2023

Study of Displaced Vertex Tagging with the CMS Experiment

Ethan Todd

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS & SCIENCES

STUDY OF DISPLACED VERTEX TAGGING WITH THE CMS EXPERIMENT

By

ETHAN TODD

A Thesis submitted to the
Department of Physics
in partial fulfillment of the
requirements for graduation with
Honors in the Major

Degree Awarded:
Spring 2023

Ethan Todd defended this thesis on March 22, 2023.

The members of the supervisory committee were:

Ted Kolberg
Thesis Director

Peng Xiong
Committee Member

Sanghyun Lee
Outside Committee Member

Signatures are on file with the Honors Program office.

TABLE OF CONTENTS

Appendix

ABSTRACT

The Standard Model (SM) of particle physics is a very successful yet incomplete theory describing fundamental particles and their interactions. Attempts to solve similar theoretical problems have historically involved the discovery of new particles, motivating the search for particles that could solve the hierarchy problem present in particle physics. Supersymmetry (SUSY) theories posit the existence of partners to the SM particles that differ in spin and are on the order of the W and Z boson masses; the existence of such particles has the potential to solve several of the problems within the SM. More recent theories such as neutral naturalness, i.e. the existence of supersymmetric partners to the SM particles that are not charged under SM QCD, attempt to explain the lack of evidence for SUSY particles at the LHC while still solving problems with the SM. Scalar long-lived particles are predicted by such theories, and are possible to observe using the Compact Muon Solenoid Experiment (CMS) at CERN. Here we present efforts to aid a search for such particles using the novel Regions of Interest (ROI) mechanism. The principle of the ROI technique is to identify displaced vertices directly, rather than identifying displaced objects (i.e. electrons, jets). This is accomplished by identifying displaced pairs of tracks within the CMS tracker, fitting them to a vertex, and forming an artificial region around the vertex. This identification of displaced vertices is beneficial as it allows for the analysis of complex final states, such as the τ of this analysis, without the need to actually reconstruct these final states as physics objects. This ultimately allows for stringent limits to be set on branching ratios to previously ignored final states.

CHAPTER 1

INTRODUCTION

The Standard Model of particle physics (SM) is an extremely successful theory of fundamental particles and their interactions. Despite this, there are phenomena which the SM fails to explain; notably, neutrino masses/oscillations and the existence of dark matter [1]. In addition, particle physics as a whole suffers from a hierarchy problem. In theories where the Higgs mass can be calculated, its value is dependent on quantum corrections which could force its value to be on the order of the Planck scale – the scale at which the forces, including gravity, are unified. However, the observed mass of the Higgs, ~ 125 GeV, is nowhere near the 10^{19} GeV Planck scale. In order to get the calculated mass to match the observed mass, an enormous amount of fine-tuning of parameters is required. This fine-tuning is considered by physicists to be "unnatural", and therefore this issue is also commonly expressed as the naturalness problem.

One potential solution is to posit the existence of new particles, which will in some way "cancel out" the diverging quantum correction terms in calculations of the Higgs mass. The most popular class of such solutions is supersymmetry (SUSY), a group of theories which posits the existence of "superpartners" to the SM particles which are identical in all aspects except for spin. That is, spin-$\frac{1}{2}$ particles have spin-0 superpartners, and spin-1 particles have spin-$\frac{1}{2}$ superpartners, etc. Clearly, the symmetry cannot be exact, or else superpartners with identical mass to their SM counterparts would have already been observed. In most SUSY theories, the masses of the superpartners are instead on the order of the masses of the W and Z bosons [2]. However, the lowest mass SUSY particles should have been observed at the LHC and have not been, an issue which has been called little hierarchy. Theories of neutral naturalness, which posit the existence of a new symmetry relating the SM quarks to partners that are colorless and thus neutral to the SM QCD, resolve the little hierarchy problem by providing a physics excuse as to why the lowest mass SUSY particles have not yet been observed at the LHC.

Such theories would indeed be able to explain the observed mass of the Higgs in a more natural way, but it is not clear how the particles of such theories would be detected. One possibility

is that this connection between the Higgs and these hidden particles would be manifest in exotic decays of the Higgs – particularly in decays of the Higgs to neutral Long-Lived Particles (LLPs) which decay back to SM particles.

Many past searches have focused on final states containing jets; one recent example comes from ATLAS [3]. In this study, decays of the form $\Phi \to SS$ were considered, where Φ is a neutral boson (that could be the Higgs) with mass from 125-1000 GeV, and S is a long-lived particle with mass between 5 GeV and 400 GeV. The analysis searched for two jets with no associated activity in the tracker, and a high ratio of energy deposited in HCAL to energy deposited in ECAL, that appear narrower than prompt (i.e. not displaced) jets when reconstructed. Strong limits were able to be set on branching ratios of previously unstudied combinations of Φ and S masses.

Production Mode	Cross Section (pb)
ggH	48.6 ± 2.8
VBF	3.78 ± 0.08
WH	1.37 ± 0.03
ZH	0.88 ± 0.03

Table 1.1: Various productions modes of the Higgs, listed with their associated cross section at the LHC operating at $\sqrt{s} = 13$ TeV. Data taken from [4].

Of particular relevance to this analysis is another relatively recent result which studied events containing a Higgs produced in conjunction with a Z boson that then decayed to a pair of leptons [5]. This dilepton final state of the Z boson allowed for very clean triggering, as the only major contribution to the background came from the Drell-Yan process[1]. After triggering, displaced jets, i.e. jets with no tracks, were identified according to several tagging variables. The number of displaced jets associated with each event was then used to distinguish signal from background. This ZH analysis was able to set the exclusion limit on the branching ratio of the Higgs to LLP to b and d-quark final states below 1, but the tau final state was not constrained at all due to its complicated decay modes/reconstruction.

As mentioned, the triggering for a Higgs produced in association with a Z boson (ZH) is very clean thanks to the dilepton final state, but the cross section to produce this Higgs/Z pair is ex-

[1] A quark/antiquark annihilate and form a photon/Z boson which then decays to a lepton/antilepton pair.

tremely small. On the other hand, the cross section to produce a Higgs through gluon gluon fusion (ggH) is comparatively very large. Table 1.1 displays the relevant values of these cross sections at the LHC. In the case of the ggH production mode, the trigger strategy becomes much less straightforward. This analysis focuses on the often ignored τ final state, so that the leptonic decay of the τ to a soft muon can be exploited and the B Parking High Level Trigger can be used. Additionally, rather than using displaced jet multiplicity to distinguish signal from background, this analysis uses the novel Regions of Interest (ROI) mechanism.

The Large Hadron Collider (LHC), the Compact Muon Solenoid (CMS) experiment, detector triggering, and data parking are described in the next chapter. In Chapter 3, the signal process and relevant background processes are discussed more thoroughly, and technical details about the data used are provided. Chapter 4 contains a brief description of the machine learning framework used in this analysis. Chapter 5 describes the ROI mechanism and the parameters of the ML model. Chapter 6 presents results of the performance of the ML model and relates them back to the underlying physics. Finally, Chapter 7 contains a discussion of the ROI technique and its potential applications to other physics processes.

CHAPTER 2

THE CMS EXPERIMENT

The LHC is a superconducting pp accelerator and collider with a circumference of 27 km and instantaneous luminosity of $L \sim 10^{34}$ cm^{-2} s^{-1}. Each proton beam has an energy of slightly less than 7 TeV, which means collisions occur with a total center-of-mass energy of $\sqrt{s} = 13$ TeV. Alternating electric fields are used to accelerate bunches of protons, which are then directed by magnetic fields to collide at multiple locations around the LHC. CMS is located at one such point; it consists of several layers of detectors embedded within a superconducting solenoid that produces a magnetic field of ~ 4 T, as well as of muon detection chambers located outside of the solenoid. The detector nearest to the collision point is the silicon tracker, which captures the tracks of charged particles in order to provide precise measurements of momentum. The electromagnetic calorimeter (ECAL) and the hadronic calorimeter (HCAL) are the next two detectors; both are primarily used to measure energy, with the ECAL responsible for electrons/photons and the HCAL responsible for hadrons. HCAL is particularly important, as it is used for the (indirect) detection of neutrinos as well as potentially new physics by identification of missing transverse energy from jets. Finally, the muon detection system is used in combination with the tracker for precise measurements of muon momentum. Figure 2.1 depicts a summary of the CMS particle detection methods, along with the scale of the experiment. A full description of CMS can be found in [6].

2.1 CMS Coordinate System

The origin of the CMS coordinate system is placed at the collision point of the protons inside the experiment. Figure 2.2 depicts the coordinate system; importantly, the z-axis is taken along the beam line. As usual, the azimuthal angle ϕ is taken as the angle from the x-axis in the xy-plane and the polar angle θ is measured from the z-axis. Often the coordinates r and η are used; r is the radial distance in the xy-plane and η, the pseudorapidity, is defined as $\eta = -\ln\tan(\theta/2)$. η is useful because differences in η are Lorentz invariant under boosts along the beam axis.

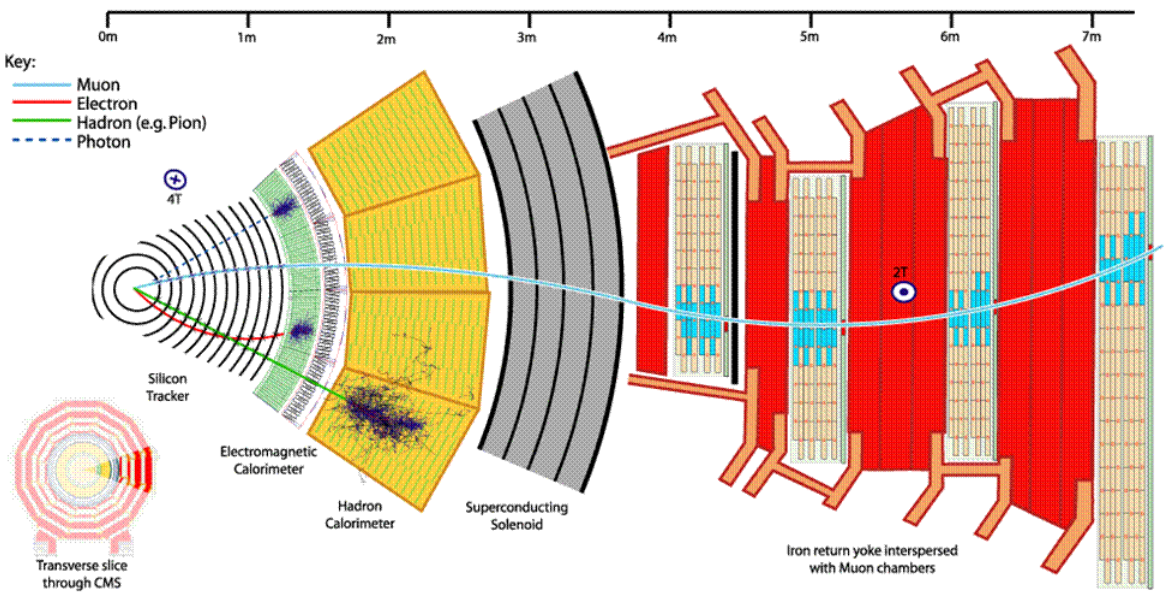


Figure 2.1: A schematic of a slice of CMS, showing the paths of various types of particles through the detector, taken from [7].

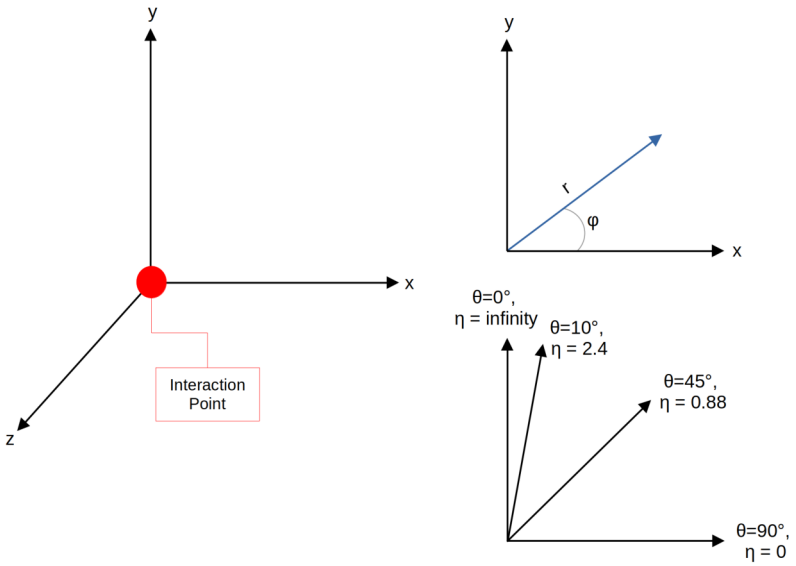


Figure 2.2: A depiction of the coordinate system used in CMS, where the z-axis is taken along the beam line and the x-axis is radially inward towards the center of the collider.

2.2 Triggering at CMS

The LHC collides proton bunches at a rate of up to 40 MHz, but CMS only stores roughly 1 kHz of physics events. This is because many of the events are uninteresting as they are well understood according to the SM, and because it would be nearly impossible to produce the computational resources necessary to store all the events. CMS therefore uses triggers to quickly decide which events are interesting and should be stored, and which should be discarded. The first trigger is the Level 1 (L1) trigger, which takes advantage of muon chamber and calorimeter information in order to reduce the output rate to a maximum of 100 kHz. The L1 trigger is hardware-based, and so the output is limited by hardware capabilities. The next trigger is the software-based High Level Trigger (HLT), which starts from the L1 candidate and combines it with tracking data to further select events. For example, the presence of/lack of tracks can be used to identify energy clusters in ECAL that passed the L1 trigger as belonging to electrons/photons respectively, and the data can then be passed through the HLT or discarded depending on the particular situation. After passing the HLT, events are stored for prompt reconstruction, which reduces the rate of collection to 1 kHz.

2.3 Data Parking

Data parking refers to the practice of selecting events at the HLT, skipping prompt reconstruction, and immediately moving them to tape storage, therefore "parking" the data. These events can then remain on tape until there are sufficient available computing resources to reconstruct them. This practice is useful because it allows for more than the standard 1 kHz of physics events (as described in the prior section) to be recorded, which potentially allows for the detection of interesting events which would otherwise have been discarded.

2.4 B Parking

B parking specifically refers to the 2018 collection of over 10 billion events containing a pair of B hadrons. This dataset was triggered by requiring a soft and displaced muon originating from the decay of a B hadron, which means that no requirements were placed on the other B in the pair. A main motivation for B parking was to collect an adequate sample of $B^0 \to K^* e^+ e^-$, in order to compare the branching fractions of $B^0 \to K^* e^+ e^-$ and $B^0 \to K^* \mu^+ \mu^-$, and to therefore test lepton universality. However, as will be elaborated on in the following section, this B parking dataset is also appropriate for our analysis, because of the selection involving the soft muon final state with a moderate displacement from the primary vertex.

CHAPTER 3

SIGNAL AND BACKGROUND

As mentioned in the introduction, the process of interest to this analysis is the decay of a Higgs produced through gluon-gluon fusion to two scalar long-lived particles, which each then decay to a $\tau\bar{\tau}$ pair (shown in Figure 3.1). The tau final state has frequently been ignored in previous

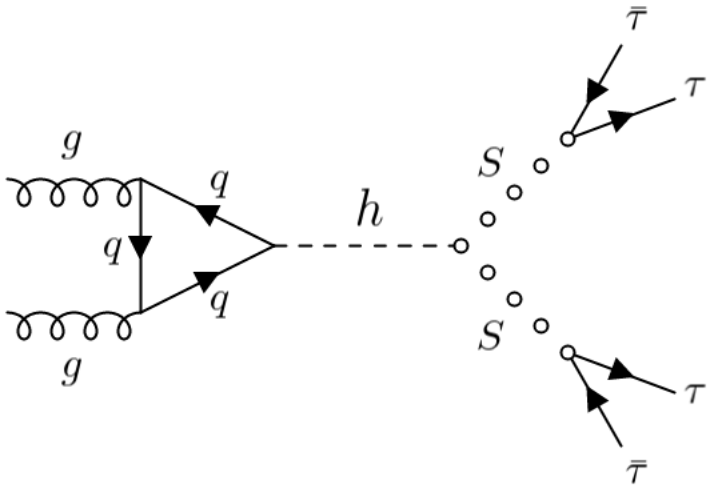


Figure 3.1: A cartoon Feynman diagram of ggH $\rightarrow$ SS $\rightarrow \tau\tau\bar{\tau}\bar{\tau}$.

analyses due to its complicated nature. In particular, the tau itself can decay both leptonically and hadronically (see Figure 3.2), which makes a triggering relatively unclear, especially in comparison to the clean dilepton final state of a Z boson.

3.1 B Parking Trigger Efficiency

For our purposes, the leptonic decay is actually a benefit, as it includes a decay to muons with $\sim 17\%$ probability. Such muons have low pT and are displaced from the primary vertex by a non-neglible amount, which conveniently matches the criteria of the B parking HLT. The B parking HLT thus has a relatively high trigger efficiency at most of the mass points and decay lengths used in this analysis [8]. However, because of the selection criteria requiring a good measurement of

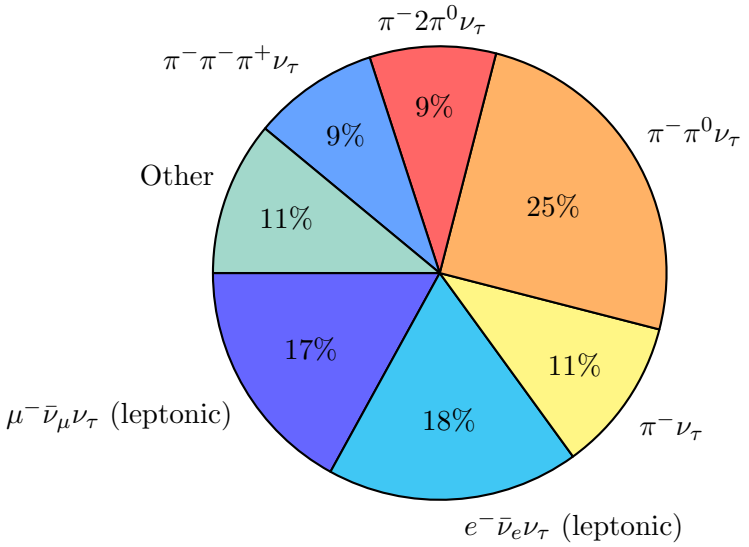


Figure 3.2: Figure displaying the decay modes of the τ and their rounded branching fraction, based on data from the PDG in [4].

impact parameter (which is a measure of displacement from the primary vertex), decay within the tracker is crucial for events passing the trigger. Signal points with $c\tau = 1000$ mm fail to leave a track within the tracker region, resulting in a lack of impact parameter information and therefore a failure to pass the trigger. Signal points with $c\tau = 1$ mm display a dependence on mass, which is due to the fact that a lower mass particle will be boosted more on average, as shown in Figure 3.3. This means that despite the decay length being less than the distance to the tracker, many of these events are able to make it to the tracker and decay there, providing an impact parameter measurement and allowing them to pass the trigger.

3.2 Background

Use of the B parking trigger, though convenient, is not without its drawbacks. Because the trigger requirements are relatively loose, the data is populated by background events. The biggest contribution comes from QCD, as many heavy flavor (loosely, containing a b quark/antiquark) final states (B-mesons) with long lifetimes are produced. $t\bar{t}$+jets (TTJets) also produce many B particles, although TTJets are produced with a lower cross-section. Other processes contribute much less to the background, and thus QCD and TTJets are the only background processes used in the ML training for this work.

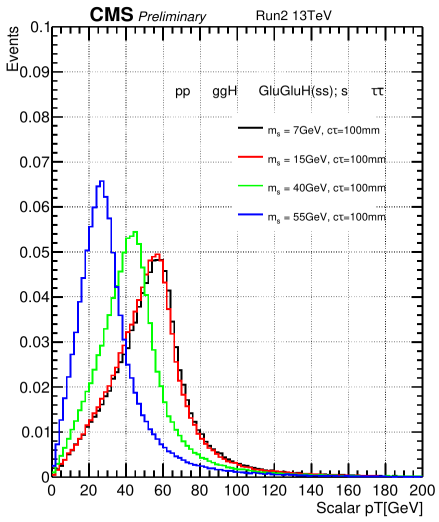


Figure 3.3: Histogram depicting the distribution of scalar transverse momentum p_T at various masses, taken from [8].

3.3 Signal

Table 3.1 lists the masses and lifetimes of the LLP studied. The cross section used in the Monte Carlo generation of the signal samples is 4.414 pb [8]. Table 3.2 lists the B Parking Trigger samples used in the analysis, as well as the HLT trigger path.

Mass (GeV)	$c\tau$ (mm)
7	10
15	10, 100, 1000
40	100
55	100

Table 3.1: Mass and lifetime of the scalar LLP used to generate signal samples.

Data Sample	Trigger
ParkingBPH*-Run2018A	HLT_Mu9_IP6_part*
ParkingBPH*-Run2018B	HLT_Mu9_IP6_part*
ParkingBPH*-Run2018C	HLT_Mu12_IP6_part*
ParkingBPH*-Run2018D	HLT_Mu12_IP6_part*

Table 3.2: Table displaying technical details of the B Parking data and the HLT trigger paths used in the analysis, as stated in [8].

CHAPTER 4

MACHINE LEARNING

The use of machine learning in this analysis is based on the Deep Sets framework described in [9]; more specifically, the ROI mechanism (see Chapter 5) is an adaptation of the "particle flow networks" described in [10]. In general, the use of ML in analyzing data from colliders is complicated due to the fact that the potential number of inputs to a ML model is not fixed, which is troublesome because many common ML techniques require a fixed length vector as input for training. In physics analyses, we rely on observables (i.e. transverse momentum, multiplicity, mass etc.) which are functions of a non-fixed number of particles in order to help us understand the physics underlying the observations. This means that if we want to use experimental data to calculate some observable, we need to have a function that is general enough to work with a variable number of inputs. Additionally, it is clear that there is no physical reason why the order of the inputs should matter, so the function in question must also be invariant under permutation of the input particles. This makes the application of ML to the calculation of an observable less than straightforward. However, as is described more thoroughly in [10], this can be accomplished through the decomposition of the observable into a function of a summation of some representation of particles.

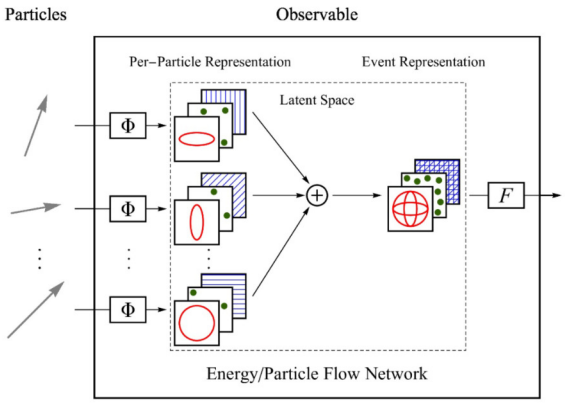
4.1 Observable Decomposition

More specifically, as in [10], we can write that an observable $\mathcal{O}$ which is a function of M particles p_i can be approximated as

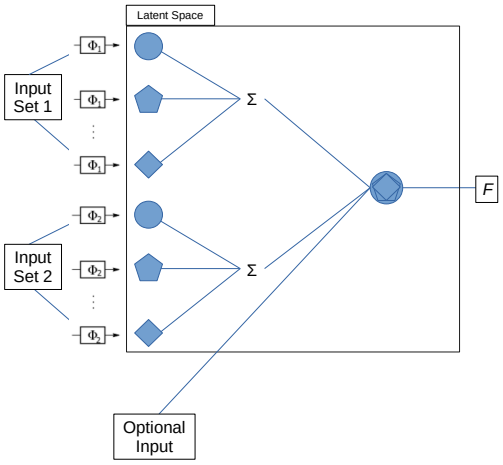
$$\mathcal{O}(\{p_1, p_2, ..., p_M\}) = F\left(\sum_{i=1}^{M} \Phi(p_i)\right), \tag{4.1}$$

where Φ is a function that maps a particle to a representation in latent space [1], and F is a function that takes the sum of the latent space representations and determines the value of the observable. In this case, a particle can be thought of a vector containing some finite number of features, which could be fundamental properties such as charge or spin, as well as kinematic properties such as p_T or position. This decomposition can be visualized in Figure 4.1a, taken from [10]. While this is

[1] A space in which the distance between objects is determined by their similarity, i.e. nearer is more similar or vice versa. These spaces are often lower-dimensional than the space from which the data points originate, which generally makes the problem appropriate for machine learning.



(a) Figure that represents the decomposition of a physical observable as written in Equation 4.1, taken from [10]. The sum of the latent space representations of the particles is taken to be the latent space representation of an event.



(b) Figure depicting a representation of the model used in this analysis. Notable differences include the use of multiple Φ functions, as well as the interpretation of F as a representation of an event. Here, it instead reprents an ROI.

Figure 4.1: Comparison of the structure described in [10] and the structure used in this work.

very abstract language, the approach can be understood by considering two familiar examples of observables: particle multiplicity and mass. In the first case, our Φ maps every particle to 1, so that our sum will simply equal M. Clearly then $F(x)$ should just be $F(x) = x$, that so we obtain the particle multiplicity M as desired. Similarly, if we map every particle to its four-momentum p^μ and define $F(x^\mu) = \sqrt{x^\mu x_\mu}$, we will obtain the mass.

4.2 Application to Displaced Vertices

As is no doubt already clear, this analysis is concerned with the identification of displaced vertices. Though this application is more complicated than those discussed in the previous section, the Deep Sets/particle flow network framework is robust enough to handle it, albeit with slight adjustments. In particular, Figure 4.1b is a more accurate representation of the framework of our model [2]. We sort our input variables into multiple sets and assign each set their own Φ. Classification of input variables into a set is dependent on the characteristics of the variables, and is described in Chapter 5. Additionally, some variables can be passed through directly to F.

[2]Note that this document contains a description of the structure of the ML model at the time the work was performed, which is no longer completely accurate.

CHAPTER 5

REGIONS OF INTEREST

As mentioned, this analysis seeks to use the presence of displaced vertices to distinguish signal events from background. In particular, the novel Regions of Interest (ROI) mechanism is applied, rather than using a variable such as the number of displaced jets. Because of the number of variables associated with each ROI (20-30), a machine learning based tagging approach is used rather than the variable tagging method used in the ZH analysis. This section details the formation of ROIs, the variables contained within them, and the role of ML in this method.

5.1 ROI Formation

The principle behind using ROIs is to identify displaced vertices directly, rather than by identifying displaced objects (jets, electrons, etc.) within them. Figure 5.1 displays a cartoon depiction of the formation process, as well as of the annulus definition. ROI formation begins by using the CMSSW V0Fitter to fit pair-wise tracks of Lost-tracks and PackedPFCandidates in MINIAOD data into a vertex. The fitted vertex (vertices) is (are) then clustered into a Region of Interest with a 1 cm radius. We then need to define the annulus of the ROI:

- start with a cone of $\Delta R < 0.3$ around the center of the ROI, where ΔR is calculated with respect to the primary vertex
- define the annulus plane as the plane passing through the center of the ROI and oriented perpendicularly to the axis of the cone
- form a circle in the annulus plane corresponding to the intersection with the cone [1]
- any tracks that pass through the annulus are also saved within the ROI data and designated as annulus tracks

All of the information associated with the tracks and annulus tracks is therefore stored within each ROI.

[1]The current version of this technique takes the circle and uses it to form a spherical shell, which is called the isolation shell [8]. Any tracks that pass through the entire shell are the annulus tracks.

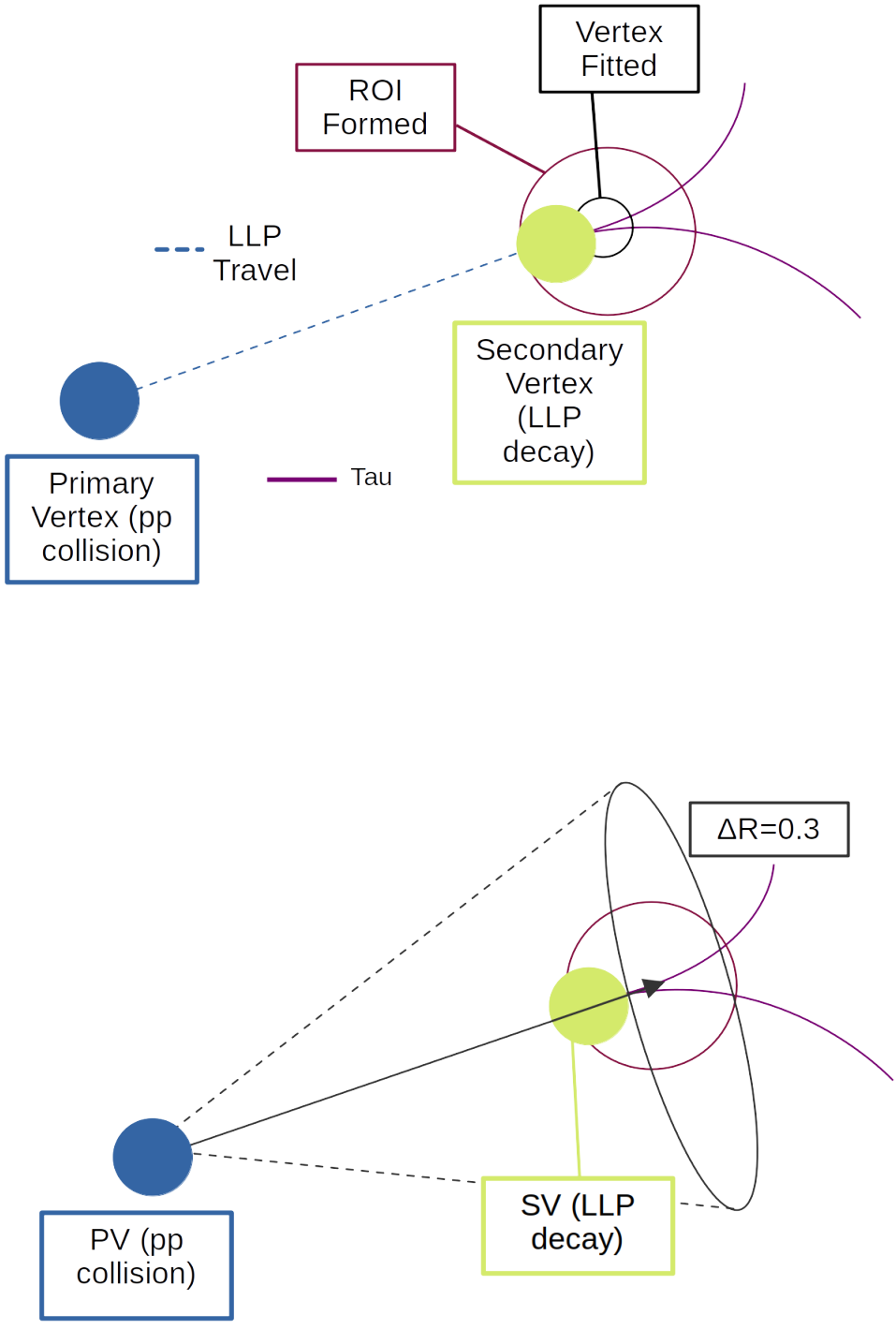


Figure 5.1: Figure depicting the formation of an ROI, as well as a definition of the annulus, adapted from figures in [8].

5.2 Machine Learning Input Variables

Appendix A contains a description of the deep neural network machine learning used in this work. Table 5.1 shows the Tensorflow parameters used in this analysis. The extensive variables for input

Table 5.1: Tensorflow information

Epoch	300
batch size	250
Phi sizes	((64,128,256),(32,64,128))
f sizes	(256,128,32)
Signal	ggHSSTo4Tau-MS15GeV-cτ100 mm
Background	QCD_Pt120-170_MuEnriched and TTJets

of our deep neural network (DNN) are described and categorized in Tables 5.2, 5.3, and 5.4.

Table 5.2: ROI (trackCluster) variables by category

TrackCluster	Position	TrackClusters.vx() - primaryVertex.X()
	Position	TrackClusters.vy() - primaryVertex.Y()
	Position	TrackClusters.vz() - primaryVertex.Z()
	Covariance	TrackClusters.vertexCovariance()(0,0)
	Covariance	TrackClusters.vertexCovariance()(0,1)
	Covariance	TrackClusters.vertexCovariance()(0,2)
	Covariance	TrackClusters.vertexCovariance()(1,0)
	Covariance	TrackClusters.vertexCovariance()(1,1)
	Covariance	TrackClusters.vertexCovariance()(1,2)
	Track0,1	Track0,1.pt
	Track0,1	Track0,1.eta
	Track0,1	Track0,1.phi
	Track0,1	Track0,1.dxy
	Track0,1	Track0,1.dz
	Track0,1	Track0,1.normalizedChi2
	Track0,1	Track0,1.HighPurityInt

The variables saved within the ROI are grouped into four categories: vertex, annulus, auxiliary, and event. The vertex category contains essentially contains the position of the vertex (and each track), the p_T for each track, and the impact parameter of each track (along with associated error information). It is important to note that the positions within the ROI are defined with respect to the location of the primary vertex. The annulus category contains the information (pT, dxy,

Table 5.3: ROI (Annulus) variables by category

Annulus	pfCandidate/LostTracks	pfCandidate/LostTracks.pt
	pfCandidate/LostTracks	pfCandidate/LostTracks.eta
	pfCandidate/LostTracks	pfCandidate/LostTracks.phi
	pfCandidate/LostTracks	pfCandidate/LostTracks.dxy
	pfCandidate/LostTracks	pfCandidate/LostTracks.dz
	pfCandidate/LostTracks	pfCandidate/LostTracks.normalizedChi2
	pfCandidate/LostTracks	pfCandidate/LostTracks.HighPurityInt
	pfCandidate/LostTracks	pfCandidate/LostTracks.DeltaR(trackMomentum)

Table 5.4: Event variables by category

ROI	Position	x
	Position	y
	Position	z
H_T		

dz, etc.) associated with each track that passed through the aforementioned annulus circle. The auxiliary information contains the location of each vertex within an ROI, as well as the number of vertices present within it. ROIs commonly have only one vertex present, so this is essentially a proxy for the location of the vertex itself. As will be shown later, both the annulus and the auxiliary variables are strong predictors of signal vs. background. For instance, signal events are not likely to produce many tracks that are captured within the annulus compared to a major source of background such as QCD. The event level info includes ROI position and H_T. The ROI position is duplicate information of vertex position in the vertex category, and the H_T variable can be thought of as describing the amount of hadronic activity in the process, with the scale set by the 125 GeV Higgs mass.

CHAPTER 6

RESULTS

6.1 Data Simulation & ML Optimization

In order to train ML models for data analysis, one must first simulate the appearance of signal and background events in CMS. Of course, we also have to form ROIs in the simulated data according to the process described in Chapter 4. This process is beyond the scope of this work, but what is important is that in order to do so, one must of course select parameters (mass, lifetime) of the LLPs (the parameters used in this analysis are given in Chapter 3). This in turn means that the ML model may perform differently depending on which parameters are chosen. Additionally, the well-understood background processes (TTJets and QCD) must also be simulated, and also depend on physical parameter choice. Again, this means that the performance of the ML model depends on the values of the parameters, so in reality performance is dependent on the combination of signal and background parameters. It is thus important to test each combination of signal and background parameters, so that a model with the best performance can be produced and the most stringent limits on branching ratios can be achieved. Table 6.1 depicts the combinations studied in this work, along with their associated AUC scores. AUC is a metric used to assess the performance of a machine learning model, where the closer the score is to 1, the better. The result of the training is a neural network which assigns each ROI in the dataset a score between zero and one, where zero indicates background and one indicates signal.

6.2 Epoch Testing

During testing of the various machine learning model parameters, the effect of varying epoch number was also considered. The number of epochs refers to the number of times the entire dataset is passed through the model. As one might expect, an increase in the epoch number typically corresponds to a large increase in computing time, and it is therefore important to know at what point the performance levels off. Table 6.2 displays the model output scores for various epoch numbers for one of the models tested. The model appears to not benefit from more than $\sim$125 epochs.

Table 6.1: Table of the AUC scores associated with each combination of signal and background for models trained with 125 epochs. The AUC scores are all very similar, indicating that the method is effective at a wide variety of signal and background parameters.

Signal	Background	AUC
ggH_HToSSTo4Tau_MH-125_MS-7_ctauS-10	QCD_Pt-20to30_MuEnrichedPt5	0.9696
	QCD_Pt-470to600_MuEnrichedPt5	0.9727
	TTJets	0.9680
ggH_HToSSTo4Tau_MH-125_MS-15_ctauS-10	QCD_Pt-20to30_MuEnrichedPt5	0.9598
	QCD_Pt-470to600_MuEnrichedPt5	0.9679
	TTJets	0.9610
ggH_HToSSTo4Tau_MH-125_MS-15_ctauS-100	QCD_Pt-20to30_MuEnrichedPt5	0.9755
	QCD_Pt-470to600_MuEnrichedPt5	0.9738
	TTJets	0.9680
ggH_HToSSTo4Tau_MH-125_MS-15_ctauS-1000	QCD_Pt-20to30_MuEnrichedPt5	0.9791
	QCD_Pt-470to600_MuEnrichedPt5	0.9742
	TTJets	0.9772
ggH_HToSSTo4Tau_MH-125_MS-40_ctauS-100	QCD_Pt-20to30_MuEnrichedPt5	0.9695
	QCD_Pt-470to600_MuEnrichedPt5	0.9698
	TTJets	0.9705
ggH_HToSSTo4Tau_MH-125_MS-55_ctauS-100	QCD_Pt-20to30_MuEnrichedPt5	0.9635
	QCD_Pt-470to600_MuEnrichedPt5	0.9674
	TTJets	0.9706

Epoch Number	Loss	Acc.	val_loss	val_acc	AUC
100	0.2161	0.9074	0.2469	0.8942	0.9388
150	0.2039	0.9120	0.2393	0.8983	0.9414
200	0.1934	0.9151	0.2523	0.8953	0.9408
250	0.1977	0.9144	0.2459	0.8982	0.9399
300	0.1738	0.9272	0.2573	0.8977	0.9387
350	0.1607	0.9332	0.2693	0.8934	0.9403
400	0.1459	0.9387	0.2823	0.8970	0.9394

Table 6.2: Model output scores for varying epoch number during training.

6.3 ML Model Performance

While AUC is a good measure of the performance of a ML model, it is far from the only metric that can be used. Additionally, other methods of evaluating ML models can help visualize differences in discriminatory power, and actually display some of the physics embedded within the analysis. One way to visualize comparisons of different ML models is through histograms such as the one in Figure 6.1, which shows the relative amount of events classified as signal and background for several signal combinations, along with a dotted line showing the same for a QCD_Pt20-30_MuEnriched back-

ground dataset. Figure 6.2 shows a similar plot for another combination of signal and background. Ideally, every signal point would lie in the bin closest to one, and similarly every background point would lie in the bin closest to zero, but in actuality sharply decreasing curves towards or away from zero for signal or background respectively indicate good performance. In the case of Figure 6.1, the model was trained on a portion of $m_S = 15$ GeV and $c\tau_S = 1$ mm signal data, so it performs quite well on the full signal and background datasets corresponding to these parameters. Clearly, however, the performance does not extend as well to the longer lifetime datasets, indicating that training with this choice of parameters is not well suited to create the main ML model for the analysis. Additionally, there appear to be several peaks present at ROI scores of ∼0.4 and ∼0.8. Similar peaks appear at slightly different ROI scores on the histogram in Figure 6.2. Further studies are necessary to investigate the origin of these peaks, as they persist in different signal and background samples. It is possible that they can be associated to specific particles; this could be investigated by comparing the results to the Monte Carlo generation parameters.

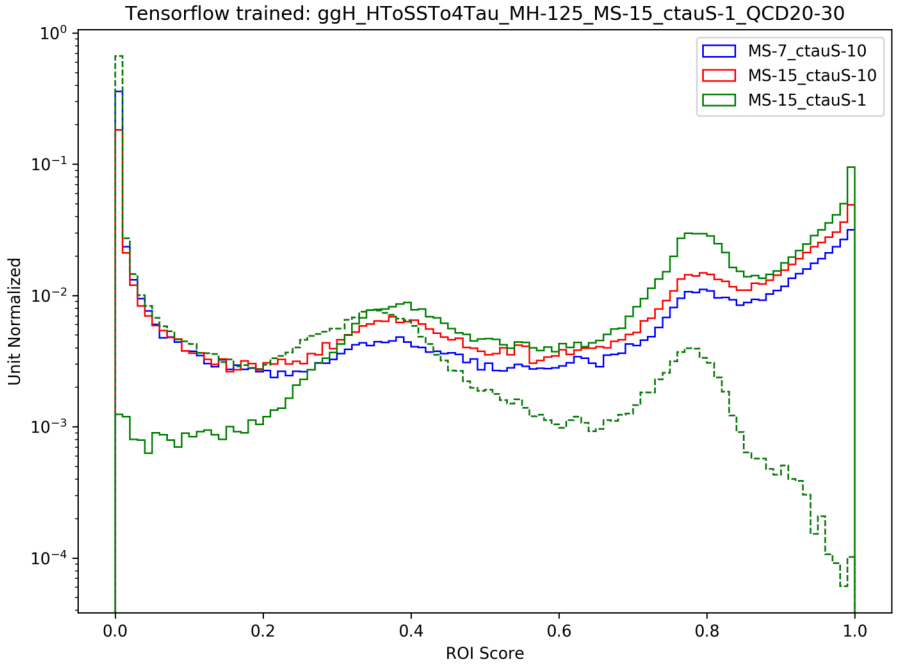


Figure 6.1: A histogram depicting the number of events at each ROI score. The dotted line indicates the performance of the model on a dataset containing purely QCD_Pt20-30_MuEnriched background events.

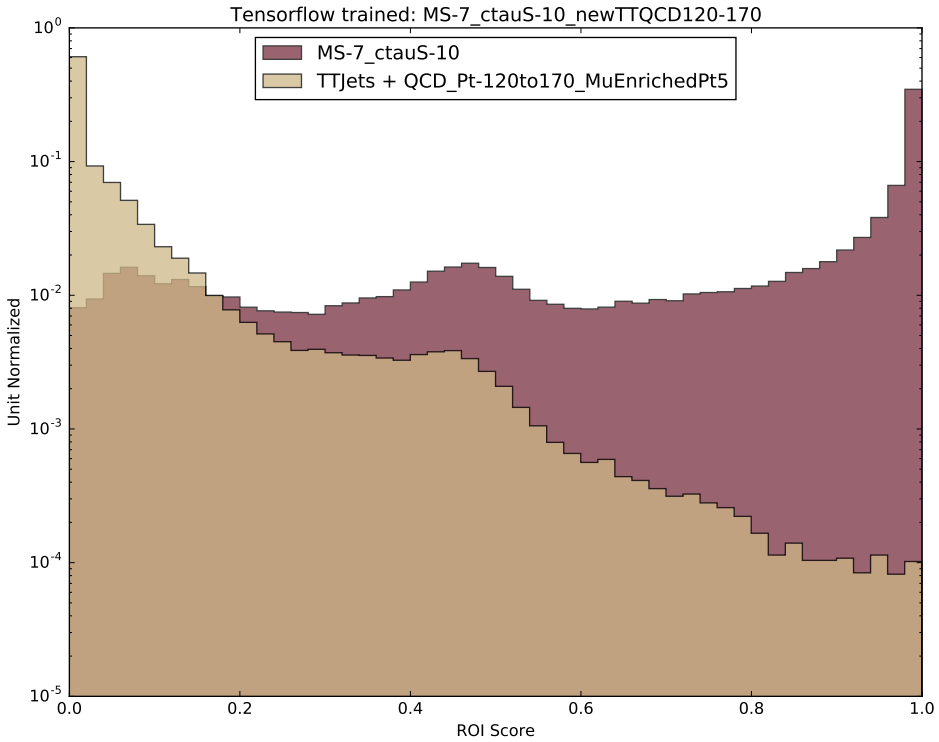


Figure 6.2: A histogram depicting the number of events at each ROI score, for a m_S= 7 GeV, $c\tau$ = 10mm signal dataset and TTJets + QCD 120-170 background dataset.

6.4 Variable Correlation

Another way to visualize ML model performance is through simple correlation plots, i.e. plots of ROI score vs. each variable used as an input to the DNN. For example, one would expect that an increasing ROI position (increasing distance from the primary vertex) would be positively correlated with increasing ROI score for signal events, since signal events by nature have a large displacement from the primary vertex. In other words, one would expect such a correlation because ROIs are designed to identify displaced vertices. Similarly, you would expect a negative correlation for the same such plot for background events, since it is unlikely that there would be many background events with large displacements from the PV. Figure 6.3 displays such plots, and the behavior is exactly as expected. Correlation plots have been produced for all of the variables used as inputs

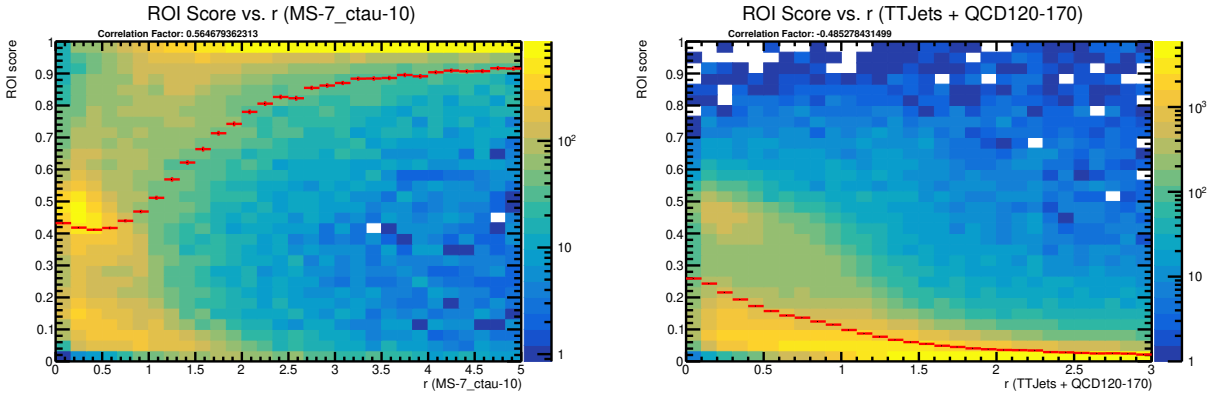


Figure 6.3: Correlation plots (ROI score vs. r) for signal and background, where $r = \sqrt{x^2 + y^2}$, and x, y refer to the ROI coordinates. The model used was trained with signal ggHSSTo4Tau-MS7GeV-cτ10 mm and background QCD Pt120-170 MuEnriched and TTJets.

Variable	Correlation Factor
ROI Distance from Primary Vertex, r	0.57
Annulus Track ΔR to ROI	-0.24
ROI Number of Constituents	-0.14

Table 6.3: Table displaying several variable inputs to the ML model, ranked by absolute value of correlation score.

to the ML model, and the complete set of plots are provided in Appendix A. The x-limits of each plot have been manually adjusted to eliminate regions where there are very few events, as the relationship between variable and ROI score can otherwise appear somewhat unclear. Several of the top performing variables have been isolated and ranked by absolute value of correlation factor;

the results of this are shown in Table 6.3. The tracks in the ROI vertices are ordered by transverse momentum, with higher tranverse momentum desginated as track 0, and lower as track 1. This is visible in Figure 6.4, where upon close inspection it is clear that higher values of pT are more populated in the track 0 plot than in the track 1 plot.

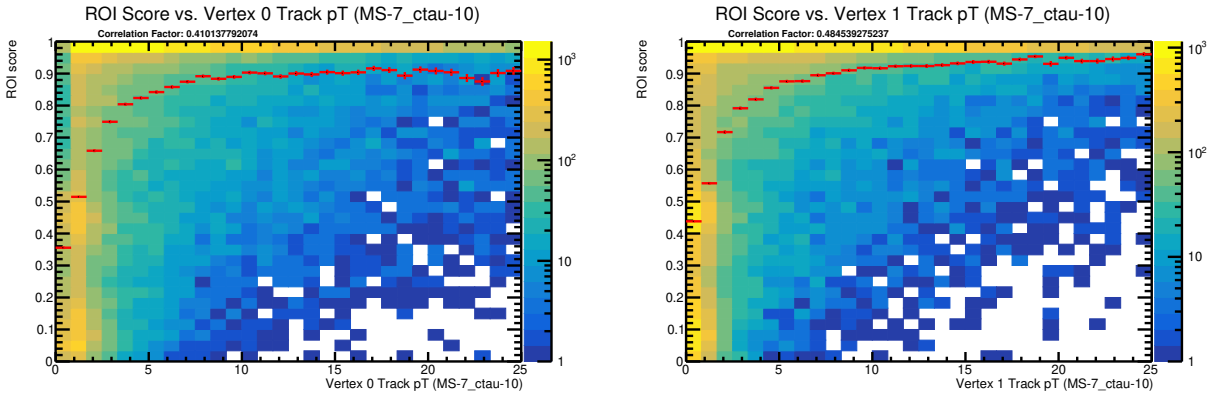


Figure 6.4: Correlation plots for vertex transverse momentum for each track in the ROI vertex.

We can understand the ordering of the results within Table 6.3 through physics. It is no surprise that the strongest performing variable is distance from the primary vertex, as this directly related to the defining characteristic of the signal signature, i.e. a long-lived particle. High transverse momentum corresponding to signal makes sense, as the model used in this case was trained with $m_S = 7$ GeV, so there is a lot of energy available for a boost when the Higgs decays. This apparent dependency on mass could be an issue if attempting to use a model trained with $m_S = 7$ GeV to set limits on $m_S = 40$ GeV or 55 GeV, but would likely not be problematic if used for $m_S = 15$ GeV (see Figure 3.3 from Chapter 2). A negative correlation with the number of constituents in an ROI is also consistent with physics. As mentioned in Chapter 4, background such as QCD is likely to produce many tracks that are near each other and therefore stored within an ROI, whereas signal is not. Thus one would expect and indeed finds that increasing number of constituents makes the model less likely to identify an event as signal. This result is also reflected in the negative correlation with annulus track ΔR to ROI, as signal events produce fewer annulus tracks that are closer to the ROI, and QCD produces many annulus tracks that can be farther away.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

It is clear from the results that the ROI mechanism accomplishes its main purpose, which is to identify displaced vertices and thereby eliminate a majority of events from consideration as possible signal. As the results summarized in Table 6.3 show, not all of the input variables are very effective, and it is possible that performance of the model could be improved by removing some of these inputs. However, this would need to be tested thoroughly as it is entirely possible that the contribution of certain variables to the performance of the model cannot be demonstrated by a simple correlation plot such as the ones produced in this work. This is particularly true for some of the variables in the vertex category, where the information of one track alone is perhaps not sufficient but instead the combined information of the two tracks is very useful for the model. Future studies can determine the minimal number of inputs to the ML model that will produce sufficient performance, thereby maximizing performance while minimizing computing time. Overall, the ROI mechanism was effectively used in this analysis to identify events as potential signal. The scores produced from this method were combined with cuts determined according to variables not used as ML inputs in order to determine the final event selection [8]. The preliminary limits produced by this analysis, taken from [8], are shown in Figure 7.1.

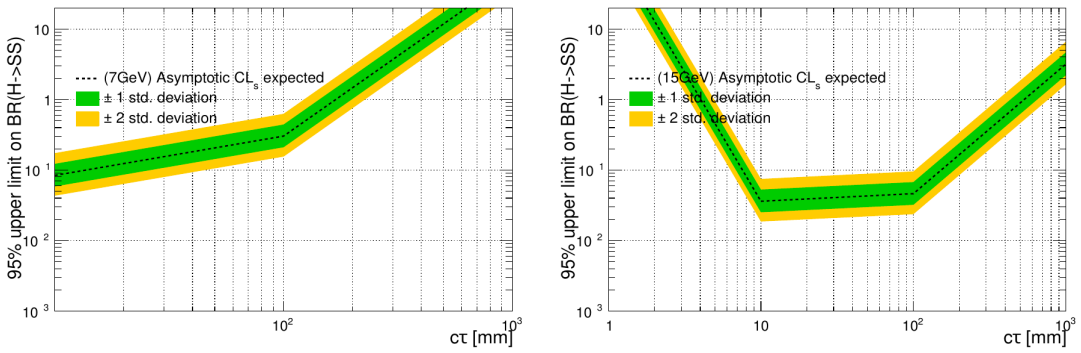


Figure 7.1: Preliminary limit plots for the branching ratio H $\rightarrow$ SS as determined in [8].

The most probable use of the ROI mechanism in future analyses lies in its ability to identify displaced vertices directly without the need to identify the physics objects within them. Any physics search that has a clear trigger strategy and revolves around decays within the tracker of CMS should be able to use the ROI mechanism. The advantage of using this method is that for signatures similar to ours, where the actual reconstruction of the final state is too complicated to allow the use of variable tagging or cuts, ROI score can be used to greatly reduce the potential number of events that need to be considered. After this is done, additional cuts according to variables not present in the ML can be applied, as in this analysis. The High Luminosity LHC (HL-LHC) will be collecting data on and off from the late 2020s through the early 2040s, so analysis techniques such as the ROI technique will be very useful in conducting searches using the massive amount of HL-LHC data in the coming years.

One potential use of ROIs in an ongoing search for new physics is in emerging jets, which are predicted by theories that postulate the existence of new fermions that have no electric charge but are instead charged under some QCD-like force in the dark sector. Since the interaction is QCD-like, these fermions, or "dark quarks", have similar confinement properties to the SM quarks, meaning they will immediately hadronize into a jet if produced in isolation. In particular, some models of these dark quarks predict jets containing long-lived dark hadrons, which are called emerging jets. These dark hadrons decay to SM particles via some mediator particle, therefore giving rise to many displaced vertices within the jet. A search for these jets based on using a trigger based on the p_T of jets in an event was performed in [11]. The analysis was able to set limits on the signal cross section over many regions of dark pion lifetime and mediator particle mass, but it relied on the use of several artificial variables in order to identify events as signal or background. Given that the signature is so messy, it is possible that using the ROI mechanism, i.e. including all of the displaced vertices present in the emerging jets inside an ROI (Figure 7.2), would allow for the ML model to pick up on patterns that are not obvious from the artificially defined variables. ROI score could then serve as a first event selection before other cuts were applied, which may allow for more stringent limits to be set.

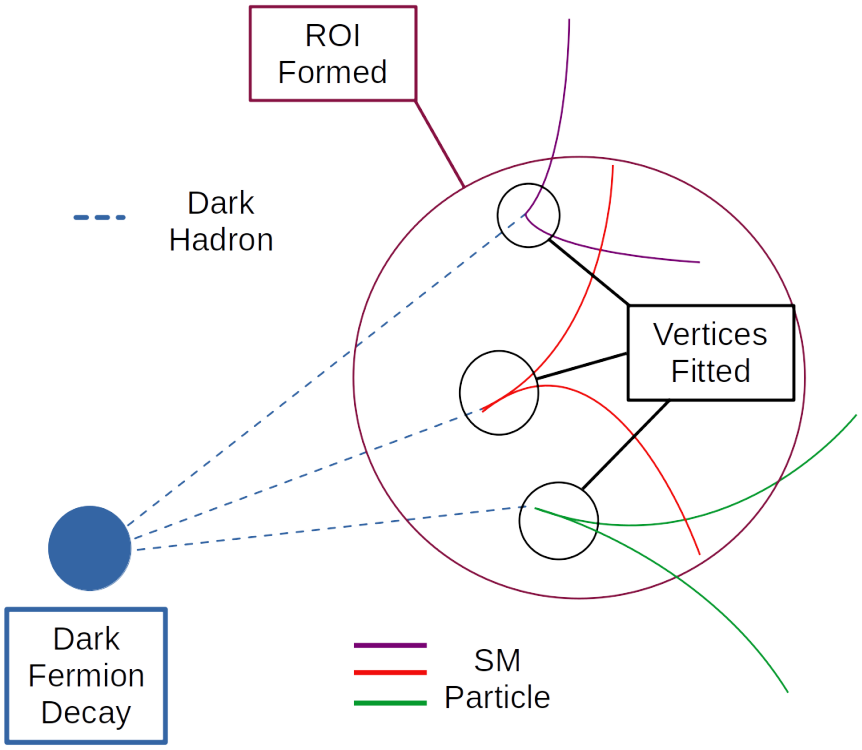
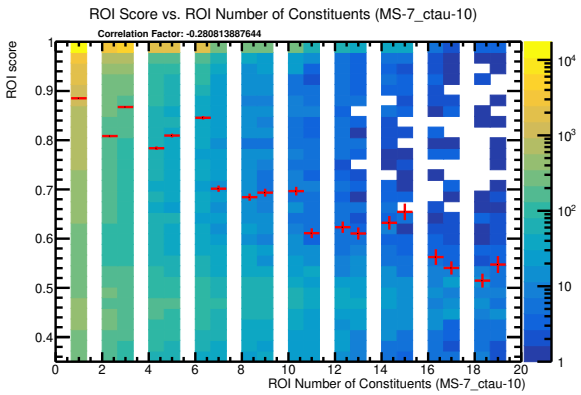


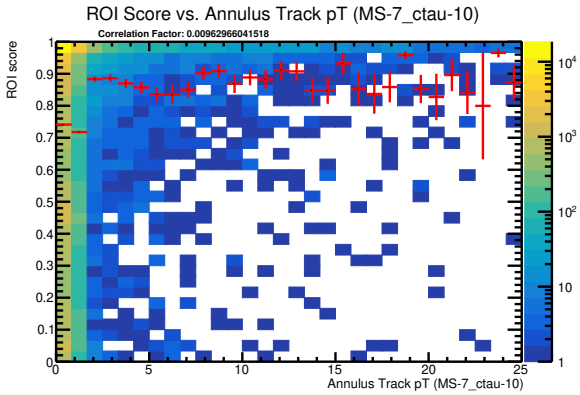
Figure 7.2: Cartoon diagram of the displaced vertices within an emerging jet being clustered into an ROI.

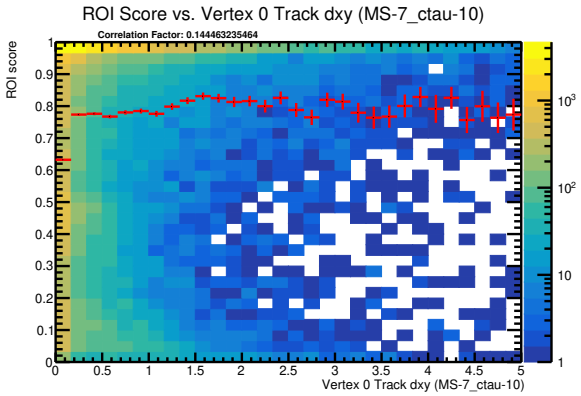
APPENDIX A

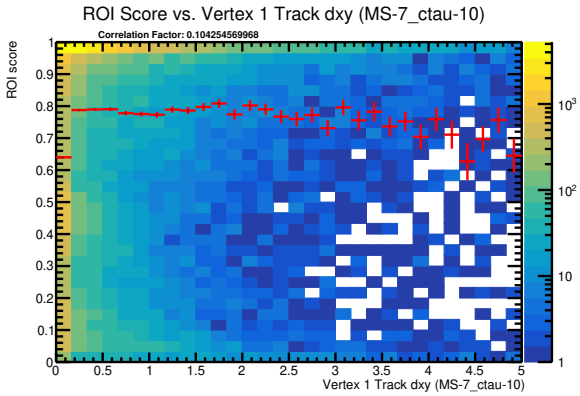
CORRELATION PLOTS

This appendix contains the correlation plots for the variables investigated that were not shown in Chapter 5. Plots for the entire set of ~ 30 variables that serve as ML inputs are available, but are not shown here[1]. The lack of a clear pattern in many of the plots (or a large correlation coefficient) demonstrates that on their own they are not contributing to the performance of the model in an obvious way. Future work should involve investigating the performance of the model with some of the low scoring variables removed.

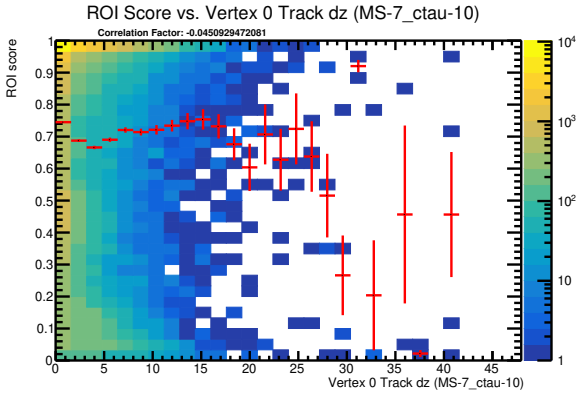


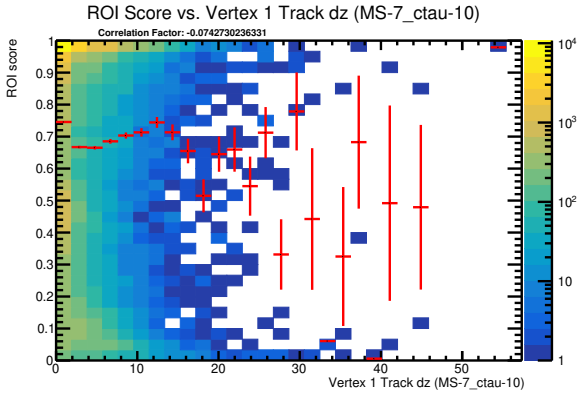


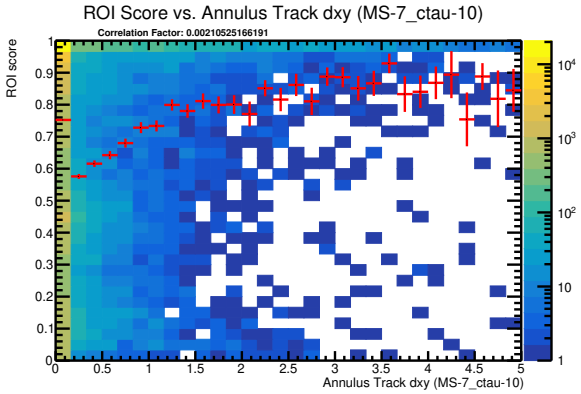


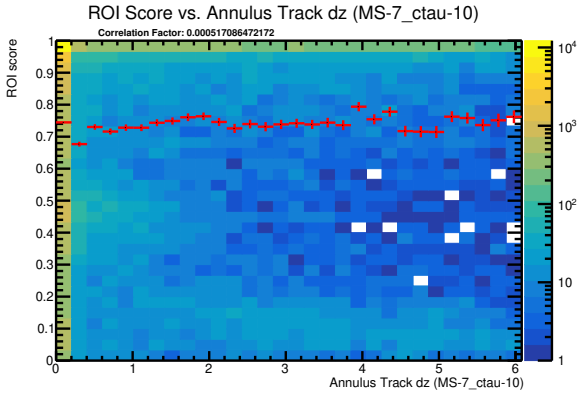


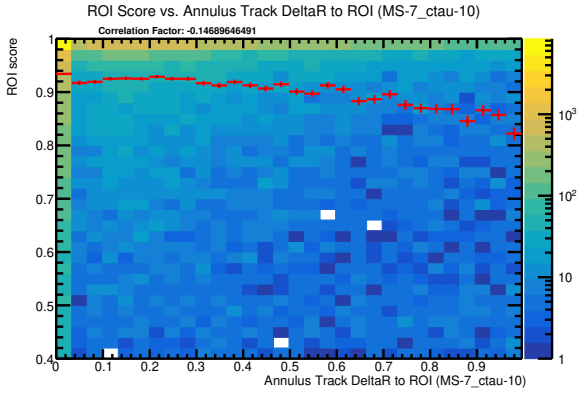
[1]See `https://drive.google.com/drive/folders/1VfCNeiuPTaBsKykkbAuYd78I5bdX3BFg?usp=share_link` for the other plots.


ROI Score vs. Vertex 0 Track dz (MS-7_ctau-10)
Correlation Factor: -0.0450929472081
ROI score
Vertex 0 Track dz (MS-7_ctau-10)


ROI Score vs. Vertex 1 Track dz (MS-7_ctau-10)
Correlation Factor: -0.0742730236331
ROI score
Vertex 1 Track dz (MS-7_ctau-10)


ROI Score vs. Annulus Track dxy (MS-7_ctau-10)
Correlation Factor: 0.00210525166191
ROI score
Annulus Track dxy (MS-7_ctau-10)


ROI Score vs. Annulus Track dz (MS-7_ctau-10)
Correlation Factor: 0.000517086472172
ROI score
Annulus Track dz (MS-7_ctau-10)


ROI Score vs. Annulus Track DeltaR to ROI (MS-7_ctau-10)
Correlation Factor: -0.14689646491
ROI score
Annulus Track DeltaR to ROI (MS-7_ctau-10)

REFERENCES

[1] D. Clowe, M. Bradač, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky, "A direct empirical proof of the existence of dark matter," *The Astrophysical Journal*, vol. 648, no. 2, pp. L109–L113, Aug 2006. [Online]. Available: https://doi.org/10.1086%2F508162

[2] B. Martin, *Nuclear and Particle Physics: An Introduction*, 2nd ed. Wiley, 2009.

[3] The ATLAS Collaboration, "Search for long-lived neutral particles in pp collisions at $\sqrt{s} = 13$ TeV that decay into displaced hadronic jets in the ATLAS calorimeter," *The European Physical Journal C*, vol. 79, no. 6, Jun 2019. [Online]. Available: https://doi.org/10.1140%2Fepjc%2Fs10052-019-6962-6

[4] R. L. Workman and Others, "Review of Particle Physics," *PTEP*, vol. 2022, p. 083C01, 2022.

[5] The CMS Collaboration, "Search for long-lived particles produced in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Journal of High Energy Physics*, vol. 2022, no. 3, Mar 2022. [Online]. Available: https://doi.org/10.1007%2Fjhep03%282022%29160

[6] The CMS Collaboration, "The CMS experiment at the CERN LHC," *JINST*, vol. 3, p. S08004. 361 p, 2008, also published by CERN Geneva in 2010. [Online]. Available: http://cds.cern.ch/record/1129810

[7] The CMS Collaboration, "CMS Particle Detection Summary," *WorkBookCMSExperiment*, Dec 2015. [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperiment

[8] S. Kim, "Search for Higgs Boson Decays to Long-Lived Scalar Particles with Regions of Interest and Machine Learning in CMS," Ph.D. dissertation, Department of Physics, Florida State University, Tallahassee, FL, USA, 2022.

[9] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, "Deep sets," 2017. [Online]. Available: https://arxiv.org/abs/1703.06114

[10] P. T. Komiske, E. M. Metodiev, and J. Thaler, "Energy flow networks: deep sets for particle jets," *Journal of High Energy Physics*, vol. 2019, no. 1, jan 2019. [Online]. Available: https://doi.org/10.1007%2Fjhep01%282019%29121

[11] The CMS Collaboration, "Search for new particles decaying to a jet and an emerging jet," *Journal of High Energy Physics*, vol. 2019, no. 2, feb 2019. [Online]. Available: https://doi.org/10.1007%2Fjhep02%282019%29179