Prospectus
by
Ali Al Kadhim

Submitted to the Physics Department
on , in partial fulfillment of the
requirements for the degree of
PhD in Physics

Abstract

We live in a very unique and singular time in the history of fundamental physics. On the one hand, we are confronted with deep unanswered questions in fundamental physics and the standard model without being able to find a satisfying answer. On the other hand, our field has developed in its ability to derive theoretical prediction based on the Standard Model (SM) to a mind-boggling precision, and our experiments have developed into the most complex and largest scientific collaborations ever organized by humankind. If we have a chance at answering these exciting questions, we must exploit all the available data and uncertainties that we can from the LHC, and we must aim at a paradigm shift in the field which makes experimental analyses fully replicable and reproducible, and we must exploit the accelerating advances in advanced statistical techniques, machine learning and computer science to aid us in the process. This is what I hope to convey in my ambitious prospectus and hope to contribute to in my coming research.

This document is organized in the following fashion: in Chapter 1 I give a brief overview of all the physics that is necessary for the understanding of this document, and give 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 be directly used in a crucial way in my upcoming thesis plan. Finally, I describe my upcoming research plans in more detail, the novel statistical models and inference that is necessary to develop, and how I can use my results to search for new physics.

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

Contents

Chapter 1

Overview and Motivation

1.1 Introduction, Jet Definition, What a jet cross section is and why it's important

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. 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 their original partons. The interested reader should read reviews on jets, such as [22].

In the CMS experiment, jets are reconstructed from energy deposits (trigger primitives) of final stable particles (the jet constituents with lifetime $c\tau > 1\ cm$) in the calorimeter towers. These energy deposits are then used as input to a clustering algorithm in which a seed particle i sets some initial direction, and one sums the momenta of all particles j within a circle ("cone") of radius R around 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.

After jets have been reconstructed, studying them can offer treasure trove of information about fundamental physics. One of the best best studied phenomena in the LHC related to jets is the inclusive jet spectrum, which is related to the momentum transfer $2 \to 2$ scattering of partons inside the proton. In this process, the energy of a the jet is closely related to that of the hard scattering of partons inside the protons, therefore the inclusive jet spectrum quantifies 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 measuren in e^+e^-, ep, pp, and $p\bar{p}$ colliders. They have been used to test the predictions of perturbative QCD, have given precise measurements 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 theproton (as well as differentiate between PDF sets), and have been used to look for possible deviations from

[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.

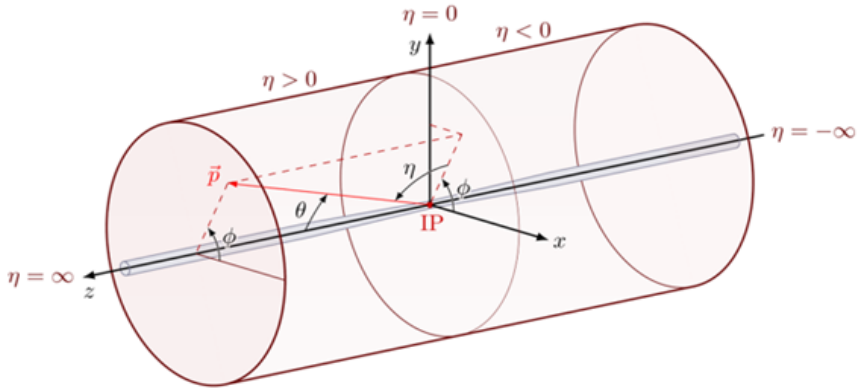


Figure 1-1: In the LHC, the beam direction is assumed to be in the z axis, and the $x-y$ plane is perpendicular to the beam. θ is the angle between the particle momentum diirection and the z direction. A particle with energy E and momentum along the beam direction p_z has rapidity $y \equiv \frac{1}{2} \ln \frac{E+p_x}{E-p_x}$ 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.

the standard model. [1, 22, 16, 12]

1.2 What is a cross section and what are PDFs?

The total scattering cross section is computed by convoluting the parton distibution function for each incoming parton from each proton with the corresponding partonic level cross section. Hence, in the language of QCD, the short-distance (high energy) part of the process can be computed from perturbation theory, and long-distance (low energy) part of the process is driven by the non-perturbative nature of QCD at low-energy scales. Collinear factorization theorem allows us to separate the perturbative (calculable) hard part of the process from the non-pertubative one, which can be described in terms of parton distribution (or fragmentation) functions. The total cross section of inelastic proton-proton scattering to produce a final state n can be calculated with the formula

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

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

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

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 n and multiplied by the flux factor $1/(2\hat{}s) = 1/(2xaxbs)$

results in the calculation of the parton-level cross section $\hat{\sigma}_{ab\ss n}$.

Hence we can intuitively say that the differential cross section in transverse momenta of the observed jet can be factorized in the following form [2]

$$\frac{d\sigma_{jet}}{d\mathcal{O}} \sim \sum_{a,b} \int dx_a f_{a/A}\left(x_A, \mu\right) \int dx_b f_{b/B}\left(x_B, \mu\right) \frac{d\sigma_{partons}}{\mathcal{O}} \tag{1.3}$$

Where $\sigma_{partons} = \int d\Phi_n \frac{1}{2\hat{s}} \left|\mathcal{M}_{ab\to n}\right|^2 \left(\Phi_n; \mu_F, \mu_R\right)$ can be seen from 1.1, and $\mathcal{O}$ is any jet observable for example the jet p_T or the rapidity $|y|$. Equation 1.1 illustrates the principle of *factorization* i.e. that short distance and long distance processes are separable such that they can be convoluted in this manner, so that the "hard part" $\sigma_{partons}$ and "normalizations" from the PDFs are on diffrent scales. Factorization also posits that the PDFs are universal, i.e. process-independent.

The parton level cross sections $d\sigma_{partons}$ has an expansion in powers of α_S

$$\frac{d\sigma_{partons}}{dP_T} \sim \sum_{N} \left(\frac{\alpha_s(\mu)}{\pi}\right)^N H_N\left(x_A, x_B, P_T; a, b; \mu\right) \tag{1.4}$$

Where the coefficients H_N are calculable in perturbative QCD. For more details on the theoretical underpinnings of this, see for example [6]. Equation **??** demonstrates the principle of *Asymptotic Freedom*, i.e. hard scattering is weak at short distances, and hence perturbatively calculable. At next-to-leading-order and beyond, however, the calculation will involve divergences that must be removed, and the dependence on the scale μ will appear in their place.

1.3 Short-term and long-term plan for my thesis

As Run 3 will start next year, the timing of my graduate studies is well-suited to analyze Run 1 and 2 data, as well as be one of the first to analyze Run 3 data. Although I have worked and still work on many different projects in ML, PDFs, HGCAL, database, 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 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 writing that is to follow, I plan to do my own inclusive jet cross section measurement using the preliminary 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 cross problematic correlations between the different channels/bins and cumbersome bin requirements that the jets must satisfy in typical cross

[2] sometimes this is called the "master formula"

section measurements. Having measured this cross section over the full p_T range, we will use that to do an Effective Field theory (EFT) 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. This would give the best chance of saying something about dimension-6 operators that might set bounds on new physics, as well as searching for contact interactions.

Chapter 2

My Previous Projects, and How They are Relevant for My Upcoming Research

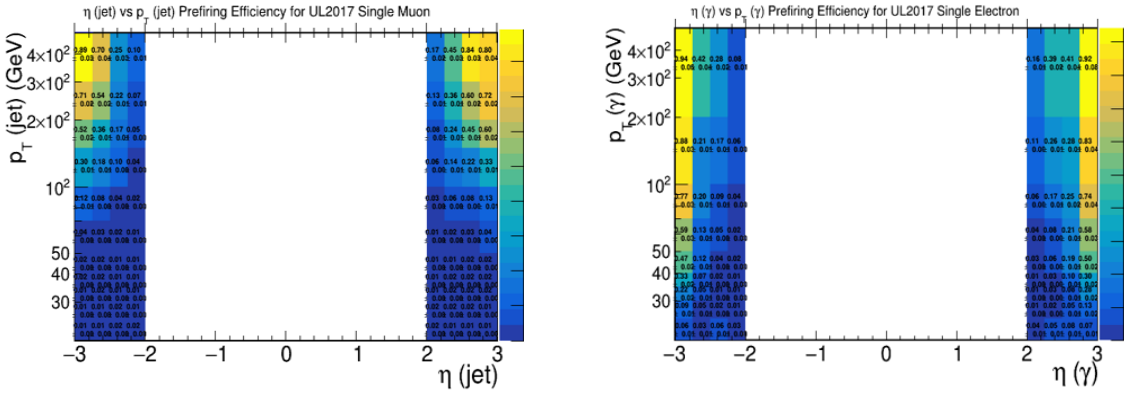
2.1 My previous projects

This measurement requires many areas of expertise and skill, 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've worked on and how they are relevant in my preparation for this measurement.

2.2 L1 Prefiring

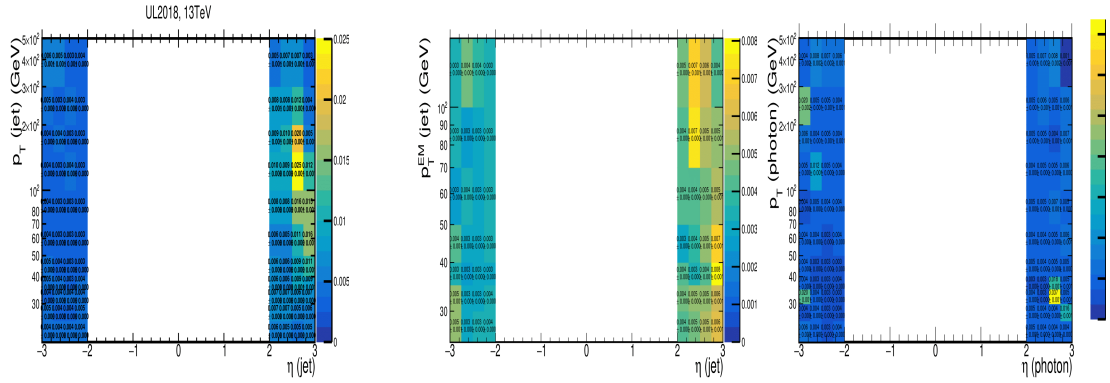
As a member of the JETMET POG [17], I have worked on the L1 Prefiring problem for data collected in 2017 and 2018, and has made contributions and discoveries that has affected the entire CMS experiment. A brief summary of this issue is the following: from the end of 2016, certain jets and photons in the forward rejoin ($2 < |\eta| < 3$) were wrongly associated by the L1 trigger to be belonging to the previous bunch crossing, resulting in a loss of these events. This effect cannot be accounted for in MC, therefore a correction to MC must be applied to compensate for this loss of efficiency. A centrally produced map for the UL 2017 dataset was produced by me to account for this inefficiency. The probability map is in the (p_T, η) plane of the jets or photons, and the probability of prefiring can be applied to MC jets/photons as weights to account for this inefficiency. Figure 2-1 shows an example of the maps that I produced, but the interested reader should read my contributions to the JETMET POG presenting these results [20]. 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 performed this study again but this time for UL 2018 dataset. All scientists that are involved in these studies expected that we would see very small probabilities of prefiring for the UL 2018 data, since this problem was assumed to have been fixed. See for example [23] 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.



(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. The interested reader should also read my JETMET Contribution on this at [20]



(a) (b) (c)

Figure 2-2: Prefiring Maps for Ultra Legacy 2017 dataset, produced and presented to CMS by myself, in the (η, p_T) plane. (a)-(b): for jets, (c): for photons. The interested reader should also read my L1 DPG Contribution on this at [21]

My studies on prefiring during 2018 showed two things that were unexpected. First, although the probabilities of prefiring were low overall, there was clear assymetry in prefiring probabilities for the $\eta > 0$ region compared to the $\eta < 0$ region (previously it was completely symmetric in η previously, as sen in figure 2-1). This can be seen in figure 2-3. The other surprising thing was that the residual evident prefiring was coming from Run A, with a clear η/ϕ structure, as seen in figure 2-3. 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 L1 DPG (see [26]) by a different method, and as a result we removed these bad runs from the Golden JSON for 2018, affecting the entire CMS experiment.

Without the need to mention, there are many uncertainties and inefficiencies that enter the jet cross section measurement, and this is one of the inefficiencies that enter the calculation which *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.

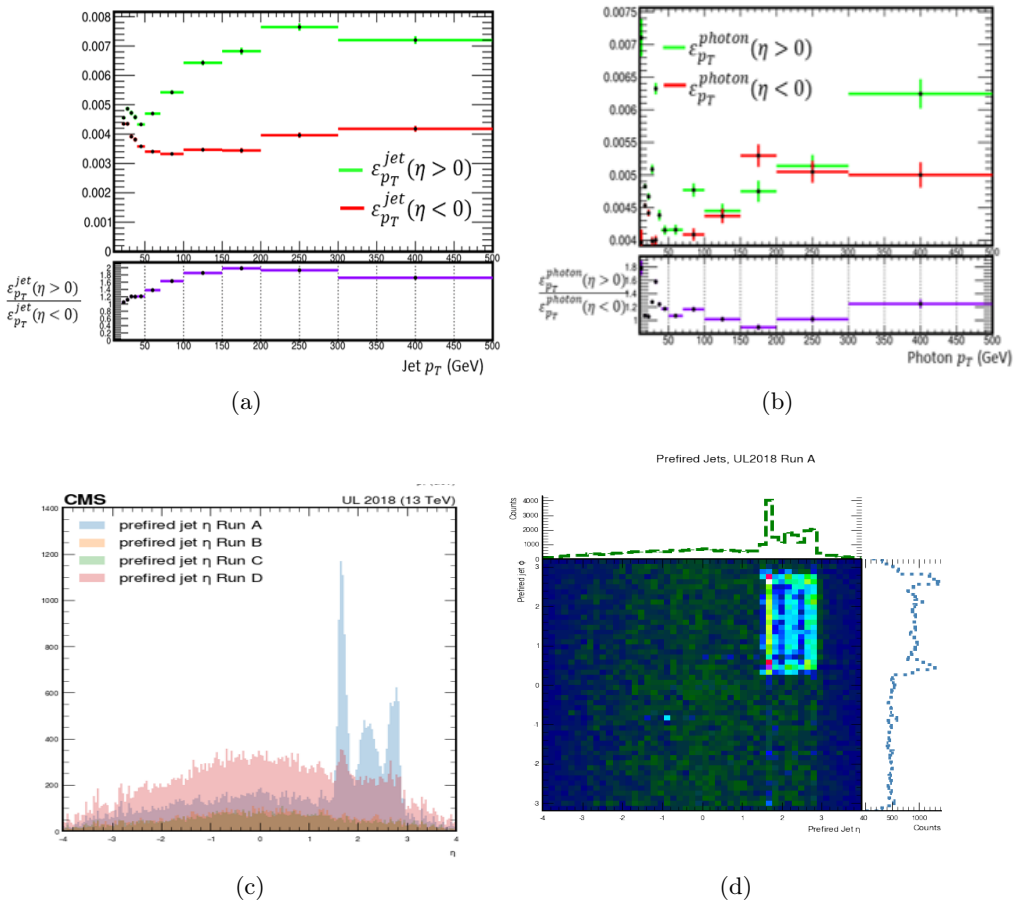


Figure 2-3: Unexpected and Intriguing L1 Prefiring that I discovered for the UL 2018 dataset. (a)-(b): Efficiencies of prefiring as a function of p_T for jets and photons, respectively. The probability of prefiring is not symmetric in η, as was expected. (c): jet η distributions for all the involved runs, run A is clearly the run resposible for the prefiring. (d): heat map of prefiring in the (η, ϕ) plane, showing the effect is contained as a detector defect. The interested reader should also read my L1 DPG Contribution on this at [21]

2.3 PDFs and xFitter

As shown in Section 1.2, PDFs are paramount in any jet cross section measurement and in any cross section measurement. Since there are many groups that routinely publish different PDF sets, a big problem in HEP is the proper characterization of these different PDFs and the quantification of their uncertainties and the uncertainties of the parameters that are included in different PDF sets, which was the aim of my studies in PDFs, where we used the PDF fitting software xFitter [11, 4].

As discussed below, the reported PDF uncertainties and confidence intervals make some assumptions regarding the parameterization of the PDFs, which may not be statistically justified. A popular method for estimating the PDF errors is by constructing confidence intervals on the parameters using $-2 \log L$: suppose that the likelihood for θ is multivariate

normal, the likelihood function of a single observation is of the form

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

Taking the log and dropping the constant in the beginning (since the constant is independent of θ), we have

$$\begin{aligned} \log L(\boldsymbol{\theta}; \boldsymbol{x}) &= -\frac{1}{2} [\boldsymbol{x} - \mathrm{g}(\boldsymbol{\theta})]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} [\boldsymbol{x} - \mathrm{g}(\boldsymbol{\theta})] \\ &= -\frac{1}{2} \chi^2 \end{aligned} \tag{2.2}$$

Thus, $-2 \log L$ is precisely the χ^2 expression in the leaset squares method. A popular method for estimating the errors in the maximum likelihood method is to look for parameters $\theta_\pm$ for which

$$-2\Delta \log L \equiv -2 \left[\log L \left(\boldsymbol{\theta}_\pm; \boldsymbol{x} \right) - \log L(\hat{\boldsymbol{\theta}}; x) \right] = 1 \tag{2.3}$$

This method yields 68 % confidence intervals on the individual parameters. These shifts in the parameters $\boldsymbol{\theta}_\pm$ can be attained with the maximum likelihood methodology. In other words, finding the points where $\Delta\chi^2 = 1$ corresponds to the method of finding the 68 % confidence interval $(\hat{\theta} - \theta_+, \hat{\theta} + \theta_+)$.

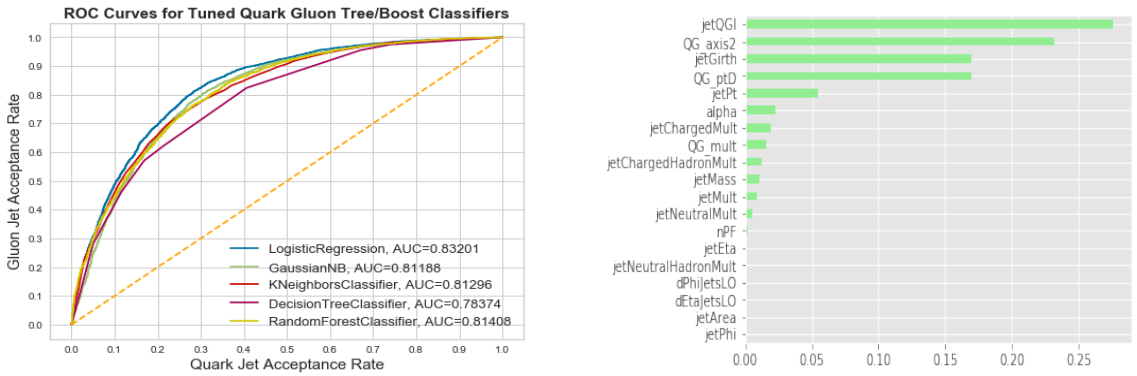
This method is widely used by all PDFs, but it is critical to remember that this assumes the normal sampling of x. If the sampling is not normal, then the probability content will be different, and must be determined for the correct sampling distribution. Since this method is used by xFitter and most other PDF fitting tools, we aim to study this assumption, whether it is true that the parameters that we aim to study are normally distributed.

Our study aims at studying to what extent the PDF parameter marginalized densities (distributions) are normal (which is what PDF groups imply as discussed above), and whether we can retreive the actual parameter likelihoods, which might not be Gaussian, 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 or approximate likelihood of our choice, which assume $\Delta\chi^2 = 1$. 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 that are returned by xFitter, respectively. Our method shows promise for working in the limit of increasing datasets included in the PDF fit, due to the descrepancies between the different experiments which would affect the noramlity of the parameter distributions. For more details, please visit my Github repository, where all the code is available [19], and to my invited talk at the 2022 xFitter Workshop [18].

2.4 Quark/Gluon Jet Discrimination with ML

The identification the origin of jets; whether they are quark or gluon jets, is an extremely important experimental test in uncovering the fundamental 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. 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 $gg \to Hgg$ (which generates a gluon jet). In this project I used CMS data to build machine learning classification models to classify between quark or gluon initiated jets. Using Baye's Theorem, $p(y \mid x) = \frac{p(x|y)p(y)}{p(x)}$ we therefore have the probability of observing a gluon jet given data x as $p(\text{ gluon } |x) = \frac{p(x|\text{ gluon })p(\text{ gluon })}{p(x|\text{ gluon })p(\text{ gluon })+p(x|quark)p(\text{ quark })}$ which is the outcome of our 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 random grid search. The best performing classifier, as indicated by the ROC curves in figure 2-4. Perhaps more interestingly, I found that the jet observable that resulted in the greatest discrimination power between quark and gluon jets was the jetQCI variable, which was constructed on the basis that gluon jets are wider than quark jets, as seen in figure 2-4. See also [7] for more details.



(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 [7].

This study was later extended to consider a more realistic model for quarks and gluons. Recall that the objective of any classifier is to approximate the function

$$f^* = P(t = 1 \mid x) = \frac{p(x \mid s)\text{prior}(s)}{p(x \mid s)\text{prior}(s) + p(x \mid b)\text{prior}(b)} \tag{2.4}$$

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

$$D(x) = \frac{s(x)}{s(x) + b(x)} \tag{2.5}$$

However, this discriminant assumes ideal and *pure* probability densities $g(x)$ and $q(x)$. The actual densities of the samples have quark densities that include mixtures of the other quark flavor, and since we are only considering binary classification of quark and gluon jets, the actual densities are

$$\begin{aligned} G(x) &= (1 - \epsilon_q)\, g(x) + \epsilon_q q(x) \\ Q(x) &= (1 - \epsilon_g)\, q(x) + \epsilon_g g(x) \end{aligned} \tag{2.6}$$

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, which includes the mixture fractions

$$D'(x) = \frac{G(x)}{G(x) + Q(x)} \tag{2.7}$$

Which, after subsitituting G and $Q(x)$ and simplifying

$$D'(x) = \frac{\epsilon_q q(x)_g(x)\,(1-\epsilon_q)}{\epsilon_g g(x) - \epsilon_g + \epsilon_q q(x) + g(x)\,(1-\epsilon_q) + 1} \tag{2.8}$$

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

$$D' = D\frac{(g+q)(eqq + g(1-eq))}{g(egg - eg + eqq + g(1-eq) + 1)} \tag{2.9}$$

This means that the realistic contaminated discriminant is a function of D only since ϵ_q and ϵ_g are constants. Furthermore, the fact that D' is one-to-one with D means that their discrimination power is the same. This means that the optimal discriminant D can be computed from D', where the contamination fractions are given in D' , and by studying the relationship between D and D' over the range of ϵ_q and ϵ_g one could identify where the relationship between them is monotonic.

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 by the contaminated classifier technique described above. It also taught me the power of machine learning in jet studies and how one has freedom to construct jet observables, which becomes all the more powerful if it is informed by intuition and machine learning. Constructing similar observable will be very important for my thesis plan.

2.5 Highest p_T Jet Observable and the Response Matrix

My goal first in this project was to explore and quantify to what extent the p_T of highest-p_T jet in an event can be used as an observable. Further discussion on why we want to use this observable in jet cross section measurement is described in section 3.2. 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 [8]

$$P\left(p_T^{\text{rec}} \mid p_T^{gen}\right) = P\left(p_T^{\text{gen}} \mid \text{SF}\right) \times (1 + \text{SF} \times \Delta_{\text{MC}} R_G) \tag{2.10}$$

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 [1] and SF are smearing factors which are extracted from the measurement of the resolution of the data, and are provided centrally at CMS by JetMET. We argue that this mapping makes unproven assumptions (such as the normality of the smearing) and

[1] meaning $\mu = 0, \sigma = 1$

it requires exact knowledge of the smearing factors, which are nuissance parameters, making it very inconvenient.

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

$$P\left(p_T^{\text{rec}} \mid p_T^{gen}\right) = \int R\left(p_T^{rec} \mid p_T^{gen}\right) f\left(p_T\right) dp_T \tag{2.11}$$

Where $f(p_T)$ is the true spectrum of the leading jet. The Data is consistent of pythia 8 (CUEITP8M1) [25] MC samples with $2 \rightarrow 2$ parton-parton interactions included in the matrix element (ME): these will be referred to as "gen" or generator (truth) samples. [2] In order to study the detector-related effects, the CMS detector response is also simulated using the GEANT4 package [24]. 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 to further 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 nto be within a cone of radius $\Delta R_{12} = \sqrt{\Delta\eta_{12}^2 + \Delta\phi_{12}^2} < 0.25$.

In order to study the kinematic variables of the reco and gen jets, we can use the p_T ordering of the jets as a means to quantify whether the p_T of the highest jet can be used as an observable. One way to tell whether it can be used as an observable is to calculated the probability of flipping of the order of the gen jets compared to rec jets, $P_{flip}(gen, rec)$: what is the probability that the order of the p_T of the rec jets is flipped when ordered according to the p_T order of the gen jets?

Hence, if we show that 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 in order to be compared to the provided theoretical distribution by multiplying the integral by a flipping probability

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

For our task, we want to find the conditional distribution $P\left(p_T^{rec} \mid p_T^{gen}\right)$. This can be done via a Machine Learning and a "likelihood ratio trick" (see for example [9] which also uses this in a HEP application). Given a classification model, which tries to find the mapping $p_T^{rec} = f(p_T^{gen})$, which can be estimated as $\hat{f}$

$$f(p_T^{rec}) \approx \hat{f}(p_T^{gen}) = \arg_{\hat{f}} \min \int L(p_T^{rec}, \hat{f})\ p(p_T^{rec}|p_T^{gen}) dy \tag{2.13}$$

Where $L(p_T^{rec}, \hat{f})$ is a (binary classification) loss function.

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

Hence our task is to approximate the optimal classifier, which, given training data x learns

$$d(x) = \frac{P\left(x|p_T^{rec}\right)}{P\left(x|p_T^{gen}\right) + P\left(x|p_T^{rec}\right)} = \frac{f(x)}{1+f(x)} \tag{2.14}$$

Where f is the classifier decision function, Which is 1-to-1 with the likelihood ratio,

$$\hat{r}(x|p_T^{gen}, p_T^{rec}) = \frac{P\left(x|p_T^{rec}\right)}{P\left(x|p_T^{gen}\right)} \tag{2.15}$$

Which can be written as

$$\hat{r}(x|p_T^{gen}, p_T^{rec}) = \frac{P\left(x|p_T^{rec}\right)\ P(p_T^{rec})}{P\left(x|p_T^{gen}\right)\ \ P(p_T^{rec})} = \frac{d(x)}{1-d(x)} \tag{2.16}$$

Our posterior density could finally be calculated as

$$P\left(p_T^{rec} \mid p_T^{gen}\right) = \frac{d}{1-d} \times f\left(p_T^{rec}\right) \tag{2.17}$$

In building our probabilistic machine learning model, we have the target class corresponding to the true p_T^{gen} and the normalized distribution of the rec jets $Z = \frac{p_T^{rec}}{p_T^{gen}}$

$$\left\{ p_T^{\text{gen}}, Z = \frac{p_T^{rec}}{p_T^{\text{gen}}} \right\},\ \text{Target}\ = 1 \tag{2.18}$$

And the other class corresponding to a distribution that is randomly sampled from p_T^{gen} and a random sampling of the Z distribution, $f(Z)$:

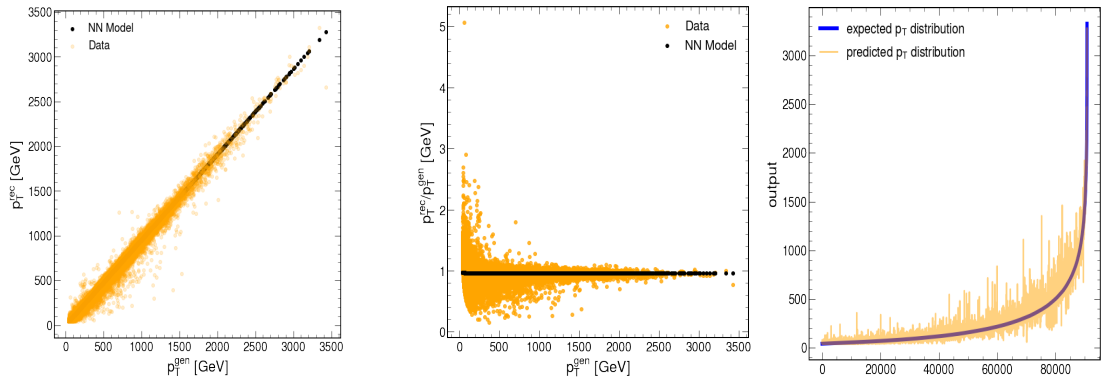
$$\left\{ f\left(p_T^{gen}\right), f(Z) = f\left(\frac{p_T^{rec}}{p_T^{gen}}\right) \right\},\ \text{Target}\ = 0 \tag{2.19}$$

We pick $f(Z)$ to be randomly sampled from a Gaussian with mean $\mu = \frac{p_T^{rec}}{p_T^{\text{gen}}}$ and standard deviation corresponding to the width of the Z distribution, i.e.

$$f(Z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{p_T^{rec} - p_T^{gen}}{\sigma}\right)^2} \tag{2.20}$$

Therefore the function $f(Z)$ can be regarded as a rough approximation of the distribution Z, and given this, $\frac{d}{1-d}$ fixes the approximation.

The posterior distribution from our model in equation 2.17 was compared to the simulated data which corresponds to the true CMS detector response. These "data" were simulated with Delphes 3 [10] by passing the truth distribution (which was generated with Pythia) to Delphes to reconstruct the jets given the detector response and smearing, using the CMS card. Results using this method is shown in Figure 2-5.



(a) (b) (c)

Figure 2-5: (a): Response matrix for higheset-p_T jet for matched jets, (b): $p_T^{\text{rec}}/p_T^{\text{gen}}$ ratio vs p_T^{gen} for data and NN model prediction, (c): Expected and NN-predicted distributions of our inference model.

Chapter 3

Measuring Jet Cross section for Run 3, Doing EFT Fit and Searching for Contact Interactions

3.1 Doing the Jet Cross section for Run 3 and The Relevant Uncertainties

As mentioned earlier, I plan to measure the inclusive jet differential cross section for the preliminary Run 3 data. The LHC will start collecting data again for Run 3 in early 2023. 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 is measuring this cross section means 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 $|y|$, hence a common measurement inclusive jet double-differential cross section is $\frac{d^2\sigma}{dp_T dy}$.

Once my samples for the preliminary Run 3 have been obtained, jets are isolated, reconstructed and corrected, ensured to pass criteria, etc., The measured inclusive double differential jet cross section looks like

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

Where Δp_T and $\Delta|y|$ are the corresponding bins widths. The rapidity intervals are chosen, for example $\Delta|y| = 0.5$, N is the number of jets in the corresponding p_T bin, $\mathcal{L}$ is the effective integrated luminosity of the data sample taking into account the trigger prescales, 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.

Any source of error that affects the p_T, M_{jj} or any other observable that might affect the studied jet cross section must be discussed as a source of uncertainty. The experimental uncertainties that come into this measurement are numerous and complex, and could constitute a PhD thesis on their own (eg see [13]). 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. These inefficiencies are taken to be factorized into multiple uncertainties that are important to discuss.

3.2 The importance of Jet Observables and Defining Useful Ones

All cross section measurements in hadron colliders can be described by the "master formula", such as equation 1.1 discussed earlier. We can write a simpler version describint a scattering process where we have beams A and B coming in and final states 1 2 3 ... n coming out. If we want to compute the cross section for some observable σ_{obs} a simplified master formula, assuming the beams are massless, is

$$\sigma_{obs} = \frac{1}{E_{CM}^2} \sum_{n=1}^{\infty} \int d\phi_n \left|\mathcal{M}_{AB\to 12\ldots n}\right|^2 f_{obs}\left(\phi_n\right) \tag{3.3}$$

where E_{CM} is the center of mass energy, the sum is over all the final states, and the integral is over the Lorentz-invariant n-body phase space ϕ_n (i.e. over everywhere these particles go). $|\mathcal{M}_{AB\to 12\ldots n}|$ is the hard scattering amplitude, and $f_{obs}\left(\phi_n\right)$ is what we choose as an observable, which depends on the phase space or kinematics of the outgoing particles. The problem for jets is that the final states we measure, 1 2 3 .... n, are hadrons, but the scattering amplitudes that we can calculate, $\mathcal{M}$ are in terms of quarks and gluons. Therefore we have to somehow bridge the divide between the types of calculations whose scattering amplitudes we can calculate perturbatively, and the types of measurements that we can take which are composed of hadronic final states. The key to bridging this gap is the choice of observables in an appropriate "factorizable" way.

Suppose we chose our observable to be something like the dijet invariant mass M_{jj} or the dijet average p_T^{avg}, then the question becomes whether will we define the spectrum such that we will restrict the jets to be in the same rapidity bin or not.

3.3 SMEFT Fit and Contact Interactions Search

The Standard Model Effective Field Theory (SMEFT) is a consistent effective field theory generalization of the SM built out $\mathrm{SU_c}(3) \times \mathrm{SU_L}(2) \times \mathrm{U_Y}(1)$ higher dimensional operators, composed of SM fields. The SMEFT is defined as

$$\mathcal{L}_{SMEFT} = \mathcal{L}_{SM} + \mathcal{L}^{(5)} + \mathcal{L}^{(6)} + \mathcal{L}^{(7)} + \ldots \tag{3.4}$$

Where

$$\mathcal{L}^{(d)} = \sum_{i=1}^{n_d} \frac{C_i^{(d)}}{\Lambda^{d-4}} Q_i^{(d)} \quad \text{for } d > 4 \tag{3.5}$$

Where $Q_i^{(d)}$ 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 e.g. [5] for a recent study of a global SMEFT fit.
Why $d = 6$ is the most interesting set of operators.

It is important to note that a constraint that physial cross sections are semi-positive definite quantities, which can be accounted for in global SMEFT analyses. Consider, for example, the SMEFT Lagrangian

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \sum_{i=1}^{n_{\text{op}}} \frac{c_i}{\Lambda^2} \mathcal{O}_i \tag{3.6}$$

Where $\mathcal{O}_i$ are the dimension-6 operators and c_i are the Wilson coefficients, which are assumed to be real. Then any observable, such as the cross section, calculated using this Lagrangian can be written as the expansion

$$\begin{aligned} \sigma = & \underbrace{c_0^2 \sigma_{00}}_{\text{SM contribution}} \\ & + \underbrace{c_0 c_1 \sigma_{01} + c_1 c_0 \sigma_{10} + c_0 c_2 \sigma_{02} + \dots}_{\text{linear } \mathcal{O}\left(\Lambda^{-2}\right) \text{ EFT contributions}} + \underbrace{c_1^2 \sigma_{11} + c_1 c_2 \sigma_{12} + c_1 c_3 \sigma_{13} + \dots}_{\mathcal{O}(\Lambda^{-4}) \text{contributions}} \\ & = \boldsymbol{c}^T \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{c} \end{aligned} \tag{3.7}$$

And since the physical cross section must be either positive or null, the matrix $\boldsymbol{\Sigma}$ must be semi-positive-definite. The Sylvester criterion requires the constraints coming from the 2×2 minors as $\left(\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2\right) \geq 0, \quad i, j = 0, \dots n_{\text{op}}$. Using this convention for the dimension-6 SMEFT, a genereic LHC cross section will be modified as

$$\sigma_{\text{LHC}} = \sigma_{\text{SM}} + \sum_{i=1}^{n_{\text{op}}} c_i \sigma_i^{\text{EFT}} + \sum_{i<j}^{n_{\text{op}}} c_i c_j \sigma_{ij}^{\text{EFT}} \tag{3.8}$$

The quality of the global fit as well as the statistical model will be discussed in detail in the next chapter, but let us suppose that we are using a χ^2 function as is typically used.

The quality of a global fit can be obtained by minimizing the log-likelihood, or the χ^2 function, which compares the theoretical predictions to the experimental data by means of a covariance matrix. In this case is defined by

$$\chi^2(\boldsymbol{c}) \equiv \frac{1}{n_{\text{dat}}} \sum_{i,j=1}^{n_{\text{dat}}} \left(\sigma_i^{(\text{th})}(\boldsymbol{c}) - \sigma_i^{(\text{exp})}\right) \left(\text{cov}^{-1}\right)_{ij} \left(\sigma_j^{(\text{th})}(\boldsymbol{c}) - \sigma_j^{(\text{exp})}\right) \tag{3.9}$$

Where $\sigma_i^{(\text{exp})}$ and $\sigma_i^{(th)}(\boldsymbol{c})$ are the central experimental data and corresponding theoretical prediction, which depends on the Wildon coefficients $\boldsymbol{c}$, for the i-th cross section. Note that equation 3.9 is precisely *not* what we want to do! We are including it here since this is what is currently done in all such fits. See Chapter 4 for a more on this.

A theoretical cross section depends on all the Wilson coefficients, and the exact dependence on each coefficient, say the EFT coefficient c_j can be found by varying c_j while setting the rest of the coefficients to zero, so that

$$\sigma_m^{(\mathrm{Th})}(c_j) = \sigma_m^{(\mathrm{SM})} + c_j \sigma_{m,j}^{(\mathrm{EFT})} + c_j^2 \sigma_{m,jj}^{(\mathrm{EFT})} \tag{3.10}$$

which results in a quartic polynomial form for the χ^2, which can be expanded as $\chi^2(c_j) = \sum_{k=0}^{4} a_k (c_j)^k$ and minimized to obtain the best-fit value of the coefficient, $c_{0,j}$. See also [14] for more details.

Note however, that equation 3.9 assumes that the data are Gaussian distributed around the true SM values, which may or may not be a good approximation. This also assumes that the covariance matrix can be expressed as a sum of separate experimental and theoretical covariance matrices, and hence that the experimental and theoretical uncertainties are uncorrelated

$$\mathrm{cov}_{ij} = \mathrm{cov}_{ij}^{(\mathrm{exp})} + \mathrm{cov}_{ij}^{(\mathrm{th}\ )} \tag{3.11}$$

Where the experimental covariance matrix is constructed from all sources of statistical and systematic uncertainties that are made available by the experiments. EFT theorists therefore use only the full covariance matrix, without details of its individual components. This complicates EFT fit reinterpretations, because the covariance matrices have to be modifidied to include more data, additional uncertainties, etc. Furthermore, the provided covariance matrix is not necessarily positive-definite or it is ill-defined. In this case, the dataset is either discarded for the fit or its covariance matrix is regularized by some ad-hoc procedure. All these problems (there's more) would be avoided if one publishes the full statistical model together with the data, since 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.

3.3.1 Search for Contact Interactions

Quark compositeness models assume that quarks are composed of more fundamental particles with new strong interactions at a composite scale Λ, much greater than the quark masses. At energy much below Λ, quark contact interactions (CI) are induced by the underlying strong dynamics, and yield observable signals at hadron colliders. It is modeled that the contact interactions are induced by an effiective $d = 6$ Lagrangian, given by

$$\mathcal{L}_{CI} = \frac{1}{2\Lambda^2}(c_1 O_1 + c_2 O_2) \tag{3.12}$$

Where here it is considered that the quark contact interactions are products of left-handed electroweak isoscalar quark currents which are assumed to be flavor-symmetric to avoid large flavor-changing neutral-current interactions [2, 15], c_1, c_2 are the Wilson coefficients, and O_1, O_2 are the relevant four-fermion operators.

1

[1] $O_1 = \delta_{ij}\delta_{kl}\left(\sum_{c=1}^{3} \bar{q}_{Lci}\gamma_\mu q_{Lcj} \sum_{d=1}^{3} \bar{q}_{Ldk}\gamma^\mu q_{Ldl}\right)$, $O_2 = \mathrm{T}^a_{ij}\ \mathrm{T}^a_{kl}\left(\sum_{c=1}^{3} \bar{q}_{Lci}\gamma_\mu q_{Lcj} \sum_{d=1}^{3} \bar{q}_{Ldk}\gamma^\mu q_{Ldl}\right)$ where c, d are generation indices, i, j, k, l, a are color indices, T^a are the Gell-Mann matrices with the nor-

In the SM, QCD predicts that jets are preferabbly produced in large rapidity bins, via small angle scattering in t-channel processes. Furthermore, it predicts that the p_T or M_{jj} spectrum falls off very fast at large values. On the other hand, jet production induced by contact interactions (such as $qq \to qq$) is expected to be much more isotropic and fall off slower with increasing values of p_T or M_{jj}. CI models also predict that the region that is most sensitive to CI are low rapidity region. This is why it was shown that the inclusive jet p_T and the dijet anglular distribution, show a great sensetivity to possible quark contact interactions induced by SMEFT (new physics) models. The dijet angular variable is $\chi_{\text{dijet}} = \exp(2y^*)$, where $y^* = \frac{1}{2}|y_1 - y_2|$ where $\pm y^*$ are the rapidities of the two jets in the parton-parton c.m. frame [3].

The theoretical jet cross sections could be calculated using the CI scale Λ and Wilson coefficients c_i, and this cross section, interestingly, could be decomposed into a cross section per each kinematic bin

$$\begin{aligned}\sigma_{bin}^{theor} = &\sum_{i=1}^{6} \left(\lambda_i \left(b_i + a_i r\right)\right)/\Lambda^2 + \sum_{i=1}^{6} \left(\lambda_i^2 \left(b_{ii} + a_{ii} r\right)\right)/\Lambda^4 \\ &+ \sum_{i=1,3,5} \left(\lambda_i \lambda_{i+1} \left(b_{ii+1} + a_{ii+1} r\right)\right)/\Lambda^4 + \sum_{i=1,2,5,6} \left(\lambda_i \lambda_4 \left(b_{i4} + a_{i4} r\right)\right)/\Lambda^4\end{aligned} \tag{3.13}$$

Where $c_i = 4\pi\lambda_i, r = \ln(\Lambda/\mu_0)$, and μ_0 is an arbitrary reference scale chosen according to the kinematic range of the bin. Choosing an observable in which there is only one instance per event, as mentioned earlier, makes the calculation of the cross section parameterized by EFT coefficients, as in equation 3.13 much more feasible, and would avoide the cumbersome requirements (e.g. whether the jets are required to be in the same rapidity bin or not).

Therefore, my plan for measuring the inclusive jet cross section could be roughly outlined by the following steps:

- Measure σ^{exp} with a good observable in which there is only one instance per event
- calculate σ^{theor} e.g. from equation 3.13 using that observable
- Fit σ^{theor} and σ^{exp} to set limits on Wilson coefficients
- Answer the question: Are there CI's? (set limits on Λ in equation 3.12).

malization $\text{Tr}\left(\text{T}^a\ \text{T}^b\right) = \delta^{ab}/2$.

Chapter 4

Statistical Model and Publishing the Likelihood

4.1 Why Publish the Likelihood?

The whole point of science is to be reproducible, replicable and verifiable ... more on motivation why this must be done.

Unfortunately, fully exploiting the information provided by the LHC on any measurement is often restrained by publicly available statistical information provided by the experimental collaboration. Such vital statistical uncertainties may sometimes even be difficult to access for collaboration members. This leads to global fits, such as PDF fits, having to use a χ^2 as a figure of merit to compare theory with experiment. However, equation 3.9 includes several downfalls. For example, 3.9 is itself an approximation that the data is normally distributed around their true theoretical values, which may or may not be a good approximation. Further, usually only the full covariance matrix is provided, without details of its individual components. This means that one cannot identify the relevant sources of systematic errors causing the problem. it does not separate separate the different sources of error, and only the total systematic error is provided and information about correlation is missing. There are more problems associated with this approach (see [9] for a longer review and discussion on this).

A much more complete, useful and general solution to this is to provide the statistical model of an analysis (or at least the likelihood function). The statistical model, given by the probability density $P(\textbf{observed}|\textbf{parameters})$, related the observed quantities **observed** to the parameters **parameters**, describing the prediction in a *model-independent way*. Then if a full mathematical description of the final likelihood is provided, one can change the **parameters** based on their predictions, to arrive at a new prediction for the same published observed quantities.

If we publish the likelihood, the the full details of the observed data, e.g. n_{dat} bin-by-bin correlated errors, both statistical and systematic, are provided, in which case the covariance

matrix can be schemattically constructed as

$$\mathrm{cov}_{ij} = \delta_{ij}\sigma_i^{(\mathrm{stat})}\sigma_j^{(\mathrm{stat})} + \sum_{k=1}^{n_{\mathrm{sys}}} \sigma_i^{(\mathrm{sys})(k)}\sigma_j^{(\mathrm{sys})(k)}, \quad i,j = 1,\ldots,n_{\mathrm{dat}} \tag{4.1}$$

And the likelihood itself will take care of all manner of non-Gaussian effects, and would avoid the problematic way fitting is done as mentioned earlier. We want to do a paradigm shift in particle physics: any group that would like to fit the experiment to theory, such as PDF groups, would simply use the the sum of the negative log-likelihoods

$$-\sum_{i=1}^{n_{\mathrm{Datasets}}} \log L_i(\theta_i) \tag{4.2}$$

where n_{Datasets} is the number of datasets (or experiment) that is 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 4.2, as opposed to a χ^2 such as the one in equation 3.9.

4.2 Statistical Model

Supose that a set of N_c independent channels for events are defined, in my case the channels could be distinct physics channels (processes leading to jets) or bins of a histogram. A certain number of events n_i is found in channel i and for every event j in that channel one measure a vector of values x_{ij}. Suppose the n_i counts can be modeled as independent and Poisson-distributted with mean counts $\nu_i(\mu,\theta)$ where μ are parameters of interest and θ are nuissance parameters. In particle physics we typically care about the cross section as a parameter of interest, and it is related to the mean count by

$$\nu = \varepsilon\mathcal{L}\sigma + b \tag{4.3}$$

Where ε is the signal efficiency, $\mathcal{L}$ is the integrated luminosity, and b is the background count. The mean counts of various physics processes (that do not interfere quantum-mechanically), indexed by k, will be

$$\nu_i(\sigma,\varepsilon,\mathrm{BR},\mathcal{L}) = \sum_k \nu_{ik}(\sigma,\varepsilon,\mathrm{b},\mathcal{L}) \tag{4.4}$$

Each process k is associated with a probability $p_{ik}(x|\sigma,\varepsilon,\mathrm{b},\mathcal{L})$ to produce outcomes x for channel i. The probability $p_i(x_{ij}|\sigma,\varepsilon,\mathrm{BR},\mathcal{L})$ to measure x_{ij} in channel i event j is therefore the weighted sum

$$p_i\left(x_{ij} \mid \sigma,\varepsilon,\mathrm{b},\mathcal{L}\right) = \sum_k \frac{\nu_{ik}(\sigma,\varepsilon,\mathrm{b},\mathcal{L})}{\nu_i(\sigma,\varepsilon,\mathrm{b},\mathcal{L})} p_{ik}\left(x_{ij} \mid \sigma,\varepsilon,\mathrm{b},\mathcal{L}\right) \tag{4.5}$$

This corresponds to what we usually refer to as a stacked histogram, which is a statistical mixture model. Given our auxiliary data, which could be estimates of the relevant parameters $\{\hat{\sigma},\hat{\varepsilon},\hat{\mathrm{b}},\hat{\mathcal{L}}\}$ with pdf $p(\hat{\sigma},\hat{\varepsilon},\hat{\mathrm{b}},\hat{\mathcal{L}} \mid \varepsilon,\mathrm{b},\mathcal{L})$, whose form we can leave open, the full statistical model can be written as

$$p(n, x, \hat{\varepsilon}, \hat{\mathrm{b}}, \hat{\mathcal{L}} \mid \sigma, \varepsilon, \mathrm{b}, \mathcal{L}) = \prod_{i=1}^{N_c} \left[\mathrm{Pois}\left(n_i \mid \nu_i(\sigma, \varepsilon, \mathrm{b}, \mathcal{L})\right) \prod_{j=1}^{n_i} p_i\left(x_{ij} \mid \sigma, \varepsilon, \mathrm{b}, \mathcal{L}\right) \right] \tag{4.6}$$

And when the observable data $\{n, x, \hat{\varepsilon}, \hat{\mathrm{b}}, \hat{\mathcal{L}}\}$ are entered into the statistical model, it becomes the likelihood, $p(n, x, \hat{\varepsilon}, \hat{\mathrm{b}}, \hat{\mathcal{L}} \mid \sigma, \varepsilon, \mathrm{b}, \mathcal{L}) = L(\sigma, \varepsilon, \mathrm{b}, \mathcal{L})$. [1]

4.2.1 Multinomial Model

If we use a multinomial distribution rather than a product of Poisson distributions, as in 4.6, then every term in the multinomial is the cross section in that bin divided by the total cross section, $\frac{\sigma_{\text{bin}}}{\sigma_{\text{total}}}$. Let us denote $L_{\mathrm{P}}(\sigma, \varepsilon, \mathrm{BR}, \mathcal{L})$ as the Poisson likelihood [2]

As described above (and is typically used), and $L_{\mathrm{MN}}(\sigma, \varepsilon, \mathrm{BR}, \mathcal{L})$ to our suggested multinomial likelihood. These are related as

$$\begin{aligned} L_{\mathrm{MN}} &= \frac{\prod_{\text{bins } i} L_{\mathrm{P},i}}{L_{\mathrm{P},total}} = \frac{\prod_{\text{bins } i} \frac{e^{\sigma_i} \sigma_i^x}{x!}}{\frac{e^{\sigma_{\text{total}}} \sigma_{\text{total}}^{x_{\text{total}}}}{x_{\text{total}}!}} \\ &= x_{\text{total}}! \cdot \prod_{\text{bins } i} \frac{1}{x_i!} \left(\frac{\sigma_i}{\sigma_{\text{total}}}\right)^x \end{aligned} \tag{4.7}$$

This has a virtue that since the number of observed events for a process $N \approx \varepsilon \mathcal{L} \sigma \rightarrow \sigma \approx \frac{N}{\varepsilon \mathcal{L}}$, using $\frac{\sigma_i}{\sigma_{\text{total}}}$ would lead to a cancellation to many of the nuissance parameters that are included in the cross section such as the efficiency and integrated luminosity, since they take the same in each bin's cross section as in the total cross section. Therefore by using a multinomial model, we become insensetive to the fluctuations in the luminosity and efficiency and potentially cancel the dependency on luminosity and efficiency.

4.3 Folding, Not Unfolding!

Since the CMS detector, like any detector, is not perfect, the measured (detector-level) jet observables are not expected to be the same as that of the corresponding true (or particle-level) jets. Therefore, to account for the detector response and construct the mapping from the truth-level to the observed detector-level, nearly all HEP analysis "unfolds" the truth-level observables to the detector-level observables by means of a response matrix [3].

In the Bayesian unfolding process, one defines the parameters $\vec{\mu} = (\mu_1, \ldots, \mu_M)$ to represent the expected number of entries in a bin (e.g. μ_2 is the expected number of events in bin $i = 2$) assuming perfect resolution. In reality, the detector has limited resolution and so an event with a true value of a variable in a bin might be measured (reconstructed) in a

[1] From here one can eliminate the nuissance parameters θ by computing the profile likelihood $L_{\mathrm{p}}(\sigma) = L(\sigma, \hat{\hat{\theta}}(\sigma))$, where $\hat{\hat{\theta}}(\sigma)$ are the values that maximize the likelihood for a given parameter of interest. One can also eliminate the nuissance parameters by marginalizing the posterior density to find the probability of the cross section, $p(\sigma \mid x) = \int p(\sigma, \theta \mid x) d\theta$

[2] Then if our data is x and our parameter of interest is σ, then $L_{\mathrm{P}} = \frac{e^{\sigma} \sigma^x}{x!}$.

[3] Sometimes the response matrix is reffered to as the unfolding matrix

different one, so that $\vec{\nu} = (\nu_1, \dots \nu_N)$ represents the number of events at the reconstructed or detector level. These parameters are related by

$$\vec{\nu} = R\vec{\mu} \tag{4.8}$$

Or $\nu_i = \sum_{j-1}^{N} R_{ij}\mu_j$, where R is an $N \times M$ the response matrix such that R_{ij} represents the probability to measure ν in bin i given that its true value μ was in bin j. Please not that 4.8 is not strictly correct as it entails an approximation. The actual formula that is meant by such an expression is

$$\nu(x) = \int \underbrace{P(x|y)}_{\text{"}R\text{"}} \mu(y)\ dy \tag{4.9}$$

And even if we discritize this integral in equation 4.9, it's still that for every bin this holds; suppose that i labels the bins in the space of observations, and j labels the bins in the space of theory, then we have $\nu_i = \int P(x_i|y_j)\mu(y_j)dy_j$. If we have the bins being very narrow, [4] then one could argue that the expression actually reaches the continuous expression in equation 4.9, or in a binned scenario, $\nu_i = R(x_i|y_j)\mu(y_j)\Delta y_j$. However, if the bins are wide, then R in equation 4.8 actually depends on μ. Therefore we argue that unfolding is misguided and mathematically ill-defined [5] and that we should avoid this unfolding completely!

Instead of unfolding the spectrum, which leads to all the issues that are mentioned above, we will publish the likelihood, and then what is required is that we publish something with it that allows a theorist that allows a theorist to take their prediction and fold it. [6]

Conclusion: Part of my research going forward is to develop techniques that do the folding, where theorists can use it to map the theory spectrum to the observed spectrum in a fast and efficient way. ML techniques, such as transformer models, offer a possibility to achieve such a mapping, which would extend my research on this topic.

[4] By that I mean $\lim_{\Delta y \to 0} \nu_i = \int P(x_i|y_j)\mu(y_j)dy_j$

[5] For example, every implementation of this kind of unfolding injects some assumptions, and different assumptions that are injected in it leads to slightly different results. Also to our knowledge, nobody has proven that this effect is small.

[6] Unfolding can be viewed as a mapping from observation space to theory space $\vec{\nu} \to \vec{\mu}$. In that sense, folding is a mapping from theory to observation, $\vec{mu} \to \vec{\nu}$.

Bibliography

[1]

[2]

[3]

[4] 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.

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

[6] J M Campbell, J W Huston, and W J Stirling. Hard interactions of quarks and gluons: a primer for lhc physics hard interactions of quarks and gluons: a primer for lhc physics 2, 2006.

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

[8] 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.

[9] Kyle Cranmer, Sabine Kraml, Harrison B. Prosper, Philip Bechtle, Florian U. Bernlochner, Itay M. Bloch, Enzo Canonero, Marcin Chrzaszcz, Andrea Coccaro, Jan Conrad, Glen Cowan, Matthew Feickert, Nahuel Ferreiro Iachellini, Andrew Fowlie, Lukas Heinrich, Alexander Held, Thomas Kuhr, Anders Kvellestad, Maeve Madigan, Farvah Mahmoudi, Knut Dundas Morå, Mark S. Neubauer, Maurizio Pierini, Juan Rojo, Sezen Sekmen, Luca Silvestrini, Veronica Sanz, Giordon Stark, Riccardo Torre, Robert Thorne, Wolfgang Waltenberger, Nicholas Wardle, and Jonas Wittbrodt. Publishing statistical models: Getting the most out of particle physics experiments. 9 2021.

[10] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. Delphes 3, a modular framework for fast simulation of a generic collider experiment. 7 2013.

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

[12] S. Chatrchyan et al. 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.

[13] V. Khachatryan et al. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12, 2017.

[14] 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.

[15] Jun Gao. Cijet: a program for computation of jet cross sections induced by quark contact interactions at hadron colliders. 1 2013.

[16] L. A. Harland-Lang, A. D. Martin, and R. S. Thorne. The impact of lhc jet data on the mmht pdf fit at nnlo. *European Physical Journal C*, 78, 3 2018.

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

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

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

[20] Ali Al Kadhim. Ultra legacy 2017 prefiring maps. https://indico.ijclab.in2p3.fr/event/7847/contributions/25471/attachments/18431/24586/xFitter_Workshop2022_Alkadhim2_.pdf.

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

[22] Gavin P. Salam. Towards jetography. 6 2009.

[23] 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.

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

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

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