Development of Multivariate Data Visualisation Software and Searches for Lepton Jets at CMS

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2013

Benjamin Charles Radburn-Smith

School of Physics and Astronomy

Contents

Total word count: 32000

List of Tables

List of Figures

Abbreviations

2D (nD): 2-dimensions (n-dimensions)
ak: Anti-Kt
ALICE: A Large Ion Collider Experiment
API: Application Programming Interface
ASIC: Application-Specific Integrated Circuits
ATLAS: A Toroidal LHC ApparatuS
BDT: Boosted Decision Tree
CDF: Collider Detector at Fermilab
CL: Confidence Level
CMB: Cosmic Microwave Background
CMS: Compact Muon Solenoid
CMSSW: CMS Software
CR: Charged Ratio
CSC: Cathode Strip Chambers
CSV: Comma separated value
DAQ: Data Acquisition System
DQM: Data Quality Monitoring
DT: Drift Tube
DY: Drell Yan
EB: ECAL Barrel
ECAL: Electromagnetic Calorimeter
EDM: Event Data Model
EE: ECAL Endcap
EJ: Electron Jets
EMF: Electromagnetic Fraction
FPGA: Field Programmable Gate Arrays

GPL: General Public License
GPU: Graphics Processing Unit
GSF: Gaussian Sum Filter
GSL: GNU Scientific Library
GSW: Glashow-Weinberg-Salam
GT: Grand tour
GUI: Graphical User Interface
GUT: Grand Unified Theory
HB: HCAL Barrel
HCAL: Hadron Calorimeter
HE: HCAL Endcap
HF: HCAL Forward
HLT: High Level Trigger
HO: HCAL Outer
IP: Interaction point
IR: Infrared
IR: Interaction regions
JCS: Jet Candidate Selection
JSON: JavaScript Object Notation
L1: Level 1 Trigger
LDA: Linear Discriminant Analysis
LEP: Large Electron Positron collider
LHC: Large Hadron Collider
LHCb: LHC beauty
LJ: Lepton Jets
LO: Leading Order
LS1: Long Shutdown 1
LSS: Long Straight Section
LUT: Look up tables
MC: Monte Carlo
MDI: Multiple Document Interface
MSSM: Minimal Supersymmetric Standard Model
MVA: Multivariate Analysis
NLO: Next to Leading Order
NNLO: Next to Next to Leading Order

OpenGL: Open Graphics Library
OS: Operating System
PAW: Physics Analysis Workstation
$PbWO_4$: Lead tungstate
PDF: Parton Distribution Function
PFJets: Particle flow jets
PSB: Proton Synchrotron Booster
PSB: Proton Synchrotron
PU: Pile up
PV: Primary vertices
PVT: Physics Validation Team
QCD: Quantum Chromodynamics
QED: Quantum Electrodynamics
QFT: Quantum Field Theories
RAM: Random Access Memory
RF: Radio Frequency
RPC: Resistive Plate Chambers
SDI: Single Document Interface
SISCone: Seedless Infrared Safe Cone
SM: Standard Model
SPS: Super Proton Synchrotron
SSS: Short Straight Section
SUSY: Supersymmetry
TEC: Tracker EndCaps
TIB: Tracker Inner Barrel
TID: Tracker Inner Disks
TMVA: Toolkit for MultiVariate Analysis
TOB: Tracker Outer Barrel
VBF: Vector boson fusion
VEV: Vacuum Expectation Value
WLCG: Worldwide LHC Computing Grid
WMAP: Wilkinson Microwave Anisotropy Probe

Abstract

Despite advances in multivariate visualisations and computer graphics, allowing for effective implementations, most particle physics analyses still rely on conventional data visualisations. The currently available software implementing these techniques has been found to be inadequate for use with the large volume of multivariate data produced from modern particle physics experiments. After a design and development period, a novel piece of software, DataViewer, was produced.

DataViewer was used as part of a physics analysis at the CMS experiment, searching for an associated Higgs decaying through a dark sector into collimated groups of electrons, called Electron Jets. Observation of such a signature could explain astrophysical anomalies found by numerous telescopes. The full 2011 dataset, equivalent to an integrated luminosity of 4.83 fb^{-1} at a centre of mass energy of $\sqrt{s} = 7$ TeV, recorded by the experiment was analysed.

DataViewer was found to be extremely powerful in rapidly identifying interesting attributes of the signature which could then be exploited in the analysis. Additionally it could be used for cross checking other complex techniques, including multivariate classifiers. No evidence was found for the production of a Higgs boson in association with a Z boson, where the Higgs subsequently decays to Electron Jets. Upper limits on the production of benchmark models were set at the 95% Confidence Level.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Benjamin Charles Radburn-Smith
School of Physics and Astronomy
University of Manchester
Oxford Road
Manchester
M13 9PL
June 2013

Copyright

The Author

The author was educated at Charters School, Sunningdale in Berkshire between 1997 and 2004, before reading physics at the University of Bristol, where he obtained a BSc (Hons). The author joined the Particle Physics Group at the University of Manchester in September 2007 where he obtained a MSc by Research in Particle Physics in 2008. The work presented in this thesis was undertaken while the author was a member of the University of Manchester and the Rutherford Appleton Laboratory, where he was based.

Acknowledgments

I would like to thank all who have supported me during this work. To my supervisors Steve Watts, Lakshmi Sastry and Claire Shepherd-Themistocleous for their invaluable help and guidance over the duration of the work. To the RAL PPD CMS group, especially Ian Tomalin, Manny Olaiya and Sam Harper for assistance with practical problems that arose during the physics analysis. To T. Gaggia for their assistance over the last year. To Sri Nagella for assistance with practical problems that arose during the software development. I would like to thank Yuri Gershtein, Eva Halkiadakis, Kristian Hahn, Stephen Mrenna and Chris Munson for their guidance on the H→EJ analysis. In a wider context I would also like to thank the CMS collaboration and in particular the CMS Exotica MultiJet group for assistance during the work.

Away from RAL and CMS I would like to thank my family for their support, encouragement and welcome distractions (usually in the form of Settlers/Puerto Rico among other games); my parents Dianne and Colin, my sister Sarah and my brothers Marcus and David. Marcus and Sarah's families have also expanded over the duration of this work to now include Katy (whom I also thank for proof reading) and James, Pete and Zac.

Chapter 1

Introduction

The use of multivariate classifiers as part of data analyses has gained momentum within the particle physics community over the past few decades, spurred on by increasingly powerful computers. The motivation for this increased use was to extract as much information as possible from data. Despite the use of these algorithms, most analyses still rely on conventional data visualisations. However, recent advances in multivariate data visualisations have shown much promise over the conventional visualisations. The currently available software implementing these techniques was found to be inadequate for use with the large volume of multivariate data produced from modern particle physics experiments. After a design and development period a novel piece of software, DataViewer, was produced. This software was then utilised within a physics analysis searching for lepton jets at the Compact Muon Solenoid (CMS) experiment.

This thesis is divided into two parts, with the first detailing the multivariate data visualisation work. Chapter 2 gives an introduction to the current visualisations and techniques available and describes in detail two interesting visualisations; parallel coordinates and the grand tour. In Chapter 3, multiple visualisation and data analysis software packages were tested and their strengths and weaknesses discussed. The requirements of a new software package were identified and the design of such a package outlined. Details of the new software package, DataViewer, are then described along with its current capabilities and potential developments.

The second part of this thesis details a search for lepton jets at the CMS experiment. Chapter 4 gives an introduction to the theoretical motivation behind the search. Chapter 5 describes the Large Hadron Collider (LHC) and CMS at CERN[1]. In Chapter 6 the search for associated Higgs decaying into electron jets is presented, in part using the DataViewer software developed in Part I. Chapter 7 draws together the conclusions from this work.

1.1 Conventions

The Cartesian coordinate system from the nominal interaction point in CMS is defined as follows: positive x points towards the centre of the LHC, positive y points vertically upwards and positive z points along the beam line towards the Jura mountains. The polar angle, θ, is defined as

$$\theta = \tan^{-1}\left(\frac{\sqrt{x^2+y^2}}{z}\right). \tag{1.1}$$

For ease of use, the η, ϕ, z coordinate system is utilised. This system is related to the Cartesian system with the following equations: Pseudorapidity, η is defined as

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right). \tag{1.2}$$

The azimuthal angle, ϕ is given by

$$\phi = \tan^{-1}\left(\frac{y}{x}\right). \tag{1.3}$$

z is equivalent to that in the Cartesian system.

A quantity used throughout this thesis is the distance in the pseudorapidity-azimuthal angle space, $\Delta R \equiv \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ [1]. A cone in $\eta - \phi$ space may be defined with radius, R. The physics motivation for pseudorapidity is given in reference [2]. Natural units were used throughout this thesis, where $\hbar = c = 1$.

[1] Conseil Européen pour la Recherche Nucléaire

Part I

Multivariate Data Visualisation

Chapter 2

Visualisation Theory

2.1 Introduction to Visualisation

Bruce McCormick in an article from 1987 stated that interactive computing would be an invaluable aid during the scientific discovery process, as well as a useful tool for gaining insight into scientific anomalies or computational errors [3]. The main aim of the visualisation techniques discussed here is to provide insights into the data, which not only includes highlighting errors or anomalies, but also in the more general sense of allowing the user to find patterns. A physicist could also use visualisation techniques to improve the purity and efficiency of a signal over a background.

Multivariate visualisations can also be referred to as multidimensional or high-dimensional visualisations. These visualisations relate to displaying multivariate data which are, among others, common within the particle physics community. An example of a simpler visualisation of multivariate data would be a 3D scatter plot; in which 3 variables are chosen from the n-dimensional data. The data is then plotted as points in three dimensions and projected onto a 2D display or printed onto a 2D medium. By using graphical techniques the number of dimensions displayed in this plot can be increased, for example, with use of colour, size and shape.

Multivariate data analysis algorithms such as neural networks have been used in particle physics since 1995. Likelihood discriminants, Boosted Decision Trees and other such analyses have been used in recent years by the Collider Detector at Fermilab (CDF) [4] and the DØ [5] experiments at the Tevatron. These techniques are regularly used by experiments at the LHC. The Toolkit for MultiVariate Analysis [6] (TMVA) group continues work in this area of research at CERN, creating a toolkit for CERN's ROOT [7] data analysis framework which implements a variety of multivariate classification algorithms.

Although multivariate data algorithms are being used with increasing frequency within particle physics analyses, the visualisation of multidimensional data is somewhat limited. In addition to this, despite advancements in computational power, highly interactive visualisations are not used to their maximum potential and instead it is usual practice for physicists to view their data using static displays. In most cases the physicist writes a piece of code using ROOT which runs over a dataset and then produces a histogram or 2D scatter plot. Following analysis of the plots, there are further rounds of code writing, running and viewing the output.

2.2 General Visualisation Techniques

A few important visualisation techniques include the use of interactive visualisations, linked plots and transparency.

Linked plots, otherwise known as linked brushing [8], allows the user to fully utilise multiple views which focus on different aspects of the dataset. When the user identifies and highlights certain data instances in one view, referred to as brushing the data, the same data instances in the other views are also automatically highlighted. This is usually performed with the use of colour. Linked brushing helps the user stratify the different plots and allows a deeper understanding of the relationship between variables.

Setting the transparency of the data instances, for example data points in a 2D scatter plot, allows the user to see data instances which would otherwise be hidden

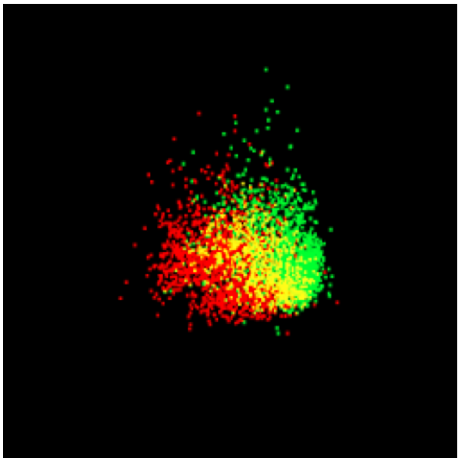

Figure 2.1: 2D scatter plot showing the use of the alpha channel. The red data points are the background, the green points are the signal and the yellow points show where the two data groups overlap.

from view. This is an important feature when used in conjunction with brushed data subgroups. For instance when studying the differences between a data signal in green and a data background in red, the user can see where the data overlaps as the colours merge to form yellow (see Figure 2.1). In computational graphics terms the alpha channel is responsible for the transparency of a display. This is described in Section 3.4.3. With transparency enabled, information from behind an object is used when rendering an object. If it was not enabled then only the last objects to be rendered from a collection of overlapping objects would be visible.

2.3 Visualisations Available

Many high-dimensional visualisations exist; a recent review of various techniques is given by G. Grinstein et al. [9]. Visualisations of interest include scatter plot matrices, heat maps, height maps, radviz and polyviz. Two popular and promising visualisation techniques, parallel coordinates and the grand tour, are described in more detail in Sections 2.4 and 2.5 respectively.

2.3.1 Scatter Plot Matrices

Scatter Plot Matrices are an array[1] of scatter plots where all the possible combinations of pairs of variables are plotted. The scatter plots are arranged so that all plots in a column have the same x-axis variable and all plots in a row have the same y-axis variable. Along the diagonal of the array, in which the same variable is plotted on both the x and y axis, a histogram or one dimensional trace display is usually used. An example of a scatter plot matrix is shown in Figure 2.2. Many variations of the scatter plot matrix exist, for example through the use of projections or by binning the plots.

While informative in showing the different combinations of scatter plots available, the scatter plot matrix has a few drawbacks. The plots on one side of the diagonal are the mirrored copies of the corresponding variables on the opposite side. This leads to a large area of redundant display space when viewing the matrix. If there are a large number of variables in the data the number of different possible scatter plots increases, leading to an unmanageable array size. The size of each scatter plot either becomes too small to view or the variables displayed are limited. There can also be problems when trying to simultaneously understand the connection between a large number of variables. This problem can be at least partially overcome through brushing the data.

2.3.2 Heat Maps and Height Maps

Heat maps can be understood as extensions to the normal 2D scatter plot where the graph area is divided into cells. The density of data points for each cell is represented through use of colour. For example a dense cell in the plot, where many data points lie, could be coloured as red while a less dense cell could be coloured blue. A heat map may be unrelated to a scatter plot, for example if attempting to understand the density of data on a hard disc, where the x and y positions of each cell has no meaning. The height map is a further extension of the heat map where the density of the cells are represented through the use

[1]Scatter plots can be arranged in a non-array format, for example in a circular or hexagonal pattern, but this is not as useful as the standard square array format.

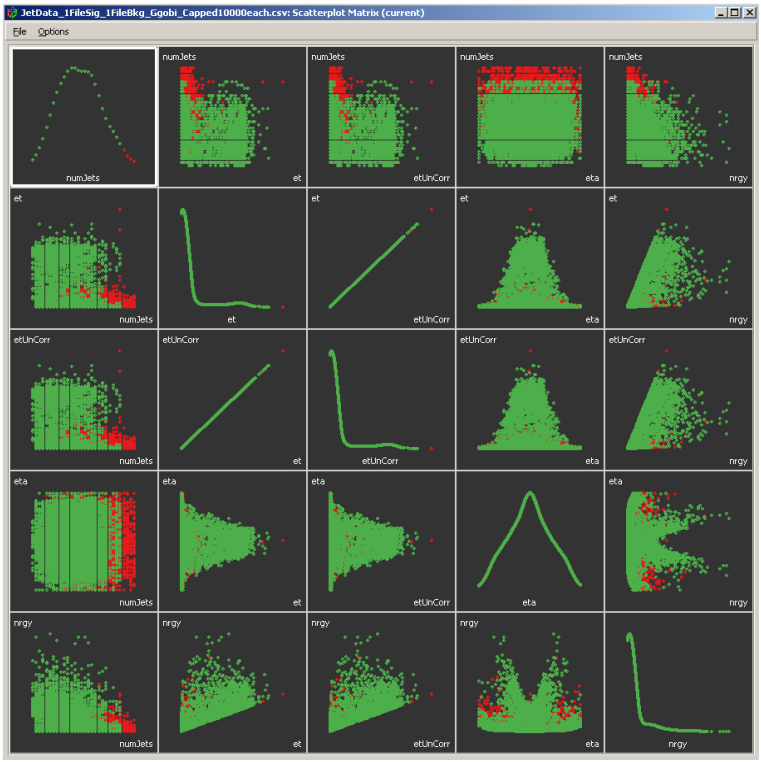

Figure 2.2: An example of a Scatter Plot Matrix (using GGobi [10]).

of a z axis, orthogonal to the x and y axis. Both colour and height can be used together to allow an easier understanding of the graph.

The traditional heat maps and height maps are fairly intuitive to understand, however they may suffer from imprecise representation of the data due to the lower resolution through the use of cells. By decreasing the cell size the plots can become almost continuous representations of the data. A similar effect of the traditional heat and height maps can be achieved without using cells by instead plotting the 2D scatter plot and adjusting the colours or height of the points if they lie on top of other points. Examples of these sorts of heat maps and height maps are shown in Figures 2.3 and 2.4 respectively. These plots are still limited to two dimensions and cannot represent higher dimensional data through single instances of the plot.

In heat maps the data cannot be brushed with colour to mark different subsets as the colour is already automatically calculated based on the density. Using transparency, the qualities of heat/height maps can be mimicked without using

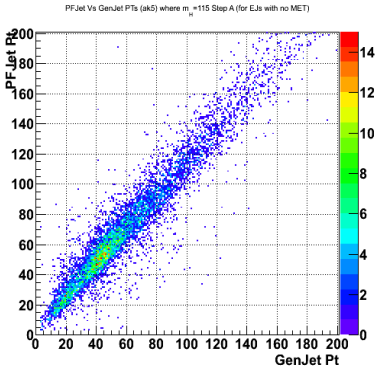

Figure 2.3: Example of a heat map.

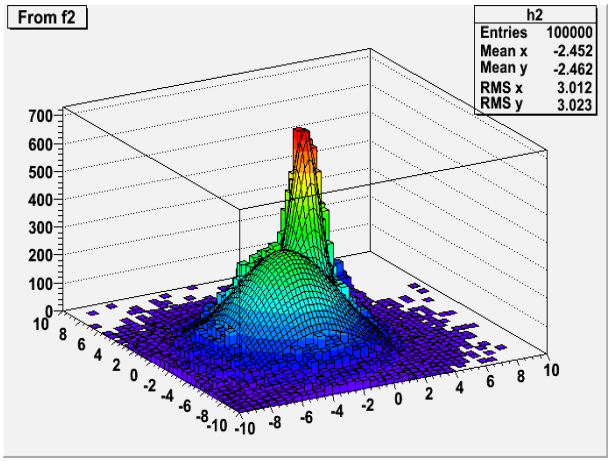

Figure 2.4: Example of a height map.

colours, therefore allowing the user to continue to brush their data and at the same time, understand the distribution of the data points. An example of the use of transparency is shown in Figure 2.1.

2.3.3 RadViz and PolyViz

RadViz, or Radial Coordinate Visualisation [11], is a relatively new style of visualisation. A circular plot is drawn with each of the variables of the data assigned a position around the circumference of the circle. These positions are referred to as the dimensional anchors. The data points are then plotted inside the circle at positions which relate the magnitude of each data point in a particular variable to the distance to the corresponding dimensional anchor. This is achieved by using

spring constants to find the correct position of a data point in relation to all the variables, where a spring is attached from the data point to each of the anchors. The data point is then drawn at the position where sum of the spring forces on that data point is equal to 0.

By changing the position of the anchors around the circumference of the plot, different patterns in the data emerge. This also leads to one of the main problems associated with this visualisation; certain features of the data only become apparent if the anchors are positioned in particular arrangements. Recent work has been conducted on vectorised RadViz [12] which attempts to automatically adjust the anchor positions in order to separate out clusters within the data. An example of RadViz is shown in Figure 2.5

PolyViz is an extension of RadViz where, instead of plotting on a circle, the graph is a polygon. In this plot the dimensional anchor is an edge of the polygon rather than just a point. Each data point is attached to value specific locations along the edge. The spring constants are then calculated from the data point to each edge and again plotted when the sum of the forces on the point is equal to 0 in a similar fashion to RadViz. PolyViz gives more information than RadViz as it displays the distributions of the data for each of the variables, as can be seen by the coloured lines drawn along the edges in Figure 2.6. These lines originate from the axis at the datum's value for that variable and points to the datum's final position in the middle of the graph.

With RadViz and PolyViz, projections of data are not required as all data and variables are seen in the graph. As such these visualisations give a global view of data, and the relations between data and patterns may be easily identified. However, the issue with the layout of the dimensional anchors, where the attributes of data only become apparent under particular configurations is still the biggest problem. Another problem may be that the plots become overcrowded when large datasets are visualised, making it difficult for the user to understand the features of data.

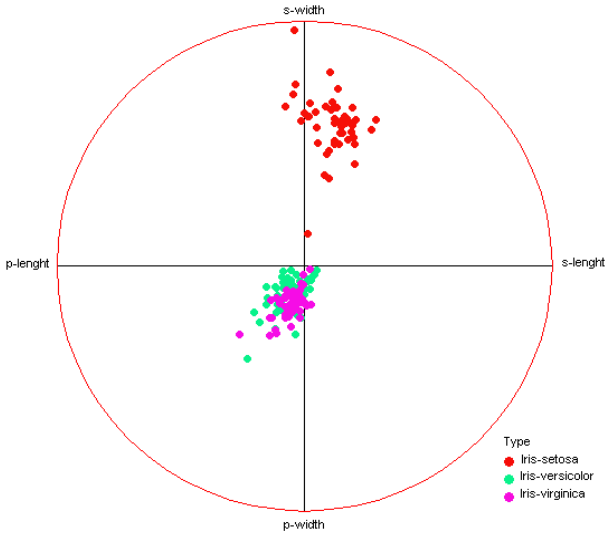

Figure 2.5: Example of the RadViz plot using the Iris dataset, which contains the geometric features of pollen grains [13].

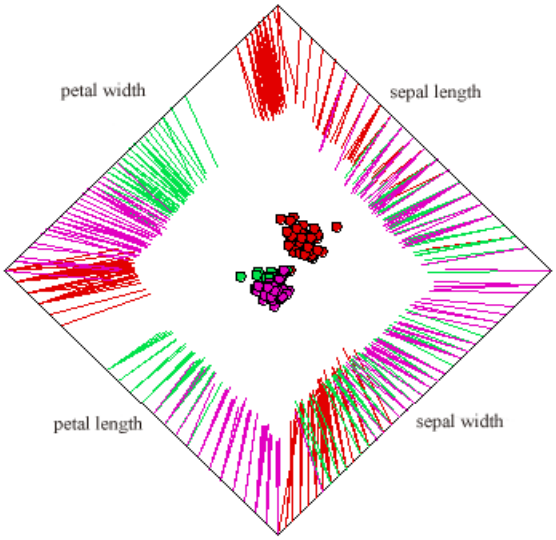

Figure 2.6: Example of the PolyViz plot using the Iris dataset, which contains the geometric features of pollen grains [13].

2.4 Parallel Coordinates

The origins and development of parallel coordinates comes from many sources. However there are three significant milestones that should be noted. The first recorded instance of the parallel coordinates concept can be found in a book written by Maurice d'Ocagne in 1885 [14]. In this work d'Ocagne describes the new coordinate system as a method to view high dimensional geometries. The line to point dualities, which are covered in Section 2.4.1 were described. Unfortunately this text was ahead of its time and the techniques described in it were not seized upon or developed any further. Computer graphics are needed to realise this visualisation effectively.

100 years later, parallel coordinates were reinvented by Alfred Inselberg [15] who was unaware of the earlier work by d'Ocagne. Inselberg, as did d'Ocagne, introduces the coordinate system as a method of working with higher dimensional geometries. In this work Inselberg describes how the plots can be used to obtain multivariate relations as well as other relations between 2D scatter plots and parallel coordinates. These relationships are detailed in Section 2.4.1.

Edward Wegman in 1990 [16] then developed Inselberg's ideas and further described how the coordinate system could be used as a high dimensional data analysis tool. Wegman describes how both one dimensional relations and n-dimensional relations can be viewed using the plots. He also describes the problems experienced in trying to understand the relations between axes which are not adjacent to each other and subsequently how permutations of the axes for pairwise comparisons are important.

The general concept of parallel coordinates is, as the title suggests, to align the axes parallel to one another instead of orthogonally as in a 2D or 3D Cartesian coordinate system. This allows the parallel coordinate system to handle more than the two or three variables/dimensions that a Cartesian system is limited to. It has been shown in other scientific research areas that the parallel coordinates visualisation technique can be a beneficial alternative to conventional methods when analysing data [17, 18] .

Figure 2.7 shows how a data point in a 2D scatter plot with a low value of A

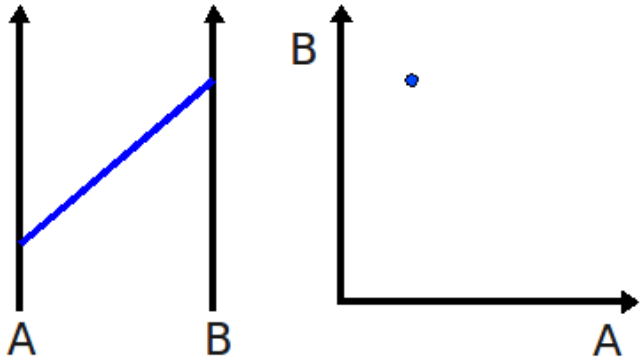

Figure 2.7: Image showing how a point in the 2D Cartesian coordinate system (RHS) relates to a line in the parallel coordinate system (LHS).

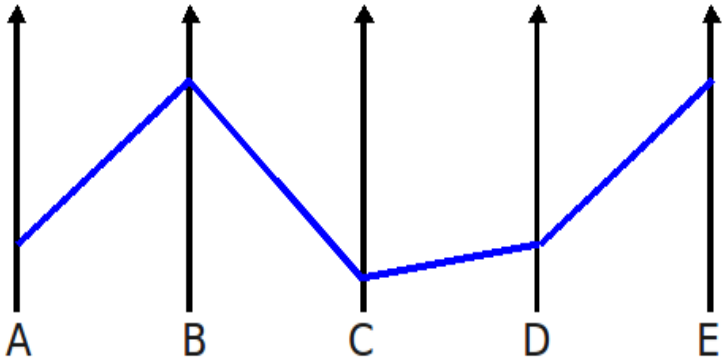

Figure 2.8: Image showing how a multivariate data instance is represented as a polyline in parallel coordinates.

and a high value of B translates in the parallel coordinate system to a line which intersects the A axis at a low value and the B axis at a high value. No information has been lost by using the new visualisation. As previously mentioned, parallel coordinates are not limited to two or three variables/dimensions. One data instance may have many attributes or variables attached to it[2]. These extra variables can be included into a parallel coordinate system by adding the relevant axes. For example the line in Figure 2.7 then becomes a polyline which intersects each axes at the related datum's value at that variable. Figure 2.8 shows a data instance represented in this parallel coordinate system as a blue polyline which has 5 variables, A to E.

More data instances can then be added to the plot, as shown in Figure 2.9. The

[2]The data is described as being multidimensional.

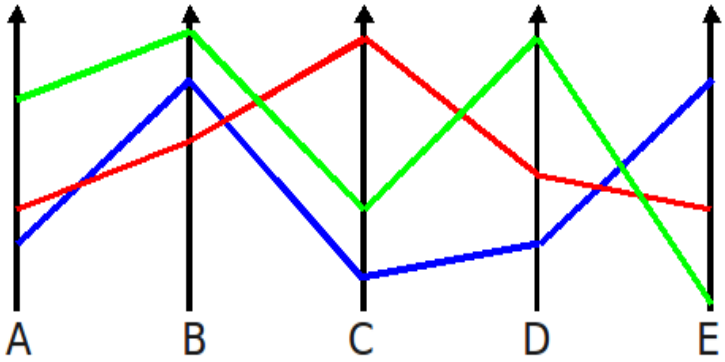

Figure 2.9: Image showing how multivariate data are represented as different polylines in parallel coordinates.

plots are normalised so that the maximum value for a variable is assigned to one end of the corresponding axis, and the minimum value is at the other end. So while each data line is independent of all the other data lines, by adding or removing data from the plots the absolute positions may change. Some implementations of this visualisation may stop this behaviour from occurring when removing data from the plot. All the data instances in the parallel coordinate plot must have the same number of variables attached to them, otherwise the plot would contain broken data lines making it extremely difficult to understand how the properties of each data instance and even the relations between variables.

A problem arises when comparing the values on one axis with those on a non-adjacent axis. This highlights the importance of the ordering of the axes and their permutations in the parallel coordinate system. These permutations are rendered unnecessary with the introduction of the grand tour [19].

2.4.1 Dualities to Scatter Plots

There are certain dualities between the parallel coordinate system and the 2D Cartesian coordinate system [20]. As has already been shown, a point in the 2D Cartesian coordinates becomes a line in parallel coordinates. It is also true that a series of points that lie on a line in the 2D Cartesian coordinate system translates to a series of lines which intersect at a point in the parallel coordinate system. This effect can be seen in Figure 2.10.

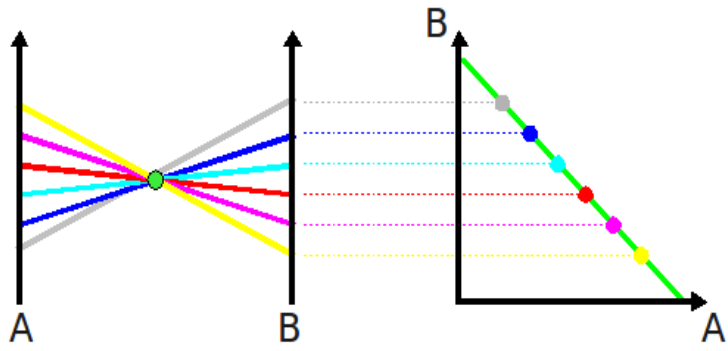

Figure 2.10: Image showing how a group of points which lie on a green line in the 2D Cartesian coordinate system translates to a series of lines which intersect at a green point in the parallel coordinate system. Where there is a negative correlation between the variables.

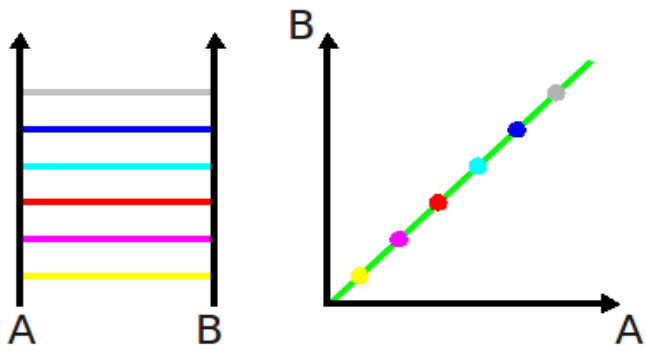

Figure 2.11: Image showing how a line of points in the 2D Cartesian coordinate system (RHS) translates to a set of parallel lines in the parallel coordinate system (LHS). Where there is a positive correlation between the variables.

If a group of data lines intersect at a point between axes in the parallel coordinates this relates to a negative correlation in a Cartesian system, as seen in Figure 2.10. On the other hand if a group of lines are parallel to one another between axes this would represent a positive correlation in the Cartesian system, as seen in Figure 2.11. If all the lines in parallel coordinates have the same value for one of the variables this would be represented by a vertical or horizontal line of points in the Cartesian system, as seen in Figure 2.12. This relationship between the 2D scatter plot and parallel coordinates can be described as a rotation to translation duality. As by rotating the line of points in the 2D scatter plot, the point of intersecting lines in the parallel coordinates plot is translated. While moving a point in the 2D scatter plot results in the corresponding line in parallel coordinates being

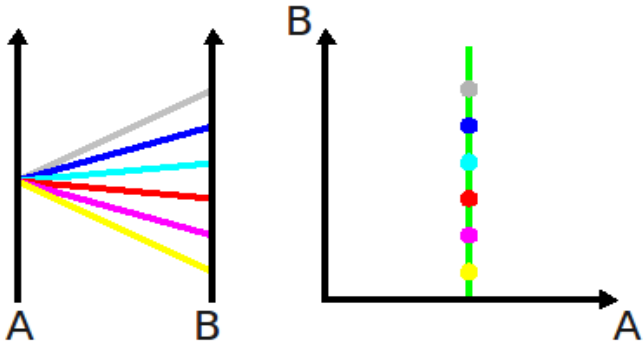

Figure 2.12: Image showing how a line of points in the 2D Cartesian coordinate system (RHS) translates to a point in the parallel coordinate system (LHS). Where there is one value for one of the variables.

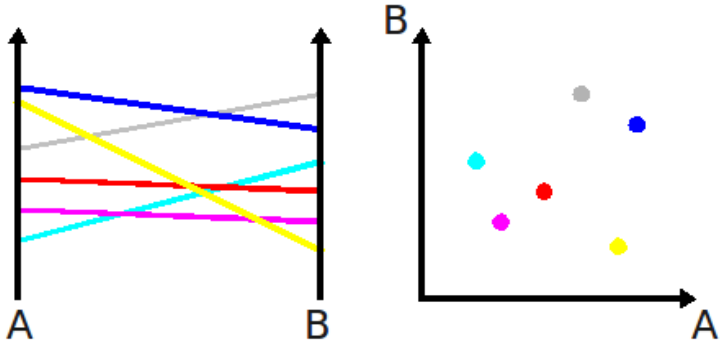

Figure 2.13: Image showing the relation of points in the 2D Cartesian coordinate system (LHS) to lines in the parallel coordinate system (RHS) where there is no correlation between the variables.

rotated.

If there is no correlation between the two variables, the lines in parallel coordinates exhibits no pattern just as points in a 2D scatter plot also shows no pattern. This effect can be seen in Figure 2.13.

An ellipse of points in the 2D scatter plot is represented as a hyperbolic curve of points in parallel coordinates. This relationship can be seen in Figure 2.14. Other dualities not discussed here include those between cusps and inflections as described in detail by Inselberg (1999) [20].

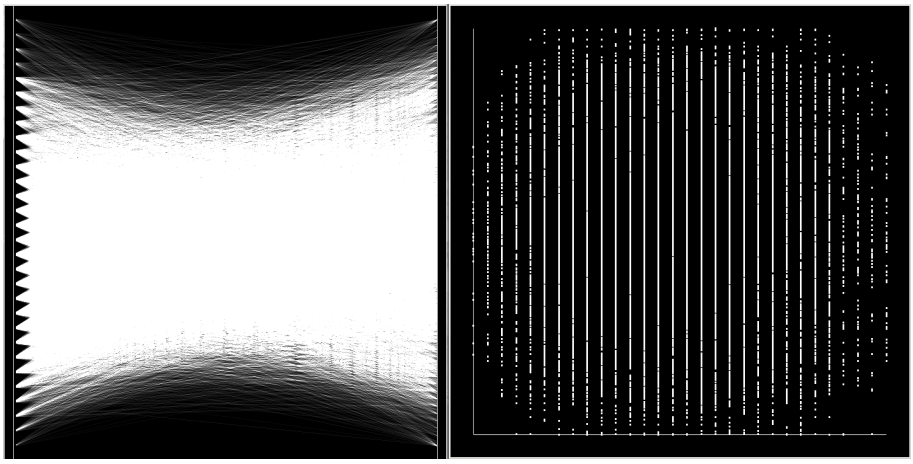

Figure 2.14: The duality between a hyperbolic curve of lines in parallel coordinates (LHS) and an ellipse of points in a 2D scatter plot (RHS).

2.4.2 Parallel Coordinate Density Plots

One of the difficulties that may be experienced with the standard parallel coordinate display is that of overplotting, where too many data lines occupy the same area in the plot and it becomes difficult for the user to understand what is happening. The parallel coordinate density plot is a powerful extension of the standard plot which overcomes this problem. Through the use of transparency, plots which have many data instances displayed can be better understood, as the plot can be adjusted so that only the areas where many data instances lie on top of each other are visible.

This technique can be seen between Figure 2.15 where low levels of transparency are used and Figure 2.16 where higher levels of transparency are employed. In the plot with increased transparency the areas of the plot where the majority of data lies are uncovered thereby giving the user a greater understanding of their data. Colour mixing helps to identify regions where two brushed subsets of the data overlap; green for signal, red for background and yellow where they overlap.

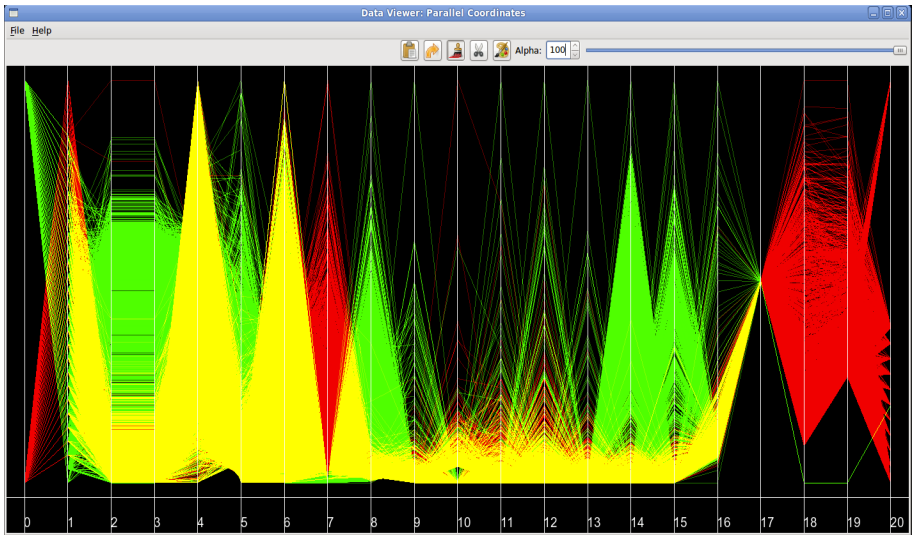

Figure 2.15: The parallel coordinate density plot with low transparency.

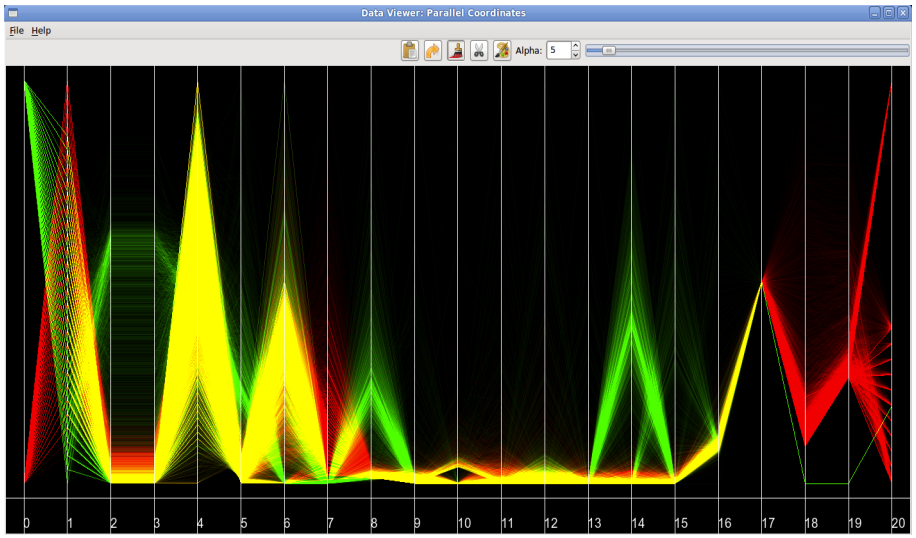

Figure 2.16: The parallel coordinate density plot with high transparency.

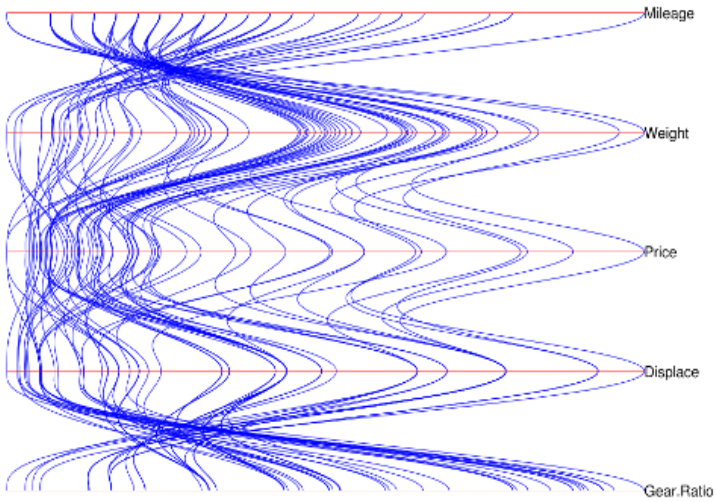

Figure 2.17: Image showing curved line interpolation in parallel coordinates.

2.4.3 Curved Line Interpolation

An interesting modification to the standard parallel coordinates plot is to use curved lines instead of the usual straight lines. This curved line interpolation can be seen in Figure 2.17. It has been shown that there can be difficulties in tracing where straight lines go if multiple instances meet at the same point on an axis [21].

2.5 The Grand Tour

Rotating views of multidimensional data are not new in the particle physics community. A computer based kinematic display of a multidimensional scatter plot was pioneered through the construction of the PRIM-9 system by Fisherkeller, Freidman and Tukey in 1974 [22] at SLAC[3]. This system allowed the user to picture, rotate, isolate and mask in up to nine dimensions. John Tukey supervised the construction of a 'Graphics Interpretation Facility' which hosted the system where bubble chamber data were used in the early analysis of the techniques. The work on the PRIM-9 led to the idea of projection pursuit which automatically

[3]Stanford Linear Accelerator Center

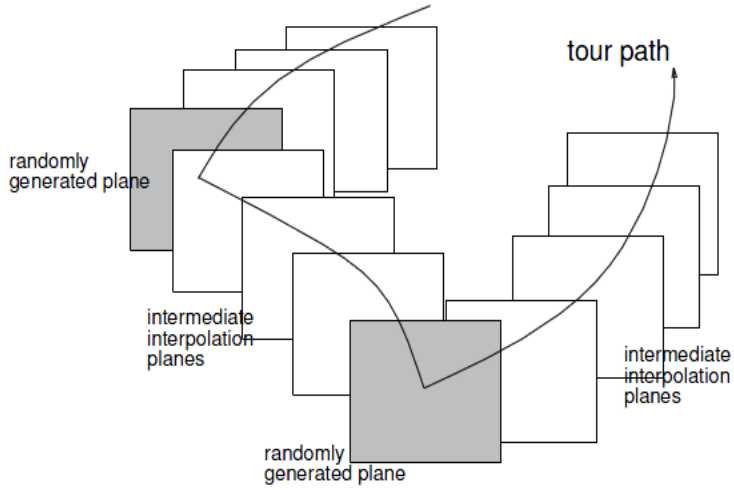

Figure 2.18: Graphical representation of the path of a tour [26].

finds interesting low-dimensional projections of multivariate data by optimising a projection index [23].

The grand tour was originated by Daniel Asimov in 1983 [24] while he worked at SLAC and then further developed by Asimov and Buja in 1986 [25]. The general strategy of the grand tour is to design 1-parameter families of 2-planes[4] in p-dimensional space, where the 1-parameter is thought of as time. The p-dimensional data is then projected onto these planes in rapid succession while increasing the time parameter thereby creating a movie of the p-dimensional plot. A representation of this process is shown in Figure 2.18.

The aim of the grand tour is to view the data from all possible perspectives while allowing the user to inspect a multitude of aspects of the data structure simultaneously and in relation to each other. This method also reduces the probability of overlooking structures within the data. Unlike the projection pursuit technique where the output is a projection of some optimisation of an index, the output of the grand tour is the movie itself.

A requirement of this visualisation is to create a smooth continuous sequence of projections which allows the user to easily track the data points and their structures. The mathematics therefore requires a continuous, space-filling path

[4]A 2-plane is the plane contained by a pair of orthogonal vectors known as 2-vectors.

through the set of 2-planes in the p-dimensional space in order to satisfy this requirement [19]. Various space-filling algorithms exist such as the Asimov-Buja winding algorithm, random curve algorithm and fractal algorithm [19, 25].

Wegman suggested in 1991 [27] replacing the manifold of 2-planes with a manifold of k-planes, where $k < p$ of a p-dimensional plot. The data could then be projected onto the k-plane and visualised using either parallel coordinates or a scatter plot matrix. This is referred to as Wegman's k-dimensional grand tour.

It is generally advisable that the user should be cautious when inferring relations in the data from projections alone, as our normal 2D and 3D geometric intuition breaks down for higher dimensional geometry.

Chapter 3

Visualisation Software

This chapter describes a selection of the visualisation software packages that were available at the start of the project. Particular focus is given to their strengths and weaknesses, found through testing the different software. There follows a description of the requirements for a new visualisation program based on this testing. Finally the DataViewer program, a novel piece of visualisation software developed for use in particle physics analyses, is described in detail along with an assessment of the current state of the software and an itinerary of possible improvements.

3.1 Existing Visualisation Software

There exists a large repository of both free and proprietary visualisation software which implements the techniques described in Chapter 2, each with their own various strengths and flaws. A selection of programs were tested to find useful features which could then be incorporated into the design of a new piece of visualisation software tailored for use in Particle Physics. The software of interest were GGobi [28], CrystalVision [29], ROOT [7], RapidMiner [30], Weka [31], Orange [32], Matlab [33] with edatoolbox, XmdvTool [34] and IRIS Explorer [35]. Testing focused on CrystalVision, GGobi and ROOT as these programs exhibited the most useful features.

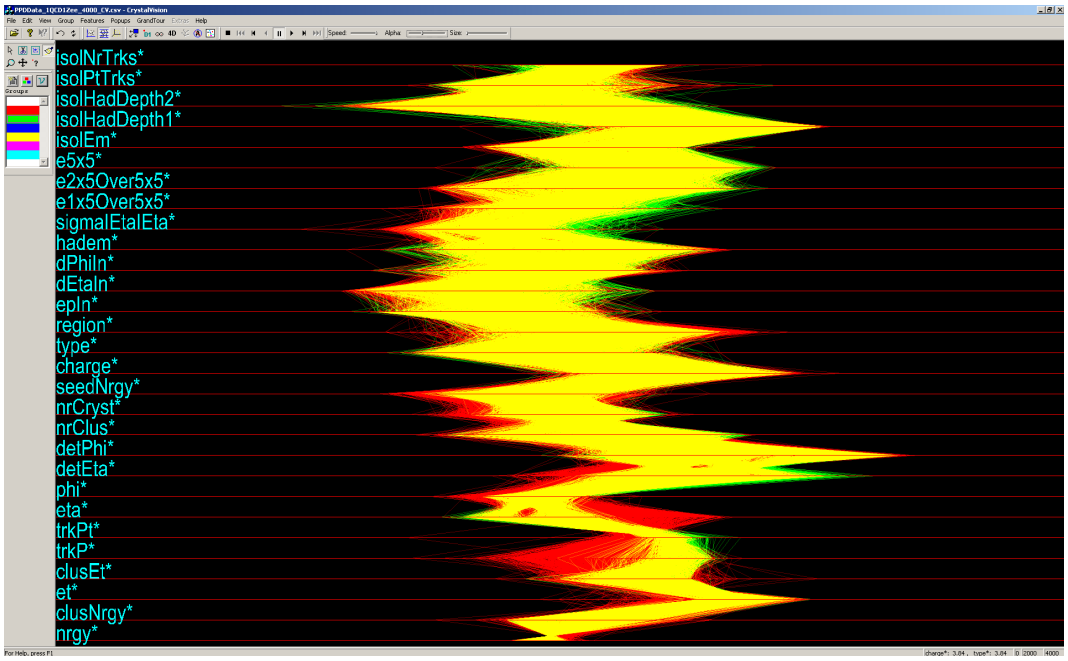

Figure 3.1: A screen shot of CrystalVision showing the parallel coordinate display (during a grand tour).

3.1.1 CrystalVision

CrystalVision was created by E. Wegman and Q. Luo circa 2000. It was developed as a commercial evolution of the freely available ExplorN by D. Carr, E. Wegman and Q. Luo. By 2003 it appeared that work stopped on CrystalVision and Wegman put a full copy of the program on his George Mason University homepage. This version has since been taken down and no version of the software is currently available.

CrystalVision allows the data to be displayed in a parallel coordinate view, as a p-dimensional scatter plot and as a scatter plot matrix. Figure 3.1 shows a screen shot of the parallel coordinate view in CrystalVision with a data set loaded. CrystalVision implements Wegman's k-dimensional grand tour of the parallel coordinates view as described in Section 2.5. It also utilises the alpha channel to set the transparency of the data instances as shown in Figure 3.1 where the red and green instances overlap to produce a yellow view. The level of transparency as well as the size of the data points and the speed of the grand tour can all be adjusted.

The program allows the user to remove, or prune, data instances from the plots with the mouse, either by cutting away the undesirable points or by cropping the points the user wants to retain. The number of data points on the plot are printed on the display which then allows the user to manually calculate the efficiencies when editing the data. CrystalVision implements the linked brushing technique, as described in Section 2.2, in a similar fashion as GGobi. The scatter plot/scatter plot matrix cannot be viewed at the same time as the parallel coordinate display and the user instead has to switch between the views.

CrystalVision allows the user to change the background to either black or white. When the background is white it allows the user to set the transparency to a low level and still allow individual instances to be viewed that would otherwise be too faint to observe with the black background. The program has a zoom function allowing the user to expand an area of the plot. CrystalVision does not allow the user to save a modified version of the data set. It does provide information of the grand tour when it is paused in the form of a rotation matrix.

The main strengths of CrystalVision are the implementations of both the grand tour and parallel coordinates within an intuitive Graphical User Interface (GUI), made possible through the use of tools and the toolbar. The main problems with CrystalVision are that the software is no longer available as well as some performance shortcomings when dealing with larger datasets. As the software was closed source, it could not be natively integrated with any other software or non comma separated value (csv) file formats. Also CrystalVision was only available as an executable for the Windows Operating System (OS), while the majority of the particle physics community uses UNIX based OS.

3.1.2 GGobi

GGobi was developed by D. Swayne, A. Buja, D. Temple Lang and D. Cook in 2001 [10]. It is a direct descendant of the XGobi system which dates back to the early 1990's [36]. GGobi is a product of AT&T labs and although they hold the copyright to the product, it is under a Common Public License. GGobi becomes increasingly slow and unresponsive when large amounts of data are loaded into

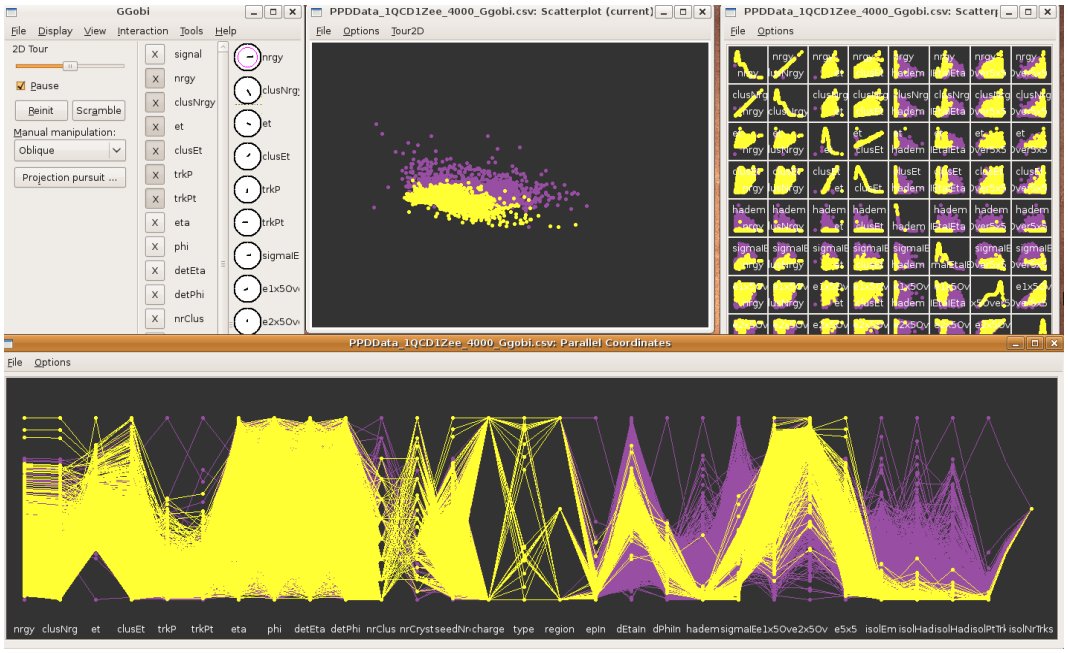

Figure 3.2: A screen shot of GGobi showing a tour, scatter plot matrix and parallel coordinate display of a particle physics dataset.

the program.

GGobi implements a Single Document Interface (SDI) which leads to multiple windows on the screen. Figure 3.2 shows a screen shot of GGobi with a particle physics data set loaded. GGobi allows the data to be viewed in 2D scatter plots, scatter plot matrices, parallel coordinates, as a time series and as a bar chart. It also implements the grand tour of the p-dimensional scatter plot.

In addition to allowing a grand tour of the data set GGobi permits the user to run a projection pursuit on the data. The projection pursuit can find an interesting projection of the data through several different options. Three options to note are the Holes [37], Central Mass [37] and Linear Discriminant Analysis (LDA) [38]. After brushing the signal and background in the data, the LDA tries to separate out the two groups of data as best it can. The Holes option tries to find the projection in which the largest gap appears inside the scatter plot. The Central Mass option tries to find the projection where the data points are most dense.

As can be seen in the top left window in Figure 3.2, GGobi shows the various orientations of the axes (the round graphics) of the p-dimensional scatter plot.

The program allows the user to click on one of the variables to manipulate that axis directly. There is also a toggle button in the grand tour window which shows how the axes are aligned for each projection. The variables can be selected and deselected from the grand tour. The speed of the tour, as defined by the number of frames processed per second, can also be varied.

The parallel coordinate display in GGobi does not implement Wegman's k-dimensional grand tour. However, GGobi does allow the user to swap axes thereby allowing correlations between variables to materialise which would otherwise go unnoticed. GGobi does not use the alpha channel and so the transparency of the data instances cannot be changed. This means that instances that are covered by other instances are hidden. Re-brushing the groups in the data set may uncover the hidden data sets, however this is an inefficient method and does not allow the user to view the characteristics of the groups simultaneously.

GGobi uses the linked brushing technique to allow the user to cut away or highlight data instances and view the consequences in the other views. However it does not allow the user to delete data instances from the plots. This means that the user cannot measure whether the efficiency of a signal over the background is increased or decreased by making a cut in the p-dimensional space.

GGobi's main strengths are the intuitive use of a wide variety of plots allowing for simultaneous inspection of the data from different perspectives. The implementations of the scatter plot grand tours along with projection pursuit allow for useful data exploration. The main weaknesses of GGobi include the slow down in performance with large amounts of data, not implementing transparency in the plots and not allowing the user to remove data instances from the plots.

3.1.3 ROOT

ROOT is an open source object oriented framework for large scale data analysis. It is a C++ replacement of the popular Physics Analysis Workstation (PAW) program developed at CERN, providing a platform independent data analysis environment. Almost all particle physics data is analysed using ROOT due to its efficient storage, query and access to data, statistical analysis algorithms, and

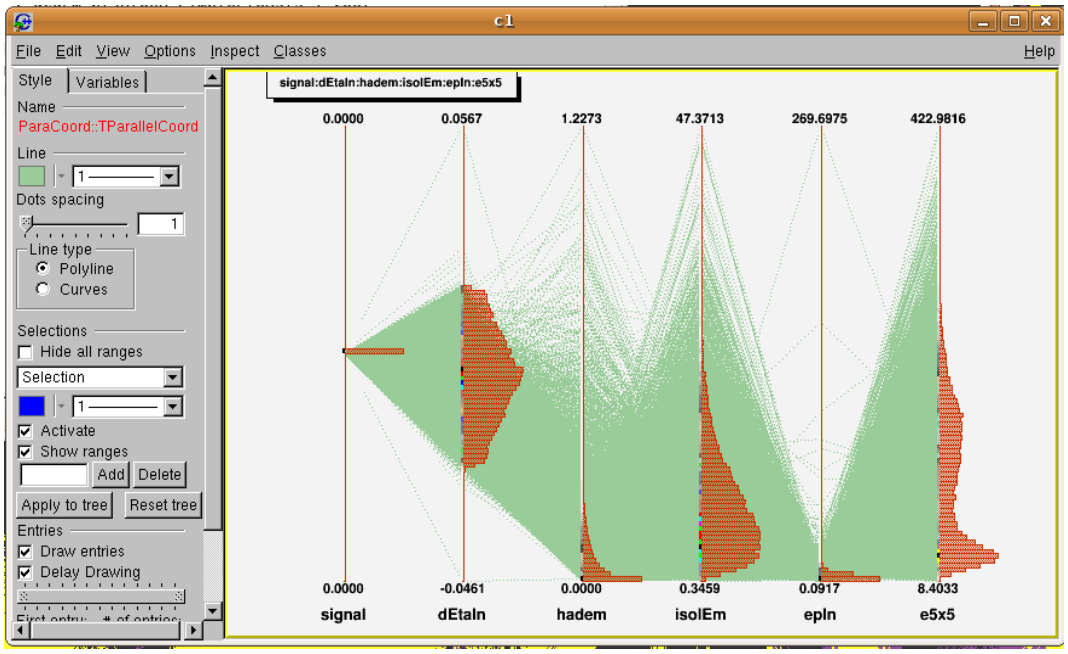

Figure 3.3: A screen shot of ROOT showing the parallel coordinate display of a dataset.

scientific visualisations.

The data format used in ROOT was designed to handle the petabytes of data produced by large experiments. This file structure allows the user to load only the data which is of interest into memory, thus providing high performance input and output. This structure was designed based on classical data analysis and visualisation techniques; however, the design of ROOT allows for new features to be incorporated into the framework.

Olivier Couet implemented some new visualisation techniques into ROOT in 2008 [39]. These visualisations include the parallel coordinate display as well as other visualisations not discussed here such as spider plots. Figure 3.3 shows the rudimentary parallel coordinate implementation within ROOT. The parallel coordinate view is initiated from a ROOT command line syntax and runs on OpenGL libraries.

The user cannot brush data instances as is the case with GGobi and CrystalVision, nor can data instances be removed from the plot. Rather than using the alpha channel as implemented in CrystalVision, ROOT allows the user to change

the style of the line to dashes of varying width. This technique does not appear to be as effective as the use of the alpha channel. The grand tour visualisation is not implemented within ROOT.

ROOT does however allow the user to change from straight lines in the parallel coordinate view to curved lines. As described in Section 2.4 this is an interesting feature that attempts to allow the user to identify data instances attributes in crowded areas of the plot. A useful feature of the parallel coordinate implementation within ROOT is the ability to plot 1-Dimensional histograms on top of the axes.

The main strengths of ROOT are the data structure format allowing for fast read and write performance along with the statistical analysis algorithms bundled within the framework. The implementation of traditional visualisations is adequate for most analysers, however more complicated visualisations such as the parallel coordinate display are not as powerful as those found in other packages such as CrystalVision and GGobi.

3.2 Motivation For New Software Development

After testing the available software described in the previous section, several useful features and problems were identified. Most particle physics analyses do not appear to make good use of the variety of data visualisations available. This is most likely due to the restrictive nature of the analysis method within the ROOT environment, where the user writes lines of code which runs over a dataset and then, after some time, produces a display representing a particular attribute. The user studies this visualisation and then modifies their code to focus on a specific aspect of the data.

This static method of producing informative scientific visualisations slows down the entire data analysis. A new software package was envisioned that would allow the user to quickly inspect their multivariate data using some of the more interesting visualisations available, such as parallel coordinates and the grand tour.

3.3 Requirements and Design

Requirements for this new software package were identified. These requirements influenced the design and implementation of the software. As well as being robust and reliable, the software should meet the conditions detailed in this section.

The program should be able to load data straight from ROOT classes which would be beneficial to particle physicists who store their data in this format. When the data is viewed showing all the attributes that belong to that data, it should also allow the user to see those for a selected subset of the data. Performance information for the data should be presented to the user in an easily understandable format. The user should be able to edit the loaded data through brushing and pruning. The software should have a mechanism that saves the sequence of actions performed on a data set.

The primary requirement of the program is to have a parallel coordinates view which utilises as much of the screen estate as possible by minimising the use and size of buttons. Once the data has been loaded, the user should have the ability to change the order of the axes within the plot. An automated function to find the best order of the axes should be included where the start order can be defined by the file which holds the data. It may also be of use to split the parallel coordinates plots, for example into two plots, one for signal and one for background. A useful tool could be included to switch between different interpolations of parallel coordinates via the use of straight lines or curved lines.

In addition to the parallel coordinates view the user should have access to a scatter plot view. It would be of value to have a scatter plot matrix view as well. The user should be able to change the size of the data points in the scatter plot using a tool with an appropriate scale.

The program needs the capability to set the transparency of the data using the alpha channel, using proper sliders that have a scale, thereby implementing the parallel coordinate density plot as described in Section 2.4.2. This transparency feature should also be implemented for the scatter plots.

The grand tour (GT), as defined in Section 2.5, of both scatter plots and parallel

coordinates needs to be included in the program. Interaction of the GT through the use of the cursor and/or keyboard should be allowed. A toggle to show the axes of the GT using scatter plots, as implemented in GGobi, should be included. A toggle to show the values of the matrix transformation should also be implemented. The ability to prune data during the GT is essential to allow the user to find interesting hyperplanes to make cuts across. A slider to change the speed of the GT (with proper scale) needs to be included. An implementation of projection pursuit of the tours could also be written into the software.

In terms of functionality, it was essential that the program have the ability to see all views at the same time and have these views linked such that changing a value in one view changes the related value in all views. There should be a function to brush data instances with specific colours (from a palette) and another to prune data instances thereby removing them from the plots. The user should be allowed to invert selections of data instances to define those they wish to keep in the plots. It would also be useful to include a performance indicator of the pruning function that shows efficiencies/purities/number of events and/or possibly confusion matrices to show the performance of classification [40].

Some informative tools should be integrated with the program. For example, a tool to indicate which variables show the most discrimination of signal from background. A mechanism should be incorporated that suggests which data mining algorithms would work well on the data set; which could then have the ability to run the algorithms on the data from within the visualisation. A function that creates new variables by multiplying other variables could also be of use.

It may be visually useful to allow different colour backgrounds for the various plots to ease the spotting of patterns in the data. A zoom function in the different views could also help the user get more detailed information from the plots, especially in areas of densely plotted points/lines.

A mechanism that logs the sequence of actions, subsequent results and saves them to disk would allow the user to save their work effectively. To increase the usability of the program it would be useful to include a history function that allows the user to go back through the actions.

3.4 DataViewer

After examining and testing the various software solutions addressing multivariate visualisations described in Section 3.1 a new program, DataViewer, was designed and implemented. The main purpose of this software was to give an additional tool for data analysis. It is hoped that the software package can be released under a GNU General Public License (GPL), or similar, to allow the community of interested users to patch and make useful contributions to the software. A screen shot of the program in use is shown in Figure 3.4.

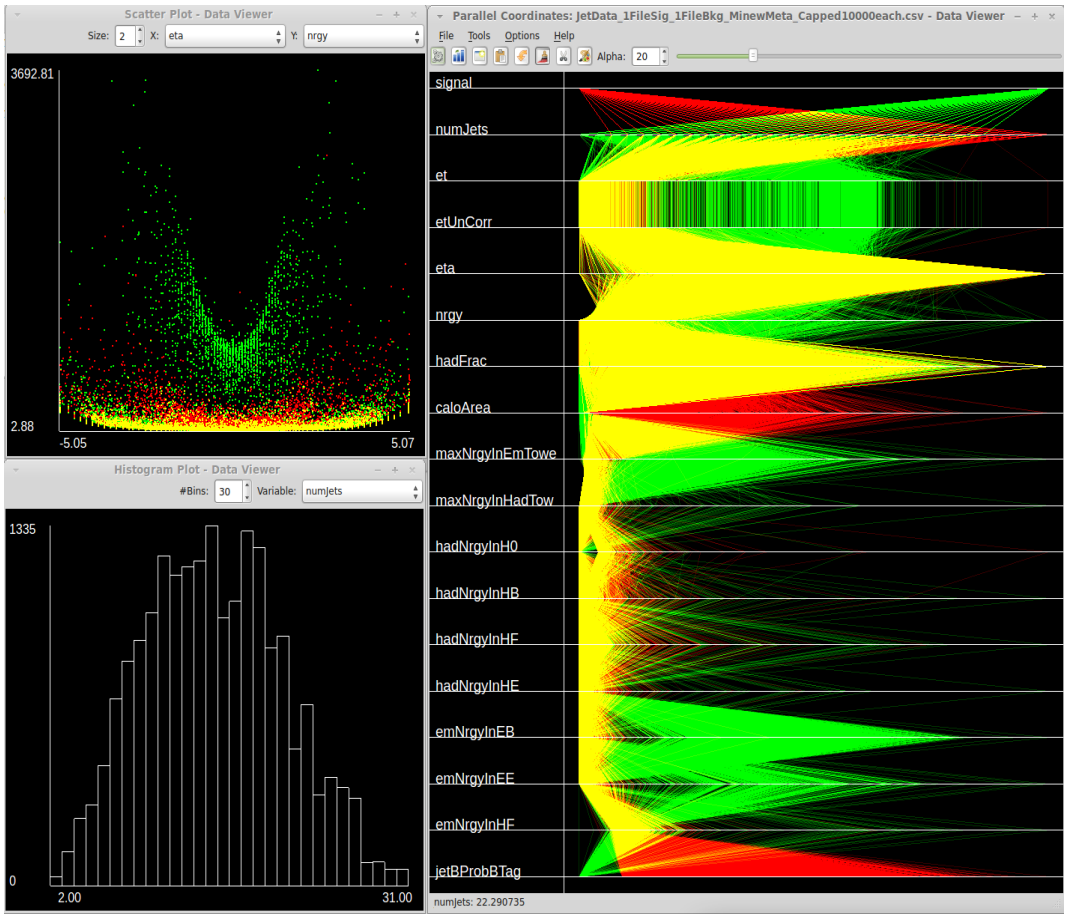

Figure 3.4: Screen shot of various plots in DataViewer once data has been loaded. Top left shows an example of the scatter plot, bottom left shows a histogram, and the main window on the right shows the parallel coordinates view.

3.4.1 Languages and Libraries Utilised

The majority of code in DataViewer was written in the C++ programming language. C++ is an multi-paradigm[1] general purpose language based on the C language. C++ was utilised due to its fast performance and modularity making it easier to enhance the program with added functionality throughout the implementation, testing period and into the future. This style also makes it easier for more than one person to improve upon the software. Additionally, the use of C++ allows for the possibility to interact directly with the ROOT data analysis framework which is used extensively within particle physics.

For the graphics intensive part of the program the Open Graphics Library (OpenGL) was used. OpenGL is a cross-language, cross-platform application programming interface (API) which is very efficient at displaying large amounts of data quickly through its use of the Graphics Processing Unit (GPU) on a computer.

The Qt framework [41] was chosen for developing the front end GUI of the program due to its ease of use and cross-platform accessibility. Therefore a single source tree of code could be created which could run on multiple Operating Systems (OS). Qt is an open source project which provides a neat and stylish windowing system that can embed OpenGL displays. The framework uses macro interpretation to enable a signal and slot messaging system between windows allowing for a powerful method of messaging in real-time without having to break the C++ paradigm.

Algorithms from the GNU Scientific Library (GSL) were deployed within the program in order to allow quick mathematical computations. In particular the statistics part of the library was of interest in testing the attributes of variables within a dataset as detailed in the 'Ordering of the axes' part of Section 3.4.5.

[1] Although multi-paradigm, C++ was used as an object orientated language throughout this project.

3.4.2 Overview of the General Layout

The Single Document Interface (SDI) system was chosen for this program in which multiple windows are managed as individual entities. This was preferred to the Multiple Document Interface (MDI) windowing system where all sub-windows are contained within one large window. The SDI windowing system has numerous advantages over the MDI system such as providing more information about which views, or windows, are open and allowing the OS to handle those windows. However, from testing other visualisation programs, the main advantage with SDI was its flexibility. On multi-monitor workstations this windowing system was designed to give more power to the user, allowing greater manipulation of the display. Drawbacks of the SDI system include a greater chance of losing a window, when it is hidden from view, usually via a mis-click. Also, adjusting the size or position of one window does not automatically adjust any nearby windows; this means greater effort is required by the user to achieve an optimal set up.

The program can be started by either double clicking the binary file, or preferably, via the terminal which provides more information during the programs use. The main window allows data to be loaded into the software, create scatter plots and histograms, perform various functions and select options. These functions and options are detailed in Sections 3.4.5, 3.4.6 and 3.4.7 respectively. Some functions relating to scatter plots and histograms are stored within the main window in order to leave screen real-estate for the actual visualisations. The main window holds the single parallel coordinate view which is generated using OpenGL.

Data can be loaded into the program by selecting Open under the File menu in the main window. This launches the Open Data File dialogue, which is a straightforward implementation of Qt's QFileDialog. Once a file is selected, the program runs a function which first scans the file to count the number of rows relating to the number of data instances, and secondly the number of columns which gives the number of variables in the dataset. It then allocates a suitable amount of memory and stores each value for each data instance in that memory as an implementation of the C++ standard library vector.

After the data has been read from file into memory, a different function transforms the values to a range between 0 and 1 for each variable. This allows the

plotting functions to place the appropriate graphics in the correct locations on the plots. This transformation process is called normalisation. The resulting normalised data is stored in a separate vector in order to increase performance when subsequently reading the data. The parallel coordinate plotting function then draws the data to the screen. During this loading process helpful information is printed to the screen to show what the program is currently doing, along with a timer to indicate how long that particular part of the sequence has taken. Information displayed includes the name of the file loaded, the number of data instances and number of variables.

A key feature of DataViewer is the provision of multiple scatter plots and histograms to display the data loaded. All of these windows are linked, so by pruning/colouring the data in one window, all other windows will be automatically updated, thereby making use of the technique described in Section 2.2.

Interaction with the software primarily occurs through use of the mouse, although there are a few commands which can be run from a combination of mouse and keyboard strokes. The rubber band method was implemented in order to select data in the program in which the cursor is clicked and dragged to make a box.

A status bar was included at the bottom of the main window in order to give useful real time information and helpful tips during the program operation, for example, the value for a variable at the position of the cursor.

3.4.3 Plots

The main feature of DataViewer are the multiple plots it can provide to display the data loaded into the program. The technique of linked plots, as described in section 2.2 is utilised in the program so that brushing data points/lines with colour in one plot will update the other plots showing the same data points/lines to reflect the change. The three types of plots which can be displayed through the program are described in more detail in the following sections.

All the plots are rendered using the OpenGL API in order to utilise the computational power of the GPU by using many parallel processes at the same time.

Graphical primitives are used to draw the data to the display the first time a plot is displayed. The display is first rendered to a buffer which is then transferred to the screen, a technique called double buffering. This buffer of rendered graphical information is then saved and used for displaying the data until a modified version is needed, either from resizing the window or brushing/pruning the data. The advantage of drawing from a saved buffer rather than drawing the graphical primitives repeatedly is to allow the creation of rubber bands for selection without having to draw the whole display from scratch which would be slow.

The default background colour for all of the plots is black which enhances the visual information that can be garnered from displays. This is especially true when using transparency. If data has been brushed with multiple colours, the rendering uses a blending technique in order to calculate the addition of the composites of each colour for each pixel and to show the correct resulting colour. For example, when a set of data is coloured green and lies on top of another dataset coloured red, the resulting overlapping data is rendered in yellow.

As part of the rendering of colours on the plots, the level of transparency for each colour is also recorded and used. This is achieved through use of the alpha channel, which tells the rendering algorithm how much of the existing pixel information should be used when colouring a pixel. Each pixel is drawn to the plot using a RGBA value, which relates to the amounts of red, green, blue and alpha to be used. These values take the range between 0.0 and 1.0. The level of transparency is changed by using the slider in the main window. This controls transparency for both the parallel coordinates and scatter plots simultaneously as described in Section 3.4.5.

The windows which contain the various plots are implementations of Qt's QMainWindow. Whenever a new window is created it is automatically connected to all the other plots already open and, if no other windows have been initialised, only the parallel coordinates plot. Modifying the data in one plot will update all other plots showing the same data. This is done through Qt's signals and slots mechanism which unlike conventional C++, allows connections to be made while the program is running. This also allows for connections inside a window between widgets such as buttons, sliders and spin boxes and their associated functions. All windows can be resized.

Parallel Coordinates Plot

The parallel coordinates plot is the centre stage of the program as it is incorporated into the main window and cannot be switched off. Only one instance of this plot is provided by the software, unlike the other plots where multiple instances can be initiated.

The parallel coordinate plot is split into two areas by a divider orthogonal to the axes; on one area the names of the variables attached to each of the axes is displayed and on the other area the data instances are drawn. Based on testing with different variables, this divider is currently set at a static distance from the edge of the plot. In future versions of the software this could be assigned by finding the longest variable name and, if below a maximal limit, setting the divider after the last character in that name.

The axes can be moved round by swapping locations between pairs of axes. The changing of axis location is important in order to show patterns between different sets of variables and therefore provide more useful information from the plot as described in Section 2.4. This is achieved through clicking and dragging the axes.

It is desirable to allow the removal of variables from the plot, for example when a variable does not show any useful information or the user wishes to be blind to a variable. This allows other variables, which are of greater interest, to take up more of the plot's real estate. An axis can be removed directly from the plot. The variables and their associated axes can also be removed and reintroduced using the 'Select Visible Variables' option described in section 3.4.7.

The area in which the data instances are drawn is considerably larger than the area designated to displaying the labels of the axes. When the parallel coordinate plot is resized, the data area adjusts in a suitable manner but the label area remains static. This stops the labels from either becoming unreadable or taking too much space away from the area where the data are rendered. There are small gaps both between the divider and the point where the data lines first cross the axis, and the point where the last data lines are drawn and the edge of the window. This maintains the display in a neat fashion and minimises the risk of mis-clicking away from the plot, for example when selecting data at the edge of

the plot.

Data lines in the plot adjust to encompass all available space along the axis. The maximum and minimum values for each of the variables are calculated, and then set as the upper and lower limits along the axis. When data are removed from the plot, new maximum and minimum values are recalculated and set to the ends of the axes. The data lines are drawn underneath the lines representing the axes, to allow the user to successfully locate the axis.

Scatter Plot

The scatter plot has two areas; one where the data points are plotted and another where the axes labels are shown. Both of these areas have a surrounding gap to the edge of the window for the same reasons outlined for the parallel coordinates plot. As the size of the characters in the label text does not change, these gaps are a fixed size so that when the plot is resized, only the area where the data points are rendered changes.

There is one spin box and two drop down menus at the top of the plot window which are implementations of Qt's QSpinBox and QComboBox respectively. The spin box changes the size of the data points which are drawn on the plot allowing for a choice of size within the possible range of 1 to 10 in arbitrary units. After testing various options, the default value of the point size upon initialising a scatter plot was set at 2. The drop down menus show a list of the variables in the data, from which the x and y axis can be chosen. The list of possible variables is automatically generated when data are loaded into the software.

On each axis the minimum and maximum values for the variable relating to that particular axis are calculated and then displayed. If data are pruned on the plot thereby changing the minimum or maximum values, the new numbers are automatically calculated and displayed. The longest minimum and maximum values from all variables are calculated, recorded and used to calculate the position of the vertical axis from the left hand side of the window. This allows the text to fit in the optimal space whilst giving more of the plot's real estate to the drawing of the data points. This value is also used when calculating the starting position

to draw the label corresponding to the maximum value in the horizontal axis, in order to place it at the end of the axis.

The data points are drawn in a similar style to the data lines in parallel coordinates, using graphical primitives drawn to a buffer which are then saved and displayed until the data is brushed or pruned at which point the buffer is then updated. The plot also uses the same technique to mix colours and show transparency as the parallel coordinate plot. It is sometimes desirable to switch off the transparency for the scatter plot, so that the slider only affects the parallel coordinate plot. This can be achieved by using the option under the 'Tools' menu in the main window.

Histogram Plot

The histogram plot is set out in a similar way to the scatter plot, with two areas; one for the labels for the axes and another for the histogram. There are two input functions at the top of the window; the spin box sets the number of bins to be used in the histogram and the drop down menu selects which variable should be plotted. The variable list is constructed in the same manner as the scatter plot.

The number of bins used in the histogram can be changed through the corresponding number between 1 and 100 in the spinbox. This range of values was chosen after testing numerous variables and appears to give the most useful spread. However, it may be increased in future incarnations if it is deemed necessary. The default value is calculated using Scott's Normal Reference Rule [42]:

$$h = \frac{3.5\sigma}{n^{1/3}}, \tag{3.1}$$

where h is the width of the bins σ is the standard deviation of the variable and n is equal to the number of instances. The standard deviation of the data in a particular variable is calculated using the gsl_stats_sd function in the GSL library. More complex methods to calculate the number of bins used in the plot could be implemented in future versions of the software.

3.4.4 Data Formats and Storage

A general purpose file format was chosen to store the data loaded in and saved out of the program in order for it be useful between various groups within the particle physics community and outside of the subject. The comma separated value (CSV) file format was selected as it provides an easy to read format which stores data in a plain text form. The first line of the CSV file holds the names of the variables which the dataset populates. Each subsequent line holds a new record of values which corresponds to a data instance. Tab separation is used as the delimiter between values, in preference to commas as it allows for easier inspection of the data file by human eye. As the values for a particular variable are usually aligned in the same vertical column, it is easier to spot input errors and corruption of the dataset.

The data is read from the CSV file into the program using a dedicated function which loads it into multidimensional arrays. These arrays are stored using memory allocated from the heap rather than the stack as this allows many different functions or threads to access the same area in memory. This is particularly useful for allowing fast parallel reading and manipulation of the memory via OpenGL. When the data file is closed in the program, which automatically occurs if the program is exited, this memory is given back to the OS.

Although putting all the data straight into memory means that the program can run very quickly compared to repeatedly reading from disk, it does put demands on certain resources. In particular, it is useful to have a decent amount of Random Access Memory (RAM), available on the computer. However, this does not appear to be an issue for the data sets used in testing the software as the hardware currently available is plentiful with large capacities of high speed RAM.

3.4.5 Toolbar

There are various functions available once data has been loaded into the program. These functions control the ordering of axes in the parallel coordinate plot, create new histograms and scatter plots, log actions, rotate the parallel coordinate

plot, brush data, prune data, set colours and set transparency. This section describes some of these functions in detail. The buttons used in the toolbar are implementations of the QToolButton, whilst for the alpha value a combination of QSpinbox and QSlider was used. The pictures used on the buttons inside the toolbar are from the Gnome 2.18 icon theme [43], and are free to use under GPL.

Ordering of the axes

The ordering of the axes tool activates a function to attempt to find the most advantageous ordering of the axes in the parallel coordinates plot. As described in section 2.4, the order of the axes in the parallel coordinates plot is important as it can show patterns between variables which otherwise could be missed.

Through testing of the parallel coordinates system in both this program and other software it was found that one powerful method of finding interesting features in the dataset was to place the axes of highly correlated variables next to each other. This correlation could be positive, so orthogonal to the axes the data lines are more parallel, as in Figure 2.11; or the correlation could be negative where the data lines would cross each other to the maximum degree, as in Figure 2.10.

An algorithm was created in order to assign positions for each of the axes in the plot based on the correlation coefficients between the variables relating to those axes. The first step of this process is to calculate all the coefficients between each variable to all the other variables. The gsl_stats_correlation function from GSL was used to calculate the coefficient for each pair of variables. As both positive and negative values of correlation are of use here, the modulus of this value was then stored in a matrix.

Rather than finding the order of axes which produces a plot with the highest average values of correlation coefficient between the variables, it is of more interest to find the pair of variables with the highest value, then find the next highest pair and so on until all variables are used up. The issue here is to make sure that after pairs of variables are used that they cannot then be reused when looking for the next highest valued pair. From the matrix of correlation coefficients calculated using GSL, the pair of variables with the highest value were found and then

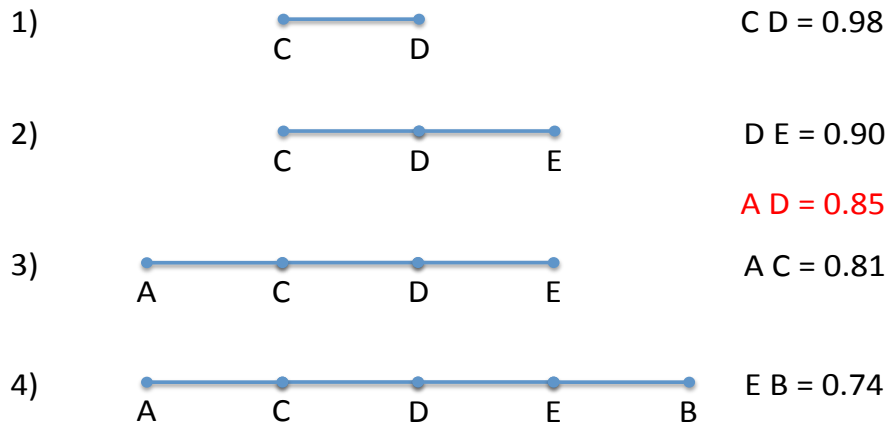

Figure 3.5: A diagram describing the algorithm which calculates the suggested order of axes for the parallel coordinate plot. Five variables: A, B, C, D, E are shown with the correlation coefficients given on the RHS.

stored in a double-ended queue referred to as a deque. Then the next highest value using one of the variables from the original pair is chosen and added to the correct end of the deque. The algorithm then processes the rest of the correlation coefficients relating to the variables, which are not already in the deque to the two variables at each end of the deque and finds the next maximum. It then adds the variable relating to this maximum to the correct end of the deque. This process is repeated until all variables are used. Figure 3.5 shows a diagram describing this algorithm.

The effect of running this algorithm on an example dataset is shown in Figure 3.6. The screen shot on the left hand side shows the order as given from the list of variables in the input file. Once the algorithm has run, the resulting order is used to display the data as shown on the right hand side of the screen shot. Variables that are closely correlated are placed next to each other. This accentuates features of the plot such as areas dominated by signal and those by background.

Log functions

The 'Log' function records the consequences of cuts upon the data thereby allowing a researcher to keep track of their work. When activated, it records a suitable output that is normally displayed in the terminal during the program operation.

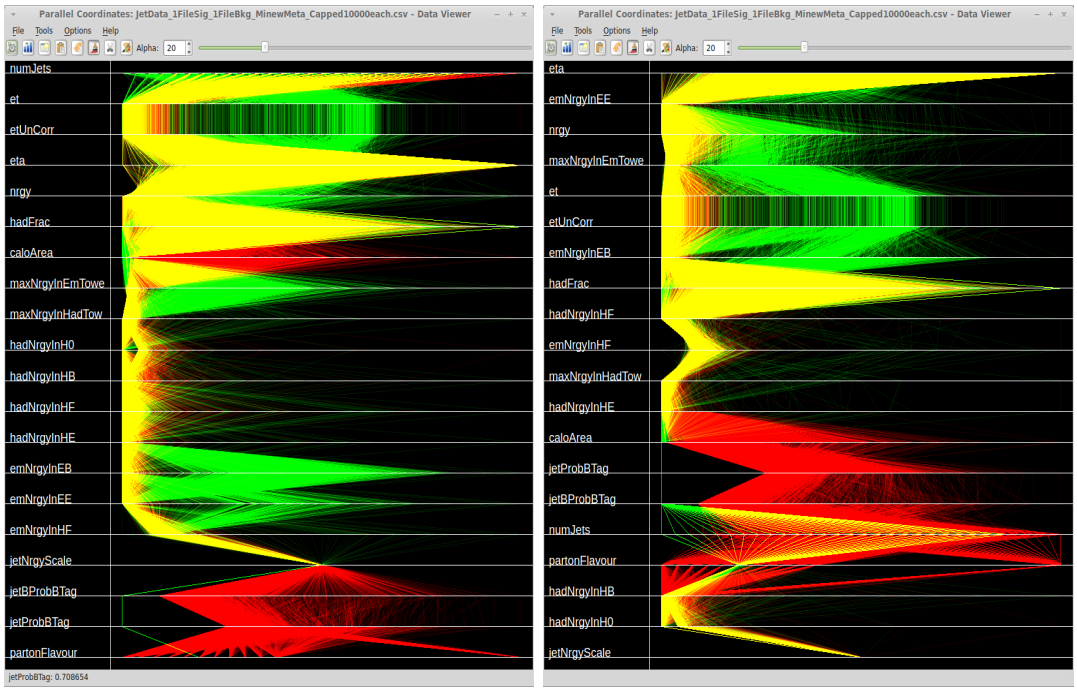

Figure 3.6: A screen shot of the program before and after running the algorithm to reorder the axes.

A log file is created and stored in the base directory of the software. The file is named using a time stamp from the point at which the application was launched to avoid overwriting other log files and to keep the logs in a logical order.

Rotate views

The orientation of the parallel coordinates plot can be changed, as seen in Figure 3.7. The rotate function changes a simple Boolean flag within the code signifying which orientation is requested. The code adjusts the way it calculates the width and height of the plotting area and then uses the in-built translate and rotate functions of OpenGL to give the desired effect. By using the in built functions of OpenGL no performance is lost when using one view over the other and the switching of views is performed in a speedy manner.

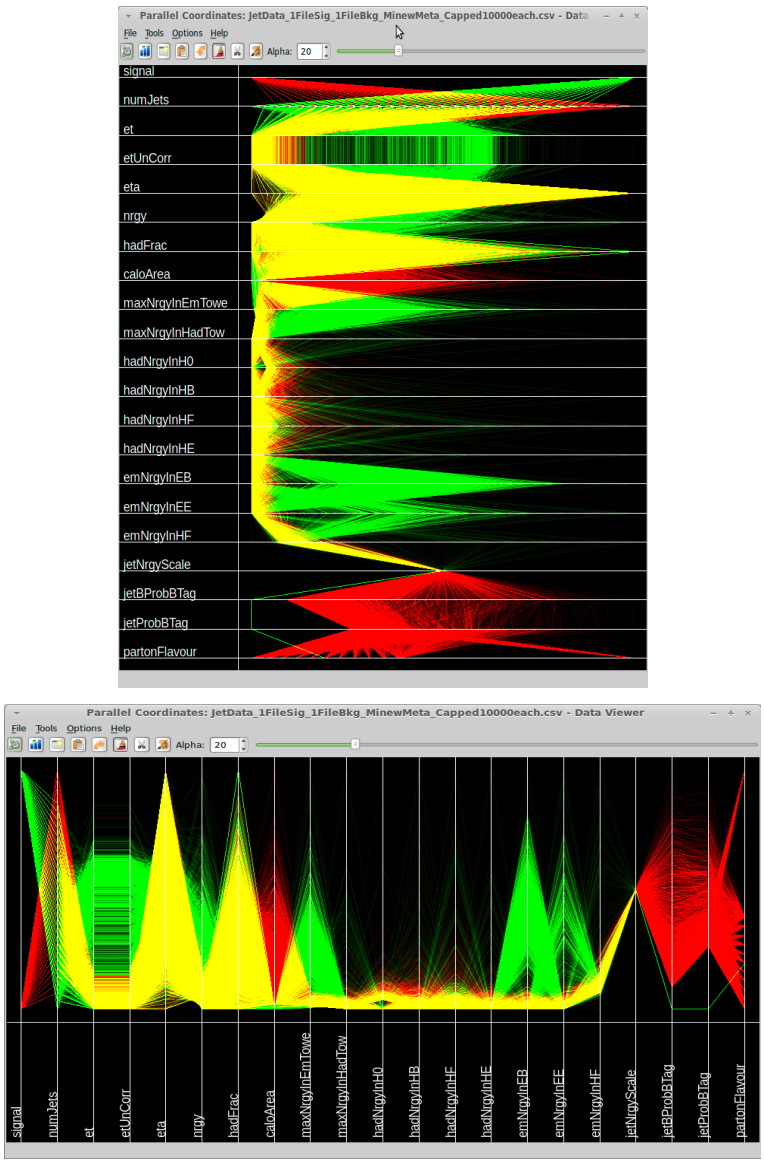

Figure 3.7: Screen shots of the parallel coordinate view in both vertical and horizontal layouts.

Brushing data

One of the two main actions used in the program is to colour the data, a technique known as brushing the data. A suitable colour should be chosen by using the set colour tool before brushing. When in brushing mode the cursor can be used to select data which the program then changes to that selected in the colour dialogue. Black has been disabled as this would cause the data instances to disappear into the background.

The brushing tool is used in a toggle state with the pruning tool, whereby selecting one deselects the other. The application was constructed in this way as there is no need to both colour in and delete data at the same time, as this would lead to confusing results. This toggle mechanism was created through the use of flags and Qt's signals and slots. By default the highlight option along with the colour white, which is the colour the of the data upon loading, is selected when the application is launched so that if data is accidentally selected no effect takes place.

As previously mentioned the various plots are linked, so if data is brushed with colour in one plot, the points or lines which represent the same data in all the other plots are automatically updated with the new colour. This is a very powerful technique as it shows how the data populates the different variables thus providing information on how the variables relate to each other.

Pruning data

The second of the two main actions used in the program is to remove data instances from the plots, through a technique called pruning activated through the prune tool. When in pruning mode, selected data is removed them from the plots.

As described in Section 3.4.2, the data are stored in a vector from memory. A map is then used which associates each data instance in that vector with a Boolean operator that marks if that instance should be visible or not. Instead of actually removing the data instances from the vector which would be a slow process, the program just alters the Boolean flag of the relevant data instances to make them

invisible. This invisibility not only stops the data from being seen in the plots but also affects the brushing and pruning of the data as well as modifying the statistics associated with those data.

If the 'Set signal and background' tool, as described in the next section, is used then the statistics relating to the plot are printed to the terminal. These statistics are based on which instances are visible in the plot, so by pruning data from the plot the statistics are updated; showing how many instances are left in the plot compared to when the signal and background were set.

Setting colours

Before brushing the data, a suitable colour must be chosen. This can be done via the colour palette tool which launches a colour selection dialogue, an implementation of Qt's QColorDialog. A colour can then be selected from this dialogue. The appearance of the colour selection dialogue is dependant on the computer OS, for example the Mac OSX version of this dialogue has five different ways to select the colour.

Only certain colours work as expected within DataViewer; red, green, blue, cyan, magenta and yellow. These are the colours which have RGB values which are combinations of 0.0 or 1.0. If colours are chosen which do not have RGB values with these combinations then when there are many data points or lines occupying the same space on the plot the colours appear white. An example of this effect is seen in Figure 3.8. This limitation is believed to be due to how the OpenGL colour blending is utilised in order to include transparency in the plots.

There are two main reasons for using colour in the program. Firstly, to highlight trends and/or patterns in the data. Secondly, to categorise the data into two sets, one for signal and one for background. The colouring allows the program to calculate the signal and background numbers. This is part of a two step process, where the data belonging to signal and background are categorised via brushing before implementing the 'Set signal and background' tool.

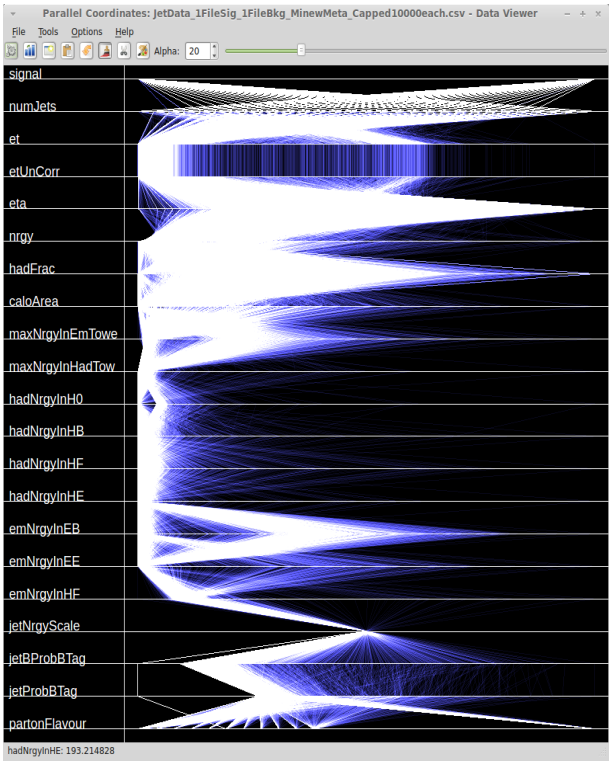

Figure 3.8: Screen shot showing the problem using certain colours in DataViewer.

Setting transparency

As described in Section 2.4.2, parallel coordinate density plots are highly advantageous over the normal parallel coordinate display. Control over the level of transparency is important in the density plots in order to utilise the plots to their full potential. For this reason a large slider and spinbox has been placed in the main window to control the alpha channel in the plot. The values of alpha which are allowed in OpenGL lie within the range of 0.0 to 1.0. Therefore, a simple map was used to translate the scale of 1 to 100 as shown on the slider and connected spinbox to the allowed range. After testing different transformations between the two ranges, a simple division by 100 was used as this gave a satisfactory effect.

The slider can also be used to change the transparency of the scatter plots to show, as with the parallel coordinate density plot, where the majority of data lies. This in effect gives a plot which is similar to the heat map display where colour shows the density of points. However, one advantage of using the alpha channel in the scatter plots over the heat map is the ability to also categorise the

data through colours.

3.4.6 Tools

Resetting views

The plots can be reset to their original state before any cuts were made. Resetting views does not affect the colouring of the data as testing showed this was usually undesired. With the lack of a history function in this version of DataViewer the resetting views is especially useful as a method of undoing previous actions. This mechanism of resetting the plots is much more efficient than closing the data file and reloading it back into the program, which would take longer whilst data are read from disk into memory.

Setting signal and background

In the situation where there are only two categories in the data, signal and background, it may be useful to know how many data instances there are for each type. The program allows the identification of signal and background instances by using the 'Set Signal and Background' function. A function is then activated which finds the number of data instances for each category and displays this information, as well as the ratio, via the terminal. Whenever a cut is made upon the data these statistics are automatically updated. This provides a measure of the efficiency and purity of cuts applied.

Finding maximum correlation

During testing of the software, an interesting function was realised which shows the other variables most correlated with a chosen variable. This is particularly relevant if a signal variable which discriminates the signal and background has been created, which can then be used in conjunction with this function. The program could give direction as to which variables might give the best discrimination

between the signal and background. This function uses the GSL correlation coefficients, in the same way as the automatic ordering of the axes, to print out the order of most correlated variables. The 'Find Maximum Correlation' function calculates the correlations to a variable and the output is presented in the terminal.

3.4.7 Options

Automatic adjustment of plot axis scale

By default whenever a cut is made in any of the plots the data range is recalculated and the plots updated in order to fill the maximal amount of space as described in Section 3.4.3. However, in some instances it may be desirable to stop this automatic rescaling of the plots, for example when directly comparing two variables and wishing to view how making a cut on one variable affects another. The 'Auto adjust axis on plot' function toggles this automatic adjustment on and off.

Use of transparency in Scatter plots

During testing of the software it was found that controlling the switching on and off the transparency in the scatter plots was desirable. This is especially true when using a high amount of transparency for the parallel coordinates plot, with low transparency in the scatter plots. Usually the data points in the scatter plots are drawn with a small point size so that the probability of many points lying on top of each other will be lower than in the parallel coordinates view. A function, 'Use Alpha in Scatter Plots' is provided to toggle on and off the use of transparency in the scatter plots.

Selecting Visible Variables in Parallel Coordinate view

It is often useful not to show all the variables in the data within the parallel coordinates view. If some variables do not provide useful information then they are detracting space and attention away from the variables which are of more interest. To make good use of the plotting space, the area of the plot given to those variables of interest should be maximised. This can be achieved by using the 'Select Visible Variables' function or alternatively through the use of a key stroke via an implementation of Qt's QKeyEvent.

There are several reasons why it is desirable to hide variables from the parallel coordinate plot. Beyond trying to maximise the amount of space given to the variables which are of most interest, it is sometimes useful to not know which data instances belong to which category. This technique is called 'blinding'. Removing unnecessary variables also increases the performance of the software when updating or refreshing the plot, for example when removing data or resizing the window. This is especially useful when the program is run on computers with low hardware specifications.

3.4.8 Current Capabilities and Future Work

Review of DataViewer

The current version of DataViewer (v1.1.0_Beta) addresses many of the requirements described in Section 3.3. The software gives very high performance with smaller datasets and satisfactory usefulness with extremely large datasets. This performance is quantified in Table 3.1 which shows the speed of various operations on different datasets consisting of different numbers of data instances as well as different numbers of variables, or dimensions. The data used was comprised of randomly generated numbers from ROOT's TRand3 function. The benchmarks were run on a laptop with an Intel C2 DUO P8600 2.40 GHz 1066 MHz Central Processing Unit (CPU), 8 GB (2 × 4 GB) DDR2 PC 6400 800 MHz RAM, WXGA+ 1440 × 900 + NVIDIA 9650M GT 512 MB GDDR3 Graphics Processing Unit (GPU), and 320 GB 5400 rpm SATA Harddisk. The tests were run

under the Ubuntu 11.10 GNU/Linux Operating System.

The laptop utilised does not have top of the range specifications and so gives a reasonable account of the performance expected from the software. From these tests, as well as extensive usage as part of the work described in later chapters, it is recommended that the software be used on datasets with around 20-Dimensions or less and approximately a million data instances or less to gain good performance and usefulness. With a lower number of dimensions more data can be viewed with higher performance. In practice it was found that typically less than 20-Dimensions of data were viewed. Also when irrelevant variables are removed from the plot the relative performance of the software increases.

There is also a limit to how much data can be shown within DataViewer due to the range of alpha values allowed. When the program was loaded with the 10,000,000 data instances used in the benchmark test, the parallel coordinate plot became completely white. The alpha value was set to 1, using the alpha slider on the main window, which corresponds to a real alpha value of 0.01. However, with the amount of data present the plot still became over saturated.

The program compares extremely well to some of the other visualisation software described in section 3.1. This is most likely due to the direct use of the GPU via the OpenGL API and through sensible memory management as discussed in section 3.4.4. The use of transparency provided more useful plots compared to those in GGobi.

During testing a number of issues and improvements were identified as described in the next section.

Future work

It has become apparent that a number of improvements could be made to DataViewer. The SDI interface was chosen during the design stages for reasons outlined in section 3.4.2. However, after testing DataViewer on a range of different OS's it may be more useful to use the MDI interface instead. The reasoning behind this conclusion is that when plotting windows are hidden from view, for

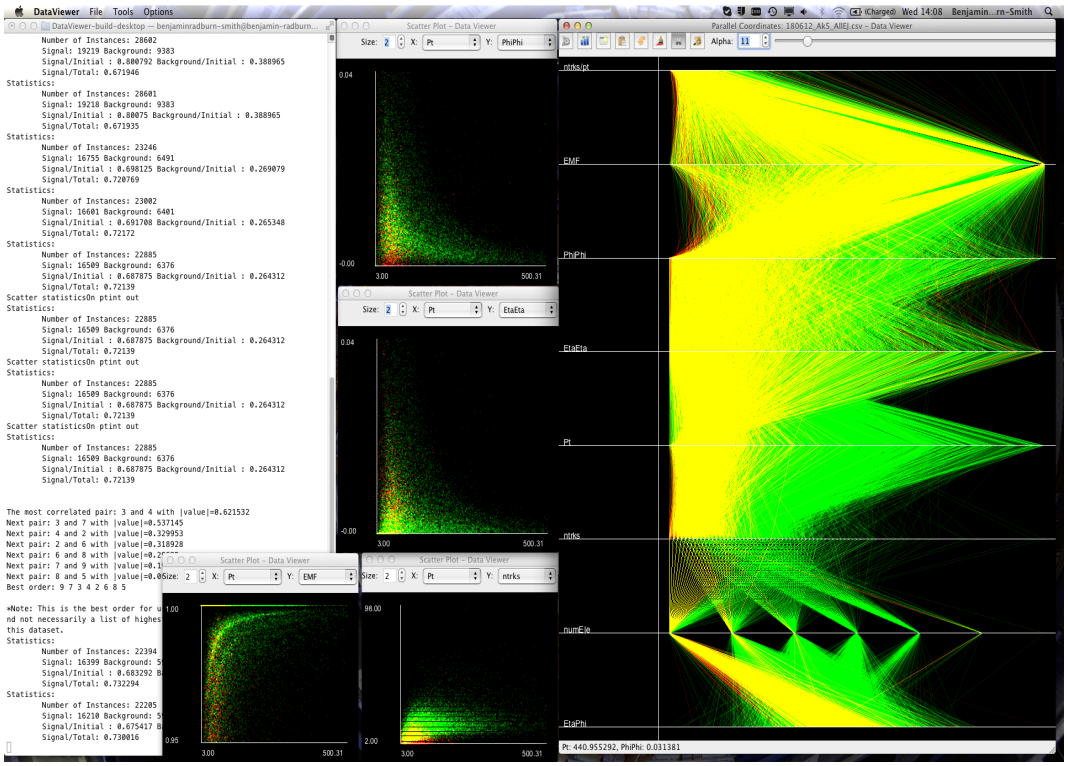

Figure 3.9: Screen shot of DataViewer while running in Mac OSX.

example if more than one application is running on the same workspace and that window is clicked, it was found that more effort was required to keep the display in a functional state than would be required if there was just one main window which held all the plots.

Additionally, when running on different OS, the appearance of DataViewer alters leading to undesirable layout effects. Figure 3.9 shows an example of this effect. In future versions of the program this should be accounted for in order to ensure a consistent experience independent of the OS. However, it should be noted that the windowing system of each OS will impose its own rules on how the software should look.

It would be greatly beneficial to DataViewer's usefulness within particle physics if it were able to read straight from ROOT files, without creating and manipulating CSV files. If variables could be selected straight from the ROOT ntuples the visualisation software could be run before the user's analysis code. This would

mean the visualisations could quickly influence the analysers design and, for example, show which variables would be most interesting to investigate. It would also allow for identification of any problems that may have occurred in the data processing up to that point. This method of working would greatly boost the efficiency of a particle physics analysis.

Some initial work has been carried out to get DataViewer to read the data straight from ROOT ntuples. However various issues arose during this process and the work was abandoned due to time constraints. The process of extracting the meta information from the ntuples prior to loading the data led to reliability issues between different datasets. Another problem exists with data in a dataset having a different number of variables or attributes attached to them. The parallel coordinates implementation insists on consistently having the same variables between all data, otherwise the polylines representing the data became broken thus making it impossible to understand the properties of the data.

DataViewer uses OpenGL which uses parallel processing when drawing the graphics for the plots. The rendering performance can be seen in the 'Time to Update' column of Table 3.1. However, other parts of the program do not utilise parallel processing. Performance gains can be achieved across the board if this technique was used throughout the program. Data could be loaded quicker by reading multiple streams simultaneously. The 'Time to Load' column of Table 3.1 shows that the current software's loading times could be optimised. Faster calculations of how the data needs to be scaled (normalisation of the data) could be achieved. The process of removing data from the plots and brushing the data would also be quicker if parallel processing was used compared to the current performance as recorded in the 'Time to Delete Data' and 'Time to Colour Data' columns of Table 3.1, respectively. Finally the calculations required to find the correlation coefficients and subsequently to automatically order the axes into an interesting order would benefit from parallel processing.

Currently only the removal of data from the plots can be saved. Data polylines can be deleted from a plot and the resulting dataset saved as a CSV file with a specific name. A more sophisticated save function is envisaged which records the removal of variables and also saves which data is brushed with the selected colour, thus allowing a quick return to a specific point in the data exploration

process.

Within this version of DataViewer there is currently no history function, which would be required to implement undo and redo actions. Qt offers a history mechanism which could be used to implement this function, however due to time constraints a stable implementation of this mechanism could not be fully investigated. Allowing the user to undo an action is of great importance and should be of high priority in any further developments in DataViewer, as often the user may misclick within the plot leading to undesired consequences and/or perform an action on purpose only to find it does not give the desired outcome. Without the undo/redo actions as part of a history function the user has to use the 'Reset' function which draws the plots back as they were initially shown upon loading the data.

Implementations of the grand tour within the scatter plot view as well as within the parallel coordinates view have not yet been incorporated into DataViewer. As described in section 2.5 these techniques would provide even more insight into the data and allow the user to find interesting hyperplanes which could be used to categorise or differentiate the data.

Section 3.4.5 detailed the limitations of which colours could be used in DataViewer. It is believed that these limitations are due to the implementation of OpenGL's blending methods which are required to enable both the transparency in the plot and mixing of colours. The limited palette of available colours should be addressed in future versions of the code to allow free choice over which colours to use to categorise the data. During testing of the software many users prefer the green and red choices for signal and background however people with red/green colour blindness found this scheme sometimes difficult to follow. With a larger palette of colours available, all users should be able to use the software without these problems.

Other possible improvements of the software include:

- Dynamically setting the divider between the variable names and data in the parallel coordinate plot, in order to maximise the plotting area.
- Implementing an exponential alpha slider to set the level of transparency.

Testing showed that greater control of the transparency level at low values is more important than at higher values. With a larger range more data can be viewed in the program.

- Adding colours to the histogram plot in order to show how differently brushed data populates the bins.
- Allow alpha numeric characters in data rather than just numbers.
- Ability to load more data after the initial load. Currently the data can only be selected at the beginning.
- A function to allow the creating of new variables by manipulating those already loaded, for example multiplying two variables together.
- Ability to change the variable names once the data has been loaded into the software.
- A query tool to find the number of instances within a selected area without removing them from the plot.
- The ability to zoom in on the plots, for example through the use of the mouse scroll wheel. This would allow for a greater accuracy when using the program.
- A mechanism to have different levels of transparency in the different plots.

Number of Data Instances	Number of Dimensions	File Size (MB)	Time to Load	Time to Delete Data	Time to Remove Variables	Time to Update	Time to Colour Data
100	5	0.004	negligible	negligible	negligible	negligible	negligible
1000	5	0.039	negligible	negligible	negligible	negligible	negligible
10000	5	0.385	negligible	negligible	negligible	negligible	negligible
100000	5	3.8	1	1	negligible	negligible	negligible
1000000	5	38	7	9	2	2	5
10000000	5	376	73	86	15	16	57
100	10	0.008	negligible	negligible	negligible	negligible	negligible
1000	10	0.077	negligible	negligible	negligible	negligible	negligible
10000	10	0.770	negligible	negligible	negligible	negligible	negligible
100000	10	7.5	2	1	1	negligible	1
1000000	10	75	13	13	3	3	7
10000000	10	752	131	135	25	25	75
100	20	0.012	negligible	negligible	negligible	negligible	negligible
1000	20	0.116	negligible	negligible	negligible	negligible	negligible
10000	20	1.1	negligible	negligible	negligible	negligible	negligible
100000	20	11	2	2	1	1	2
1000000	20	113	23	23	5	5	13
10000000	20	1100	238	247	52	53	147
100	30	0.016	negligible	negligible	negligible	negligible	negligible
1000	30	0.154	negligible	negligible	negligible	negligible	negligible
10000	30	1.5	negligible	negligible	negligible	negligible	negligible
100000	30	15	4	4	1	1	2
1000000	30	150	35	36	8	9	21
10000000	30	1500	348	329	31	83	225

Table 3.1: Benchmark results from testing DataViewer with a variety of different sized input files. Time is measured in seconds using a stop watch.

Part II

Searches for Lepton Jets at CMS

Chapter 4

Theoretical Motivation

4.1 Symmetries and Gauge Theories

One method of trying to understand the universe is to explore the symmetries of the physical laws. Conservation laws are intimately linked to symmetries.

A global transformation is one which is carried out at all points in space-time. Conversely a local transformation is carried out differently at different points in space-time. If a global transformation is invariant, the system is said to have a global symmetry. Generally, global invariant theories are not invariant under local transformations. However, through the addition of new force fields which interact in a particular fashion with elementary particles, local symmetry can be restored. A gauge theory is one in which local phase transformations of quantum fields are invariant.

Quantum Electrodynamics (QED) is a gauge theory which emerged as the prototype of modern Quantum Field Theories (QFT) in the late 1940's. In QED, Maxwell's classical electromagnetism is married with quantum mechanics. The gauge field of QED is the electromagnetic field. The gauge invariance of QED introduces a massless vector boson, the photon. QED provides extremely accurate predictions with different experiments, such as measurements of the anomalous magnetic dipole moment of the electron, showing agreement between theory and

experiment to within 10^{-8} precision[1].

The Glashow-Weinberg-Salam (GWS) [46] gauge theory is a Yang-Mills theory, based on the symmetry group $SU(2)_L \times U(1)_Y$ [47]. This theory introduces four gauge fields associated with weak isospin and weak hypercharge. Upon spontaneous symmetry breaking this theory describes the electromagnetic and weak interactions, collectively referred to as electroweak interactions, between quarks and leptons. The quanta of the gauge fields are the $W^{\pm}$ bosons, the Z^0 boson and the photon.

Quantum Chromodynamics (QCD) [48] is a QFT in which a $SU(3)_C$ gauge group acts upon a degree of freedom called colour. QCD describes the strong interactions between quarks and introduces the gluon as the mediator of the strong force.

4.2 The Standard Model

The Standard Model (SM) of particle physics is a QFT which provides a description of all known elementary particles and the interactions between them. The SM was formulated in the 1970's, and describes every particle as having an associated field with the particles observed being excitations of these fields. The SM is a combination of the GWS and QCD gauge theories, producing a $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge symmetry.

There are two families of particles within the SM; the fermions which obey Fermi-Dirac statistics and the bosons which obey Bose-Einstein statistics. Fermions consist of quarks and leptons split into three generations. There are six quarks: generation I contains up and down quarks, generation II contains charm and strange quarks whilst generation III contains top and bottom quarks. Particles in the second and third generations share similar characteristics to those in the first, but have larger masses.

[1] Calculated through separate measurements of g-2 and atom-recoil velocities allowing the determination of the fine structure constant [44, 45]

The up, charm and top quarks carry an electrical charge with a ratio of +2/3 to the modulus elementary charge $|e|$ of the electron, while the down, strange and bottom quarks carry electrical charges equal to -1/3. The up, down and strange are conventionally classified as the light quarks while the charm, top and bottom quarks are heavy. The top is heavier than the other quarks and has a Yukawa coupling $\sim$1 which leads to the theory that it has a special role to play in electroweak symmetry breaking [49].

The three generations of leptons each contain one lepton with charge -1 and one neutral lepton called a neutrino. Generation I contains the electron and electron neutrino, generation II contains the muon and muon neutrino while generation III contains the tau and tau neutrino.

Summaries of the fermions and bosons are outlined in tables 4.1 and 4.2 respectively. Each fermion also has an antimatter counterpart with equal quantum numbers and opposite values for charge and handedness.

Due to the nature of QCD, coloured particles such as quarks and gluons do not propagate over macroscopic distances but are instead confined into colour singlet bound states called hadrons. Hadrons can be observed using experimental detectors. There are two types of hadron observed in nature: mesons and baryons. Mesons are composed of a quark and an anti-quark. Protons and neutrons, which form the nuclei of ordinary matter that is visible throughout the Universe, are baryons which contain three quarks bound in a colour neutral state.

There are five gauge bosons in the SM: the force carrier of electromagnetism, the photon; the bosons of the Weak force, the Z^0 and $W^{\pm}$; the carrier of the Strong force, the gluon; and the Higgs Boson. The strong interactions are responsible for binding the quarks inside the proton and neutron; while the weak interactions are responsible for the radiation of nuclear β-decay. The purpose of the Higgs boson is described Section 4.3.

Type	Generations			Charge
	I	II	III	
Quarks	(u) up 0.003	(c) charm 1.3	(t) top 175	+2/3
	(d) down 0.006	(s) strange 0.1	(b) bottom 4.3	-1/3
Leptons	(e) electron 5.1×10^{-4}	(μ) muon 0.106	(τ) tau 1.78	-1
	(ν_e) electron neutrino $<1\times10^{-8}$	(ν_μ) muon neutrino $<2\times10^{-4}$	(ν_τ) tau neutrino <0.02	0

Table 4.1: A summary of the fermions, with the mass (in GeV) of the fermions written under their names.

Name	Charge	Mass (GeV)
photon (γ)	0	0
W^+	+1	80.4
W^-	-1	80.4
Z^0	0	91.2
gluon (g)	0	0
Higgs (H)	0	$\sim$ 125 GeV[a]

[a] See Section 4.3 for comments on the observation of a Higgs like boson by the ATLAS and CMS collaborations

Table 4.2: A summary of the gauge bosons.

4.2.1 Problems with the SM

The SM has proven to be a successful theory with predictions agreeing with experimental data. However, there exist certain phenomena that are not explained by the SM. In the SM neutrinos are massless. However, experiments with solar, atmospheric and reactor neutrinos have observed neutrino oscillations, whereby the neutrinos change flavours as they propagate through space [50–52]. The implication is that there is a non-zero mass difference between the two states that mix [53].

The SM fails to explain gravity and subsequently gravitons. The Wilkinson Microwave Anisotropy Probe's (WMAP) observation of the Cosmic Microwave Background (CMB) radiation shows the Universe is made up of ∼23% cold dark matter and ∼73% dark energy [54], which is not explained by the SM. The matter/antimatter asymmetry of the Universe is also not explained by the SM.

Other issues with the SM include the hierarchy problem. There appear to be two fundamental energy scales in nature; the electroweak scale at $\mathcal{O}(10^2)$ GeV and the Planck scale, M_P at $\mathcal{O}(10^{18})$ GeV, where gravity becomes as strong as gauge interactions [55]. The reason for this vast discrepancy in scales, or hierarchy, is unknown. One effect of this problem is the requirement of fine tuning to the mass of the Higgs boson, m_H. Loop corrections drive m_H to the Ultraviolet energy scale, Λ. If $\Lambda \sim \mathrm{M}_P$, then an equally large value of the Lagrangian parameter in the Higgs potential is needed to give a high accuracy cancellation and ensure m_H remains $\sim \mathcal{O}(10^2)$ GeV [56].

Also of note within the SM, the weak and electromagnetic forces are unified into the electroweak force. However, the electroweak and strong forces are not unified and therefore the SM is not a grand unified theory (GUT).

4.3 The Higgs Boson

Weak interactions are short range and mediated by massive W and Z bosons. However, in the Yang-Mills theory gauge bosons are required to be massless in order to preserve gauge invariance [57]. In order to overcome this problem a mechanism for spontaneously breaking the SU(2)×U(1) gauge symmetry needs to be introduced. The Brout-Englert-Higgs (BEH) mechanism [58, 59] provides a solution to this electroweak symmetry breaking (EWSB) problem.

In the BEH mechanism elementary particles interact with a scalar field, the Higgs field. This interaction occurs through the quanta of the Higgs field, referred to as the Higgs boson [59].

The Higgs field has a non-zero Vacuum Expectation Value (VEV). This non-zero value is not invariant under a (local) gauge transformation and so the gauge invariance is said to be hidden or spontaneously broken [60].

Three key effects result from the existence of the Higgs field. Firstly, the Weak bosons, $W^{\pm}$ and Z^0, acquire mass from interactions of the electroweak gauge fields with the Higgs field. Secondly, fermions gain mass from interactions with the Higgs field. Finally, a neutral particle representing the quanta of the Higgs field, the H^0 boson, would exist.

The profile of the SM Higgs boson is uniquely determined once its mass M_H is fixed [47]. The decay width, branching ratios and production cross sections are supplied by the Yukawa couplings to the fermions and bosons. These couplings depend on the masses of the particles.

On the 4th July 2012 the ATLAS and CMS experiments at CERN announced the discovery of a new boson. This new boson has been measured by the ATLAS collaboration to have a mass of 126.0 ± 0.4 (statistical error) ± 0.4 (systematic error) GeV [61] and by the CMS collaboration to have a mass of 125.3 ± 0.4 (stat.) ± 0.5 (syst.) GeV [62].

Properties of the Higgs particle now require measurement such as the mass, width, charge, spin, parity, couplings to other particles, and self-couplings. Whilst it

should be possible to determine the mass and spin of the new boson using the 2011 and 2012 datasets, significantly more data is required in order to measure precisely its couplings to other particles. The couplings measurements would test the internal consistency of the SM or could provide a portal to physics beyond the SM.

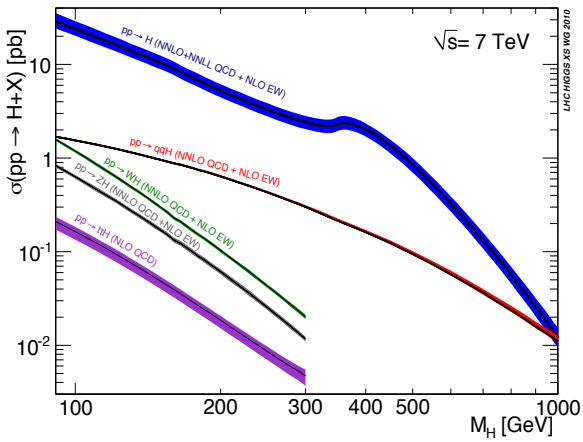

Figure 4.1: Standard Model Higgs boson production cross sections at $\sqrt{S} = 7$ TeV. The blue line represents gluon fusion, the red line represents vector boson fusion, the green and grey lines represent WH and ZH Higgstrahlung respectively, and the purple line represents tree level top coupling [63].

Figure 4.1 shows the SM Higgs boson production cross sections [63]. The gluon fusion process (shown as the blue line) provides the production mechanism with the largest cross section. This is followed by vector boson fusion (VBF) (red) and the Higgstrahlung process (as green for WH and grey for ZH) process. Other Higgs production mechanisms (e.g. the tree level top coupling (purple line)) with much smaller cross sections also exist.

The VBF and Higgstrahlung processes provide distinct signatures which help in the identification of a Higgs decay. Of particular interest is the Higgstrahlung process in which an associated vector boson is produced. This vector boson can be identified which reduces the number of background events thereby improving the purity of Higgs events.

4.4 Hidden Valley Models

One possible extension to the SM is the hidden valley class of models [64]. In these models, hidden or dark sectors exist which contain particles that are relatively light. Higher dimensional operators allow interactions between these light particles and SM fields [64]. High energies are required to produce these operators.

The hidden sectors are provided through an extra non-abelian group, which if broken at the GeV scale, produce particles with low mass and potentially long lifetimes. The masses, lifetimes and multiplicities of the hidden hadrons are sensitive to underlying parameters, such as the hidden quark masses. Heavy particles carrying charges from both the SM gauge group and a new hidden gauge group allow interactions between the new particles and SM fields. These dark sectors usually weakly interact with normal matter. Figure 4.2 shows a graphical representation of the hidden valley.

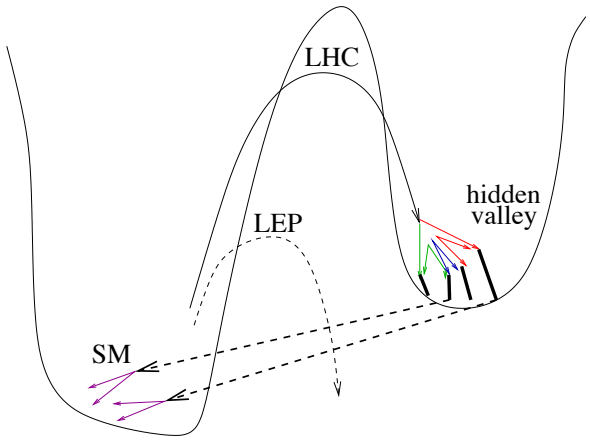

Figure 4.2: Graphical representation of the hidden valley model. The LHC may reach high enough energies to penetrate the barrier between the SM and dark sector, where LEP could not. [65]

If some of the dark sector hadrons were stable they would be candidates for dark matter. The experimental signature for dark matter at a collider would be missing momentum. Other particles would decay to neutral combinations of SM particles. Potentially the Higgs could interact with the new field and if kinematically allowed, could decay to dark sector hadrons, although this process

would be rare.

4.4.1 Lepton Jets

Within the dark sector, dark particles can decay to lighter dark particles. This process is referred to as cascade decays, which increase particle multiplicity. Particles in the dark sector can produce dark photons which are permitted to mix with the SM photon and decay into leptons. The resulting collection of collimated leptons are called Lepton Jets (LJ) [66, 67]. Figure 4.3 depicts the cascade decay producing LJ.

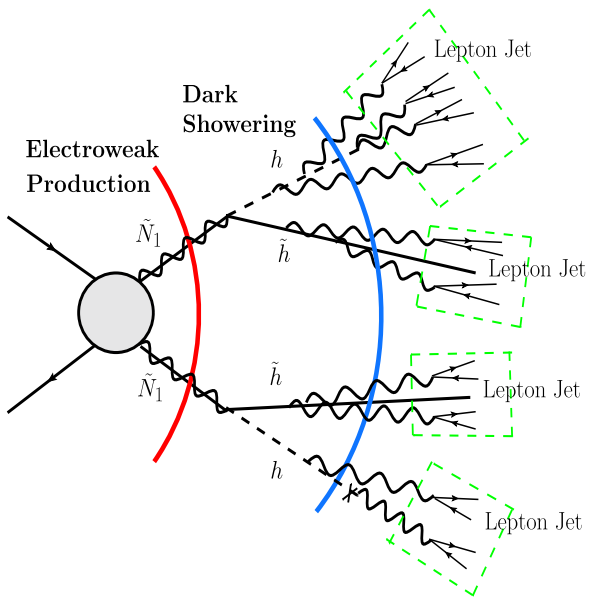

Figure 4.3: An illustration depicting a cascade decay in the dark sector producing multiple LJ. In this example Neutralinos, $\tilde{N}_1$, decay to hidden Higgs bosons, h, and hidden Higgsinos, $\tilde{h}$, producing lepton jets [67]

In order to allow the dark sector to become visible again, interactions are allowed through a kinetic mixing with the dark photon γ_d, of mass $m_{\gamma d}$ and the hypercharge field, B_μ,

$$\mathcal{L}_{mix} = \frac{1}{2}\epsilon\, \gamma_d^{\mu\nu} B_{\mu\nu} = \frac{1}{2}\epsilon\, \gamma_d^{\mu\nu} \left(\cos\theta_W A_{\mu\nu} - \sin\theta_W Z_{\mu\nu}\right) . \tag{4.1}$$

Where $\gamma_d^{\mu\nu}$ is the field strength of γ_d and $B_{\mu\nu}$ is the field strength for B_μ. θ_W is

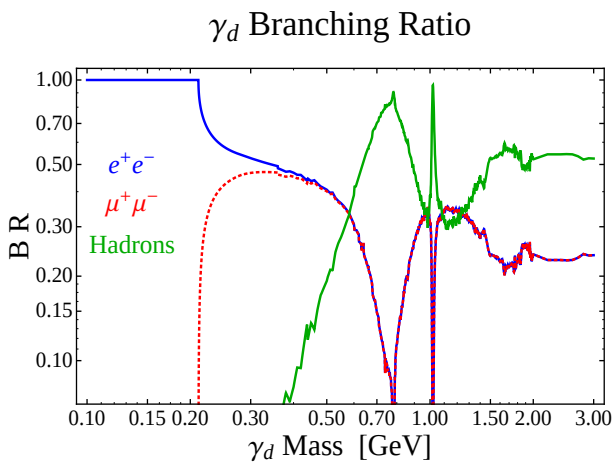

Figure 4.4: Plot of the Branching Ratio of the dark photon versus the dark photon mass, $m_{\gamma d}$. If $m_{\gamma d} < 200$ MeV the dark photon decays exclusively to electrons [68].

the Weinberg angle and ϵ is the mixing parameter which is assumed to be small at $\epsilon \leq 10^{-3}$ [68]. The photon mixing can be removed by a shift of the photon field,

$$A_\mu \rightarrow A_\mu + \epsilon \cos\theta_W \gamma_d. \tag{4.2}$$

The dark photon couples to all electrically charged particles with a strength $\epsilon e \cos\theta_W$ allowing for decays to leptons and hadrons. The decays to electrically neutral particles that couple to the Z are suppressed by $m_{\gamma d}^2/m_Z^2$ and can be ignored [68].

The particles into which the dark photon decays is dependant upon the mass of the dark photon. If $m_{\gamma d} < 200$ MeV the dark photon decays exclusively to electrons as shown in Figure 4.4. This results in Electron Jets (EJ).

Hidden Valley models producing EJ provide explanations for various astrophysical anomalies, such as the excess in the positron fraction of cosmic rays, which rise to between 20 to 200 GeV without any antiproton excesses. The positron fraction cannot be explained by secondary production processes resulting from cosmic ray nuclei interacting with the interstellar gas. This excess is seen by many experiments including the PAMELA [69] and FERMI Large Area Telescope satellites [70] and the AMS-02 experiment [71] (Figure 4.5). There has also been an excess in the electron+positron spectrum observed by, amongst others,

the MAGIC telescopes [72] (Figure 4.6).

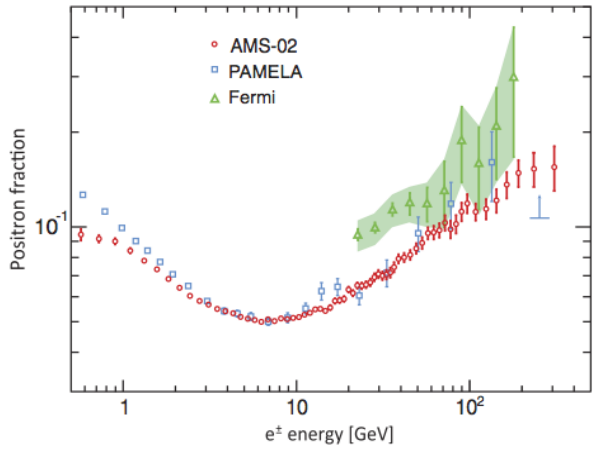

Figure 4.5: The increase in positron fraction with respect to energy as measured by PAMELA, Fermi LAT and AMS-02 experiments [71].

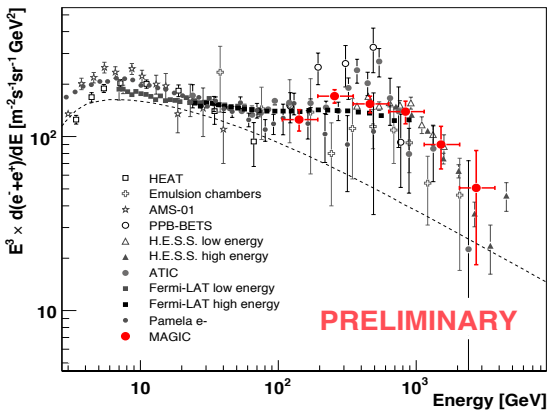

Figure 4.6: The overall electron + positron excess with respect to energy as measured by numerous telescopes including the MAGIC telescopes [72].

There are numerous motivations to search for LJ signatures. From a top down perspective, the hidden valley scenario arises in many string-theory constructions and appears to be consistent with most methods for solving the hierarchy problem [64]. The lightest stable particle in the dark sector is a dark matter candidate. For low mass dark photons, EJs are produced which can provide an explanation to astrophysical anomalies. Also the LJ signatures provide an interesting signature in their own right, which pushes the detection and reconstruction within experimental devices to the limit.

4.4.2 Hidden Higgs

A Higgs boson can decay into a light hidden sector through direct couplings. The resulting cascade decay within the dark sector produces lepton jets as described in Section 4.4.1. The idea of a hidden Higgs is more naturally expressed within the context of supersymmetry (SUSY). The Minimal Supersymmetric Standard Model (MSSM) incorporates two Higgs doublets to give five Higgs bosons. An additional singlet can couple to the Higgs doublets. This singlet has associated scalars, χ_1 and χ_1^*, which are charged under the hidden sector. Mixing between the MSSM Higgs bosons and the hidden particles can occur if these scalars obtain non-zero VEVs [68].

The lightest MSSM Higgs boson can decay to two lighter scalars, χ_1 and χ_1^* (Figure 4.7). These scalars can then decay to two dark sector Higgs bosons, h_d (Figure 4.7). The dark sector Higgs bosons can each decay to two dark photons, γ_d, which then further decay to two leptons as shown in Figure 4.8.

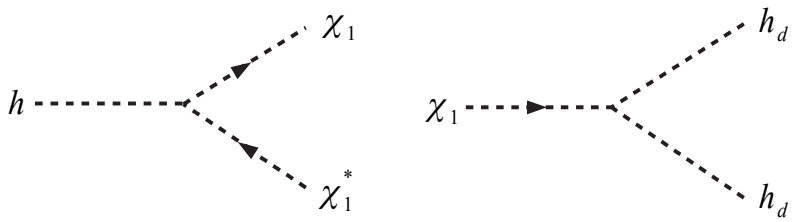

Figure 4.7: Diagrams showing the Higgs decaying into a hidden sector. The left hand diagram shows the decay of the lightest MSSM Higgs decaying into two lighter scalars, χ_1 and χ_1^*. These scalars can then decay into hidden scalars such as dark sector Higgs, h_d [68].

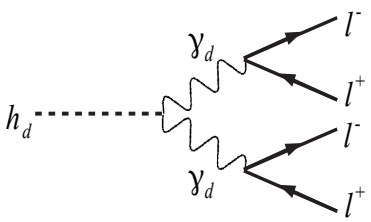

Figure 4.8: Diagram showing the decay of a hidden Higgs, h_d, into a pair of dark photons, γ_d, which then decay into a pair of leptons, l [67].

A complete Higgs decay chain using the particle notation used for the rest of the thesis is shown in Figure 4.9. This is an example of a 3 step decay representing

the three stages in the dark sector. The first dark stage contains the light scalars (h_{d2}), the second contains the dark sector Higgs (h_{d1}) and the third contains stable dark sector particles (h_{d0}) and dark photons (γ_d).

For the analysis described in Chapter 6 the mass of the dark photon is set to m_{γ_d} = 100 MeV ensuring a decay to electrons. The masses of the other three dark sector particles can be varied, changing the characteristics of the dark sector, as per Section 6.2. The branching ratio for the h_{d1} particles decaying to h_{d0} and γ_d was set at a representative benchmark of 0.8 and 0.2 respectively after consultation with experts [73, 74]. The branching ratio of the Higgs decaying into the dark sector is set to 1 [73]. As the dark photons are constrained to decay to electrons, the value of ϵ in Equation 4.2 only affects the decay length. A value corresponding to a prompt decay was used.

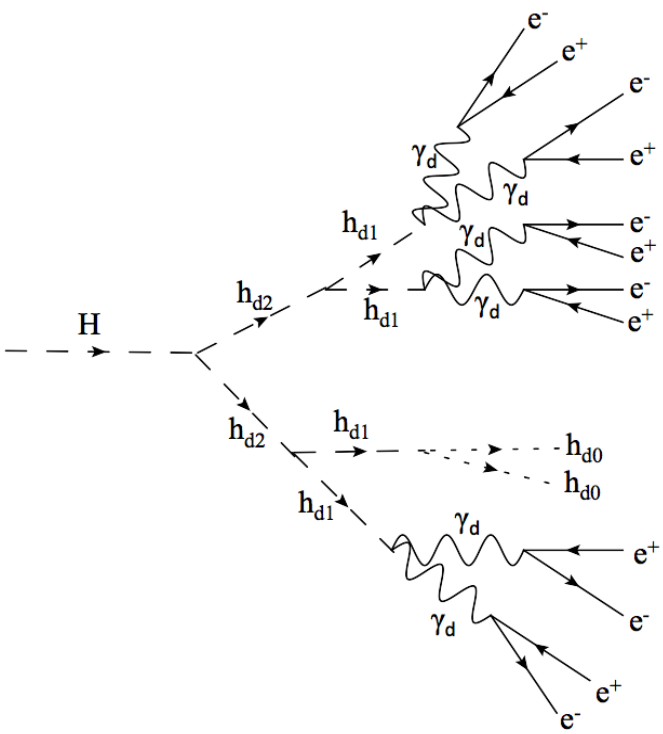

Figure 4.9: Diagram showing one possibility of a complete decay chain of a Higgs through the dark sector. The lightest MSSM Higgs (H) decays to lighter scalars, h_{d2}, which decay to dark sector Higgs, h_{d1}. The h_{d1} particles can decay to stable dark sector particles, h_{d0}, or dark photons, γ_d, which subsequently decay into a pair of electrons, e.

Chapter 5

Experimental Considerations

5.1 LHC

5.1.1 Introduction

The Large Hadron Collider (LHC) is currently the highest energy hadron collider in operation. It is designed to accelerate and collide protons at a centre of mass energy of $\sqrt{s} = 14$ Tera electron volts (TeV) with a high frequency of 40 million collisions per second. This will be achieved by colliding two counter-rotating beams containing 2808 bunches with around 1.1×10^{11} protons per bunch. This leads to a machine luminosity of L = 10^{34} cm^{-2}s^{-1}. L is defined as

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \varepsilon_n \beta *} F, \tag{5.1}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ε_n the normalised transverse beam emittance, $\beta*$ the beta function at the collision point, and F the geometric luminosity reduction factor due to the crossing angle at the interaction point (IP) [75]. The LHC is also used to accelerate heavy ions to 1.38 TeV per nucleon before colliding to produce a luminosity of L = 10^{27} cm^{-2}s^{-1}.

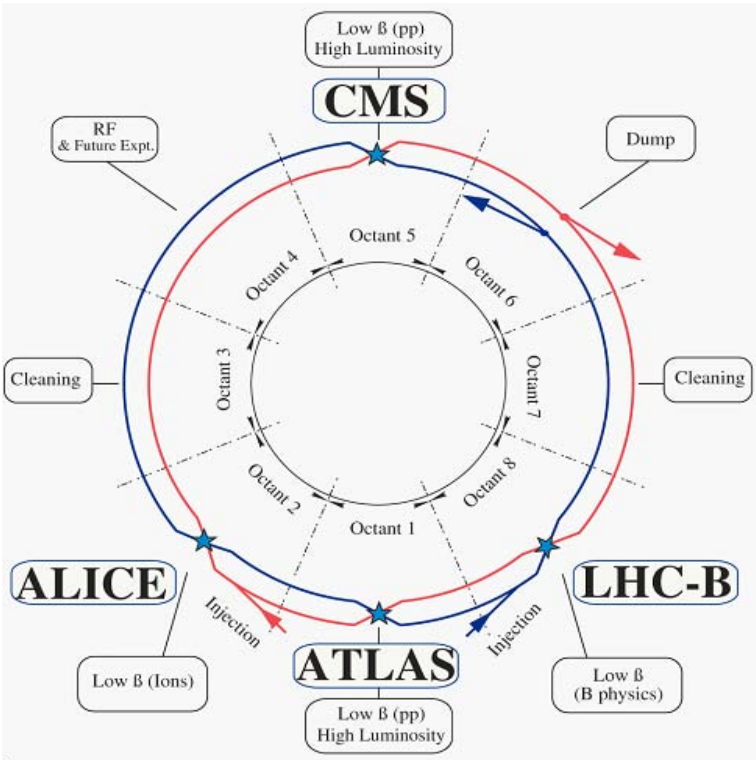

Figure 5.1: Diagram showing the layout of the LHC. The experiments are located at interaction points relating to the octant marked on the diagram [75].

The tunnel which houses the LHC at CERN was originally constructed between 1984 and 1989 for its predecessor, the Large Electron Positron collider (LEP). The tunnel is 26.7 km in circumference sprawling out underground between Switzerland and France. Due to the type of rock beneath the Jura Mountains the tunnel is inclined at an angle of 1.4% and lies 45 to 170 metres beneath the surface with the East side closer to the surface than the West side.

The LHC consists of eight straight sections and eight arcs. These straight sections are around 528 m long and each contains experimental halls which can serve as either detectors or service utilities. Four of those halls are occupied by the four main experiments; A Toroidal LHC ApparatuS (ATLAS), A Large Ion Collider Experiment (ALICE), Compact Muon Solenoid (CMS) and LHC beauty (LHCb) all of which operate at the LHC (experiments are indicated on Figure 5.1). Figure 5.1 shows the layout of the LHC and the positions of the four main experiments. The beams of hadrons are inserted into the LHC at points 2 and 8, whilst at points 3 and 7 collimators are used to clean the beams. Point 4 contains an Radio

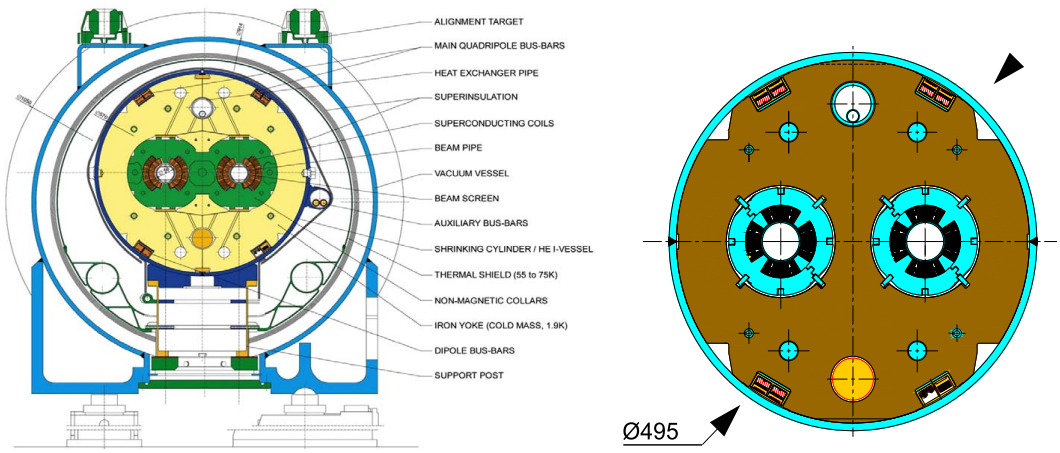

Figure 5.2: Cross sections of the LHC dipole (LHS) and a quadrupole magnet (RHS) where the two beam apertures are separated by 194 mm [75].

Frequency (RF) system for each beam and point 6 the beam dump insertion.

Quadrupole magnets provide the focusing and defocusing of the beam, whilst dipole magnets are used to steer the beam around the arc sections and to alter the separation between the beams.

A luminosity of 10^{34} cm^{-2}s^{-1} requires high beam intensities which therefore excludes the use of anti-protons. Instead, the LHC is designed to accelerate counter-rotating beams of protons which collide at specific interaction regions (IR) around the ring. The two counter-rotating beams occupy separate magnetic fields and vacuums in the arcs and only share a common beam pipe in the IRs where the experimental detectors are located.

Due to space limitations in the tunnel, the LHC uses twin bore magnets with separate coils and beam channels housed in one mechanical structure and cryostat. Consequently the magnetic flux circulates in opposite directions for the two beams. A cross section of these magnets is shown in Figure 5.2. The magnets use superconducting technology via NbTi Rutherford cables in order to reach a peak dipole field of 8.33 T for the 7 TeV beams. In order to reach this field strength, beam losses must be kept to a minimum and the temperature inside the cryostat kept constant at 1.9 K by using superfluid helium.

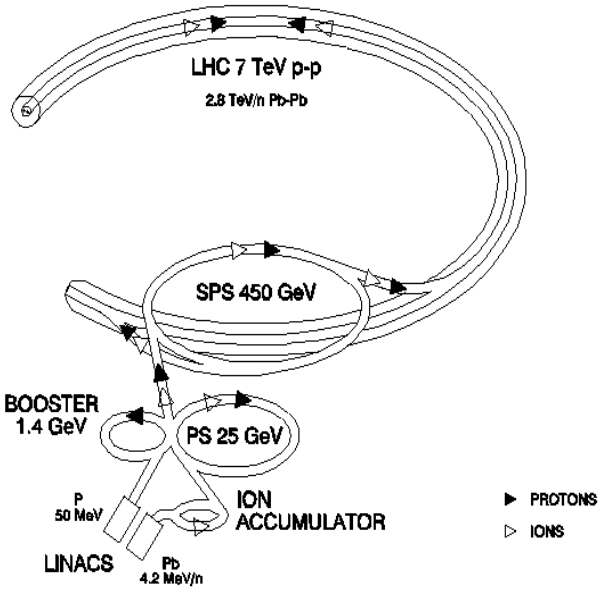

Figure 5.3: Diagram showing the injection chain used to fill the LHC with protons and lead ions [76].

5.1.2 Source and Injection Chain

The source of protons for the LHC comes from a hydrogen tank which feeds a duoplasmatron source before entering a linear accelerator, Linac2. Linac2 accelerates the protons from 140 keV to around 50 MeV. The protons then enter the Proton Synchrotron Booster (PSB) which accelerates them to 1.4 GeV. Next they enter the Proton Synchrotron (PS) which accelerates them to 25 GeV. Finally, before entering the LHC, the protons leave the PS and enter the Super Proton Synchrotron (SPS) which accelerates them to 450 GeV. An overview of this injection chain is shown in Figure 5.3.

5.1.3 Acceleration and Storage

Once beams have been injected into the LHC they are increased in energy to 7 TeV per beam. This is achieved through RF systems which capture, accelerate and store the beams using a 400 MHz superconducting cavity system located at point 6. At nominal operation the RF systems increase the energy of the protons from 450 GeV at injection from the SPS to 7 TeV with an energy gain per turn

of 485 keV. At 7 TeV the synchrotron radiation loss per turn is expected to be approximately 7 keV. The bunch length and emittance of the proton bunches is dictated by the luminosity requirements of the experiments. With the increase in beam separation provided by the separation dipoles, different RF cavities can operate on the two beams providing independent control.

The luminosity of the machine decreases during a collision run as the intensities and emittance of the beams degrade. This is mostly due to the loss of protons from collisions.

5.1.4 Dumping

The LHC stores more than 1 GJ of energy during a run which needs to be safely absorbed at the end of a scheduled run or in an emergency. A beam dumping system is installed which removes the beams from the LHC at interaction region IR6. The two beams are kicked out horizontally and deflected to carbon absorbers wrapped in steel and surrounded by radiation shielding in a separate tunnel 750 m away from the main ring. This dumping system puts additional limits on the maximum beam energies and intensities achievable by the machine.

5.1.5 Performance

The machine began operation in 2008 but, following an incident involving one of the superconducting magnets, it was put on an operation hiatus. On the 27th February 2010 proton beams were reintroduced after successfully fixing the machine and stable collisions at 3.5 TeV took place on 30th March 2010 [77]. Table 5.1 shows some of the milestones achieved by the LHC from the restart in 2010 to 2012.

Between 2010 and 2011 the LHC collided protons with a centre of mass energy of $\sqrt{s} = 7$ TeV due to concerns about the machine running at higher energies. These runs also used larger bunch separations of 50 ns instead of the designed 25 ns. For 2012 the machine successfully increased the centre of mass energy to

Date	Bunches per beam	Colliding Bunches	Luminosity ($cm^{-2}s^{-1}$)
29/08/10	50	35	1.00×10^{31}
25/10/10	368	348	2.07×10^{32}
22/03/11	200	194	2.50×10^{32}
23/05/11	912	874	1.10×10^{33}
26/10/11	1380	1331	3.65×10^{33}
06/03/12	264	194	9.00×10^{32}
06/06/12	1380	1377	6.76×10^{33}

Table 5.1: A few of the milestones in the LHC running history. Colliding bunches refers to those in IP5 and IP8, and Luminosity is the instantaneous luminosity. The machine starts colliding protons in March and ends in September before switching to heavy ion runs which are not recorded here.

$\sqrt{s} = 8$ TeV. The machine was run with the proton collision program until the end of 2012 before a heavy ion in early 2013. The LHC then went into a long shutdown (LS1) from mid February 2013 lasting until approximately the end of 2014 to perform upgrades and corrections needed to reach higher energy. It is predicted the machine will then run collisions with a centre of mass energy of approximately $\sqrt{s} = 13$ TeV.

The substantial number of collisions between 2010 and 2011 gave integrated luminosities equal to 0.04 fb^{-1} and 6.10 fb^{-1} respectively. Figure 5.4 shows the luminosity as measured by one of the two general purpose experiments, CMS.

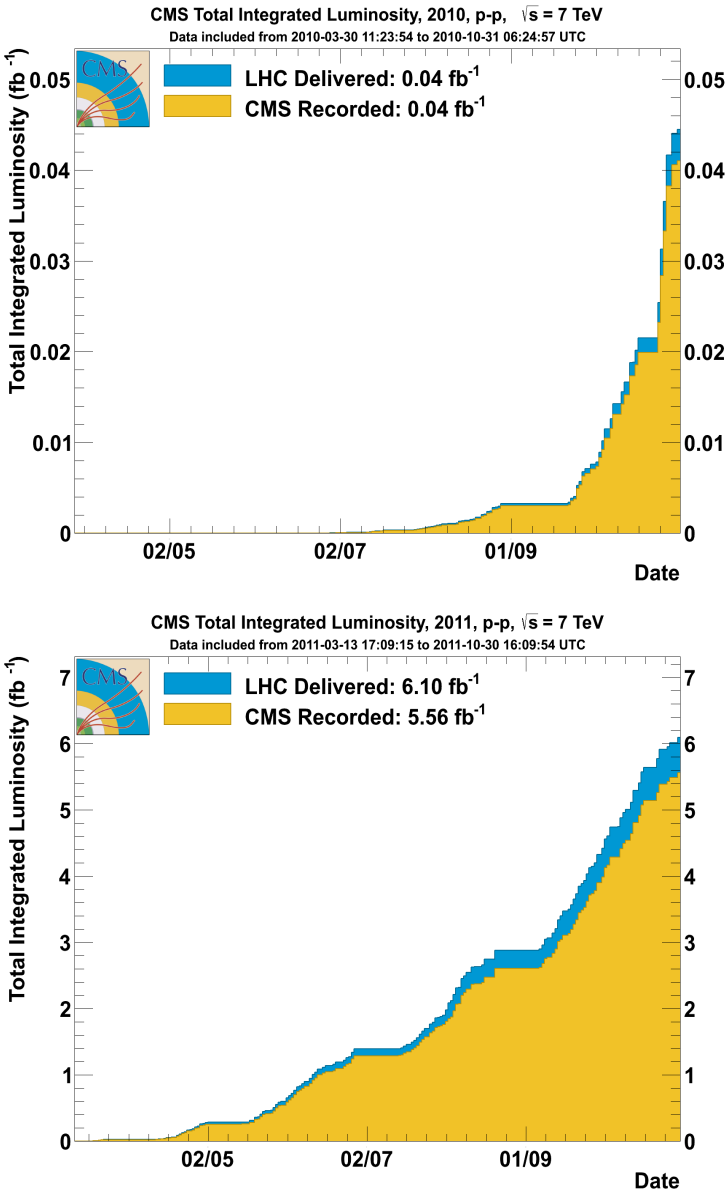

Figure 5.4: Plots showing the total integrated luminosity delivered by the LHC and recorded by CMS in 2010 and 2011 [78].

5.2 CMS

The Compact Muon Solenoid experiment (CMS) is one of two general purpose detectors at the LHC. The primary design goal of CMS is to measure the properties of highly energetic proton-proton and ion-ion collisions[1] delivered by the LHC. The layout of the detector follows a traditional configuration of concentric detectors centered upon the interaction point where the hadrons collide. The detectors at the centre of the machine focus on measuring tracks. These are then encompassed by calorimeters which measure the amount of energy deposited by particles. Surrounding both the tracking detectors and calorimeters lies a superconducting solenoidal magnet. Beyond the solenoid are the muon detectors interspersed with the iron return yoke. A schematic of the detector is shown in Figure 5.5.

Unlike ATLAS, CMS uses one large solenoidal magnet which is used to identify charged particles both within the tracking section of the detector as well as at the

[1] At the time of writing this document there are plans to also run with ion-proton collisions.

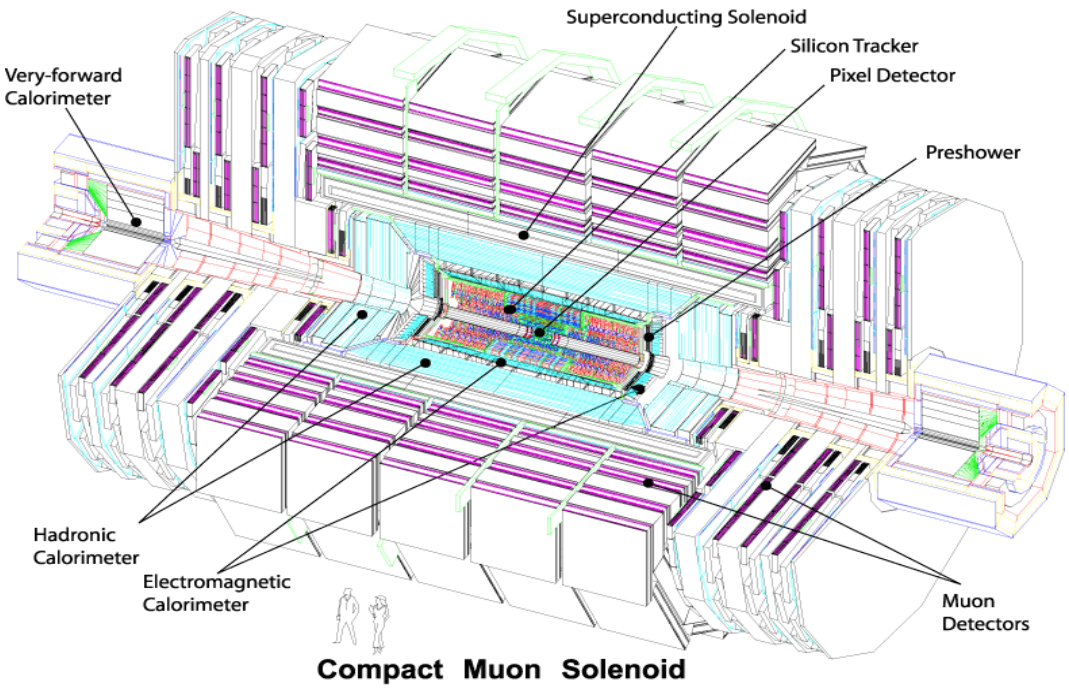

Figure 5.5: A schematic overview of the CMS experiment displaying the different subdetectors.

outer muon detectors. This setup dictates the design of the rest of the detector.

As described in Section 5.1, the LHC nominally aims to deliver proton collisions with a centre of mass energy of 14 TeV and luminosity of 10^{34} cm^{-2}s^{-1}. This places substantial requirements on the detector. Good hermicity of the detector is needed to ensure that all particles with high transverse momentum are detected and infer missing momentum from particles which minimally interact with matter. The detector components are required to have a fast response in order to separate the different collisions. Fast readouts from the components are needed to allow measurements to be taken, whilst the detectors and electronics need to be radiation hard in order to survive the harsh environment provided by the LHC.

In the following sections, details of the different components which constitute the detector are described. Special emphasis is placed on the sections of the detector which were pivotal for the analysis described in Chapter 6. The need for triggers and subsequent data storage is described in Sections 5.2.6 and 5.2.7 respectively.

5.2.1 Silicon Tracker System

The aims of the tracking system are to provide precise and efficient measurements of the tracks from the interaction point. In addition it should reconstruct secondary vertices within the densely populated environments the LHC provides. A detector technology featuring high granularity and a fast response is required in order to effectively identify and assign tracks to the correct bunch crossing [79]. However, to obtain these requirements, a large amount of on-board electronics is needed which subsequently necessitates substantial cooling.

The total tracking system occupies a length of 5.8 m and a diameter of 2.5 m. It is comprised of a Silicon Pixel Detector surrounded by the Silicon Layer Tracker. Silicon detectors were chosen for their spatial precision, granularity, reliability and also the ability to survive the severe radiation damage from the LHC over the expected lifetime of around 10 years. Figure 5.6 shows a schematic of the tracking configuration and Figure 5.7 are plots showing the different contributions to the tracker material.

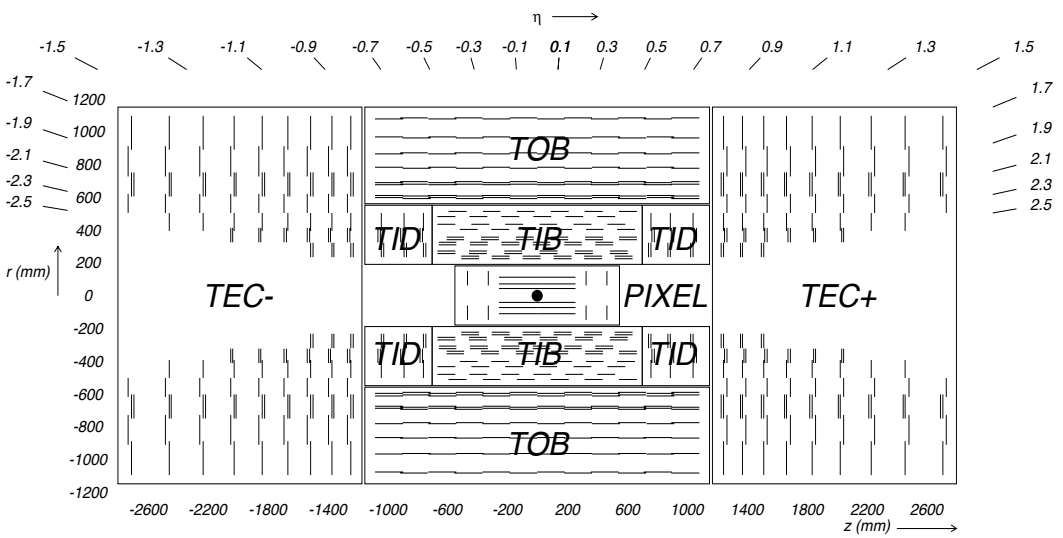

Figure 5.6: A schematic overview of the tracking detectors [79].

At nominal design luminosity each bunch crossing has ∼1000 particles hitting the tracker. At a radius of 4 cm this equates to a hit rate density of 1 MHz/mm^2; at 22 cm the rate drops to 60 kHz/mm^2 and at 115 cm the rate drops further to 3 kHz/mm^2. In order to reduce the occupancy to below 1% at the inner radii (< 10 cm) a pixelated silicon detector was used. Each pixel has a surface area of 100 × 150 μm^2 in r-ϕ and z respectively which leads to an occupancy in the order of 10^{-4} per pixel per bunch crossing. The Silicon Pixel Detector comprises three cylindrical layers with radii of 4.4, 7.3 and 10.2 cm. At each end are two disks located at a distance of 34.5 and 46.5 cm from the interaction point respectively.

Beyond the pixel detectors in the radial region, between 20 cm and 116 cm, lie the silicon micro-strip detectors. The silicon strip detector is split into three regions, the Tracker Inner Barrel and Disks (TIB/TID), which is radially surrounded by the Tracker Outer Barrel (TOB) with the Tracker EndCaps (TEC) extending the region in the z direction.

The TIB is composed of four cylindrical layers at radii of 255.0, 339.0, 418.5 and 498.0 mm and length of 1400 mm. The two inner layers of TIB use double-sided modules while the two outer layers use single sided modules. The TID is constructed of three disks at each end of the TIB between $800 < |z| < 900$ mm and span the radius of $200 < r < 500$ mm. Each disk contains three rings of strips, where the two innermost use double sided modules and the last uses single

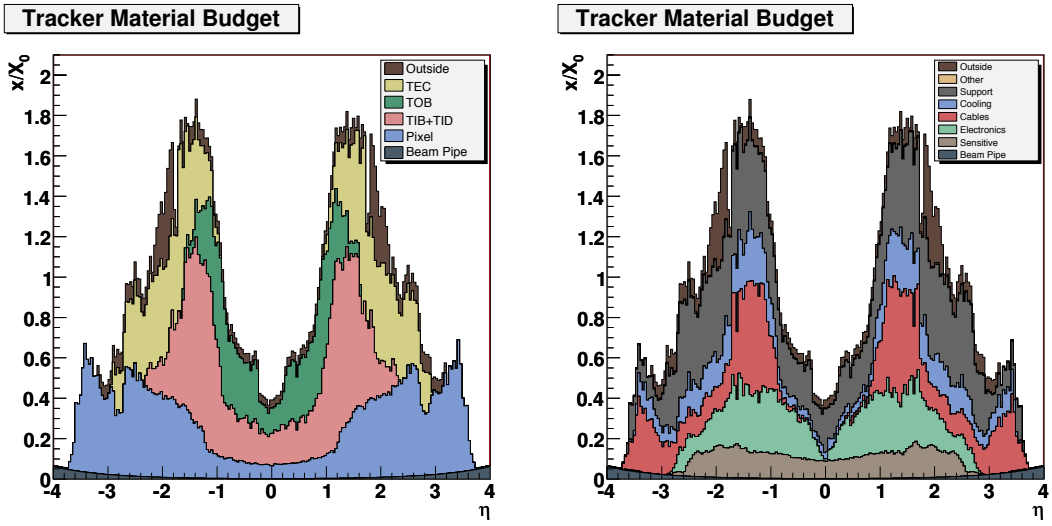

Figure 5.7: Plots showing the different contributions to the tracker material budget, divided into the different subdetectors (LHS) and different components (RHS) [79].

sided modules. The TIB/TID provides up to 4 r-ϕ measurements of a track and guarantees pseudorapidity coverage up to $\eta = 2.5$.

Surrounding the TIB/TID in the barrel is the TOB which spans the radius of 555 $< r <$ 1160 mm and has a length of 2180 mm (2360 mm) without (with) cabling. There are six detection layers at radii of 608, 692, 780, 868, 965 and 1080 mm. The two inner layers have double sided modules with the four outer using single sided modules.

The two Tracker EndCaps, TEC$\pm$, lie between $1240 < |z| < 2800$ mm and are constructed of nine disks with detector modules mounted. The inner three disks extend radially between $229 < r < 1135$ mm while the outer six extend between $309 < r < 1135$ mm to allow access for the insertion of the pixel detector. The detector modules are arranged in rings and mounted onto substructures called petals which are in turn mounted onto the disks. Disks 1 to 3 have seven rings of modules, disks 4 to 6 have six rings, disks 7 and 8 have five rings and disk 9 has four rings, as can be seen in Figure 5.6. Rings 1, 2 and 5 are constructed of double sided modules while all others are single sided.

Due to the high granularity of the tracker, in conjunction with the strong magnetic field, as described in Section 5.2.4, a momentum resolution of 1.5% can be

achieved on promptly produced charged particles with tranverse momentum p_T = 100 GeV. The hit resolution in the barrel sensors has been measured to be 10.4 μm for the transverse coordinate with high momentum particles. The resolution in the longitudinal coordinate varies as a function of the track angle with values between 20 and 45 μm [80].

The hit efficiency is the probability of finding hits in a given silicon sensor that has been traversed by a charged particle. In the pixel detector the average hit efficiency has been measured to be over 99% using particles with a transverse momentum greater than 1 GeV and tracks reconstructed with a minimum of 11 hits in the strip detector. Hits from the pixel layer under study are not removed during track reconstruction. To avoid this biasing the results, the tracks are required to have hits in the other two pixel layers. The efficiency is calculated from the fraction of traversing tracks for which either a hit is used in the reconstruction or a hit is found within 500 μm of the predicted track position [80].

The hit efficiency in the strip tracker is measured to be 99.8% using tracks which have a minimum of 8 hits. The efficiency is calculated from the fraction of traversing tracks for which a hit is found anywhere within the region of a traversed module. Defective modules are excluded from these measurements corresponding to 2.4% and 2.3% of the pixel and strip detectors respectively [80].

The tracking efficiency can be measured using a tag and probe method with muons. Z to muon candidates are reconstructed using pairs of oppositely charged muons identified using the muon chambers with an invariant mass between 50 and 130 GeV. The tag muon is reconstructed in both the tracker and muon chambers and the probe reconstructed in the muon chambers with no requirements on the tracker. The tracking efficiency can then be estimated from the fraction of probe muons with associated reconstructed track in the tracker (Figure 5.8).

5.2.2 Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) is a homogeneous absorber/detector constructed from 61,200 lead tungstate ($PbWO_4$) crystals in the barrel with 7324 crystals in the endcaps. One of the primary motivations in the design of the

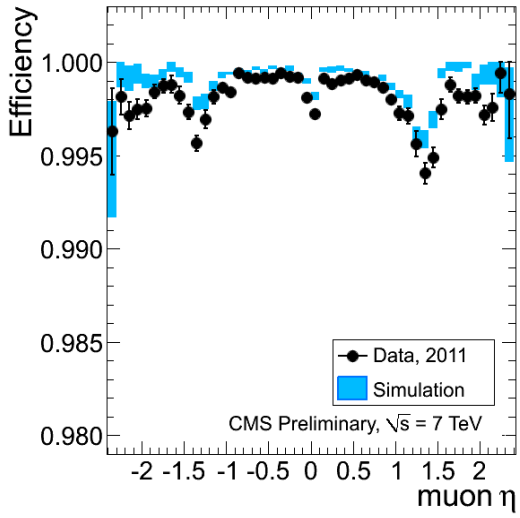

Figure 5.8: The tracking efficiency as a function of pseudorapidity using tag and probe muons from data taken in 2011 and simulations. The tag muon is reconstructed in both the tracker and muon chambers and the probe is reconstructed in the muon chambers with no requirements on the tracker [80].

ECAL was to detect the photons in the Higgs decaying to two photons channel. The calorimeter needs to be fast, radiation hard and have fine granularity in order to be useful under LHC conditions. $PbWO_4$ crystals were chosen due to their short radiation length of 0.89 cm and small Molière radius of 2.2 cm resulting in a high granularity, radiation hard scintillator. In order to measure the scintillation light emitted from the crystals as charged particles traverse the medium, photodetectors are attached to the ends of the crystals. Figure 5.9 shows the layout of the CMS ECAL.

The cylindrical barrel part of the ECAL (EB) covers the pseudorapidity region $|\eta| < 1.479$. The barrel contains 360 crystals in ϕ by 85 crystals in both positive and negative η resulting in the previously mentioned total of 61,200 crystals. The crystal is pointed at a slight angle of 3° in both η and ϕ away from the nominal interaction point. This minimises cracks apparent to the particles produced from collisions. The crystals are also tapered with a surface area 22×22 mm^2 at the front and 26×26 mm^2 at the rear. The crystals have a length of 230 mm which corresponds to 25.8 radiation lengths. Avalanche photodiodes are used as the photodetecters in the barrel section. The crystals are grouped by 5×5 crystals into modules. There are four supermodules for both positive and negative η,

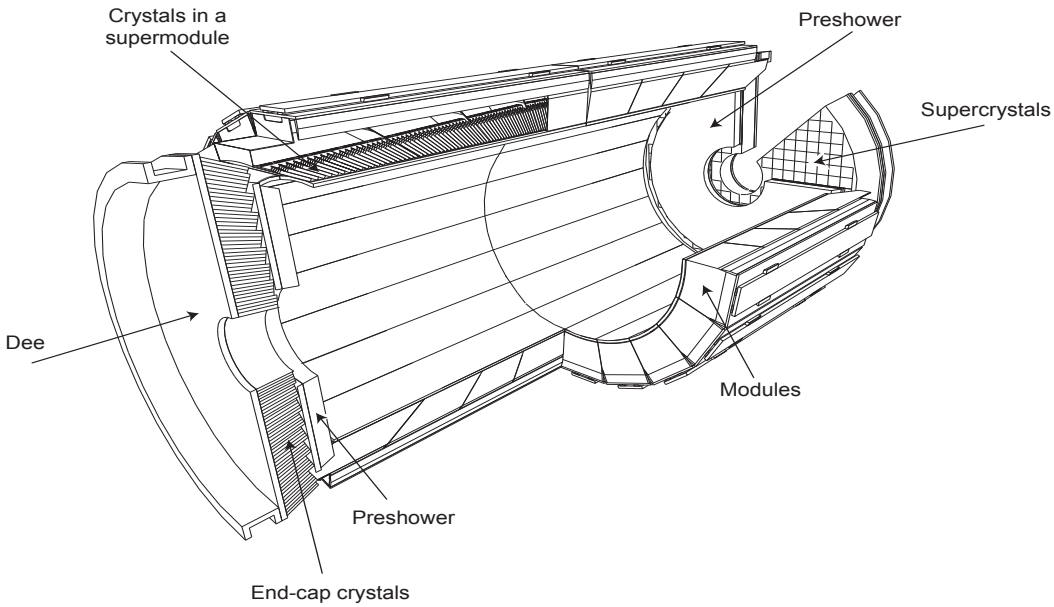

Figure 5.9: Layout of the CMS ECAL [79].

where each supermodule contains four modules except for the supermodule closest to $\eta = 0$ which contains five modules. The summed energy from each module is used for triggering, referred to as an ECAL trigger tower.

The endcaps of the ECAL (EE) cover the pseudorapidity region $1.479 < |\eta| < 3.0$. The EE is constructed of two halves called Dees. In front of the endcap crystals, within the range $1.653 < |\eta| < 2.6$, lies a preshower detector, which is primarily used to identify neutral pions. The crystals used in the EE are of a different specification to those in the EB, with greater surface area and shorter length. The front surface area is 28.62×28.62 mm^2 and 30×30 mm^2 at the rear with a length of 220 mm which corresponds to 24.7 radiation lengths. Vacuum phototriodes are glued onto the ends of the crystals for use as the photodetectors in the EE.

The optical properties of the crystals change over time in the harsh conditions present at the LHC. This is due to ionizing radiation interacting with impurities in the crystals. To compensate for these effects, a laser correction system is installed at the ECAL. Laser light is injected at the front of the crystals and the response by the photo diodes measured. Figure 5.10 shows the effect of applying transparency corrections to the ratio of an electron's energy as measured in ECAL

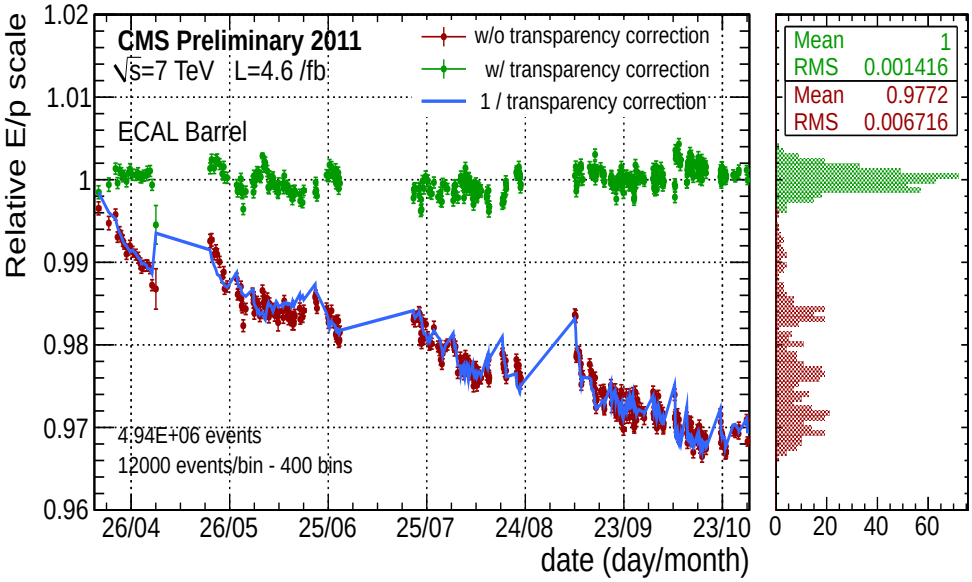

Figure 5.10: The effect of laser corrections on ratio of the electron energy measured in ECAL to the electron momentum measured in the tracker. The green points show the ratio with corrections and the red points show the ratio without corrections applied [81].

versus the momentum of the electron as measured in the tracker.

The scintillation properties of the crystals are temperature dependent with less light emitted as the temperature increases. Therefore, the crystals and photodetectors must be kept at a constant temperature of 18 ± 0.05°C. Heat from the readout electronics must be efficiently extracted. This is achieved by using a water flow cooling system with the ECAL.

The energy resolution of the ECAL can be parameterised as

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2, \tag{5.2}$$

where S is the stochastic term, N the noise term, and C the constant term. During the 2004 test beam runs the typical energy resolution as reconstructed by summing 3 × 3 crystals was measured to be:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{0.12}{E}\right)^2 + (0.30\%)^2. \tag{5.3}$$

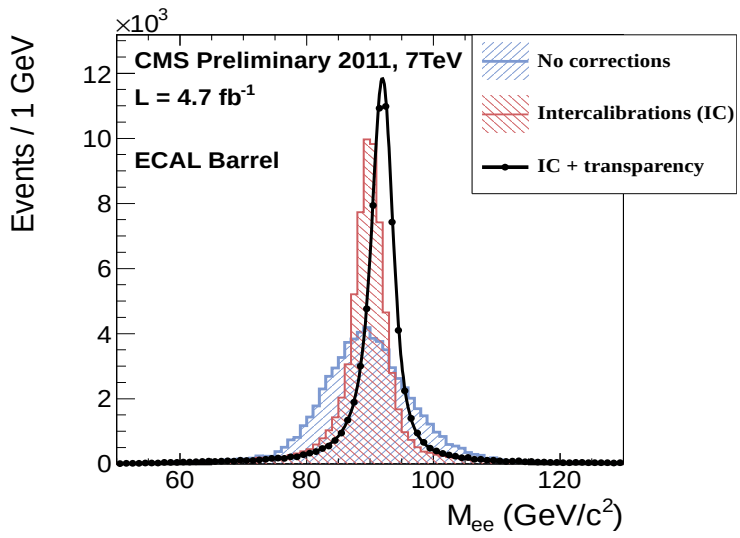

Figure 5.11: The di-electron invariant mass spectrum using electrons decaying from Z bosons in data taken in 2011. The plot shows the improvements in energy scale and resolution after applying energy scale corrections to account for the intrinsic spread in crystal and photo-detector response, and time-dependent corrections to compensate for crystal transparency loss [81].

Approximately 94% of the incident energy of an electron or photon is contained in 3×3 crystals.

The overall energy scale of the calorimeter is calibrated using decays of Z bosons to electrons from data (Figure 5.11). The instrumental resolution after preliminary energy calibration of 2011 data is obtained from the invariant mass of electrons produced from the decay of Z bosons, through ECAL energies and electron track directions. It is measured to be 1.0 GeV in the ECAL Barrel as calculated from a fit to the Z resonance in Figure 5.11 [81]. This measurement is from the width of a Crystal Ball function [82] convoluted to the Z→ee Briet-Wigner shape. Figure 5.12 shows the energy resolution as a function of pseudorapidity for Z→ee decays from data taken in 2011 and simulation, where the electron energy resolution is derived by a crystal ball function [83].

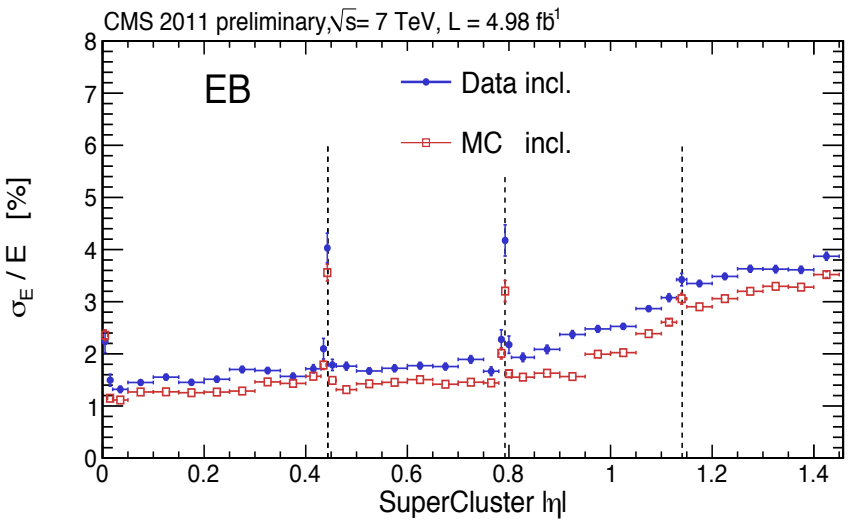

Figure 5.12: The energy resolution, σ_E/E, for di-electrons decaying from Z bosons in data taken in 2011 and simulation, unfolded in bins of pseudorapidity. The dotted lines represent the boundaries between supermodules. The relative energy resolution worsens as a function of η due to the tracker material in front of the ECAL as shown in Figure 5.7. The difference between data and MC may be due to an underestimate of the number of parasitic collisions (pile up) [83].

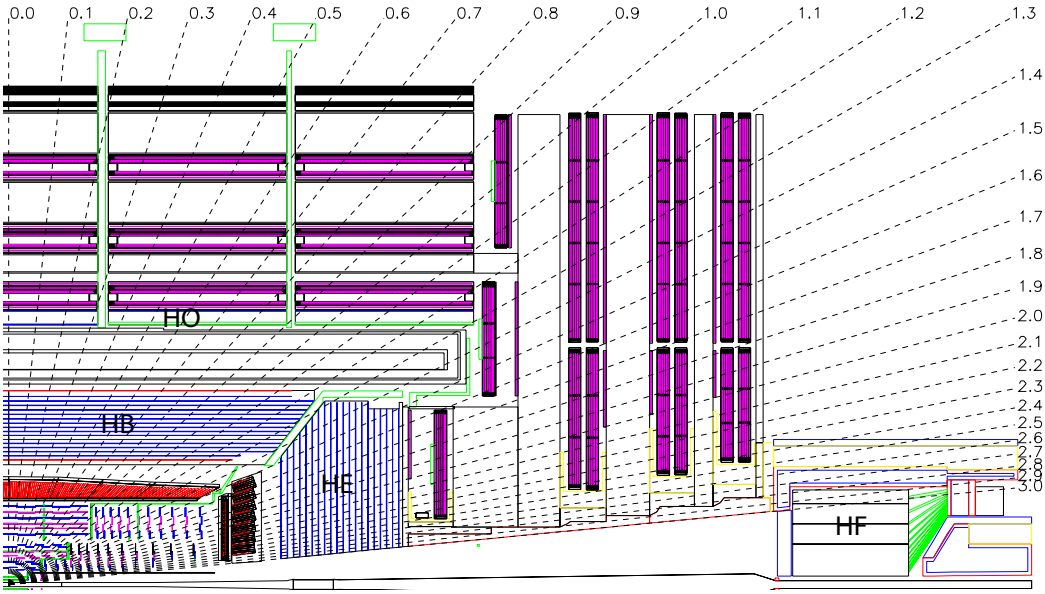

Figure 5.13: Schematic showing the layout of the CMS HCAL. Dotted lines with attached numbers represent values of pseudorapidity [79].

5.2.3 Hadron Calorimeter

The Hadron Calorimeter (HCAL) is comprised of four subdetectors; the hadron barrel (HB), hadron endcap (HE), hadron outer (HO) and hadron forward (HF). Figure 5.13 shows the layout of the various subdetectors within the CMS HCAL.

Brass was chosen as the absorber material as it is non-magnetic, has sufficient interaction lengths to contain hadronic showers, good mechanical properties and is relatively inexpensive.

Between the ECAL and the magnet lies the HB, a sampling calorimeter using brass absorbers and plastic scintillators which covers the pseudorapidity region $|\eta| < 1.3$. The HB consists of 2×18 identical wedges which corresponds to two half barrels, HB$\pm$. Each wedge is further split into four sectors in ϕ and aligned parallel to the beam axis. The wedges use brass absorber plates, except for the innermost and outermost plates which are made of stainless steel for structural support. The first steel plate is 40 mm thick, followed by 8×50.5 mm thick brass plates, then 6×56.5 mm brass plates with the final steel plate of thickness 75 mm. At $\eta = 0$ the total absorbed thickness is 5.82 interaction lengths, which

rises as η increases to 10.3 interaction lengths at $\eta = 1.3$. The EB adds around 1.1 interaction lengths worth of material.

The scintillator is constructed from plastic for long-term stability and moderate radiation hardness [79]. In total there are 17 layers of scintillator in the HB wedges. The first is 9 mm thick and positioned in front of the first steel support plate to sample hadronic showers developed in the inert material between the ECAL and steel. The next 15 layers are 3.7 mm thick interspersed between the brass absorbers. The final layer is 9 mm thick in order to correct for late developing showers leaking out from the back of the HB. Wavelength shifting fibres collect scintillation light from the plastic.

The HE covers the pseudorapidity region $1.3 < |\eta| < 3.0$ and is attached to the muon endcap yoke. The brass plates used in the HE are 79 mm thick with 9 mm gaps for the plastic scintillator. The total length of the endcap calorimeter, including the EE, is 10 interaction lengths. The granularity of the calorimeters is $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for $|\eta| < 1.6$ and $\Delta\eta \times \Delta\phi \approx 0.17 \times 0.17$ for $|\eta| \geq 1.6$. Multipixel hybrid photodiodes are used as photodetecters for the scintillation light due to their low sensitivity to strong magnetic fields and large dynamical range.

In the barrel region of the detector the EB and HB do not provide sufficient material to contain all the hadronic showers. Therefore, the HCAL is extended beyond the solenoid by the HO which can identify late showers and measure shower energies deposited after the HB. The solenoid provides approximately $1.4/\sin\,\theta$ interaction lengths, therefore at $\eta = 0$ the HB provides the minimal amount of interaction length. Consequently two layers of scintillators are positioned either side of 19.5 cm thick iron, from part of the iron return yoke, at radii 3.82 and 4.07 m. Beyond $|z| > 1.268$ m only one layer of scintillator is needed at the radial distance of 4.07 m as the HB provides greater absorber depth.

The HF lies in the very forward regions of $|\eta| \geq 1.6$ at 11.2 m away from the nominal iteraction point. In this region the amount of particle flux and energy deposited is much higher than any other part of the detector. On average 760 GeV per proton-proton collision is deposited in the two forward calorimeters compared to 100 GeV in other parts of the detector [79]. For this reason the HF needs to be

constructed of materials which can survive these harsh conditions. Quartz fibres were chosen as the active medium, where Cherenkov light is generated from the shower of charged particles above a particular threshold. These fibres are inserted in the grooves of a steel absorber. Real time measurements from the HF, as well as from a dedicated instrument called the Pixel Luminosity Telescope, can provide tools for monitoring the luminosity on a bunch by bunch basis.

5.2.4 Solenoidal Magnet

Beyond the tracking systems, EB and HB, lies the superconducting solenoidal magnet. The magnet is 6 m in diameter and has a length of 12.5 m producing a 4 T field. The magnetic flux is returned through the use of an 10,000 tonne iron yoke, between which the muon detectors are interspaced. The solenoid is a four layer winding made from stabilised reinforced NbTi Rutherford-type conducting cables. The solenoid has greater values for the stored energy (2.6 GJ) and energy over mass (11.6 kJ/kg) than any previous detector magnet. The structure therefore needs to be sufficiently strong to cope with stresses induced by energising the magnet. The coil is kept at a temperature of 4.6 K through the use of liquid helium.

During 2011 and 2012 runs the magnet was producing a magnetic induction of around 3.8 T.

5.2.5 Muon Chambers

The muon system has three main functions; muon identification, momentum measurement and triggering. Good momentum resolutions, as well as the triggering capability, are possible due to the large magnetic field provided by the solenoid and return yoke. The yoke also acts as an absorber for hadrons providing a more efficient muon identification. Three different types of gaseous detectors were used for the muon identification. The layout of the muon system is shown in Figure 5.14.

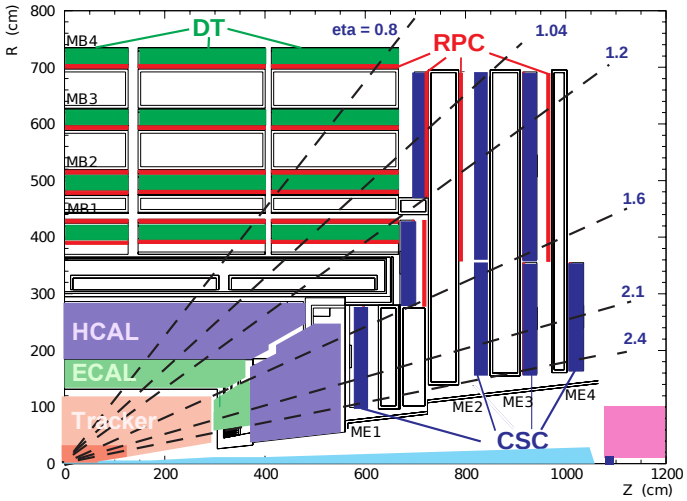

Figure 5.14: Longitudinal layout of the muon system in one quadrant of the CMS detector [84].

In the barrel region of $|\eta| < 1.2$ the low muon flux and uniform magnetic field allows the use of drift tube (DT) chambers. The DTs are arranged into 4 stations interspersed between layers of the iron return yoke. The first three stations contain 2×4 chambers which measure the muon position in the r-ϕ plane and 4 chambers measuring the position in the z direction. The last station does not contain chambers for measuring in the z direction. The drift cell in each chamber is layered, with respect to its neighbours, in order to reduce dead spots in the detector. Figure 5.15 shows the layout of the muon system in the barrel section of the detector.

In the endcaps pseudorapidity region $0.9 < |\eta| < 2.4$, where the muon rates are higher and the magnetic field non-uniform, cathode strip chambers (CSC) are used. In a similar fashion to the barrel region there are 4 CSC stations in the endcap interspersed with the return yoke. The cathode strips in the CSCs are aligned radially outwards providing a measurement in the r-ϕ plane. The anode wires in the CSCs are perpendicular to the strips providing measurements in η as well as the beam crossing time of the muon.

A third type of muon detector consists of resistive plate chambers (RPC). This is used to complement the DT and CSC triggering on the transverse momentum of muons. The RPCs are installed in both the barrel and endcap, covering a range

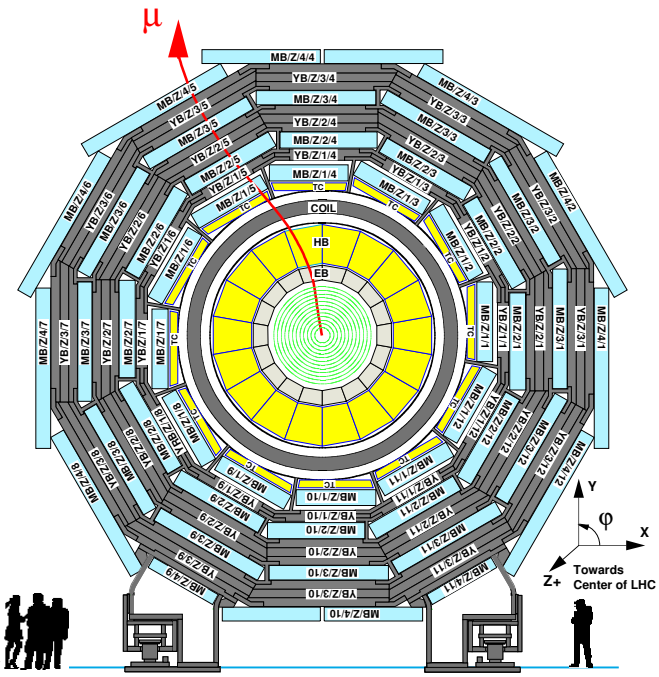

Figure 5.15: Layout of the CMS Muon system in the Barrel section [79].

of $|\eta| < 1.6$. Six layers of RPCs are installed in the barrel region with two in each of the first two muon stations and one in each of the last two stations. Three layers of RPCs are installed in the endcap, one for each of the first three stations. The RPC was designed to cover the pseudorapidity region $|\eta| < 2.1$, however not all of the chambers were installed for 2011 and 2012 running.

The momentum scale and resolution of muons are studied from cosmic-ray muons for muons with high momentum, $p_T \gtrsim 100$ GeV, and from J/ψ and Z resonances for muons with low and intermediate momentum, $p_T \lesssim 100$ GeV. Tracks from muon candidates found in the Muon chambers are matched to tracks found in the Silicon tracker as described in Section 6.6.2. For muons with a $p_T \lesssim 100$ GeV the resolution is dominated by that of the Silicon tracker.

For low momentum muons, $p_T \lesssim 10$ GeV, the muon relative transverse momentum resolution, $\sigma(p_T)/p_T$, was found to be between 0.8% and 3% depending on pseudorapidity. Fits to the Z resonance from data taken in 2010 and comparisons between data and simulations were used for intermediate-p_T muons. Corrections from comparisons between data and simulations to the Z resonance in the di-muon spectrum are shown in Figure 5.16. The relative transverse momentum

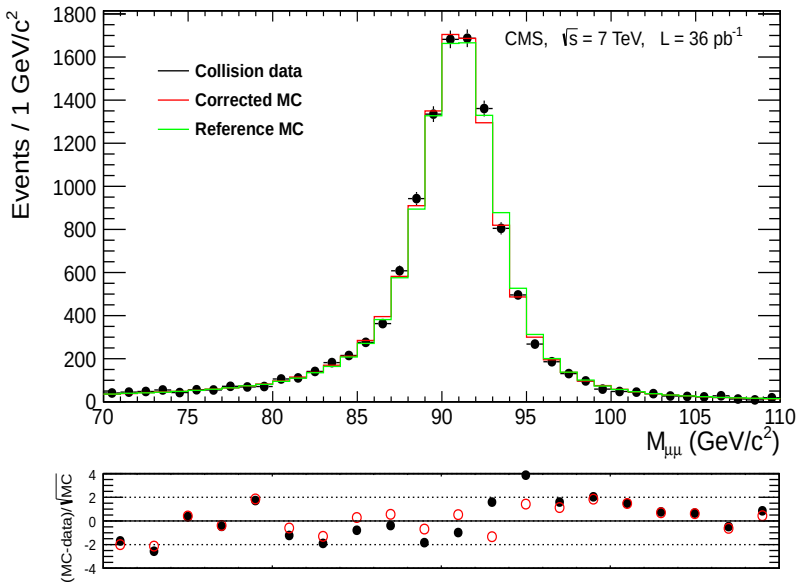

Figure 5.16: Top: the di-muon invariant mass spectrum for selected Z boson candidates decaying to muons from data taken in 2010 and simulation with and without corrections from SIDRA applied. SIDRA (SImulation DRiven Analysis) compares data and simulation of Z decaying to muons and allows modifications to the simulation through a scale shift and a worsening in resolution with respect to data. Bottom: the difference between simulation and data, divided by the expected statistical uncertainty without (black) and with (red) corrections [85].

resolution as a function of pseudorapidity from muons produced from Z decays in data taken in 2010 and simulations is shown in Figure 5.17. The relative p_T resolution was found to be in the range 1.3% to 2.0% for muons in the barrel and up to ∼6% for muons in the endcap [85].

5.2.6 Trigger

At nominal values the LHC will provide protons with a beam crossing interval of 25 ns corresponding to a frequency of 40 MHz. This rate is too great to record all events and in most cases the collisions do not produce anything of interest. CMS uses a two step triggering system called the Level 1 (L1) and the High Level Trigger (HLT) to reduce the number of events, and therefore the amount of data stored, to those of interest. The L1 consists of custom designed programmable

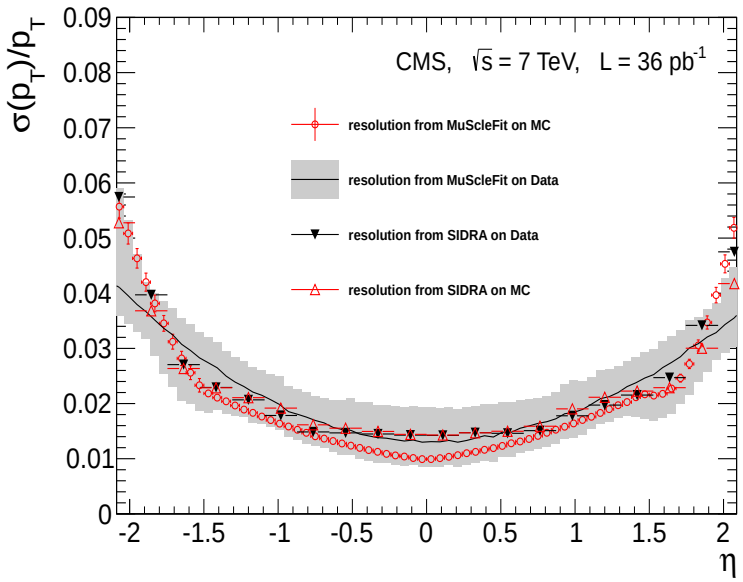

Figure 5.17: Relative transverse momentum resolution in data taken in 2010 and simulation measured by applying different methods (MuScleFit and SIDRA) to muons produced in the decay of Z bosons and passing selection, as a function of pseudorapidity. MuScleFit (Muon momentum and Scale calibration) is an absolute measurement of the momentum scale and resolution by using a reference model of the generated Z lineshape convoluted with a Gaussian function. The grey band represents the 1σ uncertainty on the measurement of MuScleFit on data [85].

electronics while the HLT is software based, using a farm of commercial processors.

The L1 reduces the rate of events from ∼40 MHz down to a limit of ∼100 kHz. During the 2011 and 2012 runs the L1 reduced the rate to $\lesssim$ 80 kHz. The L1 performs decisions using coarsely segmented data from the calorimeters and muon detectors. At the same time it keeps the more detailed high resolution data in memory pipelines on the front end electronics. For flexibility, the L1 hardware uses Field Programmable Gate Arrays (FPGA) wherever possible, which can be reprogrammed to gain improvements in performance. Where the electronics needs to be more radiation hard or faster, application-specific integrated circuits (ASIC) and programmable memory look up tables (LUT) are used. The L1 electronics are located either on the detectors themselves or in an underground control room 90 m from the experiment.

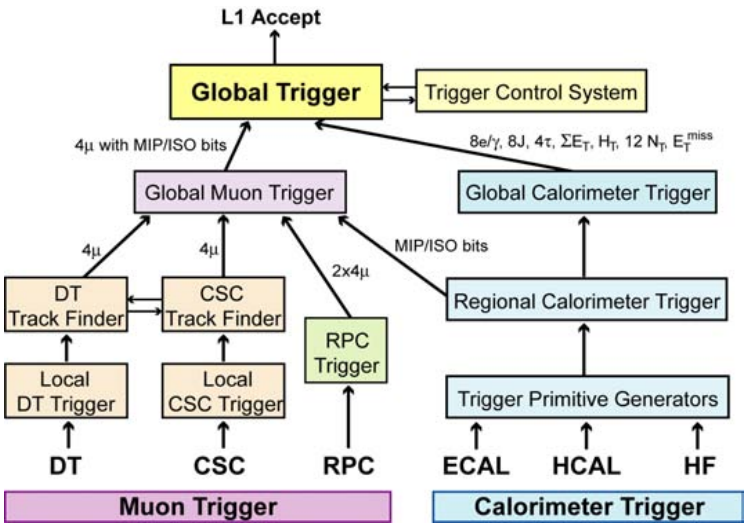

Figure 5.18: Overview of the L1 Trigger system [79].

The L1 has three components; local, regional and global. The local component, or Trigger Primitive Generators, fires off of energy deposits in calorimeters and tracks or hits in the muon chambers. These local components are then combined and ranked, in order of energy or momentum versus quality using pattern logic, into regional components. The Global Calorimeter and Global Muon Triggers then parse the highest ranked objects to the Global Trigger. The Global Trigger then passes or rejects the event based on running algorithms over the objects whilst simultaneously receiving information on the detector's readiness from the Trigger Control System. The trigger latency, as measured between the bunch crossing and distribution of the decision to the front end electronics, is 3.2 μs [79]. Figure 5.18 shows the overview of the L1 triggering system.

When the L1 passes an event, the CMS Data Acquisition system (DAQ) records data from all the subdetectors. An overview of the system is shown in Figure 5.19. The DAQ must be able to sustain an input rate of 100 kHz (equal to the limit of the L1 output), where each event is roughly 1 MB giving a total data flow of around 100 GB/s. The DAQ must also provide suitable computing power in the form of a computing farm on which the data quality monitoring (DQM) system and the HLT operates before data is transferred to a data storage centre on the CERN Meyrin site.

The HLT uses more complex software algorithms than the L1 in order to reduce

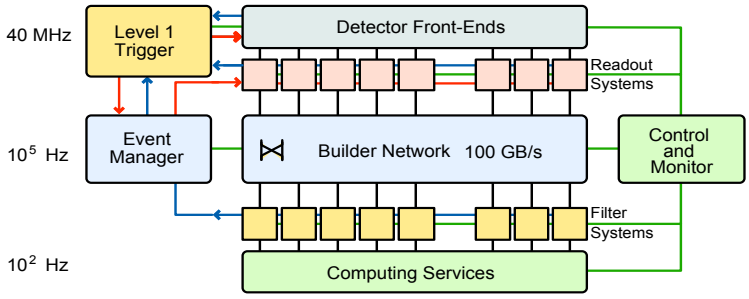

Figure 5.19: Overview of the DAQ system [79].

the rate from tens of thousands of events per second to a few hundred per second. These algorithms can be of a similar complexity to those used in offline analysis. The HLT uses a faster version of the offline reconstruction software in order to utilise more complicated physics objects whilst keeping the output rate high.

The HLT is modified throughout the running of the experiment. New triggers are constantly being developed and implemented which may trigger on new signatures or improve the performance of others. Old triggers can become deprecated and superseded, for example when replaced with a more efficient trigger that captures the same events. The prescales attached to each trigger are also varied appropriately in order to maintain bandwidth whilst keeping interesting events as physics priorities change.

5.2.7 Data Storage

A fully distributed computing model was designed in order to store and distribute the data from CMS and other LHC experiments. This system is based on grid middleware managed through the Worldwide LHC Computing Grid (WLCG). The computing system must be flexible enough to evolve over the lifetime of the experiments as priorities change. These systems must provide tools for locating, transferring and processing large collections of events. A multi-tiered computing system was developed as depicted in Figure 5.20

Using the grid, raw data from the detectors are reconstructed and stored into datasets containing physical objects of interest in a physics analysis. Additionally, simulated data are produced and distributed using the grid. The data are stored

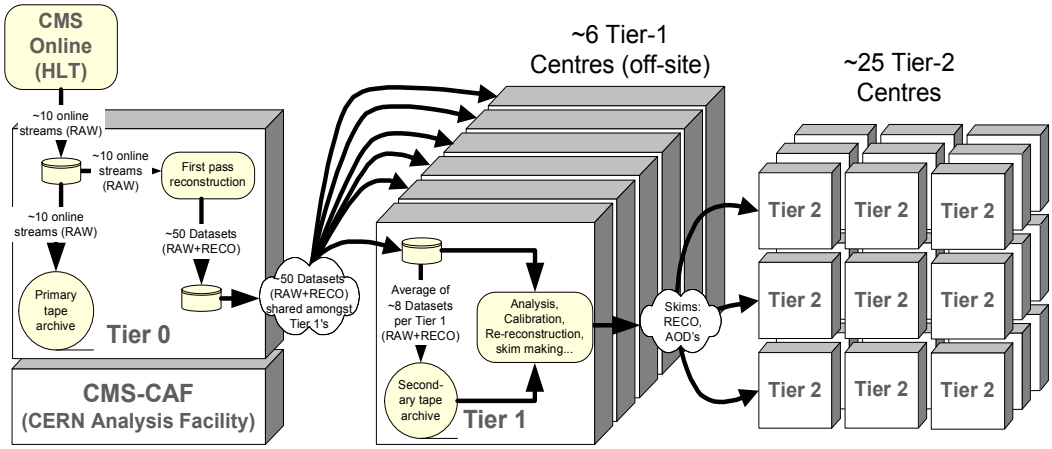

Figure 5.20: Overview of the CMS Computing system [79].

as ROOT files [7]. The collection of software used to simulate, calibrate, reconstruct and analyse the data is referred to as CMS Software (CMSSW). CMSSW is written using C++ classes with configuration files written in python. It is built around an Event Data Model (EDM) which uses the concept of an event; a C++ object which holds all the data related to a particular collision [86]. The data items in the event can be individually or collectively stored in ROOT files allowing for subsets of algorithms to be run without having to go through further processing steps.

Chapter 6

Associated Higgs Decaying to Electron Jets

6.1 Introduction

The motivation for searching for Lepton Jets was detailed in Chapter 4. Previous searches for signatures resulting from these models have been undertaken at ATLAS, CDF, CMS and DØ. The majority of these analyses focused on searching for clusters of leptons with particular requirements on the cluster size.

Using data recorded by the ATLAS detector, a search for a model producing a signature of two dark photons decaying to two or more groups of two or more muons was performed [87]. This analysis would be insensitive to models with low mass dark photons predominantly decaying to electrons.

A recent analysis with data collected by the CDF detector [88] searched for a W or Z in addition to many low energy electrons with momenta greater than 2 GeV, or muons with a momentum greater than 3 GeV. Due to the close proximity of the leptons, new data-driven identification algorithms were developed and no isolation requirements were used. The number of additional leptons were counted, with the SM predicting few events with multiple leptons. Models which produce low lepton multiplicity or no vector bosons decaying leptonically would evade this

search.

An analysis using data from the CMS detector searched for a dark photon resonance in the di-muon spectrum [89]. If the mass of the dark photon is lower than two times the mass of the muon then it will decay predominantly to electrons and will evade this search. In addition, this analysis is insensitive to processes dominated by three body decays.

An analysis using data recorded by the DØ detector focused on a particular signature producing a single photon, two leptons and missing energy [90]. Models without photons or with high multiplicity of particles would evade those limits. Another analysis at the same experiment searched for a signature of two groups of two leptons with missing energy [91]. Any models which produce wide jets or high multiplicity of particles would evade their analysis.

No evidence was found for any of these models and limits were set on each of their particular production mechanisms.

The analysis presented in this thesis attempts to broadly search for lepton jet signatures produced in association with a vector boson. The lepton jets are produced by the decay of a Higgs particle through the dark sector. Rather than counting the number of individual leptons reconstructed, this search uses the characteristics of the jets. This increases the sensitivity to different models as the search encompasses those producing varying number of leptons during the decay.

6.2 Model

6.2.1 Characteristics

Modifying the characteristics of the hidden sector, described in Section 4.4.2, has a great effect upon the properties of the resulting jets. Therefore, in order to cover as much phenomenological phase space as possible, multiple benchmark parameters of the dark sector were selected. In these benchmarks the focus was

on a three step decay, where there are three dark sector particles (h_{d0}, h_{d1}, h_{d2}), each with a particular mass in addition to the dark photon (γ_d). Using a relatively short decay chain in the dark sector reduces particle multiplicity and increases the amount of missing energy.

The masses of the three dark sector particles were chosen using simulations. In these simulations a simple mechanism which produces one particle per event with a particular set of parameters, referred to as a 'particle gun', was utilised from which the decay properties could be investigated. A Higgs particle of mass 120 GeV was created which then decayed into the dark sector. The masses in the simulations were altered and the properties of the resulting jets were recorded. The combination of masses which produced extremes of particular quantities were taken as the benchmarks. The quantities of interest include:

- MpT = Jet mass divided by jet p_T,
- N5 = Number of tracks carrying 50% of the jet energy,
- RoR = Energy in a cone of radius, R = 0.25 divided by that in a cone size of R = 0.5,
- NoN = Number of tracks carrying 50% of the jet energy divided by the number of total tracks, and
- pT = Amount of missing p_T.

6.2.2 Benchmarks

Figure 6.1 shows the results of varying the three masses in the particle gun simulation. From these results four benchmarks were selected which maximised one of the quantities (see Table 6.1). The labels Model A, Model B, Model C and Model D for each of the benchmarks will be used for the rest of this chapter.

Model	Dark Sector Particle Mass (GeV)			Quantity Maximised
	h_{d2}	h_{d1}	h_{d0}	
A	25.0	12.0	0.1	NoN
B	25.0	7.0	0.1	pT
C	15.0	6.0	0.1	MpT
D	1.0	0.21	0.1	N5

Table 6.1: Four benchmarks chosen which maximised interesting jet attributes. h_{d2}, h_{d1} and h_{d0} are the dark sector particles.

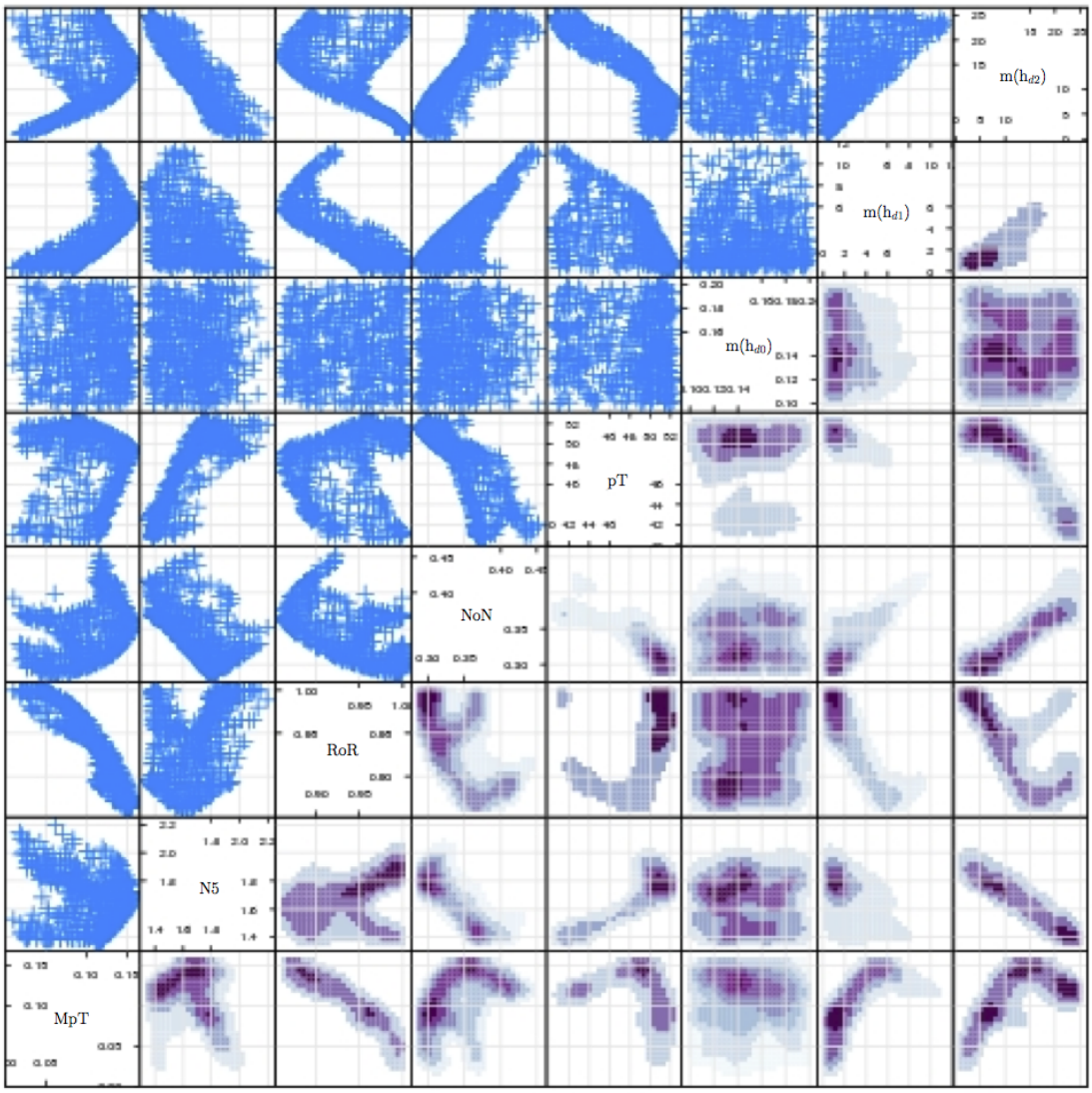

Figure 6.1: Scatter plot matrix showing the results of the Higgs particle gun simulation. Various quantities were measured (as described in the text) while altering the masses of the dark sector particles. The variable names are displayed along the diagonal squares which are used for the x-axis (y-axis) for that column (row). The uppermost three rows and three columns furthest to the right relate to the three dark sector masses m(h_{d0}), m(h_{d1}) and m(h_{d2}). For example, the bottom right square shows the effect of altering the mass dark sector particle h_{d2}, as plotted on the x axis, upon the MpT quantity, plotted on the y axis. The left hand side of the diagonal shows the traditional scatter plots while the right hand side shows a heat map representing the density of points [92].

6.2.3 Associated Production

In the model considered here, the Higgs is produced in association with a Vector Boson. The benefits of including a boson in the production model are two fold; triggering on the event becomes more efficient due to the isolated leptons from the boson decay, and secondly, the separation of the signal from the various backgrounds is enhanced. However, by including the boson in production, the cross sections for these processes are smaller than the total inclusive production (Figure 4.1). Six Higgs masses were chosen in order to increase the search space with values of m_H = 115, 125, 150, 200, 400 and 600 GeV.

The dark photon mass was set at 100 MeV. Figure 4.4 demonstrates that it will always decay into electron positron pairs. Figure 6.2 shows a graphical representation of a decay from this model. The cross sections for the different benchmarks are given in Table 6.2.

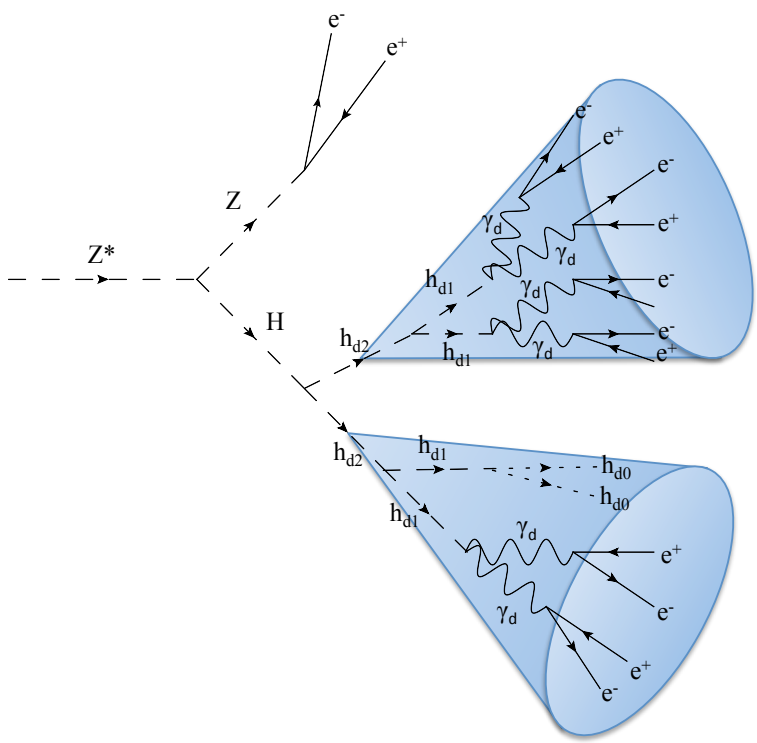

Figure 6.2: Graphical representation of the Higgs decaying into Electron Jets signature. Cones have been drawn around the collection of electrons to represent how the collimation of electrons produces two jet-like structures.

6.3 Data and Simulated Samples

6.3.1 Signal MC

The signal Monte Carlo (MC) simulations used for this analysis were produced by the CMS central production team. Around 200,000 events were produced for each of the four benchmarks and six Higgs masses with an associated Z boson. All events were generated using the PYTHIA 8 generator [93] and processed using the CMS software version, CMSSW_4_4_2_patch8[1]. The Monte Carlo generators are tuned using measurements from experiments. The signal samples were produced using the Tune4C tune. Table 6.2 lists the different signal samples along with their cross sections.

An important feature of the samples listed are the production cross sections for each process. For Higgs masses of 200, 400 and 600 GeV the cross sections are 1.25×10^{-5}, 3.65×10^{-8} and 1.58×10^{-9} pb respectively. Therefore, in order to produce any events, approximately 100 fb, 100 ab and 1 zb of data is required respectively. However, strong dynamics could enhance the cross sections and therefore these mass points were included in the analysis. Only the scenarios involving the lighter Higgs masses of 115, 125 and 150 GeV were used when optimising the search strategy.

6.3.2 Background MC

The background Monte Carlo samples used in this analysis were also produced by the official CMS central production team. The different datasets are outlined in Table 6.3. All datasets were produced using CMSSW_4_4_X. The MadGraph generator [94] was used to produce the Drell Yan (DY) plus jets[2] and $t\bar{t}$ samples. POWHEG [95] was used to generate the single top samples. The diboson samples are produced using PYTHIA 6 [96]. All background MC samples use the TuneZ2 tune which is applicable for PYTHIA 6.

[1]CMSSW is described in Section 5.2.7. The 4_4_X series was the recommended software release to be used for 2011 data and MC.

[2]Also referred to as Z+Jets in this work

All of the background processes were generated at Leading Order level (LO) and scaled using a ratio (K-factor) of the Next to Leading Order level (NLO) to LO. The exception to this was the DY plus jets sample which was generated at LO then scaled using a K-factor at Next to Next to Leading Order (NNLO), calculated with the FEWZ code [97]. For the $t\bar{t}$, diboson and single top samples the MCFM code [98] was used to calculate the cross section to NLO and subsequent K-factor.

The cross sections and associated theoretical errors are given in Table 6.3. The error has contributions from scale and total Parton Distribution Function (PDF) uncertainties. Scale uncertainty is determined by varying the factorisation and renormalisation scales. The PDF uncertainties are calculated using procedures defined by the various groups [99–101] and combined according to the PDF4LHC working group prescriptions [102].

6.3.3 Data

The data used in this analysis was recorded in 2011 by the CMS experiment in the presence of a single muon. The CMS physics validation team (PVT) creates lists of runs that are deemed good for use within an analysis as a JavaScript Object Notation (JSON) file. The particular JSON file used in this analysis was Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt. The total amount of data used corresponds to 4.83 fb^{-1} at a center of mass energy of $\sqrt{s} =$ 7 TeV. Table 6.4 shows a summary of the dataset split into two different epochs, Run A and Run B, related to specific LHC run ranges.

Model-Point	Dataset	Cross-Section (pb)
ZH115A	/ZH115_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0179
ZH115B	/ZH115_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0179
ZH115C	/ZH115_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0179
ZH115D	/ZH115_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0179
ZH125A	/ZH125_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0121
ZH125B	/ZH125_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0121
ZH125C	/ZH125_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0121
ZH125D	/ZH125_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.0121
ZH150A	/ZH150_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.00244
ZH150B	/ZH150_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.00244
ZH150C	/ZH150_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.00244
ZH150D	/ZH150_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	0.00244
ZH200A	/ZH200_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.25E-5
ZH200B	/ZH200_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.25E-5
ZH200C	/ZH200_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.25E-5
ZH200D	/ZH200_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.25E-5
ZH400A	/ZH400_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	3.65E-8
ZH400B	/ZH400_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	3.65E-8
ZH400C	/ZH400_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	3.65E-8
ZH400D	/ZH400_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	3.65E-8
ZH600A	/ZH600_Dark_3step_25_12_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.58E-9
ZH600B	/ZH600_Dark_3step_25_7_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.58E-9
ZH600C	/ZH600_Dark_3step_15_6_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.58E-9
ZH600D	/ZH600_Dark_3step_1_pt21_7TeV_Tune4C-pythia8/Fall11-PU_S6_START44_V9B-v1/AODSIM	1.58E-9

Table 6.2: The signal MC samples used in this analysis along with the naming convention used for each benchmark and their production cross sections with the dark sector and Z to leptons (e,μ,τ) branching ratios folded in.

Process	Dataset	Cross-Section (pb)	Uncertainty
Z+Jets ($m_{ll} > 50$)	/DYJetsToLL_TuneZ2_M-50_7TeV-madgraph-tauola/	3048.0	±132.0
$t\bar{t}$	/TTJets_TuneZ2_7TeV-madgraph-tauola/	157.5	+23.2-24.4
WW	/WW_TuneZ2_7TeV_pythia6_tauola/	43.0	±1.5
WZ	/WZ_TuneZ2_7TeV_pythia6_tauola/	18.2	±0.7
ZZ	/ZZ_TuneZ2_7TeV_pythia6_tauola/	5.9	±0.15
Single Top (t-ch)	/T_TuneZ2_t-channel_7TeV-powheg-tauola/	42.6	+2.4-2.3
Single Top (t-ch)	/Tbar_TuneZ2_t-channel_7TeV-powheg-tauola/	22.0	+1.0-0.8
Single Top (tW)	/T_TuneZ2_tW-channel-DR_7TeV-powheg-tauola/	10.6	±0.8
Single Top (tW)	/Tbar_TuneZ2_tW-channel-DR_7TeV-powheg-tauola/	10.6	±0.8
Single Top (s-ch)	/T_TuneZ2_s-channel_7TeV-powheg-tauola/	2.72	+0.11-0.10
Single Top (s-ch)	/Tbar_TuneZ2_s-channel_7TeV-powheg-tauola/	1.49	+0.09-0.08

Table 6.3: The background MC Samples used in this analysis along with their recalculated cross sections and the theoretical errors (as described in the text) associated with them.

Dataset	$\mathcal{L}(\mathrm{fb}^{-1})$	LHC Run range
/SingleMu/Run2011A-08Nov2011-v1/AOD	2.09	160431-168437
/SingleMu/Run2011B-19Nov2011-v1/AOD	2.74	175832-180296
Total Luminosity	4.83	

Table 6.4: 2011 datasets used in this analysis.

6.4 Search Strategy

The main focus of this analysis was to find lepton jets using the jet characteristics rather than trying to reconstruct the dark photon mass from objects within the jet. Reconstruction of the dark photons was attempted in a similar fashion to the analysis searching for a dark photon resonance in the di-muon spectrum using data from the CMS detector [89]. Figure 6.3 demonstrates that it would be possible to find the dark photon mass, however the efficiency would be too low.

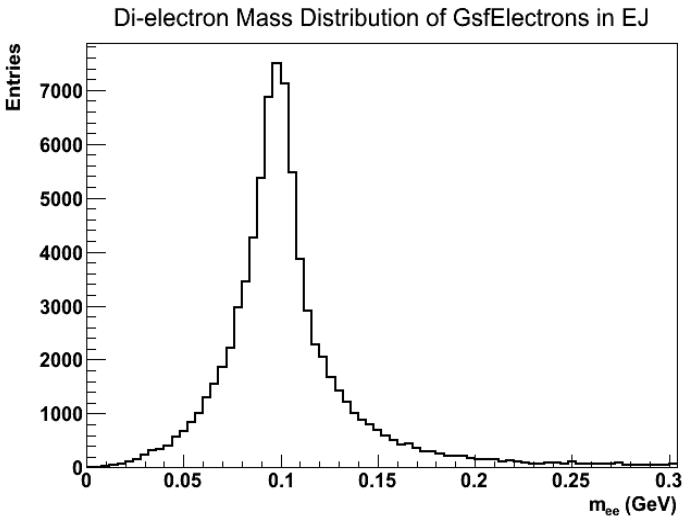

Figure 6.3: Plot showing the reconstruction of the dark photon mass.

This low efficiency is due to poor reconstruction of the electrons within the EJ. Not all the electrons in the EJ can be detected. Some electrons will be outside of the acceptance region of the detector and therefore not detectable. Other electrons will have too low a momentum to reach the ECAL as they will be diverted by the magnetic field. Figure 6.4 shows the p_T distribution of the generator level electrons. Figure 6.5 shows the number of electrons from generator level produced for each EJ and Figure 6.6 shows the number of electrons, as defined in Section 6.6.2, found inside EJs produced from the different benchmarks. The number of electrons found in Model D is lower than the other benchmarks due to problems reconstructing electrons which lay close to each other as described in Section 6.9.4.

The strategy for this analysis was to use the Z boson as a tagging tool for the

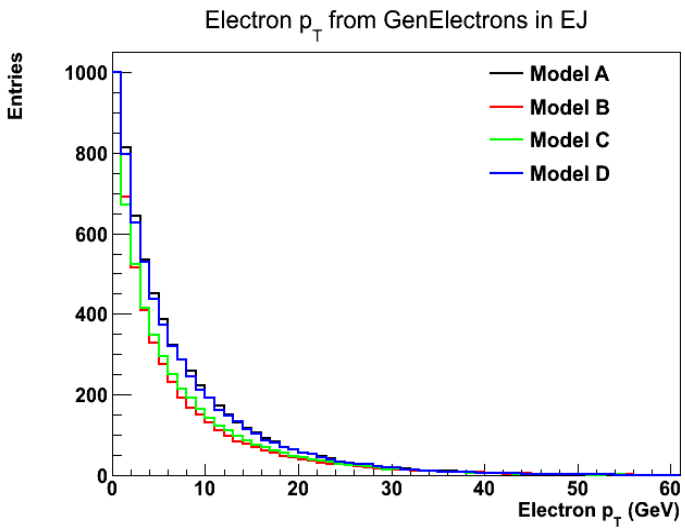

Figure 6.4: The p_T distribution of generated electrons inside an EJ for the different benchmarks: Model A, B, C and D.

event. After triggering on a muon decaying from the Z, the two muons produced from the decay are searched for and the invariant mass of the two hardest muon candidates in the event calculated. If this mass is in range of the Z invariant mass, a Z candidate is said to be found. Once a Z boson is found, an EJ identification is applied to each jet in the event. The invariant mass of the two hardest jets passing the identification cuts is then calculated. A sliding dijet mass window is applied and the number of events inside the window counted. If a signal exists, an excess of this number above the background would be observed. The details of each step in this process are described in the following sections.

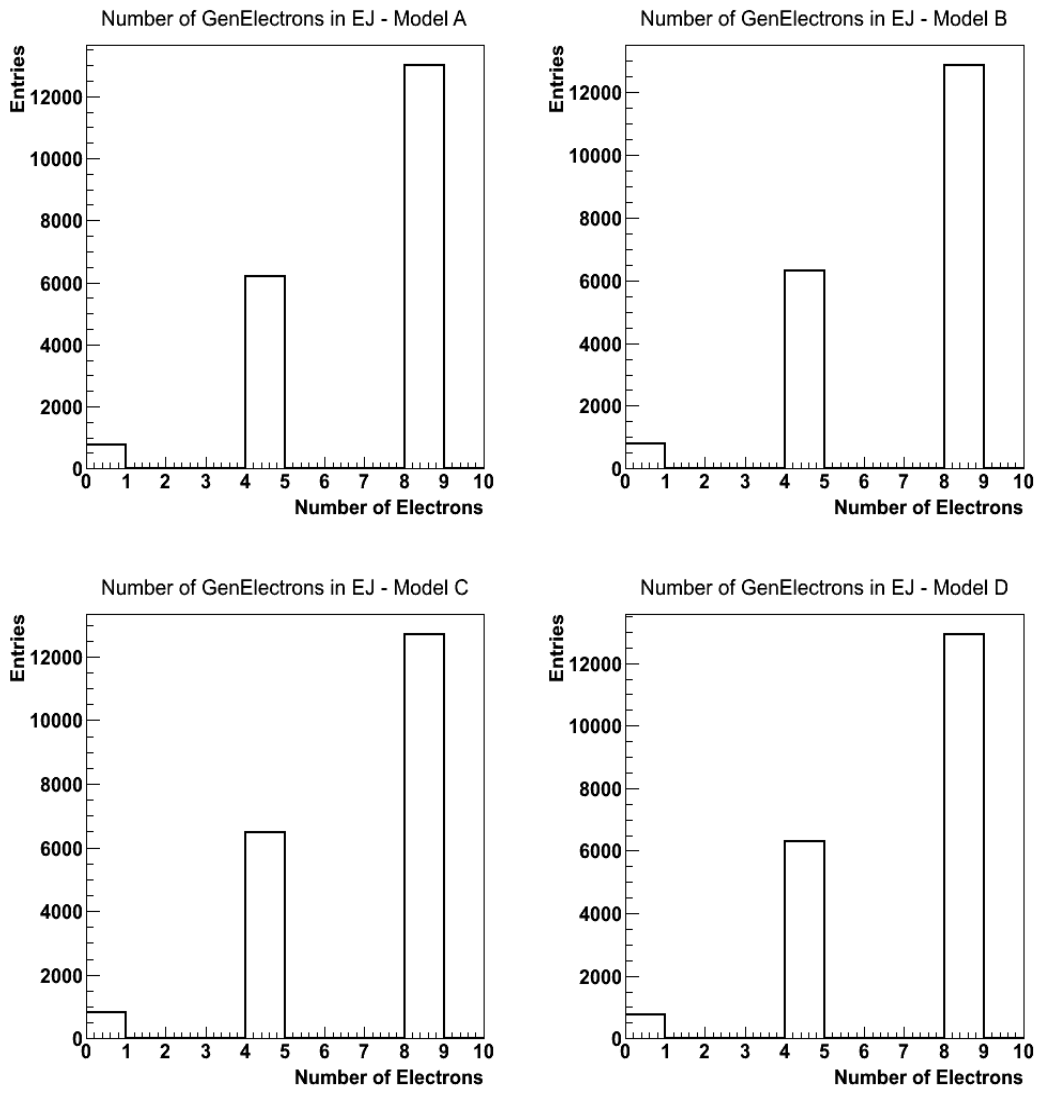

Figure 6.5: The number of generated electrons inside an EJ for the different benchmarks: Model A, B, C and D.

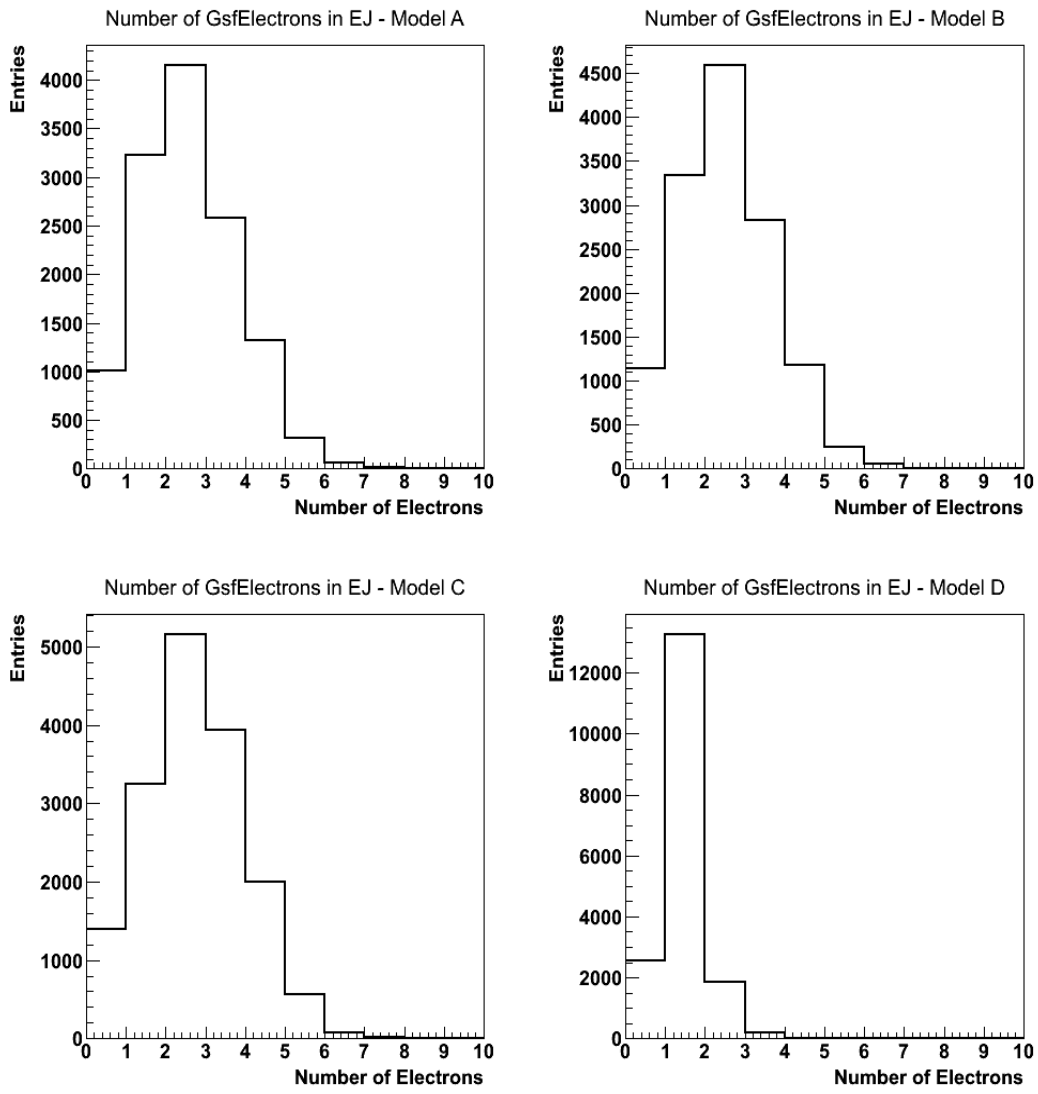

Figure 6.6: The number of reconstructed electrons found inside an EJ for the different benchmarks: Model A, B, C and D.

6.5 Trigger Requirements

The trigger used in this analysis utilised the muons produced from the decay of the Z boson. A muon with a $p_T > 40$ GeV and $|\eta| < 2.4$, corresponding to the pseudorapidity range of the muon detector, was used in the earlier period of data collection, Run A. During later runs, Run B, the instantaneous luminosity delivered to CMS was increased. Due to the fixed limit on the output rate of the number of events that the L1 trigger processes and the increased number of events delivered to CMS, a more restrictive requirement was necessary on the muon trigger. For the later runs a trigger requiring a muon with a $p_T > 40$ GeV and $|\eta| < 2.1$ was used. The OR of these two trigger paths was used in the analysis to ensure the maximal amount of data was used.

6.6 Physics Objects

6.6.1 Primary Vertices and Pile Up

The primary vertices (PV) in an event were selected using the Deterministic Annealing algorithm [103]. These vertices are required to satisfy the standard CMS selection [104]:

- The z position from the nominal detector centre is < 24 cm,
- A radial position from the beam line is < 2 cm,
- Have at least 4 associated tracks.

The primary vertex, which produced the particles that caused the event to trigger, was selected from these vertices. This vertex is required to have the largest value of $\Sigma_i p_{T_i}^2$, where p_{T_i} is the transverse momentum of the ith track associated with the vertex.

Many parasitic collisions were produced per bunch crossing due to the high luminosity of the LHC in 2011. These additional interactions are collectively known as

pile up (PU). The number of PU interactions for each triggered event is directly related to the number of primary vertices in that event. On average there were six PU interactions in the early runs; while during later runs the number of PU interactions increased to over ten.

PU leads to increased numbers of low p_T jets and tracks seen in the events. The majority of these jets are in the very forward region, so by only considering jets with $\eta < 2.2$ a large proportion are eliminated from the event. Setting the requirement that physics objects of interest originate from the primary vertex also helps in eliminating PU objects.

The Fall11 Monte Carlo used in this analysis has a PU profile similar to that of the 2011 data. However, the events in MC have to be reweighted in order for the simulations to accurately represent the processes in the data. This was achieved by first calculating the distributions of the number of primary vertices per event for both MC and data, and scaling to unity as shown in Figure 6.7. The ratio of the two plots can be used to determine the weights to be applied to each event from MC based on the number of primary vertices in each event. This PU reweighting plot is shown in Figure 6.7.

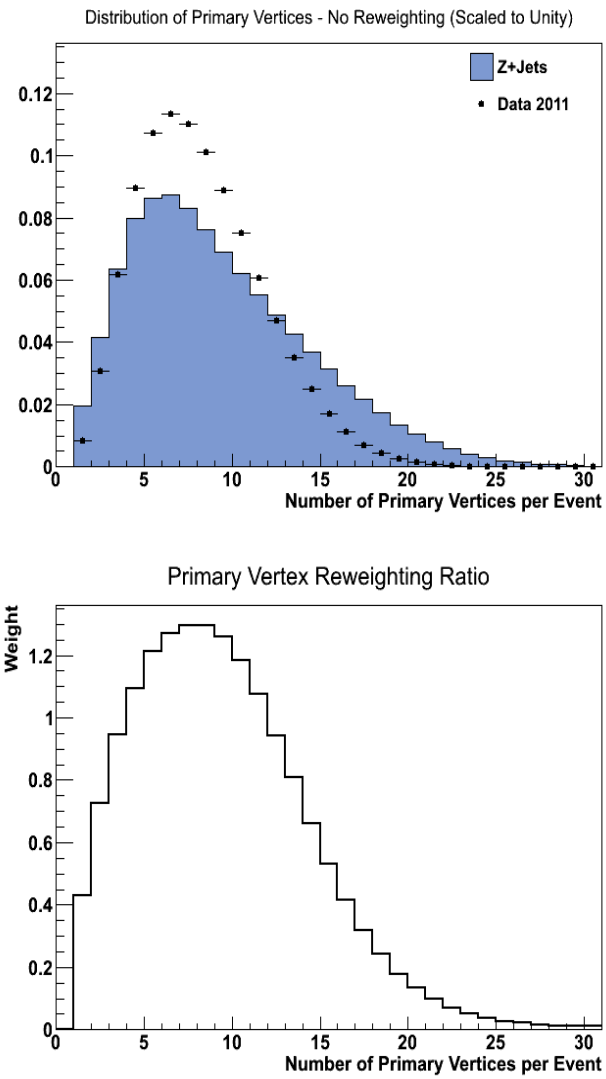

Figure 6.7: Plots showing the the distribution of the number of primary vertices per event for both MC and data after scaling to unity and the ratio used for reweighting.

6.6.2 Leptons

Muons

The muons used in this analysis were first reconstructed using standard CMS algorithms [105] and [106]. Muons were reconstructed by finding consecutive hits within the DT, CSC and RPC muon chamber layers which were then used as seeds in a Kalman track filter algorithm [105]. The collection of tracks produced using this algorithm are referred to as 'standalone muons'. After finding a standalone muon, the tracks found in the silicon tracker are searched and the best match to the muon, as defined through the use of a Kalman filter, is selected. This pairing of the track from the silicon tracker and the standalone muon in the muon detectors is referred to as a 'global muon'.

For this analysis the 'tight' muon identification selection for 2011 data as defined by the muon Physics Object Group was used. This significantly reduces the rate of muons from decays in flight, at the price of a small loss in efficiency for prompt muons such as those from W and Z decays [85]. Global muons are used with the following 'tight' muon quality requirements:

- χ^2/number of degrees of freedom of the global-muon track fit < 10,
- At least one muon chamber hit included in the global-muon track fit,
- Muon segments in at least two muon stations,
- The silicon tracker track has transverse impact parameter $\mathrm{d}_{xy} < 2$ mm with respect to the primary vertex,
- Number of pixel hits > 0, and
- Number of hits across the silicon tracker > 10.

In addition to these requirements, the muons must pass the following isolation, pseudorapidity and p_T requirements:

- Relative combined isolation, $\mathrm{R} < 0.15$,

- $|\eta| < 2.4$, and
- $p_T^\mu > 20$ GeV.

The relative combined isolation is calculated using tracker and calorimeter information. The algorithm calculates the scalar sum of the p_T of all tracks reconstructed in the tracker, as well as the sum of energies measured in ECAL and HCAL towers, within a cone of radius 0.3 centred on the muon track direction. This sum is then divided by the p_T of the muon. The track p_T and energy deposits associated with the muon itself are not included in the sum. The ratio of this sum to the deposits associated with the muon are required to be less than 0.15.

Electrons

Electron reconstruction begins by grouping energy deposits in ECAL crystals into clusters. If there were no material between the IP and ECAL and no magnetic field, 97% of the energy from a single electron or photon would be contained in a 5×5 group of crystals. However, with material present due to the tracking system, the electrons bremsstrahlung and photons undergo conversions. In the presence of the magnetic field the energy spreads in ϕ. To overcome this problem, the clusters of ECAL crystals are grouped in the ϕ direction to make superclusters using algorithms described in detail in [107].

These superclusters are used to initiate an iterative algorithm which searches the pixel tracker for associated detector hits. If two hits matching the trajectory of the energy weighted average centre point within a ϕ and z window are found, they become associated with that supercluster. The seeds are required to have an $\text{Energy}_{HCAL}/\text{Energy}_{ECAL} < 0.1$ and an $E_T > 4$ GeV.

A dedicated electron tracking algorithm, based on a combinatorial Kalman filter with a dedicated Bethe Heitler modelling of the electron energy losses, is run from the seeds [107]. This algorithm iterates over the tracker layers testing candidate trajectories with a loose χ^2 compatibility. A gaussian sum filter (GSF) fit is applied for each tracker hit to estimate the electron track parameters.

The electron candidates are built from the reconstruction of GSF tracks and their associated superclusters. These candidates are required to pass a loose set of selection cuts based upon their position of closest approach to the supercluster position as extrapolated from the innermost track position and direction. Electrons from reconstructed conversion legs, which are from photons radiated by primary electrons, are cleaned by resolving cases where several tracks are associated to the same supercluster [107].

The electrons used in this analysis are defined as objects from the gsfElectron collection with no extra identification or isolation requirements.

6.6.3 Jets

Calorimeter based jet reconstruction uses energy deposits in the calorimeter to create jet objects. The ECAL and HCAL energy deposits are combined into projective collections based on the granularity of the HCAL, called towers. Clustering algorithms are then run over these towers to produce the jets.

Numerous jet algorithms exist, which can be classified into two groups; cones and sequential recombination. In cone algorithms jets are defined geometrically. The particle with highest p_T is taken as the jet axis, upon which a cone of size R in $\eta - \phi$ space is cast around and all particles inside the cone are marked as jet constituents. The energy and direction of these constituents are used to recalculate the jet axis. The procedure is repeated until the energy of the jet changes by less than 1% between iterations and the direction of the jet changes by $\Delta R < 0.01$. The jet constituents are then removed from the list of possible inputs, the stable jet is added to the list of jets and the whole process repeats until no more objects above an assigned threshold remain [108].

Cone algorithms generally suffer from being infrared (IR) and collinear unsafe, whereby adding a new soft particle or splitting the partons in a collinear fashion leads to an extra hard stable cone being found . The Seedless Infrared Safe Cone (SISCone) algorithm attempts to overcome these problems [109]. Sequential recombination algorithms are IR and collinear safe.

Sequential recombination algorithms such as kT [110,111] and Anti-Kt (ak) [112] algorithms attempt to work backwards through branchings of quarks and gluons, repeatedly combining objects to form jets. The kT algorithm combines pairs of particles with the smallest distance, as defined by a minimisation function, which are then merged and filed as one new object in the list of input objects. This process is repeated until a defined distance is reached, at which point the resulting object is recorded as a jet and removed from the list of possible inputs. The whole procedure is repeated until all objects are included in jets [108]. Particles which have been radiated with low momentum are therefore clustered first. Anti-kT jets are formed using a different minimisation function which favours clustering of particles with high p_T. For ak jets the area that encloses the jet constituents is approximately a circle in the $\eta - \phi$ plane, unlike kT jets.

CMS uses a variety of algorithms to find jets through an interface with the experiment independent FastJet package [113]. The ak algorithm for sizes equivalent to R = 0.5 and 0.7, the kT algorithm for R = 0.4 and 0.6, and the SISCone algorithm for R = 0.5 and 0.7 are all supported.

These algorithms can be run over standard jet reconstruction objects such as generator level MC particles producing generator jets, calorimeter towers producing calorimeter jets, and Particle Flow candidates producing particle flow jets (PFJets). The most common type used in CMS are the CaloJets with Anti-Kt algorithm for sizes corresponding to radii 0.5 and 0.7, which are referred to as ak5CaloJets and ak7CaloJets for the rest of this thesis. The Anti-Kt algorithm is IR and collinear safe, provides convenient jet area shapes which are useful in pileup subtraction and has been shown to give the best performance when resolving many jets in complicated events [114].

Although CMS has recently attempted to move towards particle flow based techniques for the majority of analyses, PFJets could not be used in this analysis. The particle flow algorithms had difficulty in identifying individual electrons when they were closely positioned to other objects. Instead of reconstructing the electrons as electrons, they are identified as neutral hadrons and photons. This misreconstruction occurs due to assumptions made in the particle flow algorithms.

Usually it is necessary to apply corrections to the jets in an analysis in order

to translate the measured jet energy to the true particle or parton energy. This is due to the non-linear calorimeteric response to particles. These correct for jet response versus pseudorapidity, p_T, amount of energy measured in the ECAL divided by that in the HCAL, known as the Electromagnetic Fraction (EMF), jet flavour, and parton corrections. However in this analysis only the jets which are formed out of electrons were of interest and consequently the usual jet corrections do not apply.

6.7 Vector Boson Reconstruction

The $Z \rightarrow \mu\mu$ decay can be identified using the muons which pass the selection described in Section 6.6.2. Candidate Z bosons are selected based on the requirement that the two muons with largest transverse momentum produce an invariant mass between $75 < M_{\mu\mu} < 105$ GeV. Figure 6.8 shows the di-muon invariant mass spectrum plotted using the data recorded (points) and using simulated events (histograms), where the number of events is scaled to that expected given the integrated luminosity used. Good agreement can be seen between data and MC simulation.

6.8 Pre-Selection and Jet Candidate Selection

In order to provide a clean sample of events, at least one PV was required using the methods described in Section 6.6.1 and each event should not contain more than 31 PV. The requirement of one PV ensured that a hard collision was reconstructed. An upper limit existed in order to allow comparison between data and MC, as events with more than 31 PV were not modelled in the simulations. The relative proportion of data which contained events with > 31 PV is low and can be safely excluded.

Prior to EJ identification, the jets are filtered to exclude those produced from PU thereby creating a reasonably pure sample. Each jet must pass the following Jet Candidate Selection (JCS):

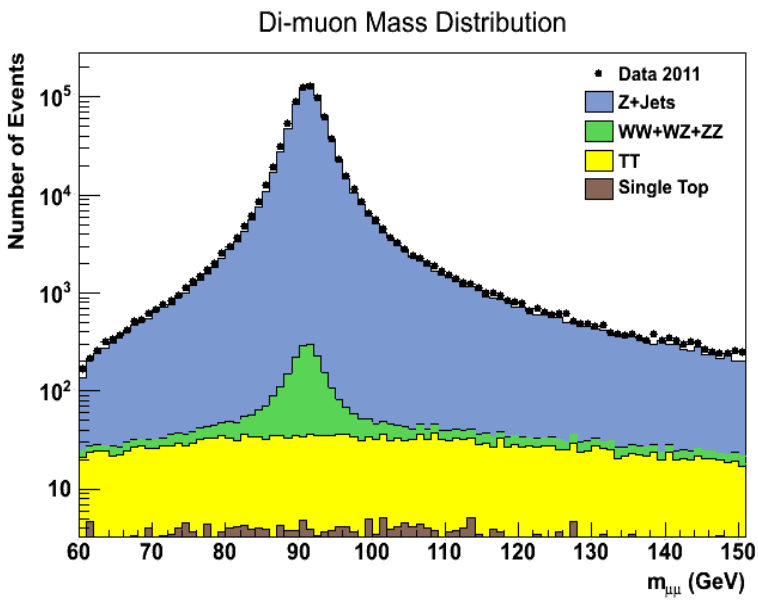

Figure 6.8: Di-muon mass distribution for events passing the selection defined in the text using the 4.83 fb^{-1} of data from 2011.

- Jet $|\eta| < 2.4$,
- Jet $p_T > 10$ GeV, and
- The jet must contain at least one track which points to the Primary Vertex.

The pseudorapidity requirement was implemented as the tracking detector can only measure tracks within $|\eta| < 2.5$. The selection uses a more restrictive range compared to the detector limits to avoid issues with tracks from the jet falling outside the sensitive region. Jets arising from PU will normally have a low momentum, therefore a minimum cut on the p_T was implemented. A requirement of at least one track pointing to the PV also excludes large numbers of jets from PU.

6.9 Electron Jets Identification

Once a Z candidate has been identified and the pre-selection satisfied, the collection of jet candidates in the event are analysed. Each jet is tested to determine if it passes a jet ID selection. This selection has been modified over several generations as described below.

6.9.1 Scheme A

EJ identification was initially explored using selection points defined in the theoretical paper [74], referred to here as Scheme A. Falkowski et al. tested variables, described in an earlier work [68], to separate EJ's from QCD jets. In their simulations the jet EMF and a quantity called the Charged Ratio (CR) was found to be particularly powerful at discriminating between different types of jets. CR is defined as E_T^{jet}/p_T^{jet}. Figure 6.9 gives an example of these two variables as constructed from their simulations.

When these cuts were applied on our samples it was found that the QCD jet background efficiency was $\mathcal{O}(10^{-3})$. However the signal only achieved an efficiency of 20% to 40%, approximately the same efficiency as the single electron background.

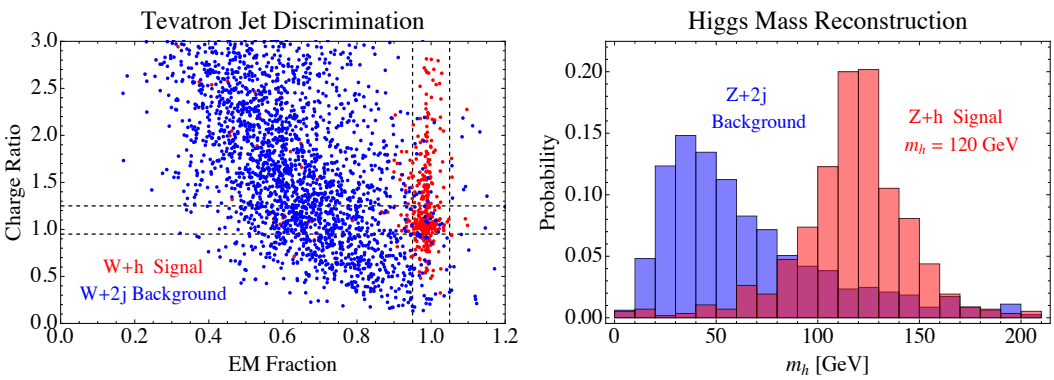

Figure 6.9: Plots showing the selection points and predicted discrimination power as suggested by Falkowski et al. [74].

6.9.2 Scheme B

Following on from the results of Scheme A, multivariate analysis techniques (MVA) such as Boosted Decision Tree (BDT) classifiers were used. The BDT used was from the TMVA software suite in ROOT. Variables such as EMF, E_T^{jet}/p_T^{jet}, $\sigma(\phi\phi)$, $\sigma(\eta\eta)$, dR, ntrks/p_T and number of electrons were used as inputs to the classifier. Figure 6.10 shows an output of the software after training the classifier on signal, QCD background and single electron background MC simulations. The signal used in classifier training was from benchmark model A. After training, a decision boundary was set which discriminated between these three different classes. These plots indicate how well the various classes can be separated.

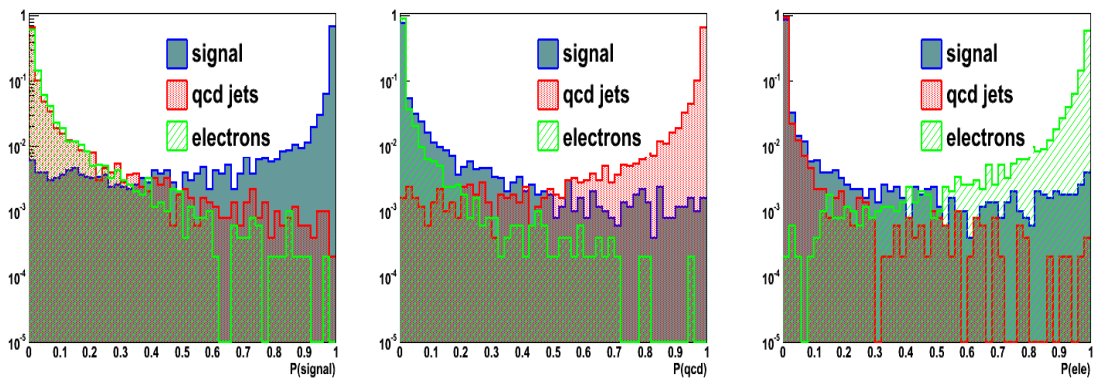

Figure 6.10: Plots showing the BDT response for the probability when selecting signal, qcd and single electron background [92].

By applying the trained classifier to signal benchmarks, such as the one used in training, a high efficiency can be achieved as shown in Figure 6.11. However, as can be seen in the same figure, there were concerns about performance when using BDTs with different benchmarks. There were also large uncertainties in modelling the dark sector to which BDTs and other MVA classifiers would likely be more sensitive than traditional cuts.

6.9.3 Scheme C

The variables which gave the most powerful discrimination in the BDTs of Scheme B were noted and a new cut-based identification was developed. The aim of this

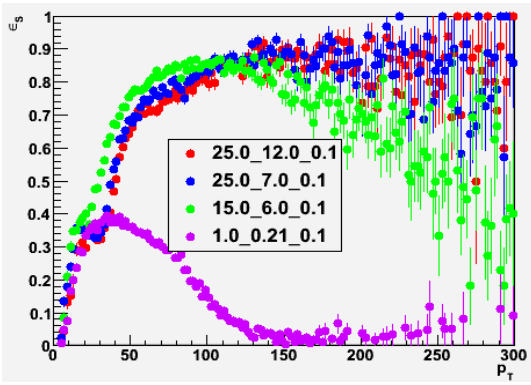

Figure 6.11: Plot showing the efficiency of Scheme B, highlighting concerns over benchmark dependence. The colours red, blue, green and purple represent the benchmarks Model A, B, C and D respectively. P_T is measured in GeV [92].

new cut-based identification was to achieve efficiencies similar to that of the BDT whilst being benchmark independent. The most effective jet discriminants used the EMF of the jet and the number of electrons in the jet. Using this information a new selection scheme was developed as follows:

- EMF > 0.95,
- Number of tracks in the jet > 2,
- Number of electrons in the jet > 0.

If the number of electrons in the jet $= 1$ the following additional requirements were necessary:

- If jet $p_T < 70$ then $\sigma(\phi\phi) > 0.005$ else $\sigma(\phi\phi) > 0.005\text{-}0.0001\times(p_T\text{-}70)$, and
- The number of tracks in the jet divided by the jet $p_T < 0.45$ GeV.

The spread of the energy in the ϕ direction within the jet, $\sigma(\phi\phi)$, is used in the case where one electron is found in the jet. This quantity is calculated from the second moment of energy deposits within the calorimeter towers in ϕ as:

$$\sigma(\phi\phi) = \frac{\sum((\Delta\phi)^2 E_T) - (\sum(\Delta\phi E_T))^2 / \sum E_T}{\sum E_T}, \tag{6.1}$$

where the summations are over the calorimeter towers associated with the jet, E_T is the energy measured in the tower and $\Delta\phi$ is the distance in ϕ between the tower and jet axis. If the number of electrons in the jet > 2 then no additional requirements were needed as the background was small. The motivation for using a cut on $\sigma(\phi\phi)$ as a function of p_T is outlined in Figure 6.12 which shows the majority of the single electron background occupying specific values in the plot. This 'sliding' cut is represented by the red dotted line. The cut on the number of tracks in the jet divided by the jet p_T was applied to reduce the amount of background from QCD jets, as the jet constituents carry a relatively smaller fraction of p_T.

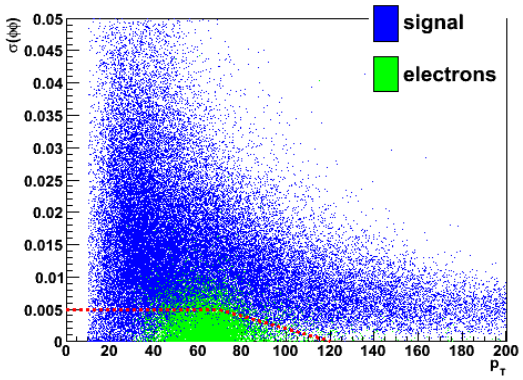

Figure 6.12: Two dimensional scatter plot of $\sigma(\phi\phi)$ vs jet p_T showing discrimination power between the signal and background in MC. P_T is measured in GeV [92].

The efficiency of Scheme C on the four benchmarks: Model A, B, C and D over the selected values for the Higgs mass is shown in Figure 6.13. The selection was tested on ak7CaloJets matched to the jets constructed from the decays at generator level for the signal efficiencies. The MC efficiency for the main backgrounds are also shown. The background efficiencies were calculated using all ak7CaloJets passing the jet candidate selection as described in Section 6.8. While the background efficiency appears to be low at 10^{-2}, the signal efficiency only peaks at 80% at high jet p_T and is quite low for the lower p_T jets.

Events with 2 or more jets passing the selection can be used to reconstruct the Higgs mass. Figure 6.14 shows the dijet invariant mass of the two jets matched and reconstructed for the various signal MC. The plots have not been scaled by the cross section for the processes. Generally the invariant mass of the dijet

system matches closely with the generated Higgs mass (shown as a purple dashed line for guidance). Multiple peaks appeared from the different decay processes allowed in the dark sector, where certain amounts of missing energy would be missed in the jet reconstruction. Figure 6.15 shows the dijet mass spectrum for the main backgrounds, constructed from the two highest p_T jets which passed selection, not scaled by cross sections. A large peak appeared around the Z pole in the ZZ background, whilst all other backgrounds were minimal.

An independent analysis of the signal MC compared with background MC using DataViewer was undertaken with the aim of confirming or disproving the discrimination power of these variables. Figure 6.16 shows DataViewer in operation while investigating the performance of Scheme C. The identification of the $\sigma(\phi\phi)$ vs p_T cut can be seen in the scatter plot at the top middle window of the screen. Through use of DataViewer, it was confirmed that these variables were useful in discrimination.

After applying the Scheme C selection within the full analysis chain, the number of events expected to be observed in the 4.83 fb^{-1} of data was too low to be useful. Table 6.5 shows the acceptance of Scheme C for signal events with Z bosons decaying to muons along with the event yield expected using the 2011 data, calculated by multiplying the cross section of the process with the integrated luminosity and then multiplying by the acceptance. The number of events expected from background is also given. However, it should be noted that the number of events corresponds to those left in the entire p_T range and not those in the final selection window.

Model	Number of events passing Scheme C	Acceptance	Number of events expected in 2011 data	
			Signal	Background
ZH115A	3291	4.93%	1.42	2.47
ZH115B	3284	4.92%	1.42	
ZH115C	3530	5.29%	1.53	
ZH115D	3882	5.82%	1.68	
ZH125A	3826	5.74%	1.12	
ZH125B	3762	5.64%	1.10	
ZH125C	4076	6.11%	1.19	
ZH125D	4427	6.64%	1.29	
ZH150A	5038	7.55%	0.30	
ZH150B	5082	7.62%	0.30	
ZH150C	5279	7.92%	0.31	
ZH150D	5568	8.38%	0.33	

Table 6.5: Scheme C acceptance for Z decaying to muons and predicted yield for the signal benchmarks and background.

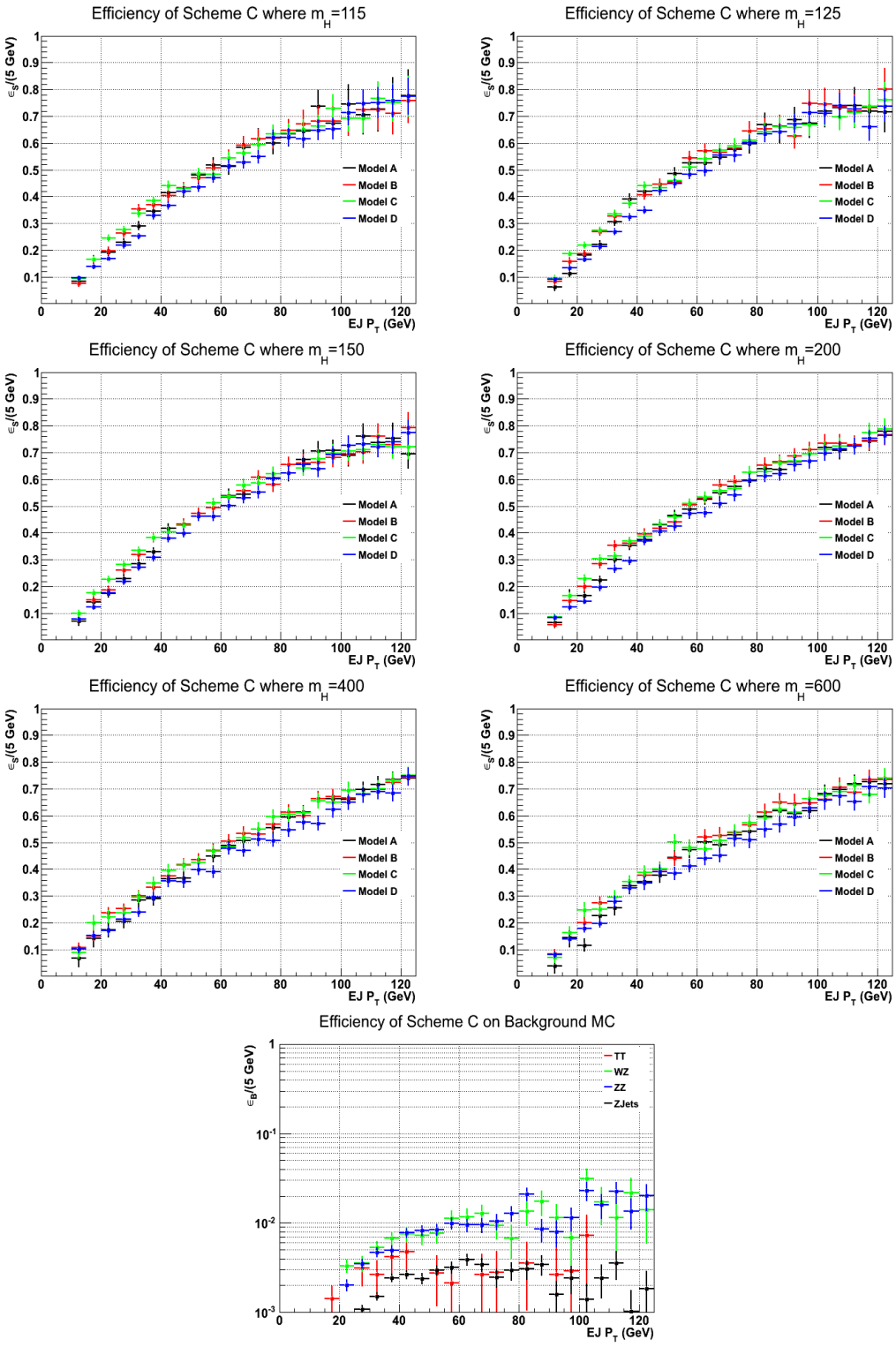

Figure 6.13: Plots showing the efficiency of Scheme C as a function of jet p_T on the background and signal MC for the benchmarks corresponding to different Higgs mass points and dark sector decays. The signal efficiencies were calculated using jets matched to the MC truth.

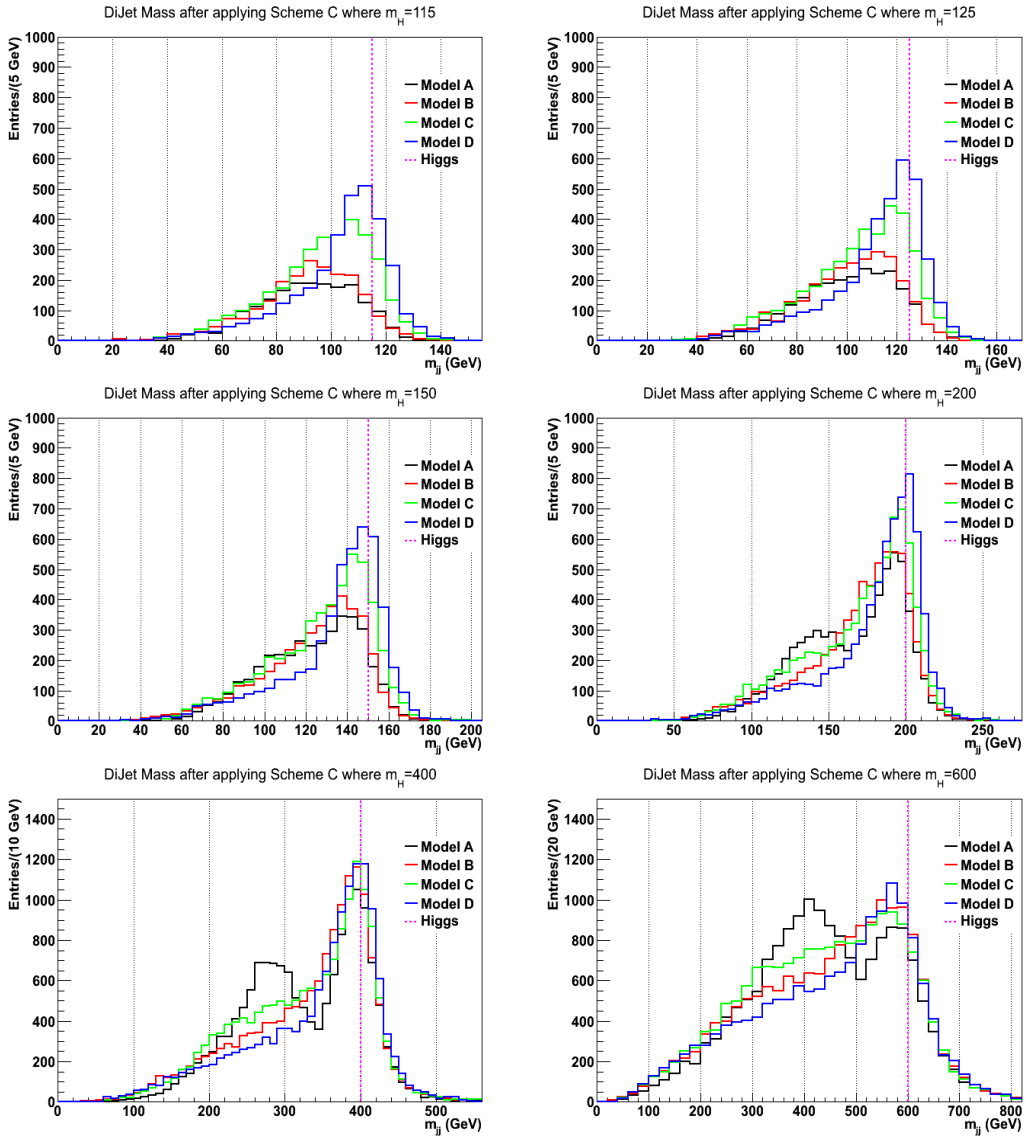

Figure 6.14: Dijet mass spectrum from the signal MC using jets passing Scheme C.

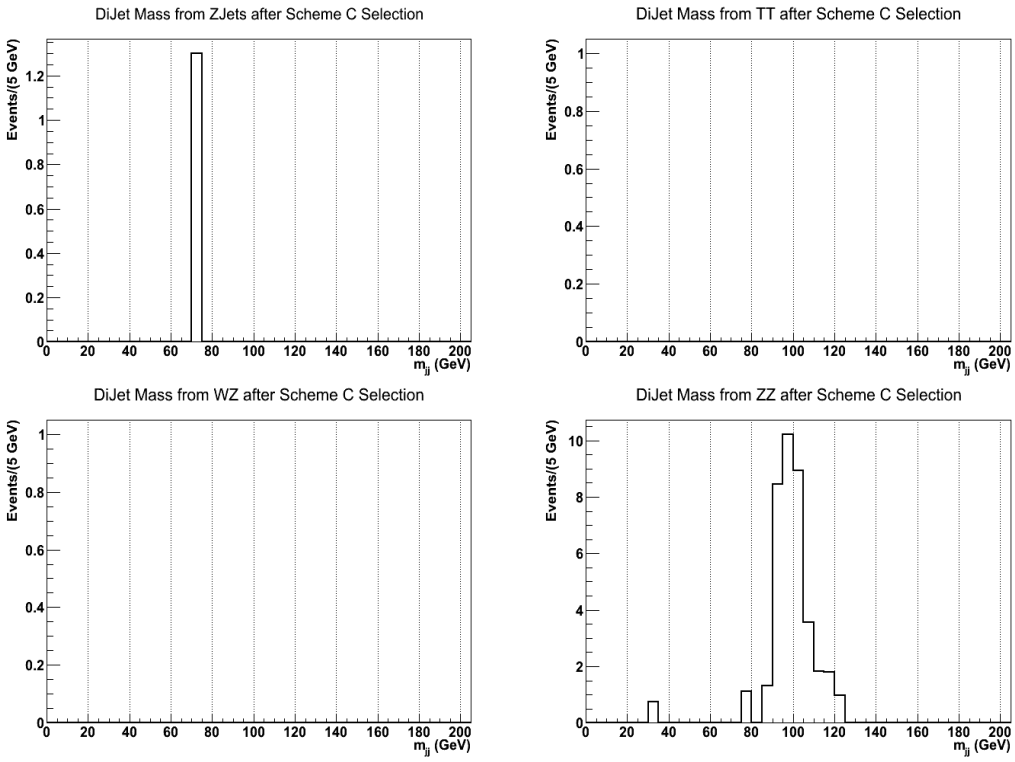

Figure 6.15: Dijet mass spectrum from the background MC using jets passing Scheme C.

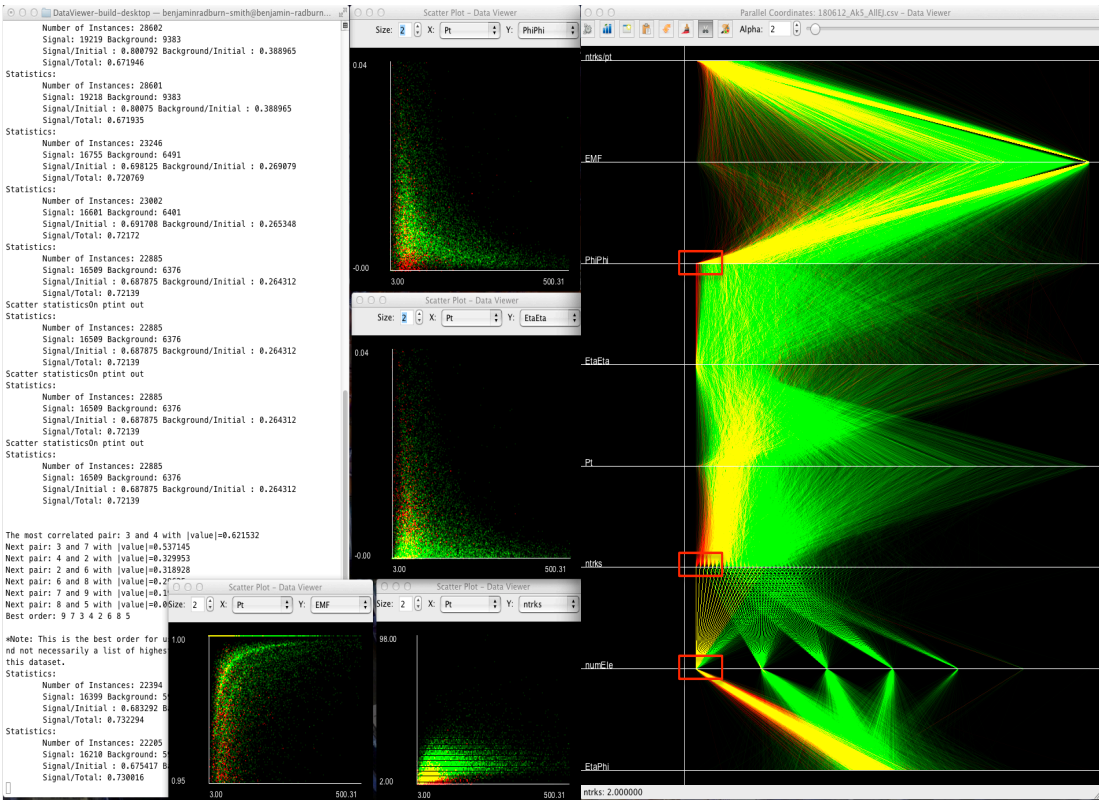

Figure 6.16: Screen shot of DataViewer in operation whilst investigating Scheme C performance. The Z+Jets background MC is coloured as red and the Signal MC is coloured as green. Red boxes have been superimposed on the image to show the areas of parallel coordinates plot exhibiting interesting features.

6.9.4 Scheme D

Concerns over benchmark dependence, due to the requirement of variables dependent on p_T in Scheme C, led to modifications in selection. In Scheme D the selection was identical to that of Scheme C except in situations where only one electron was found in the jet. Instead of using the $\sigma(\phi\phi)$ variable, an isolation cone was placed around the electron. In the case of single electrons from EWK processes it was expected that the amount of energy within a cone of R = 0.3 would be very small. In contrast, for EJs one would expect the amount of energy in this cone to be much higher as there would be electrons which have not been fully reconstructed in close proximity to the fully reconstructed electron.

This scenario can be termed an anti-isolation cut as there would be additional energy around the electron. The relative isolation of the electron was used, which is equal to the amount of energy measured in the ECAL within the isolation cone relative to the p_T of the electron. The energy associated with the electron is removed from the calculation. The cut on the relative isolation was chosen by testing different values and viewing the effect on signal efficiency. Figure 6.17 shows the effect of cutting upon differing values of relative isolation on the signal efficiency for Model A as well as for the single electron background, from Z bosons decaying to electrons MC. The efficiency for the signal remains unchanged between the different values however, the background efficiencies vary.

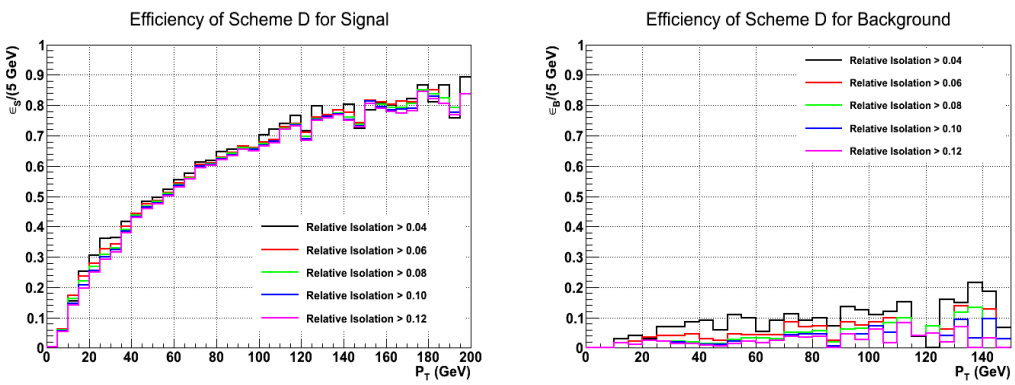

Figure 6.17: The signal (Model A) and background (Z→ee) efficiencies achieved using differing values of relative isolation as calculated by the amount of energy measured in the ECAL within the isolation cone divided by the p_T of the electron.

The Scheme D selection did not work for the highly boosted model of EJs, Model

D. This was due to the fact that all the electrons from the EJ lay too close to each other and their energies were summed into one electron which is then removed from the calculation, giving the appearance of no energy in the isolation cone. Figure 6.18 shows the efficiency of Scheme D on the signal benchmarks (MC), with the drop off for Model D. The efficiency for the other benchmarks using this selection was higher than Scheme C.

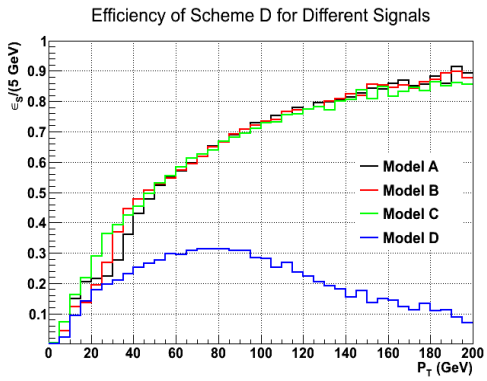

Figure 6.18: The efficiency of the Scheme D identification as a function of jet p_T on the signal benchmarks (MC) showing performance fall off for the boosted benchmark, Model D.

6.9.5 Scheme E

Given the problems Scheme D showed with model dependence and Scheme C's low acceptance for signal events a new selection, Scheme E, was developed using DataViewer. Through use of brushing in conjunction with real time performance information by pruning the dataset as described in Sections 3.4.5 and 3.4.5, a new set of cuts were determined. These cuts give a higher signal event yield predicted for 2011 data with respect to Scheme C, whilst maintaining minimal background. The cuts used were as follows:

- Jet $p_T > 15$ GeV
- EMF > 0.90,
- Number of tracks in the jet > 2,
- Number of electrons in the jet > 0.

Table 6.6 shows the predicted event yield for Scheme E with 4.83 fb^{-1} of data for signal and background. A higher acceptance was achieved than with Scheme C, producing a higher expected signal event yield using the 2011 data. Note that the event yield corresponds to the entire p_T range and not those in the final selection window. Figure 6.19 shows the performance of Scheme E on the different signal MC, using matching as before, and the background MC. The efficiency of the boosted benchmark, Model D, was recovered and the overall efficiencies were higher than both Schemes C and D for all benchmarks. However, with these less stringent cuts, more background passed the selection.

Figure 6.20 shows the dijet mass plots after running selection for the various signal MC. A higher number of dijets passed selection than with previous schemes. Figure 6.21 shows the dijet mass spectrum for the main backgrounds. Although more events had passed the Z plus jets and $t\bar{t}$ MC, these events occupy the lower dijet mass range. This property could be exploited when probing the dijet mass spectrum by creating a signal region over a particular threshold, e.g. 70 GeV. However, the ZZ background still produced a sharp peak within this signal region.

Model	Number of events passing Scheme E	Acceptance	Number of events expected in 2011 data	
			Signal	Background
ZH115A	5736	8.60%	2.48	13.28
ZH115B	5932	8.89%	2.56	
ZH115C	6657	9.98%	2.88	
ZH115D	7517	11.27%	3.25	
ZH125A	6728	10.09%	1.97	
ZH125B	6655	9.98%	1.95	
ZH125C	7606	11.41%	2.22	
ZH125D	8369	12.55%	2.45	
ZH150A	8763	13.14%	0.52	
ZH150B	8873	13.30%	0.52	
ZH150C	9487	14.23%	0.56	
ZH150D	10085	15.18%	0.60	

Table 6.6: Scheme E acceptance for Z decaying to muons and predicted yield for the signal benchmarks and background.

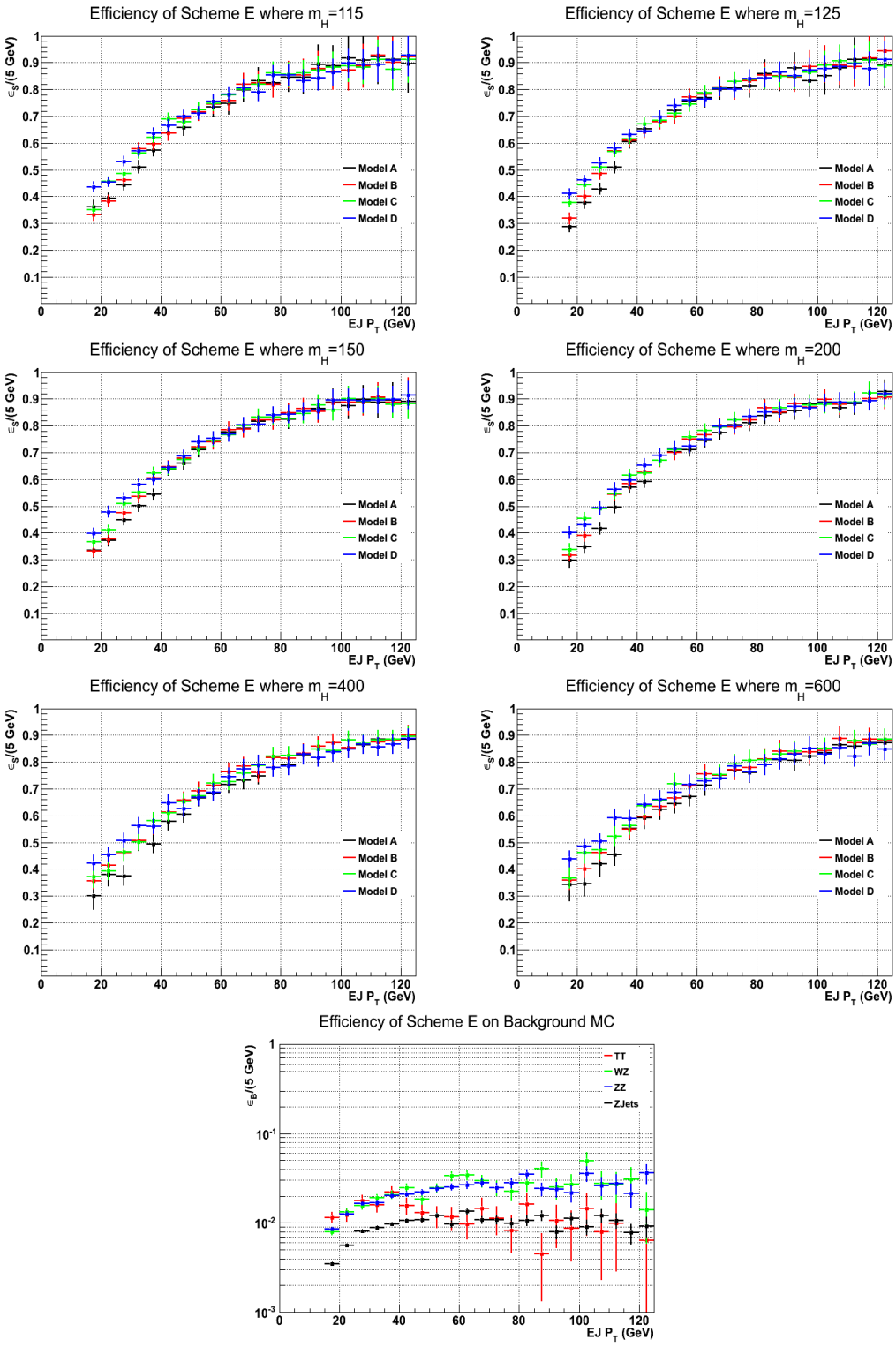

Figure 6.19: Plots showing the efficiency of Scheme E as a function of jet p_T on the background and signal MC for the benchmarks corresponding to different Higgs mass points and dark sector decays. The signal efficiencies are calculated using jets matched to the MC truth.

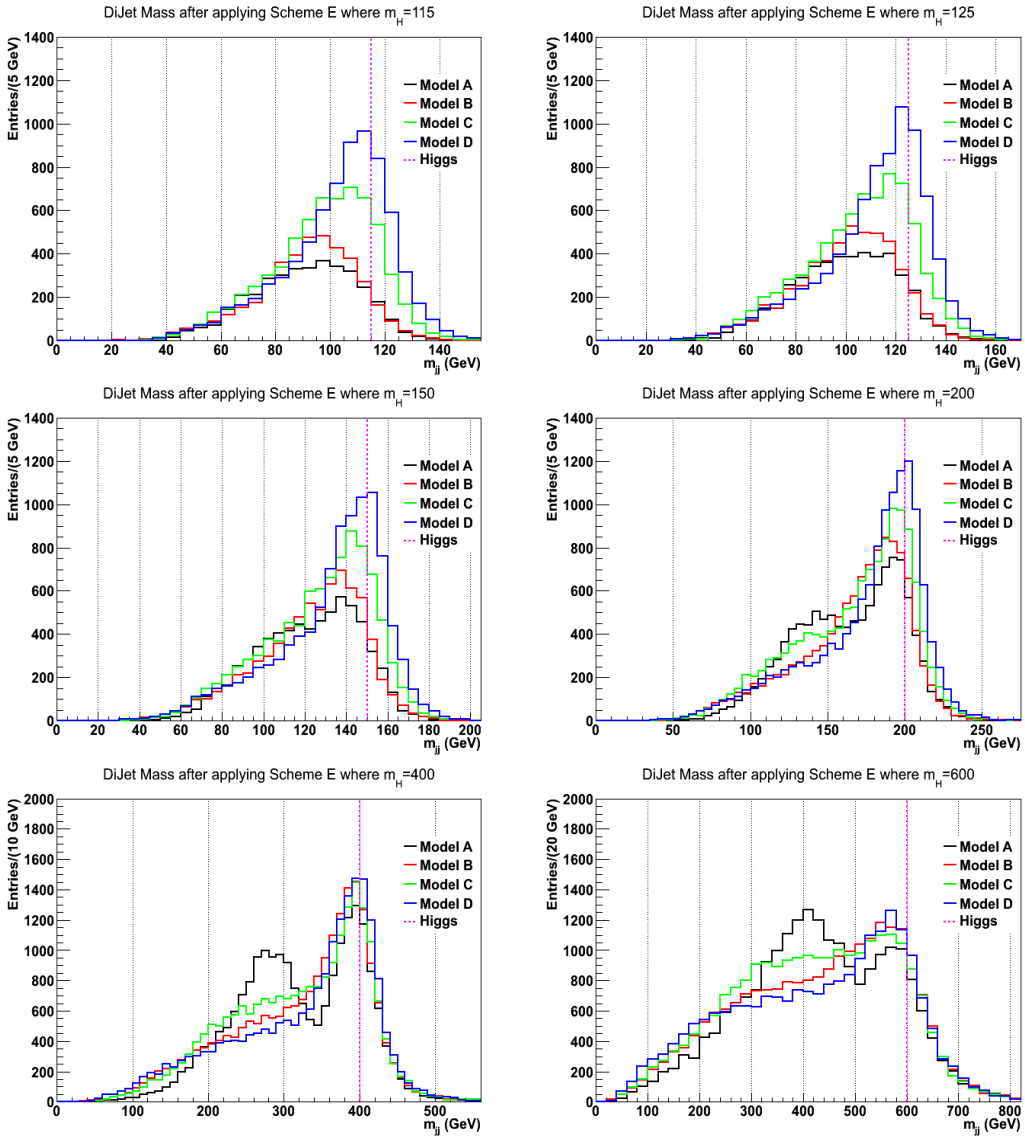

Figure 6.20: Dijet mass spectrum from the signal MC using jets passing Scheme E.

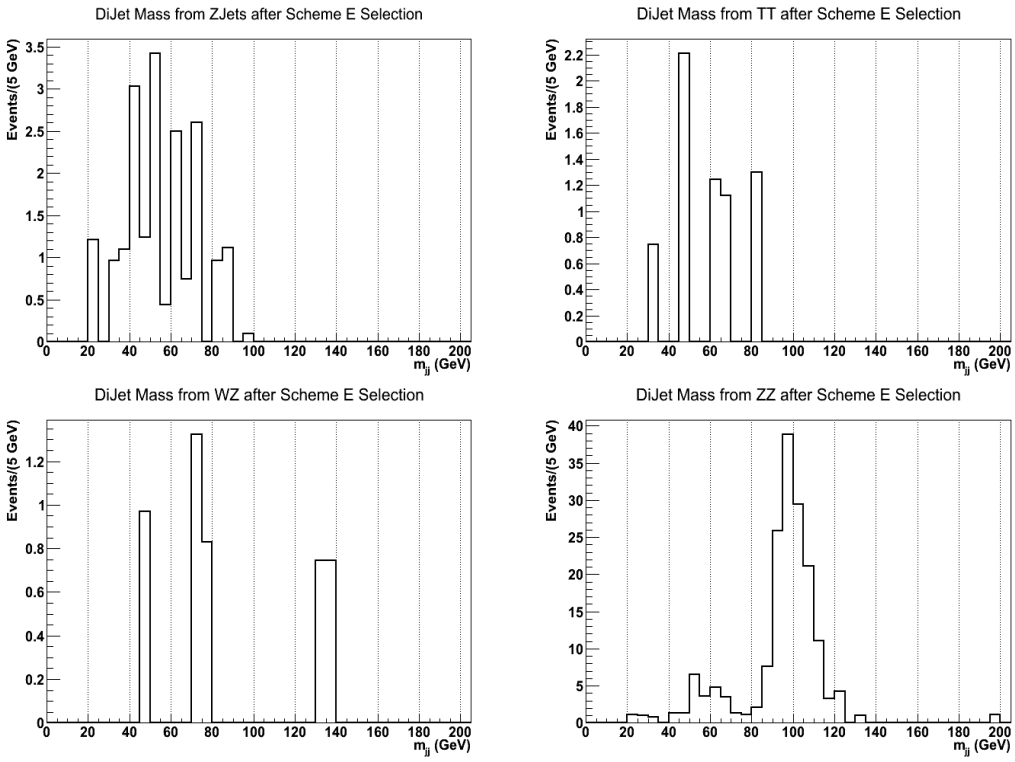

Figure 6.21: Dijet mass spectrum from the background MC using jets passing Scheme E.

6.9.6 Scheme F

While Scheme E produces an acceptable predicted event yield for 2011 data, a further EJ identification scheme was created using DataViewer. Scheme F paid greater attention to reducing the spike in the dijet mass spectrum of the ZZ background (Figure 6.21).

In order to eliminate the ZZ peak, a cut requiring that at least one jet contains two or more reconstructed electrons was introduced. The use of this cut on both of the EJs in the event would produce too low a performance from the relatively poor electron reconstruction efficiency within the EJ. Instead, a tight/loose selection was developed whereby the tight selection was defined as requiring the jet to contain two or more electrons.

The cuts used for the loose selection were as follows:

- EMF > 0.85,
- Number of tracks in the jet which point to the PV and have a momentum greater than 5 GeV ≥ 1,
- Number of electrons in the jet ≥ 1,
- Jet p_T divided by the number of tracks in the jet which point to the PV $>$ 2.5.

For an event to pass Scheme F selection, at least one jet must pass the tight selection. The other jet may either pass the tight or loose selection.

Table 6.7 shows the predicted event yield for Scheme F with 4.83 fb^{-1} of data for signal and background. Note that the event yield corresponds to the entire p_T range and not those in the final selection window. Figure 6.22 shows the efficiency of the scheme on signal and background. The efficiency for three of the dark sector benchmarks had been increased greatly over the entire jet p_T range and achieving efficiencies close to one for jets with high p_T. However, the boosted benchmark, Model D suffered a loss of the efficiency using Scheme F. This loss is explained by difficulties in reconstructing more than one electron in the boosted

jet using the CMS electron reconstruction algorithms (Figure 6.6). The efficiency of Scheme F was higher on the background compared to previous schemes, but still remained low.

Figure 6.23 shows the dijet mass spectrum from the jets passing Scheme F for the various signal MC. For Models A, B and C the number of dijet candidates passing selection is greater than previous schemes, with a smaller number for Model D. Figure 6.24 shows the dijet candidates from the main backgrounds passing selection. The number of events passing was higher than previous schemes, but as before, the majority of events occupied the lower dijet p_T range. The sharp peak in the ZZ background was eliminated by introducing the requirement of two or more electrons in Scheme F.

Model	Number of events passing Scheme F	Acceptance	Number of events expected in 2011 data	
			Signal	Background
ZH115A	13287	19.92%	5.75	72.92
ZH115B	12619	18.92%	5.46	
ZH115C	13155	19.73%	5.69	
ZH115D	3182	4.77%	1.38	
ZH125A	14330	21.49%	4.19	
ZH125B	13670	20.50%	4.00	
ZH125C	14319	21.47%	4.19	
ZH125D	2984	4.47%	0.87	
ZH150A	16631	24.94%	0.98	
ZH150B	16130	24.19%	0.95	
ZH150C	16384	24.57%	0.97	
ZH150D	2680	4.03%	0.16	

Table 6.7: Scheme F acceptance for Z decaying to muons and predicted yield for the signal benchmarks and background.

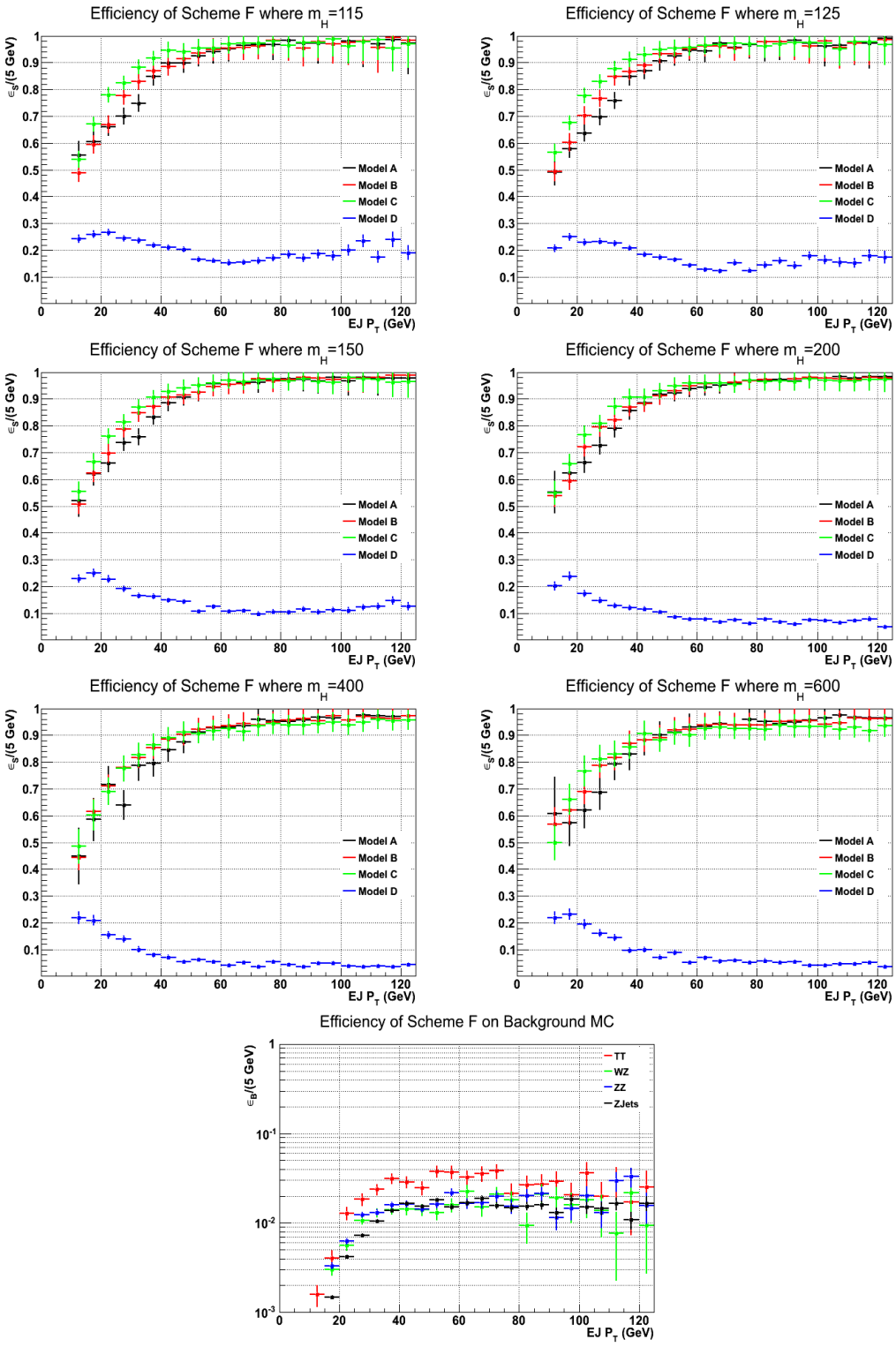

Figure 6.22: Plots showing the efficiency of Scheme F as a function of jet p_T on the background and signal MC for the benchmarks corresponding to different Higgs mass points and dark sector decays. The signal efficiencies are calculated using jets matched to the MC truth.

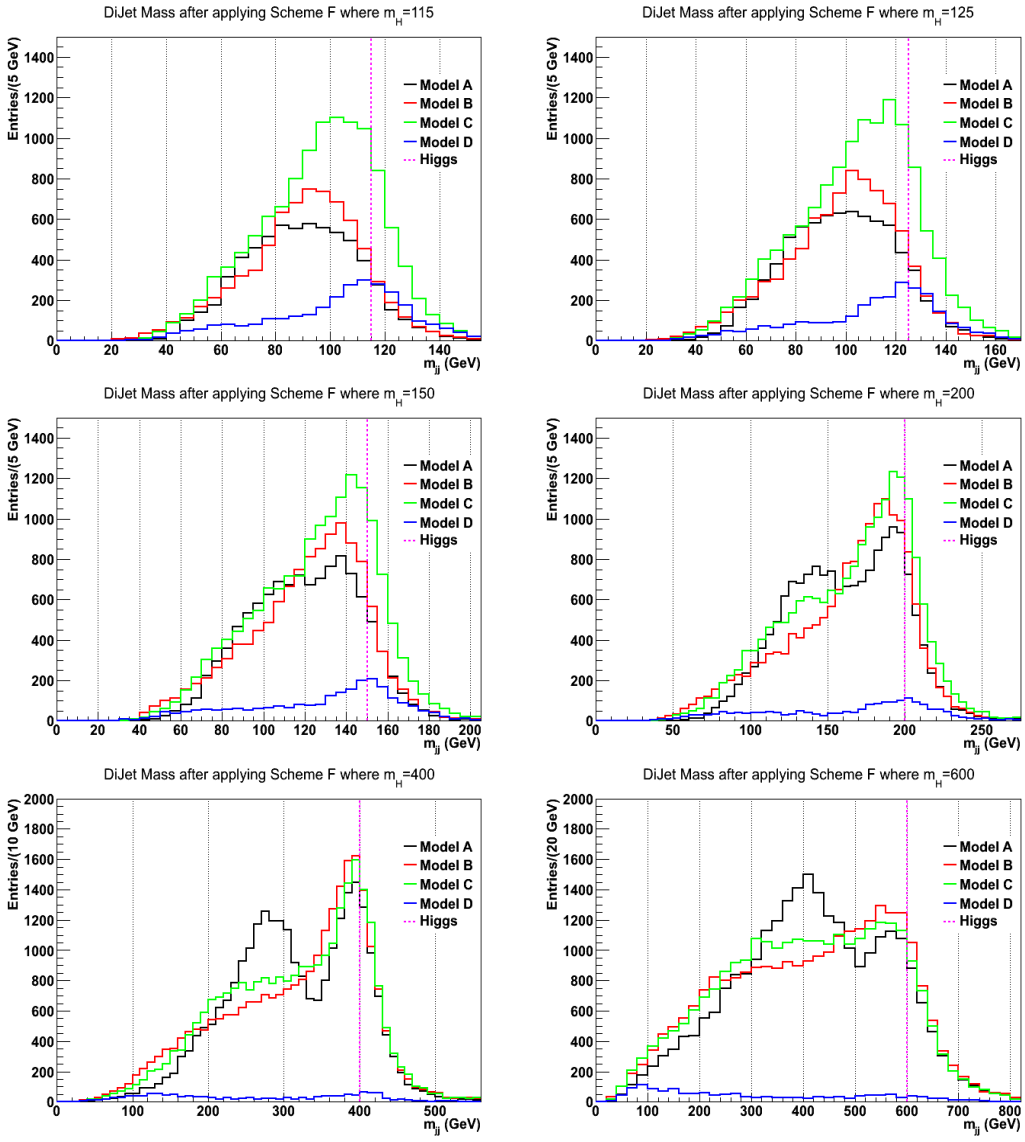

Figure 6.23: Dijet mass spectrum from the signal MC using jets passing Scheme F.

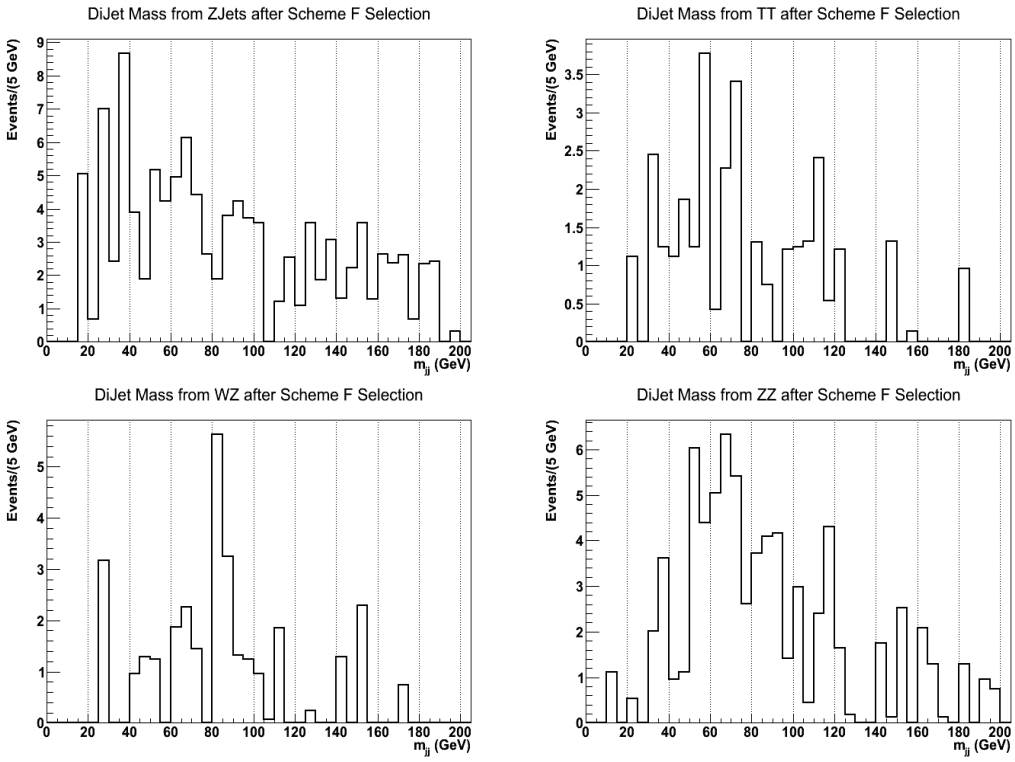

Figure 6.24: Dijet mass spectrum from the background MC using jets passing Scheme F.

6.10 Background Control Samples

To ensure that the MC is a good representation of the data, investigations were undertaken to determine that data matched within selected control regions. The control region selected a sample dominated by Z+Jets. Jets were selected using the same cuts as those used for H→EJ selection allowing assumptions from the control region to be applicable in the H→EJ analysis.

The following control region parameters were required:

- Reconstruct a Z using techniques described in Section 6.7,
- Jets $|\eta| < 2.4$,
- Jets have a $p_T > 10$ GeV, and
- Jets have at least one track pointing to the PV,

The properties of the jet with the highest p_T which passed these criteria were plotted to show differences between the Single Muon dataset and the Z+Jets MC (Figure 6.25). The number of Z candidates found within the mass range 75 to 105 GeV was measured for both MC and data. A greater number of Z candidates were found in the MC than in data. The MC was therefore renormalised by a factor of 0.978 to ensure equivalent numbers of events were present in the comparison plots. The statistical errors for these plots are small, due to the large number of events. However, theoretical errors from PDF and scale variations are dominant; for display purposes the cross section uncertainty for the Z+Jets MC is shown on the plots as a hatched area.

The distribution showing the p_T of the muon with the highest p_T produced from the Z candidate showed good agreement between the data and MC. In general the distributions for the jet properties between MC and data are within the systematic errors, however a greater than expected number of jets with low p_T was present in the data. As the MC had been renormalised using the number of Z candidates, this inconsistency could be due to problems in modelling a high PU scenario.

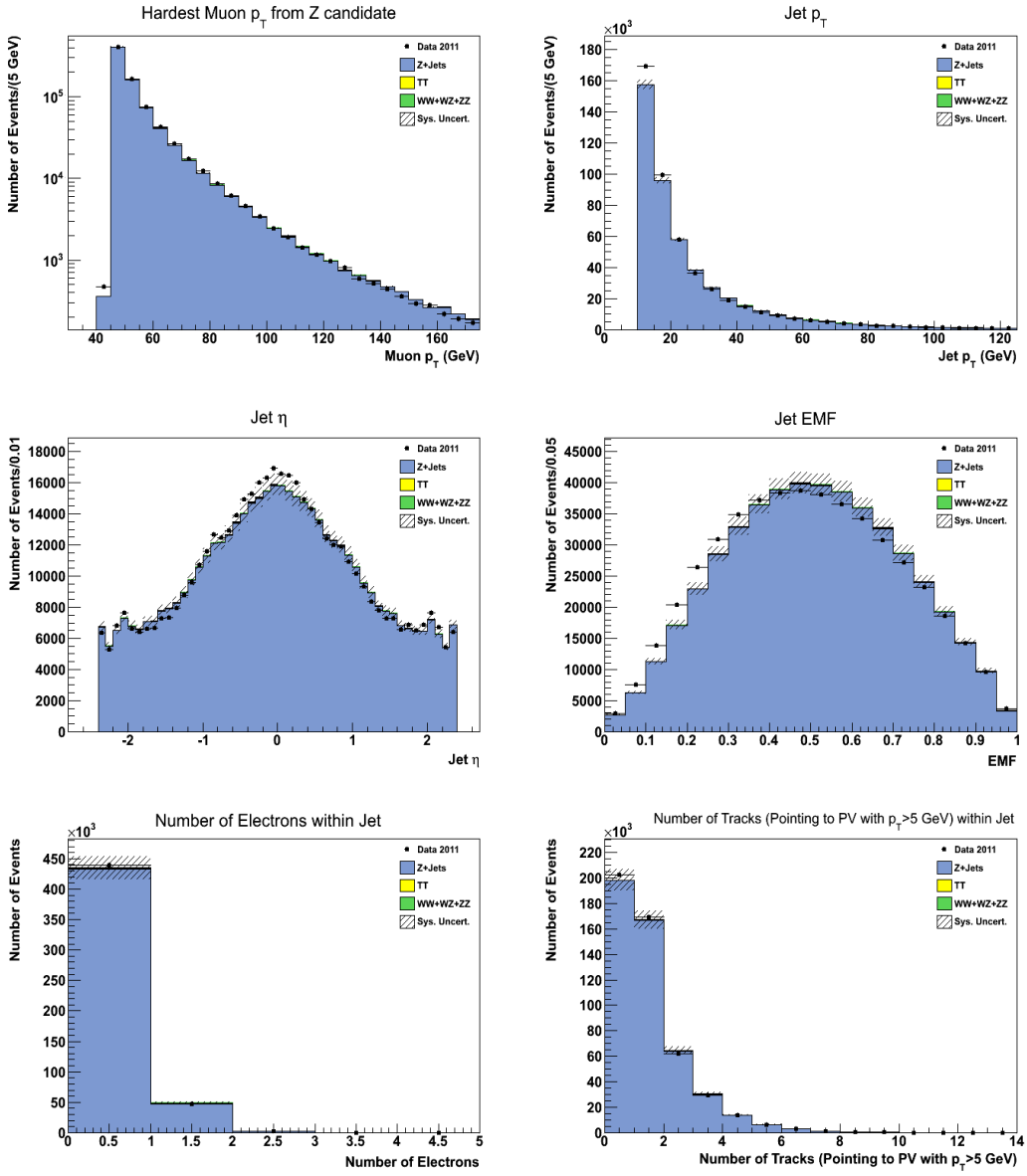

Figure 6.25: Control plots comparing data and the Z+Jets MC. The p_T of the muon with the highest p_T produced from the Z candidate is shown along with different quantities of the jet with the highest p_T which passed the cuts defined in the text. The control region selected Z+Jets with minimal contributions from VV and tt. Single top contributions were negligible and therefore not included in the plots. The uncertainty on the Z+Jets MC cross section is displayed as the systematic uncertainty.

The jet EMF (Figure 6.25) shows a discrepancy between the data and MC distributions with more jets from data containing a lower EMF value compared to MC. Figure 6.26 shows the energies of the jets measured measured in both the ECAL and HCAL, which is used to create the EMF plot. The ECAL energy distribution measured from data tends to be lower than MC.

This effect may be due to pileup effects from previous proton bunch crossings, referred to as 'Out Of Time Pileup'. The response time for the ECAL detector is of the same order as the bunch separation. Therefore when a particle produced from a hard collision is detected by the ECAL, there may still be remnants of a signal from previous particles created in previous collisions. This gives the effect of measuring lower energy in the ECAL than expected. In the MC used for this analysis the out of time pileup simulated only included contributions from the single bunch crossing before and after the event of interest [115]. The previous six bunch crossings need to be included in the simulations in order to take into account the full effects of out of time pileup [116].

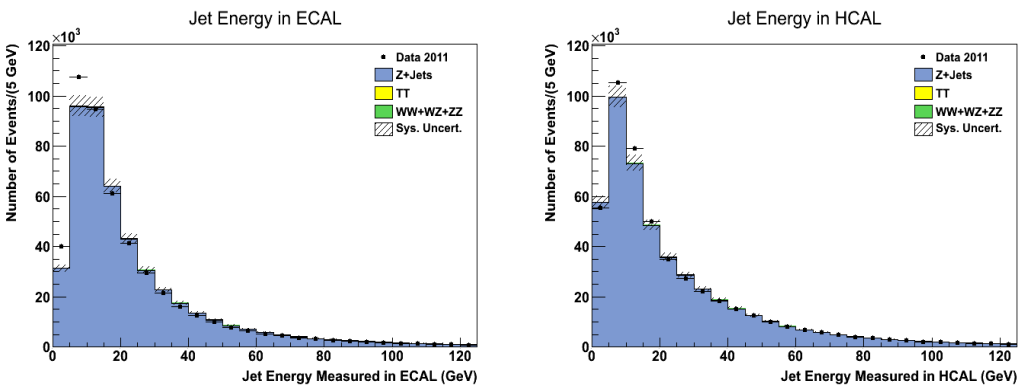

Figure 6.26: ECAL and HCAL jet energy distributions from data and Z+Jets MC.

6.11 Efficiency Scale Factors and Systematics

6.11.1 Pileup and Luminosity Uncertainty

CMS estimates the uncertainty on the measurement of the integrated luminosity used in this analysis to be 2.2% [117]. The uncertainty arising from PU reweighting has previously been measured to be less than 2% [104].

6.11.2 Muon Efficiency

The muon reconstruction and identification efficiency using the 'tight' selection was measured using 2011 data to be 96.4 ± 0.2% and 96.0 ± 0.3% for $|\eta| < 1.2$ and $1.2 < |\eta| < 2.4$ respectively [85]. These efficiencies refer to muons with $p_T > 10$ GeV. The ratio of efficiency measurements between data and simulation produced a scale factor of 0.999 and 0.983 for the two regions which was applied to correct the MC. The tracker-plus-calorimeters (combined) relative isolation efficiency was measured to be > 0.98 for a threshold of 0.15 [85].

The efficiency of the single muon trigger HLT_Mu40 has been measured by CMS using the tag and probe method to be 91.3% for data and 91.9% for MC for the pseudorapidity region $|\eta| < 2.1$. Therefore the scale factor applied to MC for trigger efficiency was 0.9935.

Other CMS analyses have measured the total uncertainty per muon to be approximately 2% using the tag and probe method [118]. An uncertainty value of 2% was also used for the current analysis.

6.11.3 Background Uncertainties

The cross-sections for each background process have uncertainties arising from scale and PDF uncertainties as described in Section 6.3.2. These uncertainties are given in Table 6.3, and have been converted into approximate percentages

with respect to the cross sections for ease in the following table, Table 6.8.

Process	Uncertainty
Z+Jets ($m_{ll} > 50$)	4.3%
$t\bar{t}$	15.5%
WW	3.5%
WZ	3.8%
ZZ	2.5%
Single Top (t-ch)	5.6%
Single Top (t-ch)	4.5%
Single Top (tW)	7.5%
Single Top (tW)	7.5%
Single Top (s-ch)	4.0%
Single Top (s-ch)	6.0%

Table 6.8: The background MC Samples used in this analysis along with the errors associated with them calculated as percentages of the total cross section.

6.11.4 Cut Efficiency and Jet Uncertainty

Efficiencies for the selection cuts, along with their statistical errors, are presented in Table 6.9. The signal MC for Model A is shown with $m_H = 115$ GeV. The Z+Jets MC is shown as the background MC. The statistical errors were calculated as follows:

$$\sigma_\epsilon = \sqrt{\frac{\epsilon(1-\epsilon)}{N_T}}. \tag{6.2}$$

N_T is the total number of events processed and ϵ is the efficiency.

To test the reliability of the variables of interest, the difference between MC and data was measured by counting the number of events remaining after applying each cut defined in Section 6.9. The relative difference is calculated as follows:

$$Rel.Diff. = \left| \frac{(NumberinMC - NumberinData)}{NumberinMC} \right|. \tag{6.3}$$

The number of events after applying the cuts and relative difference between MC and data is outlined in Table 6.10. A general difference of 10% can be seen for all jet attributes with the exception of the number of electrons which is 4%. The 10%

Selection Cut	Efficiency		
	Signal MC	Background MC	Data
Jet Candidate Selection (JCS)	90.1±0.3%	35.7±0.05%	32.1±0.1%
≥1 Jet with ≥2 Electrons	41.5±0.3%	0.01±0.00%	0.02±0.00%
Jet2 passing JCS	45.2±0.3%	1.06±0.02%	1.08±0.02%
Jet contains 1 Electron	78.9±0.4%	18.7±0.7%	18.7±0.8%
Jet number of tracks	91.5±0.3%	70.5±1.8%	77.9±1.8%
Jet pt/number of tracks	96.9±0.2%	86.3±1.6%	83.3±1.9%
Jet EMF	60.8±0.6%	12.3±1.6%	11.6±1.8%

Table 6.9: Table of selection efficiencies upon signal MC, background MC and data. Where each efficiency is derived from events passing all previous cuts. Model A signal MC is shown; m_H = 115 GeV and the Z+Jets were used as background MC.

(4%) difference between MC and data was used as an estimate of the uncertainty arising from the modelling of jet properties (number of electrons reconstructed inside the EJ).

Variable	Number in MC	Number in Data	Rel. Diff.
EMF	2540	2790	0.099
Number of Electrons	1070	1030	0.039
Number of Tracks	49700	45600	0.082
Number of Tracks (p_T >5 GeV)	41900	38800	0.075
Jet p_T/Number of Tracks	48300	44300	0.084
Jet p_T	45900	42000	0.085
Jet η	25900	23800	0.080

Table 6.10: Table showing the number of events remaining in data and MC after applying the cuts defined in the text, along with the relative difference between the two.

By varying the cuts used in the Jet Candidate Selection (JCS) and in Scheme F, an estimate on the stability of the jet properties can be calculated. The modified selection is as follows:

- For JCS the jets have $|\eta| < 2.1$ a $p_T > 15$ GeV and at least two track pointing to the PV.

Variable	Number in MC	Number in Data	Rel. Diff.
EMF	1920	2150	0.120
Number of Tracks	46900	43100	0.081
Number of Tracks (p_T >5 GeV)	30200	28000	0.073
Jet p_T/Number of Tracks	47000	43100	0.084
Jet p_T	40300	36900	0.085
Jet η	19000	17300	0.088

Table 6.11: Table showing the number of events remaining in data and MC after applying the modified cuts defined in the text, along with the relative difference between the two.

- For Scheme F loose selection the jets must have an EMF > 0.8, at least two tracks pointing to the PV with a $p_T > 5$ GeV, p_T over number of tracks pointing to the PV > 2.0 and containing at least one electron.

The standard jet energy scale and resolution uncertainty applied to jets arising from the hadronisation of quarks and gluons could not be applied in this analysis as the jets of interest contain electrons. As a benchmark, the modified p_T cut was varied by 5 GeV from the original value, corresponding to 50% of the original value. The p_T over number of tracks was varied by 20%. The pseudorapidity variable was varied from $|\eta| < 2.4$ to $|\eta| < 2.1$ in order to reduce contributions from the forward regions. The EMF cut was varied by 6%. The cut on the number of tracks was varied by 1 as this is the smallest variation possible. The number of electrons required in the jet for the modified selection remained unchanged as varying this cut would have produced drastic changes in the selection behaviour.

Table 6.11 shows the number of events after applying the modified selection and relative difference between MC and data. These values can be compared to those given from the original cuts, as shown in Table 6.10. All variables except for the EMF have a relative difference within 1% of the original cuts. The EMF has differed by 2% following a small change to the cut, which is likely due to difference in the data and MC distributions shown in Figure 6.25. This 2% is taken as an estimate of the uncertainty from this variable.

The total systematic uncertainty on the signal was estimated to be similar to that in the VV background channel at 12%. The background uncertainty varied

between processes, with the total from Z+Jets estimated to be 13%, $t\bar{t}$ to be 20%, VV to be 12% and ST to be 13%. The total uncertainties, including both statistical and systematic, are shown along with the results in Tables 6.13 to 6.16.

6.12 Results

The following selection was applied to the MC samples in addition to the 2011 dataset corresponding to 4.83 fb^{-1}:

- The muon trigger fired,
- At least one muon in the event within $|\eta| < 2.1$ has a $p_T > 45$ GeV,
- At least one PV was found and the event contains no more than 31 PV,
- A Z candidate was found using techniques described in Section 6.7,
- Jets are preselected by requiring $|\eta| < 2.4$, $p_T > 10$ GeV with at more than one track pointing to the PV,
- Scheme F selection is applied to these jets,
- The dijet invariant mass is then calculated from the two hardest jets passing selection, and finally
- A cut on the dijet mass is applied and the remaining number of events counted.

Figure 6.27 shows the resulting dijet invariant mass spectrum after running the full analysis with the Scheme F selection for data (points) and MC (histograms). The signal events passing selection (lines) have been added to the (filled) background MC. Figure 6.28 shows the data points with the background MC only. Errors due to statistical uncertainty on the MC are also shown.

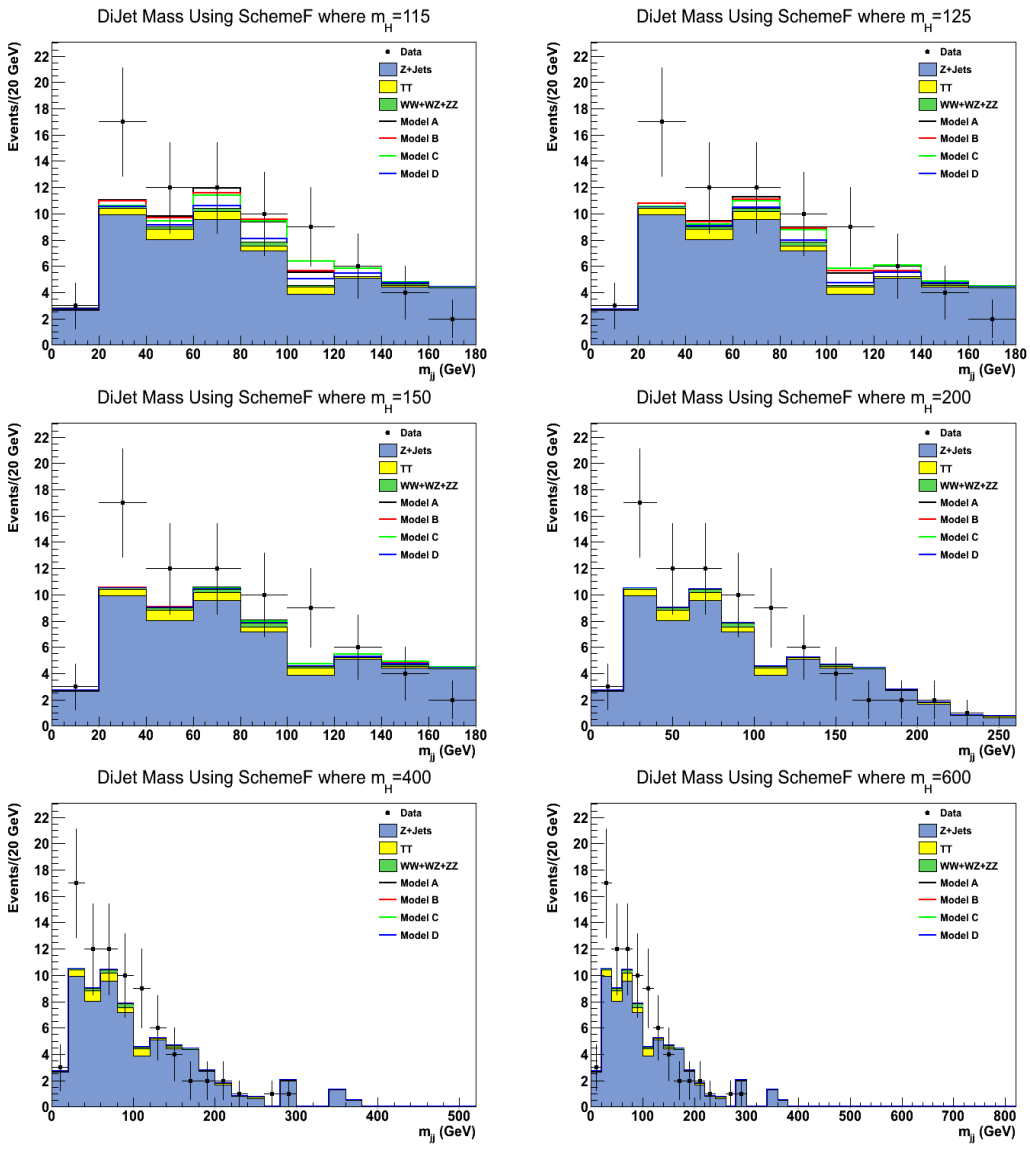

Figure 6.27: Plots showing the dijet invariant mass spectrum after running the full analysis with the Scheme F selection. Estimated signal yield is superimposed on top of the backgrounds. Each plot shows a different value of m_H with the 4 dark sector benchmarks.

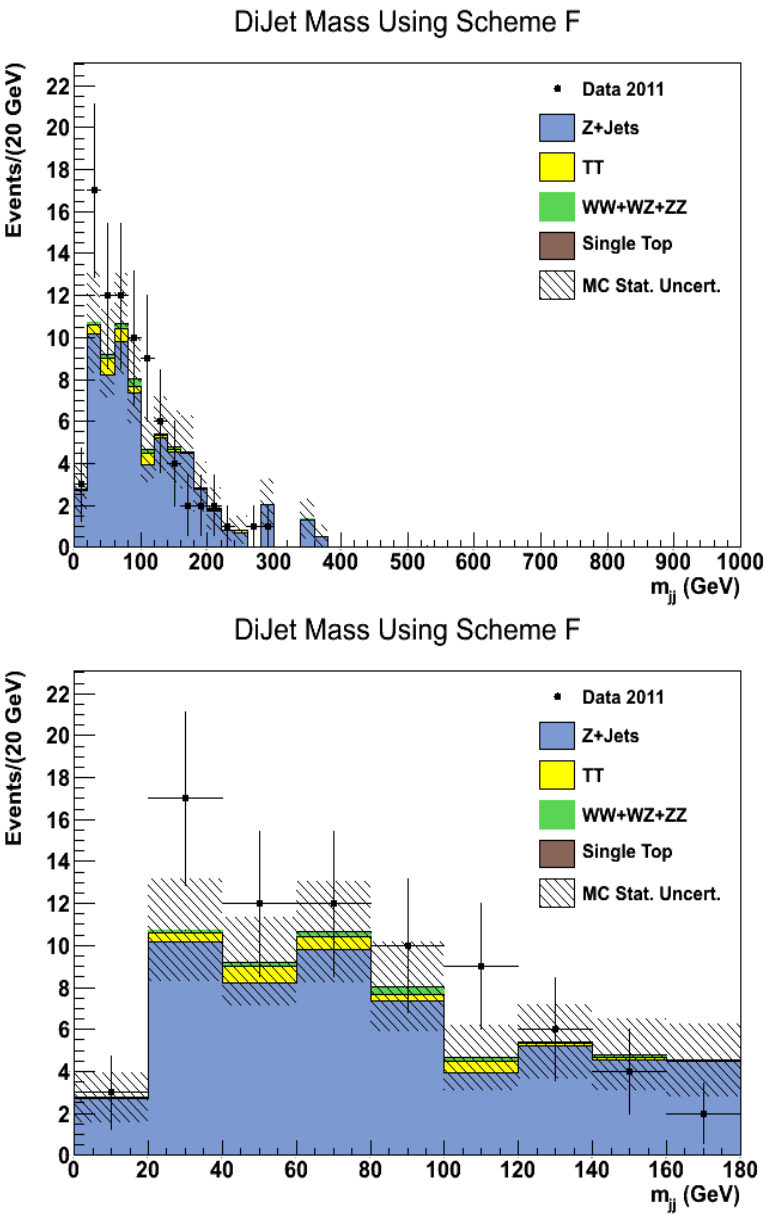

Figure 6.28: The dijet invariant mass spectrum after running the full analysis with the Scheme F selection showing the background only along with statistical uncertainty on both data and MC. The second plot highlights the mass range where any signal is more likely to appear.

6.12.1 Signal and Background Estimates

The cut applied to the dijet mass was optimised for each mass and benchmark using the signal and background MC. This cut produced a 'signal window' which maximised the number of signal events divided by the square root of the number of background events. Table 6.12 shows the values used which define the signal window in the dijet mass spectrum. Tables 6.13, 6.14, 6.15 and 6.16 give the predicted signal and background yields along with their total uncertainty for each of the different mass and model benchmarks as well as the number of observed events found in the 2011 data.

Model	Window Start	Window End	Model	Window Start	Window End
115 A	40	120	200 A	120	230
115 B	70	120	200 B	140	230
115 C	80	130	200 C	140	230
115 D	80	140	200 D	150	260
125 A	70	130	400 A	250	500
125 B	70	130	400 B	250	500
125 C	80	140	400 C	250	500
125 D	110	150	400 D	250	500
150 A	80	160	600 A	250	700
150 B	80	160	600 B	250	700
150 C	80	170	600 C	250	700
150 D	110	180	600 D	250	700

Table 6.12: Values used to define the signal window in the dijet mass spectrum, given in GeV, for different benchmarks and values of m_H.

Process	115 GeV	125 GeV	150 GeV	200 GeV	400 GeV	600 GeV
Z+Jets	28.89 ± 5.42	17.62 ± 3.91	20.58 ± 4.36	18.25 ± 4.03	3.81 ± 1.63	3.81 ± 1.63
$t\bar{t}$	2.18 ± 0.65	1.31 ± 0.46	1.12 ± 0.42	0.50 ± 0.25	0.12 ± 0.12	0.12 ± 0.12
VV	0.92 ± 0.25	0.57 ± 0.15	0.58 ± 0.15	0.19 ± 0.08	0.05 ± 0.03	0.10 ± 0.05
ST	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.05	0.05 ± 0.05	0.01 ± 0.01	0.01 ± 0.01
B_{exp}	31.99 ± 5.46	19.50 ± 3.94	22.32 ± 4.38	18.99 ± 4.04	3.99 ± 1.64	4.04 ± 1.64
ZH	4.79 ± 0.58	2.84 ± 0.34	0.75 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
N_{obs}	45	27	30	16	2	2

Table 6.13: The predicted backgrounds including the systematic and statistical errors, signal yields and observed number of events for the different Higgs mass points for Model A.

Process	115 GeV	125 GeV	150 GeV	200 GeV	400 GeV	600 GeV
Z+Jets	15.15 ± 3.53	17.62 ± 3.91	20.58 ± 4.36	13.17 ± 3.28	3.81 ± 1.63	3.81 ± 1.63
$t\bar{t}$	1.14 ± 0.42	1.31 ± 0.46	1.12 ± 0.42	0.33 ± 0.20	0.12 ± 0.12	0.12 ± 0.12
VV	0.55 ± 0.14	0.57 ± 0.15	0.58 ± 0.15	0.17 ± 0.07	0.05 ± 0.03	0.10 ± 0.05
ST	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.05	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
B_{exp}	16.84 ± 3.55	19.50 ± 3.94	22.32 ± 4.38	13.67 ± 3.28	3.99 ± 1.64	4.04 ± 1.64
ZH	3.49 ± 0.42	2.82 ± 0.34	0.73 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
N_{obs}	25	27	30	11	2	2

Table 6.14: The predicted backgrounds including the systematic and statistical errors, signal yields and observed number of events for the different Higgs mass points for Model B.

Process	115 GeV	125 GeV	150 GeV	200 GeV	400 GeV	600 GeV
Z+Jets	13.54 ± 3.32	16.16 ± 3.70	23.22 ± 4.75	13.17 ± 3.28	3.81 ± 1.63	3.81 ± 1.63
$t\bar{t}$	0.97 ± 0.38	0.97 ± 0.38	1.12 ± 0.42	0.33 ± 0.20	0.12 ± 0.12	0.12 ± 0.12
VV	0.48 ± 0.13	0.48 ± 0.13	0.60 ± 0.15	0.17 ± 0.07	0.05 ± 0.03	0.10 ± 0.05
ST	0.00 ± 0.00	0.05 ± 0.05	0.05 ± 0.05	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
B_{exp}	15.00 ± 3.35	17.65 ± 3.72	24.99 ± 4.77	13.67 ± 3.28	3.99 ± 1.64	4.04 ± 1.64
ZH	3.92 ± 0.47	3.13 ± 0.38	0.83 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
N_{obs}	22	25	31	11	2	2

Table 6.15: The predicted backgrounds including the systematic and statistical errors, signal yields and observed number of events for the different Higgs mass points for Model C.

Process	115 GeV	125 GeV	150 GeV	200 GeV	400 GeV	600 GeV
Z+Jets	16.16 ± 3.70	8.93 ± 2.53	15.87 ± 3.68	12.76 ± 3.19	3.81 ± 1.63	3.81 ± 1.63
$t\bar{t}$	0.97 ± 0.38	0.54 ± 0.27	0.55 ± 0.27	0.32 ± 0.19	0.12 ± 0.12	0.12 ± 0.12
VV	0.48 ± 0.13	0.14 ± 0.06	0.24 ± 0.09	0.13 ± 0.06	0.05 ± 0.03	0.10 ± 0.05
ST	0.05 ± 0.05	0.05 ± 0.05	0.05 ± 0.05	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
B_{exp}	17.65 ± 3.72	9.65 ± 2.54	16.71 ± 3.69	13.21 ± 3.20	3.99 ± 1.64	4.04 ± 1.64
ZH	0.99 ± 0.12	0.46 ± 0.06	0.10 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
N_{obs}	25	12	17	10	2	2

Table 6.16: The predicted backgrounds including the systematic and statistical errors, signal yields and observed number of events for the different Higgs mass points for Model D.

6.12.2 Limits

95% Confidence Level (CL) upper limits can be set on the Higgs cross section in the ZH mode with the Higgs decaying through the dark sector into Electron Jets and the Z boson decaying to muons for a dataset corresponding to an integrated luminosity of 4.83 fb^{-1}. Limits were calculated using the CMS RooStatsCL95 package implemented from the RooStats tools in ROOT [119]. The package estimates observed upper limits on the process cross section from a counting experiment along with the corresponding mean expected limit and the one and two standard deviation quantile bands.

The frequentist based CLs criterion [120, 121] was used in order to calculate the limits. Given a mean expected number of events, the probability that the number of events observed in any one experiment matches or exceeds the number that is observed is referred to as a p-value. P-values are computed for pseudoexperiments for both the signal plus background and background only models, in which the test statistic is computed for each. A frequentist approach was used to generate the pseudoexperiments with parameters containing uncertainties, referred to as nuisance parameters, in which the mean values of the nuisance parameters are generated with each pseudoexperiment. The uncertainties described in Section 6.11 were used as nuisance parameters in the calculation.

Table 6.17 summarises the expected and observed upper 95% cross section limits for the different values of m_H and dark sector benchmarks. These results are plotted in Figure 6.29 along with the 1 and 2 σ bands on the expected limits. The observed limit is within 2 σ of the expected limit for all values of m_H and dark sector benchmarks investigated.

Enhancements to the signal cross sections are possible due to dynamic strong couplings [73]. However, no evidence for deviation from the SM is observed.

Model	Expected	Observed
115 A	0.072	0.12
115 B	0.047	0.082
115 C	0.050	0.075
115 D	0.047	0.076
125 A	0.066	0.075
125 B	0.066	0.075
125 C	0.043	0.071
125 D	0.031	0.040
150 A	0.045	0.066
150 B	0.045	0.066
150 C	0.044	0.061
150 D	0.033	0.034

Table 6.17: The expected and observed upper 95% cross section limits in pb, given the predicted number of SM background events from MC, for the different values of m_H and dark sector benchmarks.

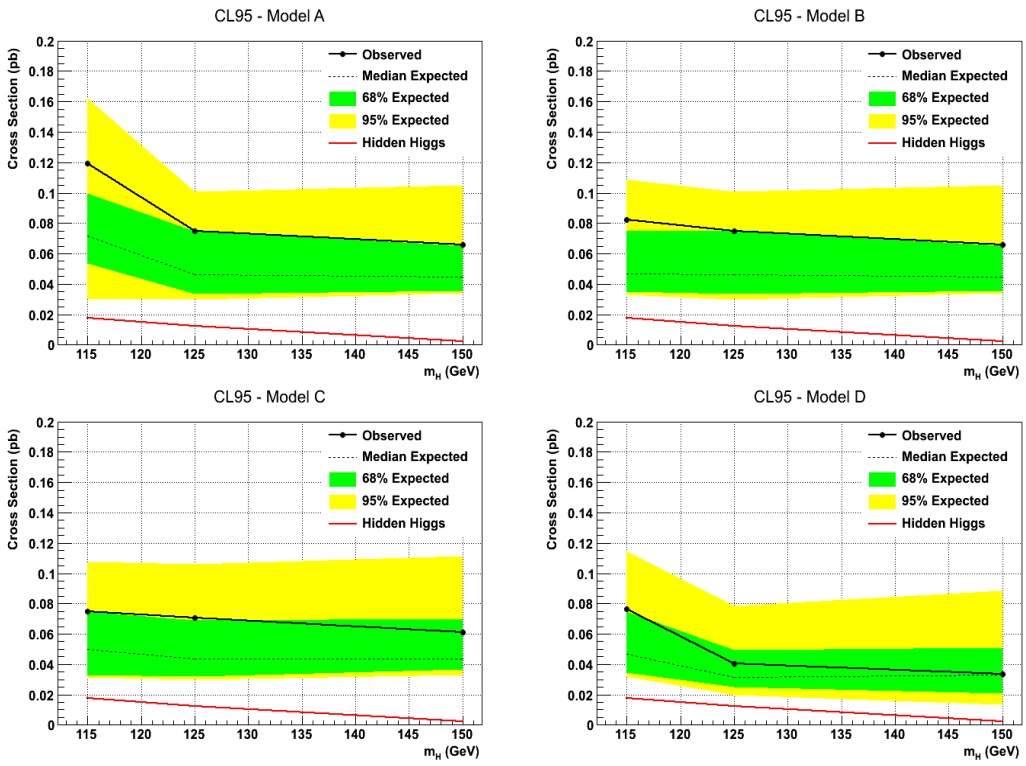

Figure 6.29: Plots showing the expected and observed 95% CL upper limits for the cross section of the associated Higgs production which then decays through the dark sector into Electron Jets for the various different dark sector benchmarks explored. The Z boson decays to muons. The shaded green and yellow areas correspond to the 68% and 95% quantiles for the expected limits respectively. The cross section for the Hidden Higgs process as given in Table 6.2 is also shown.

6.13 Conclusions and Future Work

A search for lepton jets at CMS has been presented. This analysis focused on Electron Jets produced from a Higgs decaying through a dark sector. No evidence for a signal was observed in the 2011 data with an integrated luminosity of 4.83 fb^{-1}. Upper limits at the 95% CL have therefore been set on the production of an associated Higgs decaying into the dark sector with the Z boson decaying to muons.

The primary focus was to search for an EJ signature in association with the Z decaying to muons, which provides the cleanest signature. The addition of other electrons in the same event as the EJ increases the complexity of both reconstruction and trigger efficiency.

Future modifications to this analysis could include other production channels, such as Z decaying to electrons, W decaying to a muon and neutrino and W decaying to an electron and neutrino. Data collected in 2012, corresponding to 20 fb^{-1}, still requires analysis. In 2012 the LHC ran at a higher energy of $\sqrt{s}$ = 8 TeV which led to higher cross sections for the processes of interest. However, even with the increased amount of integrated luminosity recorded and higher cross sections, it is predicted that the search using only the Z to muons decay will not be sensitive enough for a 5 σ observation in the channel studied.

Using the CL95 package it is estimated that 40 fb^{-1} of data is required at $\sqrt{s}$ = 14 TeV to be sensitive to the benchmarks with a Higgs mass of 115 GeV. It is estimated that an acceptance of 10% can be achieved and the systematic error on the background can be reduced to 7%. The ratios of LHC parton luminosities between 7 TeV and 14 TeV as calculated using MSTW2008 (NLO) parton distributions [99] were used along with the cross sections for 7 TeV.

The current analysis focused on a 3 step decay chain, where there are three dark sector particles. Longer decay chains in the dark sector, for example 5 step decays, would produce signatures with higher multiplicities and less missing energy. These longer decay chains would provide different signatures to investigate. The sensitivity to these signatures using the current analysis may be different to the 3 step decay. For example, the electron reconstruction efficiency for an increased

number of electrons within a particular area may decrease. The analysis may therefore have to be revisited.

Chapter 7

Conclusions

7.1 Work Presented

Multivariate visualisation provides many advantages over conventional visualisations. Despite this, such techniques are not well used within particle physics analyses. After detailing a number of interesting multivariate visualisations, including parallel coordinates and the grand tour, a review of current software implementing these techniques has been presented. The need for a new software package was outlined along with the requirements and design of such a package. The new software, DataViewer, was described in detail. An assessment of the program's current capabilities, limits and potential improvements have been discussed.

A theoretical overview of the Standard Model (SM) of particle physics was presented followed by possible extensions to the SM which could explain various astrophysical anomalies observed. A hidden Higgs model which could result from these extensions provides an interesting experimental signature of collimated groups of electrons, called Electron Jets (EJs). An extensive search for an associated Higgs decaying into EJs at the CMS experiment has been presented.

The novel DataViewer multivariate visualisation software package was utilised during the analysis. The software was found to be extremely powerful in cross

checking other techniques, including multivariate classifiers, in addition to suggesting other interesting attributes of the signature that could be quickly exploited. No evidence for a hidden Higgs decaying to EJs was observed using 4.83 fb^{-1} of data collected by CMS in 2011. Upper Limits on the production of an associated Higgs decaying into the dark sector with the Z boson decaying to muons were set at the 95% CL.

DataViewer has attracted attention from other particle physicists who have used the software in their own work [122]. Much interest has also been shown by other scientists during presentations of the software at both international and national conferences, such as the IoP Nuclear and Particle Physics Divisional Conference, as well as academic lectures at the CERN School of Computing.

7.2 Future Prospects

A number of improvements to DataViewer were identified in Section 3.4.8, many of which can be addressed in the immediate future but were outside the scope of this thesis. Further ahead, more studies are required to gather feedback on how other users interact with the software and identify additional improvements. More work is required investigating the marriage of automated and manual visualisations as well as connections between these and existing multivariate analysis techniques. To promote the mainstream use of the software within the particle physics community, it will most likely have to be incorporated with existing tools. The TMVA software suite would be a good candidate for this integration.

The search for an associated Higgs decaying into EJs presented here required events with the Z decaying to muons. It is predicted that 40 fb^{-1} of data at a centre of mass energy of $\sqrt{s} = 14$ TeV is required for the expected signal to be a 5 σ deviation from the SM prediction. As detailed in Section 6.13 there are numerous ways to extend this search such as including other decay modes of the Z, the W channel and adding in the 2012 data collected by CMS. Further refinements to the reconstruction and identification of the EJs are also possible which would improve the limits set.

Bibliography

[1] G. Aad et al. [ATLAS Collaboration]. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, **3**:S08003, 2008.

[2] J. Beringer et al. (Particle Data Group). Review of Particle Physics, 2012-2013. Review of Particle Properties. *Phys. Rev. D*, **86**(1):010001, 2012.

[3] B. H. McCormick. Visualization in Scientific Computing-A Synopsis. *Computer Graphics and Applications, IEEE*, **7**(7):61–70, July 1987.

[4] D. Hirschbuehl [CDF Collaboration]. Search for single top production using multivariate analyses at CDF, 2007.

[5] V. Abazov et al. [D0 Collaboration]. Multivariate searches for single top quark production with the D0 detector. *Physical Review D*, **75**:092007, 2007.

[6] A. Hocker et al. TMVA - Toolkit for Multivariate Data Analysis. *POSACAT*, **040**, 2007.

[7] R. Brun and F. Rademakers. ROOT - An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **389**(12):81 – 86, 1997. New Computing Techniques in Physics Research V.

[8] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. *Visualization, 1991. Visualization '91, Proceedings., IEEE Conference on*, pages 156–163, 419, Oct 1991.

[9] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional Visualizations. *in Proceedings of the Visual Data Mining workshop, KDD*, 2001.

[10] D. F. Swayne, D. Temple Lang, A. Buja, and D. Cook. GGobi: XGobi Redesigned and Extended. In *33th Symposium on the Interface: Computing Science and Statistics*, 2001.

[11] P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, NPIVM '99, pages 9–16, Kansas City, Missouri, USA, 1999. ACM.

[12] J. Sharko, G. Grinstein, and K. A. Marx. Vectorized Radviz and Its Application to Multiple Cluster Datasets. *Visualization and Computer Graphics, IEEE Transactions on*, **14**(6):1444 –1427, nov.-dec. 2008.

[13] Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Iris/.

[14] M. d'Ocagne. *Coordonnées parallèles et axiales: Méthode de Transformation géométriqueet Procédé nouveau de Calcul graphique, déduits de la Considération des Coordonnées parallèles.* 1885.

[15] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, **1**:69–91, 1985.

[16] E. J. Wegman. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, **85**:664–675, 1990.

[17] B. L. Pham and Y. Cai. Visualization techniques for tongue analysis in traditional Chinese medicine. In R. L. Galloway, Jr., editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5367 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 171–180, May 2004.

[18] A. F. X. Wilhelm, E. J. Wegman, and J. Symanzik. Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited, 1999.

[19] E. J. Wegman and J. L. Solka. On Some Mathematics for Visualizing High Dimensional Data. *The Indian Journal of Statistics, Selected Articles from San Antonio Conference in Honour of C. R. Rao*, **64**(2):429–452, 2002. http://www.jstor.org/stable/25051404.

[20] A. Inselberg. Don't panic ... do it in parallel! *Computational Statistics*, **14**:53–77, 1999.

[21] E. J. Wegman and R. E. A. Moustafa. On Some Generalizations of Parallel Coordinate Plots, 2002.

[22] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. PRIM-9, an interactive multidimensional data display and analysis system, 1974. Sound film, 25 minutes. Bin-88 Productions, Stanford Linear Accelerator Center.

[23] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *Computers, IEEE Transactions on*, **C-23**(9):881–890, 1974.

[24] D. Asimov. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM J. Sci. Stat. Comput.*, **6**(1):128–143, January 1985.

[25] A. Buja and D. Asimov. Grand tour methods: an outline. In *Proceedings of the Seventeenth Symposium on the interface of computer sciences and statistics on Computer science and statistics*, pages 63–67, Lexington, Kentucky, United States, 1986. Elsevier North-Holland, Inc.

[26] A. Buja, D. Cook, D. Asimov, and C. Hurley. Computational Methods for High-Dimensional Rotations in Data Visualization. **24**:391 – 413, 2005.

[27] E. J. Wegman. The Grand Tour in k-Dimensions. In *Computing Science and Statistics. Statistics of Many Parameters: Curves, Images, Spatial Models. Proc. 22nd Symposium on the Interface*, pages 127–136. Springer-Verlag, New York, 1992.

[28] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, **43**(4):423 – 444, 2003. Data Visualization.

[29] E. J. Wegman. Visual data mining. *Statistics in Medicine*, **22**(9):1383–1397, 2003.

[30] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In Lyle Ungar, Mark

Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, Philadelphia, PA, USA, August 2006. ACM.

[31] M. Hall et al. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**(1):10–18, November 2009.

[32] T. Curk et al. Microarray data mining with visual programming. *Bioinformatics*, **21**:396–398, February 2005.

[33] MATLAB. *version 8*. The MathWorks Inc., Natick, Massachusetts, United States.

[34] M. O. Ward. XmdvTool: integrating multiple methods for visualizing multivariate data. In *Proceedings of the conference on Visualization '94*, VIS '94, pages 326–333, Washinton, D.C., 1994. IEEE Computer Society Press.

[35] D. Foulser. IRIS Explorer: a framework for investigation. *SIGGRAPH Comput. Graph.*, **29**(2):13–16, May 1995.

[36] D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive Dynamic Graphics In The X Window System With A Link To S, 1992.

[37] D. Cook, A. Buja, and J. Cabrera. Projection Pursuit Indices Based On Orthonormal Function Expansions. *Journal of Computational and Graphical Statistics*, **2 (3)**:225–250, 1993.

[38] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society*, **B 10**:159–203, 1948.

[39] O. Couet. Multiple variables data sets visualization in ROOT. *J. Phys. Conf. Ser.*, **119**:042007, 2008.

[40] R. Kohavi and F. Provost. Glossary of Terms. *Machine Learning*, **30**(2-3):271–274, 1998.

[41] J. Blanchette and M. Summerfield. *C++ GUI Programming with Qt 4*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2006.

[42] D. W. Scott. On optimal and data-based histograms. *Biometrika*, **66**(3):605–610, 1979.

[43] Gnome 2.18 icon theme by AMAZIGH Aneglus : http://art.gnome.org/themes/icon.

[44] G. Gabrielse, D. Hanneke, T. Kinoshita, M. Nio, and B. Odom. Erratum: New Determination of the Fine Structure Constant from the Electron g Value and QED [Phys. Rev. Lett. 97, 030802 (2006)]. *Phys. Rev. Lett.*, **99**:039902, Jul 2007.

[45] P. Clade et al. Determination of the Fine Structure Constant Based on Bloch Oscillations of Ultracold Atoms in a Vertical Optical Lattice. *Phys. Rev. Lett.*, **96**:033001, Jan 2006.

[46] S. Glashow, A. Salam, and S. Weinberg. S. Glashow, Nucl. Phys. 22 (1961) 579; S. Weinberg, Phys. Rev. Lett. 19 (1967) 1264; A. Salam, in "Elementary Particle Theory", ed. N. Svartholm, Almqvist and Wiksells, Stockholm (1969) p. 367.

[47] A. Djouadi. The Anatomy of Electro-Weak Symmetry Breaking. I: The Higgs boson in the Standard Model. 2005. hep-ph/0503172.

[48] M. Gell-Mann, Phys. Lett. 8 (1964) 214; G. Zweig, CERN-Report 8182/TH401 (1964); H. Fritzsch, M. Gell-Mann and H. Leutwyler, Phys. Lett. B47 (1973) 365; D. Gross and F. Wilczek, Phys. Rev. Lett. 30 (1973) 1343; H. D. Politzer, Phys. Rev. Lett. 30 (1973) 1346; G. 't Hooft, Marseille Conference on Yang-Mills fields (1972).

[49] C. P. Yuan. Top quark and electroweak symmetry breaking mechanism. 1998. hep-ph/9809536.

[50] Q. R. Ahmad et al. [SNO Collaboration]. Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory. *Phys.Rev.Lett.*, **89**:011301, 2002. nucl-ex/0204008.

[51] Y. Ashie et al. [Super-Kamiokande Collaboration]. Evidence for an oscillatory signature in atmospheric neutrino oscillation. *Phys.Rev.Lett.*, **93**:101801, 2004. hep-ex/0404034.

[52] T. Araki et al. [KamLAND Collaboration]. Measurement of neutrino oscillation with KamLAND: Evidence of spectral distortion. *Phys.Rev.Lett.*, **94**:081801, 2005. hep-ex/0406035.

[53] W. M. Alberico and S. M. Bilenky. Neutrino oscillations, masses and mixing. *Phys.Part.Nucl.*, **35**:297–323, 2004. hep-ph/0306239.

[54] N. Jarosik, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, et al. Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results. *Astrophys.J.Suppl.*, **192**:14, 2011. 1001.4744.

[55] N. Arkani-Hamed, S. Dimopoulos, and G. R. Dvali. The Hierarchy problem and new dimensions at a millimeter. *Phys.Lett.*, **B429**:263–272, 1998. hep-ph/9803315.

[56] I. Aitchison. *Supersymmetry in Particle Physics*. Cambridge University Press, 2007.

[57] T. Teubner. The Standard Model. In *Proceedings of the School for Experimental High Energy Physics Students*, pages 115–209, RAL, 2009.

[58] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, **13**:321–323, Aug 1964.

[59] P. W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys.Lett.*, **12**:132–133, 1964.

[60] B. R. Martin and G. Shaw. Particle Physics (2nd Ed). UK: Wiley Blackwell, 1997.

[61] G. Aad et. al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, **716**(1):1 – 29, 2012.

[62] S. Chatrchyan et. al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, **716**(1):30 – 61, 2012.

[63] S. Dittmaier et al. [LHC Higgs Cross Section Working Group Collaboration]. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. 2011. 1101.0593.

[64] M. J. Strassler and K. M. Zurek. Echoes of a hidden valley at hadron colliders. *Phys.Lett.*, **B651**:374–379, 2007. hep-ph/0604261.

[65] M. J. Strassler. Possible effects of a hidden valley on supersymmetric phenomenology. 2006. hep-ph/0607160.

[66] N. Arkani-Hamed and N. Weiner. LHC Signals for a SuperUnified Theory of Dark Matter. *JHEP*, **0812**:104, 2008. 0810.0714.

[67] C. Cheung, J. T. Ruderman, L. Wang, and I. Yavin. Lepton Jets in (Supersymmetric) Electroweak Processes. *JHEP*, **1004**:116, 2010. 0909.0290.

[68] A. Falkowski, J. T. Ruderman, T. Volansky, and J. Zupan. Hidden Higgs Decaying to Lepton Jets. *JHEP*, **1005**:077, 2010. 1002.2952.

[69] O. Adriani et al. [PAMELA Collaboration]. An anomalous positron abundance in cosmic rays with energies 1.5-100 GeV. *Nature*, **458**:607–609, 2009. 0810.4995.

[70] M. Ackermann et al. [Fermi LAT Collaboration]. Measurement of separate cosmic-ray electron and positron spectra with the Fermi Large Area Telescope. *Phys.Rev.Lett.*, **108**:011103, 2012. 1109.0521.

[71] M. Aguilar et al. [AMS Collaboration]. First Result from the Alpha Magnetic Spectrometer on the International Space Station: Precision Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5˘350 GeV. *Phys. Rev. Lett.*, **110**:141102, Apr 2013.

[72] D. Borla Tridon, P. Colin, L. Cossio, M. Doro, and V. Scalzotto [MAGIC Collaboration]. Measurement of the cosmic electron plus positron spectrum with the MAGIC telescopes. 2011. 1110.4008.

[73] Private communications with S. Mrenna.

[74] A. Falkowski, J. T. Ruderman, T. Volansky, and J. Zupan. Discovering Higgs Decays to Lepton Jets at Hadron Colliders. *Phys.Rev.Lett.*, **105**:241801, 2010. 1007.3496.

[75] L. Evans and P. Bryant. LHC Machine. *Journal of Instrumentation*, **3**(08):S08001, 2008.

[76] R. Bailey and P. Collier. Standard Filling Schemes for Various LHC Operation Modes. Technical Report LHC-PROJECT-NOTE-323, CERN, Geneva, Sep 2003.

[77] LHC Commissioning: http://lhc-commissioning.web.cern.ch/lhc-commissioning/.

[78] Integrated Luminosity delivered by LHC and recorded by CMS: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

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

[80] [CMS Collaboration]. Description and performance of CMS track and PV reconstruction [To be published]. Feb 2012.

[81] CMS ECAL Performance on 2011 data. Feb 2012. CERN-CMS-DP-2012-002.

[82] J. Gaiser. *Charmonium spectroscopy from radiative decays of the J/psi and psi'*. PhD thesis, Calif. Univ. Stanford, Stanford, CA, 1982. SLAC-R-255.

[83] T Tabarelli de Fatis and on behalf of the CMS Collaboration). Role of the CMS Electromagnetic Calorimeter in the hunt for the Higgs boson in the two-gamma channel. *Journal of Physics: Conference Series*, **404**(1):012002, 2012.

[84] The CMS collaboration. Performance of CMS muon reconstruction in pp collision events at s**0.5 = 7 TeV. *Journal of Instrumentation*, **7**:2P, October 2012. 1206.4071.

[85] CMS Collaboration. Performance of CMS muon reconstruction in pp collision events at sqrt(s) = 7 TeV. *ArXiv e-prints*, June 2012. 1206.4071.

[86] C. D. Jones et al. The New CMS Event Data Model and Framework. In *In: CHEP 06: Computing in High Energy and Nuclear Physics*, 2006.

[87] [ATLAS Collaboration]. A Search for Lepton-Jets with Muons at ATLAS. Technical Report ATLAS-CONF-2011-076, CERN, Geneva, May 2011.

[88] T. Aaltonen et al. [CDF Collaboration]. Search for anomalous production of multiple leptons in association with W and Z bosons at CDF. *Phys.Rev.*, **D85**:092001, 2012. 1202.1260.

[89] S. Chatrchyan et al. [CMS Collaboration]. Search for Light Resonances Decaying into Pairs of Muons as a Signal of New Physics. *JHEP*, **1107**:098, 2011. 1106.2375.

[90] V.M. Abazov et al. [D0 Collaboration]. Search for dark photons from supersymmetric hidden valleys. *Phys.Rev.Lett.*, **103**:081802, 2009. 0905.1478.

[91] V. M. Abazov et al. [D0 Collaboration]. Search for events with leptonic jets and missing transverse energy in $\mathbf{p\bar{p}}$ collisions at $\sqrt{\mathbf{s}} = \mathbf{1.96}$ TeV. *Phys.Rev.Lett.*, **105**:211802, 2010. 1008.3356.

[92] Private communications with the Lepton Jet analysis team.

[93] T. Sjostrand, S. Mrenna, and P. Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput.Phys.Commun.*, **178**:852–867, 2008. 0710.3820.

[94] F. Maltoni and T. Stelzer. MadEvent: Automatic event generation with MadGraph. *JHEP*, **0302**:027, 2003. hep-ph/0208156.

[95] S. Frixione, P. Nason, and C. Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, **0711**:070, 2007. 0709.2092.

[96] T. Sjostrand, S. Mrenna, and P. Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, **0605**:026, 2006. hep-ph/0603175.

[97] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush. FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order. *Computer Physics Communications*, **182**(11):2388 – 2403, 2011.

[98] J. M. Campbell and R. K. Ellis. MCFM for the Tevatron and the LHC. *Nucl.Phys.Proc.Suppl.*, **205-206**:10–15, 2010. 1007.3492.

[99] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the LHC. *Eur.Phys.J.*, **C63**:189–285, 2009. 0901.0002.

[100] P. M. Nadolsky et al. Implications of CTEQ global analysis for collider observables. *Phys.Rev.*, **D78**:013004, 2008. 0802.0007.

[101] F. Demartin, S. Forte, E. Mariani, J. Rojo, and A. Vicini. Impact of parton distribution function and alpha s uncertainties on Higgs boson production in gluon fusion at hadron colliders. *Phys. Rev. D*, **82**:014002, Jul 2010.

[102] M. Botje et al. The PDF4LHC Working Group Interim Recommendations. 2011. 1101.0538.

[103] E. Chabanat and N. Estre. Deterministic Annealing for Vertex Finding at CMS. 2005.

[104] S. Chatrchyan et al. [CMS Collaboration]. Search in leptonic channels for heavy resonances decaying to long-lived neutral particles. *JHEP*, **1302**:085, 2013. 1211.2472.

[105] CMS Collaboration. Performance of CMS muon reconstruction in cosmic-ray events. *Journal of Instrumentation*, **5**(03):T03022, 2010.

[106] G. Abbiendi et al. [CMS Collaboration]. *Analysis Note 97*, 2008. Muon Reconstruction in the CMS Detector.

[107] W. Adam et al. [CMS Collaboration]. Electron Reconstruction in CMS. *Analysis Note 164*, 2009.

[108] G L et al. Bayatian. *CMS Physics Technical Design Report Volume I: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006.

[109] G. P. Salam and G. Soyez. A Practical Seedless Infrared-Safe Cone jet algorithm. *JHEP*, **0705**:086, 2007. 0704.0292.

[110] S. Catani, Yuri L. Dokshitzer, M. Olsson, G. Turnock, and B.R. Webber. New clustering algorithm for multi - jet cross-sections in e+ e- annihilation. *Phys.Lett.*, **B269**:432–438, 1991.

[111] S. Catani, Yuri L. Dokshitzer, M.H. Seymour, and B.R. Webber. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions. *Nucl.Phys.*, **B406**:187–224, 1993.

[112] M. Cacciari, G. P. Salam, and G. Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, **0804**:063, 2008. 0802.1189.

[113] M. Cacciari, G. P. Salam, and G. Soyez. FastJet User Manual. *Eur.Phys.J.*, **C72**:1896, 2012. 1111.6097.

[114] C. Buttar, J. D'Hondt, M. Kramer, G. Salam, M. Wobisch, et al. Standard Model Handles and Candles Working Group: Tools and Jets Summary Report. pages 121–214, 2008. 0803.0678.

[115] Private communications with D. Cockerill.

[116] Private communications with D. Petyt.

[117] [CMS Collaboration]. Absolute Calibration of Luminosity Measurement at CMS: Summer 2011 Update. 2011. CMS-PAS-EWK-11-001.

[118] Search for the Standard Model Higgs Boson Decaying to Bottom Quarks and Produced in Association with a W or a Z Boson. Technical Report CMS-PAS-HIG-11-012, CERN, Geneva, 2011.

[119] https://twiki.cern.ch/twiki/bin/viewauth/CMS/RooStatsCl95.

[120] A. L. Read. Presentation of search results: the CLs technique. *J. Phys.*, **G28**:2693, 2002.

[121] T. Junk. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Meth.*, **A434**:435, 1999. 9902006.

[122] R. Page. *Silicon Carbide Foam as a Support Structure for Silicon Sensors in a Vertex Detector*. PhD thesis, University of Bristol, 2012.