses expressed in Eqs. (4.3), (4.4), and (4.5).

In addition, we interpret the results under the hypothesis of a singly produced B singlet vector-like quark [36, 37] decaying into tW. For B quark masses above 1.2 TeV, the decay products would be heavily boosted with a similar signature to the b^* quark decay described above. We consider only narrow signals with a relative width of less than 5% and with a branching fraction of approximately 50% to tW in the model described. In contrast to the b^* model, the B quark would be produced via an electroweak interaction in association with a top or bottom quark. We consider both scenarios, but typically the associated top or bottom quark has a much lower transverse momentum than the B quark decay products, thus the effect of either on the analysis is small.

4.2 The CMS detector

The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the pseudorapidity coverage provided by the barrel and endcap detectors. Muons are detected in gas-ionization chambers embedded in the steel flux-return yoke outside the solenoid. A more detailed description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, can be found in Ref. [28].

Events of interest are selected using a two-tiered trigger system [10]. The first level, composed of custom hardware processors, uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a fixed latency of about 4 μs. The second level, known as the high-level trigger, consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to around 1 kHz before data storage.

The analysis reflects the fact that the pixel detector was changed in the winter of 2016/2017. The newer detector increased the number of barrel layers from three to four and decreased the distance of the innermost layer from the beamline in order to improve the vertex reconstruction.

4.3 Data and simulated samples

CMS data taking operates on annual cycles, and thus data collection and simulation performance can change from year to year. Therefore, we categorize both the data and simulation by year and apply dedicated scale factors before combining the distributions from all three years to derive the final result.

We analyze events from the 2016 data set recorded by a trigger that requires the scalar sum of transverse momenta, p_{T}, of all jets in the event, H_{T}, to be at least 800 or 900 GeV, or the presence of a jet with $p_{\mathrm{T}} > 450$ GeV. For 2017 and 2018 data, we analyze events recorded by a trigger that requires a minimum H_{T} of 1050 GeV or the presence of a jet with $p_{\mathrm{T}} > 500$ GeV. Additionally, 2018 data events are recorded by a trigger that requires a jet with $p_{\mathrm{T}} > 400$ GeV with a mass of at least 30 GeV, where the jet trimming algorithm [38] has been used to reconstruct the jet mass at the trigger level. This trigger did not exist for the 2016 or 2017 data collection, but the addition of events recorded by this trigger provides a higher overall selection efficiency at lower H_{T} for 2018. The choice of higher H_{T} and jet p_{T} thresholds used for 2017 and 2018 are due to an increase in the instantaneous luminosity of the LHC between 2016 and 2017. The combination of these triggers is nearly fully efficient for $m_{\mathrm{tW}} > 1200$ GeV.

The efficiency of the trigger selection is measured in data as the ratio of the number of events recorded by the combined triggers to the number of events recorded by a trigger that requires a muon candidate with $p_{\mathrm{T}} > 50$ GeV. A muon trigger is used for this measurement because it is largely uncorrelated with the triggers used for data taking.

The trigger efficiencies are parameterized as a function of dijet invariant mass (m_{jj}) and both the numerator and denominator of the ratio include events that pass the preselection described in Section 4.5. The uncertainty assigned to the efficiency measurement is one half of the trigger inefficiency.

Figure 4.1 shows the trigger efficiency derived from 2016, 2017, and 2018 data separately. Simulated samples are corrected using the efficiency measurement from the corresponding data-taking year.

A trigger inefficiency referred to as "prefiring" developed during the 2016 and 2017 data taking. Over that time period, a gradual shift in the timing of triggering systems based on the ECAL in the endcap caused certain events to not be recorded. Event corrections were calculated from data and applied to the 2016 and 2017 simulations to model this inefficiency. The uncertainties in these corrections are taken as systematic uncertainties.

The SM $t\bar{t}$ and single top quark Monte Carlo (MC) simulated samples are used as templates for background estimation in the maximum likelihood fit to data. A scale factor is applied to the generated top quark p_T spectrum to correct for the differences between data and $t\bar{t}$ simulation. It is based on a dedicated measurement [39, 40], in which the ratio of the distribution of the top quark p_T measured in data to the distribution as measured in POWHEG+PYTHIA is derived. This scale factor may be described by the expression

$$w_t(p_T) = e^{c_1 0.0615 - c_2 (0.0005/\,\mathrm{GeV}) p_T}, \tag{4.6}$$

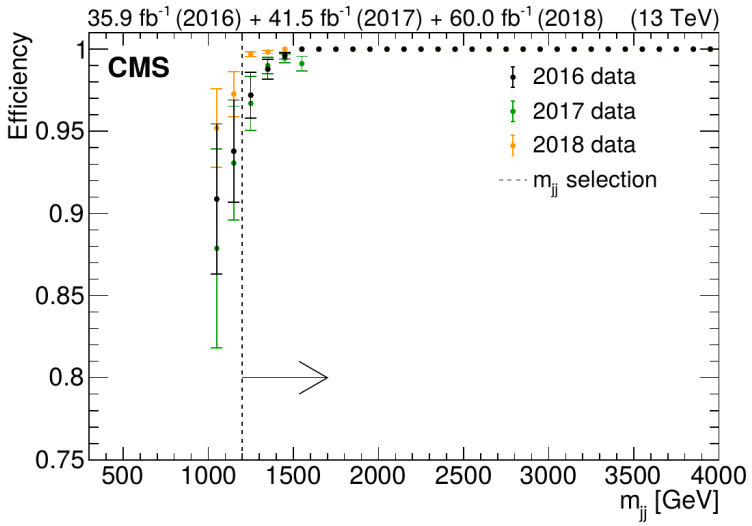

Figure 4.1: The efficiency of the full trigger selection as a function of m_{jj}, shown separately for 2016, 2017, and 2018 data. The minimum m_{jj} considered in the analysis is 1200 GeV and is marked with a dashed line and an arrow. The efficiency below m_{jj} of 1000 GeV is not measured. The points for 2017 and 2018 data are not visible in the plateau because they are overlapped by the points for 2016 data.

where c_1 and c_2 are taken to be 1, as obtained in Refs. [39, 40]. The p_T-dependent event weight is given by $\sqrt{w_t(p_T)w_{\bar{t}}(p_T)}$, where $w_t(p_T)$ and $w_{\bar{t}}(p_T)$ are evaluated using the top quark and antiquark p_T, respectively. We use the same form for the scale factor but treat c_1 and c_2 as fit parameters initialized to 1 and constrained in the fit to a Gaussian with a width of 0.5.

To simulate the SM $t\bar{t}$ and single top quark production, we use the POWHEG v2 [41, 42, 43, 44, 45] matrix element event generator. For QCD multijet simulation, we use MADGRAPH5_*a*MC@NLO version 5 [46] with subversion 2.2.2 for 2016 and 2.4.2 for 2017 and 2018. The QCD multijet simulated samples are used to derive a scale factor to the multijet background estimation procedure and for cross checks of self consistency of the background estimate. The b^* signal samples are simulated using MADGRAPH5_*a*MC@NLO version 5 over

a mass range of 1.4 to 4.0 TeV in steps of 200 GeV. Subversion 2.2.2 is used for 2016 b^* signal samples with b^* quark masses from 1.4 to 3.0 TeV and 2016 B+t and B+b signal samples. Subversion 2.4.2 is used for 2017 and 2018 b^* signal samples with b^* quark masses from 1.4 to 3.0 TeV. Subversion 2.6.5 is used for all b^* signal samples with b^* quark masses above 3.0 TeV. For the B signal simulations, we use samples based on the 2016 conditions and scale the final distributions to the luminosity of the full data set after correcting for differences in annual selection efficiencies that are measured with the b^* signal samples. We simulate B quark masses from 1.4 to 1.8 TeV in steps of 100 GeV.

Hadronization and parton showering are simulated using the PYTHIA 8 software package [47]. The NNPDF3.0 [48] parton distribution functions (PDFs) are used with the CUETP8M1 [49] underlying event tune for the 2016 simulations and the NNPDF3.1 [50] PDFs are used with the CP5 [51] underlying event tune for the 2017 and 2018 simulations. The CMS detector simulation is performed with GEANT4 [52]. Pythia version 8.212 is used for all the 2016 simulations with the exception of 2016 b^* signal samples with b^* quark masses from 1.4 to 3.0 TeV, which use version 8.226, and 2016 B signal samples, which use version 8.205. Pythia version 8.230 is used for all the 2017 and 2018 simulations.

To simulate the effect of additional pp collision data within the same or adjacent bunch crossings (pileup), additional inelastic events are superimposed using PYTHIA. Simulated samples are then reweighted to correct the pileup simulation, using the total inelastic cross section of 69 mb [53, 54] to estimate

the distribution of the number of primary vertices in data.

4.4 Event reconstruction

The candidate vertex with the largest value of summed physics-object p_{T}^2 is taken to be the primary pp interaction vertex. The physics objects are the jets, clustered using the anti-k_{T} jet finding algorithm [11, 55] with a radius parameter of $R = 0.4$ and with the tracks assigned to candidate vertices as inputs, and the associated missing transverse momentum, taken as the negative vector sum of the transverse momentum of those jets.

A particle-flow algorithm [9] aims to reconstruct and identify each individual particle in an event, with an optimized combination of information from the various elements of the CMS detector. The energy of muons is obtained from the curvature of the corresponding track. The energy of charged hadrons is determined from a combination of their momentum measured in the tracking detector and the matching ECAL and HCAL energy deposits, corrected for the response function of the calorimeters to hadronic showers. Finally, the energy of neutral hadrons is obtained from the corresponding corrected ECAL and HCAL energies. Jets are clustered with the anti-k_{T} jet finding algorithm, using all particle-flow objects as input. Jet momentum is determined as the vectorial sum of all particle momenta in the jet. Jets with a radius parameter of $R = 0.8$ are used to reconstruct the top jet and W jet candidates in an event.

The pileup per particle identification (PUPPI) algorithm [12] is used to mitigate the effect of pileup at the reconstructed particle level, making use of local shape information, event pileup properties, and tracking information.

A local shape variable is defined, which distinguishes between collinear and soft diffuse distributions of other particles surrounding the particle under consideration. The former is attributed to particles originating from the hard scatter and the latter to particles originating from pileup interactions. Charged particles originating from pileup vertices are discarded. For each neutral particle, a local shape variable is computed using the surrounding charged particles compatible with the primary vertex within the tracking detector acceptance ($|\eta| < 2.5$), and using both charged and neutral particles in the region outside of the tracking detector coverage. The momenta of the neutral particles are then rescaled according to the probability that they originate from the primary interaction vertex as deduced from the local shape variable, superseding the need for jet-based pileup corrections [56].

Jet energy corrections are derived from simulation studies so that the average measured response of jets becomes identical to that of the jets from the reconstructed particle level. In situ measurements of the momentum balance in dijet, photon+jet, Z+jet, and multijet events are used to determine any residual differences between the jet energy scale in data and in simulation, and appropriate corrections are made [57]. Additional selection criteria are applied to each jet to remove jets potentially dominated by instrumental effects or reconstruction failures [58].

4.4.1 Top quark identification

The soft-drop algorithm [14], a generalization of the modified mass drop algorithm [59, 60], with angular exponent $\beta = 0$ and soft threshold $z = 0.1$,

is applied to all jets in the event to reconstruct the jet mass and to identify subjets, and includes a grooming step to remove soft radiation, including pileup. We only consider top jets with a minimum soft-drop mass of 65 GeV.

The N-subjettiness algorithm [61] defines τ_N variables, which describe the consistency between the jet energy deposits and the number of assumed subjets, N. When compared to jets originating from a gluon or a light quark, a top jet is more consistent with three hard decay products, and the ratio of τ_3 and τ_2 allows top jets to be distinguished from QCD multijet background [62]. A lower ratio indicates the jet is more consistent with a three-pronged structure than a two-pronged structure.

The b^* signal region selection requires $\tau_3/\tau_2 < 0.65$. The N-subjettiness ratios are correlated with the jet mass, so a relatively loose selection for the signal region is used to avoid biasing the mass distribution of multijet processes.

We also require the top jet to contain a subjet from the soft-drop algorithm to be identified as a bottom quark by the DeepCSV algorithm [63]. The combination of the τ_3/τ_2 and DeepCSV selections has a QCD jet misidentification rate of approximately 1% and a top tag signal efficiency of approximately 45% [58, 64]. This selection has been chosen because it leads to an optimal sensitivity of the cross section limits.

A jet that passes both the τ_3/τ_2 and DeepCSV b tagging selection is considered "top tagged". A p_{T}-dependent correction is applied to correct for differences in the top tagging efficiency between data and simulation [64]. Separate corrections are used based upon the merging of the top quark decay products in simulation. Taking the line defined by the top quark's trajectory

as the central axis, three scenarios are considered. In the first, the three decay products are within $R < 0.8$ of the central axis and the jet is considered "merged". In the second, two out of three decay products are within $R < 0.8$ of the central axis and the jet is "semi-merged". Finally, with any other configuration of the three decay products the jet is "not merged". The merged component is the dominant contribution for the b^* signal process among these three scenarios.

4.4.2 W boson identification

Similar to top tagging, the W boson identification algorithm requires a selection based on τ_N and soft-drop mass. The W jet is required to have a soft-drop mass between 65 and 105 GeV to be consistent with the W boson mass [65]. The ratio of N-subjettiness τ_2 and τ_1 variables is used to select the characteristic two-prong structure of a hadronic W boson decay since the W jet is more consistent with having two subjets than one. The b^* signal region selection requires $\tau_2/\tau_1 < 0.4$ for 2016 data and simulation, and $\tau_2/\tau_1 < 0.45$ for 2017 and 2018 data and simulation. The combined selection on the mass and τ_2/τ_1 has a QCD jet misidentification rate of approximately 10% and a W tag signal efficiency of approximately 80%, which are consistent across the three years [58, 64]. This selection was chosen because it leads to an optimal sensitivity of the cross section limits.

A jet that passes the τ_2/τ_1 and soft-drop mass selections is considered "W tagged". Differences in the W tagging efficiency between data and simulation are corrected using simulation-to-data weights [64]. Additionally, differences

in the soft-drop mass scale and resolution between data and simulation are accounted for by scaling and smearing the soft-drop mass in simulation [58].

4.5 Event selection

To select signal-like events, two jets are required with $p_T > 400\,\text{GeV}$ and $|\eta| < 2.4$. Only the two jets with the highest p_T are considered in the following. The jets are required to satisfy that the difference in rapidity, $|\Delta y|$, be less than 1.6 and that $|\Delta\phi|$ be greater than $\pi/2$. The $|\Delta\phi|$ requirement selects back-to-back dijet events while the $|\Delta y|$ requirement suppresses multijet events with high m_{tW}, which arise from the scattering of valence quarks. These requirements comprise the "preselection", with an event then being selected as signal if one of the two jets is W tagged and the other is top tagged.

Because the background estimate relies on data in a control region defined by inverting the top tag selection, we first require that one of the two jets can be identified as a W jet. In the case that both jets can be W tagged, the jet with lower p_T is taken as the W boson candidate in the event. If neither jet can be W tagged, the event is not selected. The jet that is taken as the W boson candidate is referred to as the initially tagged or first jet and the other jet is called the remaining or second jet. If the event is selected, it is categorized in either the signal region or the multijet control region depending on whether the second jet passes the top tagging requirement. The final selection efficiency for simulated events is calculated as the number of events that pass the signal selection divided by the number of events generated. Over the range of generated b^* quark masses, signals with left-handed couplings

Table 4.1: A summary of the four selection regions considered in the likelihood fit to data. The columns indicate the possible jet tag for the jet considered in the preselection while the rows indicate the possible classification of the second jet when using the top tagging algorithm.

	W tag	Top tag
Top tag	Signal region (SR)	$t\bar{t}$ measurement region (MR)
Anti-top tag	Multijet control region (for SR)	Multijet control region (for MR)

are selected with an efficiency of 9–10%. Signals with right-handed couplings have slightly higher efficiencies, ranging from 10–11%, because of their harder jet p_T spectra.

We additionally define a dedicated $t\bar{t}$ measurement region. For this, events are required to pass the preselection but the W tag is changed to a top tag selection for the initial jet tag. This jet tag also requires a top jet mass value between 105 and 220 GeV to be consistent with the top quark mass [65]. The second top tag will be used to distribute events between the $t\bar{t}$ measurement region selection and the dedicated multijet control region for the $t\bar{t}$ measurement region. Additionally, for the initial jet tag, the subjet bottom quark requirement remains the same but a tighter selection of $\tau_3/\tau_2 < 0.54$ is required. The tighter selection on the initial jet tag increases the relative $t\bar{t}$ contribution. The τ_3/τ_2 selection on the second jet tag remains the same as for the b^* signal region selection to avoid distorting the mass distribution because of the correlation described in Section 4.4.1. If both jets fulfill the selection of the initial top tag, the jet with the lower p_T is takes as the initially top tagged jet in the event. The $t\bar{t}$ background measurement region is described in more detail in Section 4.6.2. The four tagging selection regions are summarized in Table 4.1.

Comparisons of the N-subjettiness ratio, soft-drop mass, and DeepCSV

algorithm score in simulation between signal and background events are shown in Fig. 4.2. The QCD contribution to the multijet background is shown, but the W+jets and Z+jets contributions are omitted since simulations of these processes are not used in this analysis (see Sec. 4.6.1).

4.6 Statistical model and background estimation

The background for this analysis is comprised of multijet, $t\bar{t}$, and tW-channel single top production. The multijet component is estimated from data while the $t\bar{t}$ and single top components are obtained by fitting simulation templates to data.

The m_t range considered is larger than the signal mass window of 105 to 220 GeV defined in Section 4.5. As shown in Fig. 4.2, an m_t selection is not efficient at discriminating signal from $t\bar{t}$ background. However, by using m_t as one of the two measurement dimensions, one can constrain the multijet background in the m_t sidebands while distinguishing the multijet background from the top backgrounds in the m_t signal region. Thus, the m_t range comprises both the signal peak region and the lower and upper sidebands of the peak. The signal region considers the range of 65 to 285 GeV while the $t\bar{t}$ measurement region exists between 105 and 285 GeV, where the lower mass bound of 105 GeV is used to ensure orthogonality with the W jet mass window of the signal region.

For each bin in the two-dimensional (m_t, m_{tW}) distribution, we compare the number of expected events from both the background-only and signal-plus-background hypotheses with the number of observed events in data.

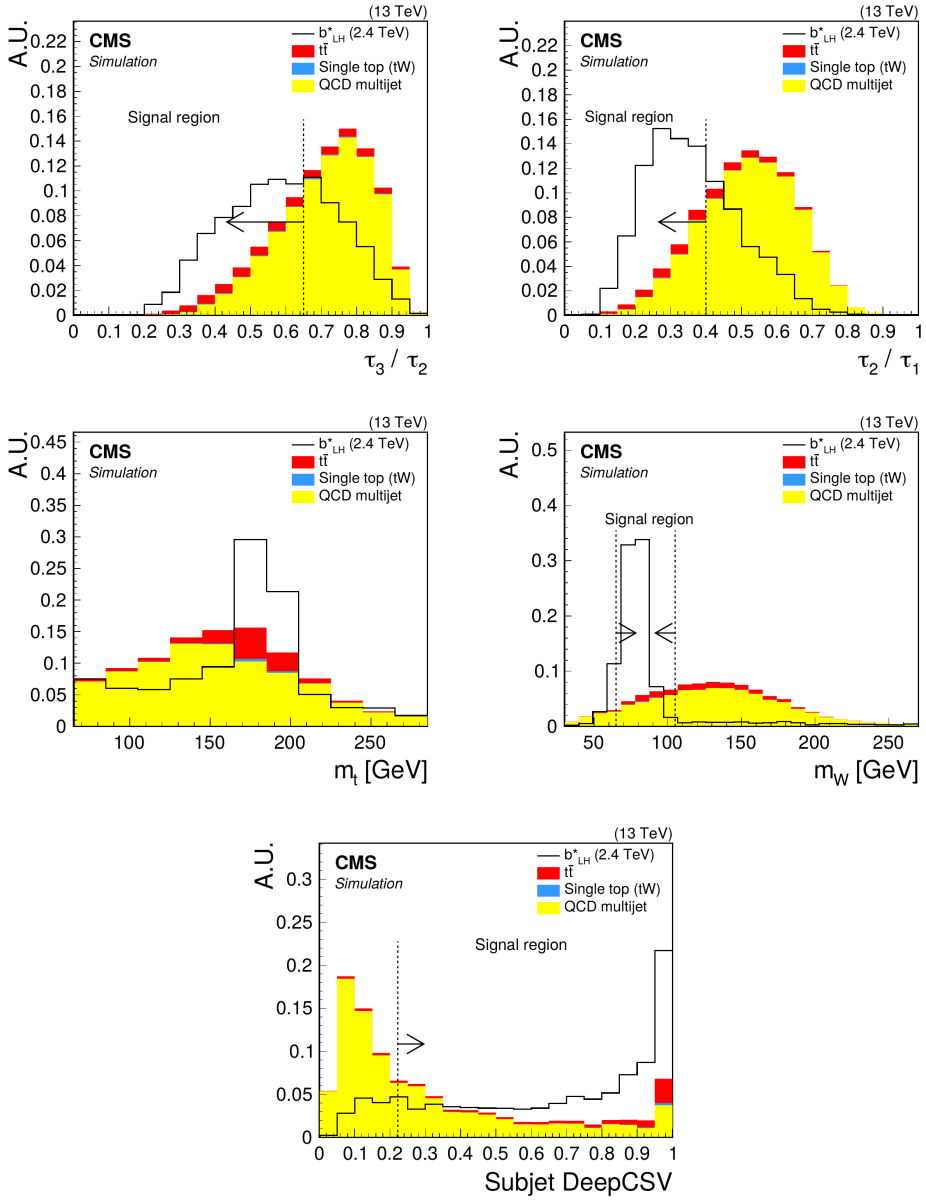

Figure 4.2: The distributions of the discrimination variables used for W and top tagging for simulation samples. These plots show the top jet τ_3/τ_2 (upper left), the W jet τ_2/τ_1 (upper right), the top tag soft-drop mass (middle left), the W tag soft-drop mass (middle right), and the subjet b-tagging discriminant (lower). The area of the total background contribution and the area of the signal component are separately normalized to unity. All analysis selections of the signal region are applied with the exception of the variable being plotted. Also shown are vertical dashed lines and arrows, which indicate the optimized selection used for events in the signal region.

The expected number of events from b^* quark production is calculated as $N_{\text{expected}} = \sigma_{b^*}\mathcal{B}(b^* \to tW \to \text{hadrons})\varepsilon L$, where σ_{b^*} is the b^* quark cross section, $\mathcal{B}(b^* \to tW \to \text{hadrons})$ is the branching fraction of $b^* \to tW$ in the fully hadronic decay mode, ε is the product of the acceptance and the efficiency, and L is the integrated luminosity of the data set.

A likelihood fit to data is used to test the signal hypothesis, where the total background model is constructed as a sum of the individual background contributions using a Poisson model for each bin of the (m_t, m_{tW}) distribution.

The number of expected events with failing, n_F, and passing, n_P, top tags in a given bin is given by

$$n_F(i,\vec{\theta}) = n_F^{\text{QCD}}(i) + n_F^{t\bar{t}}(i,\vec{\theta}) + n_F^{\text{single top}}(i,\vec{\theta}) + n_F^{\text{signal}}(i,\vec{\theta}) \tag{4.7}$$

$$n_P(i,\vec{\theta}) = n_P^{\text{QCD}}(i) + n_P^{t\bar{t}}(i,\vec{\theta}) + n_P^{\text{single top}}(i,\vec{\theta}) + n_P^{\text{signal}}(i,\vec{\theta}), \tag{4.8}$$

where i is a bin in the (m_t, m_{tW}) plane, and $\vec{\theta}$ is the set of all nuisance parameters that quantify the systematic uncertainties, as described in Section 4.7. The variable $n_F^{\text{QCD}}(i)$ is an unconstrained positive real number. Finally, $n_P^{\text{QCD}}(i)$ is given by

$$n_P^{\text{QCD}}(i) = n_F^{\text{QCD}}(i) f(m_t, m_{tW}), \tag{4.9}$$

where $f(m_t, m_{tW})$ is a transfer function defined by the ratio of top tagging pass and fail events, and is described in Section 4.6.1.

The negative log-likelihood is then

$$-\ln L(\vec{d};\vec{\theta}) = \sum_{i=1}^{N_{\text{bins,F}}} \left[n_{\text{F}}(i,\vec{\theta}) - d_{\text{F}}(i) \ln n_{\text{F}}(i,\vec{\theta}) \right] + \sum_{i=1}^{N_{\text{bins,P}}} \left[n_{\text{P}}(i,\vec{\theta}) - d_{\text{P}}(i) \ln n_{\text{P}}(i,\vec{\theta}) \right], \tag{4.10}$$

where $N_{\text{bins,F}}$ and $N_{\text{bins,P}}$ are the total number of bins and $d_{\text{F}}(i)$ and $d_{\text{P}}(i)$ are the number of observed events in a given bin, for the fail and pass distributions, respectively. Thus, there is one likelihood which combines four separate categories — signal region "pass" and "fail" and $\text{t}\bar{\text{t}}$ measurement region "pass" and "fail".

4.6.1 Multijet background estimate

After applying the kinematic selection along with the W jet identification, we define the ratio of the multijet background distributions that pass and fail the top tagging requirement in data and QCD multijet MC simulation as $R^{\text{data}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$ and $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$, respectively. Because of the combinatorial nature of multijet processes, $R^{\text{data}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$ and $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$ are both smooth as a function of m_{t} and m_{tW}. The data-to-simulation ratio of these ratios is therefore also smooth and can be used to correct for differences in simulation and data by parameterizing it with an analytic function, $R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}})$.

While $R^{\text{data}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$ could also be described by analytic functions, isolated features of the shape can be factored out by using simulation. By factoring out $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$, the fit of the analytic function to data is only responsible for describing the residual differences between data and simulation that can be parameterized with fewer parameters than the shape of $R^{\text{data}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$.

The number of events in a given bin of the passing category can then be estimated from the equation

$$n^{\text{QCD}}_{\text{P}}(i) = n^{\text{QCD}}_{\text{F}}(i) R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}}) R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}}), \tag{4.11}$$

where $f(m_{\text{t}}, m_{\text{tW}})$ has been replaced by $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}}) R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}})$ and $R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}})$ is a surface parameterized by the product of two one-dimensional polynomials in the $(m_{\text{t}}, m_{\text{tW}})$ plane with coefficients determined from the fit to data. A second-order polynomial was chosen for the m_{t} axis and a first-order polynomial was chosen for the m_{tW} axis. These choices were based on a Fisher test [20] where polynomial terms were added until the p-value obtained in the test was less than 0.95. The parameters of the two-dimensional polynomials are uncorrelated between years. The form of $R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}})$ is then

$$(p_0 + p_1 m_{\text{t}} + p_2 m_{\text{t}}^2)(1 + p_3 m_{\text{tW}}). \tag{4.12}$$

To reduce the effect of statistical fluctuations when calculating $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tW}})$ in the QCD multijet simulation, the pass and fail distributions are smoothed by using an adaptive kernel density estimate [15] (KDE) prior to calculating the ratio. Additionally, the residual contributions from the W+jets and Z+jets

backgrounds are accounted for in this analysis, as they are absorbed by the unconstrained $R_{\text{ratio}}(m_{\text{t}}, m_{\text{tW}})$ function.

4.6.2 Top quark measurement region

By performing the maximum likelihood fit to data in the signal region simultaneously with the $\text{t}\bar{\text{t}}$ background enriched measurement region, we further constrain the $\text{t}\bar{\text{t}}$ contribution to the total background estimate. In particular, this region is used to make measurements of the c_1 and c_2 fit parameters of Eq. (4.6).

The $\text{t}\bar{\text{t}}$ measurement region is evaluated in the $(m_{\text{t}}, m_{\text{tt}})$ plane, where m_{t} is the mass of the second jet when using the top tagging algorithm and m_{tt} is the invariant mass of the $\text{t}\bar{\text{t}}$ pair. Only the multijet and $\text{t}\bar{\text{t}}$ SM processes are considered in this selection since the single top quark contribution is negligible.

The strategy to estimate the multijet background in the $\text{t}\bar{\text{t}}$ measurement region is similar to the signal region. The $R^{\text{t}\bar{\text{t}}}_{\text{ratio}}(m_{\text{t}}, m_{\text{tt}})$ in this region is parameterized with the same polynomial form as in the signal region, but the parameters are uncorrelated with those of the signal region. Additionally, the $R^{\text{MC}}_{\text{P/F}}(m_{\text{t}}, m_{\text{tt}})$ is derived using QCD multijet simulation events that pass the same selection as the $\text{t}\bar{\text{t}}$ measurement region. Events from W+jets and Z+jets backgrounds are suppressed by the initial top tag requirement and any that remain are accounted for by the multijet background model as they are in the signal region.

The negative log-likelihood calculated from the $\text{t}\bar{\text{t}}$ measurement region is

constructed similarly to Eq. (4.10). The total negative log-likelihood is obtained from the sum of the negative log-likelihoods of the signal region and the $t\bar{t}$ measurement region. Because the fit to data can constrain the $t\bar{t}$ background in both selections, the values of the free parameters that determine the shape and normalization of the $t\bar{t}$ simulation are constrained by the simultaneous fit to the $t\bar{t}$- and signal-enriched selections.

4.7 Systematic uncertainties

This analysis takes into account several systematic uncertainties that can affect both the shape and normalization of the simulation.

Normalization uncertainties include those in the production cross section and in the measured integrated luminosity of the data. The uncertainties in the $t\bar{t}$ and single top tW-channel production are taken as 20 and 30%, respectively, to account for the uncertainties in the cross section and in the factorization and renormalization scales of each process. Specifically, these values were chosen based on the largest variations in yield of the simulated samples from varying the factorization and renormalization scales. The uncertainty in the measured integrated luminosity is 1.8% [66, 67, 68] for the complete Run 2 (2016–2018).

Several uncertainties exist that affect both the shape and normalization of the (m_t, m_{tW}) distributions. The uncertainties in the jet energy scale and resolution are estimated through variations in p_T and η of the PUPPI jets [57]. The uncertainty in the pileup reweighting correction is evaluated by varying the total inelastic cross section by $\pm 4.6\%$ [53]. The uncertainty in the trigger

correction is taken into account as a variation of one half of the trigger inefficiency. The uncertainty in the PDFs is derived by either evaluating the root-mean-square of the set of NNPDF MC replicas or by evaluating the contributions of eigenvectors provided in a Hessian set [69], depending on whether the PDF set represents variations as MC replicas or Hessian eigenvectors. The uncertainty due to differences in the data and simulation efficiency for the top jet tagging algorithm is evaluated by using the variations of the top tagging scale factor [64]. The scale factors and uncertainties vary depending on the merging scenarios defined in Section 4.4.1. The W tagging uncertainty is evaluated from variations in the W tagging scale factor and includes an additional uncertainty when extrapolating to jets outside of the p_{T} region used to extract the scale factor. Additionally, the uncertainty in the W tagging soft-drop mass selection is evaluated from variations in the jet mass scale and resolution [64].

Unique to the $\mathrm{t\bar{t}}$ simulation is the uncertainty in the top quark p_{T} reweighting procedure described in Section 4.3, which is extrapolated to high p_{T}. The uncertainty is represented as uncorrelated variations of $\pm 50\%$ in each of the c_1 and c_2 parameters from Eq. (4.6).

Each uncertainty affecting both the shape and normalization is Gaussian constrained where the ± 1 standard deviation of each distribution is mapped to the ± 1 standard deviation of the corresponding unit Gaussian constraint.

The uncertainty in the multijet background estimation is taken from the maximum likelihood fit to data. The parameters of each two-dimensional polynomial are uncorrelated and fitted freely with no *a-priori* constraints. An

additional uncertainty in the "bandwidth" parameter of the KDE algorithm is accounted for by varying the parameter up and down by 1, where the nominal value is 4. This parameter acts as a scale to determine the width of the adaptive kernels.

All systematic uncertainties are considered in the simultaneous fit to data such that all correlations are preserved. The uncertainties are always correlated across tW and $t\bar{t}$ selections within a given year of data and simulation. The cross section, PDF, and top quark p_T reweighting c_1 and c_2 uncertainties are individually correlated across the data-taking years. Table 4.2 summarizes the sources of uncertainty and indicates where correlations between samples exist.

Additionally, Table 4.2 includes a calculation of the "impact" of a parameter on the measurement of the final signal strength for a 2.4 TeV b^* quark signal. This value is calculated by comparing the measured signal strength in the full fit against the measured signal strength in a fit where the given nuisance parameter has been changed either "up" or "down" one standard deviation from its post-fit value in the full fit.

As can be seen in Table 4.2, the multijet estimate from data is the dominant source of background uncertainty in the measurement of the signal strength. In particular, variations of one post-fit standard deviation of the linear term in the m_{tW} axis of the signal region can change the measurement of the signal strength by approximately 19%.

4.8 Results

The (m_t, m_{tW}) and (m_t, m_{tt}) distributions are used in a simultaneous binned maximum likelihood fit to data. The signal strength is a free parameter in the model and the systematic uncertainties are accounted for as nuisance parameters as described in Section 4.6. Normalization uncertainties are modeled with log-normal priors, and uncertainties affecting simulation shapes are modeled using a template morphing approach with Gaussian priors.

While the fit is performed in two dimensions, evaluating the agreement of the background model with the data is more convenient when examining projections onto one dimension. The background estimate and measured two-dimensional distributions from the simultaneous fit of the signal region, $t\bar{t}$ measurement region, and multijet enriched regions are shown in Figs. 4.3 and 4.4, respectively, as one-dimensional projections where either the m_{tt} or m_t distribution has been separated into three regions. The lower panels show the pull, defined as the difference between the number of events observed in the data and the predicted background, divided by the systematic uncertainty in the background and the statistical uncertainty in the data, added in quadrature. All plots shown are for the signal-plus-background hypothesis post-fit, where the 2.4 TeV b^*_{LH} quark sample is normalized to the post-fit signal cross section.

In Fig. 4.3, the left column shows distributions of m_t obtained for the selection of the $t\bar{t}$ measurement region, but with a jet failing the top tagging requirement. The right column shows the same distributions, but for jets passing the top tagging requirement. The rows give the distributions for separate intervals of m_{tt}. The background estimation is observed to model the

data well in both regions, validating the estimation of the multijet background and the modeling of the $t\bar{t}$ simulation. The contribution from a possible signal is negligible in this region and therefore not visible.

In Fig. 4.4, distributions of m_{tW}, obtained for events passing the signal region selection are shown, where the distributions in the left and right columns have been obtained for jets failing and passing the top tagging requirement, respectively. Plots in the row are for separate intervals of m_t. The total background estimate agrees with the data within the uncertainties. The largest excess in data relative to the total background is observed for a left-handed b^* quark with a mass of 2.4 TeV, which results in a local significance of 2.3 standard deviations.

Additionally, the post-fit top quark p_T reweighting measurements are consistent with the pre-fit values, and are measured to be $c_1 = 1.01 \pm 0.25$ and $c_2 = 1.16 \pm 0.16$. The agreement of the background-only model is evaluated using the saturated test statistic [19, 70] and has a p-value of 0.3. Additionally, the post-fit nuisance parameter values are consistent with the pre-fit values and the nuisance parameter values from the background-only model fit are consistent with those from the signal-plus-background model fit.

Asymptotic frequentist statistics are used to derive exclusion limits on $\sigma_{b^*}\mathcal{B}(b^* \to tW \to \text{hadrons})$ at 95% CL [71]. These limits are derived separately for the b^*_{RH}, b^*_{LH}, and b^*_{LH+RH} quark signal hypotheses. The ± 1 and ± 2 standard deviations in the expected limit are derived from pseudo-experiments under the background-only hypothesis in which the nuisance

parameters are randomly varied within the post-fit constraints of the maximum likelihood fit to data.

The limits are shown in Fig. 4.5. The theoretical b^* cross sections included in the figure as a function of b^* quark mass are calculated using MADGRAPH5_*a*MC@NLO. Masses below 2.6, 2.8, and 3.1 TeV (2.9, 3.0, and 3.3 TeV) are observed (expected) to be excluded at 95% CL for the left-handed, right-handed, and vector-like hypotheses, respectively. These limits nearly doubles the mass exclusions of the previous result [31].

The sensitivity of this analysis can also be compared to the sensitivity of the CMS dijet search [72]. The branching fraction for $b^* \to bg$ approaches 20% asymptotically for high masses [29]. From the dijet search, the expected upper limit on the product of the cross section and branching fraction for a resonance decaying to a quark and a gluon is approximately 0.09 pb at 2 TeV so the cross section upper limit on b^* quark production is approximately 0.45 pb. Using the left-handed couplings result in Fig. 4.5, this analysis achieves an expected upper limit of approximately 0.015 pb at 2 TeV. With the $b^* \to tW$ branching fraction of 0.4, the cross section upper limit on b^*_{LH} quark production at 2 TeV is approximately 0.0375 pb. Thus, at 2 TeV, this search is about an order of magnitude more sensitive to the excited b^* quark than the dijet search.

The results of this search can also be used to test models of a single B quark produced via the electroweak interaction in association with a bottom or top quark and decaying into a top quark and a W boson. Because the cross section for this process is much smaller than for a b^* quark produced through the strong force, and because of the selection $m_{tW} > 1.2$ TeV, we consider

the mass range 1.4 to 1.8 TeV in this interpretation. The exclusion limits on $\sigma_{\mathrm{B}}\mathcal{B}(\mathrm{B} \rightarrow \mathrm{tW} \rightarrow \text{hadrons})$ at 95% CL are shown in Fig. 4.6. Over the mass range of 1.4–1.8 GeV, the observed upper limit ranges 0.027 to 0.009 pb when produced in association with a bottom quark and from 0.036 to 0.012 pb when produced in association with a top quark. Because of the small theoretical cross section for the model considered, no mass limit is set. When compared to the b^{*} quark in this mass range, the expected cross section upper limits for a B quark produced with an associated bottom quark are uniformly more sensitive by approximately 22%. The equivalent comparison for a B quark produced with an associated top quark shows the sensitivity is worse by no more than 7%.

These results with 137 fb^{-1} of data can be compared directly to those of Ref. [73], which analyzed the lepton+jets channel in 35.9 fb^{-1} of data recorded with the CMS experiment at $\sqrt{s} = 13$ TeV. At a B quark mass of 1.4 TeV, this analysis is less sensitive than the results from Ref. [73] by about 20% when considering B quark production with an associated top quark. However, this analysis has about 20% higher sensitivity than the previous analysis when the production is in association with a bottom quark. As the B quark mass increases, the sensitivity of this analysis increases faster than the analysis described in Ref. [73]. Thus, the sensitivity of this analysis at 1.8 TeV is about 27% higher for the associated top quark hypothesis and about a factor of two higher for the associated bottom quark hypothesis.

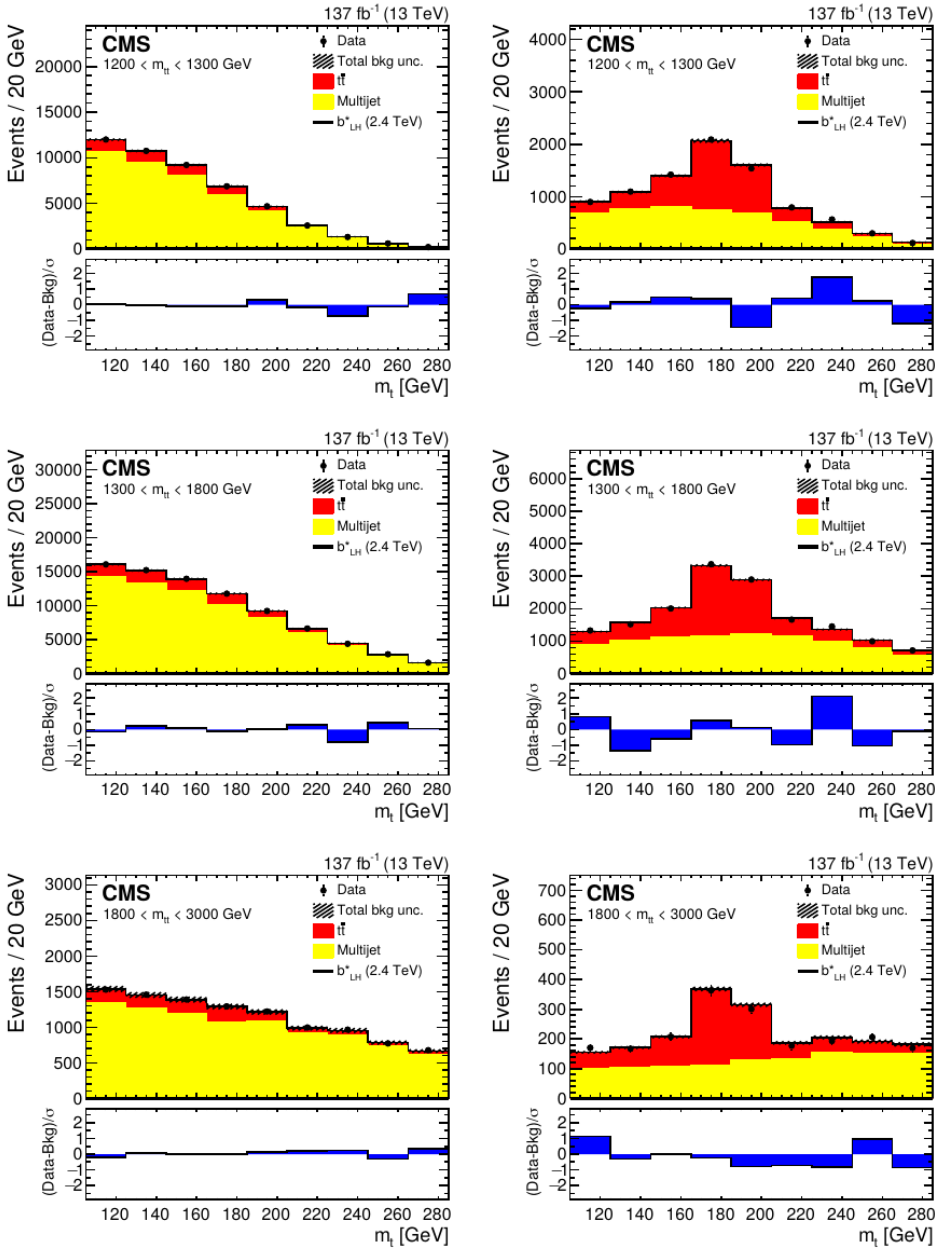

Figure 4.3: Distributions of m_{t} in the $\mathrm{t\bar{t}}$ measurement region for three intervals of m_{tt}: 1200–1300 GeV (upper), 1300–1800 GeV (middle), 1800–3000 GeV (lower). The data are shown by points with error bars and the individual background contributions by filled histograms with the barely visible shaded region indicating the uncertainty in the total background estimate. The signal is not visible because the contamination in this region is negligible. The left and right columns show distributions for events with the second jet failing and passing the top tagging requirement, respectively. The lower panels of each figure show the pull, as a function of m_{t}.

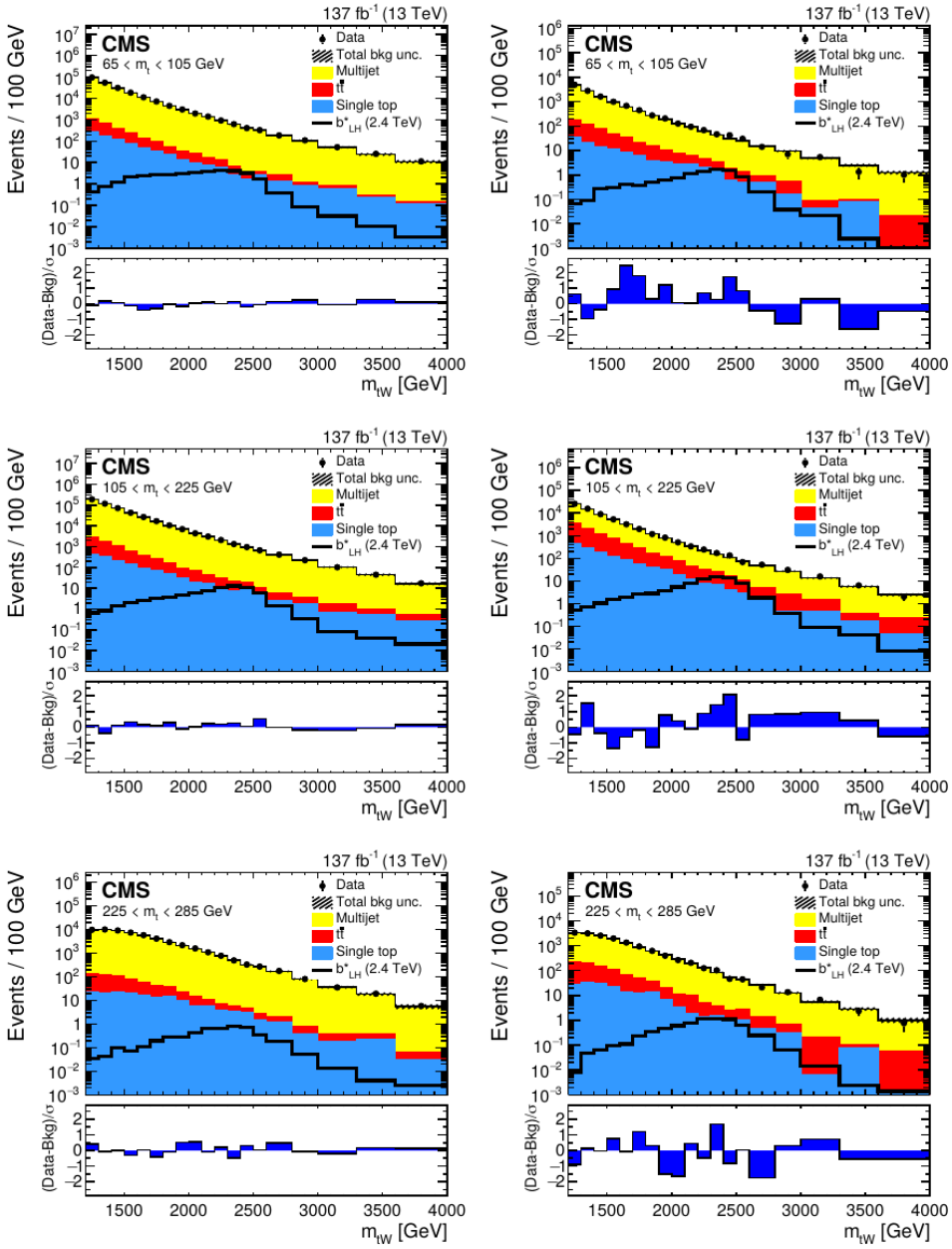

Figure 4.4: Distributions of m_{tW} in the b^* signal region for three intervals of m_{t}: 65–105 GeV (upper), 105–225 GeV (middle), and 225–285 GeV (lower). The data are shown by points with error bars, the individual background contributions by filled histograms, and a 2.4 TeV $\mathrm{b}^*_{\mathrm{LH}}$ signal is shown as a solid line. The barely visible shaded region is the uncertainty in the total background estimate. The left and right columns show distributions for events with a jet failing and passing the top tagging requirement, respectively. The lower panels of each figure show the pull, as a function of m_{tW}.

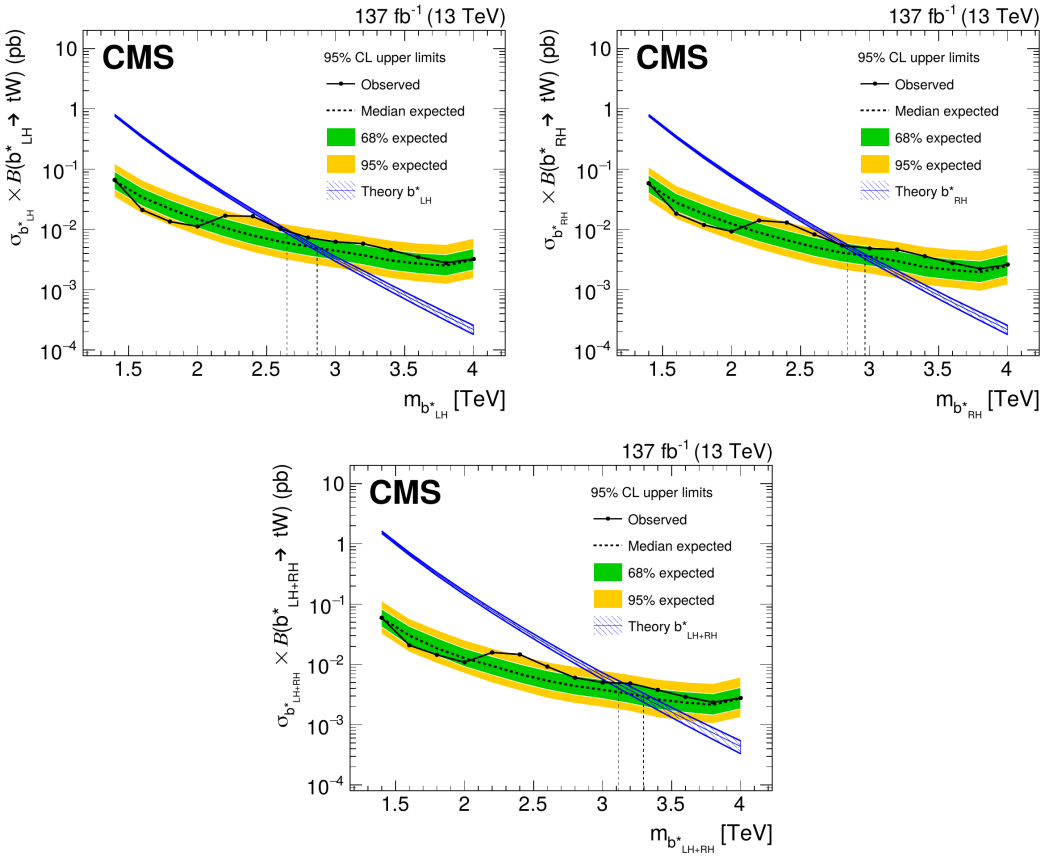

Figure 4.5: Upper limits on the product of the cross section and branching fraction at 95% CL for a b^*_{LH} (upper), b^*_{RH} (middle), and b^*_{LH+RH} (lower) quark as a function of the b^* quark mass. The expected (dashed) and observed (dot-solid) limits, as well as the b^* quark theoretical cross sections (shaded-solid), are shown. The vertical dashed lines indicate the intersection of the theoretical cross sections with the expected and observed limits. The inner and outer shaded areas around the expected limits show the 68% and 95% CL intervals, respectively.

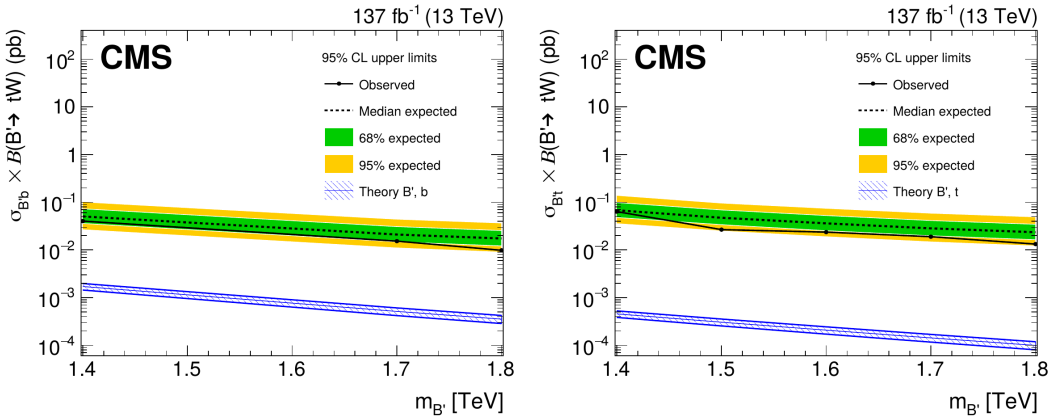

Figure 4.6: Upper limits on the product of the cross section and branching fraction at 95% CL for a B produced in association with a bottom quark (left) and top quark (right) as a function of the B quark mass. The expected (dashed) and observed (dot-solid) limits, as well as the B quark theoretical cross sections (shaded-solid), are shown. The inner and outer shaded areas around the expected limits show the 68% and 95% CL intervals, respectively.

Table 4.2: Sources of uncertainty that are taken into account in the statistical analysis of the data. The sources affecting the normalization are given with their percentage uncertainties, while the sources affecting the shape are listed as "Shape" together with the dependent parameter. The rightmost column indicates the impact of the parameter on the 2.4 TeV b^* signal strength when the parameter is changed "up" and "down" by one standard deviation from its post-fit value. For parameters where the uncertainties are uncorrelated between data-taking years, the average impact is calculated. An impact of $+0.0$ (-0.0) denotes an impact that is less (greater) than 0.1 (-0.1) but greater (less) than 0.

			Impact	
Source	Uncertainty	Samples	Up	Down
$t\bar{t}$ cross section	$\pm 20\%$	$t\bar{t}$	-4.6	$+4.4\%$
Single top cross section	$\pm 30\%$	Single top	$+1.2$	-1.4%
Integrated Luminosity	$\pm 1.8\%$	$t\bar{t}$, single top, signal	$+1.6$	-1.1%
Pileup	Shape (σ_{mb})	$t\bar{t}$, single top, signal	$+0.3$	-0.2%
Prefire	Shape (p_T, η)	$t\bar{t}$, single top, signal	$+0.0$	$+0.1\%$
Jet energy scale	Shape (p_T)	$t\bar{t}$, single top, signal	$+0.3$	-0.6%
Jet energy resolution	Shape (p_T, η)	$t\bar{t}$, single top, signal	-0.4	-0.5%
Jet mass scale	Shape (m_W)	$t\bar{t}$, single top, signal	-0.1	-0.0%
Jet mass resolution	Shape (m_W)	$t\bar{t}$, single top, signal	$+0.0$	$+0.9\%$
W tagging	Shape (p_T)	Single top, signal	$+0.9$	-0.9%
W tagging: p_T extrapolation	Shape (p_T)	Single top, signal	$+4.9$	-4.9%
Top tagging, merged	Shape (p_T)	$t\bar{t}$, single top, signal	$+0.2$	-0.2%
Top tagging, semimerged	Shape (p_T)	$t\bar{t}$, single top, signal	$+1.1$	-0.9%
Top tagging, not merged	Shape (p_T)	$t\bar{t}$, single top, signal	-0.1	$+0.1\%$
Trigger	Shape (H_T)	$t\bar{t}$, single top, signal	$+0.3$	-0.4%
Top quark p_T correction c_1	Shape (p_T)	$t\bar{t}$	-0.3	$+0.3\%$
Top quark p_T correction c_2	Shape (p_T)	$t\bar{t}$	-3.9	$+3.5\%$
PDF	Shape (m_t, m_{tW})	Signal	$+0.1$	-0.1%
KDE bandwidth	Shape (m_t, m_{tW})	Multijet (from simulation)	-1.2	$+0.2\%$
$R^{SR}_{ratio}(m_t, m_{tW})p_0$	Shape (m_t, m_{tW})	Multijet (from data)	-4.4	$+0.0\%$
$R^{SR}_{ratio}(m_t, m_{tW})p_1$	Shape (m_t, m_{tW})	Multijet (from data)	-2.0	$+2.2\%$
$R^{SR}_{ratio}(m_t, m_{tW})p_2$	Shape (m_t, m_{tW})	Multijet (from data)	$+0.9$	-0.8%
$R^{SR}_{ratio}(m_t, m_{tW})p_3$	Shape (m_t, m_{tW})	Multijet (from data)	$+18.6$	-18.8%
$R^{t\bar{t}}_{ratio}(m_t, m_{tt})p_0$	Shape (m_t, m_{tt})	Multijet (from data)	-0.4	$+0.6\%$
$R^{t\bar{t}}_{ratio}(m_t, m_{tt})p_1$	Shape (m_t, m_{tt})	Multijet (from data)	-0.4	$+0.6\%$
$R^{t\bar{t}}_{ratio}(m_t, m_{tt})p_2$	Shape (m_t, m_{tt})	Multijet (from data)	$+0.5$	-0.6%
$R^{t\bar{t}}_{ratio}(m_t, m_{tt})p_3$	Shape (m_t, m_{tt})	Multijet (from data)	-0.6	$+0.6\%$

4.9 Summary

A search for a heavy resonance decaying to a top quark and a W boson in the fully hadronic final state has been presented. The analysis uses proton-proton collision data at a center-of-mass energy of 13 TeV corresponding to an integrated luminosity of 137 fb^{-1}, collected by the CMS experiment at the LHC during 2016–2018.

This analysis considers the explicit case where the heavy resonance is an excited bottom quark, b^*. The search evaluates b^* quark masses greater than 1.2 TeV, which result in highly Lorentz-boosted top quarks and W bosons that are reconstructed as single jets. Using jet substructure algorithms designed to distinguish heavy resonance decays from light-quark and gluon jets, the top quark and W boson decays are identified as a top quark jet and a W boson jet, respectively.

The background processes in the analysis are a result of multijet processes from the strong interaction, $\text{t}\bar{\text{t}}$ production, and single top quark (tW-channel) production. The search is performed using a two-dimensional binned likelihood fit to the data that allows all backgrounds to be fit simultaneously. The multijet component in the signal region is estimated via a two-dimensional transfer function method that uses a multijet-enriched control region. The $\text{t}\bar{\text{t}}$ and single top background estimates are determined via a template fit to data. In particular, a dedicated $\text{t}\bar{\text{t}}$ measurement region is used to constrain the shape and yield of the $\text{t}\bar{\text{t}}$ background.

No statistically significant deviation from the standard model expectation

is observed. The hypotheses of b^* quarks with left-handed, right-handed, and vector-like chiralities are excluded at 95% confidence level for masses below 2.6, 2.8, and 3.1 TeV, respectively. These are the most stringent limits on the b^* quark mass to date, extending the previous best mass limits by almost a factor of two.

Chapter 5

Discussion and Conclusion

Presented in this thesis is a novel background modeling technique, 2D Alphabet, which builds a complete signal plus background model and simultaneously measures resonant backgrounds using MC simulation, non-resonant backgrounds using a data-driven technique, and the final parameter of interest for a BSM analysis, the signal strength. The resonant backgrounds are defined using Gaussian constrained binned shape templates built from MC simulation. The combinatorial non-resonant backgrounds (primarily multijet events from QCD processes) are described using a background enriched control region multiplied by an unconstrained transfer function which is measured in-situ. To account for the changing resonant contributions in the control region, the control region and signal region are also fit simultaneously.

As a software framework, 2D Alphabet constructs a model for a specific type of background estimate, provides the model as input to the Combine statistical tool, plots the 2D distributions from the fit result, and provides the infrastructure to test this result. Interacting with the framework is done via a

JSON configuration file and the 2DAlphabet API which allows for custom definitions of binned parametric distributions that can be related to one another via transfer functions. This standardization of the 2D Alphabet methodology in software makes the method accessible to other analysis teams and ensures results are reproducible.

The method was developed in parallel with the search for an excited bottom quark, b^*, in the all-hadronic decay channel. The dominant multijet component of the SM background necessitated developing a novel background estimation method which could effectively measure the data-driven multijet background simultaneous to the $t\bar{t}$ contribution. By using the two-dimensional transfer function with KDE-smoothed QCD simulation, a dedicated $t\bar{t}$ measurement region, and $137\,fb^{-1}$ of 13 TeV data collected by the CMS experiment, the analysis was able to increase the world-best limits on the mass of the b^* by a factor of two.

2D Alphabet is also in use with two other analyses which are on track for publication. The first is an analysis searching for di-Higgs resonances with both Higgs bosons decaying to $b\bar{b}$. It considers when the bottom quarks of each Higgs are merged to form two fat jets in the event. It also considers the scenario of one unmerged Higgs where the bottom quarks are resolved as two AK4 jets (with the other Higgs still identified as a fat jet). The merged Higgs jets are identified using a neural network based tagger, DeepAK8 [64]. The procedure to model the multijet and $t\bar{t}$ backgrounds is nearly the same as the procedure described for the tW analysis described in Ch. 4.

The second analysis using 2D Alphabet searches for a BSM particle, X,

decaying to a new, Y, and a Higgs boson. The Y is considered Higgs-like and so the final state considered is again composed of four bottom quarks. However, since the mass of the Y is not known, the entirety of the horizontal axis in Fig. 3.3 must be blinded. The procedure is then validated in a subset of data which is disjoint from the analysis signal selection. The multijet background estimation is again similar to the one described in Ch. 4 but the resonant background contributions are constrained with dedicated methods since the jet mass sidebands are not available for analysis while blinded.

The needs of these three analyses have informed the needs of the 2D Alphabet software implementation. For example, the newest version allows for multiple "pass" regions which can have dedicated transfer functions from the same "fail" control region. Additionally, the tools used to define a distribution of freely floating bins or a distribution defined by a parametric function have been factorized with the further addition that these objects can be combined via addition, subtraction, multiplication, and division to make most simple operations possible. For even more specific use cases, custom Python classes are easily defined via inheritance of these existing classes. This opens many possibilities, from functional parameterization of signals to custom definitions to correlate data-driven backgrounds. These features are being used by two more analyses within the Johns Hopkins CMS group, one of which searches for tH resonances and the other for HWW resonances.

As can be seen, the 2D Alphabet method is applicable to many BSM searches. Using a two-dimensional plane leads to this generalizability but provides other significant advantages over the one-dimensional counterparts,

including the ability to build a robust all-in-one model that can fit simulation shape templates simultaneous to measuring the data-driven backgrounds and the potential signal contribution. One only needs to ensure the shapes of backgrounds and signal are distinguishable in the two-dimensional plane of their measurement variables, fill the histograms and point to them in the JSON, define the data-driven model (or none at all), and allow the software to do the rest.

References

[1] Nasfarley88 MissMJ et al. *Standard Model of Elementary Particles modified version*. 2014. URL: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles_modified_version.svg.

[2] Konrad Jende Tim Herrmann. *Beta Decay*. 2010. URL: https://atlas.physicsmasterclasses.org/wpath_files/img/Feynman/betaminus.png.

[3] Niamh O'C. *Weak Decay (flipped)*. 2013. URL: https://upload.wikimedia.org/wikipedia/commons/4/4b/Weak_Decay_%28flipped%29.svg.

[4] Kurzon. *Geiger-Marsden experiment expectation and result*. 2014. URL: https://commons.wikimedia.org/wiki/File:Geiger-Marsden_experiment_expectation_and_result.svg.

[5] Ylai. *CTEQ6 parton distribution functions*. 2005. URL: https://upload.wikimedia.org/wikipedia/commons/0/0d/CTEQ6_parton_distribution_functions.png.

[6] David Barney. "CMS Detector Slice". CMS Collection. Jan. 2016. URL: https://cds.cern.ch/record/2120661.

[7] *Color Charge and Confinement*. URL: http://fafnir.phyast.pitt.edu/particles/color.html.

[8] Flip Tanedo. *When Feynman Diagrams Fail*. 2010. URL: http://blogs.uslhc.us/wp-content/uploads/2010/12/partonshower.png.

[9] A. M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[10] Vardan Khachatryan et al. "The CMS trigger system". In: *JINST* 12 (2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020. arXiv: 1609.02366 [physics.ins-det].

[11] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_T jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ex].

[12] Daniele Bertolini et al. "Pileup per particle identification". In: *JHEP* 10 (2014), p. 059. DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013 [hep-ph].

[13] CMS Collaboration and Thomas Mc Cauley. "Collisions recorded by the CMS detector on 14 Oct 2016 during the high pile-up fill". CMS Collection. Nov. 2016. URL: https://cds.cern.ch/record/2231915.

[14] Andrew J. Larkoski et al. "Soft drop". In: *JHEP* 05 (2014), p. 146. DOI: 10.1007/JHEP05(2014)146. arXiv: 1402.2657 [hep-ph].

[15] Kyle S. Cranmer. "Kernel estimation in high-energy physics". In: *Comput. Phys. Commun.* 136 (2001), p. 198. DOI: 10.1016/S0010-4655(00)00243-5. arXiv: hep-ex/0011057.

[16] J. S. Conway. *Incorporating Nuisance Parameters in Likelihoods for Multi-source Spectra*. 2011. arXiv: 1103.0354 [physics.data-an].

[17] F. James. "MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1". In: (1994).

[18] Rene Brun and Fons Rademakers. "ROOT - An object oriented data analysis framework". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.1 (1997). New Computing Techniques in Physics Research V, pp. 81–86. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(97)00048-X.

[19] Steve Baker and Robert D. Cousins. "Clarification of the use of chi square and likelihood functions in fits to histograms". In: *Nucl. Instrum. Meth.* 221 (1984), p. 437. DOI: 10.1016/0167-5087(84)90016-4.

[20] R. A. Fisher. "On the interpretation of χ^2 from contingency tables, and the calculation of p". In: *J. Royal Stat. Soc.* 85 (1922), p. 87. ISSN: 09528385. DOI: 10.2307/2340521.

[21] U. Baur, M. Spira, and P. M. Zerwas. "Excited-quark and -lepton production at hadron colliders". In: *Phys. Rev. D* 42 (1990), p. 815. DOI: 10.1103/PhysRevD.42.815.

[22] Timothy M. P. Tait and C.-P. Yuan. "Single top quark production as a window to physics beyond the standard model". In: *Phys. Rev. D* 63 (2000), p. 014018. DOI: 10.1103/PhysRevD.63.014018. arXiv: hep-ph/0007298 [hep-ph].

[23] Clifford Cheung, A. Liam Fitzpatrick, and Lisa Randall. "Sequestering CP violation and GIM-violation with warped extra dimensions". In: *JHEP* 01 (2008), p. 069. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2008/01/069. arXiv: 0711.4421.

[24] A. Liam Fitzpatrick, Gilad Perez, and Lisa Randall. "Flavor anarchy in a Randall-Sundrum model with 5D minimal flavor violation and a low Kaluza-Klein scale". In: *Phys. Rev. Lett.* 100 (2008), p. 171604. ISSN: 1079-7114. DOI: 10.1103/physrevlett.100.171604. arXiv: 0710.1869.

[25] Cesare Bini, Roberto Contino, and Natascia Vignaroli. "Heavy-light decay topologies as a new strategy to discover a heavy gluon". In: *JHEP* 01 (2012), p. 157. ISSN: 1029-8479. DOI: 10.1007/jhep01(2012)157. arXiv: 1110.6058.

[26] Natascia Vignaroli. "Discovering the composite Higgs through the decay of a heavy fermion". In: *JHEP* 07 (2012), p. 158. ISSN: 1029-8479. DOI: 10.1007/jhep07(2012)158. arXiv: 1204.0468.

[27] Natascia Vignaroli. "$\Delta F = 1$ constraints on composite Higgs models with left-right parity". In: *Phys. Rev. D* 86 (2012), p. 115011. ISSN: 1550-2368. DOI: 10.1103/physrevd.86.115011. arXiv: 1204.0478.

[28] S. Chatrchyan et al. "The CMS experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[29] Joseph W. Nutter et al. "Single top production as a probe of B′ quarks". In: *Phys. Rev. D* 86 (2012), p. 094006. ISSN: 1550-2368. DOI: 10.1103/physrevd.86.094006. arXiv: 1207.5179 [hep-ph].

[30] Georges Aad et al. "Search for single b^*-quark production with the ATLAS detector at $\sqrt{s} = 7$ TeV". In: *Phys. Lett. B* 721 (2013), p. 171. DOI: 10.1016/j.physletb.2013.03.016. arXiv: 1301.1583 [hep-ex].

[31] Sirunyan Khachatryan et al. "Search for the production of an excited bottom quark decaying to tW in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *JHEP* 01 (2016), p. 166. ISSN: 1029-8479. DOI: 10.1007/JHEP01(2016)166. arXiv: 1509.08141.

[32] Vardan Khachatryan et al. "Search for resonances and quantum black holes using dijet mass spectra in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *Phys. Rev. D* 91 (2015), p. 052009. DOI: 10.1103/PhysRevD.91.052009. arXiv: 1501.04198 [hep-ex].

[33] G. Aad et al. "Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *JHEP* 03 (2020), p. 145. ISSN: 1029-8479. DOI: 10.1007/jhep03(2020)145. arXiv: 1910.08447.

[34] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. "Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning". In: *Phys. Rep.* 841 (2020), p. 1. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2019.11.001. arXiv: 1709.04464.

[35] Roman Kogler et al. "Jet substructure at the Large Hadron Collider". In: *Rev. Mod. Phys.* 91 (2019), p. 045003. ISSN: 1539-0756. DOI: 10.1103/revmodphys.91.045003. arXiv: 1803.06991 [hep-ex].

[36] J. A. Aguilar-Saavedra et al. "Handbook of vectorlike quarks: Mixing and single production". In: *Phys. Rev. D* 88 (2013), p. 094010. ISSN: 1550-2368. DOI: 10.1103/physrevd.88.094010. eprint: 1306.0572.

[37] Andrea De Simone et al. "A first top partner hunter's guide". In: *JHEP* 04 (2013), p. 004. ISSN: 1029-8479. DOI: 10.1007/jhep04(2013)004. eprint: 1211.5663.

[38] David Krohn, Jesse Thaler, and Lian-Tao Wang. "Jet Trimming". In: *JHEP* 02 (2010), p. 084. DOI: 10.1007/JHEP02(2010)084. arXiv: 0912.1342 [hep-ph].

[39] Albert M Sirunyan et al. "Measurements of $t\bar{t}$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons". In: *JHEP* 02 (2019), p. 149. DOI: 10.1007/JHEP02(2019)149. arXiv: 1811.06625 [hep-ex].

[40] Vardan Khachatryan et al. "Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV". In: *Phys. Rev. D* 95 (2017), p. 092001. DOI: 10.1103/PhysRevD.95.092001. arXiv: 1610.04191 [hep-ex].

[41] Emanuele Re. "Single-top Wt-channel production matched with parton showers using the POWHEG method". In: *Eur. Phys. J. C* 71 (2011), p. 1547. DOI: 10.1140/epjc/s10052-011-1547-z. arXiv: 1009.2450 [hep-ph].

[42] Paolo Nason. "A new method for combining NLO QCD with shower Monte Carlo algorithms". In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146.

[43] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: The POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[44] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: The POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[45] Stefano Frixione, Giovanni Ridolfi, and Paolo Nason. "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction". In: *JHEP* 09, 126 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph].

[46] 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". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[47] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[48] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[49] Vardan Khachatryan et al. "Event generator tunes obtained from underlying event and multiparton scattering measurements". In: *Eur. Phys. J. C* 76 (2016), p. 155. DOI: 10.1140/epjc/s10052-016-3988-x. arXiv: 1512.00815 [hep-ex].

[50] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77 (2017), p. 663. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[51] Albert M Sirunyan et al. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". In: *Eur. Phys. J. C* 80 (2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179 [hep-ex].

[52] S. Agostinelli et al. "GEANT4—a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), p. 250. DOI: 10.1016/S0168-9002(03)01368-8.

[53] M. Aaboud et al. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC". In: *Phys. Rev. Lett.* 117 (2016), p. 182002. DOI: 10.1103/PhysRevLett.117.182002. arXiv: 1606.02625 [hep-ex].

[54] Albert M Sirunyan et al. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV". In: *JHEP* 07 (2018), p. 161. DOI: 10.1007/JHEP07(2018)161. arXiv: 1802.02613 [hep-ex].

[55] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet user manual". In: *Eur. Phys. J. C* 72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[56] Albert M Sirunyan et al. "Pileup mitigation at CMS in 13 TeV data". In: *JINST* 15 (2020), P09018. ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/09/p09018. arXiv: 2003.00503.

[57] Vardan Khachatryan et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[58] CMS Collaboration. *Jet algorithms performance in 13 TeV data*. CMS Physics Analysis Summary CMS-PAS-JME-16-003. 2017. URL: http://cds.cern.ch/record/2256875.

[59] Mrinal Dasgupta et al. "Towards an understanding of jet substructure". In: *JHEP* 09 (2013), p. 029. DOI: 10.1007/JHEP09(2013)029. arXiv: 1307.0007 [hep-ph].

[60] Jonathan M. Butterworth et al. "Jet substructure as a new Higgs search channel at the LHC". In: *Phys. Rev. Lett.* 100 (2008), p. 242001. DOI: 10.1103/PhysRevLett.100.242001. arXiv: 0802.2470 [hep-ph].

[61] Jesse Thaler and Ken Van Tilburg. "Identifying boosted objects with N-subjettiness". In: *JHEP* 03 (2011), p. 015. DOI: 10.1007/JHEP03(2011)015. arXiv: 1011.2268 [hep-ph].

[62] Jesse Thaler and Ken Van Tilburg. "Maximizing boosted top identification by minimizing N-subjettiness". In: *JHEP* 02 (2012), p. 093. DOI: 10.1007/JHEP02(2012)093. arXiv: 1108.2701 [hep-ph].

[63] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[64] Albert M Sirunyan et al. "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques". In: *JINST* 15 (2020), P06005. DOI: 10.1088/1748-0221/15/06/P06005. arXiv: 2004.08262 [hep-ex].

[65] Particle Data Group, P. A. Zyla, et al. "Review of particle physics". In: *Prog. Theor. Exp. Phys.* 2020 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104.

[66] *CMS luminosity measurements for the 2016 data taking period*. CMS Physics Analysis Summary CMS-PAS-LUM-17-001. 2017. URL: https://cds.cern.ch/record/2257069.

[67] *CMS luminosity measurement for the 2017 data-taking period at* $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-17-004. 2018. URL: https://cds.cern.ch/record/2621960.

[68] *CMS luminosity measurement for the 2018 data-taking period at* $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-18-002. 2019. URL: https://cds.cern.ch/record/2676164.

[69] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865.

[70] J. K. Lindsey. *Parametric statistical inference*. New York: Oxford University Press, 1966.

[71] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011), p. 1554. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727.

[72] Albert M Sirunyan et al. "Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 05 (2020), p. 033. ISSN: 1029-8479. DOI: 10.1007/jhep05(2020)033. arXiv: 1911.03947.

[73] Albert M Sirunyan et al. "Search for single production of vector-like quarks decaying to a top quark and a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 79 (2019), p. 90. DOI: 10.1140/epjc/s10052-019-6556-3. arXiv: 1809.08597 [hep-ex].