Measuring the Inclusive Jet Cross Section and its Interpretation Within the SMEFT

by

Ali Al Kadhim

PhD Prospectus
Physics Department
Florida State University
April 2022

Thesis Supervisor: Harrison B. Prosper
Kirby W. Kemper Endowed Professor of Physics

Contents

Glossary

Standard Model (SM) The Standard Model of particle physics (SM) is a theory describing three of the four fundamental forces (the electromagnetic, weak and strong interactions). This theory is based on $\mathrm{SU_c}(3) \times \mathrm{SU_L}(2) \times \mathrm{U_Y}(1)$ gauge symmetry, and it classifies all known elementary particles. 1, 5

HEP Abbreviation for High Energy Physics, another name for particle physics. 1

LHC The Large Hadron Collider (LHC) at the European Laboratory for Particle Physics (CERN) is the world's largest and most powerful particle accelerator. The LHC hosts four particle detectors. The biggest detectors and collaborations at the LHC are ATLAS and CMS. 1

HGCAL The CMS High Granularity Calorimeter (HGCAL) is a sampling calorimeter which uses silicon sensors and scintillators to which measures the energies of the particles. 1, 7

Bunch Crossing A "bunch" is a very large collection of protons. In the LHC, bunches of protons cross, or collide with each other, approximately once every 25 ns, leading to more than 600 million collisions per second. 1, 9

Golden JSON The Golden JSON is a JSON file which contains a list of run numbers - or results - of data that are deemed to be good or 'usable for analysis' to the CMS collaboration. 1, 10

Run A The LHC operates in running periods, for example, Run 2 operated from 2015-2018. Further, in CMS, data are recorded in yearly "run eras" in chronological order, for example, data in Run A was recorded prior to data in Run B, and so on. 1, 10

Ultra Legacy (UL) CMS Ultra Legacy datasets refer to the datasets which use improved detector calibrations to achieve optimal performance. Processing of these data is used for analyses requiring optimal energy resolution and will be preserved for future CMS analyses. 1, 9

Level 1 (L1) Trigger The CMS trigger ensures that potentially interesting events are recorded with high efficiency. The Level 1 (L1) trigger, comprising the calorimeter, muon and global trigger processors, uses coarse-granularity information to select (decide whether to save) the most interesting events in less than 4μs. After L1 triggering,

data are transferred from the detector readouts to the High Level Trigger (HLT) processing farm, which run reconstruction algorithms to further decrease the event rate before data storage. 1, 9

JETMET POG The JetMET Physics Object Group (POG) is responsible for monitoring, reconstructing, calibrating, and providing corrections and software tools for jets and missing energy in CMS. 1, 9

Luminosity Luminosity, $\mathcal{L}$, is a measure of the number of collisions that can be produced per cm^2 per second. 1

Monte Carlo (MC) Monte Carlo event generators, or MC, are simulations of the underlying physics based on theory, and based on the assumption that systems can be described by probability denso=ity functions which can be modeled. They are indispensible tools of particle physics, and enter every particle physics analysis, from modelling the particle interactions to simulating the response of detectors. 1, 9

Statistical Model A probability model $p(x|\mu, \theta)$ that includes the dependence on all the data x, the parameters of interest μ and the nuisance parameters θ 1

Likelihood The likelihood function is the statistical model into which data have been entered, and is a function of the parameters $L(\mu, \theta)$. 1

QCD Quantum Chromodynamics, or QCD, is an important part of the Standard Model of particle physics (SM). It describes the strong interaction between quarks and gluons, which are the constituents of hadrons such as the proton and neutron. 1

Chapter 1

Overview and Motivation

1.1 Introduction

We live in a unique time in the history of fundamental physics. On the one hand, we are confronted with deep unanswered questions. On the other hand, our field has developed an ability to derive theoretical prediction based on the (Standard Model (SM)) to a mind-boggling precision, and our experiments of teams have matured into the most complex and largest scientific collaborations ever organized by humankind. If we have a chance to answer these deep questions in the near future, we must exploit all the available data from the Large Hadron Collider (LHC) at the European Laboratory for Particle Physics (CERN). Furthermore, we should to exploit the accelerating advances in modern statistical techniques, machine learning and computer science to aid us in the process. This is what I hope to do in my PhD research.

This document is organized as follows: in Chapter 1, I give an overview of the physics to be addressed, and my short and long term plans for my PhD thesis. In Chapter 2, I summarize the previous relevant research projects that I have done at FSU so far, and how each of them will inform my upcoming thesis plan. Finally, in Chapter 3, I describe my research plans in more detail.

1.2 The CMS Detector

The Comapact Muon Solenoid (CMS) detector is a general-purpose detector at the LHC. It is designed to address a wide variety of physics questions, such as searching for physics beyond the Standard Model. The central feature of the CMS apparatus is a superconducting solenoid which provides a strong magnetic field. Within the field volume are silicon pixel and strip tracker, which tracks charged particles, a lead tungstate crystal electromagnetic calorimeter (ECAL), which measures the energy of electrons and photons, and a brass/scintillator hadron calorimeter (HCAL), which measures the energy of hadrons. The outermost sub-detecctor of the CMS experiment is the muon system, which is composed of muon chambers, is used to identify muons and measure their momenta [1].

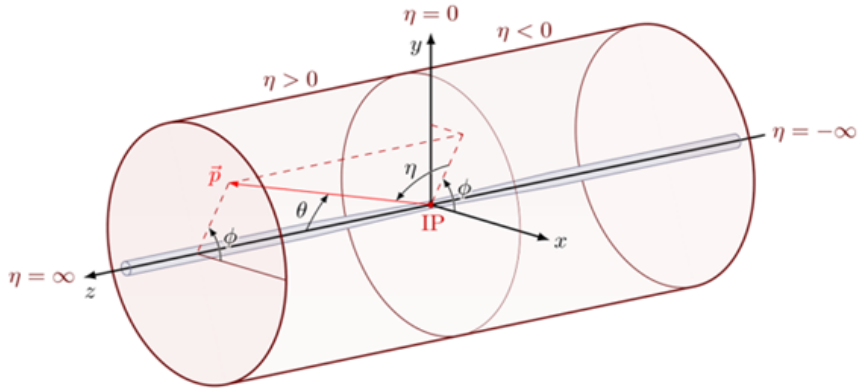


Figure 1-1: In the CMS experiment, the beam direction is taken to be in the z direction, and the $x-y$ plane is perpendicular to the beam. The polar angle θ is the angle between the particle momentum direction and the z direction. A particle with energy E and momentum along the beam direction p_z has rapidity $y \equiv \frac{1}{2} \ln \frac{E+p_x}{E-p_x}$ and pseudorapidity $\eta \equiv -\ln \tan \theta/2$. Massless particles have $y = \eta$, and differences in rapidity are invariant under longitudinal (along the beam direction) boosts.

1.3 Jets and Jet Cross Sections

Protons and neutrons are composed of quarks and gluons. Immediately after a quark or gluon, i.e. a "parton", is produced, it fragments and hadronizes into energetic hadrons [1]. The collimated spray of hadrons is called a jet.[2] Jets are produced in abundance in hadron colliders and jet production is the dominant high transverse momentum (p_T) process at the LHC, and studying it gives us the best chance of understanding the physics of partons [2].

In the CMS experiment at the LHC, jets are reconstructed from energy deposits of stable particles (the jet constituents with lifetime $c\tau > 1$ cm) in the CMS detectors. These energy deposits are input to a clustering algorithm where one sums the momenta of all particles j within a circle ("cone") of radius R around particle i in azimuthal angle ϕ and rapidity y (or pseudorapidity η) i.e. taking all particles j such that $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 < R^2$. See Figure 1-1 for a visualization (and definitions) of these coordinates.

Jets can provide information about fundamental physics. One of the best studied quantities at the LHC related to jets is the inclusive jet spectrum, which is related to the $2 \to 2$ scattering of partons inside the proton. In this process, the energy of a jet is closely related to that of the hard scattering of partons inside the proton; therefore, the inclusive jet spectrum provides information about the distributions of partons inside the proton.

Measurements of the inclusive jet and dijet cross sections are classical particle physics measurements and are benchmarks of the Standard Model at particle colliders. Such measurements have been performed in e^+e^-, ep, pp, and $p\bar{p}$ colliders. They have been used to test the predictions of perturbative Quantum Chromodynamics (QCD), have given precise measure-

[1] Almost 85% of the constituents of jets are charged hadrons, such as $\pi^+, \pi^-, \pi^0, K^+, K^-, K_L^0$ and photons γ.

ments of the strong coupling constant α_S, have been used to obtain information about the structure of the photon and neutron by constraining parton distribution functions (PDFs) of the proton (as well as differentiate between PDF sets), and they have been used to look for possible deviations from the Standard Model [2, 3, 4].

1.4 Cross Sections and PDFs

The total scattering cross section is computed by convolving the parton distibution function (PDF) for each incoming parton from each proton with the corresponding partonic level cross section. In the language of QCD, the short-distance (high energy) part of the process can be computed using perturbation theory, while long-distance (low energy) part of the process is non-perturbative and is modeled phenomenologically. The Collinear factorization theorem [5] allows us to separate the perturbative (calculable) part of the process from the non-pertubative one, which is described in terms of parton distribution and fragmentation functions. The total cross section of inelastic proton-proton scattering to produce a final state n can be calculated with formula 1.1.

$$\sigma = \sum_{a,b} \underbrace{\int_0^1 dx_a dx_b f_{a/A}\left(x_a, \mu_F\right) f_{b/B}\left(x_b, \mu_F\right)}_{\text{long-distance, non-perturbative PDF part}} \times \underbrace{\int d\Phi_n \frac{1}{2\hat{s}} \left|\mathcal{M}_{ab\to n}\right|^2 \left(\Phi_n; \mu_F, \mu_R\right)}_{\text{short-distance "hard" perturbative part}}, \quad (1.1)$$

where $f_{a/A}(x, \mu)$ denotes the parton distribution functions, which depend on the momentum fraction x of a parton a with respect to its parent hadron A, and on an arbitrary energy scale called the factorization scale μ_F. The quantity of $d\Phi_n$ is the differential phase space element over n final-state particles,

$$d\Phi_n = \prod_{i=1}^{n} \frac{d^3 p_i}{(2\pi)^3 2E_i} (2\pi)^4 \delta^{(4)} \left(p_a + p_b - \sum_{i=1}^{n} p_i\right), \quad (1.2)$$

where p_a and p_b are the initial state momenta. The convolution of the squared matrix element $\left|\mathcal{M}_{ab\to n}\right|^2$, averaged over initial-state spin and colour degrees of freedom, with the Lorentz-invariant phase space Φ_n and multiplied by the flux factor $1/(2\hat{s}) = 1/\left(2x_a x_b s\right)$ yields the parton-level cross section $\hat{\sigma}_{ab\to n}$.

Equation 1.1 illustrates the principle of *factorization* i.e. the approximate separation of short distance and long distance processes. Factorization implies that the PDFs are universal, i.e. process-independent.

1.5 Summary of Short-term and Long-Term Plan for My Thesis

Run 3 of the LHC will start this year, making the timing of my graduate studies well-suited to analyze CMS Run 1 and 2 data, and be one of the first to analyze Run 3 data. Although I have worked and still work on many different projects in Machine Learning (ML), PDFs, the CMS High-Granularity Calorimeter (HGCAL), databases, and prefiring, my PhD dissertation will, in a big picture view, be composed of two parts: a short-term plan, which I

define as my plan up to the next year, and a long-term plan, which is defined as my plan until I graduate, estimated to be 2 or 3 years.

For my short term plan, I am going to be a part of a team at DESY (The German Electron Synchrotron Laboratory in Hamburg, Germany) that will use the full Run 2 dataset to measure the inclusive jet cross section. I plan to go to DESY this summer to work on this project, and our team plans to publish a paper on this measurement next year of which I will be one of the main co-authors.

For my long term plan, which will take the bulk of the my time, I plan to do my own inclusive jet cross section measurement using the early Run 3 data, but I will be using a novel observable for which there is only a single observable per event. This is done so that we can avoid problematic correlations between the different bins and cumbersome bin requirements that the jets must satisfy in typical cross section measurements. Having measured this cross section, we will use the measurement to do a Standard Model Effective Field theory (SMEFT) fit. The novelty in our EFT fit is that this time we will do a true global fit, without making any assumptions about the EFT coefficients. This means that we will fit all the EFT (Wilson) coefficients simultaneously resulting in a full-scale modelling of the probability distributions of the Wilson coefficients. We will later use our fit to search for quark contact interactions. This would give the best chance of saying something general about dimension-6 operators that might set bounds on new physics, as well as searching for contact interactions.

Chapter 2

Previous Projects

2.1 Introduction

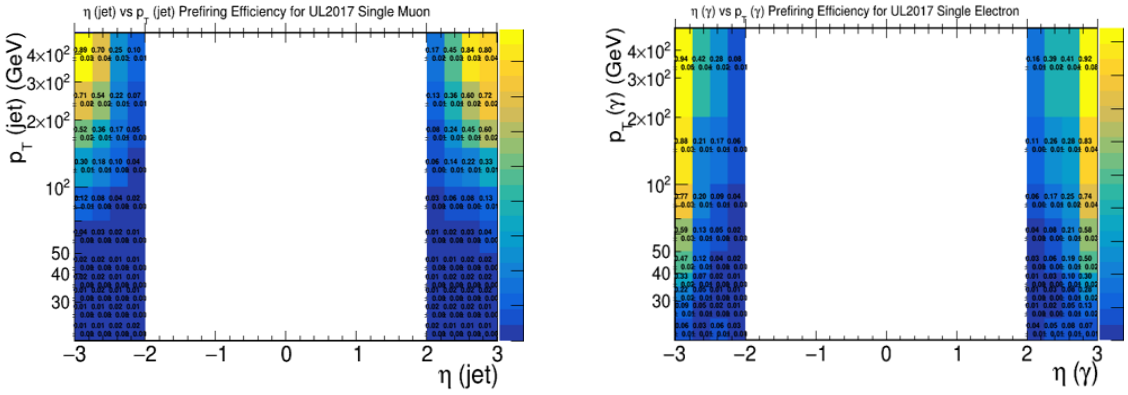
As noted in the previous chapter, I plan to do my own inclusive jet differential cross section measurement using the early Run 3 data using novel observables. This measurement requires many areas of expertise and skills, as it is one of the most difficult and ambitious measurements in modern particle physics. In preparation for this measurement, I have been building my skill sets and expertise in relevant particle physics research areas. Here are a few research projects that I have worked on and their relevance in my preparation for this measurement.

2.2 L1 Prefiring

I have worked on the a problem in the data collected by CMS in 2017 and 2018, called "L1 Prefiring", and have made contributions and discoveries that has impacted the entire CMS experiment [6]. [1] A brief summary of the "prefiring" issue is the following: from the end of 2016, certain jets and photons in the (forward) region ($2 < |\eta| < 3$) were wrongly associated by the Level 1 (L1) Trigger to the previous Bunch Crossing, resulting in a loss of these events. This effect is not accounted for in the CMS Monte Carlo (MC) simulation. Therefore, a correction to simulated events must be applied to model this loss of events. A centrally produced map for the Ultra Legacy (UL) 2017 dataset was produced by me to account for this loss of events. The probability map is in the (p_T, η) plane of the jets or photons, and the probability of "prefiring" can be applied to simulated jets/photons as weights to account for this inefficiency. Figure 2-1 shows an example of the maps that I produced [7]. Working on this issue has given me the necessary tools and expertise and experience on how triggering is done in CMS, analyzing the behaviors of the various detectors in detail, as well as contributing to the experiment in a significant way.

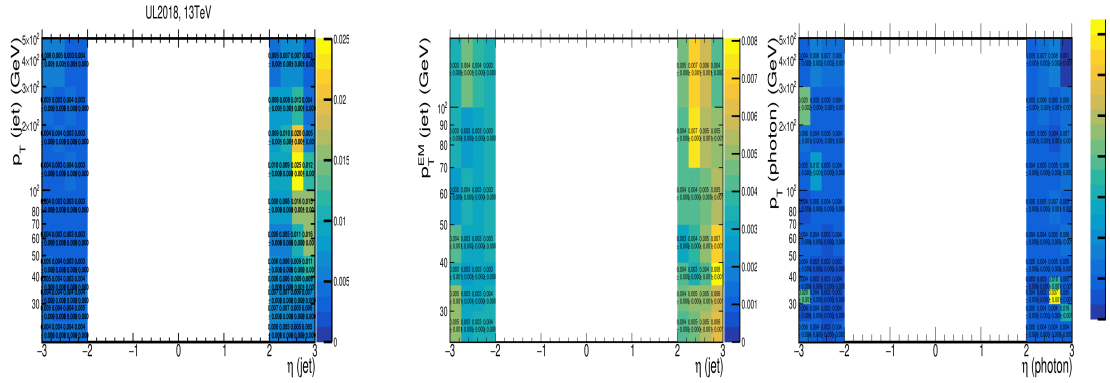
Recently, I was asked to perform this study again but this time for UL 2018 dataset. All scientists involved in these studies expected that very small probabilities of prefiring for the UL 2018 data would be observed, since this problem was assumed to have been fixed. See for example [8], which describes how it was fixed. Indeed, my study concluded that this problem was fixed in the L1 trigger as the probabilities of prefired jets and photons were very low, as shown in Figure 2-2.

[1] I have worked on this analysis as a member of the JETMET POG, defined in the glossary.



(a) (b)

Figure 2-1: Prefiring Maps for Ultra Legacy 2017 dataset, produced and presented to CMS by myself, in the (η, p_T) plane. (a): for jets, (b): for photons.[7]



(a) (b) (c)

Figure 2-2: Prefiring Maps for Ultra Legacy 2018 dataset, produced and presented to CMS by myself, in the (η, p_T) plane. (a)-(b): for jets, (c): for photons [9]

.

However, my studies on prefiring during 2018 showed two things that were unexpected. First, although the probabilities of prefiring were low overall, there was a clear asymmetry in prefiring probabilities for the $\eta > 0$ region compared to the $\eta < 0$ region (previously it was completely symmetric in η, as seen in Fig. 2-1). This asymmetry can be seen in Fig. 2-3 (a)-(b). The other surprising finding was that the residual prefiring was coming from Run A, with a clear structure contained in η/ϕ , as seen in Fig. 2-3 (c)-(d). Furthermore, I found that these effects were coming from two runs numbers in Run A where the L1 trigger misbehaved. My analysis was later fully confirmed by other CMS groups (see [10]) by a different method, and as a result we removed these bad runs from the Golden JSON for 2018, affecting the entire CMS experiment.

There are many uncertainties and inefficiencies that enter the jet cross section measurement, and this is one of the inefficiencies that *must* be accounted for in future measurements. Since I will be using Run 2 Data, correcting for this effect for my short-term plan will be crucial.

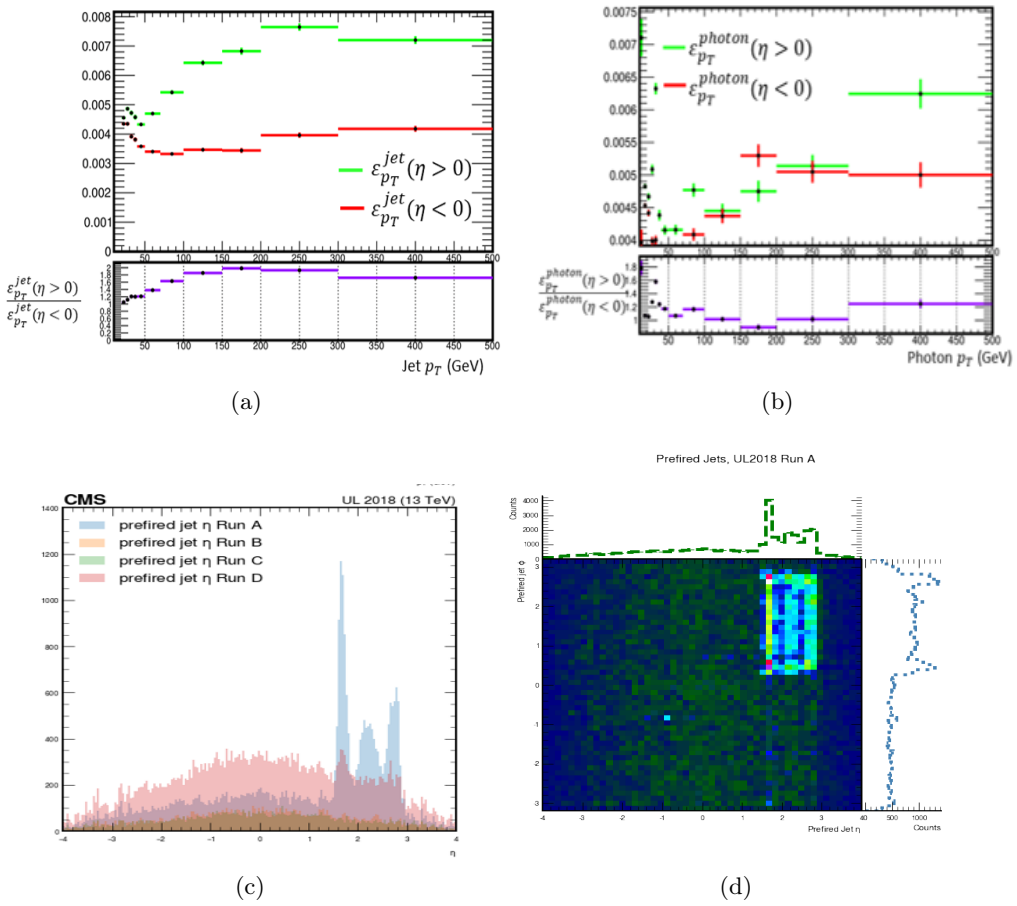


Figure 2-3: Unexpected L1 prefiring that I discovered for the UL 2018 dataset. (a)-(b): Efficiencies of prefiring as a function of p_T for jets and photons, respectively. The probability of prefiring is not symmetric in η, as was expected. (c): jet η distributions for all the involved runs, run A is clearly the run responsible for the prefiring. (d): heat map of prefiring in the (η, ϕ) plane, showing the effect is contained as a detector defect [9]

.

2.3 PDFs and xFitter

As shown in Section 1.4, PDFs are of importance in any cross section measurement at a hadron collider. Since there are many groups that publish PDF sets, a key issue in HEP is the proper characterization of these PDFs and their uncertainties. Understanding the uncertainties of the PDF parameters was the aim of my studies in PDFs, where we used the PDF fitting software xFitter [11, 12].

As discussed below, the reported PDF uncertainties and confidence intervals are based on assumptions that may not be fully justified. The standard method to estimate PDF parameters and their uncertainties is to perform a χ^2 fit. Given a collection of data $\boldsymbol{x}$, and $\boldsymbol{\theta}$, a

vector of unknown parameters, the associated likelihood function is taken to be

$$\begin{aligned} L(\boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}[\boldsymbol{x}-g(\boldsymbol{\theta})]^T\boldsymbol{\Sigma}^{-1}[\boldsymbol{x}-\mathrm{g}(\boldsymbol{\theta})]\right\}, \\ &\equiv \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\chi^2\right\}, \end{aligned} \tag{2.1}$$

where D and $\boldsymbol{\Sigma}$ are the dimension and covariance matrix of $\boldsymbol{x}$, respectively. Taking the logarithm and dropping the $\boldsymbol{\theta}$-independent constants, we have

$$\log L(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2. \tag{2.2}$$

A standard method for estimating the 68% confidence regions is to solve

$$-2\Delta\log L \equiv -2\left[\log L\left(\boldsymbol{\theta}\right) - \log L(\hat{\boldsymbol{\theta}})\right] = T^2, \tag{2.3}$$

where $\hat{\boldsymbol{\theta}}$ is the Maximum Likelihood Estimator of $\boldsymbol{\theta}$, [2] and T is a "Tolerance" factor, defined by $T \equiv \sqrt{\Delta\chi^2}$. For mutually compatible datasets, we should expect to set $T = 1$, which is the choice made by xFitter. However, other groups choose values $T > 1$ [13].

This method is widely used by PDF groups, including the DESY xFitter group. Our study aims at studying to what extent the PDF parameter marginalized densities are normal (which is what PDF groups assume), and whether we can retrieve the marginalized densities, by a reweighting technique. We reweight the parameter points θ_i by the weight $w_i = \frac{L(\theta_i)}{\pi(\theta_i)}$ where $L(\theta_i)$ is the true likelihood for parameter θ at point i and $\pi(\theta_i)$ is a prior of our choice. For example we could take it to be a multivariate Gaussian; $\pi(\theta_i) = \mathcal{N}(\mu = \hat{\theta}_i, \Sigma = \hat{\Sigma})$, where $\hat{\theta}_i$ and $\hat{\Sigma}$ are the best-fit PDF values and their covariance matrix, respectively, that are returned by xFitter. Our method shows promise for elucidating the effect of discrepancies between the different datasets which would affect the normality of the parameter densities [14, 15].

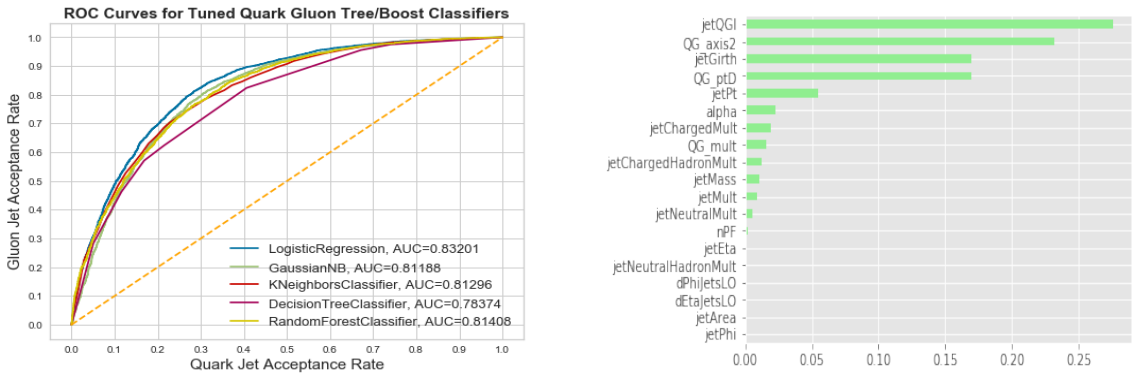
2.4 Quark/Gluon Jet Discrimination

The identification of the origin of jets; whether they are quark or gluon jets, is an extremely important experimental tool in uncovering the physics that occurs in a given event. One of the areas where quark-gluon jet identification is important is in understanding the properties of the Higgs boson. For example, if one wants to measure the Higgs boson coupling to gauge bosons, one would need to look at the weak-boson-fusion process $qq \to Hqq$ (which makes quark jets) and not the more frequent gluon-fusion process $gg \to H$ (which, typically, generates gluon jets). In this project I used CMS data to build machine learning (ML) classification models to distinguish between quark or gluon initiated-jets. Using Baye's Theorem, we can compute the probability (of observing a gluon jet given data x)

$$p(\text{gluon}|x) = \frac{p(x \mid \text{gluon})p(\text{gluon})}{p(x \mid \text{gluon})p(\text{gluon}) + p(x \mid \text{quark})p(\text{quark})}, \tag{2.4}$$

[2] The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ which maximizes the likelihood, i.e. the value of $\boldsymbol{\theta}$ that is obtained when solving $\left.\frac{\partial L(\boldsymbol{\theta};\boldsymbol{x})}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$, where i is the index of the unknown parameter.

which is the outcome of an ML classifier. An ensemble of all the state-of-the art classifiers that are available in common ML packages was studied in application to this problem, and the ML hyperparameters were all tuned to their optimal values using a random grid search. The best performing classifier, as indicated by the ROC curves is shown in Fig. 2-4. Perhaps more interestingly, I found that the jet observable that resulted in the greatest discrimination power between quark and gluon jets was a variable which was constructed on the assumption that gluon jets are wider than quark jets, as seen in Fig. 2-4. See [16] for more details on this variable.



(a) (b)

Figure 2-4: (a): ROC Curves for classifiers whose hyperparameters were tuned with grid search. (b): The jet observables that yield the greatest discrimination power between quark and gluon jets. The winner is the jetQGI variable, which is denoted as σ in [16].

The objective of any classifier is to approximate the function

$$f^* = P(t = 1 \mid x) = \frac{p(x \mid s)\text{prior}(s)}{p(x \mid s)\text{prior}(s) + p(x \mid b)\text{prior}(b)}. \tag{2.5}$$

In other words, the function f (or classifier) in this study approximates the discriminant

$$D(x) = \frac{g(x)}{g(x) + q(x)}, \tag{2.6}$$

where $g(x)$ and $q(x)$ are the gluon jet (signal) and quark jet (background) densities, respectively, assuming that the quark and gluon densities have equal priors. However, this discriminant assumes *pure* probability densities $g(x)$ and $q(x)$. The actual densities of the samples have quark densities that include mixtures of the other parton flavor, and since we are only considering binary classification of quark and gluon jets, the actual densities are the mixture models

$$\begin{aligned} G(x) &= (1 - \epsilon_q)\, g(x) + \epsilon_q q(x), \\ Q(x) &= (1 - \epsilon_g)\, q(x) + \epsilon_g g(x), \end{aligned} \tag{2.7}$$

where the fractions ϵ_g and ϵ_q are mixture fractions corresponding to the two jet flavors. Hence we can define the actual, or *contaminated discriminant* $D'(x)$, which is what we are actually approximating by the classifiers, by

$$D'(x) = \frac{G(x)}{G(x) + Q(x)}, \tag{2.8}$$

which, after substituting $G(x)$ and $Q(x)$, yields

$$D'(x) = \frac{\epsilon_q q(x) + g(x)\,(1 - \epsilon_q)}{\epsilon_g g(x) - \epsilon_g + \epsilon_q q(x) + g(x)\,(1 - \epsilon_q) + 1}. \tag{2.9}$$

Now, what is interesting is that $D'(x)$ can be written in terms of $D(x)$

$$D'(x) = \frac{(1 - 2\epsilon_q)D(x) + \epsilon_q}{1 - \epsilon_g + \epsilon_q + 2(\epsilon_g - \epsilon_q)D(x)}, \tag{2.10}$$

which implies that the contaminated discriminant is a function of D only since ϵ_q and ϵ_g are constants. By studying the relationship between $D(x)$ and $D'(x)$ over the range of ϵ_q and ϵ_g one can identify where the relationship between them is monotonic, that is, one-to-one.

This project was one of my first applications of machine learning to a particle physics problem and it yielded some interesting results, and a possibility of arriving at an optimal quark-gluon classifier using a contaminated classifier.

2.5 Highest p_T Jet Observable

My goal in this project was first to explore and quantify to what extent the transverse momentum (p_T) of highest-p_T jet in an event can be used as an observable. My second goal was to study how we can construct the "response matrix" which is a mapping from "true" or theoretical (or generated - gen for short) values to the measured (or reconstructed - rec for short) values, in the context of the jet p_T. In CMS analyses, this mapping is typically done via [17]

$$p_T^{\text{rec}} = p_T^{\text{gen}} \times (1 + \text{SF} \times \Delta_{\text{MC}} R_G)\,, \tag{2.11}$$

where $\Delta_{\text{MC}} = \frac{p_T^{\text{rec}} - p_T^{\text{gen}}}{p_T^{\text{gen}}}$ is the "resolution", R_G is a random number sampled from a standard normal distribution [3] and SF are smearing factors which are extracted from the measurement of the resolution of the data, and are provided centrally at CMS by the CMS JetMET group.

In our study, we aim to "fold" the gen jet (p_T) spectrum by the response function $R\left(p_T^{rec} \mid p_T^{\text{gen}}\right)$ by calculating the integral

$$f(p_T^{\text{rec}}) = \int R\left(p_T^{\text{rec}} \mid p_T^{\text{gen}}\right) g\left(p_T^{\text{gen}}\right) dp_T^{\text{gen}}, \tag{2.12}$$

where $g(p_T)$ is the true spectrum of the leading jet. The data in this study consists of Pythia 8 (CUEITP8M1) [18] MC samples with $2 \to 2$ parton-parton interactions included in the matrix element (ME): these will be referred to as "gen" or generator (truth) samples. [4] In order to study the detector-related effects, the CMS detector response is also simulated using

[3] Meaning $\mu = 0, \sigma = 1$.

[4] Multiparton interaction, initial state radiation, FSR and hadronization were also simulated in the MC samples.

the GEANT4 package [19]. These will be referred to as the "rec" samples. The distributions for the jet kinematic variables (p_T, η, ϕ) were constructed for ≈ 9 million events, each having a varying number of jets per event. Before we are able to analyze these samples and assess the detector response, we must have a one-to-one mapping of each gen jet to its corresponding rec jet. This is done via matching in (η, ϕ) space, where the distance between a gen (1) jet and a rec (2) jet was chosen to be within a cone of radius $\Delta R_{12} = \sqrt{\Delta\eta_{12}^2 + \Delta\phi_{12}^2} < 0.25$.

In order to determine if we can use the p_T of the highest jet as an observable, we must also calculate the probability that the order of p_T of the rec jets is flipped when ordered according to the p_T order of the gen jets. Ideally, this probability should be small.

If there is a nonzero flipping probability for the first jet, i.e. there is a probability that the leading gen jet corresponds to a non-leading reco jet, the response function has to be modified appropriately. This flipping probability was calculated to be $\approx 4\%$. Therefore, while using the highest-p_T jet observable has the virtue of simplifying the cross section calculation by avoiding problematic correlations between the different bins, it has the disadvantage that the response function needs to be modified to include this flipping probability.

Chapter 3

Detailed Plan

3.1 Traditional Measurement of the Jet Cross Section

My short-term plan entails moving to Germany for this summer and working with the DESY team that will use the full Run 2 dataset to measure the inclusive jet cross section. Furthermore, I plan to measure the inclusive jet differential cross section using the early Run 3 data on my own. The LHC will start collecting data again for Run 3 this year, and as a PhD student, I will be among the first to analyze Run 3 data in the context of an inclusive jet spectrum. In what follows I shall give a brief overview of what measuring this cross section entails from an experimental point of view, as well as all the uncertainties that are involved, and the importance of observables in this measurement.

The inclusive jet cross section is $\sigma(pp \to \text{jet} + X)$, where X signifies "anything". It is usually measured as a function of the jet transverse momentum, p_T, and absolute rapidity $|y|$. Once the samples have been obtained, jets are isolated, reconstructed and corrected, etc., the inclusive double differential jet cross section is measured in bins in the form

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_T\ \mathrm{d}|y|} = \frac{1}{\epsilon\mathcal{L}} \frac{N_{\text{jets}}}{\Delta p_T \Delta|y|}, \tag{3.1}$$

where Δp_T and $\Delta|y|$ are the corresponding p_T and rapidity bin widths, N_{jets} is the number of jets in the corresponding bin, $\mathcal{L}$ is the integrated luminosity of the data sample, and ϵ is the product of all the jet selection and trigger efficiencies. Similarly the inclusive double-differential cross section for the dijet mass is

$$\frac{d^2\sigma}{dM_{jj}d|y^*|} = \frac{1}{\epsilon\mathcal{L}} \frac{N_{\text{jets}}}{\Delta M_{jj} \Delta|y^*|}, \tag{3.2}$$

where M_{jj} is the dijet invariant mass, and $y^* \equiv \frac{|y_1 - y_2|}{2}$, where the subscripts $1, 2$ label the highest and second highest p_T jet in the event.

As discussed earlier, in my short term plan, I will help measure the cross sections using traditional observables, as defined in Eq. 3.1, whereas for my long-term plan, I plan to measure the cross section using an observable for which there is only a single observable per event, i.e. replace, p_T in Eq. 3.1 with this new observable, of which the dijet mass is an example. This is done so that we can avoid problematic correlations between the different p_T bins. It also will make finding signals for contact interactions more feasible. Furthermore, this

obsevable[1] must be one for which predictions can be made at next-to-leading-order (NLO) or next-to-next-to-leading-order (NNLO) in QCD.

Any source of error that affects the p_T, M_{jj} or any other observable is a source of uncertainty. The experimental uncertainties that come into this measurement are numerous and complex (e.g. see [20]). In summary, the experimental uncertainties come from imperfect measurement of jet energy and jet p_T, imprecise simulation of jet energy resolution, imprecise knowledge of integrated luminosity, and the uncertainties in the PDFs.

3.2 SMEFT Fit and Contact Interactions Search

The Standard Model Effective Field Theory (SMEFT) is a generalization of the SM built out of higher dimensional operators, composed of SM fields. The SMEFT is defined as

$$\mathcal{L}_{SMEFT} = \mathcal{L}_{SM} + \mathcal{L}^{(5)} + \mathcal{L}^{(6)} + \mathcal{L}^{(7)} + \ldots, \tag{3.3}$$

where

$$\mathcal{L}^{(d)} = \sum_{i=1}^{n_d} \frac{C_i^{(d)}}{\Lambda^{d-4}} \mathcal{O}i^{(d)} \quad \text{for } d > 4, \tag{3.4}$$

where $\mathcal{O}i^{(d)}$ are the operators, which are suppressed by $d-4$ powers of the cutoff scale Λ and the $C_i^{(d)}$ are the Wilson coefficients. See [21] for a review and [22] for a recent study of a global SMEFT fit. $\mathcal{L}^{(d=6)}$ is the most promising and widely-studied set of operators, [2] and shall be the SMEFT that I shall study in detail.

The cross section would be predicted as a function of the Wilson coefficients, $\boldsymbol{c}$ [22]; $\sigma^{(\text{th})} = \sigma^{(\text{th})}(\boldsymbol{c})$, and the values of the Wilson coefficients could be constrained by comparing the theoretical cross section, $\sigma^{(\text{th})}(\boldsymbol{c})$ with the measured one, $\sigma^{(\text{exp})}$.[3]

3.2.1 Search for Contact Interactions

As part of my long-term plan, I shall focus on studying the dimension-6 operators which could arise from quark compositeness. Quark compositeness models assume that quarks are composed of more fundamental particles with new strong interactions at a compositeness scale Λ, much greater than the quark interaction energies. At energy scales much below Λ,

[1] An example of such an observable is the highest jet p_T, as discussed in Section 2.5.

[2] $d = 6$ is the first higher-dimensional EFT where all the operators are allowed in the SM. Other theoretical and SM gauge symmetry arguments lead to the efficacy of this theory. See [23] for a listing of all the independent $d = 6$ operators allowed in the SM.

[3] This is typically done by minimizing a χ^2 function, which compares the theoretical predictions to the experimental data by means of a covariance matrix, cov. In this case it is defined by

$$\chi^2(\boldsymbol{c}) \equiv \frac{1}{n_{\text{dat}}} \sum_{i,j=1}^{n_{\text{dat}}} \left(\sigma_i^{(\text{th})}(\boldsymbol{c}) - \sigma_i^{(\text{exp})}\right) \left(\text{cov}^{-1}\right)_{ij} \left(\sigma_j^{(\text{th})}(\boldsymbol{c}) - \sigma_j^{(\text{exp})}\right), \tag{3.5}$$

for the i-th cross section. However, instead of using equation 3.5, we plan to use the likelihood function, directly.

4-quark interactions can be approximated as contact interactions (CI)[4], and yield observable signals at hadron colliders.

QCD predicts that jets are preferably produced in large rapidity bins, via small angle scattering in t-channel processes. On the other hand, jet production induced by contact interaction models predict that the region that is most sensitive to CI is the low rapidity region. This is why we should keep the rapidity as an observable in our cross section measurement (i.e. keep $|y|$ in Eq. 3.1), and look for deviations from SM predictions in the low rapidity region.

The theoretical jet cross sections using the jet p_T can be expressed using the CI scale Λ and Wilson coefficients c_i as follows [24]

$$\begin{aligned}\sigma_{\text{bin}}^{\text{th}} = &\sum_{i=1}^{6} \left(\lambda_i \left(b_i + a_i r\right)\right)/\Lambda^2 + \sum_{i=1}^{6} \left(\lambda_i^2 \left(b_{ii} + a_{ii} r\right)\right)/\Lambda^4 \\ &+ \sum_{i=1,3,5} \left(\lambda_i \lambda_{i+1} \left(b_{ii+1} + a_{ii+1} r\right)\right)/\Lambda^4 + \sum_{i=1,2,5,6} \left(\lambda_i \lambda_4 \left(b_{i4} + a_{i4} r\right)\right)/\Lambda^4,\end{aligned} \tag{3.6}$$

where $c_i = 4\pi\lambda_i, r = \ln\left(\Lambda/\mu_0\right)$, and μ_0 is an arbitrary reference scale chosen according to the kinematic range of the p_T bin. I plan to work with theorist Jun Gao to find an observable for which there is one instance per event, and for which NLO and NNLO cross section predictions be made with an expression like that in Eq. 3.6.

To summarize, my long-term plan for measuring the inclusive jet cross section is:

- Measure σ^{exp} with the early Run 3 data, using a good observable in which there is only one instance per event.
- Calculate $\sigma_i^{(\text{th})}(\boldsymbol{c})$ and decompose into bins, using an equation that is analogous to Eq. 3.6 using that observable.
- Fit $\sigma_i^{(\text{th})}(\boldsymbol{c})$ to σ^{exp} using a likelihood function.
- Answer the question: Are there CI's? (if not, set limits on the compositeness scale Λ.)

3.3 Statistical Model and Publishing the Likelihood

All the problems of presenting, analyzing and re-interpreting results discussed in [25] would be avoided if one publishes the full statistical model together with the data. This is because one can then construct the exact likelihood function without the need of any ad-hoc approximations and exploit the complete information offered by CMS data.

The statistical model, given by the probability density $P\left(\textbf{observed} \mid \textbf{parameters}\right)$, relates the **observed** quantities to the **parameters**, describing the prediction in a *model-independent way*, provided that the parameters are independent of the physics model.

[4]This is in analogy to the Fermi theory being a contact interaction approximation to the electroweak theory.

If we publish the full statistical model, and the observed data, the model itself will take care of all non-Gaussian effects. We want to lead a paradigm shift in particle physics: any group that would like to fit a new theory to data with the parameters of the published likelihood mapped to those of the new theory, such as PDF groups, would simply use the the sum of the negative log-likelihoods

$$-\sum_{i=1}^{n_{\text{Datasets}}} \log L_i(\theta_i), \tag{3.7}$$

where n_{Datasets} is the number of datasets considered in the fit. Each experiment/dataset i publishes a likelihood $L_i(\theta_i)$, and a group that wants to do a fit simply minimizes Eq. 3.7. This also shows why the published likelihood should be parameterized in terms of cross sections in order to be model-independent.

There currently exists archives and databases that allows experimental physicists to publish their results. An example of such a database is HEPData [26], and this is the database where I plan to publish my statistical model.

Bibliography

[1] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004–S08004, aug 2008.

[2] Gavin P. Salam. Towards Jetography. *Eur. Phys. J. C*, 67:637–686, 2010.

[3] Sanmay Ganguly and Monoranjan Guchait. Jet cross-section measurements in cms. 6 2013.

[4] The CMS Collaboration. Measurements of differential jet cross sections in proton-proton collisions at s=7 tev with the cms detector. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 87, 6 2013.

[5] John C. Collins, Davison E. Soper, and George F. Sterman. Factorization of Hard Processes in QCD. *Adv. Ser. Direct. High Energy Phys.*, 5:1–91, 1989.

[6] CMS JETMETPOG. `https://twiki.cern.ch/twiki/bin/view/CMS/JetMET`.

[7] Ali Al Kadhim. Ultra legacy 2017 prefiring maps. `https://indico.cern.ch/event/975504/contributions/4108050/attachments/2144017/3613287/JETMET_UL2017PrefiringMaps.pdf`.

[8] Nick Smith. Status and plans for ecal endcap prefiring studies, 2018. `https://indico.cern.ch/event/734407/contributions/3049707/attachments/1676211/2691277/nsmith_PrefireReview_27June2018.pdf`.

[9] Ali Al Kadhim. Ultra legacy 2018 l1 prefiring maps. `https://indico.cern.ch/event/1133830/contributions/4757791/attachments/2398918/4102127/L1DPG_UL2018PrefiringMaps.pdf`.

[10] Laurent Thomas. L1 prefiring in run 315705. `https://lathomas.web.cern.ch/lathomas/L1Prefiring/l1prefiringrun315705.pdf`.

[11] S. Alekhin et al. Herafitter: Open source qcd fit project. *European Physical Journal C*, 75, 7 2015.

[12] V. Bertone, M. Botje, D. Britzger, S. Camarda, A. Cooper-Sarkar, F. Giuli, A. Glazov, A. Luszczak, F. Olness, R. Placakyte, V. Radescu, W. Słomiński, and O. Zenaiev. xfitter 2.0.0: An open source qcd fit framework. 9 2017.

[13] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the LHC. *Eur. Phys. J. C*, 63:189–285, 2009.

[14] Ali Al Kadhim. Pdf uncertainties and extracting likelihoods from xfitter. https://github.com/AliAlkadhim/PDF_Uncertainty/.

[15] Ali Al Kadhim. Extracting the xfitter likelihood. https://indico.ijclab.in2p3.fr/event/7847/contributions/25471/attachments/18431/24586/xFitter_Workshop2022_Alkadhim2_.pdf.

[16] CMS Collaboration. Performance of quark/gluon discriminationusing pp collision data ats=8 tev. 2013.

[17] Patrick LS Connor, Luis Ignacio ESTEVEZ BANOS, Hannes Jung, and IKRadekˇRadekŽLEBRadekŽLEBˇ RadekŽLEBČ. Available on the cms information server cms draft analysis note inclusive jet production at 13 tev with 2016 data.

[18] Pythia team. Pythia. https://pythia.org/documentation/.

[19] Geant Team. Pythia. https://geant4.web.cern.ch/node/1.

[20] The CMS Collaboration. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12, 2017.

[21] Ilaria Brivio and Michael Trott. The standard model as an effective field theory. 6 2017.

[22] Jacob J. Ethier, Giacomo Magni, Fabio Maltoni, Luca Mantani, Emanuele R. Nocera, Juan Rojo, Emma Slade, Eleni Vryonidou, and Cen Zhang. Combined smeft interpretation of higgs, diboson, and top quark data from the lhc. *Journal of High Energy Physics*, 2021, 11 2021.

[23] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek. Dimension-Six Terms in the Standard Model Lagrangian. *JHEP*, 10:085, 2010.

[24] Jun Gao, Chong Sheng Li, Jian Wang, Hua Xing Zhu, and C.-P. Yuan. Next-to-leading qcd effect on the quark compositeness search at the lhc. *Phys. Rev. Lett.*, 106:142001, Apr 2011.

[25] Kyle Cranmer, Sabine Kraml, and Harrison B. Prosper et al. Publishing statistical models: Getting the most out of particle physics experiments. 9 2021.

[26] Durham University. Hepdata. https://hepdata.net.