CMS Event Simulation

Nicole A. Larsen
Department of Physics, Georgia Institute of Technology, Atlanta, GA, 30332
(Dated: August 11, 2006)

The CMS Detector located at the Large Hadron Collider at CERN recently underwent a dramatic change in software. The old framework was discarded entirely in favor of a new one, called CMSSW. In light of this transition, it is essential that new Monte Carlo data be generated and that CMSSW be tested and validated in order to become familiar with the software and provide feedback for the software developers. During this project, CMSSW was used to generate and analyze Monte Carlo samples for top pair production. Specifically, we examined jets, missing energy, and tracks for $pp \rightarrow t\bar{t}$ events.

I. INTRODUCTION

The Large Hadron Collider (LHC) is a 14 TeV proton-proton collider currently under construction at CERN. Scheduled to start running in November 2007, the LHC will hopefully answer some of particle physics' greatest unsolved questions. The main goals of the LHC are to detect the elusive Higgs boson, to look for supersymmetric particles, and to probe quark and lepton substructure, as well as to look for new physics and deviations from the Standard Model. In order to accomplish this, several particle detectors are being built at the LHC, one of which is the Compact Muon Solenoid (CMS).

In 2004, the decision was made to discard all of the old CMS software and replace it with a single new framework called CMSSW. With less than fifteen months before the LHC is planned to begin operation, it is imperative that we gain an understanding of this new software. In order to accomplish this, new Monte Carlo data must be generated and analyzed with CMSSW.

The top quark was selected as a subject for this study for a number of reasons. The top quark is unique in that it is by far the heaviest fundamental particle, and, because of its short lifetime, it does not hadronize like other quarks. In addition, because the mass of the top quark is related to the masses of the W and Higgs bosons, top measurements will place constraints on the mass of the Higgs. Top events in the CMS detector will also be a major source of background for other physics processes, as well as helping calibration by providing measurements for parameters such as jet energies. [1]

Perhaps the most important reason to examine the top quark is its large cross-section. It is anticipated that the LHC will produce eight million top events throughout its first year running at low luminosity, providing a valuable opportunity to study top physics. This is about one thousand times as many top quarks as those produced during the second run of the Tevatron. [2]

The top quark has three different decay channels (Fig. 1), all of which were examined during this project. In each channel, a top-antitop pair decays into two W bosons and two b quarks. All-hadronic decay occurs when each W boson decays into a quark-antiquark pair. Leptonic decay occurs when the W bosons each decay into a lepton and a neutrino. Finally, there is the semi-leptonic, or jets+leptons, decay channel, where one W produces a quark antiquark pair and the other decays into a lepton and a neutrino.

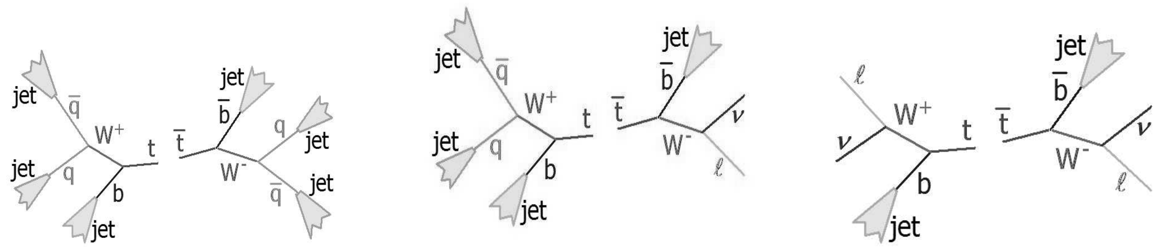

FIG. 1: $t\bar{t}$ decay channels: (Left) The hadronic channel. (Center) The semi-leptonic, or jets+leptons channel. (Right) The leptonic channel.

II. OVERVIEW OF DATA PRODUCTION AND ANALYSIS IN CMSSW

Data production begins with a configuration (.cfg) file used to configure the cmsRun executable. The file will contain several modules to be executed in a specified order, as well as parameters and configuration information for the modules. For event generation, the modules needed are an input source, a detector geometry simulator, a digitization module, a reconstruction module, and an output module. In this case, the input was a set of data produced by Pythia, a Monte Carlo generator. The digitization module simulates the detector's response to the input particles. Next, the reconstruction module uses the digitized information to recreate objects such as particle tracks, jets, electrons, and muons. Finally, the output module produces a ROOT file containing data from the events. [1]

There are multiple ways to analyze the data in a ROOT file. Analysis in bare ROOT allows one to apply cuts and data fits, but to make a more sophisticated analysis using the full framework, an Event Data Analyzer (EDAnalyzer) is required. An EDAnalyzer is a C++ object that reads in data from a ROOT file and allows a user to manipulate it using C++ code. Another configuration file is needed to call the analyzer, which typically outputs another ROOT file. During each stage of this project, we used bare ROOT to make a preliminary analysis of data. We then wrote EDAnalyzers in order to better examine the particle information.

III. COORDINATES AND CONVENTIONS FOR CMS

The coordinate system used for the CMS detector is as follows: x points south toward the center of the LHC, y points vertically upwards, and z is horizontal, in the direction of the beampipe. In polar coordinates, ϕ is the azimuthal angle, with $\phi = 0$ the positive x-axis, and θ is the polar angle, with $\theta = 0$ the positive z-axis (Fig. 2). Often, the quantity η is used to represent polar angle, where $\eta = -\ln\tan\theta/2$.

In CMS, the units used are GeV for energy, GeV/c for momentum, and GeV/c^2 for mass; position and distance are measured in cm, and time is measured in ns.

[1] ROOT is an object-oriented data analysis package that stores data in the form of ntuples or trees, which can then be viewed as histograms.

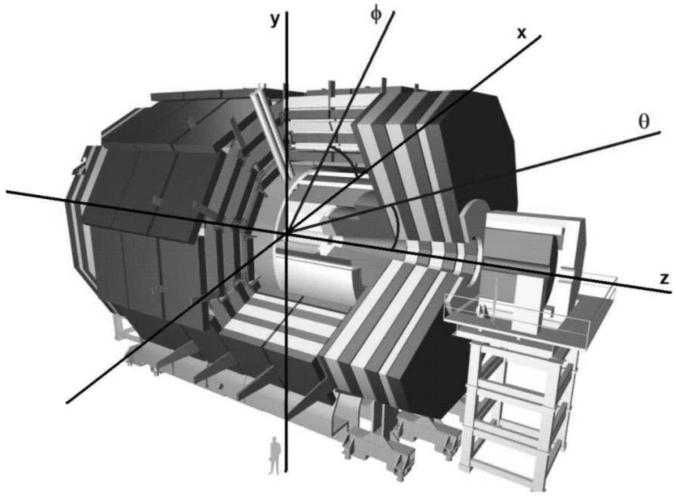

FIG. 2: Coordinates for the CMS detector

IV. JETS AND MISSING ENERGY

Using CMSSW version 0_6_1, two Monte Carlo 100-event $t\bar{t}$ production samples were generated. [3] The first of these was an inclusive sample, comprised of events from all three different top decay channels. The second contained only leptonic top decays. Each file contained only information about jets and missing energy; no tracking or lepton information was available. The quantities examined included jet momentum, energy, and coordinates; missing energy; and number of jets per event. We also wrote a jet-matching algorithm in order to pair detected jets with the generated jets that produced them. Finally, we compared and contrasted the $t\bar{t}$ inclusive data with the $t\bar{t}$ leptonic data.

The coordinates of the jets agreed well with expectations. As expected, the ϕ-distribution of the jets was fairly constant due to the cylindrical symmetry of the detector. Also as expected, most of the jets had relatively low eta. When comparing the distribution of jets detected by the calorimeter with the distribution of jets generated by the Monte Carlo, it was discovered that there were gaps at $-3 < \eta < -1.5$ and $1.5 < \eta < 3$. This can be attributed to the detector geometry, since these regions correspond to the areas where the calorimeter barrel meets the calorimeter endcap. (Fig. 3)

A notable difference between the two Monte Carlo samples was that the number of jets that occurred in each leptonic event was generally much lower than the number found for all-hadronic events (Table I). We note that this agrees with theory, since $t\bar{t}$ inclusive contains hadronic decays with six or more jets, while $t\bar{t}$ leptonic contains only leptonic decays, which are guaranteed only two jets. For both $t\bar{t}$ inclusive and $t\bar{t}$ leptonic, however, the number of jets detected per event was approximately fifty percent of the jets generated during that event. This can be explained by inefficiencies inherent in the calorimeter.

On a similar note, the jets energies were in general much lower for the jets in the $t\bar{t}$ leptonic sample than those in the inclusive sample (Table II). The reason for this is unknown.

Later, we created an EDAnalyzer that would match each jet detected in the calorimeter with the Monte Carlo generated jet that produced it. [4] This algorithm was based on a matching coefficient $dR = \sqrt{d\phi^2 + d\eta^2}$. A pair consisting of one generated and one detected

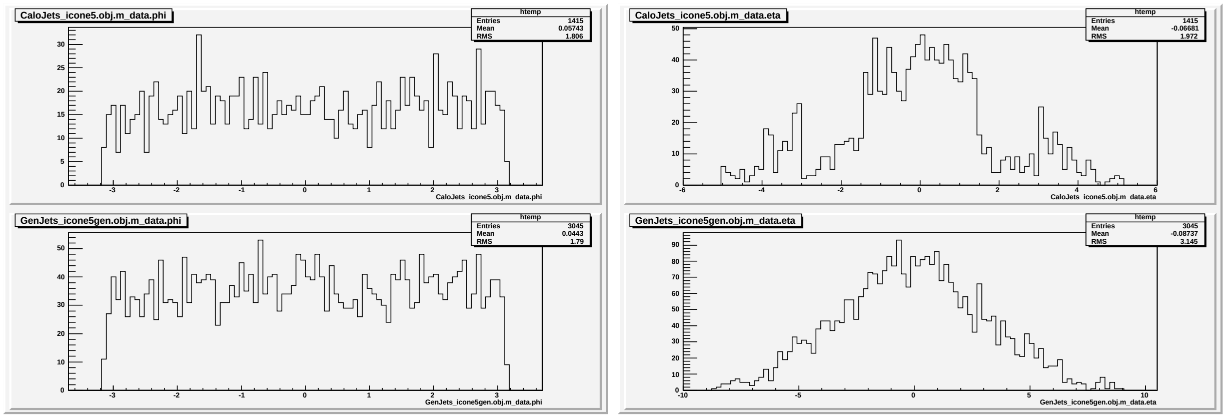

FIG. 3: ϕ- and η-distributions for $t\bar{t}$ inclusive jets: (Top left) ϕ-distribution for detected jets. (Bottom left) ϕ-distribution for generated jets. (Top right) η-distribution for detected jets. (Bottom right) η-distribution for generated jets.

TABLE I: Average number of jets per event, detected and generated.

Jet-Finding Algorithm	Energy	$t\bar{t}$ inc. Detected	$t\bar{t}$ inc. Generated	$t\bar{t}$ lep. Detected	$t\bar{t}$ lep. Generated
icone5	All energies	14.15	30.45	2.79	6.59
icone5	et> 20 GeV	3.18	6.25	0.54	1.31
mcone5	All energies	11.33	24.3	2.33	5.37
mcone5	et> 20 GeV	3.16	6.01	0.52	1.22
mcone7	All energies	8.89	18.17	1.77	3.77
mcone7	et> 20 GeV	3.1	5.55	0.54	1.11

jet was considered a match if that pair had $dR < 0.5$. This jet-matching algorithm found matches for $\sim 40\%$ of the generated jets. In general, the two jets in each pair were extremely close in mass; however, most of the matches produced had energy differences on the order of ~ 75 GeV. The reason for this is still unknown.

We also examined missing energy for each decay. Missing energy is defined as $\vec{E}_T^{miss} + \vec{E}_{sum} = 0$, where $\vec{E}_{sum} = \sum_i E_i \hat{n}_i$ and $\hat{n}_i$ is the vector pointing from $x = y = z = 0$ to the $\hat{i}$th calorimeter tower. In other words, missing energy indicates a momentum imbalance in the calorimeter where we expect momentum to be conserved. There was a large amount of missing energy detected in the calorimeter, especially for $t\bar{t}$ inclusive. (Fig. 4) Note that the missing energy generated is less than that which is detected by a magnitude of

TABLE II: Average energy per jet (GeV).

Jet-Finding Algorithm	$t\bar{t}$ inc. Detected	$t\bar{t}$ inc. Generated	$t\bar{t}$ lep. Detected	$t\bar{t}$ lep. Generated
icone5	16.67	18.33	12.93	17.02
mcone5	20.06	21.6	15.05	19.78
mcone7	41.36	50.91	20.47	28.23

approximately 10^2. At first, we believed that this discrepancy of scale was caused by an error in CMSSW. However, since it is related to total momentum, generated missing energy should be approximately zero. The question then arises why the values for generated missing energy are so large (~ 2 GeV). Further investigation is needed.

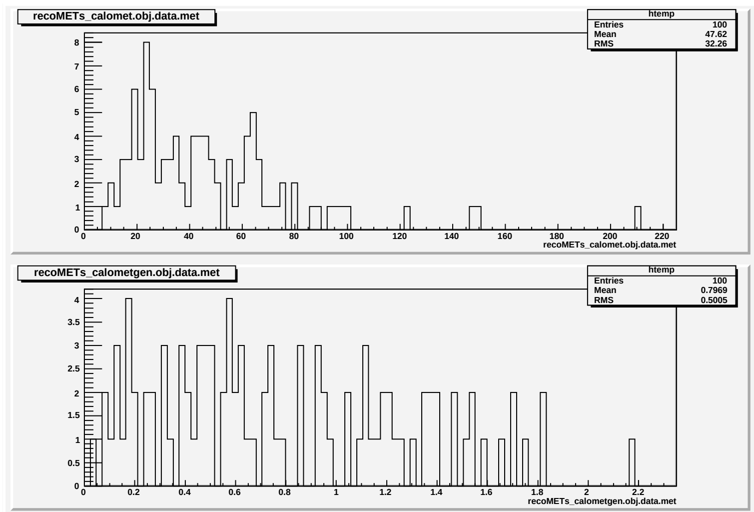

FIG. 4: Missing energy for $t\bar{t}$ inclusive: (Top) Detected missing energy. (Bottom) Generated missing energy.

V. TRACKS

For the second stage of the project, CMSSW version 0_7_0 was used to generate 20-event Monte Carlo samples for $t\bar{t}$ inclusive, $t\bar{t}$ semi-leptonic, and $t\bar{t}$ leptonic channels. In particular, we examined particle tracks produced in each sample. The samples all contained some information about tracking, but this information was incomplete. Only some of the members of the Track class could be accessed; others, such as the length-5 array of track parameters, were either not present or did not function.

The number of tracks per event was found to be between ten and sixty for each decay channel, with an average of about 35 tracks per event. On average, the leptonic decay produced fewer tracks than the other two channels.

TABLE III: Average number of tracks per event.

$t\bar{t}$ inclusive	$t\bar{t}$ semi-leptonic	$t\bar{t}$ leptonic
36.65	35.15	29

Among the Track members that were accessible were those that were also members of the TrackExtras class. These included momentum, energy, and coordinates of the outermost point in the reconstructed track. The x-,y-,and z-coordinates for the outermost point in each track can be seen in Fig. 5. While looking at track coordinates, it was discovered that the tracks produced in all three track channels were extremely forward, or had high values of η. The generator information indicates that much of the momentum of W bosons is in the z-direction. We believe that the forwardness of these tracks can be attributed to the high energy (up to 14 TeV) of the proton collisions in the LHC. In addition, the momentum at the outermost point of each track was found to be extremely low (5GeV/c on average). The reason for this is unknown. (Fig. 6)

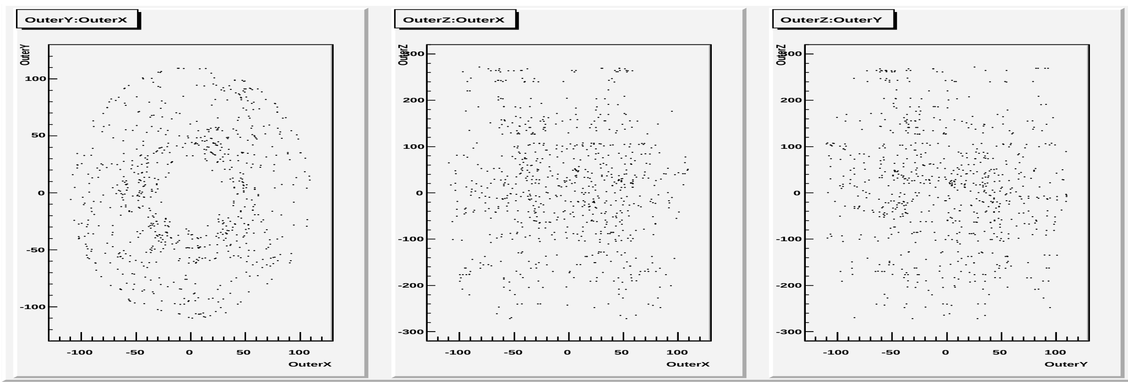

FIG. 5: The xyz coordinates of the outermost point in the track for $t\bar{t}$ inclusive: (Left) y vs x. (Center) z vs x. (Right) z vs y.

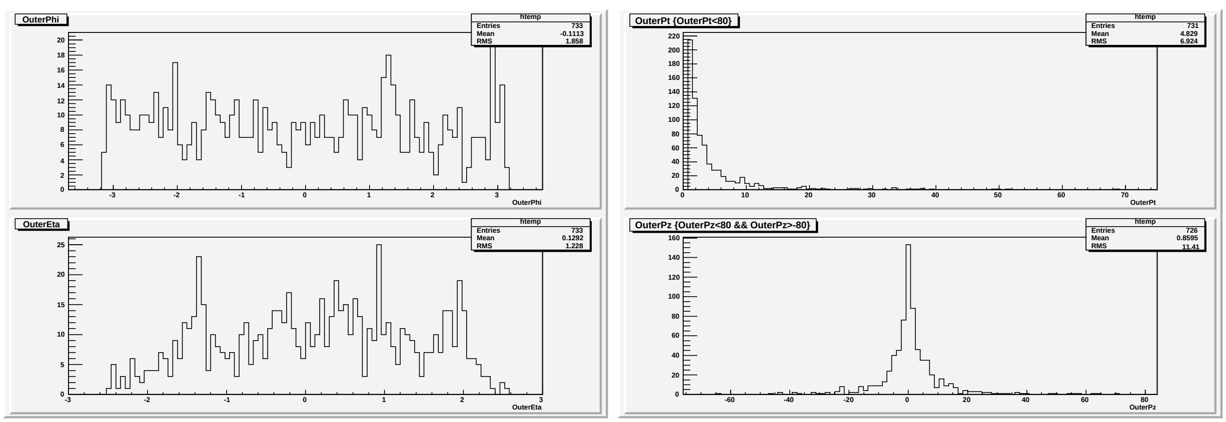

FIG. 6: Polar coordinates and momentum for $t\bar{t}$ inclusive: The ϕ- and η-coordinates for the outermost points in each track (Top and bottom left, respectively). The transverse momentum and the momentum in the z-direction for the outermost points in each track (Top and bottom right, respectively).

VI. RESULTS AND CONCLUSIONS

Jet reconstruction is operational in both CMSSW 0_6_1 and 0_7_0. We found that, for the most part, jet information agreed with expectations. However, it is not yet known why the units for detected and generated missing energy do not agree. The software needs to be be examined in order to determine precisely which quantity is being measured for missing energy. It is also necessary to further investigate the energy difference between detected and generated jets paired by the jet-matching algorithm. [5]

Track reconstruction is not yet complete in version 0_7_0. Only a few of the track variables, such as coordinates and momentum at the outermost point of the track, were accessible. The high forwardness of tracks in the CMS detector needs to be explored further, as does the low track momentum.

We recently produced Monte Carlo files for Z→ ee and Z'→ ee decays in CMSSW version 0_7_0. Unfortunately, these files contained no tracking information at all. Cluster information for these Z decays was briefly examined, but little was found, and much work remains to be done.

The next major step in this project is to install CMSSW version 0_9_0, in which tracking information will be both complete and functional, and use it to examine track reconstruction for various decays. New Monte Carlo files must be produced for both $t\bar{t}$ and Z decays, which will subsequently be analyzed for tracking information. Later, we will generate Monte Carlo data for Higgs decay and SUSY processes. By studying these processes and others, we will gain an intuition for particles in the CMS detector and we will have a better idea what to expect when the LHC goes online.

VII. ACKNOWLEDGMENTS

I would like to acknowledge Julia Thom of Cornell university for proposing and guiding this research project. I would also like to thank Peter Wittich, also of Cornell University for providing help along the way. Finally, thanks to Jennifer Vaughan (Cornell University) for producing the Monte Carlo files used in this project, and Nicholas Stone (Cornell University) for his contributions and collaborations. This work was supported by the National Science Foundation REU grant PHY-0552386 and research co-operative agreement PHY-0202078.

[1] Weiser, Christian. "Top Physics At the LHC." arXiv.org. 10 Jun 2005. http://arxiv.org/PS_cache/hep-ex/pdf/0506/0506024.pdf

[2] "Why We Care About the Top Quark: CDF Explains." Fermilab. http://www.fnal.gov/pub/news04/top_quark_cdf.html.

[3] Monte Carlo files produced by Jennifer Vaughan.

[4] Collaborated with Nicholas Stone.

[5] Currently being investigated by Nicholas Stone and Peter Wittich.