Mid Course Assessment: Validation Tests of Geant4 Electromagnetic Physics and Investigations of the Decay $B^0 \to K^* \eta$

Timothy Williams

Contents

1 Introduction **1**

2 Theory **2**

3 Detector **2**
3.1 Tracking . . . 3
3.2 Calorimetry . . . 5
3.3 Particle Identification . . . 6
3.4 Trigger . . . 6

4 Simulation **8**
4.1 Simple Calorimeter Test . . . 10
4.1.1 Comparison of GEANT4 version 9.5.p02 and 9.6.p04 . . . 10
4.1.2 Comparison of Physics Lists . . . 17
4.1.3 Simple Calorimeter Test Conclusions . . . 22
4.2 Multiple Scattering Test . . . 24
4.2.1 Results . . . 27
4.2.2 Conclusions . . . 27

5 Investigation of $B^0 \to K^*(892)\eta$ **29**
5.1 Decay Reconstruction . . . 30
5.2 Event Selection . . . 30
5.2.1 Stripping Selection . . . 30
5.2.2 Trigger . . . 32
5.2.3 Multivariate Selection . . . 33
5.2.4 Next Steps . . . 41

References **43**

1 Introduction

The LHCb experiment is one of 4 detectors making use of CERNs Large Hadron Collider (LHC). The LHC is a circular collider with a 27km circumference that accelerates protons and heavy ions to record breaking energies. To date a center of mass energy of 13TeV has been achieved with protons. The accelerated protons/ions collide at 4 interaction points around the circumference of the collider, each of which is surrounded by a detector. There are two general purpose detectors at the LHC known as ATLAS and CMS which have a wide range of physics aims. The other two detectors, however, are specialised for more specific physics goals. One of these detectors, ALICE, is specialised for studying heavy

ion collisions. LHCb is specialised for studying particles containing beauty quarks in the forward region.

Also, out of the 3 experiments focused on studying proton-proton (pp) collisions, LHCb is the only one not focused on direct searches for new physics. Rather than searching for new particles directly created in the pp collisions, LHCb searches for new physics by precisely measuring quantities that would be indirectly affected by the presence of physics beyond the standard model. These measurements include (but are certainly not limited to): Rare decay rates, CKM angles, CKM matrix elements and CP asymmetries.

The analysis section of this report studies the decay $B^0 \to K^* \eta$, which is required for the to search for the rare decay $\Lambda_b^0 \to$ p $K^- \eta$. In general, decays that are forbidden at tree level in the standard model can only proceed via loop diagrams. As theoretically a large range of particles can contribute to these loop diagrams, any new physics could contribute and change the rate at which this process happens. By measuring branching fractions and comparing these to standard model predicitons, new physics can be searched for.

This report also details work that has been carried out to validate the simulation software used at LHCb. As described in section 4 it is impossible to perform a modern particle physics analysis without simulating the production and detection of relevant particles and processes. Extensive work is carried out to tune the element of the software that generates particles with correct distributions, but this report presents validation tests of the way interations with the detector materials are simulated.

2 Theory

Some awesome theory.

3 Detector

The LHCb detector is a unique and specialised apparatus that has been optimised to meet specific physics goals, which largely involve the study of mesons and baryons containing beauty and charm quarks. The resulting design is a single arm forward spectrometer which covers the pseudorapidity range $2 < \eta < 5$. A schematic diagram of the detector is shown in 1. This limited angular coverage is motivated by the polar angle distribution of b and $\bar{b}$ quark production peaking at small angles to the beam pipe in both the forward and backward direction, as shown in figure 2. The consequence of this angular distribution is that at 8 TeV approximately 25% of b or $\bar{b}$ quarks are produced within the angular acceptance of the LHCb detector, despite the geometrical acceptance of the detector being only 0.09 steradians (0.7% of a full sphere).

Another unique characteristic of the LHCb detector is the lower luminosity that it operates at. During run 1 the ATLAS and CMS detectors made use of a peak luminosity of $7 \times 10^{33} cm^{-2} s^{-1}$ at the start of a fill which gradually decreased over the 6-8 hour lifetime of the fill as collisions took place. However, LHCb ran at a lower, levelled, luminosity of

$4 \times 10^{32} cm^{-2} s^{-1}$. The process of luminosity levelling (at LHCb) involves offestting the LHC proton beams in the transverse direction to reduce the probability of interaction [1]. As the luminosity of the LHC decreases over the lifetime of the fill, the beam offset at LHCb is decreased to keep the instantaneous luminosity delivered at LHCb constant. The result of this lower luminosity is only a mean number of proton-proton interactions per bunch crossing of 2.5 compared to 20 as experienced by the general purpose LHC detectors. Consequently, the radiation hardness requirements of the LHCb detector are reduced and backgrounds from other interactions in the same bunch crossing are also significantly lower. This lower level of background is very advantageous for rare decay searches such as $B^0 \rightarrow \mu\mu$.

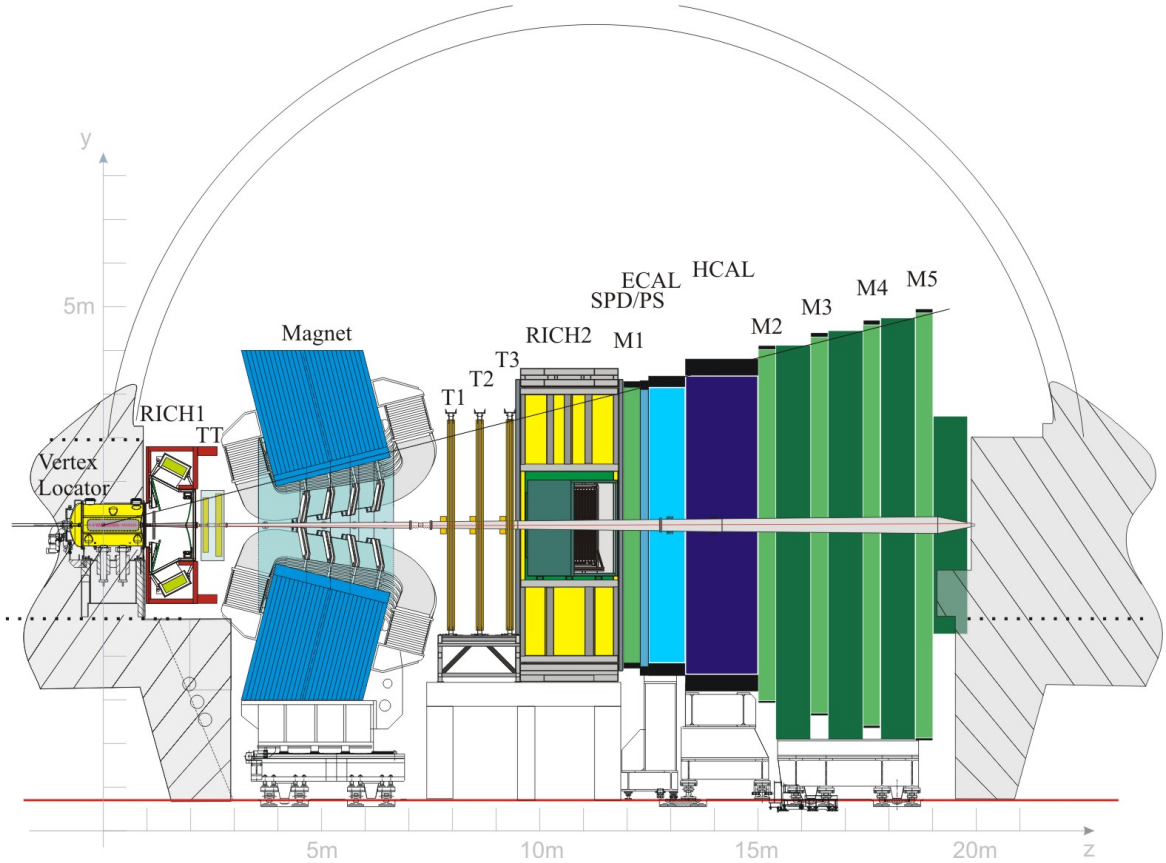


Figure 1: A schematic diagram of the LHCb detector and all its componens

3.1 Tracking

The objective of any particle physics tracking system is to provide positional information on the path taken by charged particles. At all collider based experiments the paths of charged particles are curved by a magentic field to distinguish positive and negatively charged particles, measure momentum and also aid particle identification. At LHCb this

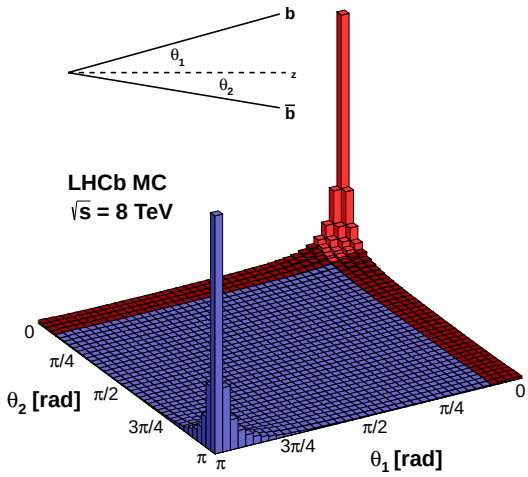


Figure 2: A diagram showing the angular distribution of $b\bar{b}$ pair production at 8TeV

is performed by a warm dipole magnet, which is subject to a polarity reversal periodically to remove systematic uncertainties.

The first stage of the LHCb tracking system is the Vertex Locator (VELO). When a B^0 meson is created it will travel a measurable distance before it decays. As the B^0 is neutral it leaves no track, therefore there is a secondary or displaced vertex present in the decay. This is also the case for many other mesons containing b and c quarks. As these make up such a large section of the LHCb phyiscs programme, the VELO was developed to detect these secondary vertices.

The VELO consists of several circular silicon sensors which are placed transverse to the beam pipe, with the beam pipe passing through the center. A diagram of this is shown in Figure 3. As shown in Figure 1 these are positioned before any other detector component as close as possible to the interaction region. There are two types of sensors, one which provides radial position information of tracks and one which provides azimuthal (ϕ) position information. These work in tandem to provide an optimal track resolution of $4\mu m$ [2], which is the best of any tracking system at the LHC.

In one year of running the VELO receives a 1MeV neutron equivalent fluence of $1.3 \times 10^{14} neq/cm^2$ which makes radiation hardness a challenge [3]. To keep this fluence to a minimum, the VELO is mechanically moved further away from the beam pipe whilst the LHC is still filling. During this stage no physics data can be taken, therefore the VELO would be undergoing irrdation for no gain in physics data.

The next stage of the LHCb tracking system is positioned just upstream of the dipole magnet, denoted "TT" in Figure 1. This also makes use of silicon microstrips to give a positional resolution of $52.6\mu m$, as measured in run I data [2].

The final stage of the tracking system is positioned downstream of the dipole magnet,

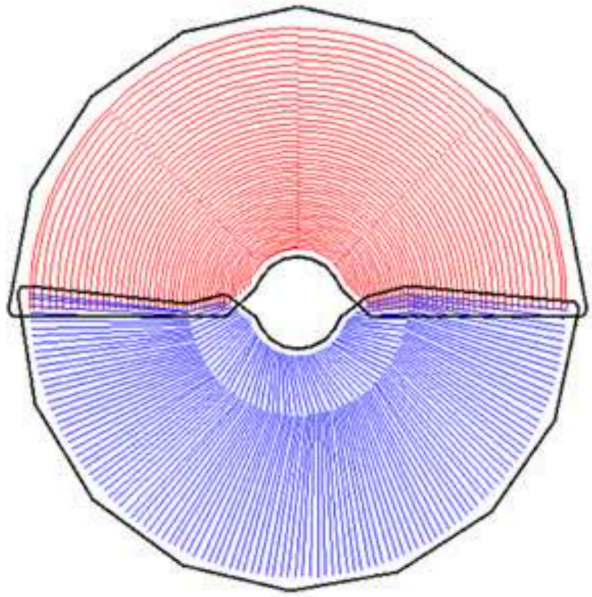


Figure 3: A diagram showing half of each type of VELO sensor in the x,y plance. The blue sensor provides azimuthal position information and the red sensor provides radial position information.

which is actually 3 seperate tracking stations labelled T1,T2 and T3 in Figure 1. The inner areas of these tracking stations are silicon microstrips which provide very similar spatial resolution to the TT. As silicon microstrips are very expensive, the outer areas of the downstream tracking stations are straw tube trackers. These provide a spatial resolution of $205\mu m$ [2].

3.2 Calorimetry

The role of calorimetry in HEP experiments is predominantly to measure the energy of particles. However they also play a role in particle identification and, particularly at the LHC, a role in triggering.

The electromagentic calorimeter used by LHCb is a tried and tested shashlik sampling calorimeter. It consists of 66 seperate layers, each consisting of 2mm of lead and 4mm of plastic scintillator. This amounts to 25 radiation lengths of material. The role of the lead is to induce an electromagnetic shower as quickly as possible and stop the incident particle within short spatial constraints. The role of the scintillator is to then measure the

energy contained in the EM shower.

The light from the plastic scintillator is transported to photo multiplier tubes by wavelength shifting fibres. These change the wavelength of the light from the plastic scintillator to a wavelength compatible with the photo multiplier tubes. The signal from the PMTs is then proportional to the energy of the incident particle, so the energy of the incident particle can be inferred.

3.3 Particle Identification

A large section of the LHCb physics programme relys on the excellent particle identification (PID) ability of the LHCb detector. There are many analyses that would not be possible if pions, kaons and protons could not be seperated as cleanly as they are. One result where this is the case is the observation of a pentaquark like state [4].

The PID performance of the LHCb detector is largely due to the presence of two ring imaging chernkov detectors (RICH1 and RICH2 in Figure 1). When a charged particle travels faster than the speed of light in a medium it will emit photons at an angle θ to the momentum direction of the charged particle. The angle at which these photons are radiated is given by,

$$cos(\theta) = \frac{1}{n\beta} = \frac{E}{np} = \frac{\sqrt{p^2 + m^2}}{np} \tag{1}$$

where n is the refractive index of the medium and p,E and m are the momentum, energy and mass of the charged particle respectively. The RICH1 detector is filled with C_4F_{10} and RICH2 is filled with CF_4. The refractive index of these mediums has been measured very accurately.

The RICH system measures the angle θ of cherenkov photons and the momentum of the charged particles can be measured independently of the RICH system with the curvature of the charged tracks. By combining these measurements, with the knowledge of the refractive index of the medium, one can determine the mass of the charged particle. It can be determined with enough precision to identify the particle.

The performance of the RICH detector varies as a function of charged particle momentum. During run I, for a momentum averaged Kaon efficiency of 95%, a 10% pion mis-identification rate was observed. This means 95% of kaons can be correctly identified as kaons whilst only 10% of pions are mis-identified as kaons [2].

3.4 Trigger

The role of a trigger in a HEP experiment is to decide which events are worth recording. At the LHC, proton bunches arrive at the interaction point every 25ns therefore collisions happen at a rate of 40MHz. It is impossible to process and record every single bunch crossing at this rate. Also, it would not be useful as only a fraction of the interactions that take place at LHCb are of interest.

The LHCb trigger system is comprised of 2 distinct levels. The first level is the L0 trigger, which is entirely implemented in hardware using field programmable gate arrays

(FPGAs). The role of the L0 trigger is to reduce the rate of events from 40MHz to 1.1MHz of potentially useful events. As these decisions have to be made at such a high rate, only information from the calorimeters and muon system are used in the L0 trigger.

The next stage is the High Level Trigger (HLT), which is an entirely software trigger that makes use of a computing farm with 26000 processor cores. The HLT reduces the overall event rate to 3kHz. However, this is split into 2 sub-triggers; the HLT1 and HLT2.

The HLT1 reduces the 1.1MHz input rate to 43kHz by partially reconstructing the event. In addition to the calorimeter and muon information, the information from the VELO and tracking systems is used. Simple selection criteria based on the information from these sub detectors can then be applied.

Now that the rate is reduced to 43kHz, the HLT2 can perform a full event reconstruction and apply more complete selection critera. These involve the information from all sub detectors and are often based on topological multivariate requirements [5]. The rate of events passing the HLT2 is 3kHz, therefore the LHCb trigger system has reduced the event rate my 4 orders of magnitude from 40MHz to 3kHz.

4 Simulation

It is crucial to any particle physics experiment to understand and be able to calculate the behavoiur of the detector and apparatus being used. No detector is perfect, therefore only a fraction of the particles created in the process being studied actually get fully measured and reconstructued. From this subset of detected particles one has to infer the total number of particles created and any relevant distributions related to the measurement in question, the only practical way to calculate these detector efficiencies is with a full detector simulation. One area of the LHCb physics programme where detector simulation is of particularly great importance is semi leptonic decays involving neutrinos. Due to the LHCb detector not being hermetic neutrinos escape the detector, consequently the only way to perform analyses involving neutrinos is with template fits based on monte carlo predictions. An example of an analysis exploiting this strategy is the measurement of V_{ub} using $\Lambda_b^0 \to p\mu\ \nu$ decays, which is a result considered to be of high importance [6]. A full detector simulation is also crucial for detector performance studies, understanding backgrounds and comparisons between theory and experiment.

The complexity of particle physics detectors and the many degrees of freedom in the underlying problem behind a detector simualtion make it unfathomably complex and highly multi-dimensional. Consequntly, the "monte carlo"(MC) or statistical sampling method is used to perform these simulations due to the fact it deals with multi dimensinal problems very well and it makes the problem conceptually far easier to understand. More specifically, direct monte carlo is used.

The first step in the monte carlo simulation is the generation of a pp interaction. This is the responsibility of the event generator, which generates particles with randomly generated properties sampled from the correct distributions. In LHCb Pythia is the defualt generator used for general purpose production, but Herwig ++ is sometimes used as a cross check [7, 8]. Both of these generators also simulate the hadronisation of the particles created in the hard pp interaction but Herwig uses a cluster model and Pythia uses a string model. There are also some specialised generators used which include: Bcvegpy for the production of B_c^+ mesons, GenXicc for Ξ_c production and Hiijing for heavy ion collisons [9–11].

The next stage of the simulation process is the transportation of the generated particles through the detector and all the interactions with the detector material that occur. In LHCb this is performed by the Geant4 toolkit which is a well established software framework solely created to simulate the interatctions of particles with matter [12]. It is this stage of the simulation process, as performed by Geant4, that is the subject of study in this report. Geant4 has the capability to simulate all relevant hadronic, electromagnetic and optical interactions over a large energy range for many long lived particles. Futhermore, Geant4 manages the geometry of the apparatus being simulated and also has the capability to simulate magnetic fields that are present at all collider based particle physics experiments.

Geant4 was first commisioned in 1993 as a result of studies that looked at how modern computing techniques could impove the existing fortran based GEANT frameworks that

date back to 1974 [12, 13]. These studies lead to GEANT4 being implemented as an object orientated framework in C++. One of the many advantages of object orientation is the ease by which improvements and refinements to the existing code can be made. Consequently, there are regular new releases of GEANT4, the latest version being 10.1 [14]. However, this is yet to be implemented in LHCb; until recently GEANT4 v9.5.p02 was implemented in the LHCb simulation package (GAUSS) but this has recently been updated to v9.6.p04.

When a new release of GEANT4 is adopted into the LHCb simulation package it is important to validate the reuslts of the new package and understand any discrepancies compared to the previous version. If changes to simulation results are not understood this could have serious consequences for physics analyses due to the unavoidable reliance on simulation. Part of the study presented in this report is a comparison of electromagnetic physics results between GEANT4 version 9.5.p02 and version 9.6.p04 for two seperate scenarios, both of which are important to LHCb.

The object orientated design of GEANT4 means that it is also very easy to have a variety of models avaiable for each interaction process; it is well known that the model giving the optimum agreement with data varies for different energy ranges and different particle types. Consequently, it is important to be able to tailor the simulation to the specific scenario under investigation. However, with the large number of interaction processes simulated and the large array of long lived particles to consider even a small selection of models for each process and particle would quickly lead to a very extensive range of possible configurations for the same scenario. Therefore, GEANT4 provides a set of physics lists (PL) each of which specifies a fixed set of models designed to be optimal for a smaller range of scenarios. For example, there are PLs available for low energy scenarios which would specify models that show the best agreement with data at lower energy. The effect of PL choice for scenarios relevant to LHCb also forms part of the study detailed in this report.

4.0.0.1 Outline of Geant4 Simulation Procedure GEANT4 simulates the passage of a particle by iteratively stepping it through the geometry being simulated. For some physical processes the step length is determined purely by cross sections; the distribution of distance travelled before the next interaction is calculated and sampled for a specific scenario. For some physical processes(such as multiple scattering) not every interaction is simulated, instead statisitcal effects are applied after a step has been taken. In this case the step length is chosen as a balance between accuracy and CPU time.

Before any step is taken, the step size is calculated for every physical process and the smallest is chosen. After this step is taken the proabilities of all physics processes for this step size are calculated, and random numbers are generated to decide if they take place. This includes but is not limited to: energy loss, changes of direction and production of secondary particles. This process is then repeated until the particle exits the geometry, loses all it's energy or is annihlated.

4.0.0.2 Production Cuts When a detector simulation takes place many secondary particle are created through processes such as bremstrahlung and pair production. Each

of these particles then has to be tracked and simulated by GEANT4 which takes a non negligable amount of CPU time. However, a lot of these particles are low energy particles that contribute very little to the overall result of the simulation. Therefore, a minimum range cut is applied in GEANT4 in order to minimise the amount of CPU time that is wasted tracking and simlutaing particles that have a negligible effect on the end result. This requires any secondary particles created to travel a minimum distance for them to be simulated by GEANT4. In LHCb this cut has been chosen as 5mm, a distance that has been found to reduce CPU time sufficiently but have a small enough impact on the simulation result. By default, production cuts are applied to the bremsstrahlung, ionisation and e+e- pair production processes. However, one can use the *applycuts* option which causes production cuts to also be applied to the photoelectric effect, compton scattering and gamma conversion.

4.1 Simple Calorimeter Test

The first test that has been setup is an extended electromagnetic GEANT4 example designed to model an electromagnetic calorimeter. This has been setup to model the LHCb electromagnetic calorimeter (ECAL) as closely as possible. As described in section 3, the LHCb ECAL is a sampling calorimeter consisiting of 66 alternating layers of lead and plastic scintillator [15]. This example has the ability to change: layer thicknesses, number of layers, lateral size and crucially the physics list that is used.

4.1.1 Comparison of Geant4 version 9.5.p02 and 9.6.p04

The basic principle behind this comparison was to run this example within both GEANT4 versions and compare selected results that could have implications for LHCb. Firstly, checks were made to ensure that the implementation of the example was unchanged between versions to ensure any discrepencies could be attributed to changes global to the GEANT4 framework and not just a change in this specific example. In order to make this a direct comparison between what is currently implemented and what will be implemented, the example was customised to run with the LHCb private PL. Therefore, the same physics models are being used in both versions, but the models themselves could have changed between versions.

4.1.1.1 Fractional Resolution Investigation The first result used as a comparison tool between GEANT4 versions was the fractional resolution of the electromagnetic calorimter. No detector is perfect, therefore the energy of an incident particle is measured with a resolution defined by a gaussian distirbution. However, the width of this gaussian distribution varies with the energy of the incident particle. The fractional resolution is defined as $\frac{\sigma(E)}{E}$ where $\sigma(E)$ is the standard deviation of a particles measured energy at a given incident energy E. The variation of the fractional resolution as a function of energy is given by,

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \bigoplus \frac{b}{E} \bigoplus c \tag{2}$$

where a, b and c are constants to be defined [16].

The 'a' term in equation 2 arises from statistical fluctuations in the electromagnetic shower induced by the sampling calorimeter. In an electromagnetic shower many particles are produced and the energy measured by the ECAL is the sum of the energies deposited by each particle. However in a sampling calorimeter only a fraction of the shower takes place in active regions, therefore only a fraction of the shower particles are actually measured. Consequently the number of shower particles that are measured is subject to poisson fluctutations with a standard deviation of $\sqrt{N}$, where N is the number of shower particles in the active region. Assuming the calorimeter is linear (which it should be) the number of shower particles produced is proportional to energy, therefore $\frac{\sigma}{E} \propto \frac{\sqrt{N}}{N} \propto \frac{1}{\sqrt{E}}$. In a sampling calorimeter it is sampling fluctiations that dominate the resolution, thus the 'a' term is usually by far the largest term in equation 2.

The 'b' term in equation 2 arises due to electronic noise which causes energy independent resolution effects. However, this is not simulated in this study and will therefore not be considered. The 'c' term in equation 2 arises due to shower leakage from the ECAL. At a given energy the amount of energy lost from the ECAL is subject to event by event fluctuations, which leads to broadening of the resolution. However, the total amount of energy lost from the ECAL is proportional to the energy of the incident particle and this dominates over the event by event effects. This leads to the standard deviation of measured energies due to shower leakage being proportinal to incident energy, and consequently energy indpendent for the fractional resolution.

The first step towards obtaining the fractional resolution of the model ECAL was to fire electrons into the ECAL at several fixed energies. At each energy a histogram was produced showing the distribution of energy deposited in active regions of the ECAL. This distribution is predicted to be gaussian, therefore a minimum chi squared fit of a gaussian function was performed at each energy to obtain the mean and standard deviation of the distribution. An example of one of these fits is shown in figure 4. This was performed at 13 different energies between 1.78GeV and 44.44GeV, this range was chosen because it includes the majority of electron energies expected at LHCb.

From these fits the standard deviation divided by the mean was plotted against the incident particle energy, as defined in the configuration of Geant4. $\frac{\sigma}{E}$ is expected to follow equation 2 but without the b term. Therefore a mimimum chi squared fit to,

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \bigoplus c \tag{3}$$

was performed with the parameters 'a' and 'c' left floating. This process was applied to both Geant4 version 9.5.p02 and 9.6.p04 with the same energies and 10000 events at each energy. The results of these fits for both Geant4 versions are shown in figure 5 and table 1.

It is clear from these fits that there is a significant discrepancy in the statistical term of the fractional resolution parameterisation between Geant4 versions, despite the same PL being used in both versions. The simulation of electromagnetic showers is a well established process that, in general, shows good agreement with data. Therefore, progression of results

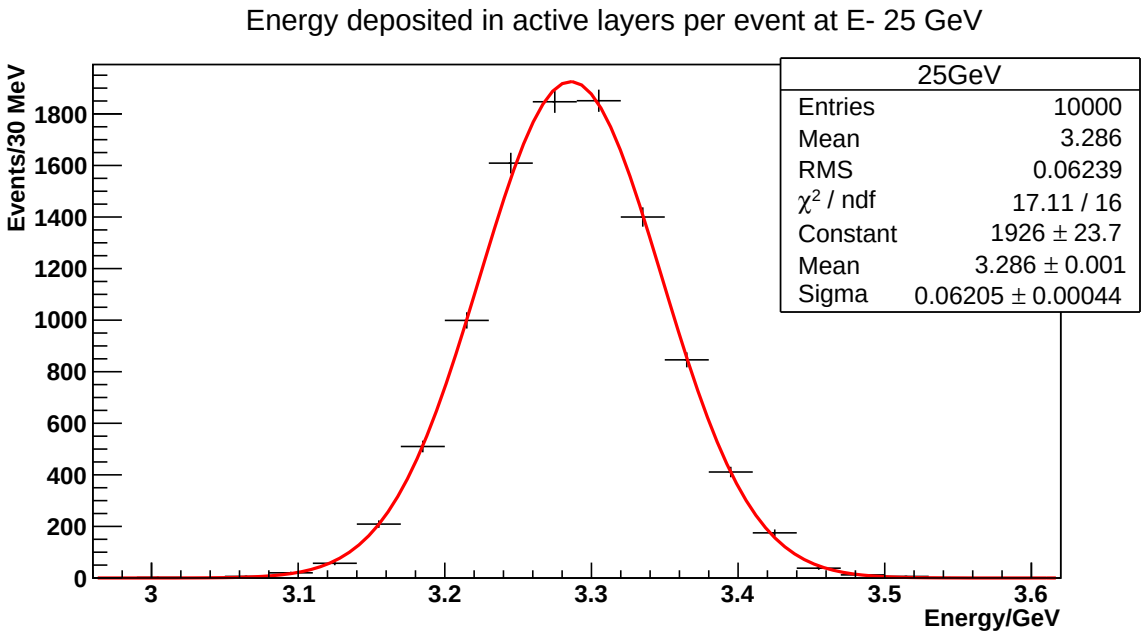


Figure 4: The distribution of energy deposited in scintillator(active) layers at 25 GeV

Version:	v9.5	v9.6	Difference
A term	0.0952±0.0003	0.0922±0.0003	8.43σ, 3.3%
C term	0.0051±0.0003	0.0049±0.0003	consistent
$\frac{\chi^2}{ndf}$	1.798	1.048	

Table 1: Fractional resolution results for comparison of GEANT4 versions. These results are obtained from the fits shown in figures 5

this significant is not expected. As many different processes contribute to electromagnetic showering more investigations were needed to understand exactly which part of the simulation is giving rise to this discrepancy.

4.1.1.2 Sampling Fraction and Shower Profile Investigations In order to understand the unexpected results described in section 4.1.1.1 it was decided that a study of the shower profiles in both active and passive layers of the calorimeter would be performed. A shower profile is a plot of dE/dX against X, where X is the longitudinal depth from the front face of the calorimeter and E is the energy deposited. Although X is often expressed in terms of radiation lengths, in this case the layer number was used because the shower profiles were in bins corresponding to the thickness of one layer of the ECAL.

Again, electrons at 13 different energies between 1.78GeV and 44.44GeV were fired into the ECAL using both GEANT4 versions. At each energy, shower profiles for active and passive layers were considered seperately and the profiles resulting from each version of GEANT4 were plotted on top of each other for comparison. The results at 1.78GeV and

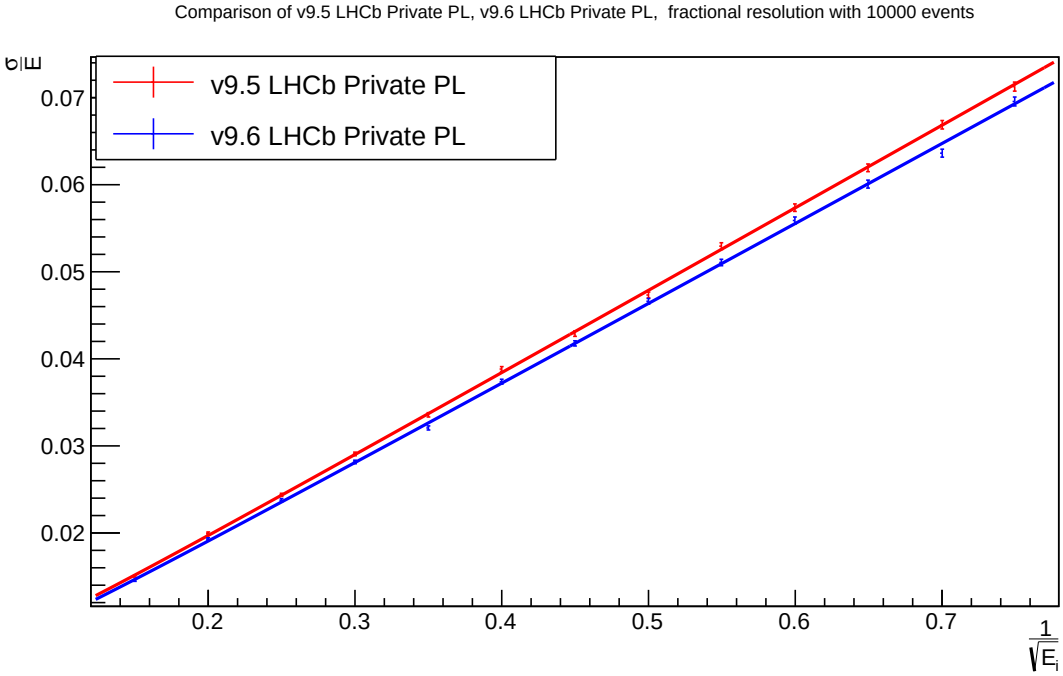


Figure 5: A plot of fractional resolution against $\frac{1}{\sqrt{E}}$ comparing versions of GEANT4. The fit is the function given in equation 3.

44.44 GeV are shown in figures 6 and 7 respectively, all other energies are shown in **??**.

It is clear from these shower profiles that GEANT4 version 9.6.p04 is depositing more energy in active layers and less energy in passive layers compared to GEANT4 9.5.p02. This pattern is observed at all energies simulated.

In order to quantify further this observed change in behaviour, the ECAL sampling fraction as a function of energy was studied for both GEANT4 versions. The sampling fraction, in this context, is defined as the energy deposited in active layers (visible energy) divided by the incident particle energy. The total energy deposited in each type of material was obtained by integrating the shower profiles over all layers. To further this quantification the ratio of energy deposited in active layers to passive layers was also plotted as a function of energy for both GEANT4 versions and the results are shown in figure 9.

4.1.1.3 Discussion of results for Geant4 version comparison The results of these sampling fraction studies are significant, both statistically and in terms of the implications for the next simulation package used by LHCb. It is very clear that the sampling fractions have changed between GEANT4 versions and this change would explain the observed discrepancy in fractional resolution seen in section 4.1.1.1. Sampling more energy would always improve the fractional resolution of an ECAL because it means any statistical fluctutations are relatively less significant. The ECAL has to be calibrated in monte carlo such that an observed value of visible energy corresponds to the correct

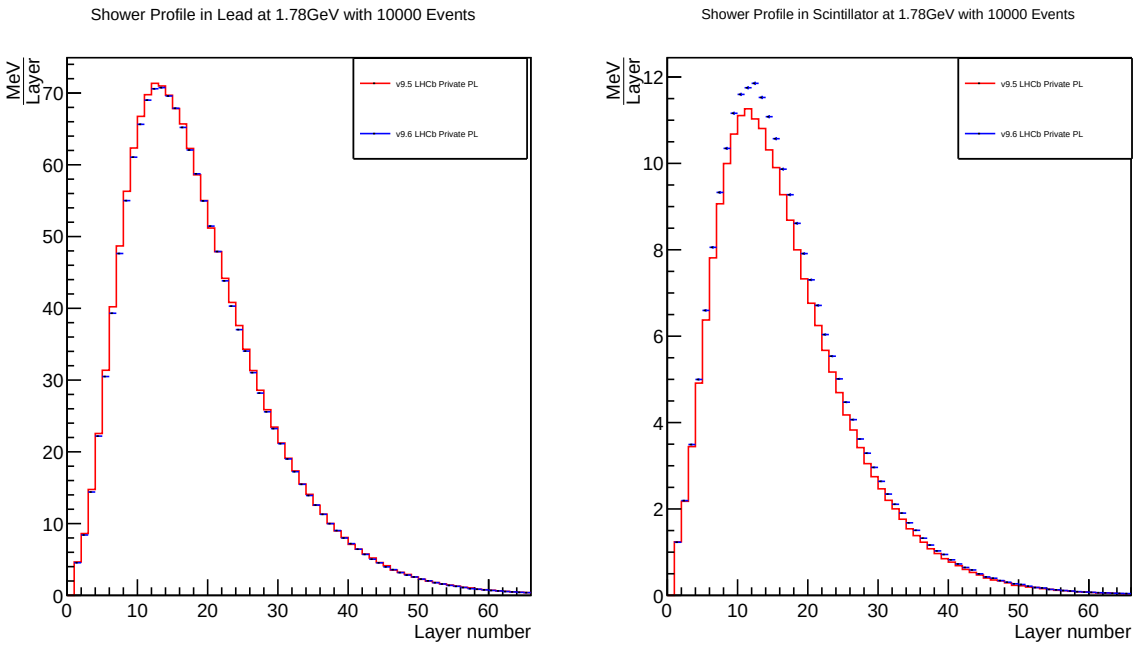


Figure 6: A comparison of shower profiles between GEANT4 version 9.5 and 9.6 using the LHCb private physics list for 1.78GeV electrons. These profiles are normalised.

incident energy. Since these results show that the visible energy of the ECAL has changed in this new version of GEANT4 this calibration will have to be performed again when the new simulation package is released.

The precise cause of these discrepancies needs to be understood because there could be implications for the simulation of other detector components.

After contacting the GEANT4 collaboration, the cause of the observed discrepancy is believed to lie with the changes to the multiple scattering model. In fact, the GEANT4 collaboration recommended that the multiple scattering models used in the LHCb private PL are changed from *UrbanMsc95* to *UrbanMsc93* below 100MeV and *WentzelVI* above 100MeV. This would bring the simulation of multiple scattering within the LHCb private PL inline with the GEANT4 *emstandard opt1* physics list.

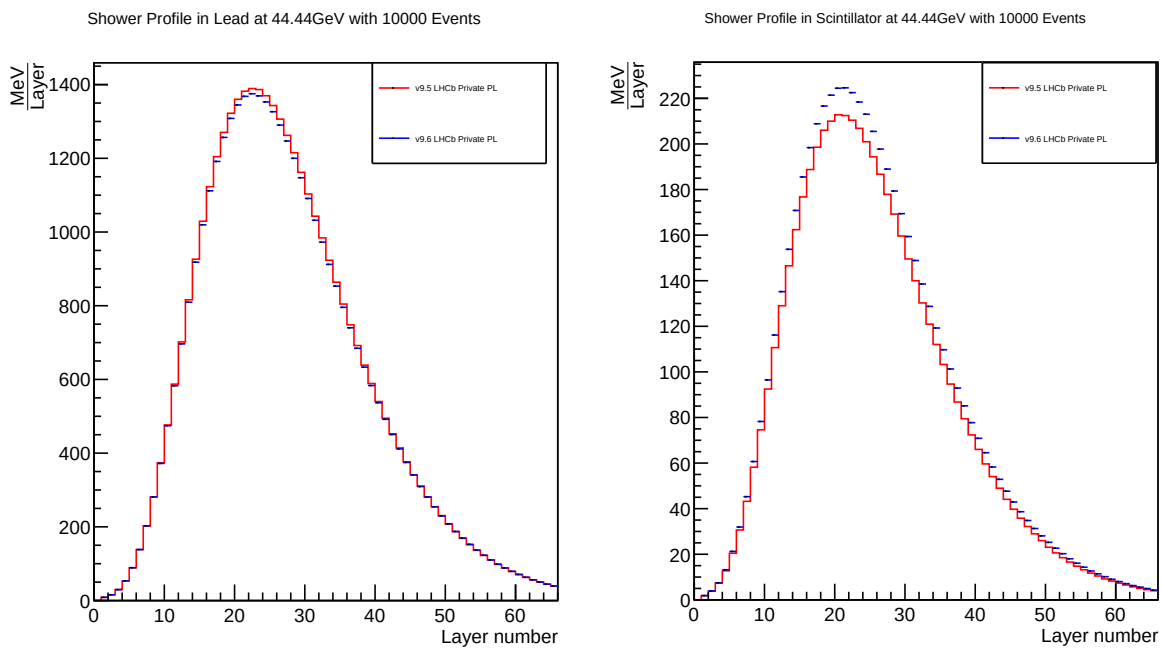


Figure 7: A comparison of shower profiles between GEANT4 version 9.5 and 9.6 using the LHCb private physics list at 44.44GeV electrons. These profiles are normalised.

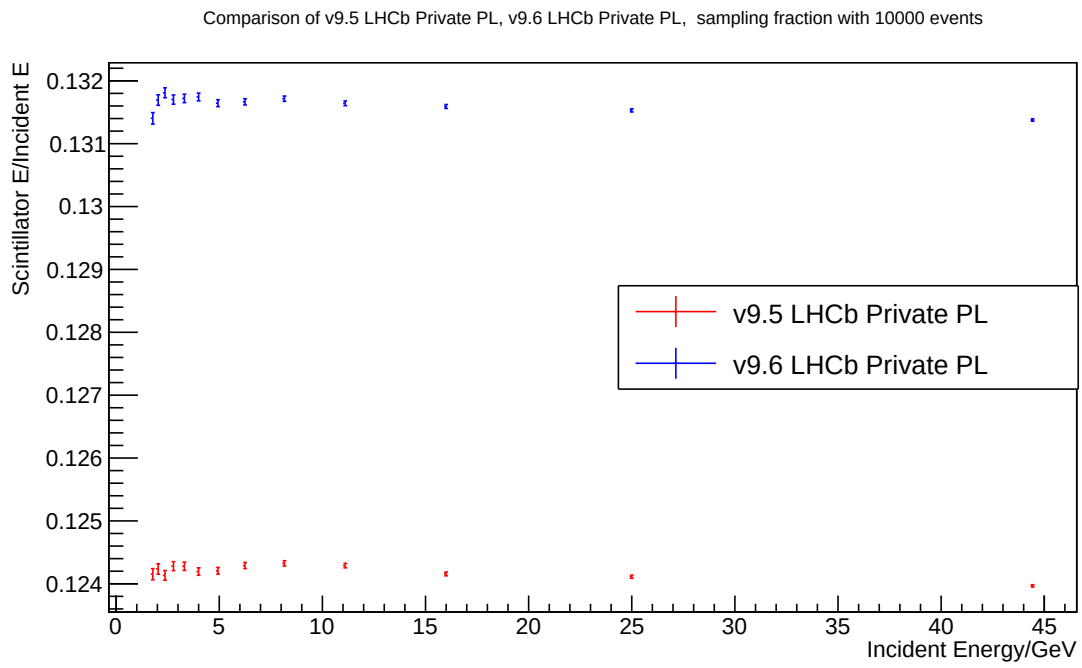


Figure 8: A plot of sampling fraction as a function of energy for both versions of GEANT4

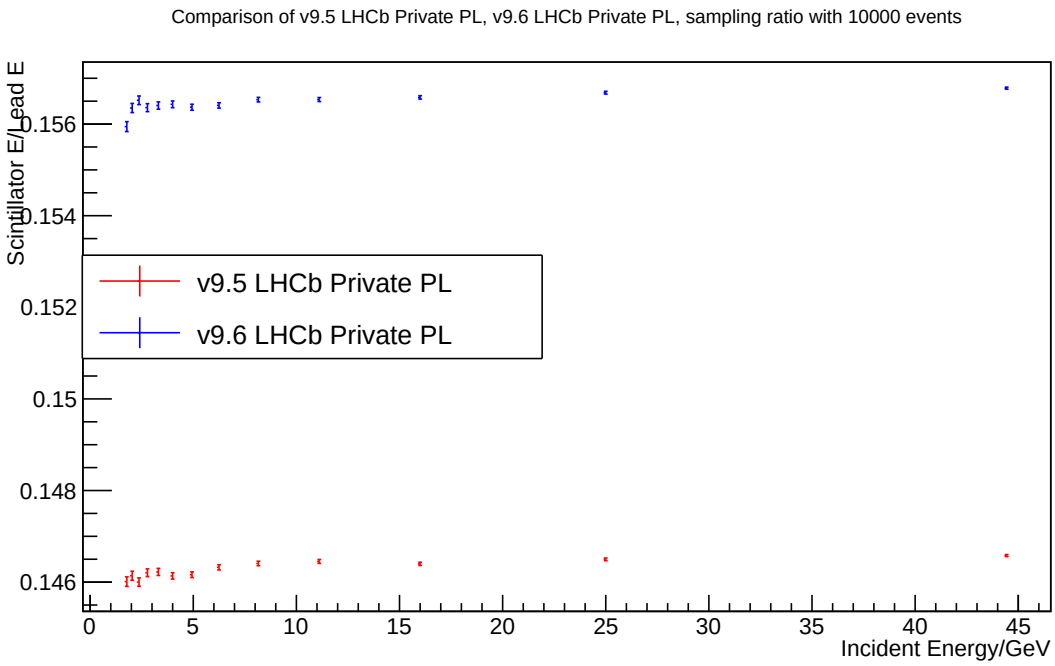


Figure 9: A plot of the ratio of energy deposited in active layers to passive layers as a function of energy for both GEANT4 versions

4.1.2 Comparison of Physics Lists

As the GEANT4 collaboration had recommended a change to the LHCb private PL it was decided a thorough investigation of other PL options would be carried out using the simple calorimeter test. This would be done using GEANT4 version 9.6.p04, so that investigations directly related to the next version of the LHCb simulation package could be carried out. The results of these investigations would help to inform the decision of which PL to use for the next release of GAUSS.

The following physics lists supplied by GEANT4 were investigated:

- *emstandard option0* This is the default electromagnetic phyics list used by GEANT4 (in general). This is designed to give optimal accuracy at medium and high energies, which are the typical energies present at HEP experiments. In version 9.6 this PL uses the *UrbanMsc95* multiple scattering model.

- *emstandard option1* This physics list is designed for HEP experiments but is described as CMS focused. This PL has moved back to using the *UrbanMsc93* multiple scattering model below 100MeV and the WentzelVI model above 100MeV in version 9.6. It also uses the *minimal* step limit type which applys looser limitations on the maximum length of a single simulation step. This results in less steps being simulated, leading to large savings in CPU time. However, this is known to cause bias in the results and this bias increases for longer production cuts [17].

- *emstandard option2* This physics list is also designed for HEP experiements but is described as LHCb focused. It uses the *UrbanMsc93* multiple scattering model and uses a different angular distribution generator for bremsstrahlung. Like *opt1* it also uses the *minimal* step limit type to save CPU time at the expense of accuracy.

- *emstandard option3* This PL provides the best accuracy of charged particle tracking in the absence of a magnetic field. It uses the *UrbanMscModel96* for multiple scattering.

- *emstandard option 4* This PL, like option 3, is designed to provide the best accuracy for charged particle tracking in the absence of a magnetic field and also uses the *UrbanMscModel96* for multiple scattering. However, it is designed to provide the highest accuracy at medium and low energies.

As well as the GEANT4 supplied PLs there is also the LHCb private physics list to consider. This is simply *emstandard option1* with the *applycuts* option disabled. However, it is based on the *opt1* PL from GEANT4 v9.4 which differs to both the v9.5 and v9.6 *opt1* PLs. Therefore, rather than just updating the multiple scattering model used in the LHCb private PL (as suggested by the GEANT4 collaboration and discussed in section 4.1.1.3) it was decided that the new version of the LHCb private PL would be the *opt1* PL from v9.6 with the *ApplyCuts* option disabled. Consequently, this study investigates both the old version of the LHCb private PL (LHCbOld) and the new version (LHCbNew).

4.1.2.1 Fractional Resolution The first quantity used to test all the PLs described above was, again, the fractional resolution of the model ECAL. Just as described in section 4.1.1.1, 10000 electrons were fired into the model ECAL at 13 different energies between 1.78GeV and 44.44GeV. The same procedure as described in section **??** was again used to extract $\frac{\sigma}{E}$ for each energy and a minimum χ^2 fit to equation 3 was performed to obtain the values of 'a' and 'c'. The simulation was run with all of the PLs described in section **??**, allowing them to all be considered as a potential PL for the next version of GAUSS.

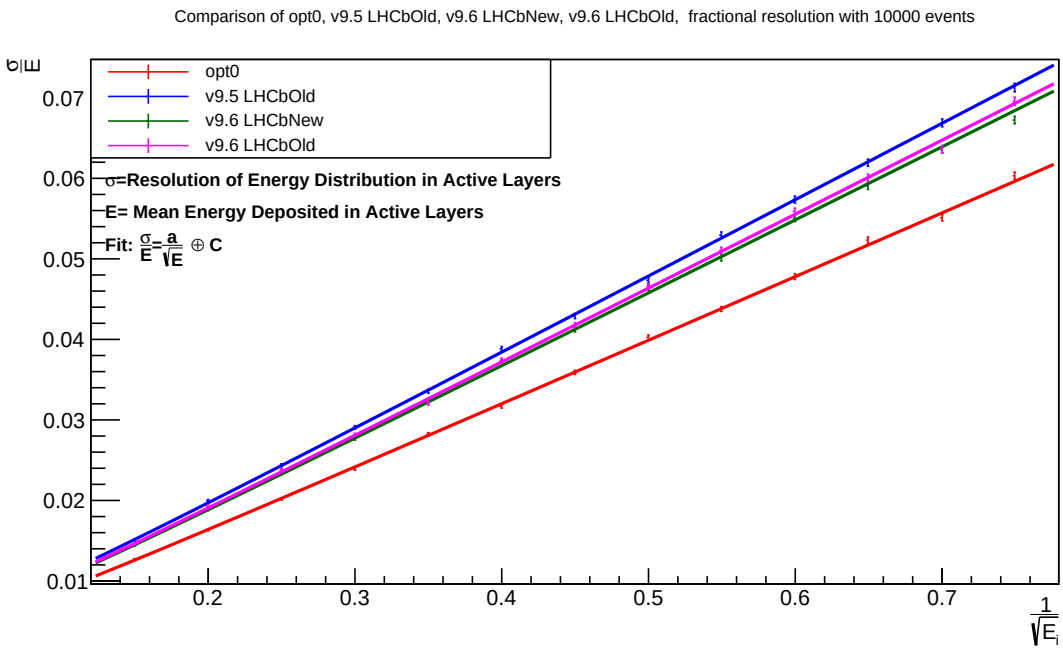


Figure 10: A plot of fractional resolution against $\frac{1}{\sqrt{E}}$ for GEANT4 v9.5 with the old PL and v9.6 with the new and old physics list options

Figure 10 shows the fractional resolution results with the two LHCb private PL options, the results from v9.5 and the opt0 physics lists are included for comparison. A similar comparison is also made between all of the standard PLs supplied by the GEANT4 collaboration, which is shown in Figure 11. A summary of all fit results is shown in Table 2

Figure 10 shows that updating the LHCb private PL to follow *emstandard opt1* has, in fact, increased the discrepancy with v9.5. Although this initially seems troublesome, one should not necessarily expect the results from the new PL to match those of the old PL and GEANT4 version because progression of results is expcected. There is nothing to say the old PL would show the best agreement with data, in fact one should expect the newer version to be the most accurate. It is, however, important that the progression of results is observed and calibrated for.

A direct comparison with data is not possible in this scenario due to limited test beam data and the absence of electronic noise in this simplifed scenario. It is known that the

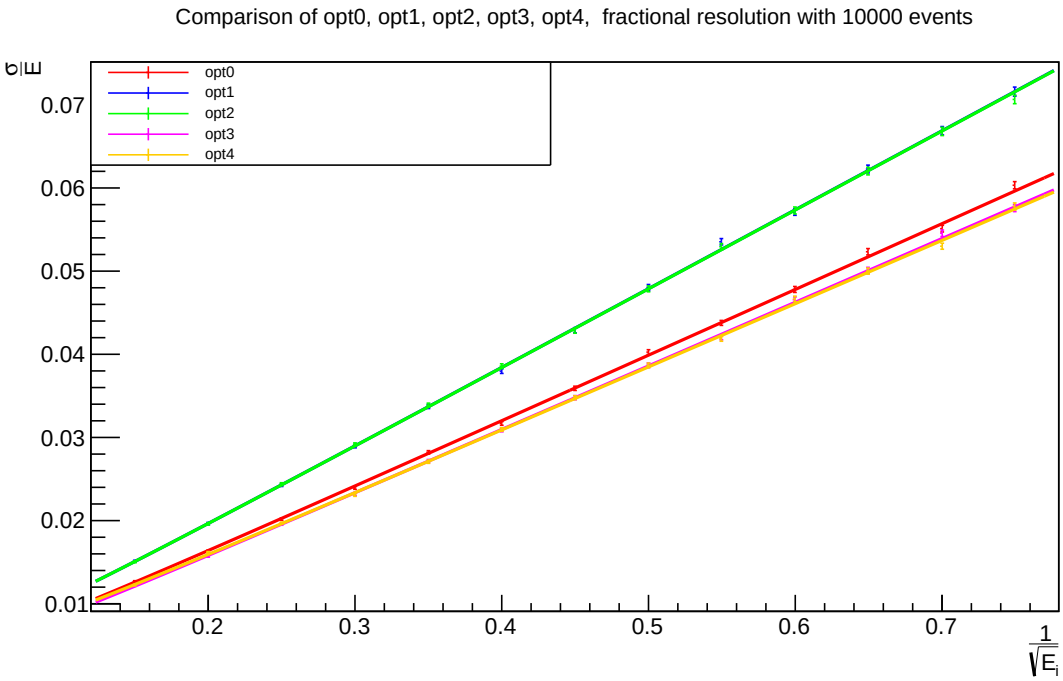


Figure 11: A plot of fractional resolution against $\frac{1}{\sqrt{E}}$ for GEANT4 v9.6 with all of the *emstandard* PLs.

PL	A term	C term	$\frac{\chi^2}{ndf}$
v9.5 LHCbOld	0.0952 ± 0.0003	0.0051 ± 0.0003	1.05
LHCbOld	0.0922 ± 0.0003	0.0049 ± 0.0003	1.80
LHCbNew	0.0910 ± 0.0002	0.0049 ± 0.0003	1.55
opt0	0.0794 ± 0.0002	0.0041 ± 0.0003	1.39
opt1	0.0953 ± 0.0003	0.0048 ± 0.0004	0.91
opt2	0.0953 ± 0.0003	0.0048 ± 0.0003	0.64
opt3	0.0768 ± 0.0002	0.0037 ± 0.0003	0.67
opt4	0.0764 ± 0.0002	0.0045 ± 0.0002	1.60

Table 2: Fractional resolution results for comparison of GEANT4 physics lists. These results are obtained from the fits shown in figures 10 and 11

LHCb private PL has to use a simplified and consequently bias multiple scattering model due to CPU time constraints. Therefore, the PLs that do not use a simplifed multiple scattering model (namely *opt0*, *opt3* and *opt4*) should provide the best agreement with data and can consequently be used to check if progress towards greater accuracy has been made. On this basis, Figure 10 shows that progress towards a more accurate fractional resolution result has been made by both moving to GEANT4 v9.6 and updating the LHCb private PL. Nevertheless, progress this large for well understood electromagnetic processes

is still surprising.

Figure 11 shows that there is a big difference in results between models that use a simplifed multiple scattering model and those that don't. There is almost no difference between *opt1* and *opt2*, both of which **do** use a simplified multiple scattering model. Likewise, *opt3*, *opt4*, and *opt0* show results that are within 5% of each other, all of which **do not** use a simpified multiple scattering model. This is useful confirmation that, as expected and observed by GEANT4 themselves, simplifying the multiple scattering model with the *minimal* step limit option introduces bias to the result [17].

4.1.2.2 Sampling Fractions and Shower Profiles

Using the same procedure as descirbed in section 4.1.1.2 shower profiles were created for all PLs under investigation.

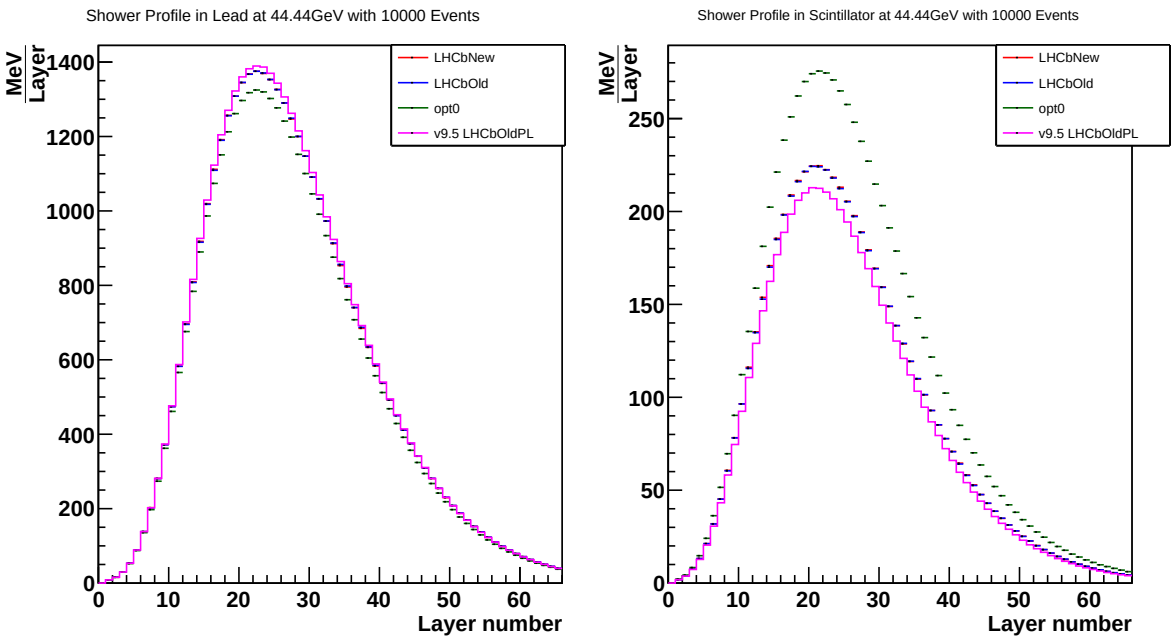


Figure 12: Shower profiles in Lead and Scintillator layers of the ECAL using LHCb private PLs and GEANT4 PLs without simplified multiple scattering models.

Figure 12 shows shower profiles in lead and scintillator layers at 44.44GeV for the LHCb private PLs; *opt0* and the v9.5 results are included for comparison. This shows that changing to GEANT4 v9.6 has moved the shower profiles towards, what is understood to be, the results with the most accurate PL. However, in the case of shower profiles the results from the new and old LHCb private PLs are too close to draw any conclusions about which is the more accurate.

Figure 13 shows a comparison of shower profiles, in both mediums, for all of the *emstandard* PLs. This shows a significant discrepancy between PLs with and without a simplified multiple scattering model. This is again useful confirmation that a simplified multiple scattering model does unfortuantely introduce bias to the results.

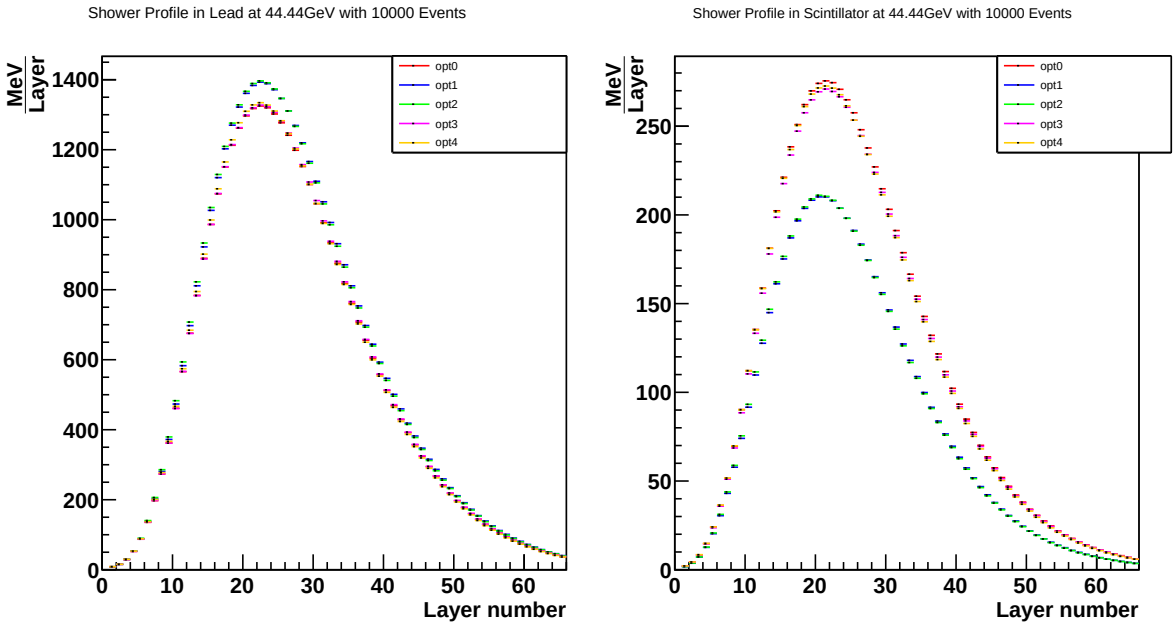


Figure 13: Shower profiles in Lead and Scintillator layers of the ECAL using GEANT4 supplied PLs

As in section 4.1.1.2 the shower profiles were integrated to obtain the sampling fraction and sampling ratio for each PL.

Figure 14 also shows that moving to GEANT4 v9.6 has made progress towards more accurate results for the LHCb private PLs. Unlike the fractional resolution, however, the LHCbOld PL and LHCbNew PL sampling fractions are within uncertainties of each other at the majority of energies. Therefore, for the purposes of choosing a new EM PL they have to be considered to be equally as accurate.

Figure 15 shows that the *opt1* and *opt2* physics lists produce very similar results. However, they are very different to the results produced by the more accurate PLs that do not use a simplified multiple scattering model. The same effect is observed under GEANT4 internal tests, which is a useful cross check [17].

These results shows that using the *fMimimal* option in the multiple scattering model has a large inpact on the amount of energy deposited, which is a quantity one would not intuitively expect to be effected by multiple scattering. However, as described in section 4.0.0.1, the length of a step taken before all particle properties are re calculated is the minimium of the step length for all physics processes. Consequently, the distance between positions where the energy loss is sampled can be effected by multiple scattering even though it plays no role in the energy loss calculations. With *fMinimal* applied the step lengths calculated for the multiple scattering process are longer, meaning that in some cases the step length used is also longer. Therefore, energy losses are sampled less often which decreases the accuracy of the simulation.

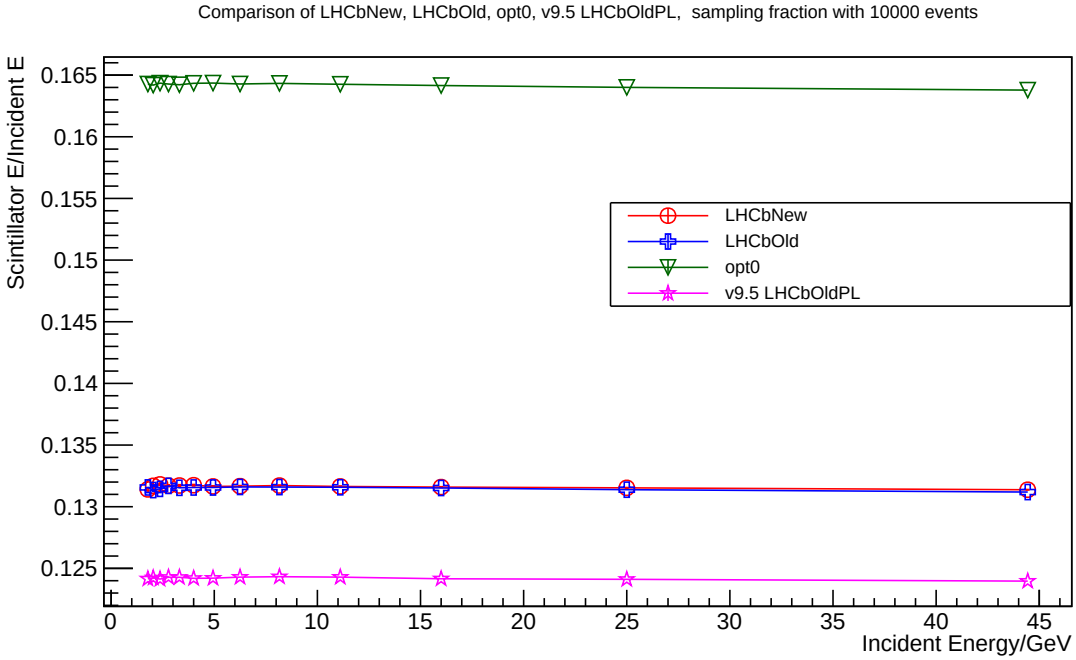


Figure 14: Sampling fraction of the ECAL as a function of energy

4.1.3 Simple Calorimeter Test Conclusions

Firstly, a significant discrepancy in fractional resolution and sampling fraction of a model LHCb ECAL has been observed between GEANT4 v9.5 and v9.6. Consequently, the LHCb calorimeter group will have to re calibrate the ECAL for the new release of GAUSS. This discrepancy is believed to be a result of the new GEANT4 version depositing more energy in scintillator and less in lead layers, most likely as a reuslt of updates to the multiple scattering models used. This discrepancy has resulted in the GEANT4 collaboration suggesting the LHCb EM PL should be updated to a new multiple scattering model. The studies in section 4.1.2 show that updating the LHCb EM PL to follow the *emstandard opt1* PL actually increases this discrepancy. The stochastic component of the fractional resolution parameterisation changes from 0.922 ± 0.0003 with the old LHCb PL to 0.910 ± 0.002 with the updated LHCb PL. However, the fractional resolution study using the most accurate PL available for the energy ranges considered here (*emstandard opt0*) yields a stochastic term of 0.794 ± 0.0002. Therefore, updating the LHCb PL would cause progression of results towards a more accurate measurement.

A choice of PL has to be made for the next release of GAUSS. Throughout this study CPU time was monitored because it is a valuable and limited resource in particle physics. It was noted that there was a siginificant increase in CPU time required when a PL without a simplified multiple scattering model was used. This increase varied from a factor of 2, to a factor of 6 compared to v9.5. Other studies have shown that approximately 45% of the time taken to simulate an event in GAUSS is spent simulating the electromagnetic shower.

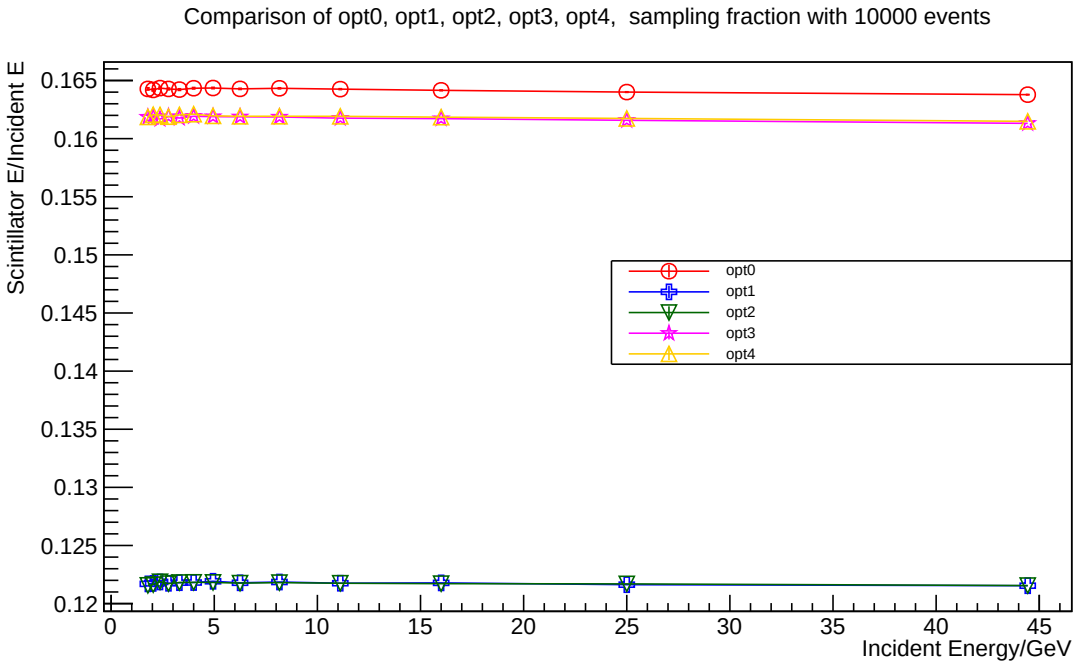


Figure 15: Comparison of sampling fractions from GEANT4 supplied PLs

Therefore, even a factor of 2 increase in CPU time is not feasible. The consequences of this are that only *emstandard opt1*, *emstandard opt2*, *LHCbOld* and *LHCbNew* are viable choices for the new PL.

Section 4.1.2 shows that *LHCbNew* produces results closest to a non simplified multiple scattering models for fractional resolution studies. Moreover, *LHCbNew* and *LHCbOld* both show closer agreement to the unbias PLs for sampling fractions and sampling resolutions compared to *opt1* and *opt2*. Therefore, either of the LHCb private PLs were deemed to be a better choice.

Furthermore, LHCb has previously experienced issues with the agreement between data and MC for the impact parameter resolution. The conclusion of an extensive investigation into this issue was to use the PL described here as *LHCbOld*. Therefore, it was decided that the new PL should stay as close as possible to *LHCbOld* to avoid introducing disagreements between data and MC elsewhere in the detector simulation. Therefore, it was decided that the updated LHCb private PL *LHCbNew* should be adopted as the new EM PL.

4.2 Multiple Scattering Test

Since the multiple scattering model has been updated in the EM PL, it was decided a direct study of multiple scattering was necessary. When a charged particle traverses a material there is a non-zero probability that it will undergo elastic coulomb scattering from a nuclei within the material. The differential cross section for this process is given by,

$$\frac{d\sigma}{d\Omega} = (\frac{1}{4\pi\epsilon_0})^2 \frac{z^2 e^4}{M^2 c^4 \beta^4} \frac{1}{sin^4(\theta/2)} \tag{4}$$

where Ω is solid angle, z is the atomic number of the material, M is the mass of the charged particle and θ is the angle through which the charged particle is scattered [18].

Except for cases where the scattering material is a very thin film, the charged particle will scatter multiple times before exiting the material. Hence, multiple coulomb scattering occurs which is more commonly known as just multiple scattering. The net effect is a lateral displacement as well as a scattering angle, as depicted in Figure 16. In this case a statistical treatment has to be used to obtain a distibrution for the scattering angle, which is defined as θ in Figure 16. This was first carried out successfully by Moliere theory, which has been shown to give very good agreement with data over a wide range of particles, materials and energies [19, 20]. Several other theories have been shown to produce consistent results but the most complete is Lewis theory, which also provides moments for the spatial displacement distribution [21].

Both the Moliere and Lewis theory give a scattering angle distribution that is gaussian for the central 98% of scattering angle values, but the tails of the distibrution fall off slower than a gaussian due to the $\frac{1}{sin^4(\theta/2)}$ term in equation 4. The width of the central gaussian is defined as θ_0 which can be approximated by the Highland formula,

$$\theta_0 = \frac{14.1MeV}{pv} z \sqrt{\frac{L}{L_R}} [1 + \frac{1}{9} log_{10}(\frac{L}{L_R})] \tag{5}$$

where p is the incident particles momentum, v is the incdient particles velocity, L is the length of the material and L_r is the radiation length of the material. This formula is an emperical formula that arises from fits to moliere theory [22].

When one simulates multiple scattering, in a similar way to the theoretical models, it is rarely possible to simulate every individual collision. It is only possible if the number of scatters is small and large amount of CPU time is available. For the latter reason, this type of simulation based on simulating every single scatter was not implemented until 2005. Unfortuantely, there is still a limited number of applications where this is a viable option. The *UrbanMsc* models get around this by using a "condensed" simulation of multiple scattering, which involves simulating a step of the particles path at a time and applying net effects at the end of each step [**?**]. More sepcifically, the angle through which the particle has been scattered and the lateral displacement are applied at the end of each step as part of the multiple scattering simulation. The scattering angle is sampled from distributions calculated using Lewis theory but no theory of a full displacement

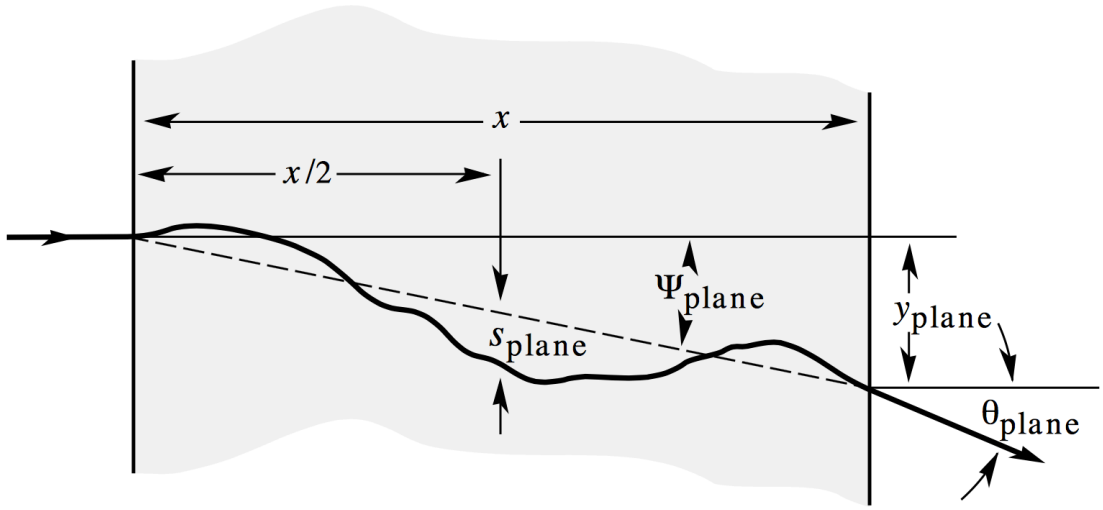


Figure 16: A diagram showing the lateral displacement and angular dispersion when a charged particles traverses a medium.

distribution exists. Therefore, GEANT4 uses its own, non-exact, algorithms to calculate the lateral displacement after each step

Another approach to simulating multiple scattering, that has been implemented more recently, is to use a "mixed" approach. This involves sampling scatterers, where the scattering angle θ is below a threshold θ_{max}, in a similar way to the "condensed" approach discussed previously. However, if the scattering angle is above θ_{max} then a single scattering approach is used. This is implemented in GEANT4 as the *WentzelVI* model.

In moving to the new EM PL the multiple scattering model for electrons and positrons has changed from *UrbanMsc95* to *WentzelVI* above 100MeV, therefore the model has moved from being condensed to mixed. However, below 100MeV the model has moved from *UrbanMsc95* to the older *UrbanMsc93* which involves some subtle changes.

To carry out a direct investigation of multiple scattering and how it is effected by these changes, the GEANT4 example TestEm5 was used. The example provides the facility to fire particles into a square sheet of material at normal incidence and study the angle of the scattered particle, to the normal, as it exits the material on the other side. A diagram of this is shown in Figure 17. The type of material used, the width of the material and

thickness of the material can all be specified. In this case the setup used was a $300\mu m$ thick sheet of silicon designed to model the LHCb VELO as closely as possible, because this is the area of the detector where the most precise tracking measurements sensitive to multiple scattering take place.

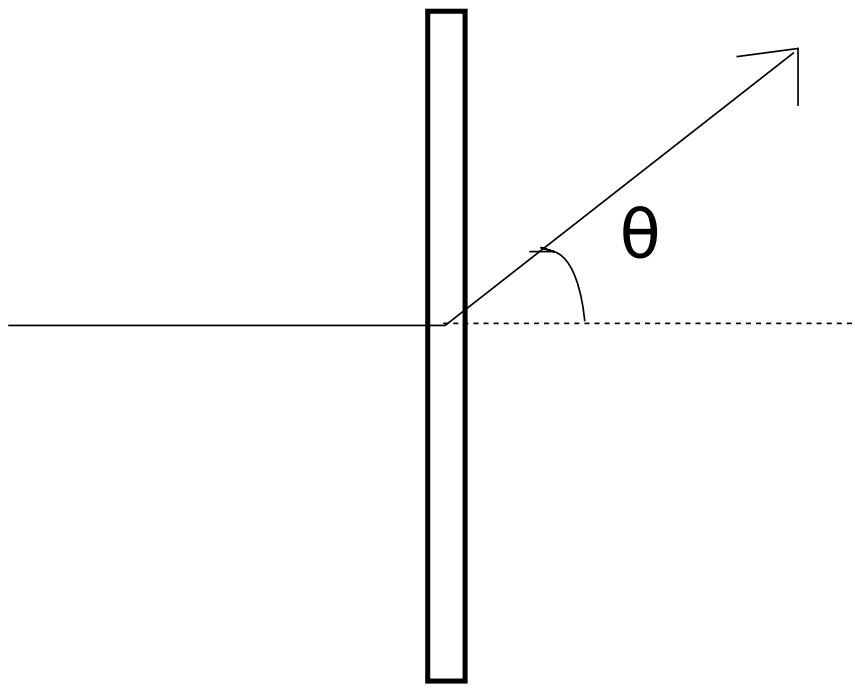


Figure 17: A diagram showing the setup of TestEm5. θ is the angle under investigation by this test

The aim of this investigation was to measure the parameter θ_0 of the scattered particles angular distribtuion for electrons at a range of energies and use it as a comparison tool between PL configurations and GEANT4 versions. The GEANT4 example is pre-configured to calculate θ_0 by taking the standard deviation of the central 98% of values after each run. However, no estimation of the uncertainty is performed by the example. As the time to simulate 10000 electrons with this minimal geometry was ¡1 second, the simulation was re-run 10000 times at each energy and configuration. θ_0 was recorded for each run and the end result was assigned as the mean of all runs with an uncertainty given by the RMS of all 10000 runs.

The energies chosen for this investigation were 1,2,3,4,5,7,9,12,15,20,25,30 and 40 GeV as these were considered representative of the expected electron energies at LHCb. The following configurations were tested with these energies: GEANT4 version 9.5 with the *LHCbOld* PL and version 9.6 with *LHCbNew* PL and verision 9.6 *LHCbOld* PL. These were chosen because it allows a direct comparison between the old release of GEANT4 and the new release of GEANT4 to be made. v9.6 *LHCbOld* was included to offer insight into the cause of any discrepancies observed.

4.2.1 Results

The results for θ_0 from running the simulation with v9.6 *LHCbOld* and *LHCbNew* were divided by the results from running the simulation with v9.5 and *LHCbOld* to make a direct comparison with the previous release of GEANT4.

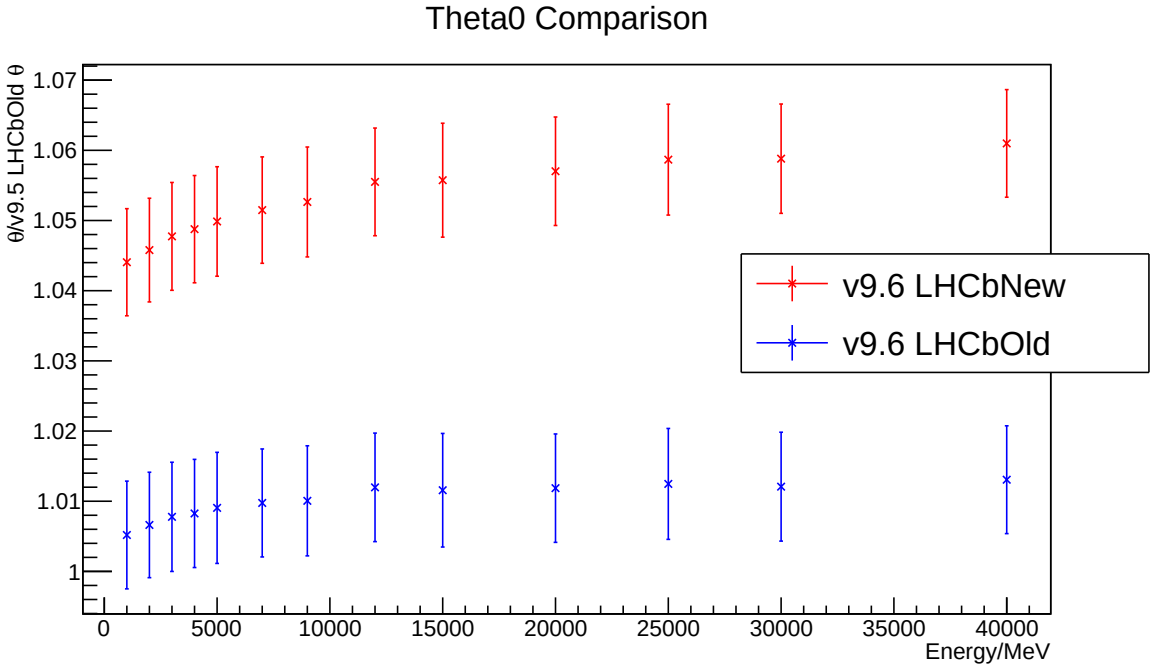


Figure 18: θ_0 as a function of energy normalised to GEANT4 v9.5 *LHCbOld*

Firstly, and most importantly, Figure 18 shows that the width of the angular distribution from the new version of GEANT4 with the new physics list has increased by $4-6\%$ compared to what was used in the previous release of GEANT4. The statistical significance of this change varies from 5.77σ at 1GeV to 7.95σ at 40GeV, meaning there is no doubt that this is due to a systematic change in the simulation process.

Secondly, Figure 18 also shows that this change in θ_0 must be largely attributed to the changes made to the multiple scattering model rather than moving to a new version of GEANT4. Although the results with the new version of GEANT4 and the old PL are systematiclly $0.5 - 1.3\%$ and upto 1.7σ larger, this is still considerably smaller than when the new PL is used with v9.6.

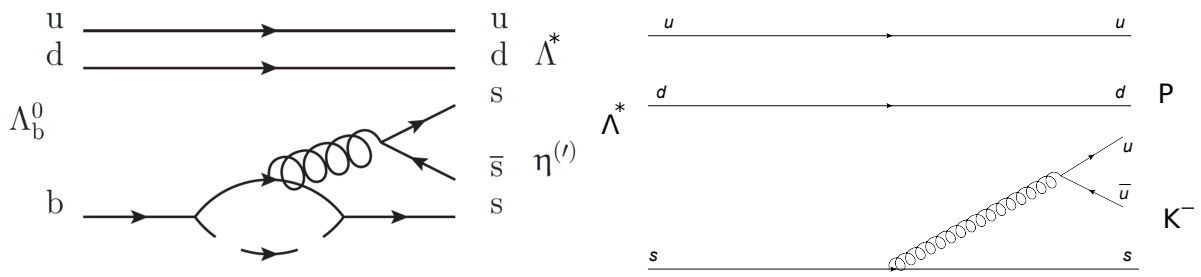
4.2.2 Conclusions

Clearly the angular distribution with the new GEANT4 version and PL has changed when a LHCb VELO like situation is considered. Unfortuantely, due to limitations of the GEANT4 example it is only possible to consider scattering from one sheet of silicon which is not completely representative of the entire VELO system (see section detector).

Therefore a full detector simulation using GAUSS is required to study this discrepancy and, importantly, the effect on the impact parameter resolution. This simulation will be carried out by the LHCb VELO group when the new release of GAUSS is made.

5 Investigation of $B^0 \to K^*(892)\eta$

The intention of this analysis is to use the decay $B^0 \to K^* \eta$ as a control channel in the search for $\Lambda_b^0 \to pK^- \eta$. As of yet a Λ_b^0 baryon has never been observed to decay to a final state involving an η' or η particle. However, evidence for the decay $\Lambda_b^0 \to \Lambda \eta$ at the level of 3σ has been observed by LHCb [23]. Unfortunately, the search for $\Lambda_b^0 \to \Lambda \eta$ suffered from low trigger efficiencies due to the long lived nature of the Λ baryon. However if the Λ_b^0 instead deacys to a short lived Λ resonance and an η particle, trigger efficiencies are expected to be higher. If this intermediate resonance is the $\Lambda^*(1520)$ or heavier, it can then decay to pK^- giving the final state that is the subject of this analysis. The dominant feynman diagram for this decay is shown in Figure 19a, where the b$\to$ s transition is mediated by an electroweak loop.



(a) Λ_b^0 decaying to Λ^* and $\eta^{(\prime)}$

(b) Λ^* decaying to pK^-.

Figure 19: Feynman diagrams showing how the rare decay often proceeds.

In order to measure (or set a limit on) the branching fraction of $\Lambda_b^0 \to pK^-$ the number of events observed has to be measured relative to the yield of a decay with a know branching fraction. This is because it is incredibly difficult to measure the efficiency of the detector well enough to make an absolute measurement. $B^0 \to K^* \eta$ is a good control channel for this search because it has only a pion/proton difference in the final state meaning minimal efficiency differences between the decays will need to accounted for when the relative branching fraction is calculated. The world average of the branching fraction is currently $1.59 \pm 0.10 \times 10^{-5}$, which means the uncertainty on this branching fraction should not be dominant in the measurement of the branching fraction for the rare mode. This branching fraction measurement has been made, most recently, at BaBar and Belle [24, 25].

The intention of this analysis is to use the full $3fb^{-1}$ dataset collected by LHCb in run I of the LHC. However, the $1fb^{-1}$ of data collected at a 7TeV centre of mass energy (CoM) and the $2fb^{-1}$ of data collected at an 8TeV centre of mass energy will be treated seperateley at least for the selection stage of this analysis. This is because the differences in kinematic variables between these two CoM energies would make a common selection for both sub-optimal. So far only the $2fb^{-1}$ of 8TeV data has been considered, therefore all results in this report are based on that dataset.

5.1 Decay Reconstruction

The K^* will be reconstructed in the channel $K^* \to K^+\pi^-$ which has a branching fraction of $99.901 \pm 0.009\%$ [26]. The η will be reconstructed in the channel $\eta \to \pi^+\pi^-\pi^0$ where the π^0 decays to 2 photons, which has a combined branching fraction of $21.83 \pm 0.28\%$ [26]. This is not the dominant decay mode of the η particle but the 2 decay modes with higher branching fractions contain only neutral particles in the final state. This means there is no tracking information available and the LHCb calorimeter (unlike the ATLAS calorimeter) does not provide any directional information for the photons it detects. Consequently, reconstruction efficiencies for all neutral final states are significantly lower and backgrounds are higher, which makes the $\pi^+\pi^-\pi^0$ channel the optimal choice.

Each of the final state particles in the decay chain are assigned a mass hypothesis based on information from all of the LHCb sub detectors before being combined to form their respective mother particles. The properties of the mother particle (e.g Energy, momentum, flight distance) are then determined with a fit to the daughter particles; the χ^2 of this fit can then aid the choice of candidate decays.

The combination of particles to form a B^0 candidate described is all done by the LHCb reconstruction packages and the initial, loose, selection of B^0 candidates is based on this reconstruction. However, a better mass resolution can be achieved by refitting the entire decay tree of B^0 candidates whilst applying extra constraints. This technique was first used by the BaBar and is practically implemented with a Kalman fitter [27]. Most commonly, any intermediate daughter particles are constrained to have their known masses and all tracks are constrained to come from the position where the hard pp interaction took place (primary vertex). In this analysis all daughter particles are constrained to come from the primary vertex and the η mass is constrained to the current world average $(547.826 \pm 0.018 MeV)$ [26]. The K^* mass is not constrained because it is a wide resonance; the natrual width of the K^* is around 50MeV. It is impossible to reconstruct a particle with a better resolution than what is dictated by it's natural width. Consequently, one can expect correctly reconstructed K^* particles to still have a mass a considerable distance from the PDG value. Therefore, constraining the K^* candidate to its PDG value can bias the refitted B^0 mass.

5.2 Event Selection

The first stage of any analysis is selecting a subset of the LHCb dataset that contains a high purity of events of interest to the analysis. Essentially, one has to mimise the number of background events that pass a set of selection criteria whilst keeping as many signal events as possible. These selection criteria are composed of several stages.

5.2.1 Stripping Selection

As the data collected by LHCb in run I is of the order of 10s of Petabytes it is impossible for every user to start with the full dataset. Therefore, loose selection criteria need to be applied to the full dataset to produce a subset of data that can be worked with. A set

of these loose selection criteria are refered to as a stripping line, some of which are quite general (e.g dimuon final state) whereas others are for a specific analysis. The amount of CPU time taken to run over the full run I dataset is very large, therefore stripping takes place when the data is taken and then incremental re-strippings take place no more than a few times a year. The stripping process is always run centrally. The cuts that go into a stripping line are determined such that they give maximal signal efficiency whilst keeping within the amount of per event CPU time allocated to that stripping line. The signal efficiencies are assessed by applying the same selection to simulated (MC) signal events. In the case of this analysis, the stripping line was already completed before this analysis started otherwise the deadline for submitting stripping lines would not have been met.

The cuts applied as part of the stripping selection used for this analysis are shown in Table 3.

Particle	Cut	Value
B^0	$\chi^2_{\rm vtx}$ <	15
	DOCA χ^2 <	20
	$p_{\rm T}$ >	1500 MeV
	DIRA >	0.9995
	IP χ^2 <	20
	m(B^0)±	750 MeV
η	$\chi^2_{\rm vtx}$ <	15
	DOCA χ^2 <	20
	$p_{\rm T}$ >	2000 MeV
	m(η)±	100 MeV
Track	$p_{\rm T}$ >	300 MeV
	Ghost Prob <	0.5
K^*	$\chi^2_{\rm vtx}$ <	9
	IP χ^2 <	5
	$p_{\rm T}$ >	1200 MeV
	m(K^*)±	100MeV
	Daughter $p_{\rm T}$ >	500 MeV

Table 3: The stripping selection applied

The meaning of these variables is as follows:

- χ^2_{vtx} is the χ^2 per degree of freedom of the initial vertex fit used to determine the properties of the daughter particles (see section).
- **DOCA** χ^2 requires the χ^2 of the distance of closest approach between all possible pairs of particles that form the B^0 to be less than the given value. See Figure 20.
- P_t: Momentum transverse to the beam pipe.
- **DIRA**: The cosine of the angle between the B^0 momentum direction and the vector between the most likely primary vertex and the B^0 decay vertex.
- **IP** χ^2: The χ^2 of the impact parameter (see Figure 20).
- **m(X)**$\pm$: Requires the mass of the candidate to lie no more than this value from the PDG mass.

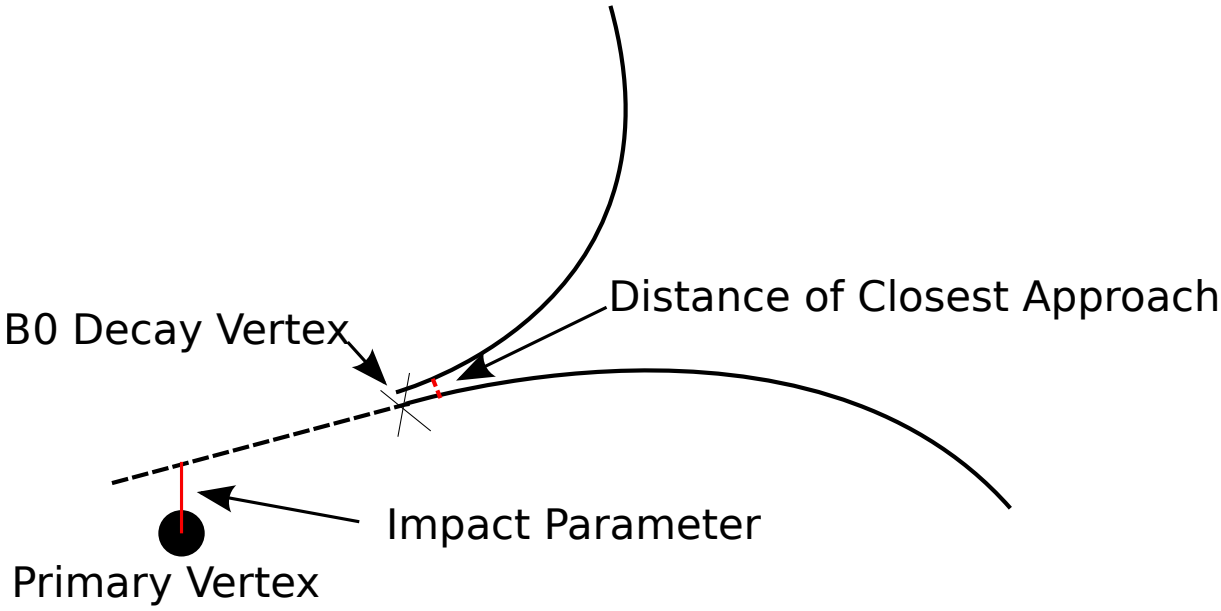


Figure 20: A schematic diagram showing DOCA and IP in a typical B^0 decay.

- **Ghost Prob**: The probability that a track is fake, as determined by the LHCb tracking system.

In addition to the cuts shown in Table 3, candidates are required to pass either of the "HLT1TrackAllL0Decision" trigger lines which are explained in section 5.2.2.

The efficiency of this selction can be assessed by applying it to signal MC; the efficiency is then simply $\frac{No.TrueSignalEventsPassingStripping}{No.Generated}$.

The stripping efficiency is then found to be 0.00317 ± 0.00002, which is a quite typical stripping efficiency at LHCb.

5.2.2 Trigger

LHCb has many trigger selections (trigger lines) that are applied as part of the triggering process, each of which is tailored towards a set of analyses with common properties. If an event passes any of these trigger lines it is stored, but the specific trigger lines that it passed are flagged. Therefore, the next step of the analyses is to require events to pass trigger lines more specific to this analysis.

When an event passes a trigger line it can be described as triggered on signal (TOS) or triggered independent of signal (TIS). TOS is when a candidate signal particle, or any component of the event contributing to it's reconstruction, causes the event to pass the trigger line. More specifically, a minimum of 70% of the trigger objects (e.g track, calorimeter cluster) that caused the event to pass the trigger line have to belong to the signal candidate. An event is classified as TIS if objects not used to reconstruct the signal candidate caused the caused the event to pass the trigger line [28]. This TIS and TOS

information can be used to perform a data driven assessment of triggering efficiencies, although this is not performed for this analysis(yet) [29].

In order to pass the trigger selection, an event must satisfy the following criteria: *(L0GlobalTIS* ***OR*** *L0HadronDecision TOS)* ***AND*** *(Hlt1TrackAllL0Decision TOS)* ***AND*** *(Hlt2Top2BodyBBDTDecision TOS* ***OR*** *B0 Hlt2Topo3BodyBBDTDecision TOS* ***OR*** *Hlt2Topo4BodyBBDTDecision TOS)*

The L0 Hadron Decision TOS line is required because the final state of this analysis contains 4 hadrons, therefore events that do not pass this trigger line are unlikely to be signal. The Hlt2 trigger lines make use of the fact that this stage of the trigger is applied offline and apply a boosted decision tree (BDT) using topological variables, more description of BDTs is given in section 5.2.3.

As part of the monte carlo production process the LHCb trigger system is emulated, therefore MC events are also flagged with the trigger lines they pass. This means the signal efficiecny of this trigger selection can be assessed simply by applying it to MC. The signal efficiency of this trigger selection is found to be 0.715 ± 0.006. This is a good trigger efficiency for LHCb, and considerably higher than the approximately 20% efficiency that hampered the analysis of $\Lambda_b^0 \to \Lambda\ \eta$ [23].

5.2.3 Multivariate Selection

There are many kinematic and detector variables associated with every event that can be used to discriminate signal events from background events. However, applying cuts to these variables individually does not provide optimal signal and background seperation. Furthermore, there are often correlations between the variables that are difficult to account for when individual cuts are used. For both of these reasons a multivariate selection procedure is applied in this analysis which makes use of all the available variables at the same time. It has been seen in many other LHCb analyses that the most effective type of multivariate analysis is a boosted decision tree(BDT). BDT's were first developed as a way of perforiming particle identification at the MiniBooNE neutrino experiment but have since been used very widely within HEP experiments as well [30].

5.2.3.1 Decision Tree A simple decision tree (DT) is just a set of binary cuts applied one after the other to decide if an event is signal or background. Each cut is known as a node. The positions of these cuts are chosen (trained) based on control samples of data that are known to be signal and background. The gini index is given by, $G = P(1 - P)$ where P is the purity of a subset of events. At each node the positon of the cut is chosen to optimise the criterion C which is given by,

$$C = G_{node} - G_{passing} - G_{notpassing} \tag{6}$$

An example of a trained DT is shown in Figure 21. Every event in this tree has the initial selection criteria at the top of the tree applied, and then the further cuts are applied depending on the output of the first. Every event is then classified as signal or background by this tree.

The specific tree shown in Figure 21 has been used to classify artificial events just based on the variables x and y. Figure 22 shows a scatter plot of x against y populated by both "signal" and "background" events. The positions of the cuts applied by this DT are also shown; it can be seen they seperate signal and background reasonably well. If more variables are used, more nodes can be added to the decision tree and the cuts become multi dimensonial.

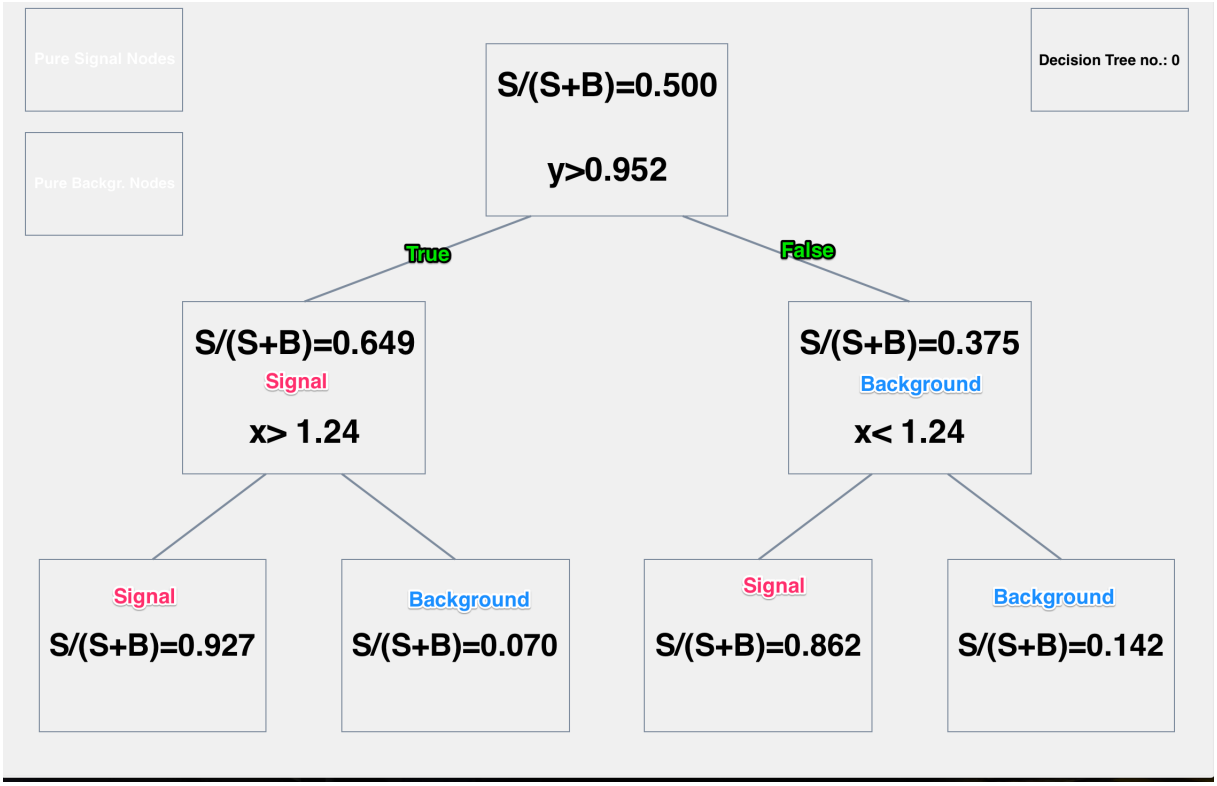


Figure 21: A single decision tree that is classifying events as signal or background based on the artificial variables x and y

It should be noted that the artificial variables provide better seperation than any found in real data but despite this there are still background events that are classified as signal events and vica versa. Also, it is not unusual for some analyses to use as many as 30 variables in a multi variate analysis which would make a single decision tree very complex. The biggest reason why a single decision tree is not used, however, is because the training of a single tree is not stable; it is highly susceptible to statistical fluctuations.

5.2.3.2 Boosting These issues are solved by the boosting process, which involves training many decision trees (400-1200), scoring the effectiveness of the tree and returning a single value that is the weighted average of the response from all the trees. In this analysis the TMVA package is used to train and apply BDTs which uses the Adaboost algorithm [31].

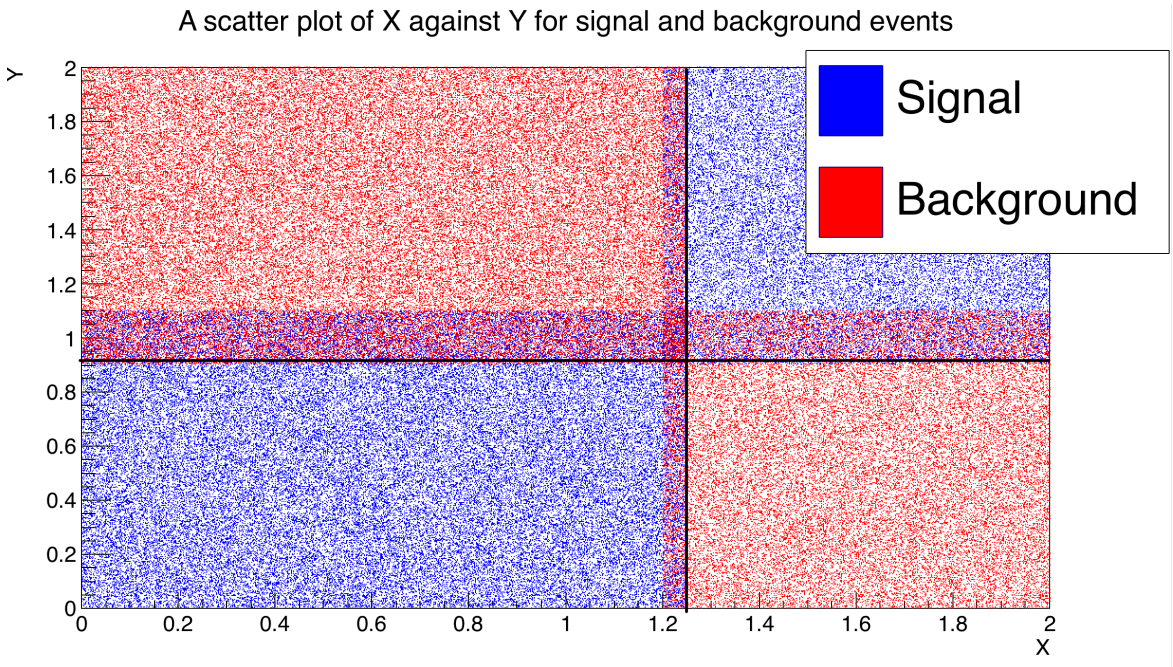


Figure 22: A scatter plot of x against y for the artificially generated events. The black lines show the cuts applied by the trained decision tree in Figure 21

After all the DTs are trained, the first step in the boosting process is to assign all events a weight $W = \frac{1}{m}$ where m is the number of events in the control sample. To caluclate the score, α, of each DT the fraction of events that are mis identified,E, is calcualted for every DT. This is another reason why control samples of known signal and background events are required for training. A DT with a low value of E is chosen as the first DT to score and the value of α for that DT is calculated as,

$$\alpha = \beta ln(\frac{1-E}{E}) \tag{7}$$

where β is a constant. Next, the weights of events mis identified by the first DT are re assigned to be,

$$W = We^{\alpha} \tag{8}$$

and all event weights are renormalised. This lends more weight to mis idenitifed events. The same scoring process is then iteratively applied to every DT trained, but mis identified events are only re weighted until the total weight of all mis identified events reaches 50%.

For a given event the final BDT value assigned is given by,

$$V = \sum_{i}^{N_{Trees}} \alpha_i V_i \tag{9}$$

where α_i is the score assigned to the i'th decision tree and V_i is the output of the i_{th} decision tree which is -1 if the DT classifys the event as background and 1 if it is classified

as signal. This means the final value assigned to an event by the BDT is between -1 and 1 with 1 being more likely to be signal. After the BDT is applied, there is a single BDT variable that can be cut on.

5.2.3.3 Application for $B^0 \to K^* \eta$ A BDT was trained, and applied to candidates that passed the trigger and stripping selections to discriminate signal from combinatorial background. Truth matched signal MC was used as a signal control sample and data with a refitted B^0 mass greater than 5500MeV was used as a background control sample to train the BDT. This mass was chosen because it is far enough away from the real B^0 mass to definitely be background whilst still keeping a reasonable sample size. Data from the lower mass sideband was not used for training as it is highly likely to contain partially reconstructed backgrounds rather than pure combinatorial background. This would be a problem because BDTs are only effective for discriminating between two classes of events.

The variables used in the BDT are shown in Table 4. The TMVA package ranks all input variables in order of importance based on the gini value and the number of times it is the variable in the top node of the tree. These variables were chosen by adding all variables that could possibly discriminate signal from background and then iteratively removing the lowest ranked variables. This was done until the lowest ranked variable still made a significant contribution to the signal and background seperation. Care was also taken to ensure variables that are poorly modelled in MC were not used, as this could lead to false seperation. Plots comparing a selection of these input variables in the signal and background control samples are shown in Figure 23.

The overall ouptut of the trained BDT for the control samples is shown in Figure 24. Firstly, this plot shows that the BDT is able to provide good seperation between signal and background. This plot also shows the overtraining check that TMVA performs. Overtraining occurs when the BDT becomes too sensitive to statistical fluctuations in the input variable distributions. In the case of overtraining, the BDT shows excellent seperation for the specific control samples used but worse seperation for the input variables in general. TMVA checks for this by splitting the control samples in half before training, and only one half is used to train the BDT. When the BDT is trained, it is applied to the remaining half of the control sampeles, and in the case of no overtraining the distirbution of the BDT response varaible should be the same for both halves of the control samples. Figure 24 shows the BDT response distributions for both the training and test samples. These are considered to be consistent, there are no systematic disrepencies between the distributions. The Kolomogrov-Smirnov test shown on the plot should be ignored, as internal LHCb studies have shown this to be unreliable [**?**]. This trained BDT was then applied to all events that passed the stripping and trigger cuts.

Variable	Explanation
B0 PT	Transverse momentum of B^0 candidate
log(1-B0 DIRA OWNPV)	natural log of one minus the DIRA angle (see section 5.2.1)
B0 ENDVERTEX CHI2	The χ^2 of the B^0 endvertex from the least squares fit used to assign the daughters to the B0 endvertex
log(B0 PVFit Chi2)	The natural logarithm of the χ^2 of the fit used to refit the entire decaytree described in section 5.2.1
B0 eta	The pseudorapidity of the B^0 candidate
B0 FD OWNPV	The flight distance of the B^0 between the primary vertex and decay vertex
log(B0 TAU CHI2)	The χ^2 of the B^0 lifetime.
Kstar CosTheta	The cosine of the angle between the K^* candidates momentum in the B^0 rest frame and the B^0 momentum in the lab frame.
Kstar eta	The pseudorapidity of the K^*
piminus0 OWNPV CHI2	The χ^2 of the π^- belonging to the η 'vertex
pi0 CosTheta	The cosine of the angle between the π^0 candidates momentum in the η rest frame and the η momentum in the lab frame.
gamma PT and gamma0 PT	The transverse momentum of the photons from the π^0
log B0RF KPlusKst costheta	The logarithm of the cosine of the angle between the K^+ and the K^* candidate in the rest frame of the B^0.
etaRF PiPlusPiMinus0 costheta	The cosine of the angle between the π^+ and π^- from the η in the rest frame of the η.
gamma0 CL and gamma CL	The probability that the photons from the π^0 are actually photons. This is assigned by the LHCb calorimeter system.
B0 Sum Track PT	The sum of the transverse momentum for all tracks in a candiate event (π^-,K^+,π^-,π^+).
B0 Sum PCHI2	The sum of the chi^2 probabilities for each track belonging to the event.
B0RF KPlusPiMinus costheta	The cosine of the angle between the K^+ and π^- in the rest frame of the B^0

Table 4: The variables used in the BDT, a diagram of the decay is shown in Figure 20

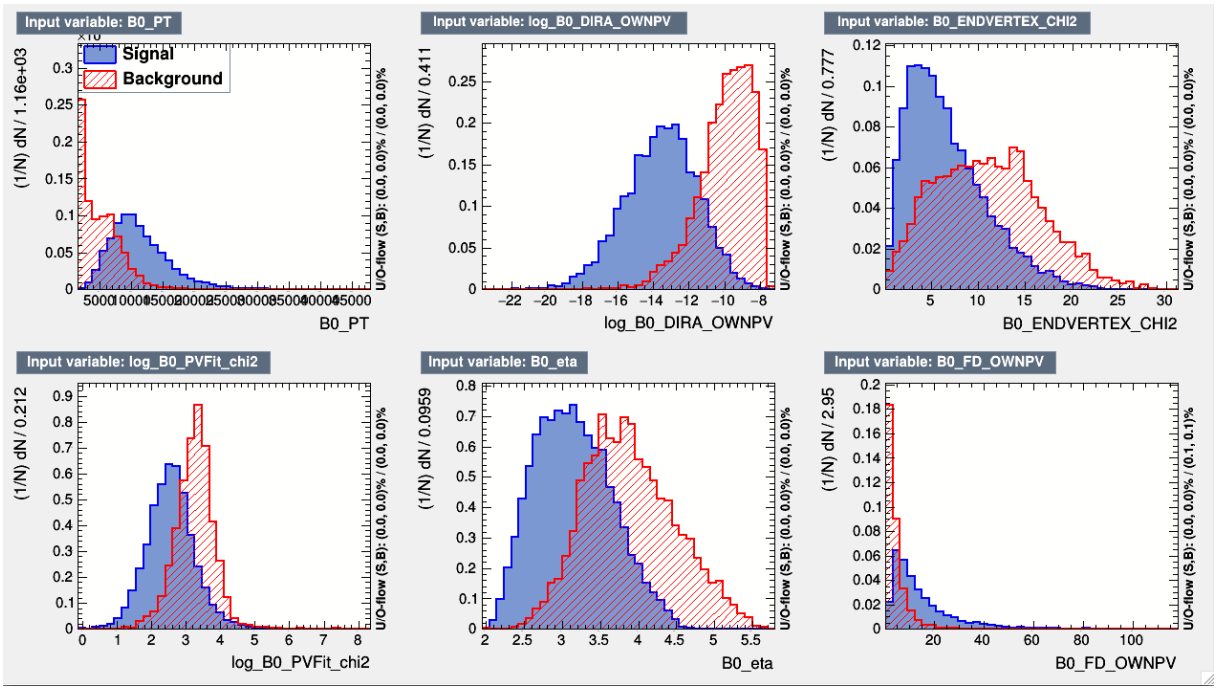


Figure 23: Plots of 6 of the input variables showing the seperation between signal and background they provide

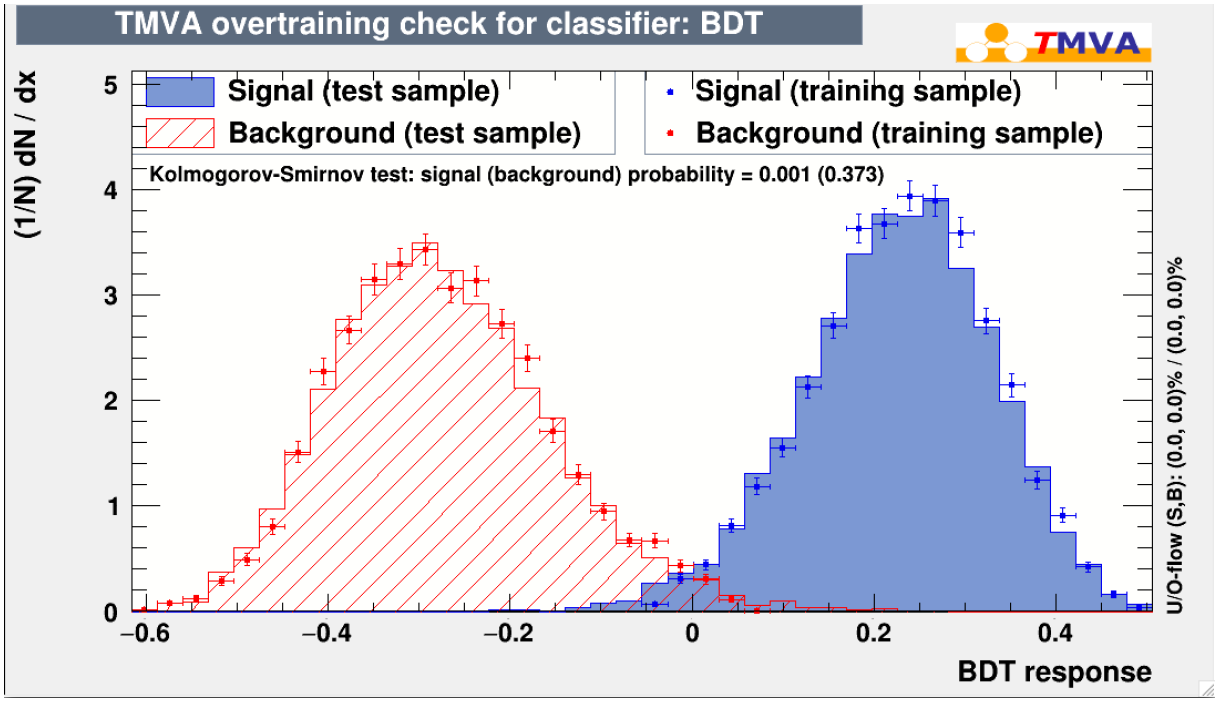


Figure 24: The output of the BDT as applied to the control samples. This also shows the overtraining check

5.2.3.4 Optimisation of BDT Cut A cut on the BDT response now needs to be applied. As the B^0 meson has been observed to decay to K^* η previously, this analysis is not performed blinded. The consequences of this is that signal significance can be used to optimise the BDT cut. The signal significance, σ, is defined as,

$$\sigma = \frac{S}{\sqrt{S+B}} \tag{10}$$

where S/B is the number of signal/background events present in a defined signal window. The values of S and B have to be extracted with a mass fit to data at each BDT cut.

In order to paramaterise the signal shape, a mass fit to truth matched MC was first performed using Crystal Ball probability density functions(PDF). The Crystal Ball experiment was the first to use this PDF, which is centrally a gaussian with a power law tail after a given cut off [32]. It is a given by the function,

$$f(x;\alpha,n,\bar{x},\sigma) = N \cdot \begin{cases} \exp(-\frac{(x-\bar{x})^2}{2\sigma^2}), & \text{for } \frac{x-\bar{x}}{\sigma} > -\alpha \\ A \cdot (B - \frac{x-\bar{x}}{\sigma})^{-n}, & \text{for } \frac{x-\bar{x}}{\sigma} \leqslant -\alpha \end{cases} \tag{11}$$

where

$$A = \left(\frac{n}{|\alpha|}\right)^n \cdot \exp\left(-\frac{|\alpha|^2}{2}\right) \tag{12}$$

$$B = \frac{n}{|\alpha|} - |\alpha| \tag{13}$$

$$N = \frac{1}{\sigma(C+D)} \tag{14}$$

$$C = \frac{n}{|\alpha|} \cdot \frac{1}{n-1} \cdot \exp\left(-\frac{|\alpha|^2}{2}\right) \tag{15}$$

$$D = \sqrt{\frac{\pi}{2}}\left(1 + \text{erf}\left(\frac{|\alpha|}{\sqrt{2}}\right)\right) \tag{16}$$

The parameter α describes the cut off where the function changes from a gaussian to a power law. A positive value of α gives a low side tail and a negative value gives a high side tail. $\bar{x}$ is the mean of the gaussian function, σ is the width of the gaussian and n can be interpreted as the power for the tail. In this fit, two crystal ball functions with opposite side tails and a common mean were used. More specifically, an extended unbinned maximum likelihood fit was used, for which the likelihood function was given by,

$$L = e^{-N_{C1}-N_{C2}} \prod_i^D N_{C1}C1(x_i,\alpha_1,n_1,\bar{x},\sigma_1) + N_{C2}C2(x_i,\alpha_2,n_2,\bar{x},\sigma_2) \tag{17}$$

where N_c is the number of signal events in each of the crystal ball functions, C1 and C2 are the two crystal ball functions with opposite side tails. x_i is the mass of the i'th

B^0 candidate. The parameters $N_{C1}, N_{C2}, \alpha_1, n_1, \bar{x}, \sigma_1, \alpha_2, n_2, \sigma_2$ are free parameters to be determined by the fit. Practically, this was performed using the RooFit package and the MINOS method was used to determine uncertainties [33].

Figure 25 shows the results of the fit performed to truth matched MC with the values of the free parameters determined by the fit. The fit fraction in this Figure is N_{C2}/N_{C1}. This shows a good quality fit to monte carlo and, crucially, it recovers the known signal yield within uncertainties.

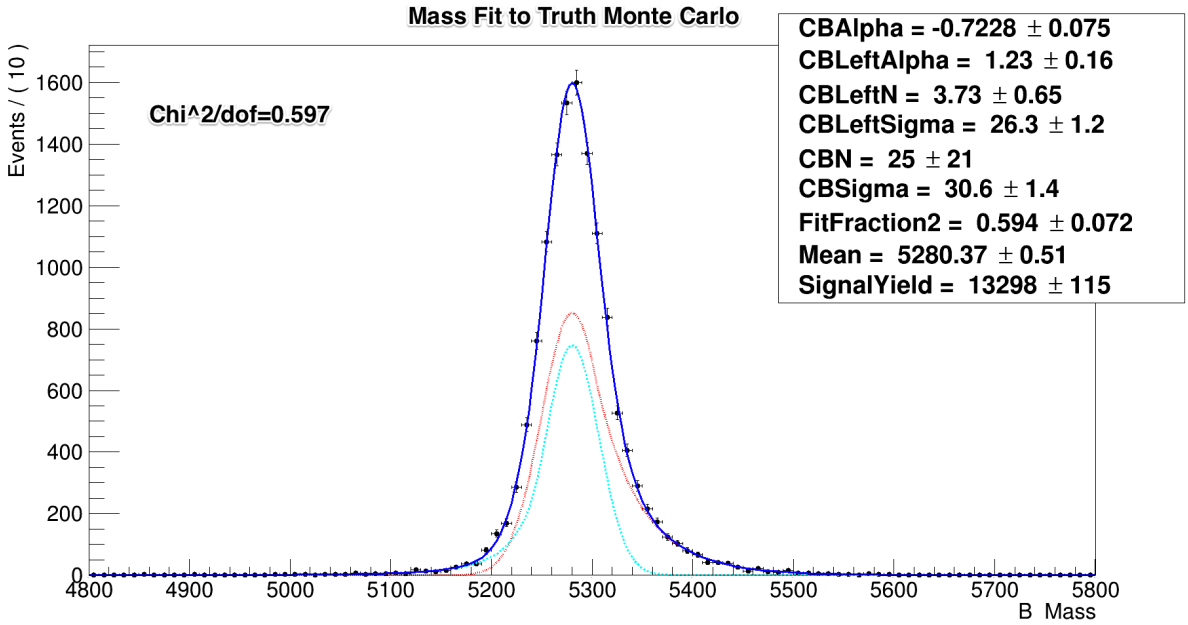


Figure 25: A plot of the truth matched monte carlo with no BDT cut and the fitted double crystal ball function. The red and blue dashed lines are the inidividual crystal ball functions, the solid blue line is the overall fit

Now that the expected signal shape is parameterised, a mass fit to data can be carried out. The combinatorial background was modelled with an exponential PDF, because this has described it well in analyses with similar final states [23]. This exponential function, with its one free parameter k, was then added to the double crystal ball signal PDF. However, for the fit to data all of the free parameters except the common mean and signal yields were fixed to those obtained in the mass fit to MC. As these values change with BDT cut the mass fit to MC was re-performed after applying every BDT cut investigated.

The signal window was then defined to be within $\pm 2\sigma$ of the B^0 mass, and the value of σ was taken to be the value obtained for the widest crystal ball function in the mass fit to MC. An extended unbinned maximum likelihood fit was then performed at each BDT cut and the number of signal and background events within the signal window were extracted at each BDT cut. This allowed the signal significance as a function of BDT cut to be determined, a plot of this is shown in Figure 26. The uncertainty on significance was calculate by propagating through the uncertainties on signal and background yields from

the fit. Figure 26 shows that the optimum BDT cut is at a value of 0.07. By applying the BDT to MC it is found that the signal efficiency of this cut is 0.934 ± 0.008, which is pleasingly high.

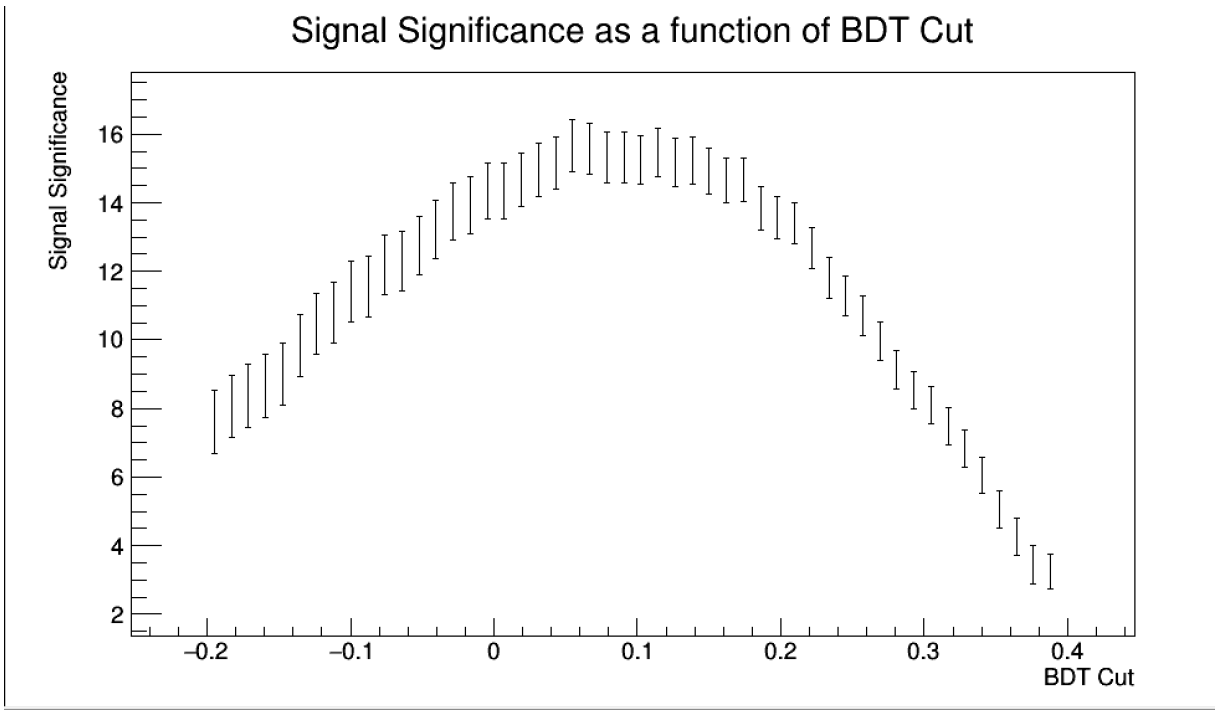


Figure 26: Signal significance as a function of bdt cut

Figure 27 shows the mass fit to data at this BDT cut. Firstly the signal peak can clearly be seen with a significance of 15.9σ and a healthy signal yield of 452 ± 29 events. However, the position of the signal peak (mean of crystal ball functions) is pulled to a higher value that is actually the limit set in the fitting procedure. Also, there appears to be features in the background other than pure (exponential) combinatorial background which are causing data points to be systematically higher/lower than the fit in several places. Therefore, the absolute values extracted from this fit can not be trusted, but this fit was deemed to provide an accurate enough estimate of the signal significane to optimise the BDT. Consequently, a BDT cut of 0.07 was applied.

5.2.4 Next Steps

The next stage of the selection is a PID selection. This uses the PID variables from the LHCb detector to remove particles that are mis identified. However, before this can take place the features in the background that are not just combinatorial need to be understood and ideally paramaterised. After the PID selection is complete multiple candidates need to be considered, which is when there is more than one B^0 candidate per bunch crossing. The way multiple candidates are dealt with will be decided after they have been investigated,

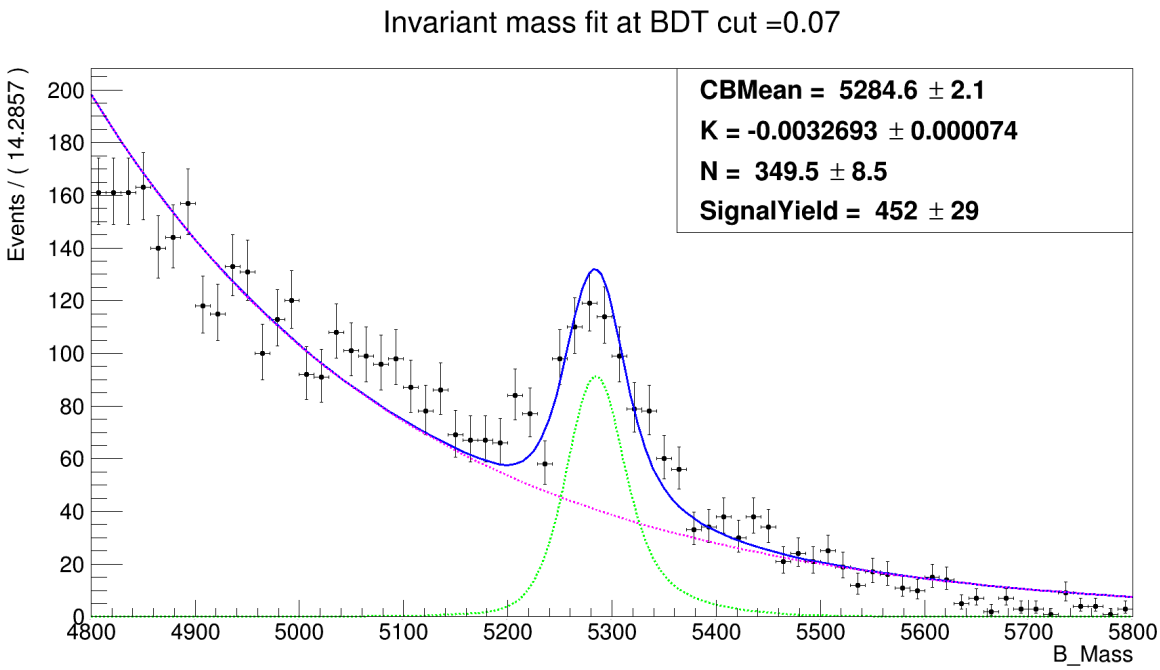


Figure 27: The mass fit to data at the optimum BDT cut of 0.07. The pink dashed line is the exponential (combinatorial) background and the green dashed line is the signal peak. K is the constant in the exponential power, CBMean is the mean of the two crystal balls used to model signal and N/SignalYield is the number of background/signal events within the 2σ signal window.

as they can arise in several ways. A selection can then also be developed for 2011 data and, if detector effects are the same, combined with 2012 data. A single mass fit can then be performed to extract the total signal yield in all $3fb^{-1}$ of data. This can then be used as a control channel in the search for $\Lambda_b^0 \to$ p K^- η.

References

[1] F. Follin and D. Jacquet, *Implementation and experience with luminosity levelling with offset beam*, in *Proceedings, ICFA Mini-Workshop on Beam-Beam Effects in Hadron Colliders (BB2013)*, pp. 183–187, 2014. `arXiv:1410.3667`. [,183(2014)], doi: 10.5170/CERN-2014-004.183.

[2] LHCb collaboration, R. Aaij *et al.*, *LHCb Detector Performance*, Int. J. Mod. Phys. A **30** (2014) 1530022. 82 p.

[3] LHCb Collaboration, A. A. Alves *et al.*, *The LHCb Detector at the LHC*, J. Instrum. **3** (2008), no. LHCb-DP-2008-001. CERN-LHCb-DP-2008-001 S08005, Also published by CERN Geneva in 2010.

[4] LHCb collaboration, R. Aaij *et al.*, *Evidence for pentaquark-charmonium states in* $\Lambda_b \to J/\psi p K^-$ *decays*, LHCb-PAPER-2015-029, in preparation.

[5] V. V. Gligorov and M. Williams, *Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree*, JINST **8** (2013) P02013, `arXiv:1210.6861`.

[6] LHCb collaboration, R. Aaij *et al.*, *Determination of the quark coupling strength* $|V_{ub}|$ *using baryonic decays*, `arXiv:1504.01568`, to appear in Nature Physics.

[7] T. Sjöstrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, `arXiv:0710.3820`.

[8] B. M. et al. *Herwig++ physics and manual*, Eur. Phys. J. C 58 639 (2008).

[9] C.-H. Chang, C. Driouichi, P. Eerola, and X. G. Wu, *BCVEGPY: An Event generator for hadronic production of the B_c meson*, Comput. Phys. Commun. **159** (2004) 192.

[10] M. Gyulassy and X.-N. Wang, *HIJING 1.0: A Monte Carlo program for parton and particle production in high-energy hadronic and nuclear collisions*, Comput. Phys. Commun. **83** (1994) 307, `arXiv:nucl-th/9502021`.

[11] C.-H. Chang, J.-X. Wang, and X.-G. Wu, *GENXICC: A Generator for hadronic production of the double heavy baryons Xi(cc), Xi(bc) and Xi(bb)*, Comput. Phys. Commun. **177** (2007) 467, `arXiv:hep-ph/0702054`.

[12] Geant4 collaboration, S. Agostinelli *et al.*, *Geant4: A simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250.

[13] R. Brun, R. Hagelberg, M. Hansroul, and J. C. Lassalle, *Simulation program for particle physics experiments, GEANT: user guide and reference manual*, CERN, Geneva, 1978.

[14]

[15] LHCb collaboration, *LHCb calorimeters: Technical Design Report*, CERN-LHCC-2000-036. LHCb-TDR-002.

[16] R. Wigmans, *Calorimetry: Energy Measurement in Particle Physics*, International series of monographs on physics, Clarendon Press, 2000.

[17] V. N. Ivanchenko, O. Kadri, M. Maire, and L. Urban, *Geant4 models for simulation of multiple scattering*, Journal of Physics: Conference Series **219** (2010), no. 3 032045.

[18] R. M. Eisberg and R. Resnick, *Quantum physics of atoms, molecules, solids, nuclei, and particles*, Wiley, 1974.

[19] H. A. Bethe, *Molière's theory of multiple scattering*, Phys. Rev. **89** (1953) 1256.

[20] B. Gottschalk *et al.*, *Multiple coulomb scattering of 160 mev protons*, Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms **74** (1993), no. 4 467 .

[21] H. W. Lewis, *Multiple scattering in an infinite medium*, Phys. Rev. **78** (1950) 526.

[22] V. L. Highland, *Some practical remarks on multiple scattering*, Nuclear Instruments and Methods **129** (1975), no. 2 497 .

[23] LHCb collaboration, R. Aaij *et al.*, *Search for the $\Lambda_b^0 \to \Lambda\eta$ and $\Lambda_b \to \Lambda\eta'$ decays with the LHCb detector*, `arXiv:1505.03295`, submitted to JHEP.

[24] Belle, C. H. Wang *et al.*, *Measurement of charmless B Decays to eta K* and eta rho*, Phys. Rev. **D75** (2007) 092005, `arXiv:hep-ex/0701057`.

[25] BaBar, B. Aubert *et al.*, *Measurement of branching fractions and charge asymmetries in B decays to an eta meson and a K* meson*, Phys. Rev. Lett. **97** (2006) 201802, `arXiv:hep-ex/0608005`.

[26] Particle Data Group, K. A. Olive *et al.*, *Review of particle physics*, Chin. Phys. **C38** (2014) 090001.

[27] W. D. Hulsbergen, *Decay chain fitting with a Kalman filter*, Nucl. Instrum. Meth. **A552** (2005) 566, `arXiv:physics/0503191`.

[28] R. Aaij *et al.*, *The lhcb trigger and its performance in 2011*, Journal of Instrumentation **8** (2013), no. 04 P04022.

[29] S. Tolk, J. Albrecht, F. Dettori, and A. Pellegrino, *Data driven trigger efficiency determination at LHCb*, Tech. Rep. LHCb-PUB-2014-039. CERN-LHCb-PUB-2014-039, CERN, Geneva, May, 2014.

[30] H.-J. Yang, B. P. Roe, and J. Zhu, *Studies of boosted decision trees for MiniBooNE particle identification*, Nuclear Instruments and Methods in Physics Research A **555** (2005) 370, `arXiv:physics/0508045`.

[31] A. Hoecker *et al.*, *TMVA: Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040, `arXiv:physics/0703039`.

[32] J. E. Gaiser, *Charmonium Spectroscopy From Radiative Decays*, PhD thesis, Stanford University, 1982.

[33] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, eConf **C0303241** (2003) MOLT007, `arXiv:physics/0306116`, [,186(2003)].