Convergence
A Journal for Young Researchers

Received: [for official use only] ; Accepted: [for official use only];
Managing Editor: [for official use only];
Action Editor: [for official use only]

Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data

Mihir Tare

St. Mark's School of Texas, Dallas, Texas, USA

Abstract

The Standard Model of particle physics, while highly successful, leaves several unanswered questions that motivate searches for undiscovered heavy gauge bosons predicted by Beyond-Standard-Model theories. Using machine-learning classifiers trained on low-mass sideband events, this paper presents a search for the Z' boson, one such hypothetical heavy neutral gauge boson that decays into dilepton pairs. A gradient boosted decision tree (BDT) and a multi-layered perceptron deep neural network (DNN) are trained to distinguish real Z boson events from Drell-Yan background events using only 11 raw kinematic features, with dilepton invariant mass deliberately excluded to avoid background sculpting of artificial signals during the search. The analysis uses real proton-proton collision data recorded by the ATLAS experiment during LHC Run 2 in 2015, released through the CERN Open Data portal in 2024. After both classifiers achieve accurate discrimination on the training set, events in a blinded high-mass search region are scored by the classifiers, a three-parameter parametric fit is used to model the background, and an Asimov significance scan is performed across mass bins. We observed no statistically significant excess, with the largest local deviation at a mass of 1120 GeV with a significance of 2.40σ in the DNN's highest-score category, which is well below the 3σ conventionally required for evidence of new physics. This analysis demonstrates how machine-learning can be used in high-energy physics "bump hunt" searches while preserving the background model, and it provides a reproducible methodology built on publicly available data.

Keywords:

Z′ boson, dilepton resonance search, machine learning, boosted decision tree, deep neural network, ATLAS Open Data, bump hunt, beyond the standard model

1. Goals and Theory

The Standard Model (SM) of particle physics provides a remarkably successful description of the fundamental constituents of matter and the forces that govern their interactions. Despite its predictive power, the SM is widely understood to be an incomplete theory. It does not account for observed matter-antimatter asymmetry of the universe, provides no dark matter candidate, fails to incorporate gravity at the quantum level, and leaves the mass hierarchy of fermions unexplained.

A variety of Beyond Standard Model (BSM) theories address these shortcomings by hypothesizing the existence of new heavy gauge bosons that come from extensions of the Standard Model's electroweak symmetry, the mathematical structure that organizes how particles interact with the weak nuclear force. These hypothetical particles are called Z' bosons (Z-prime bosons). A Z' boson is generally understood to be a heavier counterpart of the Z boson, a known particle that mediates the weak nuclear force and was discovered at CERN in 1983. Many BSM extensions predict that the Z' would couple to leptons and quarks just like the Z but at a higher mass scale.

Previous searches at the LHC have set lower mass limits on Z' bosons in the multi-TeV range for the most common benchmark models. Both ATLAS and CMS have performed high-mass dilepton searches using the full Run 2 dataset of 139 fb-1 and found no evidence of a Z'. An undiscovered resonance remains possible, particularly at masses or coupling strengths below the sensitivity of these previous analyses.

Our search utilizes the dilepton channel because of its simplicity and clarity. As a hypothetical "cousin" of the Z boson, the Z' should decay into a pair formed by a lepton and its corresponding anti-lepton (e.g. an electron and positron). Utilizing this channel, often called the "golden channel," allows for a clean experimental signature and a well-understood background from the Drell-Yan process. If a Z' exists, it would appear as a localized excess of events, or a "bump", in the dilepton invariant mass spectrum at the Z' mass, sitting on top of the smooth Drell-Yan continuum. This "bump-hunt" strategy is the standard approach to discovering new resonances at colliders.

Machine learning is now a standard tool in particle physics for classifying collision events. However, applying machine learning to bump-hunt searches comes with a concern: if a classifier learns to reconstruct invariant mass (even indirectly through correlated features), it can distort the smooth background spectrum and create artificial peaks where no real signal exists. This problem, known as background sculpting, is the primary failure this analysis is designed to avoid.

The goal of this work is to develop and validate a machine learning pipeline for a Z' boson search using LHC data, following the methodological structure of real collider analyses, and to demonstrate a statistically valid null result or identify candidate excesses.

2. Experimental Background

2.1. The LHC

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator. It accelerates two beams of protons in opposite directions to nearly the speed of light, then brings them into head-on collision at designated interaction points. During Run 2 (2015-2018), the LHC operated at a center-of-mass energy of $\sqrt{s}$ = 13 TeV, meaning the total energy available in each proton-proton collision is 13 trillion electron-volts. This analysis uses data collected in 2015 during the first year of Run 2.

2.2. ATLAS Detector

The ATLAS (A Toroidal LHC Apparatus) detector is a large general-purpose particle detector at the LHC. It is designed to record the debris from high-energy proton-proton collisions and reconstruct what happened in each event using software and sensors. The detector is built in concentric layers around the collision point, with each layer optimized to measure different types of particles:

- Inner Detector: The innermost layer immersed in a strong magnetic field. Charged particles curve as they pass through the field, and the curvature is used to measure their momentum. This is the primary system for tracking leptons such as electrons and muons.
- Calorimeters: These layers, surrounding the inner detector, absorb particles and measure their energy. The electromagnetic calorimeter measures electrons and photons, while the hadronic calorimeter measures jets (sprays of particles produced by quarks and gluons).
- Muon spectrometer: Muons pass through calorimeters unaffected (because they are heavy leptons, their momentum is too high for calorimeters to stop them), so this dedicated system in the outermost layer identifies and measures them with high precision using a toroidal magnetic field.

2.3. Dataset

This analysis uses Run 2 2015 proton-proton collision data made publicly available through the CERN Open Data Portal in 2024. The data is provided in DAOD-PHYSLITE format, an analysis-ready format in which the ATLAS experiment collaboration has already performed computationally heavy work like identifying particles. This full dataset contains 10,049 files with a total 9.3 TiB of approximately 1.69 x 10^9 events and has a luminosity of 3.2 fb^{-1}.

3. Data Preparation

3.1. Event Selection

We select events containing exactly two same-flavour leptons: dielectron events (with exactly two electrons) and dimuon events (with exactly two muons). For each dilepton event, the following reconstructed objects are extracted:

- Lepton Kinematics: Using analysis-level lepton objects (AnalysisElectronsAuxDyn and AnalysisMuonsAuxDyn), the transverse momentum pT (the component of momentum perpendicular to the beam), pseudorapidity η, and azimuthal angle φ of each lepton is gathered.
- Jet Kinematics: The three same kinematic variables are collected for the leading reconstructed jet (AnalysisJetsAuxDyn), plus the jet multiplicity n_jets (how many jets were found in the event).
- Missing Transverse Energy (MET): The apparent imbalance in the total transverse momentum of the event, the MET can signal potential undetected particles like neutrinos. It is reconstructed from the x and y components of the missing momentum vector (MET_Core_AnalysisMETAuxDyn.mpx/mpy).

All momenta are provided in MeV by ATLAS and are converted to GeV for this analysis.

3.2. Invariant Mass

The dilepton invariant mass $m\ell\ell$ is an important value for this search. If two leptons originate from a parent particle decay (like a Z' boson decaying into two leptons), then the dilepton invariant mass should equal the mass of the parent particle. Thus, if a Z' boson exists, it would produce a peak of events at a particular value in the invariant mass spectrum corresponding to its mass. This is what physics "bump hunts" typically look for. Dilepton Invariant mass is calculated using the three kinematic variables that we collected for each lepton (pT, η, φ):

$$m\ell\ell = \sqrt{(2\, pT^1\, pT^2\, [\, \cosh(\eta^1 - \eta^2) - \cos(\varphi^1 - \varphi^2)\,]\,)}$$

3.3. Quality Cuts

We apply quality cuts that are standard to the LHC. A pseudorapidity cut $|\eta| \leq 2.5$ is applied to electrons and muons, reflecting the acceptance of the ATLAS inner detector and calorimeter. For muons, this cut reflects limitations of the Semiconductor Tracker (SCT) of the ATLAS Muon Spectrometer and the Inner Detector, which are used to measure muon momentum. For electrons, the ATLAS electromagnetic calorimeter precision region is $|\eta| \leq 2.47$, which makes this analogous cut slightly conservative but still valid. Jet variables are undefined for events where n_jets = 0, so the pseudorapidity cut is conditionally applied to jet objects when n_jets > 0.

A hard cut of $m\ell\ell \leq 3{,}500$ GeV is applied because the background spectrum falls very steeply, and so past this boundary, there are hardly any events in the 20 GeV bins that our fit uses. We account for this problem with a minimum of 10 events for a bin to be used, but this hard cut prevents us from even looking where we expect zero events.

4. Machine Learning Methodology

The machine learning methodology has three stages. First, we train two classifiers, a boosted decision tree (BDT) and a deep neural network (DNN), to distinguish Z-peak events from low-mass Drell-Yan background events using only kinematic features, with the dilepton invariant mass deliberately excluded so that the classifiers cannot learn the search variable. Both are trained and validated on events drawn exclusively from the low-mass sidebands and the Z-peak region, well below the search region, so that the search region itself remains blinded during model development.

Second, we apply the trained classifiers to the high-mass search region ($m\ell\ell > 250$ GeV) and sort every event into categories by classifier score, concentrating the most Z-like events into a single high-scoring category. Within that category, we model the smooth Drell-Yan background with a three-parameter parametric fit:

$$f(m) = p_0 \bullet (m/\sqrt{s})\hat{}p_1 \bullet \exp(-p_2 \bullet m/\sqrt{s})$$

This function uses a power law combined with an exponential decay and is a standard parametrization used in ATLAS analyses. The three parameters (p_0, p_1, p_2) are determined by fitting the function to the observed mass spectrum in 20 GeV bins using χ^2 minimization, which is a statistical procedure that adjusts the parameters to minimize the sum of squared residuals between the fit and the data, weighted by the uncertainty in each bin.

We evaluate the accuracy of the fit with reduced chi-squared statistic, χ^2/ndf, which is a sum over all bins of squared deviations. χ^2 is evaluated as $\chi^2 = \Sigma_i\, (n_i - b_i)^2 \,/\, \sigma_i^2$, where n_i is the number of observed events in a bin, b_i is the number of events predicted to be in the bin by the background fit, and σ_i is a variance calculation. Ndf (number of degrees of

freedom) is calculated by subtracting the number of parameters in the fit (three in our case) from the number of bins we search. This background spectrum should remain smooth and falling, and a good fit will confirm that the classifiers have not sculpted artificial structure into it.

Finally, we perform a bin-by-bin Asimov significance scan within the high-score category, comparing the observed number of events in each bin against the background fit to test for localized excesses. The remainder of this section describes the input features, the two classifier architectures, and their training process.

4.1. Features

The core idea of this methodology is to train machine-learning classifiers to distinguish Z-like events from non-Z-like events using only kinematic features. Specifically, the momenta, angles, jet activity, and missing transverse energy of each event. The trained classifiers are then applied to the search region to identify which high-mass events look most like a Z boson and then concentrate these potential Z' signals into a high-score category.

One critical design choice for the input features is that dilepton invariant mass is excluded, as including it would allow the classifiers to directly learn the search variable and potentially sculpt the background to create artificial "bumps" in the mass spectrum where no real signal exists.

The same 11 input features are used for both classifiers. They are listed with their descriptions in Table 1.

Table 1: Input kinematic features for the models and their descriptions

Feature	**Meaning**
pT^1, pT^2	Transverse momenta of the leading and subleading leptons (electron or muon)
η^1, η^2	Pseudorapidities of both leptons
φ^1, φ^2	Azimuthal angles of both leptons
n_jets	Reconstructed jet multiplicity
leading_jet_pT	Transverse momentum of the leading jet (zero if n_jets = 0)
leading_jet_η	Pseudorapidity of the leading jet (zero if n_jets = 0)
leading_jet_φ	Azimuthal Angle of the leading jet (zero if n_jets = 0)
MET	Missing transverse energy

Both classifiers are trained as binary classifiers, with Z events labeled as signal (class 1) and low-mass Drell-Yan events labeled as background (class 0). The labeled sample of training events is split 80/20 into training and test sets using

stratified sampling (preserving class ratio in both splits), giving 8,950,298 training events and 2,237,575 held-out test events for validation.

4.2. Boosted Decision Tree (BDT)

The BDT, an ensemble of gradient-boosted decision trees that has become a standard classifier in particle physics, uses binary cross-entropy (BCE), a loss function for binary classification where loss increases exponentially when the difference between the model's prediction and the event's actual label increases. For an event with true label y (either 0 or 1) and predicted probability p, the BCE is:

$$L = -[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)]$$

Loss is small when the model's prediction matches the correct label of the event and large when the model is incorrect with high confidence. Each successive tree is fit to the gradient of this loss so that the ensemble progressively reduces its overall error. This loss calculation is done automatically using XGBoost, a scalable open-source gradient-boosting library implemented in Python through its xgboost package, which handles tree construction, gradient computation, and parallelized training internally. All preprocessing of input features, train/test splitting, and evaluation are carried out in Python using NumPy, pandas, and scikit-learn.

To prevent the model from overfitting, we apply several forms of regularization using XGBoost parameters.

- Number of trees and learning rate: We train 300 trees with a low learning rate of 0.03. The learning rate scaled down each tree's contribution, so that each tree makes a small correction rather than aggressively fitting the training data. A low rate combined with many trees sacrifices training speed for better generalization.
- Maximum depth of 5: Each tree is limited to 5 levels deep, which produces at most 32 distinct regions. Unlike deep trees, which tend to memorize individual events, shallow trees learn broad patterns.
- Minimum child weight of 10: In each tree, a leaf node can only be split if both children would contain at least 10 units of statistical weight. This prevents the model from learning patterns from small groups of unusual events.
- Minimum split gain of 1.0: Splits only happen if they reduce the loss by at least 1.0. This "prunes" each tree by removing redundant branches, which helps reduce overfitting.
- L1 regularization/Lasso regression (reg_alpha = 0.1): This essentially adds a penalty proportional to the absolute value of each leaf's prediction. This drives the weight of less important features to zero and simplifies the model by essentially performing a feature selection.
- L2 regularization (reg_lambda = 5.0): Adds a penalty proportional to the square of a leaf's prediction. It shrinks large weights toward zero and strongly discourages overconfident predictions.
- Stochastic gradient boosting (70%): Each tree is trained on a random 70% of events using a random 70% of features. This randomness prevents all trees from making the same mistakes.

The class imbalance between labeled signal events (5,811,472) and labeled background events (5,376,401) is handled by a scale_pos_weight parameter of 0.93 (calculated ratio), which decreases the weight of signal class during training.

4.3. Deep Neural Network

Neural networks, and in particular deep architectures applied to collider classification problems, are built from computational units called neurons. Each neuron takes in a set of numbers and computes a weighted sum based on weights that are learned during training. The result is passed through a nonlinear activation function. Our model uses the ReLU (Rectified Linear Unit) activation function, which outputs the input if it is positive and zero otherwise. ReLU and the DNN architecture is implemented with PyTorch, an open-source deep-learning framework in Python that provides automatic differentiation, GPU-accelerated tensor operations, and prebuilt neural-network modules. The network is defined as a torch.nn.Module subclass, with linear layers, batch normalization, dropout, and ReLU activations composed using the torch.nn API.

Neurons are arranged in multiple layers. In a fully connected layer, every neuron receives input from every neuron in the previous layer. Unlike graph-based architectures designed to find structural relations in inputs, this multi-layered perceptron (MLP) does not explicitly encode invariant relationships like mass between particles. Despite this, correlated kinematic information may still allow partial reconstruction of invariant mass, motivating our validation against background sculpting (which is discussed in the background modelling section). We initially tried to use a Graph Neural Network (GNN), but it was able to reconstruct invariant mass and scored events almost solely based on mass. The switch to this multi-layered perceptron eliminates this background sculpting problem which we were trying to avoid.

Our network consists of an input standardization layer, three fully connected hidden layers of 128 neurons each, and a single output neuron.

- Standardization layer: Before entering the network, each of the 11 input features is transformed to have zero mean and unit standard deviation by subtracting the mean and dividing by the standard deviation calculated only for training sets. Since the raw features have different scales (pT can be hundreds of GeV, while φ is bounded between $-\pi$ and π), this ensures that features with larger magnitudes do not distort the learning process with large weights.
- Hidden layers: three layers of 128 neurons each. Each layer applies a linear transformation, batch normalization (which recenters and rescales the activations across each training batch to stabilize learning), and ReLU activation. Each step is implemented using PyTorch. The depth and width of the network give it the capacity to learn complex nonlinear combinations of input features.
- Dropout (p=0.3): After each hidden layer, 30% of neurons are randomly "switched off" during each training step. This prevents individual neurons from becoming dependent on one another and improves their individual reliability. Dropout prevents the model as a whole from overfitting and memorizing training noise. When testing, however, all neurons are active.
- Output: In the last layer, a single neuron produces an unbounded score, which is passed through a sigmoid function to predict a probability between 0 (background) and 1 (signal).

Training occurs in epochs, where one epoch is a full pass through the training data. For each epoch, the data is divided into batches of 4,096 events. For each batch, the network makes predictions, the loss is computed, and the gradient of the loss with respect to every weight is computed using chain rule (backpropagation). The weights are then updated according to this gradient. We do clip gradients at 1.0 so that large updates to weights do not cause instability. Over many epochs, the DNN converges toward a set of weights that minimizes the loss.

We optimize our DNN with AdamW, a variant of the Adam optimizer that decouples weight decay from the gradient update.

We use a learning rate of 10^{-3} and weight decay of 10^{-4} (an L2 like penalty on all weights that discourages overfitting). AdamW is provided directly by the torch.optim module, and gradients are computed via PyTorch's automatic differentiation engine during backpropagation. We also use two additional techniques which are also implemented using the PyTorch library:

- Cosine annealing with warm restarts: The learning rate decreases smoothly following a cosine curve over the course of 10 epochs, then resets back to its initial value and begins decaying again. These periodic resets allow the optimizer to escape shallow local minima that it might otherwise get trapped in, while the cosine decay within each cycle provides a fine-grained optimization.
- Early stopping (patience = 10): After each epoch, the model is evaluated on the held-out test set. If the test-set performance, measured by AUC, does not improve for 10 consecutive epochs, training stops and the best-performing model is used. This prevents the network from training long enough to memorize training data. Our final model stopped after 29 epochs.

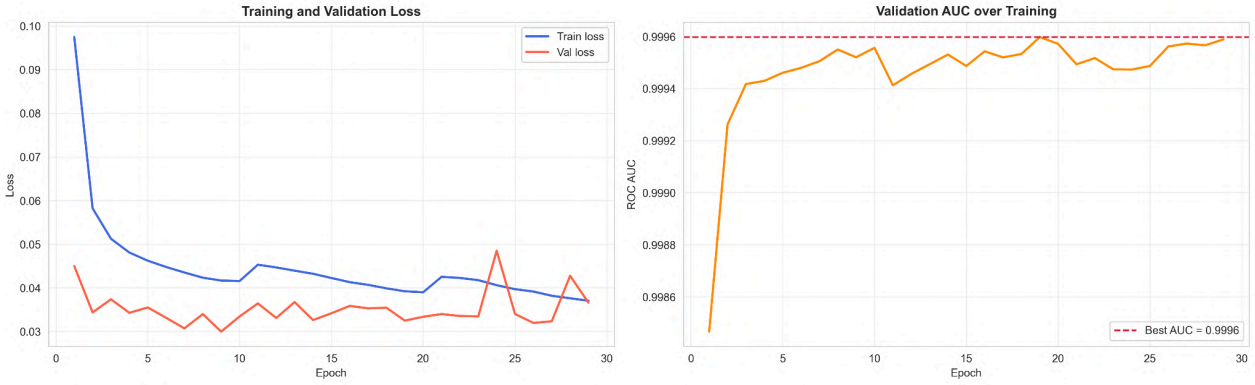


Figure 1: DNN training and validation loss (left) and validation AUC (right) per epoch. Training stopped after 29 epochs because of the early stopping implementation.

4.4. Mass Regions

During training and analysis, we divide the dilepton invariant mass spectrum into four non-overlapping Regions.

The first region used for training, the background region, is set between 40 and 71.2 GeV. The lower bound was chosen to get a clean Drell-Yan dominated region without resonance peaks from the J/ψ at 3.1 GeV and the ϒ (upsilon) at 9.5 GeV. The upper bound was chosen to avoid the Z peak at 91.2 GeV. While the Z boson has a natural width of 2.5 GeV, we use a larger buffer to avoid any Z contamination during analysis. These bounds also narrow down our background events so that the ratio of background to signal is much closer to 1. During training and analysis, we expect our models to associate these events with label=0 (not Z like).

The signal region, also used for training, is centered at the Z boson mass of 91.2 GeV and extends to ±20 GeV threshold (71.2-111.2 GeV). Since we do not use simulated Z' events, this region provides our only Z-like resonance for our search. As previously stated, this width is wide enough so that we do not rely on nearly perfect reconstructions and narrow enough so

that events from the Drell- Yan tail do not contaminate our signal classification. The lower bound matches the background region upper bound so that there is no overlap between regions but also so that all events with mass greater than 40 GeV are used.

The validation region from 111.2 GeV to 250 GeV is not used for training or search. Its purpose is to verify that the model generalizes reasonably before looking at the actual search region. The lower bound of 111.2 GeV is just above the signal region. At this mass, events are still from the high-mass tail of the Z, so they look kinematically alike to Z decays except for their higher invariant mass caused by detector effects and final state radiation. Since the model has not seen these events during training, how it scores them tell us how the learned features generalize. The upper bound of 250 GeV was determined empirically. The goal of this upper bound is to ensure that the Z tail sufficiently dies out so that a smoother power law can fit the remaining mass spectrum. Using a χ^2/ndf fit, we found that below a bound of 250 GeV, residual Z contributions create an excess that a power law cannot fit (χ^2/ndf = 3.20 at 200 GeV, 0.99 at 250 GeV). This bound is consistent with other ATLAS analyses which use 225-250 GeV as a lower bound for a high-mass search region. So, this validation region acts as a transition zone between the signal region and a smooth background. Events in this region are separated from search events specifically because they are not clean enough for a reliable background fit.

The search region, everything above 250 GeV, is where we look for a Z'. In this region, the Drell-Yan background should be smooth and falling, so that a Z' would appear as a localized narrow bump of observed events against a background fit. While there is not a specific upper bound, the invariant masses of our events do not exceed 3,000 GeV, which means there simply aren't enough events to see anything meaningful above this mass.

After training on the labeled background and signal regions, the performance of the models is evaluated with the validation region before being applied to score events in the blinded search region.

After event selection and quality cuts, there are 5,376,401 events in the background region; 5,811,472 events in the signal region; 359,621 events in the validation region; and 19,042 events in the search region.

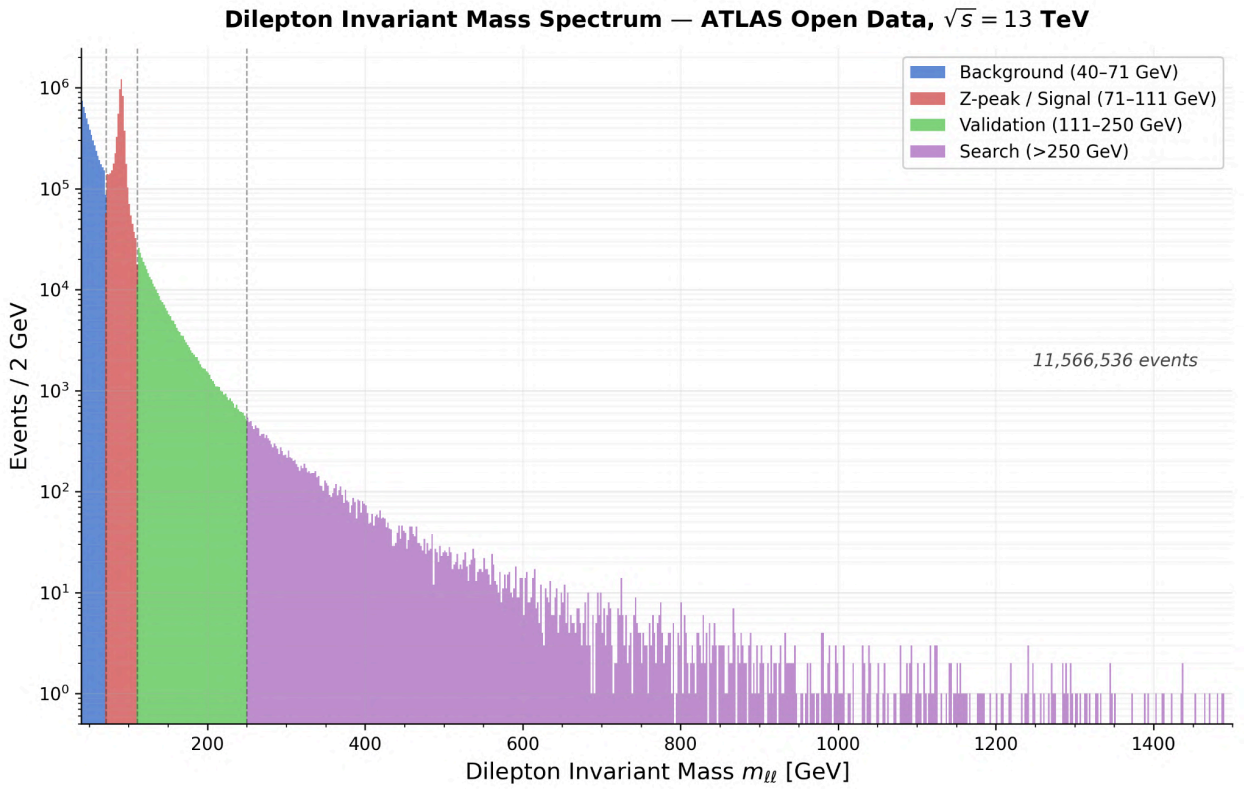


Figure 2: Dilepton invariant mass spectrum across a partial mass range, shown on a logarithmic scale. Events are color-coded by analysis region. The Z peak is at 91.2 GeV with the spectrum falling steeply at higher masses.

4.5. ROC Curve

The first metric we use to evaluate our classifiers after training is the Receiver Operating Characteristic Area Under the Curve (ROC AUC). The ROC curve plots the true positive rate (the fraction of signal events correctly identified by the classifier) against the false positive rate (the fraction of background events incorrectly labeled as signal by the classifier) as the classification threshold (the minimum score needed for an event to be classified as a signal) varies. The AUC is the area under the ROC curve and it ranges from 0.5 (essentially random guessing) to 1.0 (perfect accuracy). An AUC of 0.98, for example, means that a random signal event scores higher than a random background event 98% of the time.

The BDT achieved a test-set AUC of 0.9869, and the DNN achieved a 0.9996. These values are based on the classifiers' performance on a 20% hold-out set of the total training data, which included everything in the background and signal region. Both these values are high, which reflects the genuine kinematic differences between Z events and low-mass Drell-Yan sidebands and demonstrates that the classifiers very accurately separate these two groups. This is important, as our methodology relies on using the classifiers to find high-mass events reflecting Z-peak patterns that look very different from the background spectrum. The reason for this large difference in characteristic is because the Z boson produces

leptons with higher transverse momenta than sideband events, and this is a dominant feature driving the classification. This is confirmed by the BDT's feature importance analysis: the subleading lepton's pT accounts for 47% of the model's discriminating power, the leading lepton's pT accounts for 24%, and the other features contribute much less.

It is important to note that these AUC values reflect the classifiers' ability to distinguish the two training classes (Z-peak vs. low-mass Drell-Yan) and do not directly measure sensitivity to a Z' signal in the search region. That sensitivity is established separately through a background-fit quality and the significance scan described later.

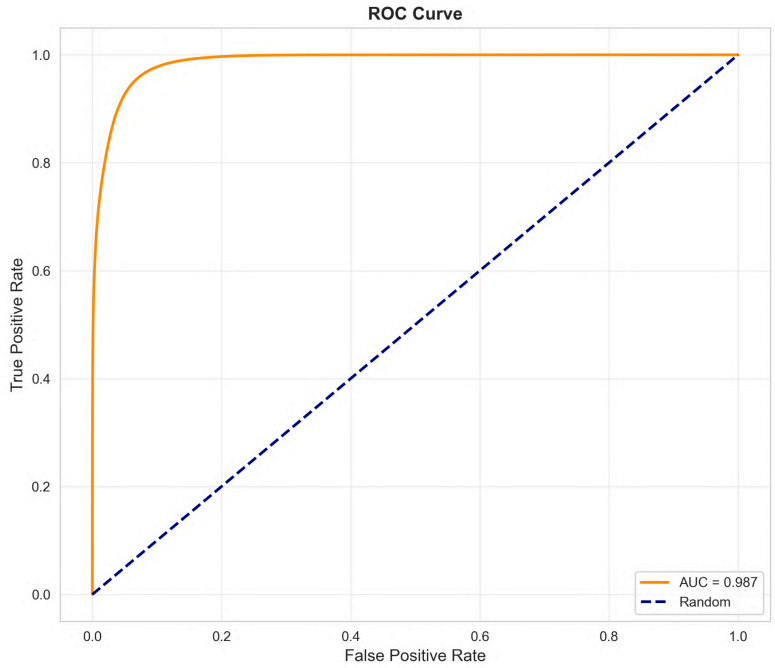


Figure 3: ROC curve on the held-out test set for the BDT classifier.

4.6. Overtraining Check

We also used performance checks for overtraining, which occurs when a model memorizes training data rather than learning general patterns, leading to good performance on training data but bad performance on unseen data. First, we verified that the model's performance on the test data mirrors its performance on training data. Then, we confirmed that score distributions evaluated on the test sets were consistent with training sets, meaning the model scores events similarly regardless of whether it has or has not seen them before.

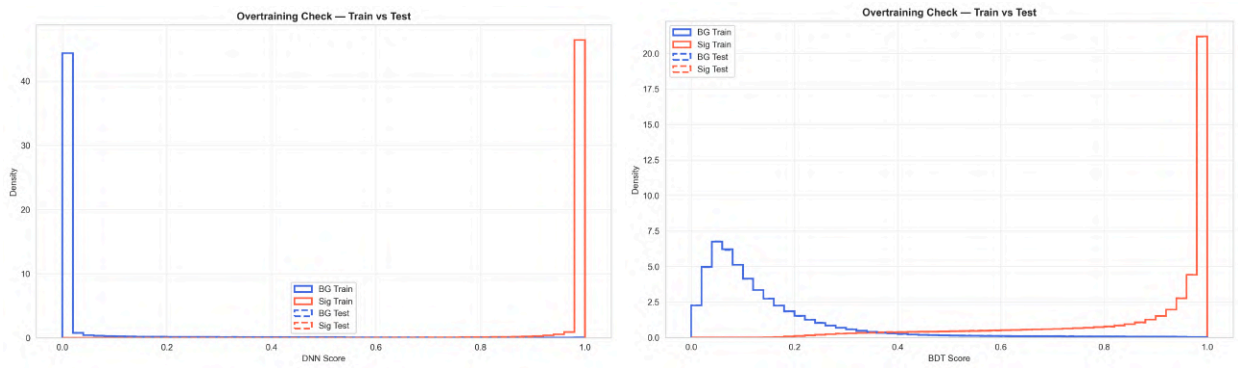


Figure 4: Overtraining check: classifier score distributions on the training and test sets for both the DNN (left) and BDT (right). Train and test distributions agree, indicating no overtraining.

4.7. Validation

Before unblinding the search region, we apply both classifiers to the 359,621 events in the validation region (111.2-250 GeV), which were not used during training. Because these events are primarily from the high-mass tail of the Z boson signal at 91.2 GeV, they look kinematically similar to Z-peak events but at higher invariant masses. The classifier's behavior in this region acts as an honest test of whether it has learned genuine Z-like features rather than simply memorizing the training mass window. In the validation region, the BDT assigns a mean score of 0.872 and the DNN assigns a mean score of 0.998, indicating that both classifiers recognize these events as strongly Z-like. Z decays are characterized by high transverse momentum, which explains why validation-region events score high (especially for the DNN): a heavier parent imparts more energy to its decay products, so validation-region events have higher lepton pT than even some Z-peak events. The score distributions transition smoothly from the signal region into the validation region with no discontinuity at the 111.2 GeV boundary, which gives us confidence that the classifiers' behavior in the search region (>250 GeV) is physically reasonable.

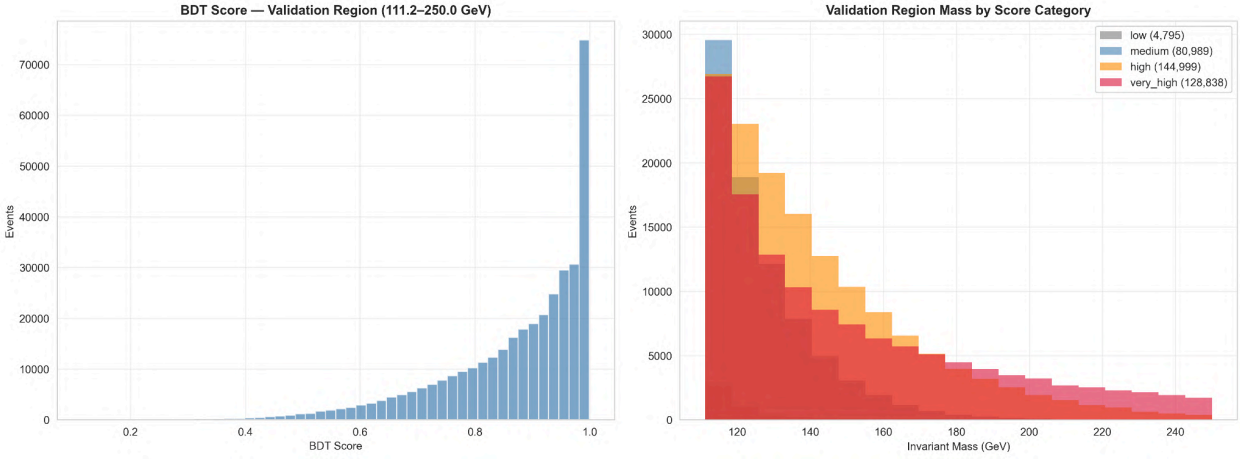


Figure 5: BDT classifier scores in the validation region (111.2-250 GeV): Distribution of BDT scores (left) and invariant mass distribution broken down by score category (right), showing a smooth transition into the search region.

5. Search and Results

5.1. Score Categories

After training and validating the classifiers, we apply them to the blinded search region starting at $m\ell\ell$ = 250 GeV. Each event is scored between 0 and 1, indicating how Z-like the classifier judges it to be. Scored events are then sorted into four categories:

- Low: score < 0.5
- Medium: 0.5 ≤ score < 0.8
- High: 0.8 ≤ score < 0.95
- Very High: score ≥ 0.95

The "very high" category includes events that the classifier considers most similar to a Z boson decay. If a Z' exists in this data and decays into lepton pairs like the Z, its events would concentrate in this category and stand out against the reduced background. This category is the primary channel for our significance scan.

5.2. Background Modeling

As explained in Section 4, we perform a three-parameter fit on the events in the "very high" category and evaluated the accuracy of the fit with a χ^2/ndf value. A χ^2/ndf close to 1 indicates the fit's acceptable agreement within statistical fluctuations. A value much greater than 1 would indicate that the fit does not accurately describe the data, probably because the classifier has distorted the mass spectrum and has started background sculpting. Our fits yield χ^2/ndf = 0.99 for the BDT and χ^2/ndf = 1.11 for the DNN in the very-high category, which confirms that neither classifier sculpts the background or creates false signatures. The likely reason that the DNN's value is further from 1 is because it flags much more events as high-scoring, which makes it harder for it to create a smooth fit with only three parameters.

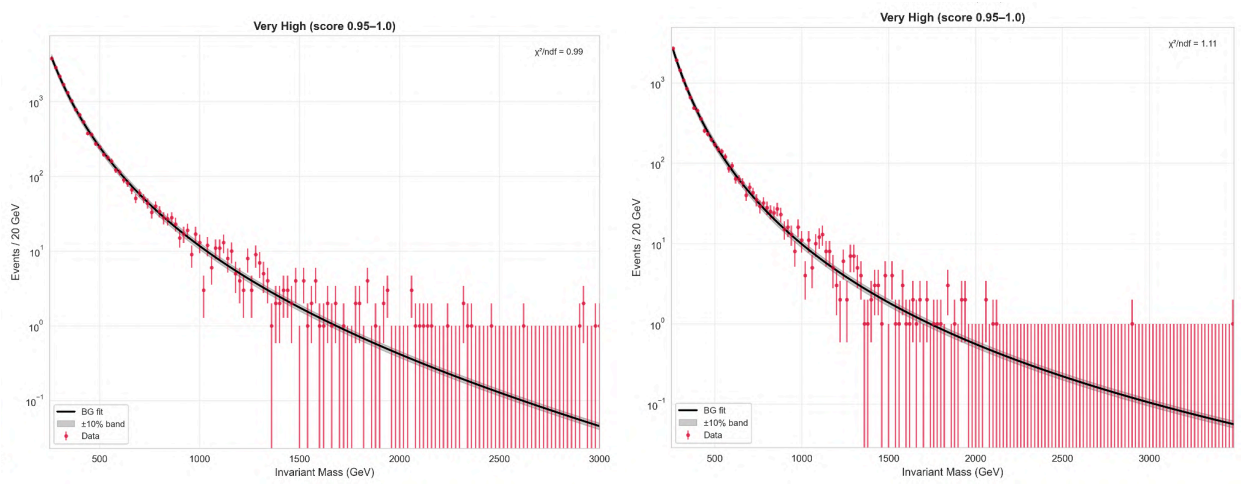


Figure 6: Three-parameter background fits in the very-high score category (score ≥ 0.95) for the BDT (left) and DNN (right), shown overlaid on the observed mass spectrum in 20 GeV bins. The fits yield $\chi^2/\mathrm{ndf} = 0.99$ and 1.11 respectively.

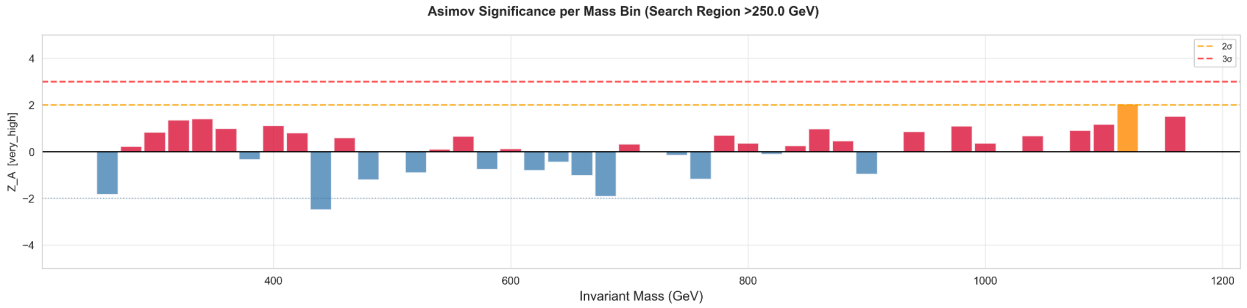
5.3. Significance Scan

A χ^2/ndf is a global evaluation of our background fit: it evaluates the success of the whole fit. To evaluate statistical significance of an observed excess in individual bins, we use an Asimov significance scan:

$$Z_A = \sqrt{\{ 2 [n \ln(n/b) - (n - b)] \}}$$

Where n is the number of observed events in a bin and b is the number of events predicted by the background fit. For both classifiers, we compute the Asimov significance per 20 GeV mass bins across the search region.

Across the search region, no bin in either classifier exceeded the conventionally required 5σ threshold for a discovery or even the 3σ threshold for evidence of a new particle. The largest significance we observed was at $m\ell\ell$ = 1120 GeV, which both classifiers marked as the most significant bin, with the BDT yielding 2.04σ and the DNN yielding 2.40σ.



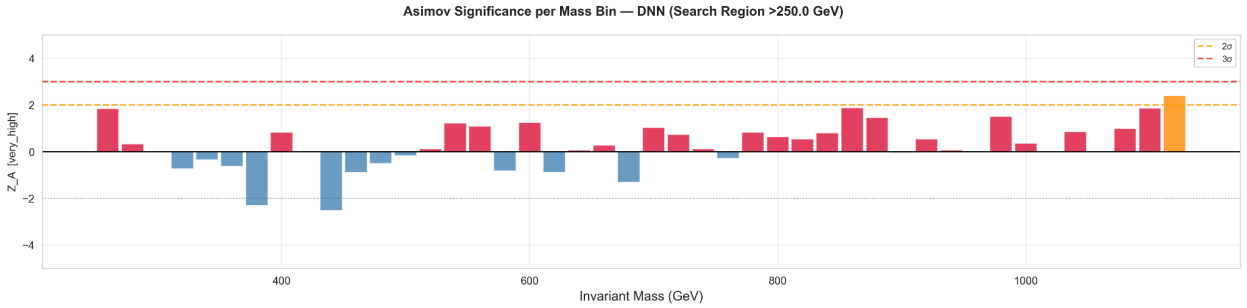


Figure 7: Asimov significance per 20 GeV bin in the very-high score category. Dashed lines indicate the 2σ and 3σ thresholds. The largest deviation was 2.40σ at mℓℓ = 1120 GeV in the DNN.

5.4. Look-Elsewhere Effect and p-value Calculations

This particular excess may be attributed to the look-elsewhere effect in particle physics, which is a tendency to find a statistically significant event simply because of a large search range. Because we scan 60 independent mass bins (20 GeV bins from 250 to 1450 GeV, which was the largest mass bin with sufficient statistics), a fluctuation of 2σ somewhere along the mass spectrum is not very surprising. We can calculate the one-sided p-values to understand the probability of finding fluctuations this large purely by chance using this formula:

$$P_global = 1 - \Phi(z)\hat{}N$$

Where P_global is the probability of finding a particular excess of significance z across N number of bins and Φ(z) is the standard normal cumulative distribution function (CDF), defined as the integral of the Gaussian. CDF values are common in statistics textbooks and we did not calculate them by hand. Performing these calculations, we find a 74.94% chance that a 2σ excess appears and a 38.98% chance that a 2.40σ excess appears at least once across all 60 bins because of statistical noise. Thus, the global significance, or the significance after considering how many bins were searched, is therefore realistically much lower than 2.40σ and even 2σ. The deviation we observed is attributed to statistical noise from the large amount of data we used, and we can safely report a null result with no evidence for a Z' boson in the mℓℓ = 250-3,000 GeV range.

6. Discussion

The main methodological goal of this analysis was to use ML classifiers in a bump-hunt search without distorting the background. We achieved this with three important design choices: excluding the invariant mass from features, avoiding any mass-normalised or mass-derived features, and training on non-overlapping mass sidebands. The $\chi^2/ndf \approx 1.00$ for both background fits in the very-high-score category provides confirmation that this strategy works well, as a classifier that learned from mass features would produce a distorted spectrum that could not be fit with a smooth three-parameter function.

While traditional approaches to bump-hunt searches involve simple kinematic cuts to select events, this ML-based approach concentrates signal-like events into higher scoring categories while preserving the smooth background shape.

This suggests that ML classifiers could be useful for improving statistical sensitivity in future searches that include systematic uncertainties.

6.1. Classifier Comparison

The DNN's test AUC (0.9996) substantially outperforms the BDT's (0.9869), reflecting the neural network's greater ability to model complex input feature combinations. This resulted in the DNN's higher sensitivity in the search region, as it yields a 2.40σ significance at mℓℓ = 1120 GeV compared to the BDT's 2.04σ. Despite the agreement between classifiers, this significance threshold is not sufficient enough for this to be considered evidence or discovery.

The two classifiers both offer their own strengths. The BDT's feature importance analysis provides insight into the classification, with a lepton pT dominated scoring (expected from a Z decay). It is more complicated to determine the decision process of the DNN because it uses more complex feature combinations. Nevertheless, presenting both classifiers and their independent results provides cross-check validations and strengthens the credibility of the result. The fact that both models agreed on a null result at the same particular excess increases the confidence in this conclusion.

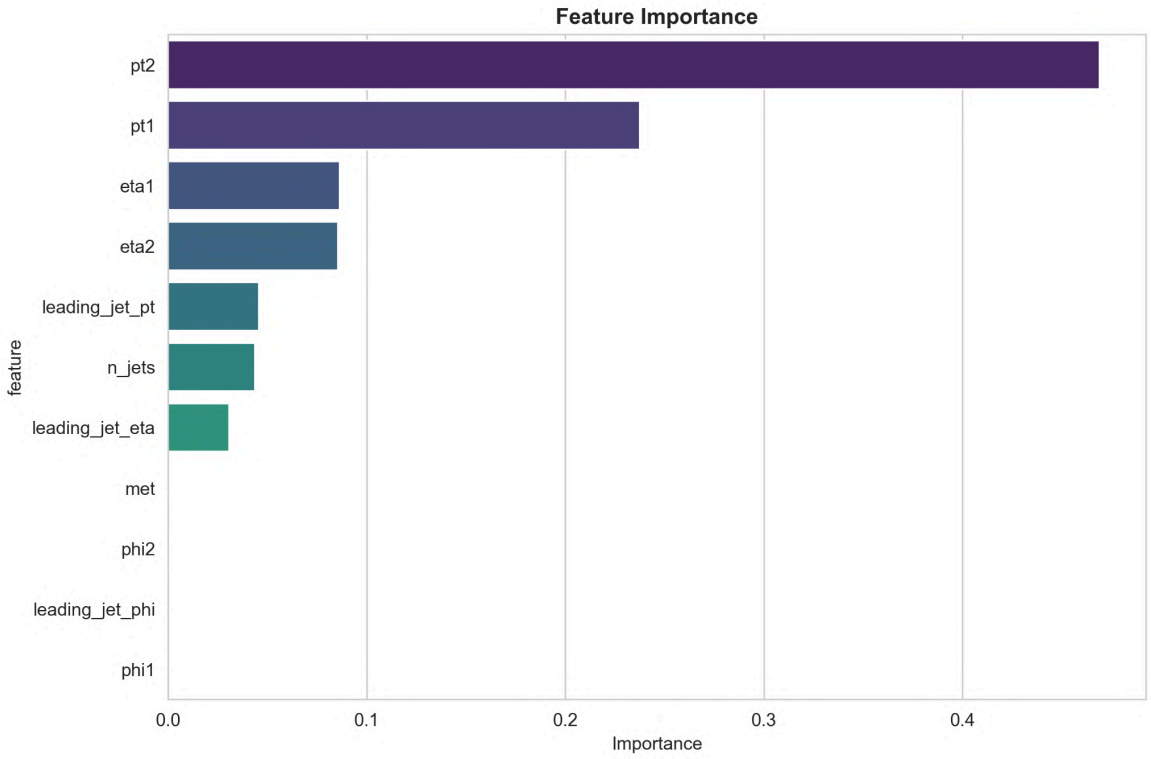


Figure 8: BDT feature importance. Lepton transverse momenta dominate, with pT^2 contributing 47% and pT^1 contributing 24%.

6.2. Limitations

This analysis is designed as a methodological demonstration rather than a full experimental result, and improvements are needed before producing results on par with official CERN and ATLAS publications:

- No comparison to theoretical models: We do not compare our results against existing benchmark theories and simulated signal events. A complete analysis would interpret the search within one or more theoretical frameworks, predict the expected signal yield for each model, and set quantitative exclusion limits on their parameters.
- Signal injection studies: To evaluate how sensitive the search is, researchers would inject simulated Z' signals at various mass points to check how often the analysis would detect them.
- Systematic uncertainties: We use only statistical uncertainties from the limited number of events in our selected bins. A complete and proper analysis would include systematic uncertainties on factors like luminosity, lepton identification efficiencies, and choice of background function.
- Alternative background functions: We use a single three-parameter function, but a professional analysis might compare several functional forms and include any variation as an additional uncertainty.
- Mass decorrelation: The DNN's very high AUC (>0.999) warrants some caution about mass correlation in the training data set. While the background fit confirms no sculpting, techniques such as adversarial training or distance-correlation penalties could provide stronger guarantees against mass correlations.

7. Conclusion

We have presented a machine-learning based search for a Z' boson using ATLAS Run 2 2015 open data, comparing a gradient boosted decision tree and a three-layer MLP trained on mass sideband kinematics. By intentionally excluding dilepton invariant mass from the input features and training on mass sidebands outside of the search region, both classifiers achieve strong discrimination between background and signal events (AUC = 0.9869 and 0.9996 respectively) while describing the high-mass background spectrum with a simple three-parameter function.

Our Asimov significance scan across bins in the search region produces no bins exceeding the 3σ evidence threshold. Our largest local deviation is 2.40σ at $m_{\ell\ell}$ = 1120 GeV using the DNN, which can be considered a statistical fluctuation when the look-elsewhere effect is considered. Thus, we report a null result with no evidence for a Z' boson in the mass range 250-3,000 GeV. It is worth emphasizing that a null result is not a failed result. Finding no excess and confirming that the Drell-Yan spectrum remains smooth across this mass range narrows the space of where new physics could be found and adds to our understanding of the universe by confirming that presently there is no evidence of a Z' boson.

The results of our methodology were successful. Using only publicly available data and simple kinematic features, we demonstrate that ML classifiers can be used in bump hunt resonance searches while preserving the background model. The full experimental process is designed to be reproducible, providing a template for future bump hunt analyses using LHC open data.

8. References

ATLAS Collaboration. (2008). The ATLAS experiment at the CERN Large Hadron Collider. Journal of Instrumentation, 3(8), S08003. https://doi.org/10.1088/1748-0221/3/08/S08003

ATLAS Collaboration. (2016). Muon reconstruction performance of the ATLAS detector in proton–proton collision data at √s = 13 TeV. European Physical Journal C, 76(5), 292. https://doi.org/10.1140/epjc/s10052-016-4120-y

ATLAS Collaboration. (2019a). Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at √s = 13 TeV. European Physical Journal C, 79(8), 639. https://doi.org/10.1140/epjc/s10052-019-7140-6

ATLAS Collaboration. (2019b). Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at √s = 13 TeV with the ATLAS detector. Physics Letters B, 796, 68–87. https://doi.org/10.1016/j.physletb.2019.07.016

ATLAS Collaboration. (2024). ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data [Data set]. CERN Open Data Portal. https://doi.org/10.7483/OPENDATA.ATLAS.AOQL.8TT3

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. Nature Communications, 5, 4308. https://doi.org/10.1038/ncomms5308

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

CMS Collaboration. (2021). Search for resonant and nonresonant new phenomena in high-mass dilepton final states at √s = 13 TeV. Journal of High Energy Physics, 2021(7), 208. https://doi.org/10.1007/JHEP07(2021)208

Collins, J. H., Howe, K., & Nachman, B. (2019). Extending the search for new resonances with machine learning. Physical Review D, 99(1), 014038. https://doi.org/10.1103/PhysRevD.99.014038

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. European Physical Journal C, 71(2), 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0

Dolen, J., Harris, P., Marzani, S., Rappoccio, S., & Tran, N. (2016). Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure. Journal of High Energy Physics, 2016(5), 156. https://doi.org/10.1007/JHEP05(2016)156

Drell, S. D., & Yan, T.-M. (1970). Massive lepton-pair production in hadron-hadron collisions at high energies. Physical Review Letters, 25(5), 316–320. https://doi.org/10.1103/PhysRevLett.25.316

Evans, L., & Bryant, P. (2008). LHC machine. Journal of Instrumentation, 3(8), S08001. https://doi.org/10.1088/1748-0221/3/08/S08001

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feickert, M., & Nachman, B. (2021). A living review of machine learning for particle physics. arXiv. https://arxiv.org/abs/2102.02770

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. European Physical Journal C, 70(1–2), 525–530. https://doi.org/10.1140/epjc/s10052-010-1470-8

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 448–456). PMLR. https://arxiv.org/abs/1502.03167

Langacker, P. (2009). The physics of heavy Z′ gauge bosons. Reviews of Modern Physics, 81(3), 1199–1228. https://doi.org/10.1103/RevModPhys.81.1199

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1711.05101

Metodiev, E. M., Nachman, B., & Thaler, J. (2017). Classification without labels: Learning from mixed samples in high energy physics. Journal of High Energy Physics, 2017(10), 174. https://doi.org/10.1007/JHEP10(2017)174

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807–814). Omnipress.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Terao, K., & Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. Nature, 560(7716), 41–48. https://doi.org/10.1038/s41586-018-0361-2

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., & McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research Section A, 543(2–3), 577–584. https://doi.org/10.1016/j.nima.2004.12.018

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Data & Code Availability Statement

The data used for this experimental setup was accessed via the ATLAS Open Data portal. Link to dataset: https://opendata.cern.ch/record/80000

Acknowledgements

I would like to thank **Professor Kaushik De** from the University of Texas at Arlington Department of Physics for his continued support and guidance during this project.

I acknowledge the work of the ATLAS Collaboration to record or simulate, reconstruct, and distribute the Open Data used in this paper, and to develop and support the software with which it was analysed.

AI Use Statement

Claude Opus 4.7 (Anthropic) was used for minor editing support and minor coding support during this project. All methodological decisions, analysis, results, and scientific content are the author's own.

Author Biography

Mihir Tare is a student at the St. Mark's School of Texas, where he enjoys studying physics and math. He leads his school's Physics Club and has represented his school in local and national competitions, including F=MA and Physics Bowl. He is particularly interested in high-energy particle physics and the Standard Model, which is why he took on this research to investigate a hypothetical particle using real CERN data and modern machine-learning strategies. Outside of physics, he enjoys playing flute in his school's band, his local youth orchestra, and in the many All-State ensembles which he has been a part of. He is also an avid Latin student who likes to compete in Classics competitions and learn about topics like Ancient Roman life, Latin vocabulary, and Latin and Greek derivatives. He was a club swimmer and water polo player and has played on his high school team, but now he likes to stay active in his free time by lifting weights, riding the bike that he built, or playing recreational sports like spikeball and volleyball. In the future, he hopes to study engineering or applied physics and wants to help discover and innovate at the forefront of science.

Mentor Contribution Statement

Professor Kaushik De of the University of Texas at Arlington served as the primary academic mentor for this project and contributed substantially to its intellectual foundation, methodological design, and scholarly presentation. He guided the author through weekly meetings throughout the project, beginning with foundational instruction in experimental particle physics. During this initial phase, Professor De provided suggested readings on the Standard Model, collider phenomena, and new-physics searches, and answered the author's conceptual questions in detail. The instruction directly shaped the project's theoretical framing and motivated the choice of the dilepton search channel.

Professor De introduced the author to the ATLAS Open Data release from LHC Run 2 and provided technical guidance on accessing, handling, and interpreting the dataset. During the computational phase of the project, he offered advice on machine learning models relevant to high-energy physics and suggested techniques to use. In weekly meetings, he reviewed results with the author and gave advice on how to interpret them.

Finally, Professor De reviewed and revised multiple drafts of the final paper, providing feedback on technical accuracy, clarity, and structure prior to submission.

Initial Editorial Assessment for *"*Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data*"*

The manuscript presents an interesting contribution and appears to have the potential to proceed to external peer review. However, before the paper is sent to reviewers, several issues related to scholarly presentation and readability should be addressed.

1. Missing in-text citations:

The manuscript currently provides a reference list but does not include inline citations within the main text. Appropriate in-text citations should be added throughout the manuscript to clearly identify the sources supporting specific statements and to place the work within the context of the relevant literature.

2. Figure readability

The axis labels in all figures, with the exception of Figure 2, are difficult to read due to their small font size. Increasing the size of the axis labels and ensuring consistent formatting across all figures would improve the overall readability of the manuscript.

These revisions are relatively straightforward but are important for ensuring that the manuscript meets the journal's standards for presentation. The editorial team therefore recommends that the authors address the points above and submit a revised version for further evaluation prior to external peer review.

Convergence
A Journal for Young Researchers

Received: [for official use only] ; Accepted: [for official use only];
Managing Editor: [for official use only];
Action Editor: [for official use only]

Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data

[name redacted]

[redacted]

Abstract

The Standard Model of particle physics, while highly successful, leaves several unanswered questions that motivate searches for undiscovered heavy gauge bosons predicted by Beyond-Standard-Model theories. Using machine-learning classifiers trained on low-mass sideband events, this paper presents a search for the Z' boson, one such hypothetical heavy neutral gauge boson that decays into dilepton pairs. A gradient boosted decision tree (BDT) and a multi-layered perceptron deep neural network (DNN) are trained to distinguish real Z boson events from Drell-Yan background events using only 11 raw kinematic features, with dilepton invariant mass deliberately excluded to avoid background sculpting of artificial signals during the search. The analysis uses real proton-proton collision data recorded by the ATLAS experiment during LHC Run 2 in 2015, released through the CERN Open Data portal in 2024. After both classifiers achieve accurate discrimination on the training set, events in a blinded high-mass search region are scored by the classifiers, a three-parameter parametric fit is used to model the background, and an Asimov significance scan is performed across mass bins. We observed no statistically significant excess, with the largest local deviation at a mass of 1120 GeV with a significance of 2.40σ in the DNN's highest-score category, which is well below the 3σ conventionally required for evidence of new physics. This analysis demonstrates how machine-learning can be used in high-energy physics "bump hunt" searches while preserving the background model, and it provides a reproducible methodology built on publicly available data.

Keywords:
Z′ boson, dilepton resonance search, machine learning, boosted decision tree, deep neural network, ATLAS Open Data, bump hunt, beyond the standard model

1. Goals and Theory

The Standard Model (SM) of particle physics provides a remarkably successful description of the fundamental constituents of matter and the forces that govern their interactions. Despite its predictive power, the SM is widely understood to be an incomplete theory. It does not account for observed matter-antimatter asymmetry of the universe, provides no dark matter candidate, fails to incorporate gravity at the quantum level, and leaves the mass hierarchy of fermions unexplained.

A variety of Beyond Standard Model (BSM) theories address these shortcomings by hypothesizing the existence of new heavy gauge bosons that come from extensions of the Standard Model's electroweak symmetry, the mathematical structure that organizes how particles interact with the weak nuclear force. These hypothetical particles are called Z' bosons (Z-prime bosons). A Z' boson is generally understood to be a heavier counterpart of the Z boson, a known particle that mediates the weak nuclear force and was discovered at CERN in 1983. Many BSM extensions predict that the Z' would couple to leptons and quarks just like the Z but at a higher mass scale (Langacker, 2009).

Previous searches at the LHC have set lower mass limits on Z' bosons in the multi-TeV range for the most common benchmark models. Both ATLAS and CMS have performed high-mass dilepton searches using the full Run 2 dataset of 139 fb-1 and found no evidence of a Z' (ATLAS Collaboration, 2019; CMS Collaboration, 2021). An undiscovered resonance remains possible, particularly at masses or coupling strengths below the sensitivity of these previous analyses.

Our search utilizes the dilepton channel because of its simplicity and clarity. As a hypothetical "cousin" of the Z boson, the Z' should decay into a pair formed by a lepton and its corresponding anti-lepton (e.g. an electron and positron). Utilizing this channel, often called the "golden channel," allows for a clean experimental signature and a well-understood background from the Drell-Yan process (Drell & Yan, 1970). If a Z' exists, it would appear as a localized excess of events, or a "bump", in the dilepton invariant mass spectrum at the Z' mass, sitting on top of the smooth Drell-Yan continuum. This "bump-hunt" strategy is the standard approach to discovering new resonances at colliders.

Machine learning is now a standard tool in particle physics for classifying collision events (Feickert & Nachman, 2021; Radovic et al., 2018). However, applying machine learning to bump-hunt searches comes with a concern: if a classifier learns to reconstruct invariant mass (even indirectly through correlated features), it can distort the smooth background spectrum and create artificial peaks where no real signal exists. This problem, known as background sculpting (Dolen et al., 2016), is the primary failure this analysis is designed to avoid.

The goal of this work is to develop and validate a machine learning pipeline for a Z' boson search using LHC data, following the methodological structure of real collider analyses, and to demonstrate a statistically valid null result or identify candidate excesses.

2. Experimental Background

2.1. The LHC

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator (Evans & Bryant, 2008). It accelerates two beams of protons in opposite directions to nearly the speed of light, then brings them into head-on

collision at designated interaction points. During Run 2 (2015-2018), the LHC operated at a center-of-mass energy of $\sqrt{s}$ = 13 TeV, meaning the total energy available in each proton-proton collision is 13 trillion electron-volts. This analysis uses data collected in 2015 during the first year of Run 2.

2.2. ATLAS Detector

The ATLAS (A Toroidal LHC Apparatus) detector is a large general-purpose particle detector at the LHC (ATLAS Collaboration, 2008). It is designed to record the debris from high-energy proton-proton collisions and reconstruct what happened in each event using software and sensors. The detector is built in concentric layers around the collision point, with each layer optimized to measure different types of particles:

- Inner Detector: The innermost layer immersed in a strong magnetic field. Charged particles curve as they pass through the field, and the curvature is used to measure their momentum. This is the primary system for tracking leptons such as electrons and muons.
- Calorimeters: These layers, surrounding the inner detector, absorb particles and measure their energy. The electromagnetic calorimeter measures electrons and photons, while the hadronic calorimeter measures jets (sprays of particles produced by quarks and gluons).
- Muon spectrometer: Muons pass through calorimeters unaffected (because they are heavy leptons, their momentum is too high for calorimeters to stop them), so this dedicated system in the outermost layer identifies and measures them with high precision using a toroidal magnetic field.

2.3. Dataset

This analysis uses Run 2 2015 proton-proton collision data made publicly available through the CERN Open Data Portal in 2024 (ATLAS collaboration, 2024). The data is provided in DAOD-PHYSLITE format, an analysis-ready format in which the ATLAS experiment collaboration has already performed computationally heavy work like identifying particles. This full dataset contains 10,049 files with a total 9.3 TiB of approximately 1.69 x 10^9 events and has a luminosity of 3.2 fb^{-1}.

3. Data Preparation

3.1. Event Selection

We select events containing exactly two same-flavour leptons: dielectron events (with exactly two electrons) and dimuon events (with exactly two muons). For each dilepton event, the following reconstructed objects are extracted:

- Lepton Kinematics: Using analysis-level lepton objects (AnalysisElectronsAuxDyn and AnalysisMuonsAuxDyn), the transverse momentum pT (the component of momentum perpendicular to the beam), pseudorapidity η, and azimuthal angle φ of each lepton is gathered.
- Jet Kinematics: The three same kinematic variables are collected for the leading reconstructed jet (AnalysisJetsAuxDyn), plus the jet multiplicity n_jets (how many jets were found in the event).
- Missing Transverse Energy (MET): The apparent imbalance in the total transverse momentum of the event, the MET can signal potential undetected particles like neutrinos. It is reconstructed from the x and y components of the missing momentum vector (MET_Core_AnalysisMETAuxDyn.mpx/mpy).

All momenta are provided in MeV by ATLAS and are converted to GeV for this analysis.

3.2. Invariant Mass

The dilepton invariant mass $m\ell\ell$ is an important value for this search. If two leptons originate from a parent particle decay (like a Z' boson decaying into two leptons), then the dilepton invariant mass should equal the mass of the parent particle. Thus, if a Z' boson exists, it would produce a peak of events at a particular value in the invariant mass spectrum corresponding to its mass. This is what physics "bump hunts" typically look for. Dilepton Invariant mass is calculated using the three kinematic variables that we collected for each lepton (pT, η, φ):

$$m\ell\ell = \sqrt{(2\, pT^1\, pT^2\, [\, \cosh(\eta^1 - \eta^2) - \cos(\varphi^1 - \varphi^2)\,]\,)}$$

3.3. Quality Cuts

We apply quality cuts that are standard to the LHC. A pseudorapidity cut $|\eta| \leq 2.5$ is applied to electrons and muons, reflecting the acceptance of the ATLAS inner detector and calorimeter. For muons, this cut reflects limitations of the Semiconductor Tracker (SCT) of the ATLAS Muon Spectrometer and the Inner Detector, which are used to measure muon momentum (ATLAS Collaboration, 2016). For electrons, the ATLAS electromagnetic calorimeter precision region is $|\eta| \leq 2.47$, which makes this analogous cut slightly conservative but still valid (ATLAS Collaboration, 2019a). Jet variables are undefined for events where n_jets = 0, so the pseudorapidity cut is conditionally applied to jet objects when n_jets > 0.

A hard cut of $m\ell\ell \leq 3{,}500$ GeV is applied because the background spectrum falls very steeply, and so past this boundary, there are hardly any events in the 20 GeV bins that our fit uses. We account for this problem with a minimum of 10 events for a bin to be used, but this hard cut prevents us from even looking where we expect zero events.

4. Machine Learning Methodology

The machine learning methodology has three stages. First, we train two classifiers, a boosted decision tree (BDT) and a deep neural network (DNN), to distinguish Z-peak events from low-mass Drell-Yan background events using only kinematic features, with the dilepton invariant mass deliberately excluded so that the classifiers cannot learn the search variable. Both are trained and validated on events drawn exclusively from the low-mass sidebands and the Z-peak region, well below the search region, so that the search region itself remains blinded during model development.

Second, we apply the trained classifiers to the high-mass search region ($m\ell\ell > 250$ GeV) and sort every event into categories by classifier score, concentrating the most Z-like events into a single high-scoring category. Within that category, we model the smooth Drell-Yan background with a three-parameter parametric fit:

$$f(m) = p_0 \bullet (m/\sqrt{s})\hat{}p_1 \bullet \exp(-p_2 \bullet m/\sqrt{s})$$

This function uses a power law combined with an exponential decay and is a standard parametrization used in ATLAS analyses (ATLAS Collaboration, 2019b). The three parameters (p_0, p_1, p_2) are determined by fitting the function to the observed mass spectrum in 20 GeV bins using χ^2 minimization, which is a statistical procedure that adjusts the parameters to minimize the sum of squared residuals between the fit and the data, weighted by the uncertainty in each

bin.

We evaluate the accuracy of the fit with reduced chi-squared statistic, χ^2/ndf, which is a sum over all bins of squared deviations. χ^2 is evaluated as $\chi^2 = \Sigma_i (n_i - b_i)^2 / \sigma_i^2$, where n_i is the number of observed events in a bin, b_i is the number of events predicted to be in the bin by the background fit, and σ_i is a variance calculation. Ndf (number of degrees of freedom) is calculated by subtracting the number of parameters in the fit (three in our case) from the number of bins we search. This background spectrum should remain smooth and falling, and a good fit will confirm that the classifiers have not sculpted artificial structure into it.

Finally, we perform a bin-by-bin Asimov significance scan within the high-score category, comparing the observed number of events in each bin against the background fit to test for localized excesses. The remainder of this section describes the input features, the two classifier architectures, and their training process.

4.1. Features

The core idea of this methodology is to train machine-learning classifiers to distinguish Z-like events from non-Z-like events using only kinematic features. Specifically, the momenta, angles, jet activity, and missing transverse energy of each event. The trained classifiers are then applied to the search region to identify which high-mass events look most like a Z boson and then concentrate these potential Z' signals into a high-score category.

One critical design choice for the input features is that dilepton invariant mass is excluded, as including it would allow the classifiers to directly learn the search variable and potentially sculpt the background to create artificial "bumps" in the mass spectrum where no real signal exists.

The same 11 input features are used for both classifiers. They are listed with their descriptions in Table 1.

Table 1: Input kinematic features for the models and their descriptions

Feature	Meaning
pT^1, pT^2	Transverse momenta of the leading and subleading leptons (electron or muon)
η^1, η^2	Pseudorapidities of both leptons
φ^1, φ^2	Azimuthal angles of both leptons
n_jets	Reconstructed jet multiplicity
leading_jet_pT	Transverse momentum of the leading jet (zero if n_jets = 0)
leading_jet_η	Pseudorapidity of the leading jet (zero if n_jets = 0)
leading_jet_φ	Azimuthal Angle of the leading jet (zero if n_jets = 0)

MET	Missing transverse energy

Both classifiers are trained as binary classifiers, with Z events labeled as signal (class 1) and low-mass Drell-Yan events labeled as background (class 0) (Collins et al., 2019; Metodiev et al., 2017). The labeled sample of training events is split 80/20 into training and test sets using stratified sampling (preserving class ratio in both splits), giving 8,950,298 training events and 2,237,575 held-out test events for validation.

4.2. Boosted Decision Tree (BDT)

The BDT, an ensemble of gradient-boosted decision trees that has become a standard classifier in particle physics (Friedman, 2001; Roe et al., 2005), uses binary cross-entropy (BCE), a loss function for binary classification where loss increases exponentially when the difference between the model's prediction and the event's actual label increases. For an event with true label y (either 0 or 1) and predicted probability p, the BCE is:

$$L = -[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)]$$

Loss is small when the model's prediction matches the correct label of the event and large when the model is incorrect with high confidence. Each successive tree is fit to the gradient of this loss so that the ensemble progressively reduces its overall error. This loss calculation is done automatically using XGBoost, a scalable open-source gradient-boosting library implemented in Python through its xgboost package, which handles tree construction, gradient computation, and parallelized training internally (Chen & Guestrin, 2016). All preprocessing of input features, train/test splitting, and evaluation are carried out in Python using NumPy, pandas, and scikit-learn.

To prevent the model from overfitting, we apply several forms of regularization using XGBoost parameters.

- Number of trees and learning rate: We train 300 trees with a low learning rate of 0.03. The learning rate scaled down each tree's contribution, so that each tree makes a small correction rather than aggressively fitting the training data. A low rate combined with many trees sacrifices training speed for better generalization.
- Maximum depth of 5: Each tree is limited to 5 levels deep, which produces at most 32 distinct regions. Unlike deep trees, which tend to memorize individual events, shallow trees learn broad patterns.
- Minimum child weight of 10: In each tree, a leaf node can only be split if both children would contain at least 10 units of statistical weight. This prevents the model from learning patterns from small groups of unusual events.
- Minimum split gain of 1.0: Splits only happen if they reduce the loss by at least 1.0. This "prunes" each tree by removing redundant branches, which helps reduce overfitting.
- L1 regularization/Lasso regression (reg_alpha = 0.1): This essentially adds a penalty proportional to the absolute value of each leaf's prediction. This drives the weight of less important features to zero and simplifies the model by essentially performing a feature selection.
- L2 regularization (reg_lambda = 5.0): Adds a penalty proportional to the square of a leaf's prediction. It shrinks large weights toward zero and strongly discourages overconfident predictions.
- Stochastic gradient boosting (70%): Each tree is trained on a random 70% of events using a random 70% of features. This randomness prevents all trees from making the same mistakes.

The class imbalance between labeled signal events (5,811,472) and labeled background events (5,376,401) is handled by a scale_pos_weight parameter of 0.93 (calculated ratio), which decreases the weight of signal class during training.

4.3. Deep Neural Network

Neural networks (LeCun et al., 2015), and in particular deep architectures applied to collider classification problems (Baldi et al., 2014), are built from computational units called neurons. Each neuron takes in a set of numbers and computes a weighted sum based on weights that are learned during training. The result is passed through a nonlinear activation function. Our model uses the ReLU (Rectified Linear Unit) activation function, which outputs the input if it is positive and zero otherwise. ReLU and the DNN architecture is implemented with PyTorch, an open-source deep-learning framework in Python that provides automatic differentiation, GPU-accelerated tensor operations, and prebuilt neural-network modules. The network is defined as a torch.nn.Module subclass, with linear layers, batch normalization, dropout, and ReLU activations composed using the torch.nn API (Nair & Hinton, 2010; Paszke et al., 2019).

Neurons are arranged in multiple layers. In a fully connected layer, every neuron receives input from every neuron in the previous layer. Unlike graph-based architectures designed to find structural relations in inputs, this multi-layered perceptron (MLP) does not explicitly encode invariant relationships like mass between particles. Despite this, correlated kinematic information may still allow partial reconstruction of invariant mass, motivating our validation against background sculpting (which is discussed in the background modelling section). We initially tried to use a Graph Neural Network (GNN), but it was able to reconstruct invariant mass and scored events almost solely based on mass. The switch to this multi-layered perceptron eliminates this background sculpting problem which we were trying to avoid.

Our network consists of an input standardization layer, three fully connected hidden layers of 128 neurons each, and a single output neuron.

- Standardization layer: Before entering the network, each of the 11 input features is transformed to have zero mean and unit standard deviation by subtracting the mean and dividing by the standard deviation calculated only for training sets. Since the raw features have different scales (pT can be hundreds of GeV, while φ is bounded between -π and π), this ensures that features with larger magnitudes do not distort the learning process with large weights.
- Hidden layers: three layers of 128 neurons each. Each layer applies a linear transformation, batch normalization (which recenters and rescales the activations across each training batch to stabilize learning) (Ioffe & Szegedy, 2015), and ReLU activation. Each step is implemented using PyTorch. The depth and width of the network give it the capacity to learn complex nonlinear combinations of input features.
- Dropout (p=0.3): After each hidden layer, 30% of neurons are randomly "switched off" during each training step. This prevents individual neurons from becoming dependent on one another and improves their individual reliability. Dropout prevents the model as a whole from overfitting and memorizing training noise (Srivastava et al., 2014). When testing, however, all neurons are active.
- Output: In the last layer, a single neuron produces an unbounded score, which is passed through a sigmoid function to predict a probability between 0 (background) and 1 (signal).

Training occurs in epochs, where one epoch is a full pass through the training data. For each epoch, the data is divided into batches of 4,096 events. For each batch, the network makes predictions, the loss is computed, and the gradient of the loss

with respect to every weight is computed using chain rule (backpropagation). The weights are then updated according to this gradient. We do clip gradients at 1.0 so that large updates to weights do not cause instability. Over many epochs, the DNN converges toward a set of weights that minimizes the loss.

We optimize our DNN with AdamW (Loshchilov & Hutter, 2019), a variant of the Adam optimizer that decouples weight decay from the gradient update. We use a learning rate of 10^{-3} and weight decay of 10^{-4} (an L2 like penalty on all weights that discourages overfitting). AdamW is provided directly by the torch.optim module, and gradients are computed via PyTorch's automatic differentiation engine during backpropagation. We also use two additional techniques which are also implemented using the PyTorch library:

- Cosine annealing with warm restarts (Loshchilov & Hutter, 2017): The learning rate decreases smoothly following a cosine curve over the course of 10 epochs, then resets back to its initial value and begins decaying again. These periodic resets allow the optimizer to escape shallow local minima that it might otherwise get trapped in, while the cosine decay within each cycle provides a fine-grained optimization.
- Early stopping (patience = 10): After each epoch, the model is evaluated on the held-out test set. If the test-set performance, measured by AUC, does not improve for 10 consecutive epochs, training stops and the best-performing model is used. This prevents the network from training long enough to memorize training data. Our final model stopped after 29 epochs.

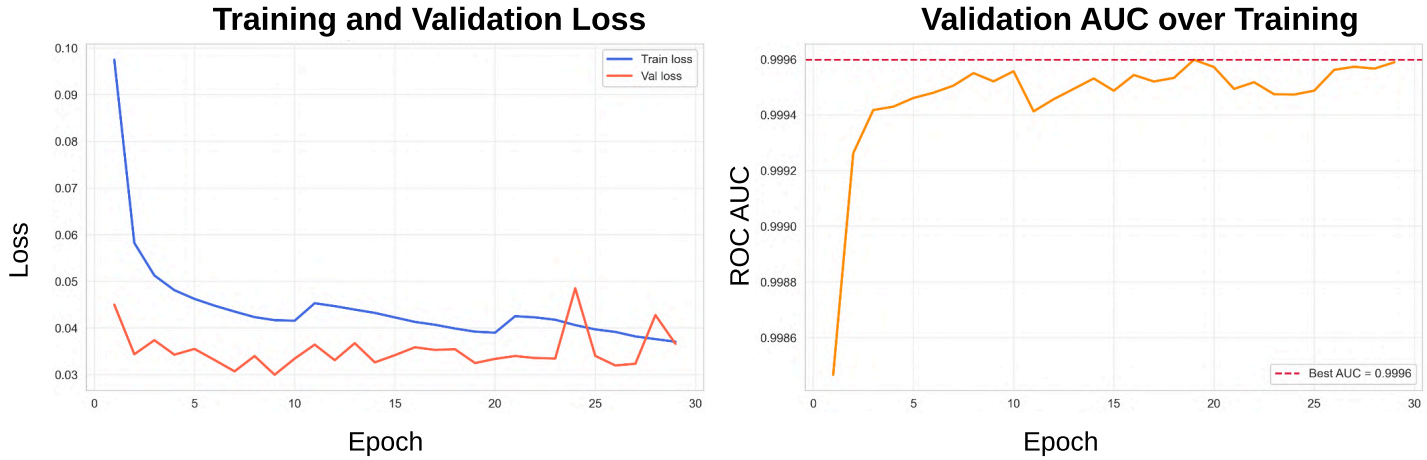


Figure 1: DNN training and validation loss (left) and validation AUC (right) per epoch. Training stopped after 29 epochs because of the early stopping implementation.

4.4. Mass Regions

During training and analysis, we divide the dilepton invariant mass spectrum into four non-overlapping Regions.

The first region used for training, the background region, is set between 40 and 71.2 GeV. The lower bound was chosen to

get a clean Drell-Yan dominated region without resonance peaks from the J/ψ at 3.1 GeV and the Y (upsilon) at 9.5 GeV. The upper bound was chosen to avoid the Z peak at 91.2 GeV. While the Z boson has a natural width of 2.5 GeV, we use a larger buffer to avoid any Z contamination during analysis. These bounds also narrow down our background events so that the ratio of background to signal is much closer to 1. During training and analysis, we expect our models to associate these events with label=0 (not Z like).

The signal region, also used for training, is centered at the Z boson mass of 91.2 GeV and extends to ±20 GeV threshold (71.2-111.2 GeV). Since we do not use simulated Z' events, this region provides our only Z-like resonance for our search. As previously stated, this width is wide enough so that we do not rely on nearly perfect reconstructions and narrow enough so that events from the Drell- Yan tail do not contaminate our signal classification. The lower bound matches the background region upper bound so that there is no overlap between regions but also so that all events with mass greater than 40 GeV are used.

The validation region from 111.2 GeV to 250 GeV is not used for training or search. Its purpose is to verify that the model generalizes reasonably before looking at the actual search region. The lower bound of 111.2 GeV is just above the signal region. At this mass, events are still from the high-mass tail of the Z, so they look kinematically alike to Z decays except for their higher invariant mass caused by detector effects and final state radiation. Since the model has not seen these events during training, how it scores them tell us how the learned features generalize. The upper bound of 250 GeV was determined empirically. The goal of this upper bound is to ensure that the Z tail sufficiently dies out so that a smoother power law can fit the remaining mass spectrum. Using a χ^2/ndf fit, we found that below a bound of 250 GeV, residual Z contributions create an excess that a power law cannot fit (χ^2/ndf = 3.20 at 200 GeV, 0.99 at 250 GeV). This bound is consistent with other ATLAS analyses which use 225-250 GeV as a lower bound for a high-mass search region. So, this validation region acts as a transition zone between the signal region and a smooth background. Events in this region are separated from search events specifically because they are not clean enough for a reliable background fit.

The search region, everything above 250 GeV, is where we look for a Z'. In this region, the Drell-Yan background should be smooth and falling, so that a Z' would appear as a localized narrow bump of observed events against a background fit. While there is not a specific upper bound, the invariant masses of our events do not exceed 3,000 GeV, which means there simply aren't enough events to see anything meaningful above this mass.

After training on the labeled background and signal regions, the performance of the models is evaluated with the validation region before being applied to score events in the blinded search region.

After event selection and quality cuts, there are 5,376,401 events in the background region; 5,811,472 events in the signal region; 359,621 events in the validation region; and 19,042 events in the search region.

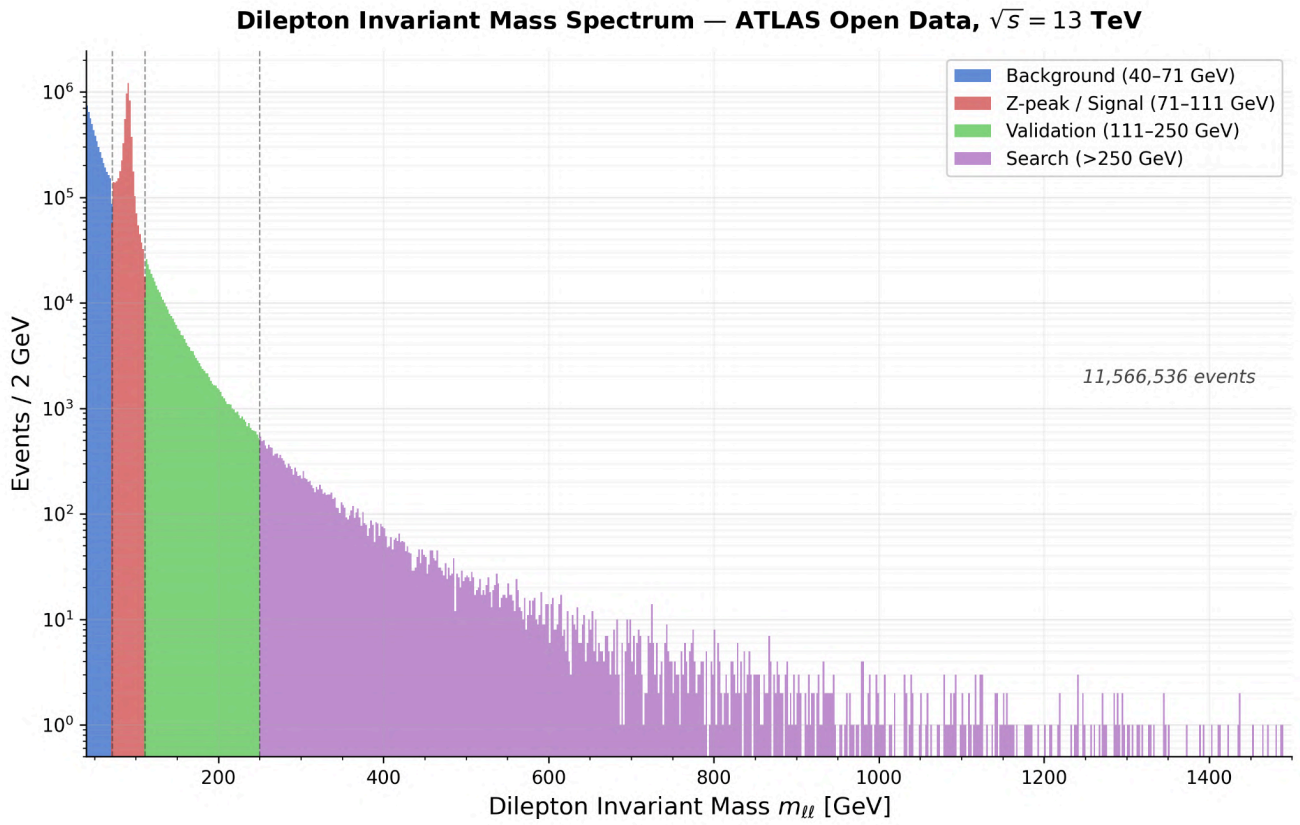


Figure 2: Dilepton invariant mass spectrum across a partial mass range, shown on a logarithmic scale. Events are color-coded by analysis region. The Z peak is at 91.2 GeV with the spectrum falling steeply at higher masses.

4.5. ROC Curve

The first metric we use to evaluate our classifiers after training is the Receiver Operating Characteristic Area Under the Curve (ROC AUC) (Fawcett, 2006). The ROC curve plots the true positive rate (the fraction of signal events correctly identified by the classifier) against the false positive rate (the fraction of background events incorrectly labeled as signal by the classifier) as the classification threshold (the minimum score needed for an event to be classified as a signal) varies. The AUC is the area under the ROC curve and it ranges from 0.5 (essentially random guessing) to 1.0 (perfect accuracy). An AUC of 0.98, for example, means that a random signal event scores higher than a random background event 98% of the time.

The BDT achieved a test-set AUC of 0.9869, and the DNN achieved a 0.9996. These values are based on the classifiers' performance on a 20% hold-out set of the total training data, which included everything in the background and signal region. Both these values are high, which reflects the genuine kinematic differences between Z events and low-mass Drell-Yan sidebands and demonstrates that the classifiers very accurately separate these two groups. This is important, as our methodology relies on using the classifiers to find high-mass events reflecting Z-peak patterns that look very different

from the background spectrum. The reason for this large difference in characteristic is because the Z boson produces leptons with higher transverse momenta than sideband events, and this is a dominant feature driving the classification. This is confirmed by the BDT's feature importance analysis: the subleading lepton's pT accounts for 47% of the model's discriminating power, the leading lepton's pT accounts for 24%, and the other features contribute much less.

It is important to note that these AUC values reflect the classifiers' ability to distinguish the two training classes (Z-peak vs. low-mass Drell-Yan) and do not directly measure sensitivity to a Z' signal in the search region. That sensitivity is established separately through a background-fit quality and the significance scan described later.

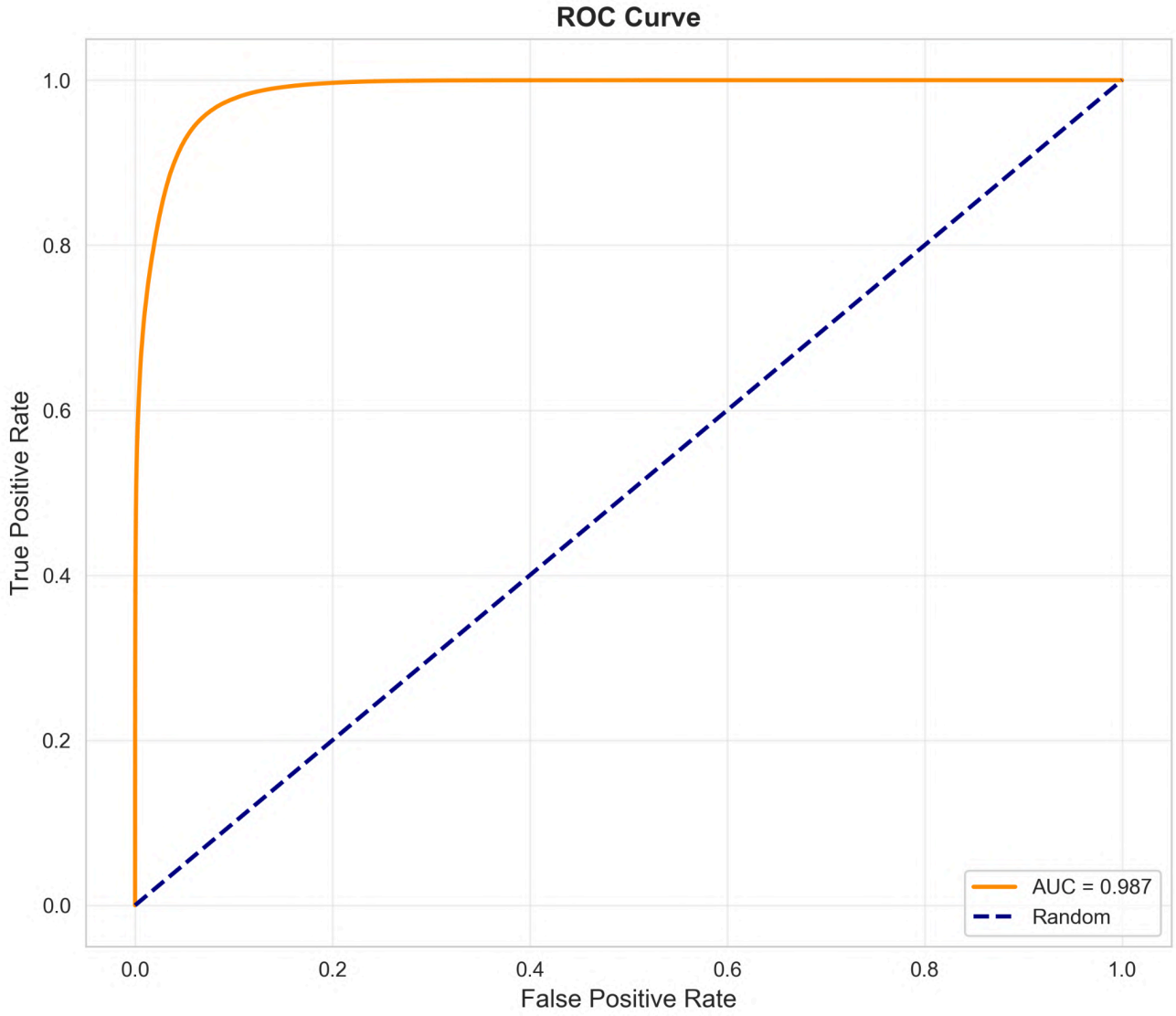
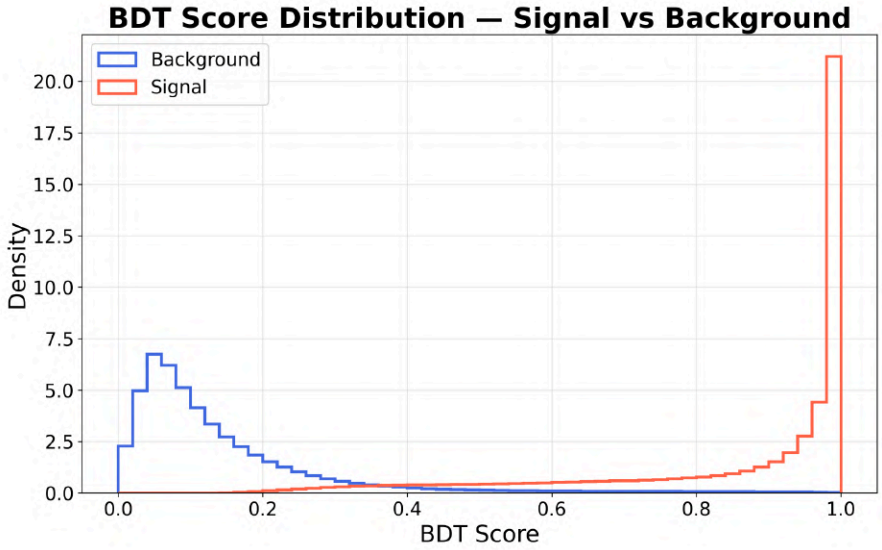


Figure 3: ROC curve on the held-out test set for the BDT classifier.

4.6. Overtraining Check

We also used performance checks for overtraining, which occurs when a model memorizes training data rather than learning general patterns, leading to good performance on training data but bad performance on unseen data. First, we verified that the model's performance on the test data mirrors its performance on training data. Then, we confirmed that score distributions evaluated on the test sets were consistent with training sets, meaning the model scores events similarly regardless of whether it has or has not seen them before.



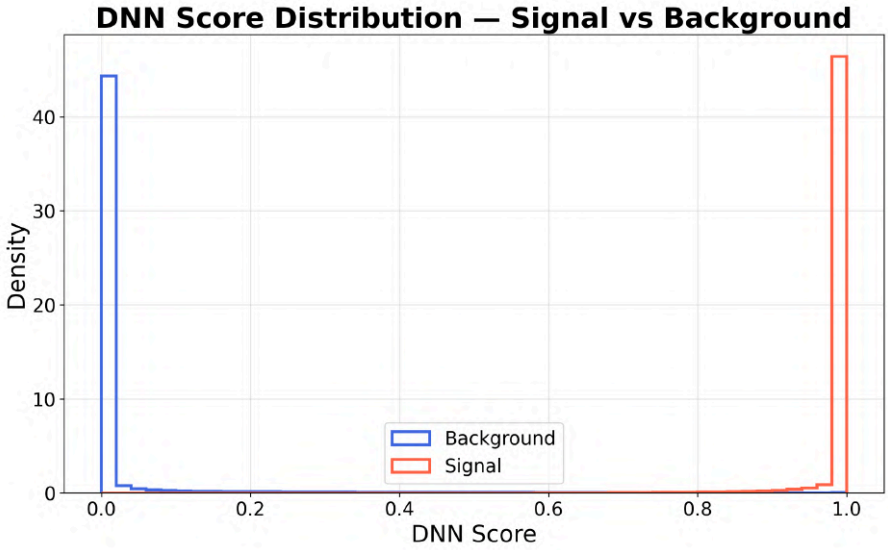


Figure 4: Overtraining check: classifier score distributions on the training and test sets for both the DNN (right) and BDT (left). Train and test distributions agree, indicating no overtraining.

4.7. Validation

Before unblinding the search region, we apply both classifiers to the 359,621 events in the validation region (111.2-250 GeV), which were not used during training. Because these events are primarily from the high-mass tail of the Z boson signal at 91.2 GeV, they look kinematically similar to Z-peak events but at higher invariant masses. The classifier's behavior in this region acts as an honest test of whether it has learned genuine Z-like features rather than simply memorizing the training mass window. In the validation region, the BDT assigns a mean score of 0.872 and the DNN assigns a mean score of 0.998, indicating that both classifiers recognize these events as strongly Z-like. Z decays are characterized by high transverse momentum, which explains why validation-region events score high (especially for the DNN): a heavier parent imparts more energy to its decay products, so validation-region events have higher lepton pT than even some Z-peak events. The score distributions transition smoothly from the signal region into the validation region with no discontinuity at the 111.2 GeV boundary, which gives us confidence that the classifiers' behavior in the search region (>250 GeV) is physically reasonable.

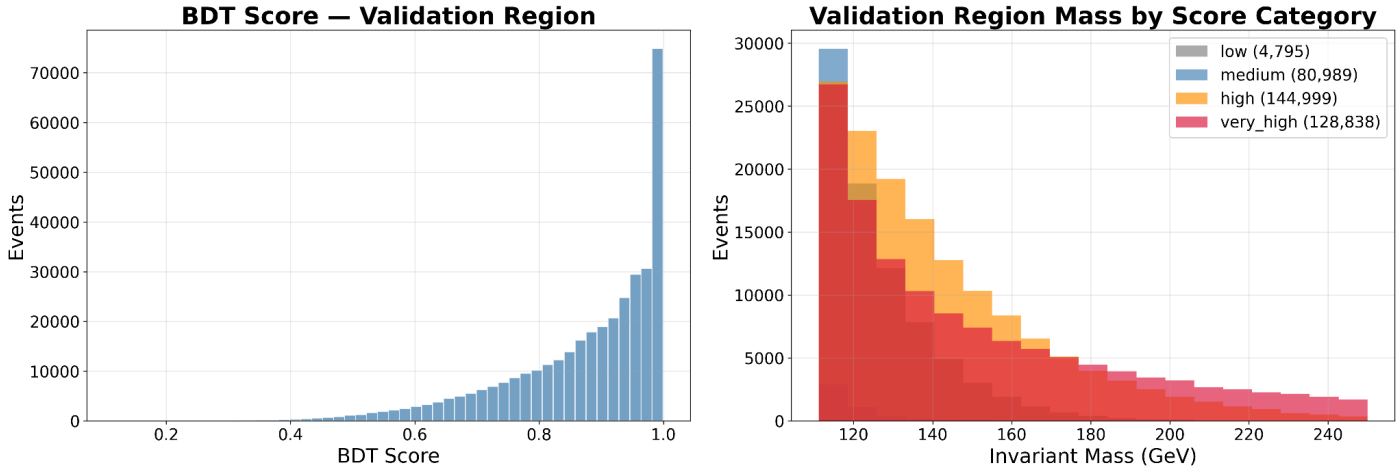


Figure 5: BDT classifier scores in the validation region (111.2-250 GeV): Distribution of BDT scores (left) and invariant mass distribution broken down by score category (right), showing a smooth transition into the search region.

5. Search and Results

5.1. Score Categories

After training and validating the classifiers, we apply them to the blinded search region starting at mℓℓ = 250 GeV. Each event is scored between 0 and 1, indicating how Z-like the classifier judges it to be. Scored events are then sorted into four categories:

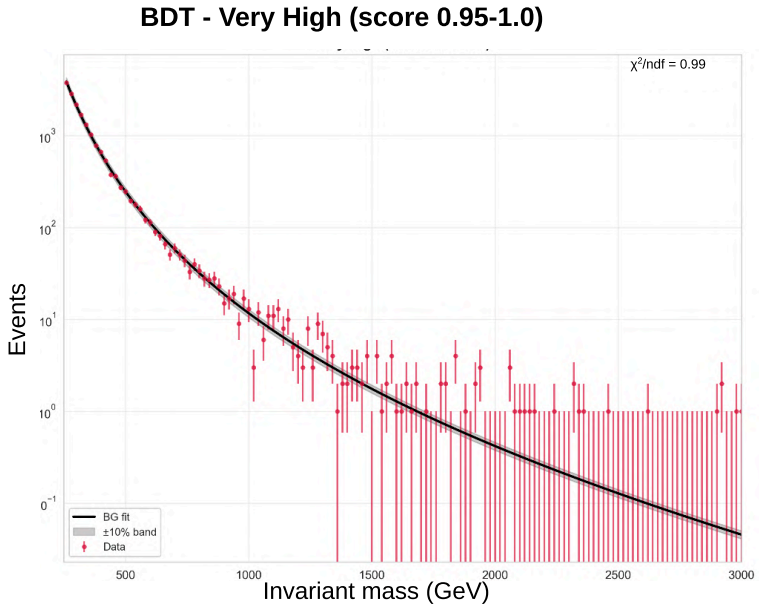
- Low: score < 0.5

- Medium: 0.5 ≤ score < 0.8
- High: 0.8 ≤ score < 0.95
- Very High: score ≥ 0.95

The "very high" category includes events that the classifier considers most similar to a Z boson decay. If a Z' exists in this data and decays into lepton pairs like the Z, its events would concentrate in this category and stand out against the reduced background. This category is the primary channel for our significance scan.

5.2. Background Modeling

As explained in Section 4, we perform a three-parameter fit on the events in the "very high" category and evaluated the accuracy of the fit with a χ^2/ndf value. A χ^2/ndf close to 1 indicates the fit's acceptable agreement within statistical fluctuations. A value much greater than 1 would indicate that the fit does not accurately describe the data, probably because the classifier has distorted the mass spectrum and has started background sculpting. Our fits yield χ^2/ndf = 0.99 for the BDT and χ^2/ndf = 1.11 for the DNN in the very-high category, which confirms that neither classifier sculpts the background or creates false signatures. The likely reason that the DNN's value is further from 1 is because it flags much more events as high-scoring, which makes it harder for it to create a smooth fit with only three parameters.



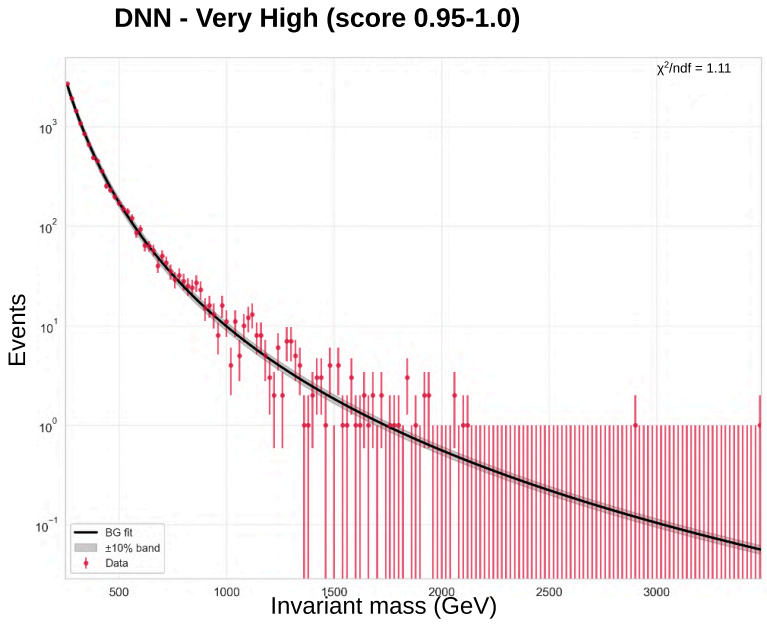


Figure 6: Three-parameter background fits in the very-high score category (score ≥ 0.95) for the BDT (top) and DNN (bottom), shown overlaid on the observed mass spectrum in 20 GeV bins. The fits yield $\chi^2/\mathrm{ndf} = 0.99$ and 1.11 respectively.

5.3. Significance Scan

A χ^2/ndf is a global evaluation of our background fit: it evaluates the success of the whole fit. To evaluate statistical significance of an observed excess in individual bins, we use an Asimov significance scan (Cowan et al., 2011):

$$Z_A = \sqrt{\{ 2 [n \ln(n/b) - (n - b)] \}}$$

Where n is the number of observed events in a bin and b is the number of events predicted by the background fit. For both classifiers, we compute the Asimov significance per 20 GeV mass bins across the search region.

Across the search region, no bin in either classifier exceeded the conventionally required 5σ threshold for a discovery or even the 3σ threshold for evidence of a new particle. The largest significance we observed was at $m\ell\ell = 1120$ GeV, which both classifiers marked as the most significant bin, with the BDT yielding 2.04σ and the DNN yielding 2.40σ.

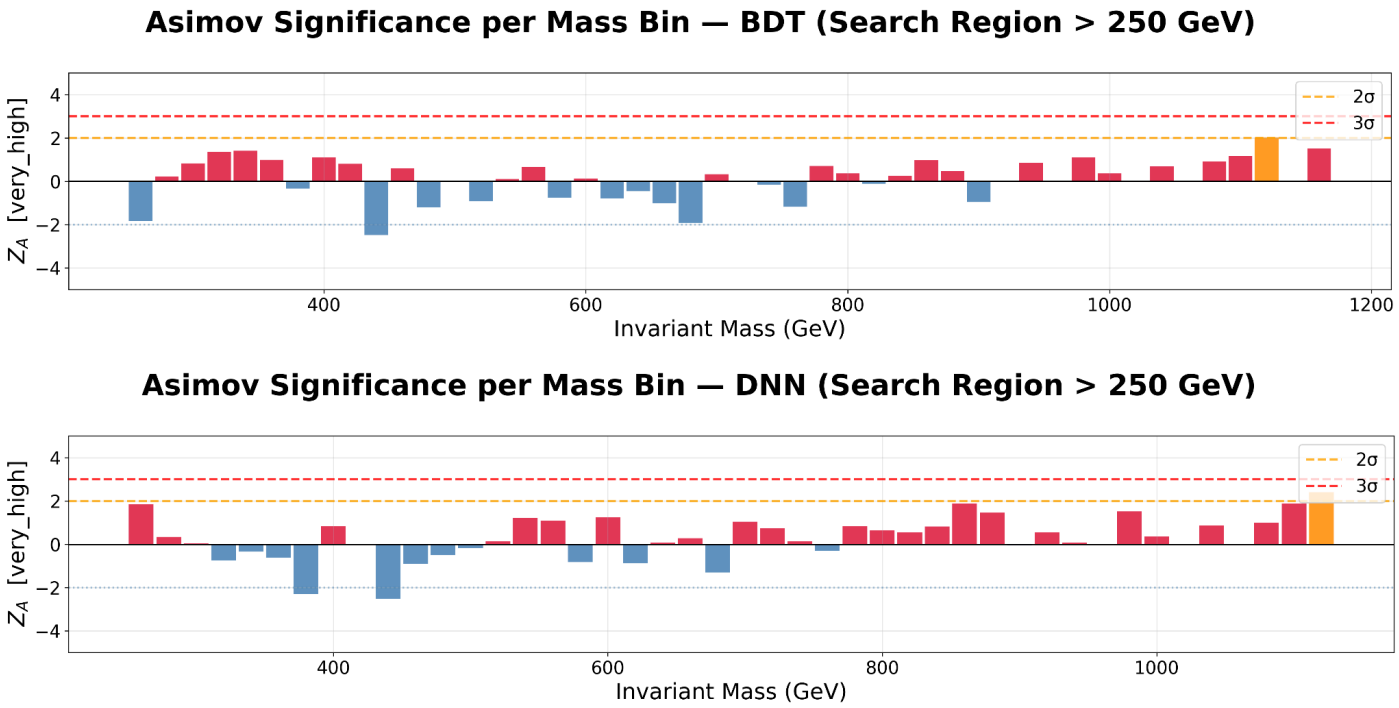


Figure 7: Asimov significance per 20 GeV bin in the very-high score category for BDT (top) and DNN (bottom). Dashed lines indicate the 2σ and 3σ thresholds. The largest deviation was 2.40σ at mℓℓ = 1120 GeV in the DNN.

5.4. Look-Elsewhere Effect and p-value Calculations

This particular excess may be attributed to the look-elsewhere effect in particle physics (Gross & Vitells, 2010), which is a tendency to find a statistically significant event simply because of a large search range. Because we scan 60 independent mass bins (20 GeV bins from 250 to 1450 GeV, which was the largest mass bin with sufficient statistics), a fluctuation of 2σ somewhere along the mass spectrum is not very surprising. We can calculate the one-sided p-values to understand the probability of finding fluctuations this large purely by chance using this formula:

$$P_global = 1 - \Phi(z)\hat{}N$$

Where P_global is the probability of finding a particular excess of significance z across N number of bins and Φ(z) is the standard normal cumulative distribution function (CDF), defined as the integral of the Gaussian (Cowan et al., 2011). CDF values are common in statistics textbooks and we did not calculate them by hand. Performing these calculations, we find a 74.94% chance that a 2σ excess appears and a 38.98% chance that a 2.40σ excess appears at least once across all 60 bins because of statistical noise. Thus, the global significance, or the significance after considering how many bins were searched, is therefore realistically much lower than 2.40σ and even 2σ. The deviation we observed is attributed to statistical noise from the large amount of data we used, and we can safely report a null result with no evidence for a Z' boson in the mℓℓ = 250–3,000 GeV range.

6. Discussion

The main methodological goal of this analysis was to use ML classifiers in a bump-hunt search without distorting the background. We achieved this with three important design choices: excluding the invariant mass from features, avoiding any mass-normalised or mass-derived features, and training on non-overlapping mass sidebands. The $\chi^2/\text{ndf} \approx 1.00$ for both background fits in the very-high-score category provides confirmation that this strategy works well, as a classifier that learned from mass features would produce a distorted spectrum that could not be fit with a smooth three-parameter function.

While traditional approaches to bump-hunt searches involve simple kinematic cuts to select events, this ML-based approach concentrates signal-like events into higher scoring categories while preserving the smooth background shape. This suggests that ML classifiers could be useful for improving statistical sensitivity in future searches that include systematic uncertainties.

6.1. Classifier Comparison

The DNN's test AUC (0.9996) substantially outperforms the BDT's (0.9869), reflecting the neural network's greater ability to model complex input feature combinations. This resulted in the DNN's higher sensitivity in the search region, as it yields a 2.40σ significance at $m\ell\ell$ = 1120 GeV compared to the BDT's 2.04σ. Despite the agreement between classifiers, this significance threshold is not sufficient enough for this to be considered evidence or discovery.

The two classifiers both offer their own strengths. The BDT's feature importance analysis provides insight into the classification, with a lepton pT dominated scoring (expected from a Z decay). It is more complicated to determine the decision process of the DNN because it uses more complex feature combinations. Nevertheless, presenting both classifiers and their independent results provides cross-check validations and strengthens the credibility of the result. The fact that both models agreed on a null result at the same particular excess increases the confidence in this conclusion.

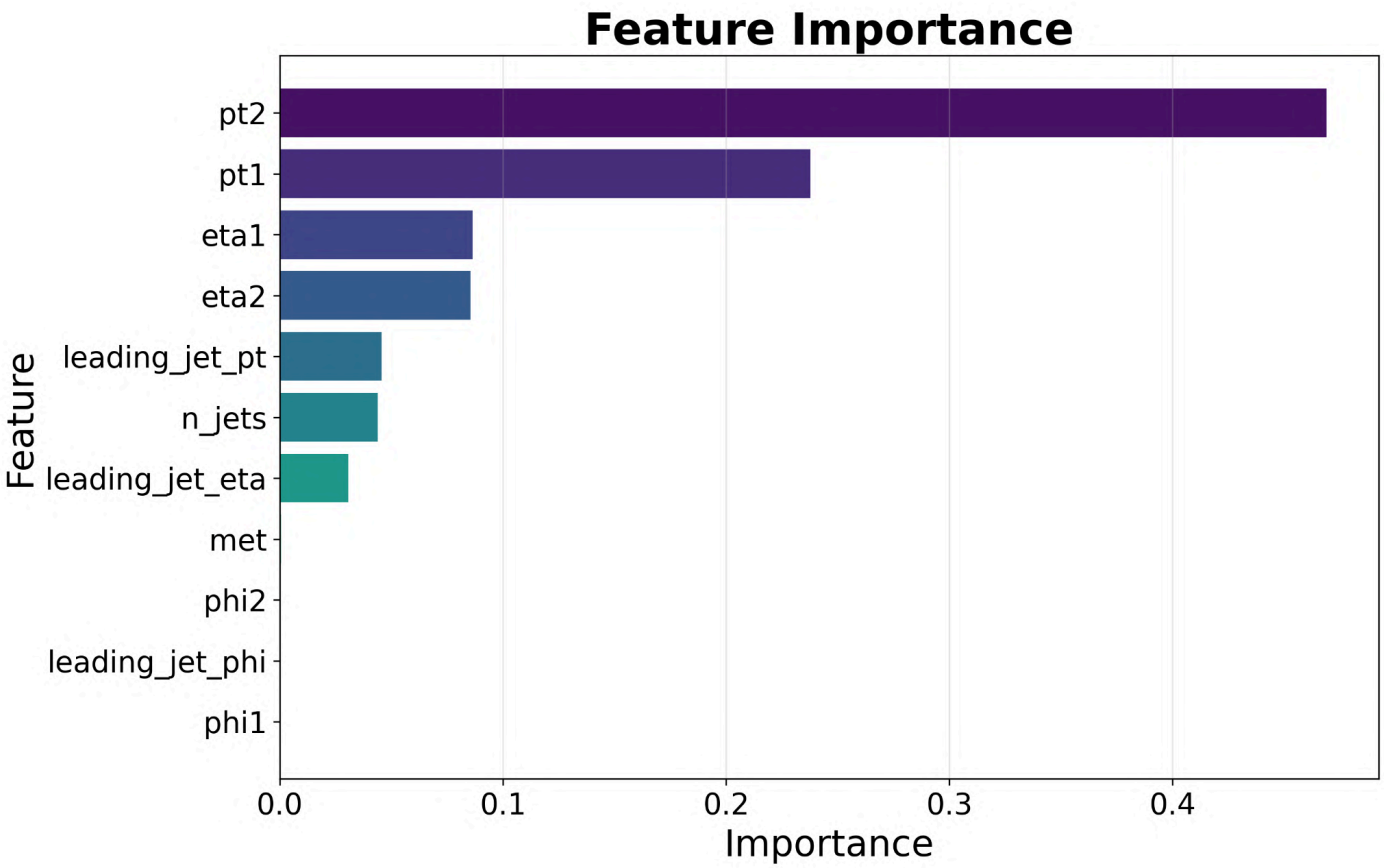


Figure 8: BDT feature importance. Lepton transverse momenta dominate, with pT^2 contributing 47% and pT^1 contributing 24%.

6.2. Limitations

This analysis is designed as a methodological demonstration rather than a full experimental result, and improvements are needed before producing results on par with official CERN and ATLAS publications:

- No comparison to theoretical models: We do not compare our results against existing benchmark theories and simulated signal events. A complete analysis would interpret the search within one or more theoretical frameworks, predict the expected signal yield for each model, and set quantitative exclusion limits on their parameters.
- Signal injection studies: To evaluate how sensitive the search is, researchers would inject simulated Z' signals at various mass points to check how often the analysis would detect them (Belis et al., 2023).
- Systematic uncertainties: We use only statistical uncertainties from the limited number of events in our selected bins. A complete and proper analysis would include systematic uncertainties on factors like luminosity, lepton

identification efficiencies, and choice of background function.

- Alternative background functions: We use a single three-parameter function, but a professional analysis might compare several functional forms and include any variation as an additional uncertainty.
- Mass decorrelation: The DNN's very high AUC (>0.999) warrants some caution about mass correlation in the training data set. While the background fit confirms no sculpting, techniques such as adversarial training or distance-correlation penalties could provide stronger guarantees against mass correlations.

7. Conclusion

We have presented a machine-learning based search for a Z' boson using ATLAS Run 2 2015 open data, comparing a gradient boosted decision tree and a three-layer MLP trained on mass sideband kinematics. By intentionally excluding dilepton invariant mass from the input features and training on mass sidebands outside of the search region, both classifiers achieve strong discrimination between background and signal events (AUC = 0.9869 and 0.9996 respectively) while describing the high-mass background spectrum with a simple three-parameter function.

Our Asimov significance scan across bins in the search region produces no bins exceeding the 3σ evidence threshold. Our largest local deviation is 2.40σ at $m\ell\ell$ = 1120 GeV using the DNN, which can be considered a statistical fluctuation when the look-elsewhere effect is considered. Thus, we report a null result with no evidence for a Z' boson in the mass range 250-3,000 GeV. It is worth emphasizing that a null result is not a failed result. Finding no excess and confirming that the Drell-Yan spectrum remains smooth across this mass range narrows the space of where new physics could be found and adds to our understanding of the universe by confirming that presently there is no evidence of a Z' boson.

The results of our methodology were successful. Using only publicly available data and simple kinematic features, we demonstrate that ML classifiers can be used in bump hunt resonance searches while preserving the background model. The full experimental process is designed to be reproducible, providing a template for future bump hunt analyses using LHC open data.

8. References

ATLAS Collaboration. (2008). The ATLAS experiment at the CERN Large Hadron Collider. Journal of Instrumentation, 3(8), S08003. https://doi.org/10.1088/1748-0221/3/08/S08003

ATLAS Collaboration. (2016). Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 76(5), 292. https://doi.org/10.1140/epjc/s10052-016-4120-y

ATLAS Collaboration. (2019a). Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 79(8), 639. https://doi.org/10.1140/epjc/s10052-019-7140-6

ATLAS Collaboration. (2019b). Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s}$ = 13 TeV with the ATLAS detector. Physics Letters B, 796, 68–87. https://doi.org/10.1016/j.physletb.2019.07.016

ATLAS Collaboration. (2024). ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data [Data set]. CERN Open

Data Portal. https://doi.org/10.7483/OPENDATA.ATLAS.AOQL.8TT3

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. Nature Communications, 5, 4308. https://doi.org/10.1038/ncomms5308

Belis, V., Odagiu, P., & Aarrestad, T. K. (2023). Machine learning for anomaly detection in particle physics. Reviews in Physics, 12, 100091. https://doi.org/10.1016/j.revip.2024.100091

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

CMS Collaboration. (2021). Search for resonant and nonresonant new phenomena in high-mass dilepton final states at $\sqrt{s} = 13$ TeV. Journal of High Energy Physics, 2021(7), 208. https://doi.org/10.1007/JHEP07(2021)208

Collins, J. H., Howe, K., & Nachman, B. (2019). Extending the search for new resonances with machine learning. Physical Review D, 99(1), 014038. https://doi.org/10.1103/PhysRevD.99.014038

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. European Physical Journal C, 71(2), 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0

Dolen, J., Harris, P., Marzani, S., Rappoccio, S., & Tran, N. (2016). Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure. Journal of High Energy Physics, 2016(5), 156. https://doi.org/10.1007/JHEP05(2016)156

Drell, S. D., & Yan, T.-M. (1970). Massive lepton-pair production in hadron-hadron collisions at high energies. Physical Review Letters, 25(5), 316–320. https://doi.org/10.1103/PhysRevLett.25.316

Evans, L., & Bryant, P. (2008). LHC machine. Journal of Instrumentation, 3(8), S08001. https://doi.org/10.1088/1748-0221/3/08/S08001

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feickert, M., & Nachman, B. (2021). A living review of machine learning for particle physics. arXiv. https://arxiv.org/abs/2102.02770

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. European Physical Journal C, 70(1–2), 525–530. https://doi.org/10.1140/epjc/s10052-010-1470-8

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 448–456). PMLR.

https://arxiv.org/abs/1502.03167

Langacker, P. (2009). The physics of heavy Z′ gauge bosons. Reviews of Modern Physics, 81(3), 1199–1228. https://doi.org/10.1103/RevModPhys.81.1199

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1711.05101

Metodiev, E. M., Nachman, B., & Thaler, J. (2017). Classification without labels: Learning from mixed samples in high energy physics. Journal of High Energy Physics, 2017(10), 174. https://doi.org/10.1007/JHEP10(2017)174

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807–814). Omnipress.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Terao, K., & Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. Nature, 560(7716), 41–48. https://doi.org/10.1038/s41586-018-0361-2

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., & McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research Section A, 543(2–3), 577–584. https://doi.org/10.1016/j.nima.2004.12.018

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Data & Code Availability Statement

The data used for this experimental setup was accessed via the ATLAS Open Data portal. Link to dataset: https://opendata.cern.ch/record/80000

Acknowledgements

I would like to thank [redacted] for his continued support and guidance during this project.

I acknowledge the work of the ATLAS Collaboration to record or simulate, reconstruct, and distribute the Open Data used in this paper, and to develop and support the software with which it was analysed.

AI Use Statement

Claude Opus 4.7 (Anthropic) was used for minor editing support and minor coding support during this project. All methodological decisions, analysis, results, and scientific content are the author's own.

Author Biography

[redacted]

Mentor Contribution Statement

[redacted]

1. In-text parenthetical citations were added throughout the manuscript so that every reference is cited at least once.
    a. One reference that should have been included was added: Belis, V., Odagiu, P., & Aarrestad, T. K. (2023). Machine learning for anomaly detection in particle physics. Reviews in Physics, 12, 100091. https://doi.org/10.1016/j.revip.2024.100091
2. Increased font size of axis labels in all figures except Figure 2. Figure 3 was simply made bigger so the text is made more readable.

Convergence
A Journal for Young Researchers

Received: [for official use only] ; Accepted: [for official use only];
Managing Editor: [for official use only];
Action Editor: [for official use only]

Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data

Mihir Tare

St. Mark's School of Texas, Dallas, Texas, USA

Abstract

The Standard Model of particle physics, while highly successful, leaves several unanswered questions that motivate searches for undiscovered heavy gauge bosons predicted by Beyond-Standard-Model theories. Using machine-learning classifiers trained on low-mass sideband events, this paper presents a search for the Z' boson, one such hypothetical heavy neutral gauge boson that decays into dilepton pairs. A gradient boosted decision tree (BDT) and a multi-layered perceptron deep neural network (DNN) are trained to distinguish real Z boson events from Drell-Yan background events using only 11 raw kinematic features, with dilepton invariant mass deliberately excluded to avoid background sculpting of artificial signals during the search. The analysis uses real proton-proton collision data recorded by the ATLAS experiment during LHC Run 2 in 2015, released through the CERN Open Data portal in 2024. After both classifiers achieve accurate discrimination on the training set, events in a blinded high-mass search region are scored by the classifiers, a three-parameter parametric fit is used to model the background, and an Asimov significance scan is performed across mass bins. We observed no statistically significant excess, with the largest local deviation at a mass of 1120 GeV with a significance of 2.83σ in the DNN's highest-score category, which is well below the 3σ conventionally required for evidence of new physics. This analysis demonstrates how machine-learning can be used in high-energy physics "bump hunt" searches while preserving the background model, and it provides a reproducible methodology built on publicly available data.

Keywords:
Z′ boson, dilepton resonance search, machine learning, boosted decision tree, deep neural network, ATLAS Open Data, bump hunt, beyond the standard model

1. Goals and Theory

The Standard Model (SM) of particle physics provides a remarkably successful description of the fundamental constituents of matter and the forces that govern their interactions. Despite its predictive power, the SM is widely understood to be an incomplete theory. It does not account for observed matter-antimatter asymmetry of the universe, provides no dark matter candidate, fails to incorporate gravity at the quantum level, and leaves the mass hierarchy of fermions unexplained.

A variety of Beyond Standard Model (BSM) theories address these shortcomings by hypothesizing the existence of new heavy gauge bosons that come from extensions of the Standard Model's electroweak symmetry, the mathematical structure that organizes how particles interact with the weak nuclear force. These hypothetical particles are called Z' bosons (Z-prime bosons). A Z' boson is generally understood to be a heavier counterpart of the Z boson, a known particle that mediates the weak nuclear force and was discovered at CERN in 1983. Many BSM extensions predict that the Z' would couple to leptons and quarks, though the coupling strength varies between models. In the Sequential Standard Model benchmark, the couplings are thought to be identical to the Z, while other models predict substantially different couplings (Langacker, 2009).

Previous searches at the LHC have set lower mass limits on Z' bosons in the multi-TeV range for the most common benchmark models. Both ATLAS and CMS have performed high-mass dilepton searches using the full Run 2 dataset of 139 fb-1 and found no evidence of a Z' (ATLAS Collaboration, 2019; CMS Collaboration, 2021). An undiscovered resonance remains possible, particularly at masses or coupling strengths below the sensitivity of these previous analyses.

Our search utilizes the dilepton channel because of its simplicity and clarity. As a hypothetical "cousin" of the Z boson, the Z' should decay into a pair formed by a lepton and its corresponding anti-lepton (e.g. an electron and positron). Utilizing this channel, often called the "golden channel," allows for a clean experimental signature and a well-understood background from the Drell-Yan process (Drell & Yan, 1970). If a Z' exists, it would appear as a localized excess of events, or a "bump", in the dilepton invariant mass spectrum at the Z' mass, sitting on top of the smooth Drell-Yan continuum. This "bump-hunt" strategy is the standard approach to discovering new resonances at colliders.

Machine learning is now a standard tool in particle physics for classifying collision events (Feickert & Nachman, 2021; Radovic et al., 2018). However, applying machine learning to bump-hunt searches comes with a concern: if a classifier learns to reconstruct invariant mass (even indirectly through correlated features), it can distort the smooth background spectrum and create artificial peaks where no real signal exists. This problem, known as background sculpting (Dolen et al., 2016), is the primary failure this analysis is designed to avoid.

The goal of this work is to develop and validate a fully reproducible machine learning workflow for a search in the dilepton channel using ATLAS Open Data, following the methodological structure of real collider analyses. We apply this workflow to a Z' boson search as a concrete test case and find no statistically significant localized excess across the probed mass range.

2. Experimental Background

2.1. The LHC

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator (Evans & Bryant, 2008). It accelerates two beams of protons in opposite directions to nearly the speed of light, then brings them into head-on collision at designated interaction points. During Run 2 (2015-2018), the LHC operated at a center-of-mass energy of $\sqrt{s}$ = 13 TeV, meaning the total energy available in each proton-proton collision is 13 trillion electron-volts. This analysis uses data collected in 2015 during the first year of Run 2.

2.2. ATLAS Detector

The ATLAS (A Toroidal LHC Apparatus) detector is a large general-purpose particle detector at the LHC (ATLAS Collaboration, 2008). It is designed to record the debris from high-energy proton-proton collisions and reconstruct what happened in each event using software and sensors. The detector is built in concentric layers around the collision point, with each layer optimized to measure different types of particles:

- Inner Detector: The innermost layer immersed in a strong magnetic field. Charged particles curve as they pass through the field, and the curvature is used to measure their momentum. This is the primary system for tracking leptons such as electrons and muons.
- Calorimeters: These layers, surrounding the inner detector, absorb particles and measure their energy. The electromagnetic calorimeter measures electrons and photons, while the hadronic calorimeter measures jets (sprays of particles produced by quarks and gluons).
- Muon spectrometer: Muons pass through calorimeters relatively undetected, depositing only a small amount of energy through ionization. They are too heavy to radiate the electromagnetic showers that electrons do, and because they are leptons that don't participate in the strong force, muons also do not produce hadronic showers like protons and neutrons. So, this dedicated system in the outermost layer identifies and measures them with high precision using a toroidal magnetic field.

2.3. Dataset

This analysis uses Run 2 2015 proton-proton collision data made publicly available through the CERN Open Data Portal in 2024 (ATLAS Collaboration, 2024). The data is provided in DAOD-PHYSLITE format, an analysis-ready format in which the ATLAS experiment collaboration has already performed computationally heavy work like identifying particles. This full dataset contains 10,049 files with a total 9.3 TiB of approximately 1.69 x 10^9 events and has a luminosity of 3.2 fb^{-1}.

3. Data Preparation

3.1. Event Selection

For event selection, we do not apply any additional trigger requirements, as the PHYSLITE Open Data is already trigger-filtered. We select events containing exactly two same-flavour leptons: dielectron events (with exactly two electrons) and dimuon events (with exactly two muons). While lepton flavor is recorded, it is not used as an input for our classifiers. We do not impose an opposite-sign charge requirement, and this limitation is discussed in Section 6.2. Limitations and Future Improvements. For each dilepton event, the following reconstructed objects are extracted:

- Lepton Kinematics: Using analysis-level reconstructed lepton objects without further ID/isolation cuts (AnalysisElectronsAuxDyn and AnalysisMuonsAuxDyn), we collect the transverse momentum pT (the component of momentum perpendicular to the beam), pseudorapidity η, and azimuthal angle φ of each lepton.
- Jet Kinematics: The three same kinematic variables are collected for the leading reconstructed jet (AnalysisJetsAuxDyn), plus the jet multiplicity n_jets (how many jets were found in the event).
- Missing Transverse Energy (MET): The apparent imbalance in the total transverse momentum of the event, the MET can signal potential undetected particles like neutrinos. It is reconstructed from the x and y components of the missing momentum vector (MET_Core_AnalysisMETAuxDyn.mpx/mpy).

All momenta are provided in MeV by ATLAS and are converted to GeV for this analysis. We do not apply any pT-based cuts while selecting events and processing data, although quality cuts are applied prior to training the model, as discussed in Section 3.3.

3.2. Invariant Mass

The dilepton invariant mass $m_{\ell\ell}$ is an important value for this search. If two leptons originate from a parent particle decay (like a Z' boson decaying into two leptons), then the dilepton invariant mass should equal the mass of the parent particle. Thus, if a Z' boson exists, it would produce a peak of events at a particular value in the invariant mass spectrum corresponding to its mass. This is what physics "bump hunts" typically look for. Dilepton Invariant mass is calculated using the three kinematic variables that we collected for each lepton (pT, η, φ):

$$m_{ll} = \sqrt{2pT^1pT^2[\cos cos\left(\eta^1 - \eta^2\right) - \cos cos\left(\phi^1 - \phi^2\right)]}$$

3.3. Quality Cuts

We apply quality cuts that are standard to the LHC. A pseudorapidity cut $|\eta| \leq 2.5$ is applied to electrons and muons, reflecting the acceptance of the ATLAS inner detector and calorimeter. For muons, this cut reflects limitations of the ATLAS Inner Detector and its Semiconductor Tracker (SCT), which work with the ATLAS Muon Spectrometer to measure muon momentum (ATLAS Collaboration, 2016). For electrons, the ATLAS electromagnetic calorimeter precision region is $|\eta| \leq 2.47$, which makes this analogous cut slightly conservative but still valid. However, standard ATLAS analyses also exclude the barrel-endcap transition region ($1.37 < |\eta| < 1.52$), where dead material degrades electron energy resolution. We do not apply this exclusion in this simplified Open Data analysis, which may slightly broaden the reconstructed electron resolution (ATLAS Collaboration, 2019a). Jet variables are undefined for events where n_jets = 0, so the pseudorapidity cut is conditionally applied to jet objects when n_jets > 0.

A hard cut of $m_{\ell\ell} \leq 3{,}500$ GeV is applied because the background spectrum falls very steeply, and so past this boundary, there are hardly any events in the 20 GeV bins that our fit uses. We account for this problem with a minimum of 10 events for a bin to be used during the significance scan which restricts the scan to $m_{\ell\ell}$ = 250-1450 GeV.

4. Machine Learning Methodology

The machine learning methodology has three stages. First, we train two classifiers, a boosted decision tree (BDT) and a

deep neural network (DNN), to distinguish Z-peak events from low-mass Drell-Yan background events using only kinematic features, with the dilepton invariant mass deliberately excluded so that the classifiers cannot learn the search variable. Both are trained and validated on events drawn exclusively from the low-mass sidebands and the Z-peak region, well below the search region, so that the search region itself remains blinded during model development. These Z-peak events are not meant to stand in for a Z' signal, but are used as a Z-like proxy-signal and real resonances that a Z' would be kinematically similar to.

Second, we apply the trained classifiers to the high-mass search region ($m_{\ell\ell}$ > 250 GeV) and sort every event into categories by classifier score, concentrating the most Z-like events into a single high-scoring category. Within that category, we model the smooth Drell-Yan background with a three-parameter parametric fit:

$$f(m) = p_0\left(1 - \frac{m}{\sqrt{s}}\right)^{p_1} m^{(-p_2)}$$

This function combines a falling power law in m with a factor that suppresses the spectrum as m approaches the kinematic limit $\sqrt{s}$ (ATLAS Collaboration, 2019b). The three parameters (p_0, p_1, p_2) are determined by fitting the function to the observed mass spectrum in 20 GeV bins using χ^2 minimization, which is a statistical procedure that adjusts the parameters to minimize the sum of squared residuals between the fit and the data, weighted by the uncertainty in each bin.

We evaluate the accuracy of the fit with reduced chi-squared statistic, χ^2/ndf, which is a sum over all bins of squared deviations. χ^2 is evaluated as $\chi^2 = \Sigma_i (n_i - b_i)^2 / \sigma_i^2$, where n_i is the number of observed events in a bin, b_i is the number of events predicted to be in the bin by the background fit, and σ_i is a variance calculation. Ndf (number of degrees of freedom) is calculated by subtracting the number of parameters in the fit (three in our case) from the number of bins we search. This background spectrum should remain smooth and falling, and a good fit ~~will confirm~~ provides evidence against ~~that~~ the classifiers ~~have not~~ sculpting ~~sculpted~~ artificial structure into it.

Finally, we perform a bin-by-bin Asimov significance scan within the high-score category, comparing the observed number of events in each bin against the background fit to test for localized excesses. The remainder of this section describes the input features, the two classifier architectures, and their training process.

4.1. Features

The core idea of this methodology is to train machine-learning classifiers to distinguish Z-like events from non-Z-like events using only kinematic features. Specifically, the momenta, angles, jet activity, and missing transverse energy of each event. While we record lepton flavor during event selection, it is not used as an input for the classifiers, so electron events and muon events are treated the same despite their different kinematic behaviors. This limitation is acknowledged in Section 6.2. The trained classifiers are then applied to the search region to identify which high-mass events look most like a Z boson and then concentrate these potential Z' signals into a high-score category.

One critical design choice for the input features is that dilepton invariant mass is excluded, as including it would allow the

classifiers to directly learn the search variable and potentially sculpt the background to create artificial "bumps" in the mass spectrum where no real signal exists. Of course, kinematic inputs such as pT, η, and φ are used to calculate invariant mass, so classifiers may still indirectly learn mass-correlated information. However, excluding the invariant mass is a step taken to mitigate this mass correlation.

The same 11 input features are used for both classifiers. They are listed with their descriptions in Table 1.

Table 1: Input kinematic features for the models and their descriptions

Feature	**Meaning**
pT^1, pT^2	Transverse momenta of the leading and subleading leptons (electron or muon)
η^1, η^2	Pseudorapidities of both leptons
φ^1, φ^2	Azimuthal angles of both leptons
n_jets	Reconstructed jet multiplicity
leading_jet_pT	Transverse momentum of the leading jet (zero if n_jets = 0)
leading_jet_η	Pseudorapidity of the leading jet (zero if n_jets = 0)
leading_jet_φ	Azimuthal Angle of the leading jet (zero if n_jets = 0)
MET	Missing transverse energy

Both classifiers are trained as binary classifiers, with Z-peak events labeled as proxy-signal (class 1) and low-mass Drell-Yan events labeled as background (class 0) (Collins et al., 2019; Metodiev et al., 2017). The labeled sample of training events is split 80/20 into training and test sets using stratified sampling (preserving class ratio in both splits), giving 8,950,298 training events and 2,237,575 held-out test events for validation.

4.2. Boosted Decision Tree (BDT)

The BDT, an ensemble of gradient-boosted decision trees that has become a standard classifier in particle physics (Friedman, 2001; Roe et al., 2005), uses binary cross-entropy (BCE), a loss function for binary classification where loss increases exponentially when the difference between the model's prediction and the event's actual label increases. For an event with true label y (either 0 or 1) and predicted probability p, the BCE is:

$$L = -[y * \log log(p) + (1 - y) * \log log(1 - p)]$$

Loss is small when the model's prediction matches the correct label of the event and large when the model is incorrect with high confidence. Each successive tree is fit to the gradient of this loss so that the ensemble progressively reduces its overall error. This loss calculation is done automatically using XGBoost, a scalable open-source gradient-boosting library

implemented in Python through its xgboost package, which handles tree construction, gradient computation, and parallelized training internally (Chen & Guestrin, 2016). All preprocessing of input features, train/test splitting, and evaluation are carried out in Python using NumPy, pandas, and scikit-learn.

To prevent the model from overfitting, we apply several forms of regularization using XGBoost parameters.

- Number of trees and learning rate: We train 300 trees with a low learning rate of 0.03. The learning rate scaled down each tree's contribution, so that each tree makes a small correction rather than aggressively fitting the training data. A low rate combined with many trees sacrifices training speed for better generalization.
- Maximum depth of 5: Each tree is limited to 5 levels deep, which produces at most 32 distinct regions. Unlike deep trees, which tend to memorize individual events, shallow trees learn broad patterns.
- Minimum child weight of 10: In each tree, a leaf node can only be split if both children would contain at least 10 units of statistical weight. This prevents the model from learning patterns from small groups of unusual events.
- Minimum split gain of 1.0: Splits only happen if they reduce the loss by at least 1.0. This "prunes" each tree by removing redundant branches, which helps reduce overfitting.
- L1 regularization (reg_alpha = 0.1): This essentially adds a penalty proportional to the absolute value of each leaf's output weight. It pushes small leaf weights toward zero, which can prune branches from the trees and simplify the model. Unlike Lasso regularization in linear models, this does not zero out an input feature's global importance. It regularizes the leaf values rather than performing input-feature selection.
- L2 regularization (reg_lambda = 5.0): Adds a penalty proportional to the square of a leaf's prediction. It shrinks large weights toward zero and strongly discourages overconfident predictions.
- Stochastic gradient boosting (70%): Each tree is trained on a random 70% of events using a random 70% of features. This randomness prevents all trees from making the same mistakes.

The class imbalance between labeled signal events (5,811,472) and labeled background events (5,376,401) is handled by a scale_pos_weight parameter of 0.93 (calculated ratio), which decreases the weight of signal class during training.

4.3. Deep Neural Network

Neural networks (LeCun et al., 2015), and in particular deep architectures applied to collider classification problems (Baldi et al., 2014), are built from computational units called neurons. Each neuron takes in a set of numbers and computes a weighted sum based on weights that are learned during training. The result is passed through a nonlinear activation function. Our model uses the ReLU (Rectified Linear Unit) activation function, which outputs the input if it is positive and zero otherwise. ReLU and the DNN architecture is implemented with PyTorch, an open-source deep-learning framework in Python that provides automatic differentiation, GPU-accelerated tensor operations, and prebuilt neural-network modules. The network is defined as a torch.nn.Module subclass, with linear layers, batch normalization, dropout, and ReLU activations composed using the torch.nn API (Nair & Hinton, 2010; Paszke et al., 2019).

Neurons are arranged in multiple layers. In a fully connected layer, every neuron receives input from every neuron in the previous layer. Unlike graph-based architectures designed to find structural relations in inputs, this multi-layered perceptron (MLP) does not explicitly encode invariant relationships like mass between particles. Despite this, correlated kinematic information may still allow partial reconstruction of invariant mass, motivating our validation against background

sculpting (which is discussed in the background modelling section). We initially tried to use a Graph Neural Network (GNN), but it was able to reconstruct invariant mass and scored events almost solely based on mass. The switch to this multi-layered perceptron mitigates this background sculpting problem (later confirmed in section 5.2), though it does not fully rule out indirect mass learning.

Our network consists of an input standardization layers, three fully connected hidden layers of 128 neurons each, and a single output neuron.

- Standardization layer: Before entering the network, each of the 11 input features is transformed to have zero mean and unit standard deviation by subtracting the mean and dividing by the standard deviation calculated only for training sets. Since the raw features have different scales (pT can be hundreds of GeV, while ϕ is bounded between $-\pi$ and π), this ensures that features with larger magnitudes do not distort the learning process with large weights.
- Hidden layers: three layers of 128 neurons each. Each layer applies a linear transformation, batch normalization (which recenters and rescales the activations across each training batch to stabilize learning) (Ioffe & Szegedy, 2015), and ReLU activation. Each step is implemented using PyTorch. The depth and width of the network give it the capacity to learn complex nonlinear combinations of input features.
- Dropout (p=0.3): After each hidden layer, 30% of neurons are randomly "switched off" during each training step. This prevents individual neurons from becoming dependent on one another and improves their individual reliability. Dropout prevents the model as a whole from overfitting and memorizing training noise (Srivastava et al., 2014). When testing, however, all neurons are active.
- Output: In the last layer, a single neuron produces an unbounded score, which is passed through a sigmoid function to predict a probability between 0 (background) and 1 (signal).

Training occurs in epochs, where one epoch is a full pass through the training data. For each epoch, the data is divided into batches of 4,096 events. For each batch, the network makes predictions, the loss is computed, and the gradient of the loss with respect to every weight is computed using chain rule (backpropagation). The weights are then updated according to this gradient. We do clip gradients at 1.0 so that large updates to weights do not cause instability. Over many epochs, the DNN converges toward a set of weights that minimizes the loss.

We optimize our DNN with AdamW (Loshchilov & Hutter, 2019), a variant of the Adam optimizer that decouples weight decay from the gradient update. We use a learning rate of 10^{-3} and weight decay of 10^{-4} (an L2 like penalty on all weights that discourages overfitting). AdamW is provided directly by the torch.optim module, and gradients are computed via PyTorch's automatic differentiation engine during backpropagation. We also use two additional techniques which are also implemented using the PyTorch library:

- Cosine annealing with warm restarts (Loshchilov & Hutter, 2017): The learning rate decreases smoothly following a cosine curve over the course of 10 epochs, then resets back to its initial value and begins decaying again. These periodic resets allow the optimizer to escape shallow local minima that it might otherwise get trapped in, while the cosine decay within each cycle provides a fine-grained optimization.
- Early stopping (patience = 10): After each epoch, the model is evaluated on the held-out test set. If the test-set performance, measured by AUC, does not improve for 10 consecutive epochs, training stops and the best-performing model is used. This prevents the network from training long enough to memorize training data.

Our final model stopped after 29 epochs.

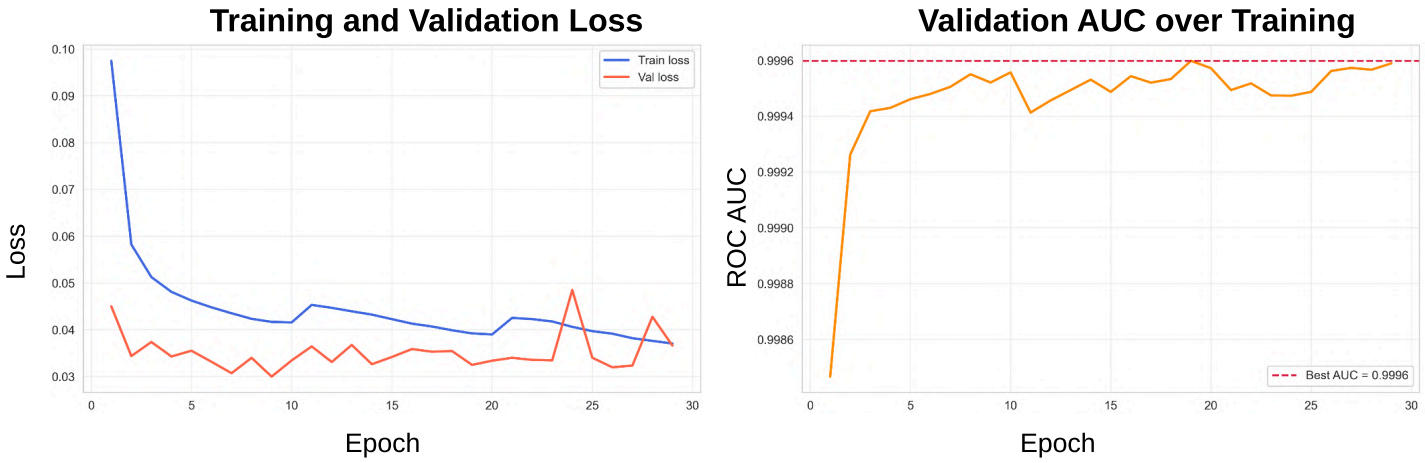


Figure 1: DNN training and validation loss (left) and validation AUC (right) per epoch. Training stopped after 29 epochs because of the early stopping implementation.

4.4. Mass Regions

During training and analysis, we divide the dilepton invariant mass spectrum into four non-overlapping regions.

The first region used for training, the background region, is set between 40 and 71.2 GeV. The lower bound was chosen to get a clean Drell-Yan dominated region without resonance peaks from the J/ψ at 3.1 GeV and the ϒ (upsilon) at 9.5 GeV. The upper bound was chosen to avoid the Z peak at 91.2 GeV. While the Z boson has a natural width of 2.5 GeV, the reconstructed peak is broader because of detector smearing by the calorimeters and muon spectrometer, so we use a larger buffer that accounts for both the natural width and detector resolution effects and avoids Z contamination. These bounds also narrow down our background events so that the ratio of background to signal is much closer to 1. These low-mass Drell-Yan sideband events serve as a background proxy rather than a signal proxy, and during training and analysis, we expect our models to associate these events with label=0 (not alike to a Z resonance).

The signal (or proxy signal) region, also used for training, is centered at the Z boson mass of 91.2 GeV and extends to ±20 GeV threshold (71.2-111.2 GeV). As previously stated, this width is wide enough so that we do not rely on nearly perfect reconstructions and narrow enough so that events from the Drell-Yan tail do not contaminate our signal classification. The lower bound matches the background region upper bound so that there is no overlap between regions but also so that all events with mass greater than 40 GeV are used. Since we do not use simulated Z' events, this region provides our only Z-like resonance for our search. To be clear, these Z-peak events are not real Z' signal events, and are instead used as a Z-like proxy signal, or a real resonance that stands in for a hypothetical Z'. This proxy carries an implicit assumption that a multi-TeV Z' is kinematically similar to 91 GeV Z boson. This assumption is imperfect, since the available phase space, the

parton distribution functions governing the initial-state quark momenta, and the resulting angular distributions of the decay leptons differ between the Z pole and the multi-TeV scale, so a classifier optimized on Z pole kinematics may not be optimal for a high-mass resonance. Despite this reduced signal efficiency, avoiding simulation removes dependence on any particular Z' model.

The validation region from 111.2 GeV to 250 GeV is not used for training or search. Its purpose is to verify that the model generalizes reasonably before looking at the actual search region. The lower bound of 111.2 GeV is just above the signal region. Above this mass, the spectrum is dominated by the high-mass Drell-Yan continuum rather than the on-shell Z peak. Detector resolution and radiative effects smear a small number of Z events into the lower edge of this window, but they are a minor contribution across the region as a whole. Continuum events are kinematically similar to Z decays, since both produce two high-pT leptons from a neutral parent, but they populate higher invariant masses. ~~At this mass, events are still from the high-mass tail of the Z, so they look kinematically alike to Z decays except for their higher invariant mass caused by detector effects and final state radiation.~~ Since the model has not seen these events during training, how it scores them tell us how the learned features generalize. The upper bound of 250 GeV was determined empirically. The goal of this upper bound is to ensure that the contribution from residual Z ~~tail~~ events ~~sufficiently dies out~~ has fallen away so that a smoother power law can fit the remaining mass spectrum. Using a χ^2/ndf fit, we found that below a bound of 250 GeV, residual Z contributions create an excess that a power law cannot fit (χ^2/ndf = 3.20 at 200 GeV, 0.99 at 250 GeV). This bound is consistent with other ATLAS analyses which use 225-250 GeV as a lower bound for a high-mass search region. So, this validation region acts as a transition zone between the signal region and a smooth background. Events in this region are separated from search events specifically because they are not clean enough for a reliable background fit.

The search region, everything above 250 GeV, is where we look for a Z'. In this region, the Drell-Yan background should be smooth and falling, so that a Z' would appear as a localized narrow bump of observed events against a background fit. Although a cut is applied to events with $m_{\ell\ell}$ > 3,500 GeV, our significance scan and statistical interpretation are restricted to 250-1450 GeV, the range over which each 20 GeV bin retains 10 events, which is what we consider sufficient for a reliable fit.

After training on the labeled background and signal regions, the performance of the models is evaluated with the validation region before being applied to score events in the blinded search region.

After event selection and quality cuts, there are 5,376,401 events in the background region; 5,811,472 events in the signal region; 359,621 events in the validation region; and 19,042 events in the search region.

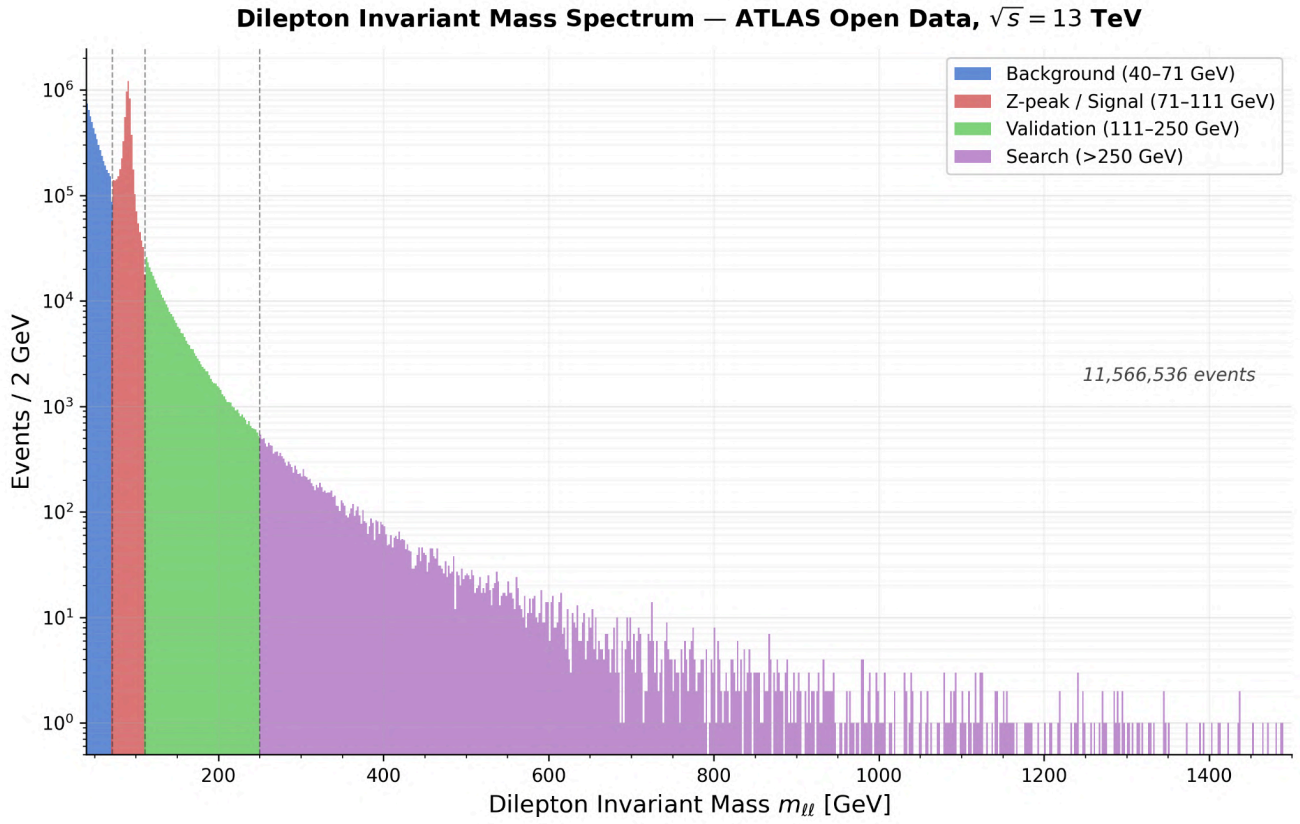


Figure 2: Dilepton invariant mass spectrum across a partial mass range, shown on a logarithmic scale. Events are color-coded by analysis region. The Z peak is at 91.2 GeV with the spectrum falling steeply at higher masses.

4.5. ROC Curve

The first metric we use to evaluate our classifiers after training is the Receiver Operating Characteristic Area Under the Curve (ROC AUC) (Fawcett, 2006). The ROC curve plots the true positive rate (the fraction of signal events correctly identified by the classifier) against the false positive rate (the fraction of background events incorrectly labeled as signal by the classifier) as the classification threshold (the minimum score needed for an event to be classified as a signal) varies. The AUC is the area under the ROC curve and it ranges from 0.5 (essentially random guessing) to 1.0 (perfect accuracy). An AUC of 0.98, for example, means that a random signal event scores higher than a random background event 98% of the time.

The BDT achieved a test-set AUC of 0.9869, and the DNN achieved a 0.9996. These values are based on the classifiers' performance on a 20% hold-out set of the total training data, which included everything in the background and signal region. Both these values are high, which reflects the genuine kinematic differences between Z events and low-mass Drell-Yan sidebands and demonstrates that the classifiers very accurately separate these two groups. This is important, as our methodology relies on using the classifiers to find high-mass events reflecting Z-peak patterns that look very different

from the background spectrum. The reason for this large difference in characteristic is because the Z boson produces leptons with higher transverse momenta than sideband events, and this is a dominant feature driving the classification. This is confirmed by the BDT's feature importance analysis: the subleading lepton's pT accounts for 47% of the model's discriminating power, the leading lepton's pT accounts for 24%, and the other features contribute much less.

It is important to note that these AUC values reflect the classifiers' ability to distinguish the two training classes (Z-peak vs. low-mass Drell-Yan) and do not directly measure sensitivity to a Z' signal in the search region. That sensitivity is established separately through a background-fit quality and the significance scan described later.

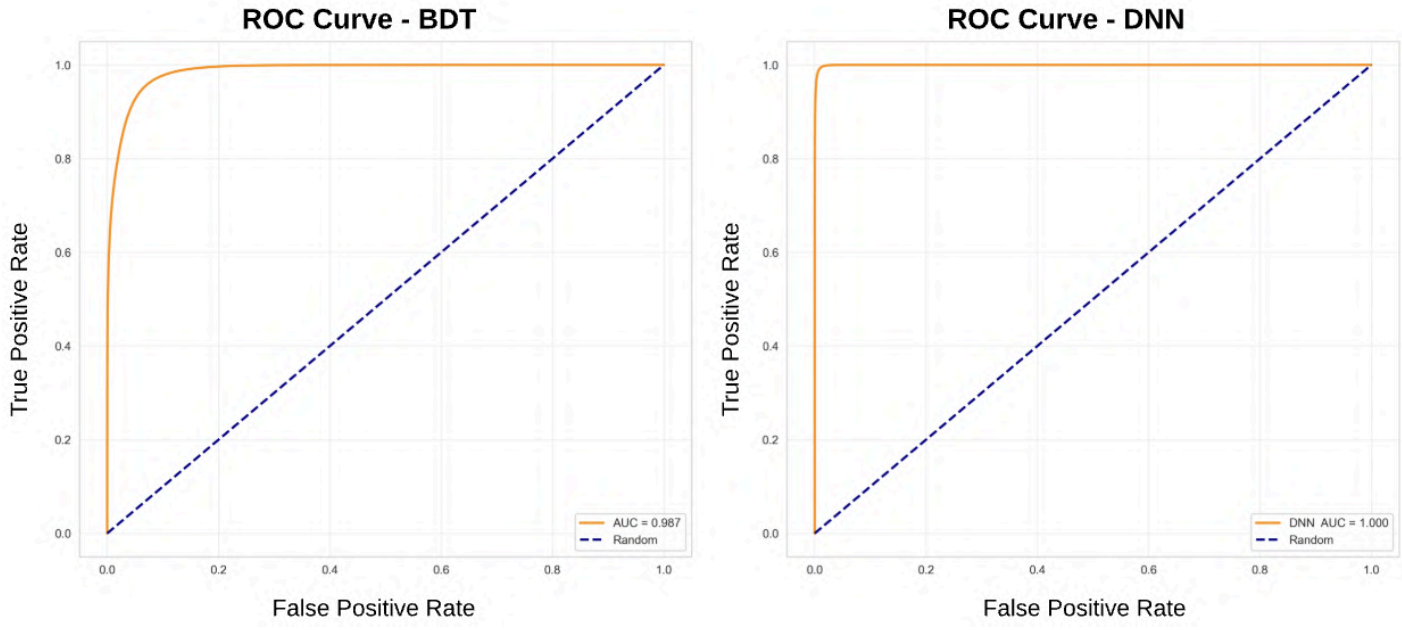


Figure 3: ROC curve on the held-out test set for the BDT (left) and DNN (right) classifiers.

4.6. Overtraining Check

We also used performance checks for overtraining, which occurs when a model memorizes training data rather than learning general patterns, leading to good performance on training data but bad performance on unseen data. First, we verified that the model's performance on the test data mirrors its performance on training data. Then, we confirmed that score distributions evaluated on the test sets were consistent with training sets, meaning the model scores events similarly regardless of whether it has or has not seen them before.

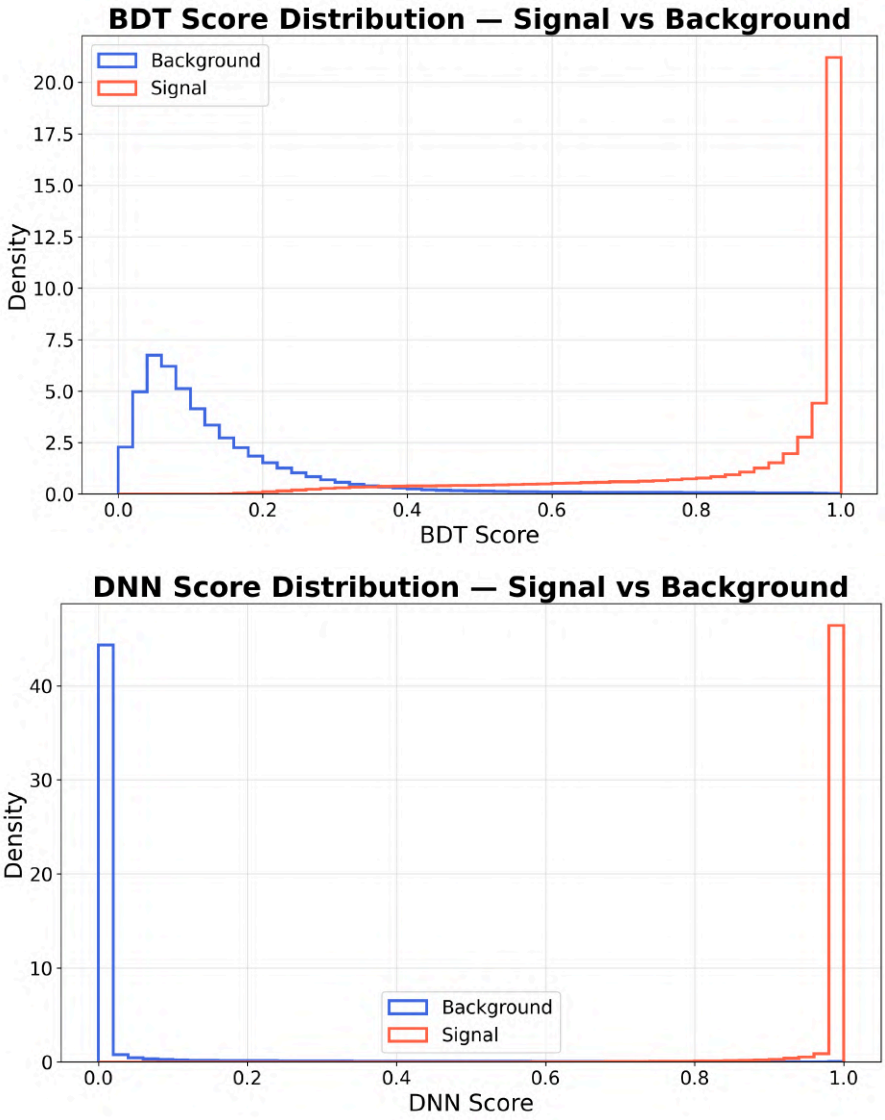


Figure 4: Overtraining check: classifier score distributions on the training and test sets for both the BDT (top) and DNN (bottom). Train and test distributions agree, indicating no overtraining.

4.7. Validation

Before unblinding the search region, we apply both classifiers to the 359,621 events in the validation region (111.2-250 GeV), which were not used during training. Because these events are primarily ~~from the~~ high-mass ~~tail~~ Drell-Yan continuum production rather than the smeared Z peak~~of the Z boson signal at 91.2 GeV~~, they ~~look~~ are kinematically similar to Z-peak events (with two high-pT leptons) but at higher invariant masses. The classifier's behavior in this region acts as an honest

test of whether it has learned genuine Z-like features rather than simply memorizing the training mass window. In the validation region, the BDT assigns a mean score of 0.872 and the DNN assigns a mean score of 0.998, indicating that both classifiers recognize these events as strongly Z-like. Z decays are characterized by high transverse momentum, which explains why validation-region events score high (especially for the DNN): a heavier parent imparts more energy to its decay products, so validation-region events have higher lepton pT than even some Z-peak events. The score distributions transition smoothly from the signal region into the validation region with no discontinuity at the 111.2 GeV boundary, which gives us confidence that the classifiers' behavior in the search region (>250 GeV) is physically reasonable.

This behavior also clarifies what the classifiers can and cannot do in the blinded search region. Because they were trained to separate a 91 GeV resonance from a 40 GeV continuum, they rank events largely by pT, and in the search region, like the validation region, both the Drell-Yan background and a hypothetical Z' produce high-pT leptons. The classifiers therefore assign high scores to most validation and search region events and do not, at least efficiently, distinguish high-mass background from high-mass signal. This is expected, as a high-mass Drell-Yan event is genuinely Z-like in its kinematics, but the classifier's role in this workflow is not to identify Z' candidates but to concentrate events with Z-like topology into a category with a smooth $m_{\ell\ell}$ spectrum that is well described by a simple parametrization. The discrimination between signal and background is performed by the fit to the mass spectrum, not by initial classifier score.

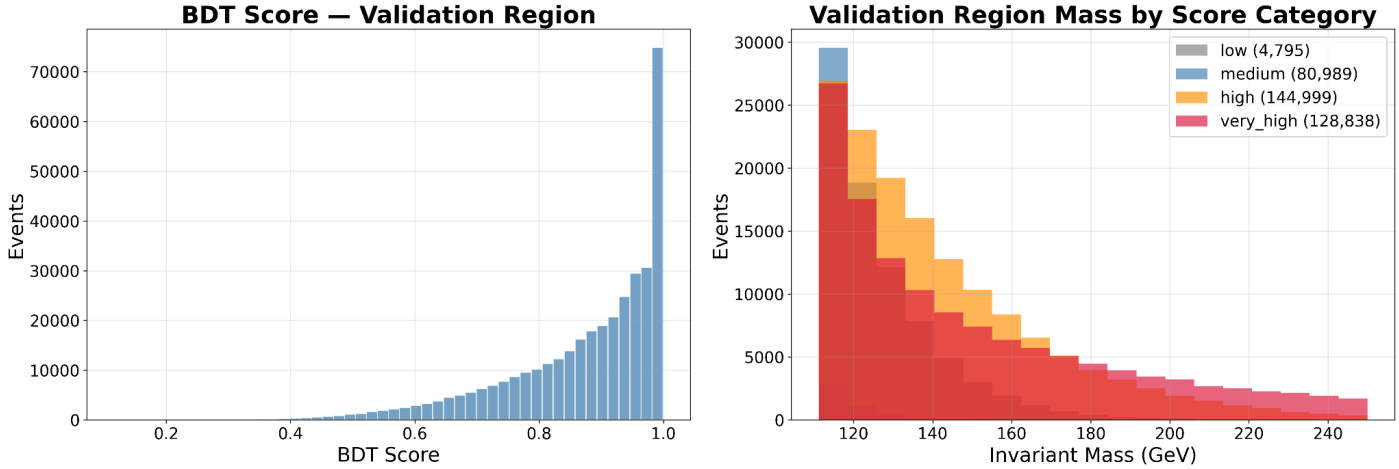


Figure 5: BDT classifier scores in the validation region (111.2-250 GeV): Distribution of BDT scores (left) and invariant mass distribution broken down by score category (right), showing a smooth transition into the search region.

5. Search and Results

5.1. Score Categories

After training and validating the classifiers, we apply them to the blinded search region starting at $m_{\ell\ell}$ = 250 GeV. Each event is scored between 0 and 1, indicating how Z-like the classifier judges it to be. Scored events are then sorted into four categories:

- Low: score < 0.5

- Medium: 0.5 ≤ score < 0.8
- High: 0.8 ≤ score < 0.95
- Very High: score ≥ 0.95

The "very high" category includes events that the classifier considers most similar to a Z boson decay. If a Z' exists in this data and decays into lepton pairs like the Z, its events would concentrate in this category and stand out against the reduced background. This category is the primary channel for our significance scan.

5.2. Background Modeling

As explained in Section 4, we perform a three-parameter fit on the events in the "very high" category and evaluated the accuracy of the fit with a χ^2/ndf value. A χ^2/ndf close to 1 indicates the fit's acceptable agreement within statistical fluctuations. A value much greater than 1 would indicate that the fit does not accurately describe the data, probably because the classifier has distorted the mass spectrum and has started background sculpting. Our fits yield χ^2/ndf = 0.99 for the BDT and χ^2/ndf = 1.38 for the DNN in the very-high category, which ~~confirms that neither classifier sculpts the background or creates false signatures~~shows no sign of a false signature or distortion that sculpting would produce. The likely reason that the DNN's value is further from 1 is because it flags ~~much~~ many more events as high-scoring, which makes it harder for it to create a smooth fit with only three parameters.

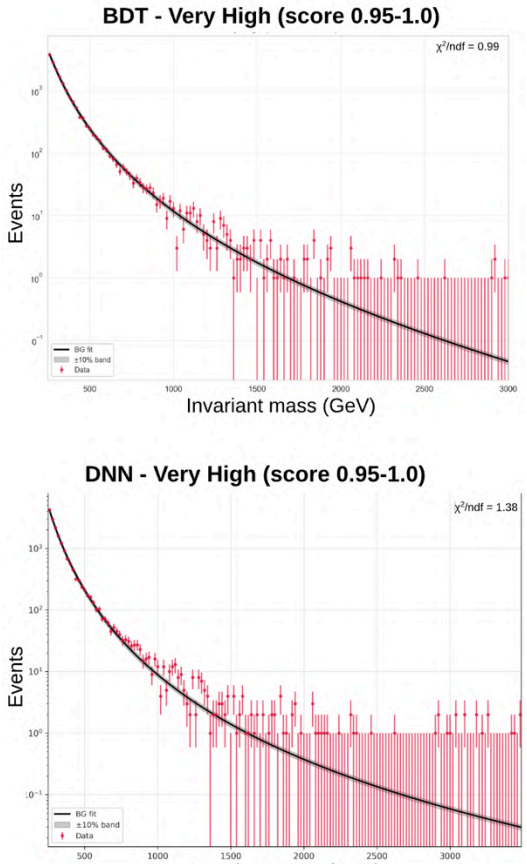


Figure 6: Three-parameter background fits in the very-high score category (score ≥ 0.95) for the BDT (top) and DNN (bottom), shown overlaid on the observed mass spectrum in 20 GeV bins. The fits yield $\chi^2/\text{ndf} = 0.99$ and 1.38 respectively.

Because a χ^2 close to 1 is a global measure, we supplement it with an examination of the score-mass relationship, where we expect background sculpting to appear. The left part of Figure 7 profiles the mean DNN scores against $m_{\ell\ell}$ across the search region. The profile is essentially flat, drifting slightly downward with mass rather than rising. The right part shows the fraction of events in each score category as a function of mass, which remains smooth with no category spiking at any specific mass. We quantify the relationship with the Pearson and Spearman correlation coefficients between DNN score and $m_{\ell\ell}$, which are -0.12 and -0.23 over the full search region and -0.15 and -0.20 within the fitted very-high category. Both are small in magnitude and negative, which means the relationship between classifier score and mass is a weak, slightly negative linear trend, so the classifier is not artificially concentrating events at a particular mass to fake a peak. A two-sample Kolmogorov-Smirnov comparison of the mass shapes in the high (n = 1,647) and very-high (n = 17,389) categories gives D = ~~a value of~~ 0.236 with $p = 1.8 \times 10^{-74}$, so the two categories do not share an identical mass distribution. This is the expected consequence of the score-mass trend quantified above rather than a sign of sculpting, since selection efficiency varies slowly across the search region and at these sample sizes the KS test resolves even a small smooth

difference. What matters for a bump hunt is whether any difference is localized, and the category fractions in Figure 7 show no category rising or falling sharply at any particular mass. ~~indicating that the categories differ smoothly and don't have any drastic divergent shapes, which means that there is no localized bump within either category.~~ Together with the $\chi^2/ndf \approx 1$ background fit, these checks find no evidence ~~establish~~ that the classifier ~~concentrates events by kinematics without forging~~ forges ~~significant~~ artificial structure into the invariant mass spectrum while concentrating events by kinematics.

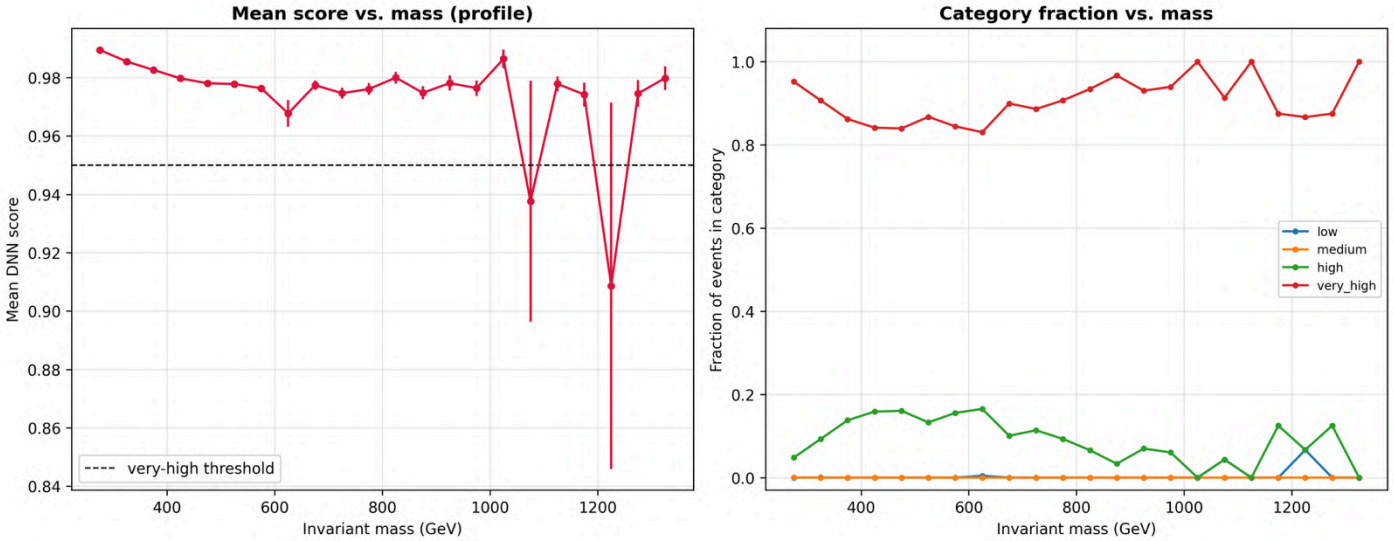


Figure 7: Classifier score to mass relationship in the DNN search region. Left: mean DNN in invariant mass bins. Right: fraction of events in each score category versus $m_{\ell\ell}$.

5.3. Significance Scan

A χ^2/ndf is a global evaluation of our background fit: it evaluates the success of the whole fit. To evaluate statistical significance of an observed excess in individual bins, we use an Asimov significance scan (Cowan et al., 2011):

$$Z_A = \sqrt{2\left[n \ln ln \left(\frac{n}{b}\right) - (n - b)\right]}$$

Where n is the number of observed events in a bin and b is the number of events predicted by the background fit. For both classifiers, we compute the Asimov significance per 20 GeV mass bins across the search region. This calculation does not rigourously treat factors such as the fitted background uncertainty, correlation across bins, or bin width, and so it can be considered an approximate local significance estimate rather than a full statistical analysis.

Across the search region, no bin in either classifier exceeded the conventionally required 5σ threshold for a discovery or even the 3σ threshold for evidence of a new particle. However, a discovery was not expected within this workflow anyway: the 3.2 fb^{-1} sample is far smaller than the full Run 2 dataset already used to exclude Z' bosons across this mass range (see Section 7 Conclusion), so this scan tests the workflow's statistical behavior rather than its discovery capability. The largest

significance we observed was at $m_{\ell\ell}$ = 1120 GeV, which both classifiers marked as the most significant bin, with the BDT yielding 2.04σ and the DNN yielding 2.83σ.

It is important to note that the reported significances use only statistical uncertainties. This should not be interpreted as an official ATLAS result as systematic uncertainties are not considered, and this is mentioned in Section 6.2 Limitations.

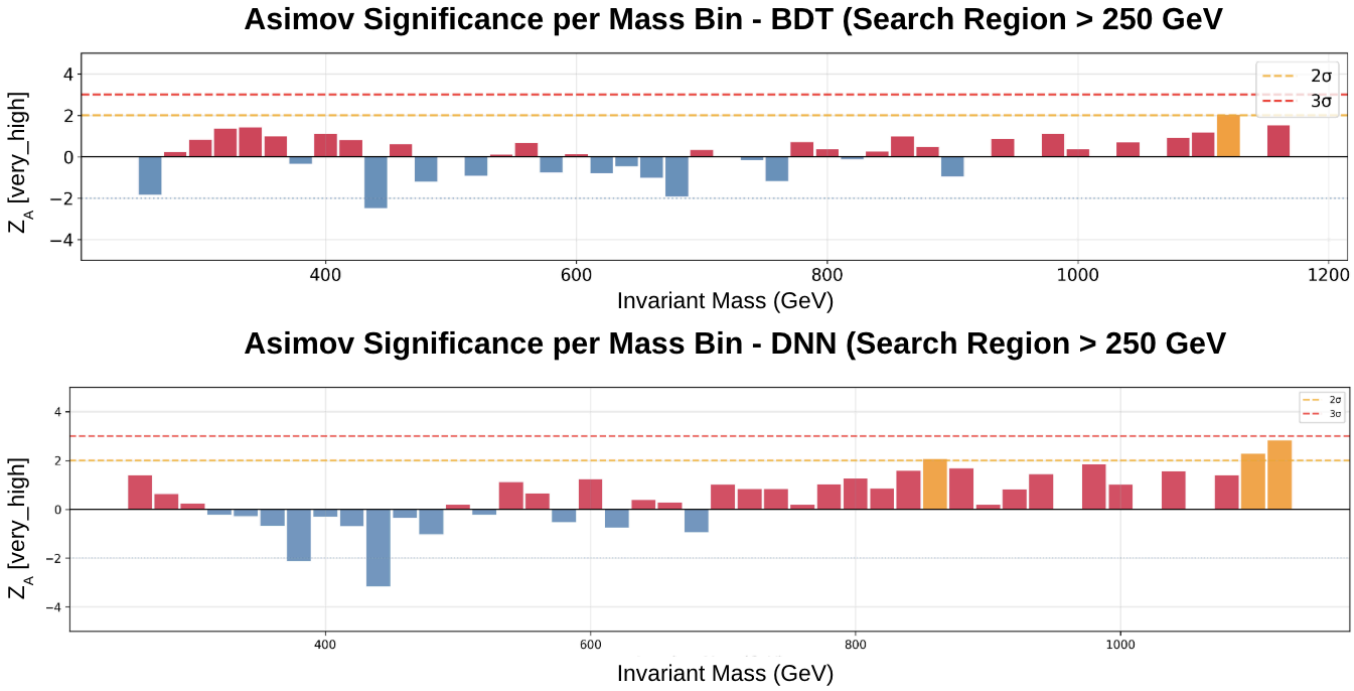


Figure 8: Asimov significance per 20 GeV bin in the very-high score category for BDT (top) and DNN (bottom). Dashed lines indicate the 2σ and 3σ thresholds. The largest deviation was 2.83σ at $m_{\ell\ell}$ = 1120 GeV in the DNN.

5.4. Look-Elsewhere Effect and p-value Calculations

This particular excess may be attributed to the look-elsewhere effect in particle physics (Gross & Vitells, 2010), which is a tendency to find a statistically significant event simply because of a large search range. Because we scan 60 mass bins (20 GeV bins from 250 to 1450 GeV, which was the largest mass bin with sufficient statistics), a fluctuation of 2σ somewhere along the mass spectrum is not very surprising. We can calculate the one-sided p-values to understand the probability of finding fluctuations this large purely by chance using this formula:

$$P_{global} = 1 - \Phi(z)^N$$

Where P_global is the probability of finding a particular excess of significance z across N number of bins and Φ(z) is the standard normal cumulative distribution function (CDF), defined as the integral of the Gaussian (Cowan et al., 2011). CDF

values are common in statistics textbooks and we did not calculate them by hand. Performing these calculations, we find an approximately 74.94% chance that a 2σ excess appears and a 14.46% chance that a 2.80σ excess appears at least once across all 60 bins because of statistical noise. This calculation treats the 60 mass bins as independent trials, which is an approximation. Because the experimental dilepton mass resolution is comparable to or broader than a single 20 GeV bin, neighbouring bins are correlated, so the effective number of independent trials is smaller than 60. A rigorous treatment would count the expected "upcrossings" of the continuous test-statistic spectrum above a threshold rather than assuming independent bins. We do not perform this continuous-trials calculation, so the probabilities above should be read as an approximate estimate rather than rigorous p-values. Nevertheless, it demonstrates that the global significance, or the significance after considering how many bins were searched, is therefore realistically much lower than 2.83σ and even 2σ. The deviation we observed is attributed to statistical noise from the large amount of data we used, and we can safely report no statistically significant Z' excess across the $m_{\ell\ell}$ = 250-1450 GeV range probed by our significance scan.

5.5 Injection Study

Because no excess is detected, we cannot determine the sensitivity (how large a resonance would have to be to be detected) of this workflow, so we perform a signal-injection study. Gaussian resonances of width equal to 2% of the resonance mass are added to the fitted background expectation of the very-high score category at 500, 1000, and 1500 GeV (Belis et al., 2023). Injecting onto the fitted expectation rather than onto the observed counts prevents whatever fluctuation already exists in a given window from being folded into the result, so the sensitivity curve begins at zero significance for zero injected events. For each injected sample, we refit the background and record the median local Asimov significance as a function of number of injected events. For the DNN, reaching the 5σ discovery threshold requires approximately 147, 38, and 17 injected events at 500, 1000, 1500 GeV respectively. For the BDT, reaching 5σ requires 157, 50, and 22 events for the same masses.

This tests the background fit and significance scan procedure rather than the classifiers themselves. The injected events are added to the fitted spectrum of the very-high score category, so the study measures how many events must appear in that category to produce a given local significance. It does not measure what fraction of genuine Z' events either classifier would place there, which would require simulated Z' sample that we do not use. So, these yields should be read as the number of selected events needed for a 5σ local excess, not as a number of produced Z' events, and the small difference between the BDT and DNN curves reflects the different background levels in their very-high categories rather than a difference in their signal efficiency.

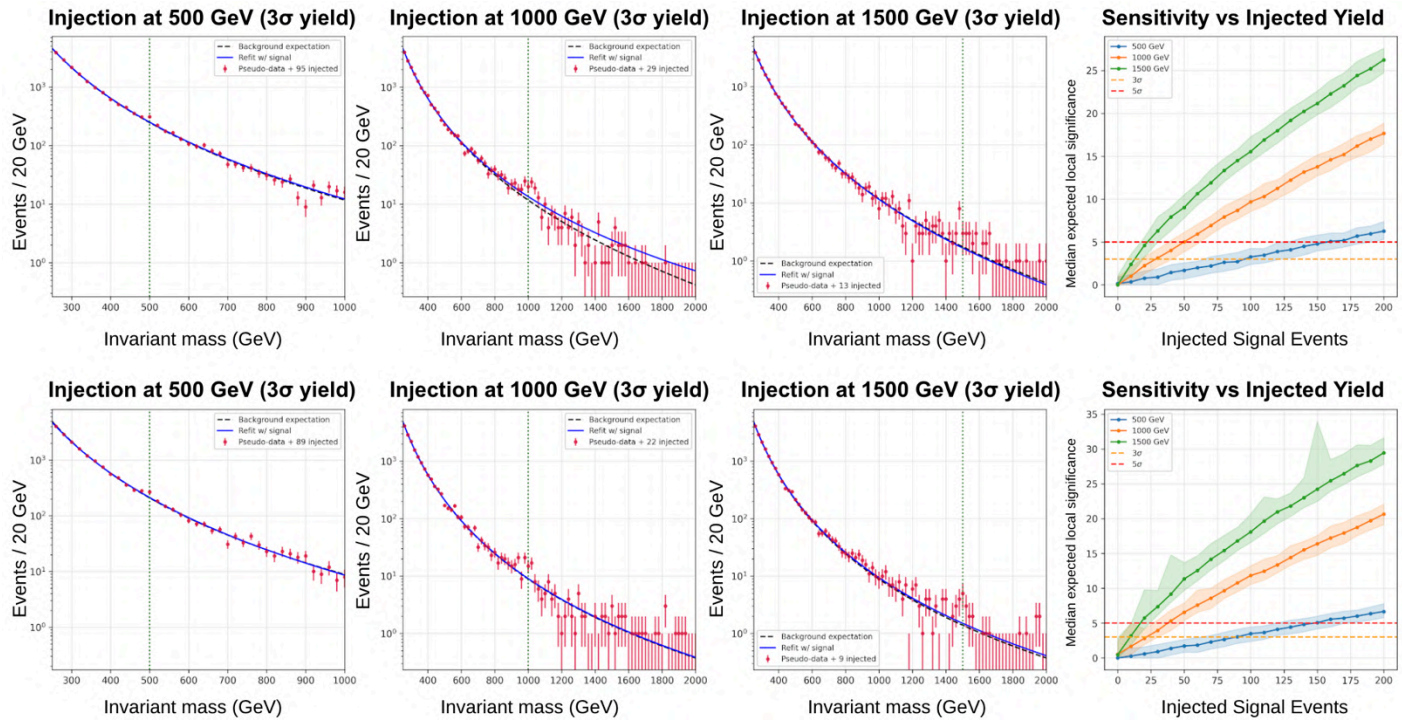


Figure 9: Signal-injection sensitivity study (top: BDT, bottom: DNN). Gaussian resonances of width 2% of the resonance mass are injected onto the fitted background in the very-high score category at 500, 1000, and 1500 GeV. The background is refit in each case. The first three plots of each row show the dataset for a 3σ yield. The right plot shows median local Asimov significance versus number of injected events.

6. Discussion

The main methodological goal of this analysis was to use ML classifiers in a bump-hunt search without distorting the background. We achieved this with three important design choices: excluding the invariant mass from features, avoiding any mass-normalized or mass-derived features, and training on non-overlapping mass sidebands. The $\chi^2/\text{ndf} \approx 1.00$ for both background fits in the very-high-score category ~~provides confirmation that this strategy works well~~is consistent with this strategy working as intended, as a classifier that learned from mass features would produce a distorted spectrum that could not be fit with a smooth three-parameter function.

While traditional approaches to bump-hunt searches involve simple kinematic cuts to select events, this ML-based approach concentrates signal-like events into higher scoring categories while preserving the smooth background shape. This suggests that ML classifiers could be useful for improving statistical sensitivity in future searches that include systematic uncertainties.

6.1. Classifier Comparison

The DNN's test AUC (0.9996) ~~substantially outperforms~~ is higher than the BDT's (0.9869), reflecting the neural network's greater ability to model complex input feature combinations. This comparison is confined to the training classes, which means the AUC measures separation of Z-peak events from low-mass sidebands, and it does not translate to a measured Z' sensitivity in the search region. The two classifiers also report different local significances in the same mass bin, but this is an observed fluctuation rather than a sensitivity measurement, and with no signal present the size of the deviation is expected to vary between classifiers. ~~This resulted in the DNN's higher sensitivity in the search region, as it~~ The DNN yields a 2.83σ significance at $m_{\ell\ell}$ = 1120 GeV compared to the BDT's 2.04σ. Despite the agreement between classifiers, ~~this~~ neither value approaches a significance threshold ~~is not~~ sufficient ~~enough~~ for this to be considered evidence or discovery. Also, determining which classifier is genuinely more sensitive to a Z′ would require simulated signal samples scored by each model, which this analysis does not use.

The two classifiers both offer their own strengths. The BDT's feature importance analysis provides insight into the classification, with a lepton pT dominated scoring (expected from a Z decay). It is more complicated to determine the decision process of the DNN because it uses more complex feature combinations. Nevertheless, presenting both classifiers and their independent results provides cross-check validations and strengthens the credibility of the result. The fact that both models agreed on a null result at the same particular excess increases the confidence in this conclusion.

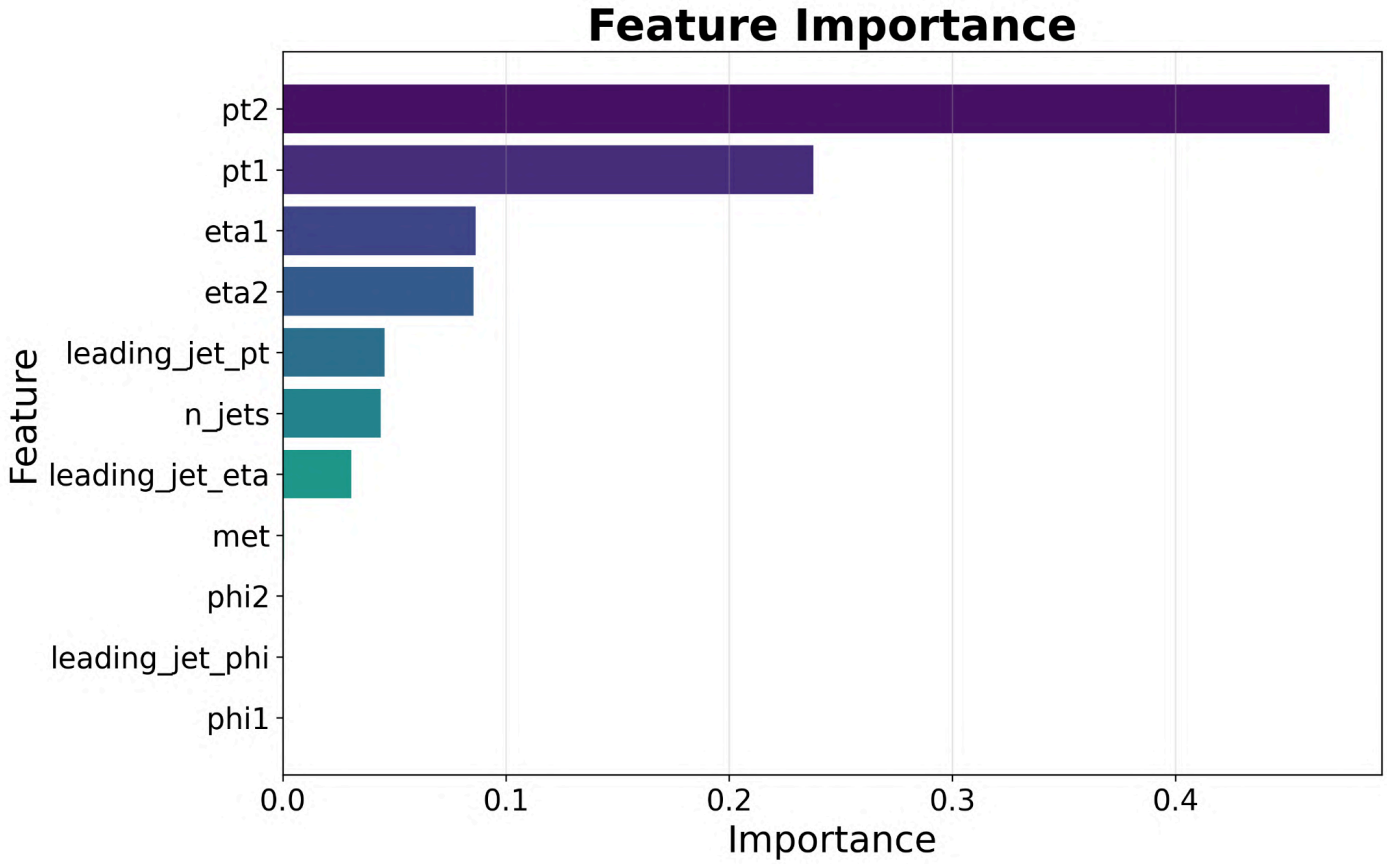


Figure 10: BDT feature importance. Lepton transverse momenta dominate, with pT^2 contributing 47% and pT^1 contributing 24%.

6.2. Limitations and Future Improvements

This analysis is designed as a methodological demonstration rather than a full experimental result, and improvements are needed before producing results on par with official CERN and ATLAS publications:

- No comparison to theoretical models: We do not compare our results against existing benchmark theories and simulated signal events. A complete analysis would interpret the search within one or more theoretical frameworks, predict the expected signal yield for each model, and set quantitative exclusion limits on their parameters.
- Systematic uncertainties: We use only statistical uncertainties from the limited number of events in our selected bins. A complete and proper analysis would include systematic uncertainties on factors like luminosity, lepton identification efficiencies, and choice of background function.
- Alternative background functions: We use a single three-parameter function, but a professional analysis might compare several functional forms and include any variation as an additional uncertainty.
- Mass decorrelation: The DNN's very high AUC (>0.999) warrants some caution about mass correlation in the training data set. While the background fit ~~confirms~~ shows no evidence of sculpting, techniques such as adversarial training or distance-correlation penalties could provide stronger guarantees against mass correlations.
- Look-elsewhere estimate: The global significance reported in Section 5.4 is computed by treating the mass bins as independent trials. In practice, the detector mass resolution correlates neighbo~~u~~ring bins, so this overstates the trials factor. A rigorous global p-value would instead use the continuous-trials (upcrossings) formalism of Gross and Vitells (2010). Because our largest local deviation is already well below the 3σ evidenc~~e~~ threshold, this does not affect the null result, but it would need to be addressed in any analysis aiming for a discovery-level statistical statement.
- Absolute azimuthal angle (phi): Because of rotational symmetry about the beam, absolute azimuthal angle is relatively meaningless, as indicated by the BDT feature importance (Figure 10). A model utilizing it may be learning detector artifacts. In the future, using relative angles rather than absolute angles may be more valuable.
- Indirect mass learning: As mentioned, invariant mass is excluded as an input feature for both classifiers to mitigate mass correlation. However, since other input variables are used to calculate invariant mass, indirect mass-correlated learning is still possible for the classifiers.
- Combined lepton channels: Our event selection collects dielectron and dimuon events into the same dataset with the same 11 features. While lepton flavor is recorded, it isn't used as an input feature for the classifiers. Because electrons and muons are reconstructed by different subsystems and their momentum resolutions degrade differently, a single model trained on the combined sample is forced to average channel-specific kinematic behavior. A future analysis should either train separate classifiers for each channel or include lepton flavor as an input feature.
- Top-quark background and b-jet veto: Our classifiers are trained to separate Z-peak events from low-mass Drell-Yan events, both of which are characterized by low missing transverse energy and low jet multiplicity. The models therefore have no opportunity to learn MET or jet multiplicity as discriminators, as indicated by their low

to zero contribution in the feature importance (Figure 10). The second most significant background in high-mass dilepton resonance searches, top-quark pair production, produces two high-pT leptons together with genuine MET from neutrinos and two b-jets. Because our classifiers rank events primarily by lepton pT, these events might be scored high despite not being Z-like. We do not apply a b-jet veto, the standard method for suppressing top-quark contamination, as b-tagging information was not collected. Future analyses should include b-tagging information and apply a b-jet veto.

- Absence of an opposite-sign requirement: The Z and Z' are electrically neutral, so charge conservation requires their decay products to be an oppositely charged lepton pair, a standard element of the dilepton "golden channel" selection. We do not apply this requirement, because lepton charge was not retained during our event selection and is therefore unavailable in our processed dataset. Our sample therefore contains some same-sign contamination. These contributions, however, are smooth and non-resonant in $m_{\ell\ell}$, so they inflate the background under the fitted spectrum and reduce sensitivity, rather than introducing a localized excess or biasing the search. The effect on training is also limited, as the Z-peak and low-mass Drell-Yan regions are mainly populated by opposite-sign events. Nevertheless, applying an opposite-sign requirement is the first correction we would make in a future analysis.

7. Conclusion

We have presented a machine-learning based search for a Z' boson using ATLAS Run 2 2015 open data, comparing a gradient boosted decision tree and a three-layer MLP trained on mass sideband kinematics. By intentionally excluding dilepton invariant mass from the input features and training on mass sidebands outside of the search region, both classifiers achieve strong discrimination between the sideband background proxy and the Z-like proxy signal events (AUC = 0.9869 and 0.9996 respectively) while describing the high-mass background spectrum with a simple three-parameter function.

Our Asimov significance scan across bins in the search region produces no bins exceeding the 3σ evidence threshold. Our largest local deviation is 2.83σ at $m_{\ell\ell}$ = 1120 GeV using the DNN, which can be considered a statistical fluctuation when the look-elsewhere effect is considered. We found no statistically significant localized excess across the 250-1450 GeV range probed with sufficient statistics; however, this does not imply a definitive null result for a Z' boson, even though a null result is expected in this mass range. Indeed, our 3.2 fb^{-1} sample is a subset of the Run 2 dataset with a full 139 fb^{-1} luminosity that has already been used to exclude Z' bosons across this mass range. A genuine resonance large enough to produce a visible excess in our sample would have appeared at overwhelming significance in the complete dataset, so any localized deviation we observe is almost certainly a statistical fluctuation rather than new physics just by nature of the dataset.

Nevertheless, the value of this work lies in the workflow rather than the search outcome, as the methodology was largely successful despite what the search yielded. We were able to utilize public data and simple kinematic features to formulate a completely reproducible pipeline for an ML-driven bump hunt search that ~~successfully mitigates~~ shows no evidence of harmful background sculpting and can be used as a template for future bump hunt analyses using LHC open data. The Z' search presented here should therefore be read as a demonstration of the workflow rather than a competitive experimental search, and the lack of a significant excess reflects the limited integrated luminosity of the dataset rather than a new exclusion of Z' models.

Many improvements would sharpen this workflow from a methodological demonstration into a more rigorous search. The most important corrections concern selection: applying an opposite-sign charge requirement and a b-jet veto would remove the same-sign and top-quark contamination that inflates the background, and training separate classifiers for dielectron and dimuon channel (or using lepton flavor as an additional input feature) would stop the models from averaging over the differing momentum resolutions. Beyond these fixes, there are statistical and interpretive extensions to be made. Considering systematic uncertainties on luminosity and the choice of background function, comparing alternative background parametrization, and replacing the independent-trials look-elsewhere estimate with the more accurate continuous-trials formalism would give any reported significance a more rigorous meaning. Taken together and with other limitations addressed in Section 6.2, these steps define a clear path from a reproducible template toward an analysis with professional-level statistical conclusions.

8. References

ATLAS Collaboration. (2008). The ATLAS experiment at the CERN Large Hadron Collider. Journal of Instrumentation, 3(8), S08003. https://doi.org/10.1088/1748-0221/3/08/S08003

ATLAS Collaboration. (2016). Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 76(5), 292. https://doi.org/10.1140/epjc/s10052-016-4120-y

ATLAS Collaboration. (2019a). Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 79(8), 639. https://doi.org/10.1140/epjc/s10052-019-7140-6

ATLAS Collaboration. (2019b). Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s}$ = 13 TeV with the ATLAS detector. Physics Letters B, 796, 68–87. https://doi.org/10.1016/j.physletb.2019.07.016

ATLAS Collaboration. (2024). ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data [Data set]. CERN Open Data Portal. https://doi.org/10.7483/OPENDATA.ATLAS.AOQL.8TT3

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. Nature Communications, 5, 4308. https://doi.org/10.1038/ncomms5308

Belis, V., Odagiu, P., & Aarrestad, T. K. (2023). Machine learning for anomaly detection in particle physics. Reviews in Physics, 12, 100091. https://doi.org/10.1016/j.revip.2024.100091

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

CMS Collaboration. (2021). Search for resonant and nonresonant new phenomena in high-mass dilepton final states at $\sqrt{s}$ = 13 TeV. Journal of High Energy Physics, 2021(7), 208. https://doi.org/10.1007/JHEP07(2021)208

Collins, J. H., Howe, K., & Nachman, B. (2019). Extending the search for new resonances with machine learning. Physical Review D, 99(1), 014038. https://doi.org/10.1103/PhysRevD.99.014038

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. European Physical Journal C, 71(2), 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0

Dolen, J., Harris, P., Marzani, S., Rappoccio, S., & Tran, N. (2016). Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure. Journal of High Energy Physics, 2016(5), 156. https://doi.org/10.1007/JHEP05(2016)156

Drell, S. D., & Yan, T.-M. (1970). Massive lepton-pair production in hadron-hadron collisions at high energies. Physical Review Letters, 25(5), 316–320. https://doi.org/10.1103/PhysRevLett.25.316

Evans, L., & Bryant, P. (2008). LHC machine. Journal of Instrumentation, 3(8), S08001. https://doi.org/10.1088/1748-0221/3/08/S08001

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feickert, M., & Nachman, B. (2021). A living review of machine learning for particle physics. arXiv. https://arxiv.org/abs/2102.02770

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. European Physical Journal C, 70(1–2), 525–530. https://doi.org/10.1140/epjc/s10052-010-1470-8

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 448–456). PMLR. https://arxiv.org/abs/1502.03167

Langacker, P. (2009). The physics of heavy Z′ gauge bosons. Reviews of Modern Physics, 81(3), 1199–1228. https://doi.org/10.1103/RevModPhys.81.1199

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1711.05101

Metodiev, E. M., Nachman, B., & Thaler, J. (2017). Classification without labels: Learning from mixed samples in high energy physics. Journal of High Energy Physics, 2017(10), 174. https://doi.org/10.1007/JHEP10(2017)174

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807–814). Omnipress.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Terao, K., & Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. Nature, 560(7716), 41–48. https://doi.org/10.1038/s41586-018-0361-2

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., & McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research Section A, 543(2–3), 577–584. https://doi.org/10.1016/j.nima.2004.12.018

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Data & Code Availability Statement

The data used for this experimental setup was accessed via the ATLAS Open Data portal. Link to dataset: https://opendata.cern.ch/record/80000

Cleaned-up scripts used for this methodology are available at this public GitHub repository: https://github.com/TheRealM1M1/zprime_search.

Acknowledgements

I would like to thank **Professor Kaushik De** from the University of Texas at Arlington Department of Physics for his continued support and guidance during this project.

I acknowledge the work of the ATLAS Collaboration to record or simulate, reconstruct, and distribute the Open Data used in this paper, and to develop and support the software with which it was analyzed.

AI Use Statement

Claude Opus 4.7/4.8 (Anthropic) was used for minor editing support and minor coding support during this project. All methodological decisions, analysis, results, and scientific content are the author's own.

Author Biography

Mihir Tare is a student at the St. Mark's School of Texas, where he enjoys studying physics and math. He leads his school's

Physics Club and has represented his school in local and national competitions, including F=MA and Physics Bowl. He is particularly interested in high-energy particle physics and the Standard Model, which is why he took on this research to investigate a hypothetical particle using real CERN data and modern machine-learning strategies. Outside of physics, he enjoys playing flute in his school's band, his local youth orchestra, and in the many All-State ensembles which he has been a part of. He is also an avid Latin student who likes to compete in Classics competitions and learn about topics like Ancient Roman life, Latin vocabulary, and Latin and Greek derivatives. He was a club swimmer and water polo player and has played on his high school team, but now he likes to stay active in his free time by lifting weights, riding the bike that he built, or playing recreational sports like spikeball and volleyball. In the future, he hopes to study engineering or applied physics and wants to help discover and innovate at the forefront of science.

Mentor Contribution Statement

Professor Kaushik De of the University of Texas at Arlington served as the primary academic mentor for this project and contributed substantially to its intellectual foundation, methodological design, and scholarly presentation. He guided the author through weekly meetings throughout the project, beginning with foundational instruction in experimental particle physics. During this initial phase, Professor De provided suggested readings on the Standard Model, collider phenomena, and new-physics searches, and answered the author's conceptual questions in detail. The instruction directly shaped the project's theoretical framing and motivated the choice of the dilepton search channel.

Professor De introduced the author to the ATLAS Open Data release from LHC Run 2 and provided technical guidance on accessing, handling, and interpreting the dataset. During the computational phase of the project, he offered advice on machine learning models relevant to high-energy physics and suggested techniques to use. In weekly meetings, he reviewed results with the author and gave advice on how to interpret them.

Finally, Professor De reviewed and revised multiple drafts of the final paper, providing feedback on technical accuracy, clarity, and structure prior to submission.

Summary

The manuscript presents an ambitious and well-motivated machine-learning-assisted search for a possible Z′ boson in the dilepton channel using publicly available ATLAS Run 2 open data. The analysis follows the logic of a resonance search: a heavy neutral boson decaying to an electron or muon pair would be expected to appear as a localized excess in the dilepton invariant-mass spectrum over the smoothly falling Drell–Yan background.

The author develops a complete analysis workflow, beginning with same-flavour dilepton event selection and reconstruction of relevant kinematic variables. The dilepton invariant mass is used to define training, validation, and search regions, while the mass itself is excluded from the classifier inputs to reduce the risk of sculpting artificial structures in the search spectrum. Two classifiers, a boosted decision tree and a deep neural network, are trained to distinguish Z-peak events from lower-mass Drell–Yan sideband events using lepton kinematics, jet information, and missing transverse energy.

The trained classifiers are then applied to the high-mass region above 250 GeV. Events are divided into score categories, with the highest-score category used as the primary search sample. In this region, the author models the background with a smooth three-parameter fit and performs a bin-by-bin significance scan. The reported background fits have reduced chi-squared values close to unity, and neither classifier finds an excess reaching the conventional threshold for evidence or discovery. The largest local deviation is reported near 1120 GeV, with a significance of 2.40σ for the neural network and 2.04σ for the boosted decision tree, which the manuscript interprets as consistent with a statistical fluctuation after accounting for the look-elsewhere effect.

Overall, the manuscript is technically ambitious and clearly organized. It brings together collider physics motivation, event selection, machine-learning classification, background modeling, and statistical interpretation in a coherent analysis. The work is especially strong in its effort to address background sculpting and to compare two independent classifiers. The main result is a null search outcome, but the broader contribution of the paper is the construction of a clear and reproducible framework for applying machine-learning methods to a realistic dilepton resonance search using open collider data.

Requested Changes

1. **Reframe the main claim more carefully.**
   The manuscript should present the analysis as an ML-assisted dilepton resonance-search workflow using ATLAS Open Data, rather than as a fully validated Z′ search. The current wording sometimes suggests a stronger experimental conclusion than the analysis can support.
2. **Clarify the meaning of “signal” and “background” in training.**
   The Z-peak events used for training are Standard Model Z events, not actual Z′ signal events. The author should describe these as a “Z-like proxy signal” or “resonance-like proxy class,” and the low-mass Drell–Yan events as a sideband/background proxy. This would make the methodology more precise without weakening the paper.
3. **Address mass leakage more directly.**
   Excluding dilepton invariant mass from the input features is a good first step, but the input variables still include the lepton (p_T), (\eta), and (\phi), from which invariant mass is calculated. The author should explicitly acknowledge that the classifiers may still learn mass-correlated information indirectly.

4. **Strengthen the background-sculpting validation.**
   The paper should not rely only on (\chi^2/\mathrm{ndf}) values close to one to claim that background sculpting has been avoided. The author should add score-versus-mass plots, mass distributions in each classifier-score category, and a quantitative check of the correlation between classifier score and (m_{\ell\ell}).
5. **Soften the statement that the MLP "eliminates" background sculpting.**
   The manuscript says the switch from a GNN to an MLP eliminated the sculpting problem. That is too strong. A better wording would be that the MLP "reduced" or "mitigated" the issue, pending further validation.
6. **Add a simple signal-injection study.**
   Even a basic toy injection of narrow Z′-like bumps at a few masses would greatly improve the paper. This would show whether the method could actually recover a resonance if one were present, rather than only showing that no large bump appears in the data.
7. **Improve the statistical interpretation.**
   The bin-by-bin significance scan should be described as an approximate local-significance estimate, not a full discovery-level statistical analysis. The author should mention that the fitted background uncertainty, bin-to-bin correlations, choice of bin width, and trials factor are not fully treated.
8. **Revise the look-elsewhere discussion.**
   The look-elsewhere calculation is useful pedagogically, but the assumption of 60 independent bins is too simple. The author should present it as an approximate estimate and avoid treating the resulting global probabilities as rigorous final p-values.
9. **Clarify the searched mass range.**
   The paper should be consistent about the usable search range. It mentions no evidence from 250–3000 GeV, but the significance scan and look-elsewhere discussion appear to use a smaller statistically meaningful range. The conclusion should match the range actually analyzed with sufficient statistics.
10. **Give more detail on event selection.**
    The analysis should specify whether opposite-charge lepton pairs are required, what lepton (p_T) thresholds are used, what trigger requirements are applied, whether lepton identification/isolation criteria are used, and whether electron and muon channels are combined or treated separately.
11. **Make the reproducibility claim stronger.**
    The data source is cited, but the paper should also provide the analysis code, exact cuts, trained-model settings, random seeds, binning choices, and scripts or notebooks needed to reproduce the figures and results.
12. **Clarify the role of systematics.**
    The limitations section mentions missing systematic uncertainties, which is good. The main text should also make clear that the reported significances use statistical uncertainties only and therefore should not be interpreted like an official ATLAS/CMS result.
13. **Check figure consistency and completeness.**
    The paper reports both BDT and DNN AUC values, but the ROC figure shown appears to be only for the BDT. The author should either include the DNN ROC curve or explain why it is omitted. The overtraining-check caption should also be corrected if the plots are stacked rather than shown left/right.
14. **Tighten a few physics explanations.**
    Some background explanations should be made more precise. For example, not all Z′ models couple to quarks and leptons "just like the Z," and the description of why muons reach the muon spectrometer should be revised.
15. **Moderate the conclusion.**
    The final conclusion should say that no statistically significant localized excess is found in this

analysis, rather than implying a broad exclusion or definitive null result for Z′ bosons. A strong ending would emphasize the successful construction of the workflow and the clear next steps needed to make it more rigorous.

Final Decision: Revise and resubmit

Review of Submission 100196

("Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data")

Final Recommendation: Accept with minor revisions

Thank you for submitting this exceptional manuscript to Convergence Journal! Approaching this review with a mentorship mindset, I want to first commend you on an incredibly thorough, well-reasoned, and scientifically rigorous paper. You have successfully bridged theoretical particle physics, experimental detector mechanics, and modern data science in a way that is accessible yet technically profound. The structure is highly logical, and your execution of the "bump hunt" methodology while actively avoiding background sculpting demonstrates a deep understanding of the exact challenges modern physicists face. This is a fantastic contribution, and with just a few minor formatting and structural adjustments, it will be fully ready for publication.

Here is the evaluation of your manuscript based on the journal's key criteria, along with specific recommendations for revision:

1. Originality & Significance

The paper contributes a novel, reproducible methodological pipeline for conducting a "bump hunt" using public ATLAS 13 TeV Open Data. By successfully implementing a Boosted Decision Tree (BDT) and a Deep Neural Network (DNN) without relying on dilepton invariant mass as an input feature, you demonstrate how to improve statistical sensitivity while safely preserving the smooth Drell-Yan background model. The scientific contribution and methodological demonstration are excellent and highly significant for educational and prototyping purposes.

2. Clarity & Structure

The argument is exceptionally well-organised, guiding the reader logically from the theoretical shortcomings of the Standard Model through complex detector mechanics and sophisticated ML training paradigms. The main points and research goals are evident throughout the text.

3. Use of Evidence & Research Methods

Your methodology is sound, carefully constructed, and meticulously explained. You provide robust justification for feature selection, hyperparameter tuning (such as learning rates, L1/L2 regularisation, and dropout), and statistical evaluation methods like the Asimov significance scan and χ^2 fitting. All sources appear real and are appropriately cited in APA format.

Equation Formatting: Currently, mathematical formulas such as the invariant mass calculation, the background fit function, the binary cross-entropy loss, and the Asimov significance are written in plain text characters. The Style Guide mandates that all equations must be written using the built-in math tools in Microsoft Word, Google Docs, or similar software. Please reformat these mathematically (e.g., ensuring square roots and fractions are properly typeset). The format could be improved by using a dedicated tool to write the equations to help the reader navigate faster through them. With the current chosen formatting, it is rather difficult to quickly understand, for example, where a square root ends or whether superscripts represent indices or powers. Taking a brief moment to properly typeset these formulas will eliminate this ambiguity and ensure your excellent methodology is as readable as possible.

4. Engagement with Literature

The paper demonstrates a deep understanding of high-energy physics literature. You perfectly contextualise the search within both theoretical frameworks (citing Langacker's work on heavy gauge bosons) and contemporary experimental limits from the ATLAS and CMS collaborations. You also expertly engage with statistical literature, correctly applying and citing concepts like the look-elsewhere effect.

5. Grammar & Language

The writing is remarkably clear, professional, and free of major grammatical errors. You effectively communicate highly technical concepts in a way that is accessible to a broad academic audience. Once you update your equations using a math editor, please review the manuscript to ensure that technical notation remains visually consistent in the prose. For example, ensure the formatting of variables like p_T and $m_{\ell\ell}$ is uniform throughout the text blocks.

Overall, the manuscript is of a very high standard. Nonetheless, further refinements are still possible, and please consider the recommendations in the list below.

1. You noted a pseudorapidity cut for electrons at a rapidity of less than 2.47, which correctly aligns with the ATLAS electromagnetic calorimeter precision region. However, standard ATLAS analyses typically exclude the transition region between the barrel and endcap cryostats, known as the "crack region" (1.37 < rapidity < 1.52). This region is heavily instrumented with cables and dead material, which significantly degrades electron energy resolution. While this might be a minor effect in this simplified Open Data analysis, acknowledging this standard exclusion (or explicitly stating why you chose not to apply it) would demonstrate a deeper understanding of the ATLAS detector's physical limitations.
2. You made the thoughtful design choice to exclude the dilepton invariant mass from your training features to prevent the models from sculpting artificial bumps into the background. However, you provided the exact constituent variables of the invariant mass equation (pT^1, pT^2, η^1, η^2, φ^1, φ^2) to a Deep Neural Network. Due to the universal approximation theorem, the DNN is highly capable of reconstructing a proxy for mass

using these variables. This is evidenced by your finding that p_T dominates the classification and that validation-region events score exceptionally high. While your χ^2 fit confirms the resulting background remains smooth, the network is still indirectly shaping the background by acting as a dynamic, highly complex p_T cut. Adding a paragraph in your "Limitations" section discussing how providing highly correlated constituent variables can still lead to indirect mass learning would greatly elevate the scientific maturity of your discussion.

3. Following from the previous point, your classifiers correctly learned that "higher p_T = more Z-like (signal)" to separate the 91 GeV Z-peak from the 40 GeV Drell-Yan background during training. Consider what this means physically in your >250 GeV search region. In this high-mass regime, both the irreducible Drell-Yan background and the hypothetical Z' signal will produce leptons with extremely high p_T. If your classifier primarily relies on high p_T to assign a "very high" score, it may simply pass nearly all search-region events into that top category, effectively losing its discriminative power between high-mass background and high-mass signal. Acknowledging this physical limitation of training on low-mass kinematics and extrapolating to high-mass searches is a vital nuance that professional physicists grapple with, and discussing it will make your paper much stronger.
4. You mention that L1 regularisation (Lasso) "drives the weight of less important features to zero... essentially performing a feature selection". While this is perfectly true for linear regression models, in the context of tree-based algorithms like XGBoost, L1 regularisation (alpha) penalises the absolute value of the leaf weights. It encourages leaf values to be exactly zero, which can lead to tree pruning, but it does not directly zero out an input feature's global importance in the way Lasso regression does. Consider slightly rewording this explanation for technical exactness.
5. In Table 1, the author lists the absolute azimuthal angles and absolute pseudorapidities of the leptons as input features for the BDT and DNN. In particle physics, the absolute azimuthal angle (phi) of an event is generally physically meaningless because the proton-proton collisions have rotational symmetry around the beam axis. If a machine learning model learns to rely on absolute phi, it might actually be learning the location of dead sensors or localised noise in the ATLAS detector rather than underlying physics. It is a standard best practice to use relative variables instead, such as the angular separation ($\Delta\eta$ or ΔR or $\Delta\phi$). I recommend that the author note this in their "Limitations" section, or better yet, and suggest they swap absolute angles for relative angles in future analyses.
6. To account for the Look-Elsewhere Effect, the author calculates a global p-value by treating the search region as N = 60 perfectly independent mass bins. This is an excellent educational approximation, but it is an oversimplification of how the Look-Elsewhere Effect works in continuous mass spectra. Because the experimental mass resolution is often wider than a single bin, neighbouring bins are highly correlated. The very paper the author cites for this (Gross & Vitells) specifically provides a methodology based on the "upcrossings" of a continuous χ^2 process, rather than assuming independent discrete bins. While the basic probability calculation in the manuscript successfully illustrates the concept of the LEE, a professional analysis would compute the global significance using the asymptotic formulas for continuous trial factors. This potentially more thorough calculation is not needed here, but it could be maybe discussed in the limitations section.

7. When justifying the bounds of the signal region, the author states, "While the Z boson has a natural width of 2.5 GeV, we use a larger buffer to avoid any Z contamination during analysis". The author is encouraged to double-check that there is no confusion between the theoretical natural decay width and the experimentally reconstructed width. While the quantum mechanical Breit-Wigner width of the Z boson is indeed about 2.5 GeV, the actual "bump" observed in the ATLAS detector is a convolution of this natural width and the Gaussian energy/momentum resolution of the detector subsystems (the calorimeters and muon spectrometer). The observable peak is physically wider than 2.5 GeV because of detector smearing. The author should correct this sentence to acknowledge that the buffer accounts for both the natural width and detector resolution effects.
8. In your Data Preparation section, you define your event selection as requiring exactly two same-flavour leptons (dielectron or dimuon). You must explicitly state whether you required these leptons to have opposite electrical charges. Because the Z and Z' bosons are electrically neutral, conservation of charge dictates that they must decay into an oppositely charged pair (i.e., electron+positron or muon+ anti-muon). If you did not apply an opposite-sign (OS) requirement, your data likely contain unnecessary contamination from same-sign (SS) fake backgrounds and diboson processes. If you did apply this cut, be sure to add a sentence to explicitly mention it, as it is a fundamental requirement of the "golden channel."
9. Your machine learning framework trains the models to specifically distinguish Z-peak events (signal) from low-mass Drell-Yan events (background). By training exclusively on Drell-Yan, your classifier is blind to the second most dominant background in high-mass dilepton searches: top-quark pair production. Because top quarks decay into W bosons and b-quarks, top+top-bar events produce two high-pT leptons, missing transverse energy (MET) from neutrinos, and multiple jets. Since you are already passing n_jets, jet kinematics, and MET into your classifiers, your model might be inadvertently scoring top+top-bar events as highly Z-like simply because they are highly energetic. Noting the lack of a b-jet veto (the standard method for removing top-quark contamination) in your Limitations section will demonstrate a mature understanding of background topologies.
10. You note that your event selection aggregates both dielectron and dimuon events, and you apply the same 11 input features to both sets. In the ATLAS detector, electrons and muons have fundamentally different detection mechanisms (calorimeters vs. spectrometers), and their momentum resolutions scale differently at high energies. By combining them into one dataset without providing the classifier a feature to distinguish them (such as an is_muon boolean), the model is forced to average out channel-specific kinematic nuances. Consider adding a note in your limitations that future analyses should either train separate models for the electron-positron and muon-anti-muon channels or include lepton flavour as a designated input feature.
11. You creatively used the Standard Model Z boson peak (71.2-111.2 GeV) as your "signal" during training because you opted not to use simulated Z' events. While this brilliantly avoids relying on theoretical simulations, it inherently assumes that a 2000 GeV Z' looks kinematically identical to a 91 GeV Z boson, just scaled up. In reality, the phase space, angular momentum conservation constraints, and parton distribution functions at the multi-TeV scale produce topologies that differ from a 91 GeV decay. Adding a sentence in your limitations acknowledging that a classifier

optimized for the Z pole might be sub-optimal for capturing the exact topology of a multi-TeV Z' will demonstrate a highly sophisticated understanding of collider kinematics.

12. You accurately state that your 2015 dataset has an integrated luminosity of 3.2 fb^{-1}, and you note that official full Run 2 analyses utilised 139 fb^{-1}. Because the full 139 fb^{-1} dataset has already ruled out a Z' in this mass range, finding a true, physical Z' in a 3.2 fb^{-1} subset of that exact same data is physically impossible; any localised excess would strictly be a statistical fluctuation. While you perfectly frame your paper as a "methodological demonstration", explicitly mentioning that the limited luminosity intrinsically precludes a novel discovery adds a mature, grounded perspective to your final discussion.

To conclude, this manuscript is an outstanding example of early-career research that tackles a genuinely complex challenge in high-energy physics data analysis. Your creative approach to mitigating background sculpting through careful feature selection demonstrates a level of critical thinking that will serve you incredibly well in your future academic endeavours. The feedback provided above, ranging from minor formatting adjustments to deeper considerations of detector physics and statistical framing, is intended strictly to help you polish an already excellent foundation into a publication-ready manuscript. I highly encourage you to incorporate these nuances, particularly in your limitations and discussion sections, as they reflect the mature, rigorous mindset of a professional physicist. Thank you for submitting such a compelling and well-executed paper to Convergence Journal; I very much look forward to reading your revised manuscript!

Convergence
A Journal for Young Researchers

Received: [for official use only] ; Accepted: [for official use only];
Managing Editor: [for official use only];
Action Editor: [for official use only]

Machine-Learning-Driven Search for a Z' Boson in the Dilepton Channel Using ATLAS Open Data

[name redacted]

[redacted]

Abstract

The Standard Model of particle physics, while highly successful, leaves several unanswered questions that motivate searches for undiscovered heavy gauge bosons predicted by Beyond-Standard-Model theories. Using machine-learning classifiers trained on low-mass sideband events, this paper presents a search for the Z' boson, one such hypothetical heavy neutral gauge boson that decays into dilepton pairs. A gradient boosted decision tree (BDT) and a multi-layered perceptron deep neural network (DNN) are trained to distinguish real Z boson events from Drell-Yan background events using only 11 raw kinematic features, with dilepton invariant mass deliberately excluded to avoid background sculpting of artificial signals during the search. The analysis uses real proton-proton collision data recorded by the ATLAS experiment during LHC Run 2 in 2015, released through the CERN Open Data portal in 2024. After both classifiers achieve accurate discrimination on the training set, events in a blinded high-mass search region are scored by the classifiers, a three-parameter parametric fit is used to model the background, and an Asimov significance scan is performed across mass bins. We observed no statistically significant excess, with the largest local deviation at a mass of 1120 GeV with a significance of 2.40σ in the DNN's highest-score category, which is well below the 3σ conventionally required for evidence of new physics. This analysis demonstrates how machine-learning can be used in high-energy physics "bump hunt" searches while preserving the background model, and it provides a reproducible methodology built on publicly available data.

Keywords:
Z′ boson, dilepton resonance search, machine learning, boosted decision tree, deep neural network, ATLAS Open Data, bump hunt, beyond the standard model

1. Goals and Theory

The Standard Model (SM) of particle physics provides a remarkably successful description of the fundamental constituents of matter and the forces that govern their interactions. Despite its predictive power, the SM is widely understood to be an incomplete theory. It does not account for observed matter-antimatter asymmetry of the universe, provides no dark matter candidate, fails to incorporate gravity at the quantum level, and leaves the mass hierarchy of fermions unexplained.

A variety of Beyond Standard Model (BSM) theories address these shortcomings by hypothesizing the existence of new heavy gauge bosons that come from extensions of the Standard Model's electroweak symmetry, the mathematical structure that organizes how particles interact with the weak nuclear force. These hypothetical particles are called Z' bosons (Z-prime bosons). A Z' boson is generally understood to be a heavier counterpart of the Z boson, a known particle that mediates the weak nuclear force and was discovered at CERN in 1983. Many BSM extensions predict that the Z' would couple to leptons and quarks, though the coupling strength varies between models. In the Sequential Standard Model benchmark, the couplings are thought to be identical to the Z, while other models predict substantially different couplings ~~just like the Z but at a higher mass scale~~ (Langacker, 2009).

Previous searches at the LHC have set lower mass limits on Z' bosons in the multi-TeV range for the most common benchmark models. Both ATLAS and CMS have performed high-mass dilepton searches using the full Run 2 dataset of 139 fb-1 and found no evidence of a Z' (ATLAS Collaboration, 2019; CMS Collaboration, 2021). An undiscovered resonance remains possible, particularly at masses or coupling strengths below the sensitivity of these previous analyses.

Our search utilizes the dilepton channel because of its simplicity and clarity. As a hypothetical "cousin" of the Z boson, the Z' should decay into a pair formed by a lepton and its corresponding anti-lepton (e.g. an electron and positron). Utilizing this channel, often called the "golden channel," allows for a clean experimental signature and a well-understood background from the Drell-Yan process (Drell & Yan, 1970). If a Z' exists, it would appear as a localized excess of events, or a "bump", in the dilepton invariant mass spectrum at the Z' mass, sitting on top of the smooth Drell-Yan continuum. This "bump-hunt" strategy is the standard approach to discovering new resonances at colliders.

Machine learning is now a standard tool in particle physics for classifying collision events (Feickert & Nachman, 2021; Radovic et al., 2018). However, applying machine learning to bump-hunt searches comes with a concern: if a classifier learns to reconstruct invariant mass (even indirectly through correlated features), it can distort the smooth background spectrum and create artificial peaks where no real signal exists. This problem, known as background sculpting (Dolen et al., 2016), is the primary failure this analysis is designed to avoid.

The goal of this work is to develop and validate a fully reproducible machine learning ~~pipeline~~ workflow for a ~~Z' boson~~ search in the dilepton channel using ATLAS Open Data~~LHC data~~, following the methodological structure of real collider analyses. We apply this workflow to a Z' boson search as a concrete test case and find no statistically significant localized excess across the probed mass range.~~, and to demonstrate a statistically valid null result or identify candidate excesses.~~

2. Experimental Background

2.1. The LHC

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator (Evans & Bryant, 2008). It accelerates two beams of protons in opposite directions to nearly the speed of light, then brings them into head-on collision at designated interaction points. During Run 2 (2015-2018), the LHC operated at a center-of-mass energy of $\sqrt{s}$ = 13 TeV, meaning the total energy available in each proton-proton collision is 13 trillion electron-volts. This analysis uses data collected in 2015 during the first year of Run 2.

2.2. ATLAS Detector

The ATLAS (A Toroidal LHC Apparatus) detector is a large general-purpose particle detector at the LHC (ATLAS Collaboration, 2008). It is designed to record the debris from high-energy proton-proton collisions and reconstruct what happened in each event using software and sensors. The detector is built in concentric layers around the collision point, with each layer optimized to measure different types of particles:

- Inner Detector: The innermost layer immersed in a strong magnetic field. Charged particles curve as they pass through the field, and the curvature is used to measure their momentum. This is the primary system for tracking leptons such as electrons and muons.
- Calorimeters: These layers, surrounding the inner detector, absorb particles and measure their energy. The electromagnetic calorimeter measures electrons and photons, while the hadronic calorimeter measures jets (sprays of particles produced by quarks and gluons).
- Muon spectrometer: Muons pass through calorimeters -relatively undected, depositing ~~unaffected~~ only a small amount of energy through ionization. They are too heavy to radiate the electromagnetic showers that electrons do, and because they are leptons that don't participate in the strong force, muons also do not produce hadronic showers like protons and neutrons. ~~(because they are heavy leptons, their momentum is too high for calorimeters to stop them),~~ So,~~so~~ this dedicated system in the outermost layer identifies and measures them with high precision using a toroidal magnetic field.

2.3. Dataset

This analysis uses Run 2 2015 proton-proton collision data made publicly available through the CERN Open Data Portal in 2024 (ATLAS collaboration, 2024). The data is provided in DAOD-PHYSLITE format, an analysis-ready format in which the ATLAS experiment collaboration has already performed computationally heavy work like identifying particles. This full dataset contains 10,049 files with a total 9.3 TiB of approximately 1.69 x 10^9 events and has a luminosity of 3.2 fb^{-1}.

3. Data Preparation

3.1. Event Selection

For event selection, we do not apply any additional trigger requirements, as the PHYSLITE Open Data is already trigger-filtered. We select events containing exactly two same-flavour leptons: dielectron events (with exactly two electrons) and dimuon events (with exactly two muons). While lepton flavor is recorded, it is not used as an input for our classifiers. We do not impose an opposite-sign charge requirement, and this limitation is discussed in Section 6.2. Limitations and Future Improvements. For each dilepton event, the following reconstructed objects are extracted:

- Lepton Kinematics: Using analysis-level reconstructed lepton objects without further ID/isolation cuts (AnalysisElectronsAuxDyn and AnalysisMuonsAuxDyn), we collect the transverse momentum pT (the component of momentum perpendicular to the beam), pseudorapidity η, and azimuthal angle φ of each lepton ~~is gathered~~.
- Jet Kinematics: The three same kinematic variables are collected for the leading reconstructed jet (AnalysisJetsAuxDyn), plus the jet multiplicity n_jets (how many jets were found in the event).
- Missing Transverse Energy (MET): The apparent imbalance in the total transverse momentum of the event, the MET can signal potential undetected particles like neutrinos. It is reconstructed from the x and y components of the missing momentum vector (MET_Core_AnalysisMETAuxDyn.mpx/mpy).

All momenta are provided in MeV by ATLAS and are converted to GeV for this analysis. We do not apply any pT-based cuts while selecting events and processing data, although quality cuts are applied prior to training the model, as discussed in Section 3.3.

3.2. Invariant Mass

The dilepton invariant mass $m_{\ell\ell}$ is an important value for this search. If two leptons originate from a parent particle decay (like a Z' boson decaying into two leptons), then the dilepton invariant mass should equal the mass of the parent particle. Thus, if a Z' boson exists, it would produce a peak of events at a particular value in the invariant mass spectrum corresponding to its mass. This is what physics "bump hunts" typically look for. Dilepton Invariant mass is calculated using the three kinematic variables that we collected for each lepton (pT, η, φ):

$$m_{ll} = \sqrt{2pT^1 pT^2 [(\eta^1 - \eta^2) - (\phi^1 - \phi^2)]}$$

~~$m_{\ell\ell} = \sqrt{(2\ pT^1\ pT^2\ [\cosh(\eta^1 - \eta^2) - \cos(\varphi^1 - \varphi^2)])}$~~¶

3.3. Quality Cuts

We apply quality cuts that are standard to the LHC. A pseudorapidity cut $|\eta| \leq 2.5$ is applied to electrons and muons, reflecting the acceptance of the ATLAS inner detector and calorimeter. For muons, this cut reflects limitations of the ATLAS Inner Detector and its Semiconductor Tracker (SCT) ~~of the~~, which work with the ATLAS Muon Spectrometer ~~and the Inner Detector, which are used~~ to measure muon momentum (ATLAS Collaboration, 2016). For electrons, the ATLAS electromagnetic calorimeter precision region is $|\eta| \leq 2.47$, which makes this analogous cut slightly conservative but still valid. However, standard ATLAS analyses also exclude the barrel-endcap transition region ($1.37 < |\eta| < 1.52$), where dead material degrades electron energy resolution. We do not apply this exclusion in this simplified Open Data analysis, which may slightly broaden the reconstructed electron resolution (ATLAS Collaboration, 2019a). Jet variables are undefined for events where n_jets = 0, so the pseudorapidity cut is conditionally applied to jet objects when n_jets > 0.

A hard cut of $m_{\ell\ell} \leq 3{,}500$ GeV is applied because the background spectrum falls very steeply, and so past this boundary, there are hardly any events in the 20 GeV bins that our fit uses. We account for this problem with a minimum of 10 events for a bin to be used during the significance scan ~~,~~ which restricts the scan to $m_{\ell\ell}$ = 250-1450 GeV.~~but this hard cut prevents us from even looking where we expect zero events.~~

4. Machine Learning Methodology

The machine learning methodology has three stages. First, we train two classifiers, a boosted decision tree (BDT) and a deep neural network (DNN), to distinguish Z-peak events from low-mass Drell-Yan background events using only kinematic features, with the dilepton invariant mass deliberately excluded so that the classifiers cannot learn the search variable. Both are trained and validated on events drawn exclusively from the low-mass sidebands and the Z-peak region, well below the search region, so that the search region itself remains blinded during model development. These Z-peak events are not meant to stand in for a Z' signal, but are used as a Z-like proxy-signal and real resonances that a Z' would be kinematically similar to.

Second, we apply the trained classifiers to the high-mass search region ($m_{\ell\ell}$ > 250 GeV) and sort every event into categories by classifier score, concentrating the most Z-like events into a single high-scoring category. Within that category, we model the smooth Drell-Yan background with a three-parameter parametric fit:

$$f(m) = p_0\left(\frac{m}{\sqrt{s}}\right)^{p_1} e^{(-p_2 * \frac{m}{\sqrt{s}})}$$

~~f(m) = p_0 • (m/√s)^ p_1 • exp(−p_2 • m/√s)¶~~

This function uses a power law combined with an exponential decay and is a standard parametrization used in ATLAS analyses (ATLAS Collaboration, 2019b). The three parameters (p_0, p_1, p_2) are determined by fitting the function to the observed mass spectrum in 20 GeV bins using χ^2 minimization, which is a statistical procedure that adjusts the parameters to minimize the sum of squared residuals between the fit and the data, weighted by the uncertainty in each bin.

We evaluate the accuracy of the fit with reduced chi-squared statistic, χ^2/ndf, which is a sum over all bins of squared deviations. χ^2 is evaluated as $\chi^2 = \Sigma_i (n_i - b_i)^2 / \sigma_i^2$, where n_i is the number of observed events in a bin, b_i is the number of events predicted to be in the bin by the background fit, and σ_i is a variance calculation. Ndf (number of degrees of freedom) is calculated by subtracting the number of parameters in the fit (three in our case) from the number of bins we search. This background spectrum should remain smooth and falling, and a good fit will confirm that the classifiers have not sculpted artificial structure into it.

Finally, we perform a bin-by-bin Asimov significance scan within the high-score category, comparing the observed number of events in each bin against the background fit to test for localized excesses. The remainder of this section describes the input features, the two classifier architectures, and their training process.

4.1. Features

The core idea of this methodology is to train machine-learning classifiers to distinguish Z-like events from non-Z-like events using only kinematic features. Specifically, the momenta, angles, jet activity, and missing transverse energy of each event. While we record lepton flavor during event selection, it is not used as an input for the classifiers, so electron events

and muon events are treated the same despite their different kinematic behaviors. This limitation is acknowledged in Section 6.2. The trained classifiers are then applied to the search region to identify which high-mass events look most like a Z boson and then concentrate these potential Z' signals into a high-score category.

One critical design choice for the input features is that dilepton invariant mass is excluded, as including it would allow the classifiers to directly learn the search variable and potentially sculpt the background to create artificial "bumps" in the mass spectrum where no real signal exists. Of course, kinematic inputs such as pT, η, and φ are used to calculate invariant mass, so classifiers may still indirectly learn mass-correlated information. However, excluding the invariant mass is a step taken to mitigate this mass correlation.

The same 11 input features are used for both classifiers. They are listed with their descriptions in Table 1.

Table 1: Input kinematic features for the models and their descriptions

Feature	**Meaning**
pT^1, pT^2	Transverse momenta of the leading and subleading leptons (electron or muon)
η^1, η^2	Pseudorapidities of both leptons
φ^1, φ^2	Azimuthal angles of both leptons
n_jets	Reconstructed jet multiplicity
leading_jet_pT	Transverse momentum of the leading jet (zero if n_jets = 0)
leading_jet_η	Pseudorapidity of the leading jet (zero if n_jets = 0)
leading_jet_φ	Azimuthal Angle of the leading jet (zero if n_jets = 0)
MET	Missing transverse energy

Both classifiers are trained as binary classifiers, with Z-peak -events labeled as proxy-signal (class 1) and low-mass Drell-Yan events labeled as background (class 0) (Collins et al., 2019; Metodiev et al., 2017). The labeled sample of training events is split 80/20 into training and test sets using stratified sampling (preserving class ratio in both splits), giving 8,950,298 training events and 2,237,575 held-out test events for validation.

4.2. Boosted Decision Tree (BDT)

The BDT, an ensemble of gradient-boosted decision trees that has become a standard classifier in particle physics (Friedman, 2001; Roe et al., 2005), uses binary cross-entropy (BCE), a loss function for binary classification where loss increases exponentially when the difference between the model's prediction and the event's actual label increases. For an event with true label y (either 0 or 1) and predicted probability p, the BCE is:

$$L = \; -[y * (p) \; + (1 - y) * (1 - p)]$$

~~L = -[y • log(p) + (1 − y) • log(1 − p)]¶~~

Loss is small when the model's prediction matches the correct label of the event and large when the model is incorrect with high confidence. Each successive tree is fit to the gradient of this loss so that the ensemble progressively reduces its overall error. This loss calculation is done automatically using XGBoost, a scalable open-source gradient-boosting library implemented in Python through its xgboost package, which handles tree construction, gradient computation, and parallelized training internally (Chen & Guestrin, 2016). All preprocessing of input features, train/test splitting, and evaluation are carried out in Python using NumPy, pandas, and scikit-learn.

To prevent the model from overfitting, we apply several forms of regularization using XGBoost parameters.

- Number of trees and learning rate: We train 300 trees with a low learning rate of 0.03. The learning rate scaled down each tree's contribution, so that each tree makes a small correction rather than aggressively fitting the training data. A low rate combined with many trees sacrifices training speed for better generalization.
- Maximum depth of 5: Each tree is limited to 5 levels deep, which produces at most 32 distinct regions. Unlike deep trees, which tend to memorize individual events, shallow trees learn broad patterns.
- Minimum child weight of 10: In each tree, a leaf node can only be split if both children would contain at least 10 units of statistical weight. This prevents the model from learning patterns from small groups of unusual events.
- Minimum split gain of 1.0: Splits only happen if they reduce the loss by at least 1.0. This "prunes" each tree by removing redundant branches, which helps reduce overfitting.
- L1 regularization~~/Lasso regression~~ (reg_alpha = 0.1): This essentially adds a penalty proportional to the absolute value of each leaf's ~~prediction~~output weight. It pushes small leaf weights toward zero, which can prune branches from the trees and simplify the model. Unlike Lasso regularization in linear models, this does not zero out an input feature's global importance. It regularizes the leaf values rather than performing input-feature selection.~~ This drives the weight of less important features to zero and simplifies the model by essentially performing a feature selection.~~
- L2 regularization (reg_lambda = 5.0): Adds a penalty proportional to the square of a leaf's prediction. It shrinks large weights toward zero and strongly discourages overconfident predictions.
- Stochastic gradient boosting (70%): Each tree is trained on a random 70% of events using a random 70% of features. This randomness prevents all trees from making the same mistakes.

The class imbalance between labeled signal events (5,811,472) and labeled background events (5,376,401) is handled by a scale_pos_weight parameter of 0.93 (calculated ratio), which decreases the weight of signal class during training.

4.3. Deep Neural Network

Neural networks (LeCun et al., 2015), and in particular deep architectures applied to collider classification problems (Baldi et al., 2014), are built from computational units called neurons. Each neuron takes in a set of numbers and computes a weighted sum based on weights that are learned during training. The result is passed through a nonlinear activation function. Our model uses the ReLU (Rectified Linear Unit) activation function, which outputs the input if it is positive and

zero otherwise. ReLU and the DNN architecture is implemented with PyTorch, an open-source deep-learning framework in Python that provides automatic differentiation, GPU-accelerated tensor operations, and prebuilt neural-network modules. The network is defined as a torch.nn.Module subclass, with linear layers, batch normalization, dropout, and ReLU activations composed using the torch.nn API (Nair & Hinton, 2010; Paszke et al., 2019).

Neurons are arranged in multiple layers. In a fully connected layer, every neuron receives input from every neuron in the previous layer. Unlike graph-based architectures designed to find structural relations in inputs, this multi-layered perceptron (MLP) does not explicitly encode invariant relationships like mass between particles. Despite this, correlated kinematic information may still allow partial reconstruction of invariant mass, motivating our validation against background sculpting (which is discussed in the background modelling section). We initially tried to use a Graph Neural Network (GNN), but it was able to reconstruct invariant mass and scored events almost solely based on mass. The switch to this multi-layered perceptron ~~eliminates~~ mitigates this background sculpting problem ~~which we were trying to avoid~~(later confirmed in section 5.2), though it does not fully rule out indirect mass learning.

Our network consists of an input standardization layers, three fully connected hidden layers of 128 neurons each, and a single output neuron.

- Standardization layer: Before entering the network, each of the 11 input features is transformed to have zero mean and unit standard deviation by subtracting the mean and dividing by the standard deviation calculated only for training sets. Since the raw features have different scales (pT can be hundreds of GeV, while ~~ϕ~~ is bounded between -π and π), this ensures that features with larger magnitudes do not distort the learning process with large weights.
- Hidden layers: three layers of 128 neurons each. Each layer applies a linear transformation, batch normalization (which recenters and rescales the activations across each training batch to stabilize learning) (Ioffe & Szegedy, 2015), and ReLU activation. Each step is implemented using PyTorch. The depth and width of the network give it the capacity to learn complex nonlinear combinations of input features.
- Dropout (p=0.3): After each hidden layer, 30% of neurons are randomly "switched off" during each training step. This prevents individual neurons from becoming dependent on one another and improves their individual reliability. Dropout prevents the model as a whole from overfitting and memorizing training noise (Srivastava et al., 2014). When testing, however, all neurons are active.
- Output: In the last layer, a single neuron produces an unbounded score, which is passed through a sigmoid function to predict a probability between 0 (background) and 1 (signal).

Training occurs in epochs, where one epoch is a full pass through the training data. For each epoch, the data is divided into batches of 4,096 events. For each batch, the network makes predictions, the loss is computed, and the gradient of the loss with respect to every weight is computed using chain rule (backpropagation). The weights are then updated according to this gradient. We do clip gradients at 1.0 so that large updates to weights do not cause instability. Over many epochs, the DNN converges toward a set of weights that minimizes the loss.

We optimize our DNN with AdamW (Loshchilov & Hutter, 2019), a variant of the Adam optimizer that decouples weight decay from the gradient update. We use a learning rate of 10^{-3} and weight decay of 10^{-4} (an L2 like penalty on all weights that discourages overfitting). AdamW is provided directly by the torch.optim module, and gradients are computed via

PyTorch's automatic differentiation engine during backpropagation. We also use two additional techniques which are also implemented using the PyTorch library:

- Cosine annealing with warm restarts (Loshchilov & Hutter, 2017): The learning rate decreases smoothly following a cosine curve over the course of 10 epochs, then resets back to its initial value and begins decaying again. These periodic resets allow the optimizer to escape shallow local minima that it might otherwise get trapped in, while the cosine decay within each cycle provides a fine-grained optimization.
- Early stopping (patience = 10): After each epoch, the model is evaluated on the held-out test set. If the test-set performance, measured by AUC, does not improve for 10 consecutive epochs, training stops and the best-performing model is used. This prevents the network from training long enough to memorize training data. Our final model stopped after 29 epochs.

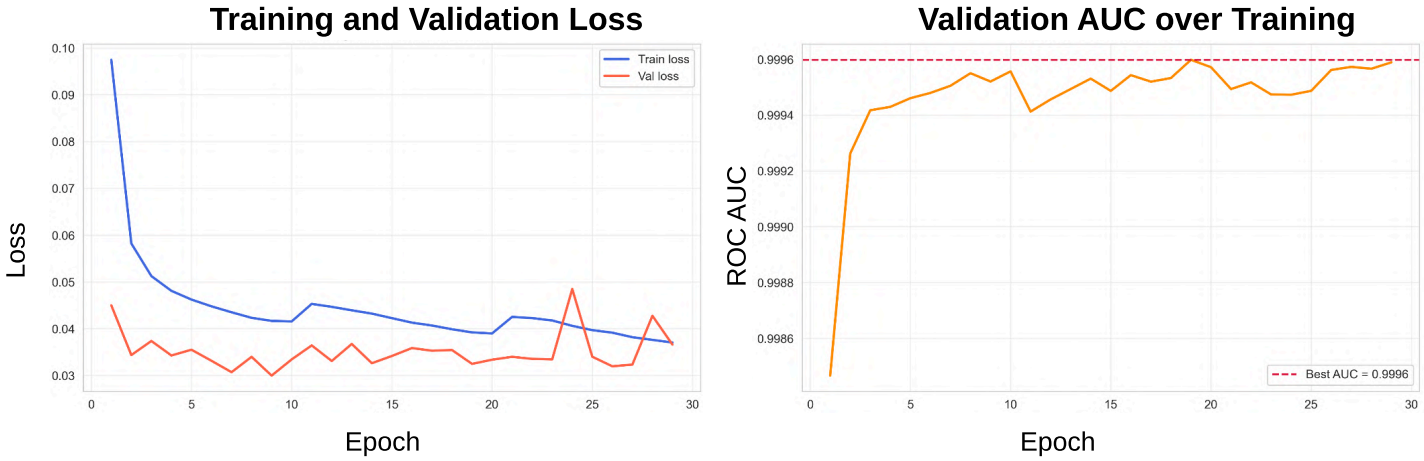


Figure 1: DNN training and validation loss (left) and validation AUC (right) per epoch. Training stopped after 29 epochs because of the early stopping implementation.

4.4. Mass Regions

During training and analysis, we divide the dilepton invariant mass spectrum into four non-overlapping Regions.

The first region used for training, the background region, is set between 40 and 71.2 GeV. The lower bound was chosen to get a clean Drell-Yan dominated region without resonance peaks from the J/ψ at 3.1 GeV and the ϒ (upsilon) at 9.5 GeV. The upper bound was chosen to avoid the Z peak at 91.2 GeV. While the Z boson has a natural width of 2.5 GeV, the reconstructed peak is broader because of detector smearing by the calorimeters and muon spectrometer, so we use a larger buffer ~~to avoid any Z contamination during analysis~~that accounts for both the natural width and detector resolution effects and avoids Z contamination. These bounds also narrow down our background events so that the ratio of background to signal is much closer to 1. These low-mass Drell-Yan sideband events serve as a background proxy rather than a signal

proxy, and dDuring training and analysis, we expect our models to associate these events with label=0 (not ~~Z-like~~alike to a Z resonance).

The signal (or proxy signal) region, also used for training, is centered at the Z boson mass of 91.2 GeV and extends to ±20 GeV threshold (71.2-111.2 GeV). ~~Since we do not use simulated Z' events, this region provides our only Z-like resonance for our search.~~ As previously stated, this width is wide enough so that we do not rely on nearly perfect reconstructions and narrow enough so that events from the Drell- Yan tail do not contaminate our signal classification. The lower bound matches the background region upper bound so that there is no overlap between regions but also so that all events with mass greater than 40 GeV are used. Since we do not use simulated Z' events, this region provides our only Z-like resonance for our search. To be clear, these Z-peak events are not real Z' signal events, and are instead used as a Z-like proxy signal, or a real resonance that stands in for a hypothetical Z'. This proxy carries an implicit assumption that a multi-TeV Z' is kinematically similar to 91 GeV Z boson. This assumption is imperfect, since the available phase space, the parton distribution functions governing the initial-state quark momenta, and the resulting angular distributions of the decay leptons differ between the Z pole and the multi-TeV scale, so a classifier optimized on Z pole kinematics may not be optimal for a high-mass resonance. Despite this reduced signal efficiency, avoiding simulation removes dependence on any particular Z' model.

The validation region from 111.2 GeV to 250 GeV is not used for training or search. Its purpose is to verify that the model generalizes reasonably before looking at the actual search region. The lower bound of 111.2 GeV is just above the signal region. At this mass, events are still from the high-mass tail of the Z, so they look kinematically alike to Z decays except for their higher invariant mass caused by detector effects and final state radiation. Since the model has not seen these events during training, how it scores them tell us how the learned features generalize. The upper bound of 250 GeV was determined empirically. The goal of this upper bound is to ensure that the Z tail sufficiently dies out so that a smoother power law can fit the remaining mass spectrum. Using a χ^2/ndf fit, we found that below a bound of 250 GeV, residual Z contributions create an excess that a power law cannot fit (χ^2/ndf = 3.20 at 200 GeV, 0.99 at 250 GeV). This bound is consistent with other ATLAS analyses which use 225-250 GeV as a lower bound for a high-mass search region. So, this validation region acts as a transition zone between the signal region and a smooth background. Events in this region are separated from search events specifically because they are not clean enough for a reliable background fit.

The search region, everything above 250 GeV, is where we look for a Z'. In this region, the Drell-Yan background should be smooth and falling, so that a Z' would appear as a localized narrow bump of observed events against a background fit. ~~While there is not a specific upper bound~~Although a cut is applied to events with $m_{\ell\ell}$ > 3,500 GeV, ~~, the invariant masses of our events do not exceed 3,000 GeV, which means there simply aren't enough events to see anything meaningful above this mass.~~our significance scan and statistical interpretation are restricted to 250-1450 GeV, the range over which each 20 GeV bin retains 10 events, which is what we consider sufficient for a reliable fit.

After training on the labeled background and signal regions, the performance of the models is evaluated with the validation region before being applied to score events in the blinded search region.

After event selection and quality cuts, there are 5,376,401 events in the background region; 5,811,472 events in the signal region; 359,621 events in the validation region; and 19,042 events in the search region.

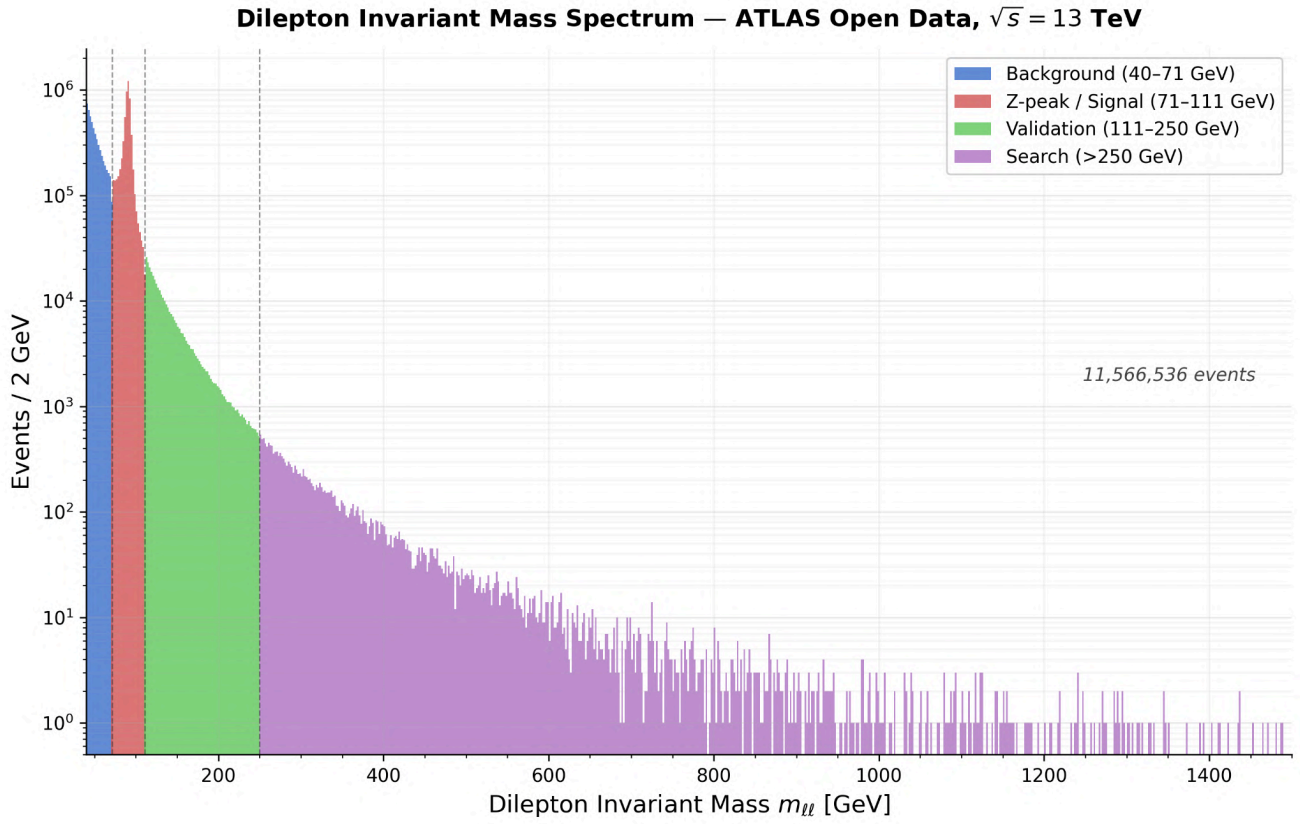


Figure 2: Dilepton invariant mass spectrum across a partial mass range, shown on a logarithmic scale. Events are color-coded by analysis region. The Z peak is at 91.2 GeV with the spectrum falling steeply at higher masses.

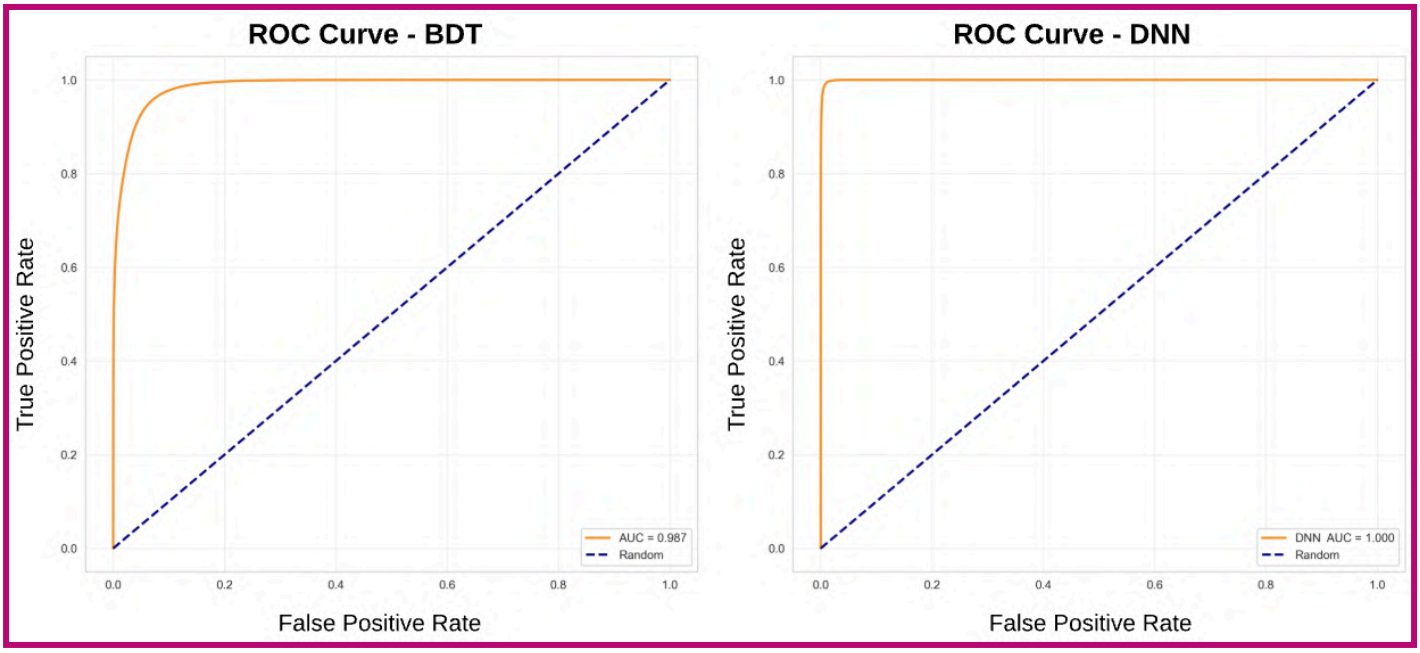
4.5. ROC Curve

The first metric we use to evaluate our classifiers after training is the Receiver Operating Characteristic Area Under the Curve (ROC AUC) (Fawcett, 2006). The ROC curve plots the true positive rate (the fraction of signal events correctly identified by the classifier) against the false positive rate (the fraction of background events incorrectly labeled as signal by the classifier) as the classification threshold (the minimum score needed for an event to be classified as a signal) varies. The AUC is the area under the ROC curve and it ranges from 0.5 (essentially random guessing) to 1.0 (perfect accuracy). An AUC of 0.98, for example, means that a random signal event scores higher than a random background event 98% of the time.

The BDT achieved a test-set AUC of 0.9869, and the DNN achieved a 0.9996. These values are based on the classifiers' performance on a 20% hold-out set of the total training data, which included everything in the background and signal region. Both these values are high, which reflects the genuine kinematic differences between Z events and low-mass Drell-Yan sidebands and demonstrates that the classifiers very accurately separate these two groups. This is important, as our methodology relies on using the classifiers to find high-mass events reflecting Z-peak patterns that look very different

from the background spectrum. The reason for this large difference in characteristic is because the Z boson produces leptons with higher transverse momenta than sideband events, and this is a dominant feature driving the classification. This is confirmed by the BDT's feature importance analysis: the subleading lepton's pT accounts for 47% of the model's discriminating power, the leading lepton's pT accounts for 24%, and the other features contribute much less.

It is important to note that these AUC values reflect the classifiers' ability to distinguish the two training classes (Z-peak vs. low-mass Drell-Yan) and do not directly measure sensitivity to a Z' signal in the search region. That sensitivity is established separately through a background-fit quality and the significance scan described later.


ROC Curve - BDT
True Positive Rate
False Positive Rate
AUC = 0.987
Random
ROC Curve - DNN
True Positive Rate
False Positive Rate
DNN AUC = 1.000
Random

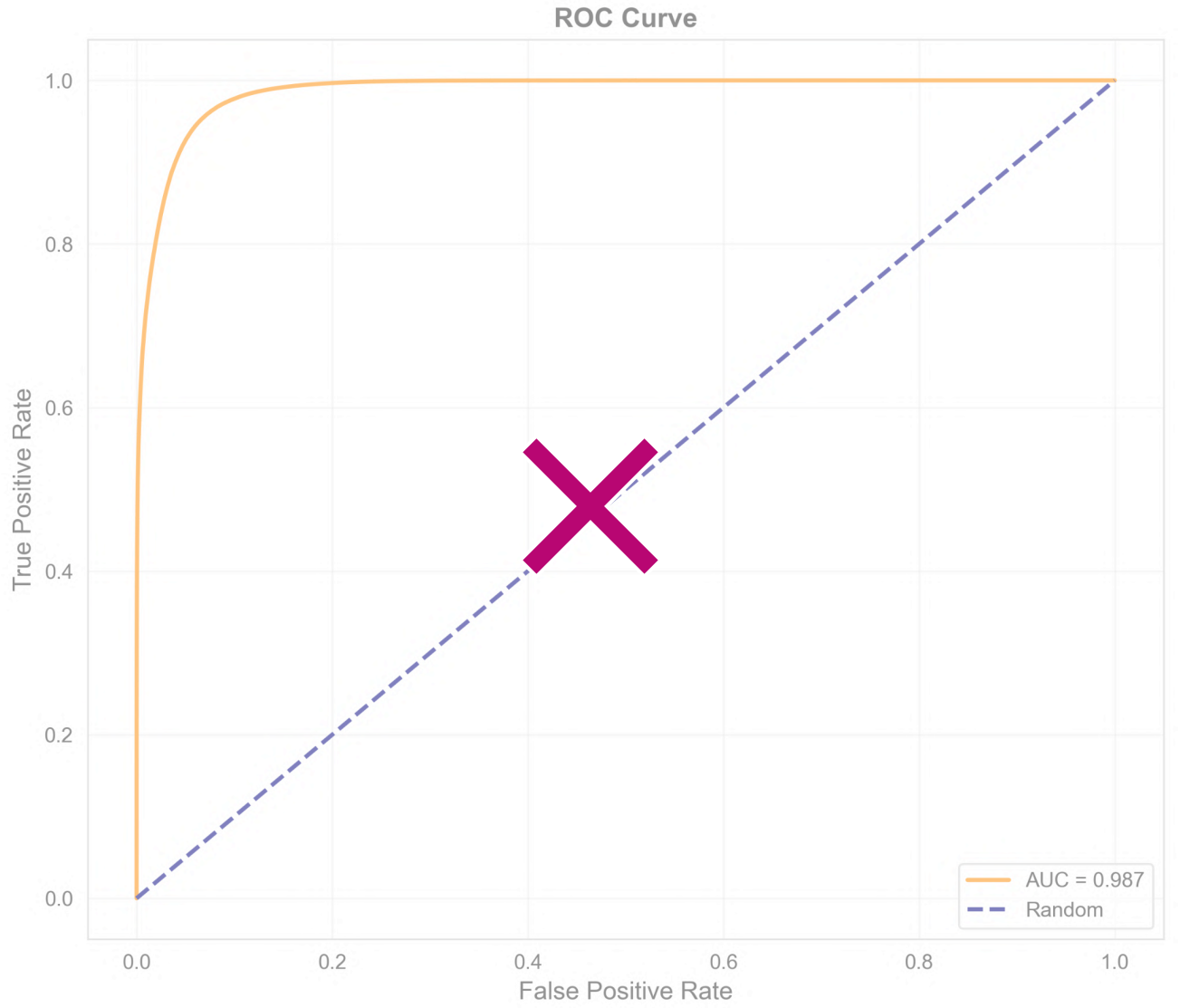


Figure 3: ROC curve on the held-out test set for the BDT (left) and DNN (right) classifiers.

4.6. Overtraining Check

We also used performance checks for overtraining, which occurs when a model memorizes training data rather than learning general patterns, leading to good performance on training data but bad performance on unseen data. First, we verified that the model's performance on the test data mirrors its performance on training data. Then, we confirmed that score distributions evaluated on the test sets were consistent with training sets, meaning the model scores events similarly

regardless of whether it has or has not seen them before.

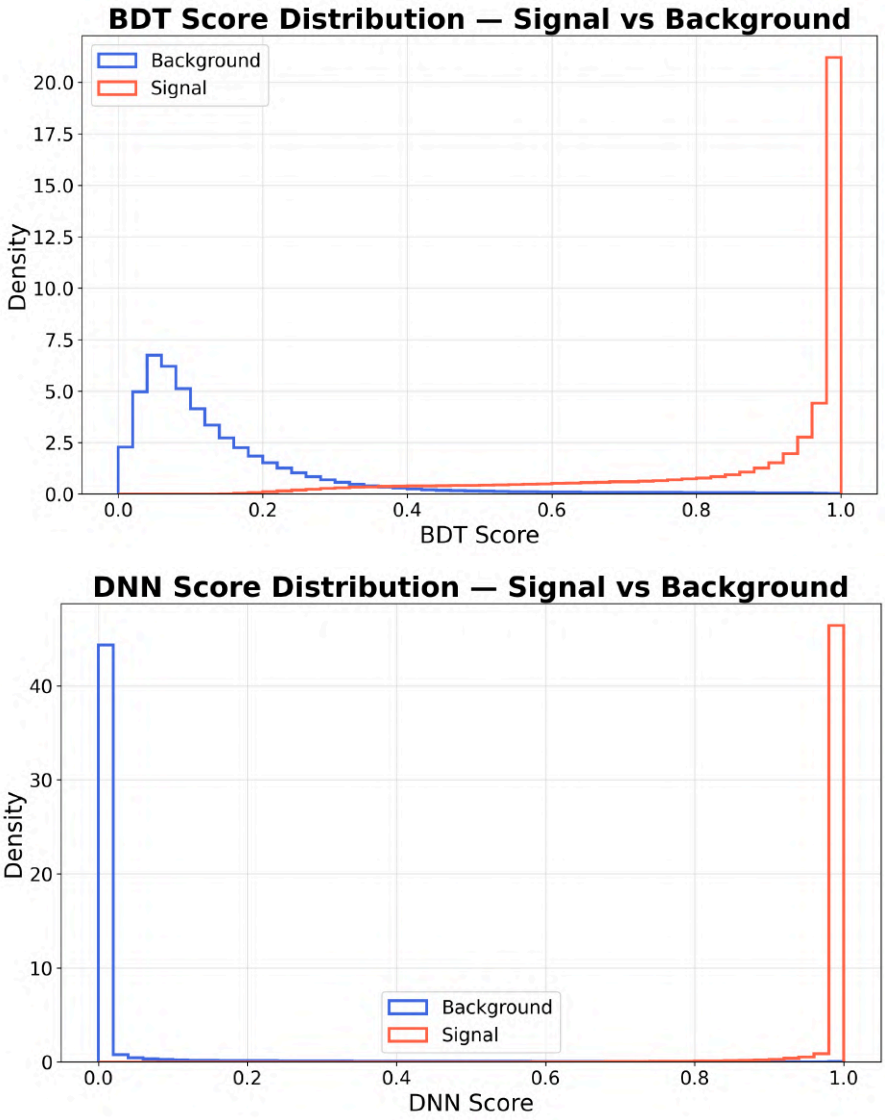


Figure 4: Overtraining check: classifier score distributions on the training and test sets for both the ~~DNN (right) and BDT (left)~~BDT (top) and DNN (bottom). Train and test distributions agree, indicating no overtraining.

4.7. Validation

Before unblinding the search region, we apply both classifiers to the 359,621 events in the validation region (111.2-250 GeV), which were not used during training. Because these events are primarily from the high-mass tail of the Z boson signal at 91.2 GeV, they look kinematically similar to Z-peak events but at higher invariant masses. The classifier's behavior in this

region acts as an honest test of whether it has learned genuine Z-like features rather than simply memorizing the training mass window. In the validation region, the BDT assigns a mean score of 0.872 and the DNN assigns a mean score of 0.998, indicating that both classifiers recognize these events as strongly Z-like. Z decays are characterized by high transverse momentum, which explains why validation-region events score high (especially for the DNN): a heavier parent imparts more energy to its decay products, so validation-region events have higher lepton pT than even some Z-peak events. The score distributions transition smoothly from the signal region into the validation region with no discontinuity at the 111.2 GeV boundary, which gives us confidence that the classifiers' behavior in the search region (>250 GeV) is physically reasonable.

This behaviour also clarifies what the classifiers can and cannot do in the blinded search region. Because they were trained to separate a 91 GeV resonance from a 40 GeV continuum, they rank events largely by pT, and in the search region, like the validation region, both the Drell-Yan background and a hypothetical Z' produce high-pT leptons. The classifiers therefore assign high scores to most validation and search region events and do not, at least efficiently, distinguish high-mass background from high-mass signal. This is expected, as a high-mass Drell-Yan event is genuinely Z-like in its kinematics, but the classifier's role in this workflow is not to identify Z' candidates but to concentrate events with Z-like topology into a category with a smooth $m_{\ell\ell}$ spectrum tht is well described by a simple parametrization. The discrimination between signal and background is performed by the fit to the mass spectrum, not by initial classifier score.

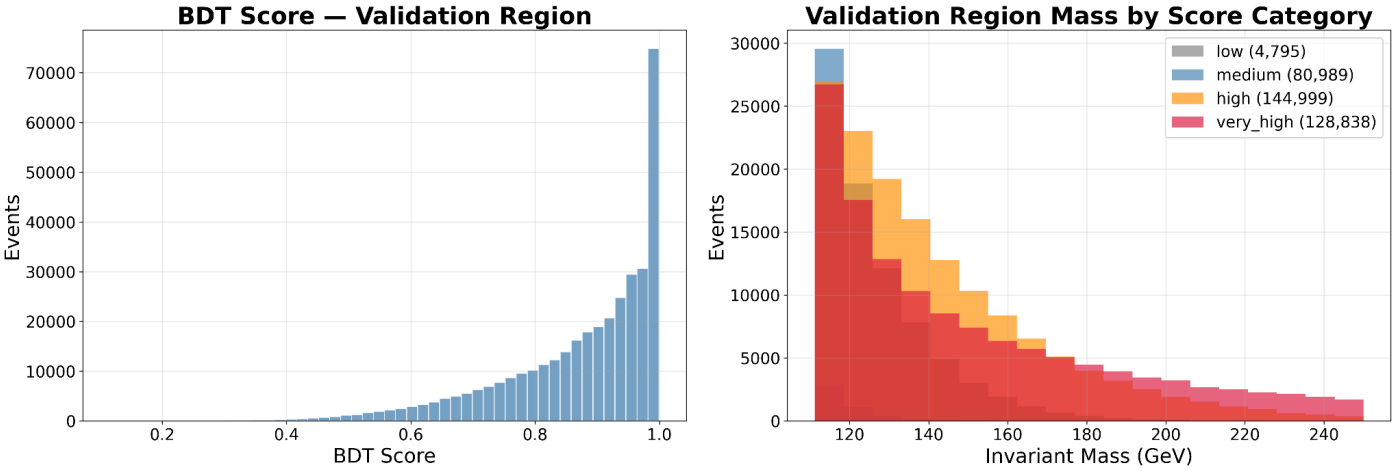


Figure 5: BDT classifier scores in the validation region (111.2-250 GeV): Distribution of BDT scores (left) and invariant mass distribution broken down by score category (right), showing a smooth transition into the search region.

5. Search and Results

5.1. Score Categories

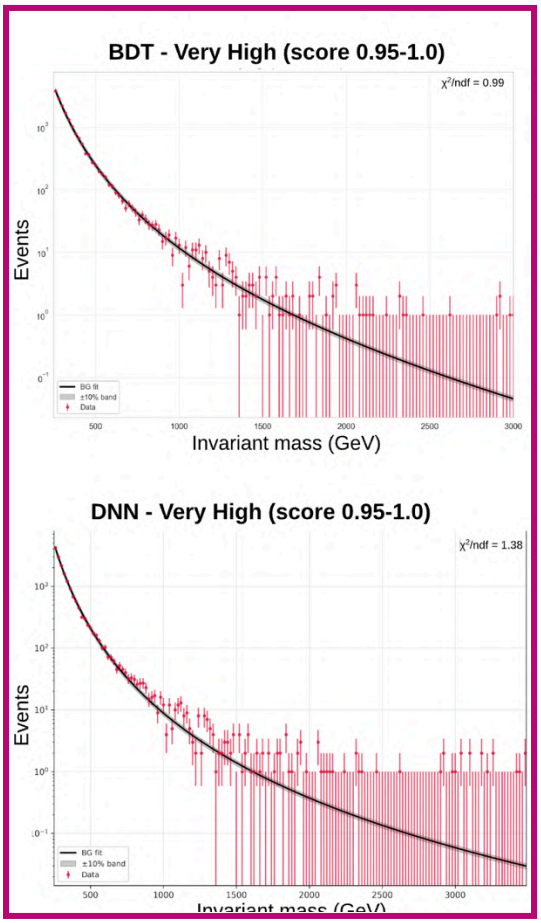
After training and validating the classifiers, we apply them to the blinded search region starting at $m_{\ell\ell}$ = 250 GeV. Each event is scored between 0 and 1, indicating how Z-like the classifier judges it to be. Scored events are then sorted into four categories:

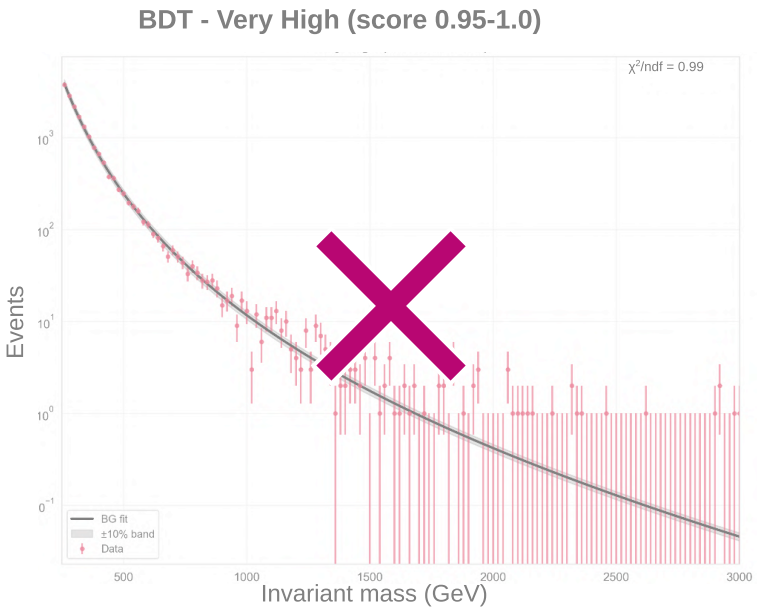
- Low: score < 0.5
- Medium: 0.5 ≤ score < 0.8
- High: 0.8 ≤ score < 0.95
- Very High: score ≥ 0.95

The "very high" category includes events that the classifier considers most similar to a Z boson decay. If a Z' exists in this data and decays into lepton pairs like the Z, its events would concentrate in this category and stand out against the reduced background. This category is the primary channel for our significance scan.

5.2. Background Modeling

As explained in Section 4, we perform a three-parameter fit on the events in the "very high" category and evaluated the accuracy of the fit with a χ^2/ndf value. A χ^2/ndf close to 1 indicates the fit's acceptable agreement within statistical fluctuations. A value much greater than 1 would indicate that the fit does not accurately describe the data, probably because the classifier has distorted the mass spectrum and has started background sculpting. Our fits yield χ^2/ndf = 0.99 for the BDT and χ^2/ndf = 1.3811 for the DNN in the very-high category, which confirms that neither classifier sculpts the background or creates false signatures. The likely reason that the DNN's value is further from 1 is because it flags much more events as high-scoring, which makes it harder for it to create a smooth fit with only three parameters.


BDT - Very High (score 0.95-1.0)
χ²/ndf = 0.99
Events
Invariant mass (GeV)
BG fit
±10% band
Data
DNN - Very High (score 0.95-1.0)
χ²/ndf = 1.38
Events
Invariant mass (GeV)


BDT - Very High (score 0.95-1.0)
χ²/ndf = 0.99
Events
Invariant mass (GeV)
BG fit
±10% band
Data

370

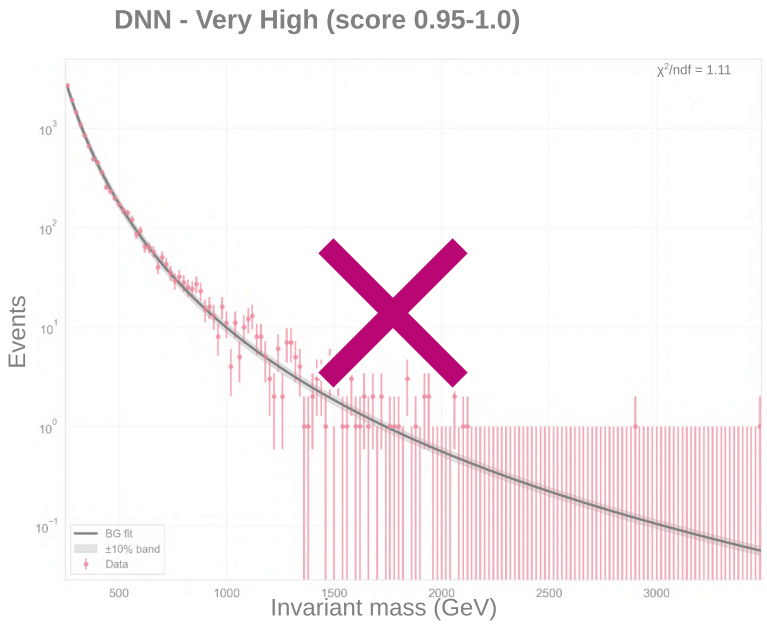


Figure 6: Three-parameter background fits in the very-high score category (score ≥ 0.95) for the BDT (top) and DNN (bottom), shown overlaid on the observed mass spectrum in 20 GeV bins. The fits yield $\chi^2/\mathrm{ndf} = 0.99$ and 1.38~~11~~ respectively.

Because a χ^2 close to 1 is a global measure, we supplement it with an examination of the score-mass relationship, where we expect background sculpting to appear. The left part of Figure 7 profiles the mean DNN scores against $m_{\ell\ell}$ across the search region. The profile is essentially flat, drifting slightly downward with mass rather than rising. The right part shows the fraction of events in each score category as a function of mass, which remains smooth with no category spiking at any specific mass. We quantify the relationship with the Pearson and Spearman correlation coefficients between DNN score and $m_{\ell\ell}$, which are -0.12 and -0.23 over the full search region and -0.15 and -0.20 within the fitted very-high category. Both are small in magnitude and negative, which means the relationship between classifier score and mass is a weak, slightly negative linear trend, so the classifier is not artificially concentrating events at a particular mass to fake a peak. A Kolmogorov-Smirnov comparison of the mass shapes in the high and very-high categories gives a value of 0.24, indicating that the categories differ smoothly and don't have any drastic divergent shapes, which means that there is no localized bump within either category. Together with the $\chi^2/\mathrm{ndf} \approx 1$ background fit, these checks establish that the classifier concentrates events by kinematics without forging significant artificial structure into the invariant mass spectrum.

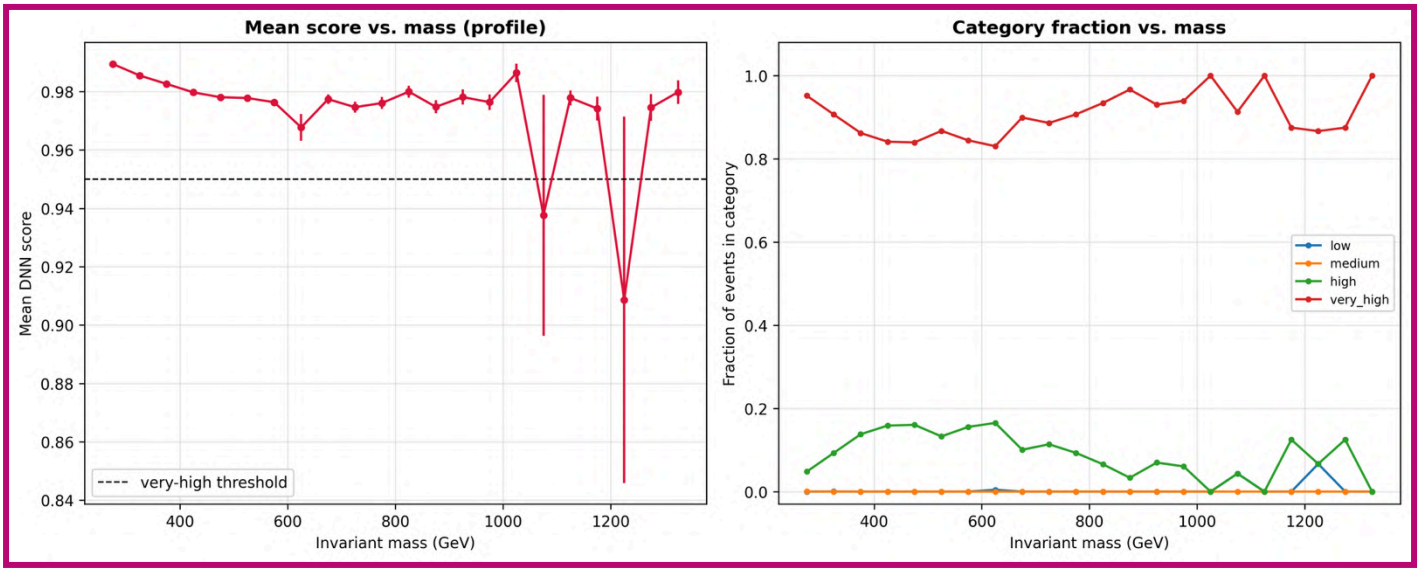


Figure 7: Classifier score to mass relationship in the DNN search region. Left: mean DNN in invariant mass bins. Right: fraction of events in each score category versus $m_{\ell\ell}$.

5.3. Significance Scan

A χ^2/ndf is a global evaluation of our background fit: it evaluates the success of the whole fit. To evaluate statistical significance of an observed excess in individual bins, we use an Asimov significance scan (Cowan et al., 2011):

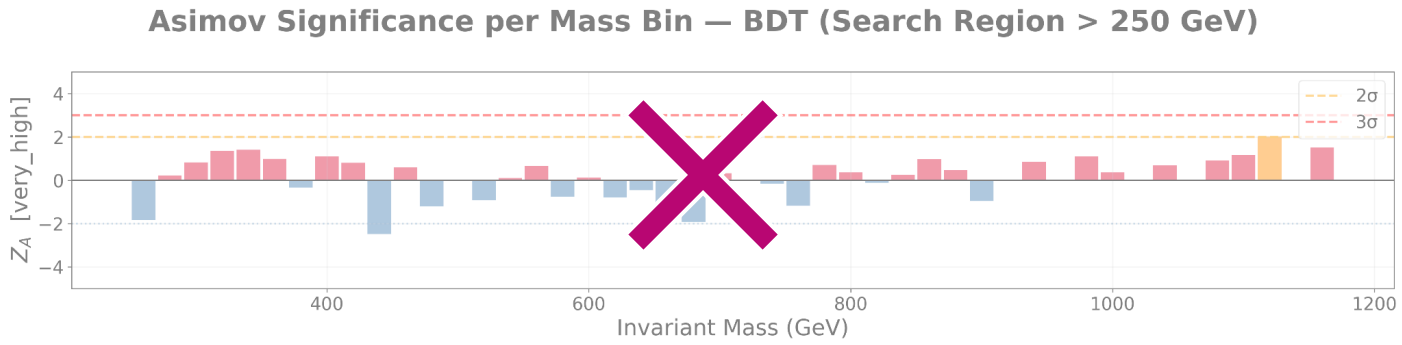
$$Z_A = \sqrt{2\left[n\left(\frac{n}{b}\right) - (n - b)\right]}$$

~~Z_A = √{ 2 [n ln(n/b) − (n − b)] }¶~~

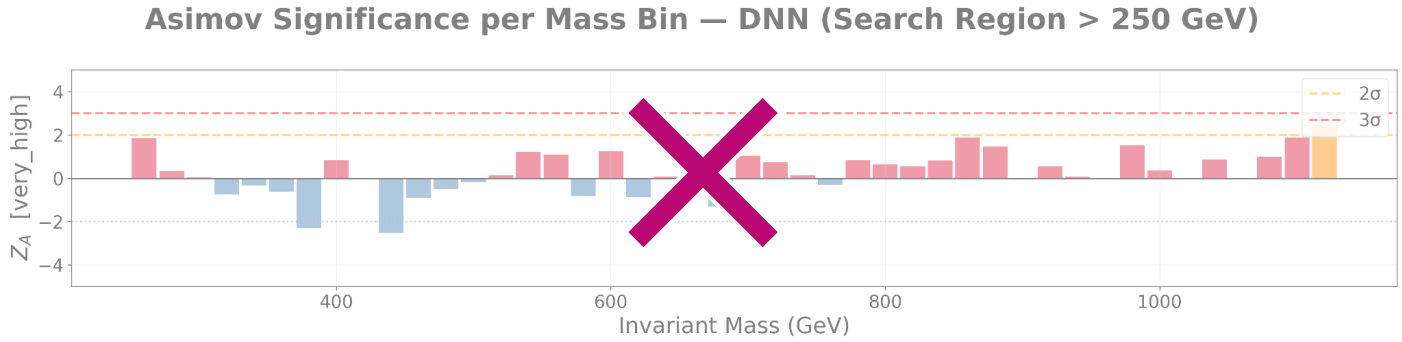
Where n is the number of observed events in a bin and b is the number of events predicted by the background fit. For both classifiers, we compute the Asimov significance per 20 GeV mass bins across the search region. This calculation does not rigourously treat factors such as the fitted background uncertainty, correlation across bins, or bin width, and so it can be considered an approximate local significance estimate rather than a full statistical analysis.

Across the search region, no bin in either classifier exceeded the conventionally required 5σ threshold for a discovery or even the 3σ threshold for evidence of a new particle. However, a discovery was not expected within this workflow anyway: the 3.2 fb^{-1} sample is far smaller than the full Run 2 dataset already used to exclude Z' bosons across this mass range (see Section 7 Conclusion), so this scan tests the workflow's statistical behavior rather than its discovery capability. The largest significance we observed was at $m_{\ell\ell}$ = 1120 GeV, which both classifiers marked as the most significant bin, with the BDT yielding 2.04σ and the DNN yielding 2.83~~40~~σ.

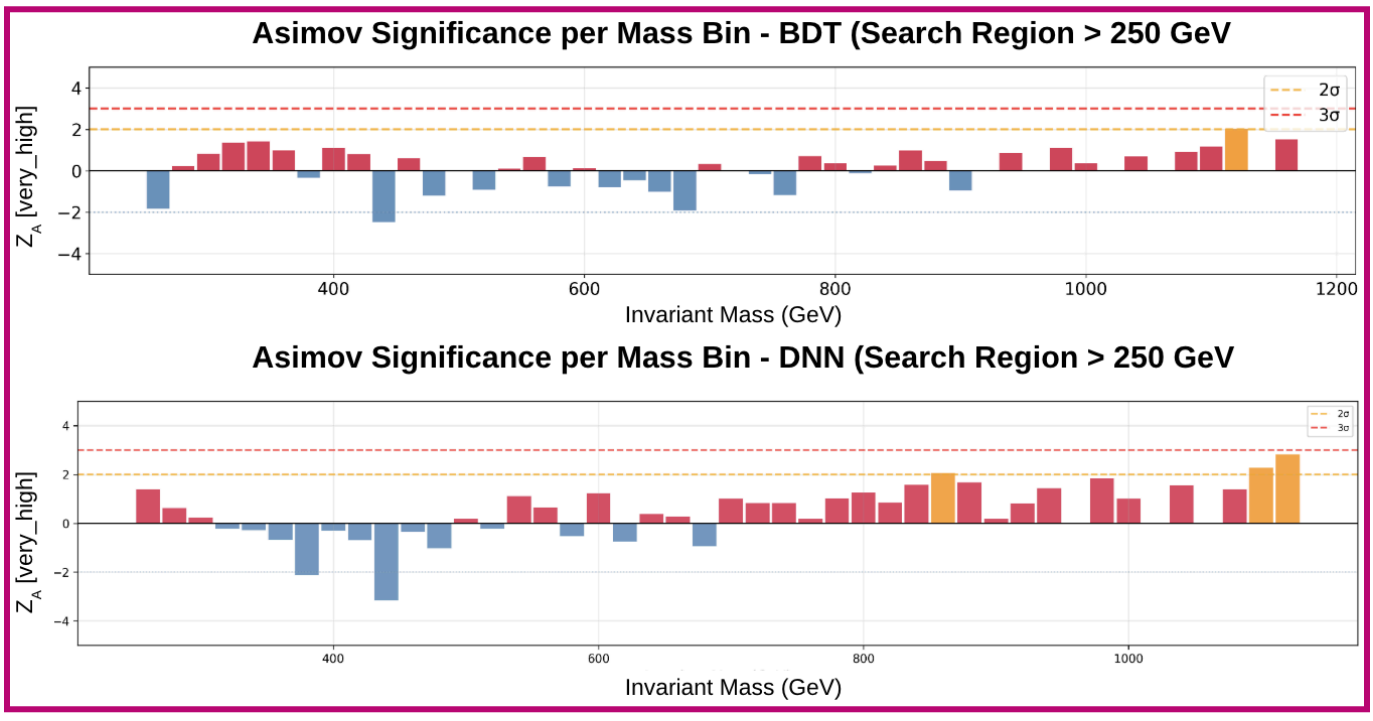
It is important to note that the reported significances use only statistical uncertainties. This should not be interpreted as an official ATLAS result as systematic uncertainties are not considered, and this is mentioned in Section 6.2 Limitations.


Asimov Significance per Mass Bin — BDT (Search Region > 250 GeV)
Z_A [very_high]
Invariant Mass (GeV)
2σ
3σ

406


Asimov Significance per Mass Bin — DNN (Search Region > 250 GeV)
Z_A [very_high]
Invariant Mass (GeV)
2σ
3σ

407


Asimov Significance per Mass Bin - BDT (Search Region > 250 GeV
Z_A [very_high]
Invariant Mass (GeV)
2σ
3σ
Asimov Significance per Mass Bin - DNN (Search Region > 250 GeV
Z_A [very_high]
Invariant Mass (GeV)

408

Figure 87: Asimov significance per 20 GeV bin in the very-high score category for BDT (top) and DNN (bottom). Dashed lines indicate the 2σ and 3σ thresholds. The largest deviation was 2.8340σ at $m_{\ell\ell}$ = 1120 GeV in the DNN.

5.4. Look-Elsewhere Effect and p-value Calculations

This particular excess may be attributed to the look-elsewhere effect in particle physics (Gross & Vitells, 2010), which is a tendency to find a statistically significant event simply because of a large search range. Because we scan 60 independent mass bins (20 GeV bins from 250 to 1450 GeV, which was the largest mass bin with sufficient statistics), a fluctuation of 2σ somewhere along the mass spectrum is not very surprising. We can calculate the one-sided p-values to understand the probability of finding fluctuations this large purely by chance using this formula:

$$P_{global} = 1 - \Phi(z)^N$$

P_global = 1 − Φ(z)^N¶

Where P_global is the probability of finding a particular excess of significance z across N number of bins and Φ(z) is the standard normal cumulative distribution function (CDF), defined as the integral of the Gaussian (Cowan et al., 2011). CDF values are common in statistics textbooks and we did not calculate them by hand. Performing these calculations, we find an approximately 74.94% chance that a 2σ excess appears and a 14.4638.98% chance that a 2.8040σ excess appears at least once across all 60 bins because of statistical noise. This calculation treats the 60 mass bins as independent trials, which is an approximation. Because the experimental dilepton mass resolution is comparable to or broader than a single 20 GeV bin, neighbouring bins are correlated, so the effective number of independent trials is smaller than 60. A rigorous treatment would count the expected "upcrossings" of the continuous test-statistic spectrum above a threshold rather than assuming independent bins. We do not perform this continuous-trials calculation, so the probabilities above should be read as an approximate estimate rather than rigorous p-values. Nevertheless, it demonstrates that Thus, the global significance, or the significance after considering how many bins were searched, is therefore realistically much lower than 2.8340σ and even 2σ. The deviation we observed is attributed to statistical noise from the large amount of data we used, and we can safely report a null result with no evidence for a no statistically significant Z' boson excess in the across the $m_{\ell\ell}$ = 250-14503,000 GeV range probed by our significance scan.

5.5 Injection Study

Because no excess is detected, we can not determine the sensitivity (how large a resonance would have to be to be detected) of this workflow, So, we perform a signal-injection study by injecting Gaussian resonances of width equal to 2% of the resonance mass at 500, 1000, and 1500 GeV (Belis et al., 2023). For each injected sample, we refit the background. We then record the median local Asimov significance as a function of number of injected events. For the DNN, reaching the 5σ discovery threshold requires approximtely 147, 38, and 17 injected events at 500, 1000, 1500 GeV respectively. For the BDT, reaching 5σ requires 157, 50, and 22 events for the same masses. This confirms that both classifiers in this workflow could actually recover a resonance if one is present, and an excess large enough to be detectable would have produced a clear signal.

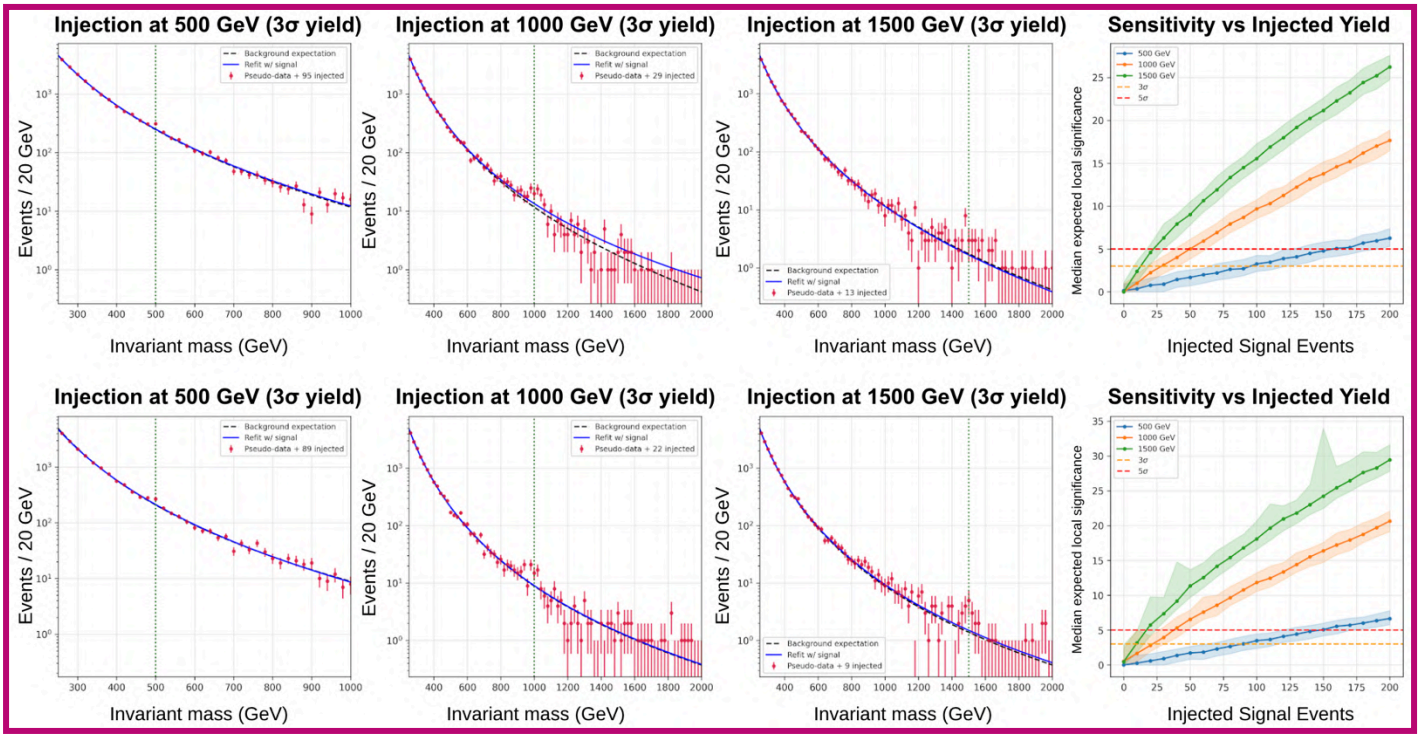


Figure 9: Signal-injection sensitivity study (top: BDT, bottom: DNN). Gaussian resonances of width 2% of the resonance mass are injected onto the fitted background at 500, 1000, and 1500 GeV. The background is refit in each case. The first three plots of each row show the dataset for a 3σ yield. The right plot shows median local Asimov significance versus number of injected events.

6. Discussion

The main methodological goal of this analysis was to use ML classifiers in a bump-hunt search without distorting the background. We achieved this with three important design choices: excluding the invariant mass from features, avoiding any mass-normalised or mass-derived features, and training on non-overlapping mass sidebands. The $\chi^2/\text{ndf} \approx 1.00$ for both background fits in the very-high-score category provides confirmation that this strategy works well, as a classifier that learned from mass features would produce a distorted spectrum that could not be fit with a smooth three-parameter function.

While traditional approaches to bump-hunt searches involve simple kinematic cuts to select events, this ML-based approach concentrates signal-like events into higher scoring categories while preserving the smooth background shape. This suggests that ML classifiers could be useful for improving statistical sensitivity in future searches that include systematic uncertainties.

6.1. Classifier Comparison

The DNN's test AUC (0.9996) substantially outperforms the BDT's (0.9869), reflecting the neural network's greater ability to

model complex input feature combinations. This resulted in the DNN's higher sensitivity in the search region, as it yields a 2.8340σ significance at $m_{\ell\ell}$ = 1120 GeV compared to the BDT's 2.04σ. Despite the agreement between classifiers, this significance threshold is not sufficient enough for this to be considered evidence or discovery.

The two classifiers both offer their own strengths. The BDT's feature importance analysis provides insight into the classification, with a lepton pT dominated scoring (expected from a Z decay). It is more complicated to determine the decision process of the DNN because it uses more complex feature combinations. Nevertheless, presenting both classifiers and their independent results provides cross-check validations and strengthens the credibility of the result. The fact that both models agreed on a null result at the same particular excess increases the confidence in this conclusion.

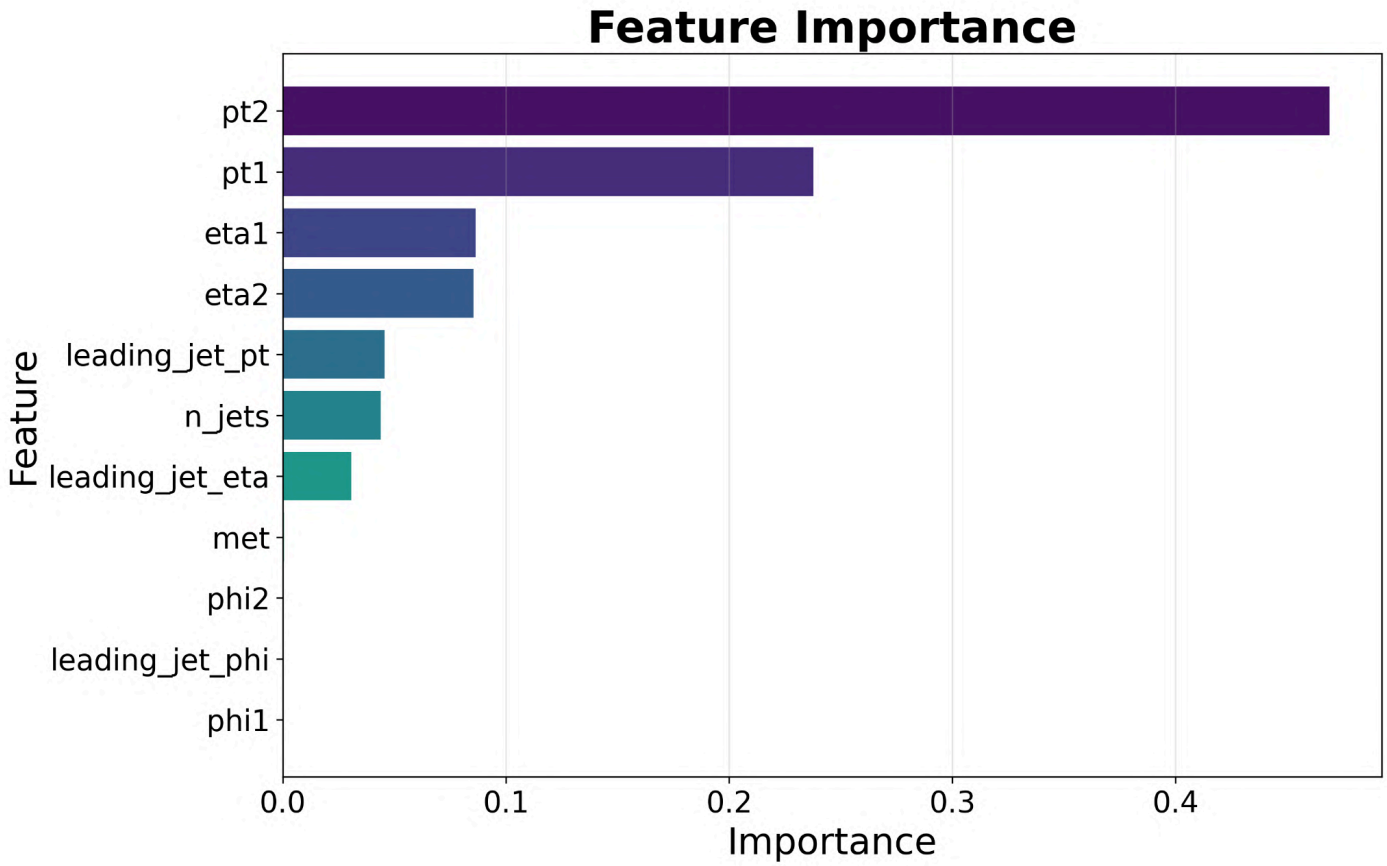


Figure 108: BDT feature importance. Lepton transverse momenta dominate, with pT^2 contributing 47% and pT^1 contributing 24%.

6.2. Limitations and Future Improvements

This analysis is designed as a methodological demonstration rather than a full experimental result, and improvements are needed before producing results on par with official CERN and ATLAS publications:

- No comparison to theoretical models: We do not compare our results against existing benchmark theories and simulated signal events. A complete analysis would interpret the search within one or more theoretical frameworks, predict the expected signal yield for each model, and set quantitative exclusion limits on their parameters.
- ~~Signal injection studies: To evaluate how sensitive the search is, researchers would inject simulated Z' signals at various mass points to check how often the analysis would detect them (Belis et al., 2023).~~ ¶
- Systematic uncertainties: We use only statistical uncertainties from the limited number of events in our selected bins. A complete and proper analysis would include systematic uncertainties on factors like luminosity, lepton identification efficiencies, and choice of background function.
- Alternative background functions: We use a single three-parameter function, but a professional analysis might compare several functional forms and include any variation as an additional uncertainty.
- Mass decorrelation: The DNN's very high AUC (>0.999) warrants some caution about mass correlation in the training data set. While the background fit confirms no sculpting, techniques such as adversarial training or distance-correlation penalties could provide stronger guarantees against mass correlations.
- Look-elsewhere estimate: The global significance reported in Section 5.4 is computed by treating the mass bins as independent trials. In practice, the detector mass resolution correlates neighbouring bins, so this overstates the trials factor. A rigorous global p-value would instead use the continuous-trials (upcrossings) formalism of Gross and Vitells (2010). Because our largest local deviation is already well below the 3σ evident threshold, this does not affect the null result, but it would need to be addressed in any analysis aiming for a discovery-level statistical statement. ▪
- Absolute azimuthal angle (phi): Because of rotational symmetry about the beam, absolute azimuthal angle is relatively meaningless, as indicated by the BDT feature importance (Figure 10). A model utilizing it may be learning detector artifacts. In the future, using relative angles rather than absolute angles may be more valuable.
- Indirect mass learning: As mentioned, invariant mass is excluded as an input feature for both classifiers to mitigate mass correlation. However, since other input variables are used to calculate invariant mass, indirect mass-correlated learning is still possible for the classifiers.
- Combined lepton channels: Our event selection collects dielectron and dimuon events into the same dataset with the same 11 features. While lepton flavor is recorded, it isn't used as an input feature for the classifiers. Because electrons and muons are reconstructed by different subsystems and their momentum resolutions degrade differently, a single model trained on the combined sample is forced to average channel-specific kinematic behavior. A future analysis should either train separate classifiers for each channel or include lepton flavor as an input feature.
- Top-quark background and b-jet veto: Our classifiers are trained to separate Z-peak events from low-mass Drell-Yan events, both of which are characterized by low missing transverse energy and low jet multiplicity. The models therefore have no opportunity to learn MET or jet multiplicity as discriminators, as indicated by their low to zero contribution in the feature importance (Figure 10). The second most significant background in high-mass dilepton resonance searches, top-quark pair production, produces two high-pT leptons together with genuine MET from neutrinos and two b-jets. Because our classifiers rank events primarily by lepton pT, these events might be scored high despite not being Z-like. We do not apply a b-jet veto, the standard method for suppressing top-quark contamination, as b-tagging information was not collected. Future analyses should include b-tagging information and apply a b-jet veto.

- Absence of an opposite-sign requirement: The Z and Z' are electrically neutral, so charge conservation requires their decay products to be an oppositely charged lepton pair, a standard element of the dilepton "golden channel" selection. We do not apply this requirement, because lepton charge was not retained during our event selection and is therefore unavailable in our processed dataset. Our sample therefore contains some same-sign contamination. These contributions, however, are smooth and non-resonant in $m_{\ell\ell}$, so they inflate the background under the fitted spectrum and reduce sensitivity, rather than introducing a localized excess or biasing the search. The effect on training is also limited, as the Z-peak and low-mass Drell-Yan regions are mainly populated by opposite-sign events. Nevertheless, applying an opposite-sign requirement is the first correction we would make in a future analysis.

7. Conclusion

We have presented a machine-learning based search for a Z' boson using ATLAS Run 2 2015 open data, comparing a gradient boosted decision tree and a three-layer MLP trained on mass sideband kinematics. By intentionally excluding dilepton invariant mass from the input features and training on mass sidebands outside of the search region, both classifiers achieve strong discrimination between the sideband background proxy and the Z-like proxy signal events (AUC = 0.9869 and 0.9996 respectively) while describing the high-mass background spectrum with a simple three-parameter function.

Our Asimov significance scan across bins in the search region produces no bins exceeding the 3σ evidence threshold. Our largest local deviation is 2.83~~.40~~σ at $m_{\ell\ell}$ = 1120 GeV using the DNN, which can be considered a statistical fluctuation when the look-elsewhere effect is considered. We found no statistically significant localized excess across the 250-1450 GeV range probed with sufficient statistics; however, this does not imply a definitive null result for a Z' boson, even though a null result is expected in this mass range. Indeed, our 3.2 fb^{-1} sample is a subset of the Run 2 dataset with a full 139 fb^{-1} luminosity that has already been used to exclude Z' bosons across this mass range. A genuine resonance large enough to produce a visible excess in our sample would have appeared at overwhelming significance in the complete dataset, so any localized deviation we observe is almost certainly a statistical fluctuation rather than new physics just by nature of the dataset. ~~Thus, we report a null result with no evidence for a Z' boson in the mass range 250-3,000 GeV. It is worth emphasizing that a null result is not a failed result. Finding no excess and confirming that the Drell-Yan spectrum remains smooth across this mass range narrows the space of where new physics could be found and adds to our understanding of the universe by confirming that presently there is no evidence of a Z' boson.~~

Nevertheless, the value of this work lies in the workflow rather than the search outcome, as the methodology was largely successful despite what the search yielded. We were able to utilize public data and simple kinematic features to formulate a completely reproducible pipeline for a ML-driven bump hunt search that successfully mitigates background sculpting and can be used as ¶

~~The results of our methodology were successful. Using only publicly available data and simple kinematic features, we demonstrate that ML classifiers can be used in bump hunt resonance searches while preserving the background model. The full experimental process is designed to be reproducible, providing~~ a template for future bump hunt analyses using LHC open data. The Z' search presented here should therefore be read as a demonstration of the workflow rather than a competitive experimental search, and the lack of a significant excess reflects the limited integrated luminosity of the

dataset rather than a new exclusion of Z' models.▪

Many improvements would sharpen this workflow from a methodological demonstration into a more rigorous search. The most important corrections concern selection: applying an opposite-sign charge requirement and a b-jet veto would remove the same-sign and top-quark contamination that inflates the background, and training separate classifiers for dielectron and dimuon channel (or using lepton flavor as an additional input feature) would stop the models from averaging over the differing momentum resolutions. Beyond these fixes, there are statistical and interpretive extensions to be made. Considering systematic uncertainties on luminosity and the choice of background function, comparing alternative background parametrization, and replacing the independent-trials look-elsewhere estimate with the more accurate continuous-trials formalism would give any reported significance a more rigorous meaning. Taken together and with other limitations addressed in Section 6.2, these steps define a clear path from a reproducible template toward an analysis with professional-level statistical conclusions.

8. References

ATLAS Collaboration. (2008). The ATLAS experiment at the CERN Large Hadron Collider. Journal of Instrumentation, 3(8), S08003. https://doi.org/10.1088/1748-0221/3/08/S08003

ATLAS Collaboration. (2016). Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 76(5), 292. https://doi.org/10.1140/epjc/s10052-016-4120-y

ATLAS Collaboration. (2019a). Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s}$ = 13 TeV. European Physical Journal C, 79(8), 639. https://doi.org/10.1140/epjc/s10052-019-7140-6

ATLAS Collaboration. (2019b). Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s}$ = 13 TeV with the ATLAS detector. Physics Letters B, 796, 68–87. https://doi.org/10.1016/j.physletb.2019.07.016

ATLAS Collaboration. (2024). ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data [Data set]. CERN Open Data Portal. https://doi.org/10.7483/OPENDATA.ATLAS.AOQL.8TT3

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. Nature Communications, 5, 4308. https://doi.org/10.1038/ncomms5308

Belis, V., Odagiu, P., & Aarrestad, T. K. (2023). Machine learning for anomaly detection in particle physics. Reviews in Physics, 12, 100091. https://doi.org/10.1016/j.revip.2024.100091

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

CMS Collaboration. (2021). Search for resonant and nonresonant new phenomena in high-mass dilepton final states at $\sqrt{s} = 13$ TeV. Journal of High Energy Physics, 2021(7), 208. https://doi.org/10.1007/JHEP07(2021)208

Collins, J. H., Howe, K., & Nachman, B. (2019). Extending the search for new resonances with machine learning. Physical Review D, 99(1), 014038. https://doi.org/10.1103/PhysRevD.99.014038

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. European Physical Journal C, 71(2), 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0

Dolen, J., Harris, P., Marzani, S., Rappoccio, S., & Tran, N. (2016). Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure. Journal of High Energy Physics, 2016(5), 156. https://doi.org/10.1007/JHEP05(2016)156

Drell, S. D., & Yan, T.-M. (1970). Massive lepton-pair production in hadron-hadron collisions at high energies. Physical Review Letters, 25(5), 316–320. https://doi.org/10.1103/PhysRevLett.25.316

Evans, L., & Bryant, P. (2008). LHC machine. Journal of Instrumentation, 3(8), S08001. https://doi.org/10.1088/1748-0221/3/08/S08001

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feickert, M., & Nachman, B. (2021). A living review of machine learning for particle physics. arXiv. https://arxiv.org/abs/2102.02770

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. European Physical Journal C, 70(1–2), 525–530. https://doi.org/10.1140/epjc/s10052-010-1470-8

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 448–456). PMLR. https://arxiv.org/abs/1502.03167

Langacker, P. (2009). The physics of heavy Z′ gauge bosons. Reviews of Modern Physics, 81(3), 1199–1228. https://doi.org/10.1103/RevModPhys.81.1199

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization [Conference paper]. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1711.05101

Metodiev, E. M., Nachman, B., & Thaler, J. (2017). Classification without labels: Learning from mixed samples in high energy physics. Journal of High Energy Physics, 2017(10), 174. https://doi.org/10.1007/JHEP10(2017)174

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807–814). Omnipress.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Terao, K., & Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. Nature, 560(7716), 41–48. https://doi.org/10.1038/s41586-018-0361-2

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., & McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research Section A, 543(2–3), 577–584. https://doi.org/10.1016/j.nima.2004.12.018

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Data & Code Availability Statement

The data used for this experimental setup was accessed via the ATLAS Open Data portal. Link to dataset: https://opendata.cern.ch/record/80000

Cleaned-up scripts used for this methodology are available at this public GitHub repository: https://github.com/TheRealM1M1/zprime_search.

Acknowledgements

I would like to thank [redacted] for his continued support and guidance during this project.

I acknowledge the work of the ATLAS Collaboration to record or simulate, reconstruct, and distribute the Open Data used in this paper, and to develop and support the software with which it was analysed.

AI Use Statement

Claude Opus 4.7/4.8 (Anthropic) was used for minor editing support and minor coding support during this project. All methodological decisions, analysis, results, and scientific content are the author's own.

Author Biography

[redacted]

Mentor Contribution Statement

[redacted]

General Notes

General notes:

- While implementing the additional background-sculpting validation requested in Reviewer 1 Comment 4, I identified a boundary-handling error in the score-categorization step of our DNN analysis. Classifier scores were sorted into categories using an open interval so scored of exactly 1.0 were ignored. Because the network's sigmoid output saturates to exactly 1.0 in single-precision floating point for high-confidence events, and because this saturation correlates with dilepton mass, the very-high category was omitting a mass-dependent subset of events. From the same trained model, I have corrected the category boundary to be inclusive at its upper edge and re-run the full downstream analysis, including the background fit, significance scan, and all category-based figures. No retraining was required, as the error was in post-scoring event selection rather than in the model itself. The corrected background fit yields $\chi^2/\mathrm{ndf} = 1.38$ for the DNN, and the largest local excess is still at $m_{\ell\ell} \approx 1120$ GeV, now at 2.83σ (previously 2.40σ). All figures and values in the revised manuscript have been corrected accordingly. The BDT had a similar bounding error; however, 0 events were affected, so no re-running was necessary. The fixes are applied to the both classifiers' main training/analysis scripts that is shared at the end of the manuscript for reproducibility.
- I rewrote all equations using the Word equation feature to fix the format

Reviewer 1

1. **Reframe the main claim more carefully.**
   The manuscript should present the analysis as an ML-assisted dilepton resonance-search workflow using ATLAS Open Data, rather than as a fully validated Z′ search. The current wording sometimes suggests a stronger experimental conclusion than the analysis can support.

Made edits in the introduction and conclusion to emphasize the workflow and acknowledge that our search is mainly a demonstration of the workflow rather than a competitive search. These edits fall in line with revisions responding to Comment 15.

2. **Clarify the meaning of "signal" and "background" in training.**
   The Z-peak events used for training are Standard Model Z events, not actual Z′ signal events. The author should describe these as a "Z-like proxy signal" or "resonance-like proxy class," and the low-mass Drell–Yan events as a sideband/background proxy. This would make the methodology more precise without weakening the paper.

Changed wording in section 4 and 4.4 to clarify that the signal region is populated by Z-peak events that are meant to act as a Z-like proxy rather than actual Z' signal events and that background region has background proxy events that are not associated with Z-like resonances. Also mentioned limitations of using Z-peak proxy signals when it comes to searching for a multi-TeV Z' whose topology differs from the regular Z boson.

3. **Address mass leakage more directly.**
   Excluding dilepton invariant mass from the input features is a good first step, but the input variables still include the lepton (p_T), (\eta), and (\phi), from which invariant mass is calculated. The author should explicitly acknowledge that the classifiers may still learn mass-correlated information indirectly.

Added a sentence in section 4.1 to acknowledge that mass-correlation is still possible because some of the kinematic inputs we use are used for calculating invariant mass, but excluding invariant mass as a direct input should help mitigate mass correlation. Also added a point in the Limitations section to address this.

4. **Strengthen the background-sculpting validation.**
   The paper should not rely only on (\chi^2/\mathrm{ndf}) values close to one to claim that background sculpting has been avoided. The author should add score-versus-mass plots, mass distributions in each classifier-score category, and a quantitative check of the correlation between classifier score and (m_{\ell\ell}).

Added a paragraph in section 5.2 Background Modeling with mass correlation plots. One plot is a mean score v mass across the DNN's search region and another plot is the fraction of events in each score category across the DNN's search region. I performed a Pearson and Spearman comparison between DNN score and mass for the full search region and the very-high score category and reported and discusses those values (-0.12/-0.23 and -0.15/-0.20). Also added a Kolmogorov-Smirnov comparison value between the high and very-high categories (0.24) to show that the shapes do not drastically diverge which means there is no localized bump in either category caused by sculpting.

5. **Soften the statement that the MLP "eliminates" background sculpting.**
   The manuscript says the switch from a GNN to an MLP eliminated the sculpting problem. That is too strong. A better wording would be that the MLP "reduced" or "mitigated" the issue, pending further validation.

I altered the wording in section 4.3 to replace "eliminates" with "mitigates," and I also acknowledged that background sculpting cannot be fully ruled out.

6. **Add a simple signal-injection study.**
   Even a basic toy injection of narrow Z′-like bumps at a few masses would greatly improve the paper. This would show whether the method could actually recover a resonance if one were present, rather than only showing that no large bump appears in the data.

Added a new section 5.5 for the Injection study. Injected events at 500, 1000, and 1500 GeV for DNN and BDT and refit the background. Then, tested median local significances at different yields and plotted them. I reported how large excesses would have to be to reach 5 sigma at each mass for both classifiers.

7. **Improve the statistical interpretation.**
   The bin-by-bin significance scan should be described as an approximate local-significance estimate, not a full discovery-level statistical analysis. The author should mention that the fitted background uncertainty, bin-to-bin correlations, choice of bin width, and trials factor are not fully treated.

Added a sentence in section 5.3 acknowledging that some factors were not fully treated and that the significance scan should be considered an approximation rather than a full analysis.

8. **Revise the look-elsewhere discussion.**
   The look-elsewhere calculation is useful pedagogically, but the assumption of 60 independent bins is too simple. The author should present it as an approximate estimate and avoid treating the resulting global probabilities as rigorous final p-values.
   Added a line in section 5.4 to acknowledge that the real p-values are more complex and that our calculations are considered general estimations. Also added a paragraph in the Limitations section.
9. **Clarify the searched mass range.**
   The paper should be consistent about the usable search range. It mentions no evidence from 250–3000 GeV, but the significance scan and look-elsewhere discussion appear to use a smaller statistically meaningful range. The conclusion should match the range actually analyzed with sufficient statistics.

Clarified that we search events up until a mass of 3,500 GeV, but because we require bins to have at least 10 events in the significance analysis, our scan only completes across the 250-1450 GeV range.

10. **Give more detail on event selection.**
    The analysis should specify whether opposite-charge lepton pairs are required, what lepton (p_T) thresholds are used, what trigger requirements are applied, whether lepton identification/isolation criteria are used, and whether electron and muon channels are combined or treated separately.

Acknowledged that same-sign cuts were not made in section 3.1. Since correcting this error would require a lot of time to fix (because I would have to reprocess all of the data), I simply acknowledge this oversight in the limitations section and emphasize that it would be the first correction in a future analysis.

The Open Data is already trigger-filtered, so I added in 3.1 that we do not apply any trigger selection/requirement. I mention in section 3.1, 4.1, and Section 6.2 Limitations that lepton flavor is not used as an input to differentiate between electrons and muons and that this is a limitation and future improvement.

We do not apply any pT cuts while selecting events and processing data, however we do apply a cut at 3,500 GeV right before training the model. Despite this cut, the significance scan only spans 250-1450 GeV because we require a minimum of 10 events per 20 GeV bin. Both of these features are explained throughout the manuscript.

Added a brief sentence in section 3.1 explaining that i use reconstructed objects provided by the dataset and no additional ID/isolation cuts are used.

11. **Make the reproducibility claim stronger.**
    The data source is cited, but the paper should also provide the analysis code, exact cuts, trained-model settings, random seeds, binning choices, and scripts or notebooks needed to reproduce the figures and results.

In the Data & Code Availability Statement, I linked a public GitHub repository with the four main scripts used: the data collection file, BDT training and analysis file, DNN training and analysis file, and the newest file used for the mass correlation and signal injection studies. These files are cleaned up but remain the same for

12. **Clarify the role of systematics.**
    The limitations section mentions missing systematic uncertainties, which is good. The main text should also make clear that the reported significances use statistical uncertainties only and therefore should not be interpreted like an official ATLAS/CMS result.

Added a sentence in section 5.3 to address that only statistical uncertainties are used.

13. **Check figure consistency and completeness.**
    The paper reports both BDT and DNN AUC values, but the ROC figure shown appears to be only for the BDT. The author should either include the DNN ROC curve or explain why it is omitted. The overtraining-check caption should also be corrected if the plots are stacked rather than shown left/right.

Included the BDT and DNN ROC curve side-by-side + fixed the caption. Also fixed Figure 4 caption.

14. **Tighten a few physics explanations.**
    Some background explanations should be made more precise. For example, not all Z′ models couple to quarks and leptons “just like the Z,” and the description of why muons reach the muon spectrometer should be revised.

Revised the explanation of Z’ models in the introduction, the description of why muons reach the muon spectrometer (Section 2.2), and the description of the muon pseudorapidity cuts (Section 3.3).

15. **Moderate the conclusion.**
    The final conclusion should say that no statistically significant localized excess is found in this analysis, rather than implying a broad exclusion or definitive null result for Z′ bosons. A strong ending would emphasize the successful construction of the workflow and the clear next steps needed to make it more rigorous.

Made revisions to the concluding paragraph to emphasize the workflow over the search result and to avoid claiming an exclusion of a Z’ boson as a result of our search. Also added “Future Improvements” to the Section 6.2 Limitations title and added a final paragraph in Conclusion to mention the most important next steps that would make the search more acceptable and rigorous.

Reviewer 2

1. You noted a pseudorapidity cut for electrons at a rapidity of less than 2.47, which correctly aligns with the ATLAS electromagnetic calorimeter precision region. However, standard ATLAS analyses typically exclude the transition region between the barrel and endcap cryostats, known as the "crack region" (1.37 < rapidity < 1.52). This region is heavily instrumented with cables and dead material, which significantly degrades electron energy resolution. While this might be a minor effect in this simplified Open Data analysis, acknowledging this standard exclusion (or explicitly stating why you chose not to apply it) would demonstrate a deeper understanding of the ATLAS detector's physical limitations.

Added sentences in Section 3.3 Quality Cuts to acknowledge that we did not use this exclusion.

2. You made the thoughtful design choice to exclude the dilepton invariant mass from your training features to prevent the models from sculpting artificial bumps into the background. However, you provided the exact constituent variables of the invariant mass equation (p_T^1, p_T^2, η^1, η^2, φ^1, φ^2) to a Deep Neural Network. Due to the universal approximation theorem, the DNN is highly capable of reconstructing a proxy for mass using these variables. This is evidenced by your finding that p_T dominates the classification and that validation-region events score exceptionally high. While your χ^2 fit confirms the resulting background remains smooth, the network is still indirectly shaping the background by acting as a dynamic, highly complex p_T cut. Adding a paragraph in your "Limitations" section discussing how providing highly correlated constituent variables can still lead to indirect mass learning would greatly elevate the scientific maturity of your discussion.

Added a sentence in section 4.1 to acknowledge that mass-correlation is still possible because some of the kinematic inputs we use are used for calculating invariant mass, but excluding invariant mass as a direct input should help mitigate mass correlation. Also added a point in the Limitations section to address this.

3. Following from the previous point, your classifiers correctly learned that "higher p_T = more Z-like (signal)" to separate the 91 GeV Z-peak from the 40 GeV Drell-Yan background during training. Consider what this means physically in your >250 GeV search region. In this high-mass regime, both the irreducible Drell-Yan background and the hypothetical Z' signal will produce leptons with extremely high p_T. If your classifier primarily relies on high p_T to assign a "very high" score, it may simply pass nearly all search-region events into that top category, effectively losing its discriminative power between high-mass background and high-mass signal. Acknowledging this physical limitation of training on low-mass kinematics and extrapolating to high-mass searches is a vital nuance that professional physicists grapple with, and discussing it will make your paper much stronger.

In Section 4.7 Validation Region, added a paragraph explaining that the classifier score does not immediately determine whether an event is Z' or not. I acknowledge that the classifiers simply concentrate events with Z-like topology and therefore end up scoring many Drell-Yan background events highly along with potential signal events because of their high pT.

4. You mention that L1 regularisation (Lasso) "drives the weight of less important features to zero... essentially performing a feature selection". While this is perfectly true for linear regression models, in the context of tree-based algorithms like XGBoost, L1 regularisation (alpha) penalises the absolute value of the leaf weights. It encourages leaf values to be exactly zero, which can lead to tree pruning, but it does not directly zero out an input feature's global importance in the way Lasso regression does. Consider slightly rewording this explanation for technical exactness.

Reworded the explanation for L1 regularization in section 4.2.

5. In Table 1, the author lists the absolute azimuthal angles and absolute pseudorapidities of the leptons as input features for the BDT and DNN. In particle physics, the absolute azimuthal angle (phi) of an event is generally physically meaningless because the proton-proton collisions have rotational symmetry around the beam axis. If a machine learning model learns to rely on absolute phi, it might actually be learning the location of dead sensors or localised noise in the ATLAS detector rather than underlying physics. It is a standard best practice to use relative variables instead, such as the angular separation ($\Delta\eta$ or ΔR or $\Delta\phi$). I recommend that the author note this in their "Limitations" section, or better yet, and suggest they swap absolute angles for relative angles in future analyses.

Acknowledged this as a potential next step in the limitations section. Also, referred to the BDT feature importance, which indicated that phi was not utilized by the model.

6. To account for the Look-Elsewhere Effect, the author calculates a global p-value by treating the search region as N = 60 perfectly independent mass bins. This is an excellent educational approximation, but it is an oversimplification of how the Look-Elsewhere Effect works in continuous mass spectra. Because the experimental mass resolution is often wider than a single bin, neighbouring bins are highly correlated. The very paper the author cites for this (Gross & Vitells) specifically provides a methodology based on the "upcrossings" of a continuous χ^2 process, rather than assuming independent discrete bins. While the basic probability calculation in the manuscript successfully illustrates the concept of the LEE, a professional analysis would compute the global significance using the asymptotic formulas for continuous trial factors. This potentially more thorough calculation is not needed here, but it could be maybe discussed in the limitations section.

Added an explanation acknowledging this in Section 5.4 and added a paragraph to the Limitations section.

7. When justifying the bounds of the signal region, the author states, "While the Z boson has a natural width of 2.5 GeV, we use a larger buffer to avoid any Z contamination during analysis". The author is encouraged to double-check that there is no confusion between the theoretical natural decay width and the experimentally reconstructed width. While the quantum mechanical Breit-Wigner width of the Z boson is indeed about 2.5 GeV, the actual "bump" observed in the ATLAS detector is a convolution of this natural width and the Gaussian energy/momentum resolution of the detector subsystems (the calorimeters and muon spectrometer). The observable peak is physically wider than 2.5 GeV because of detector smearing. The author should correct this sentence to

acknowledge that the buffer accounts for both the natural width and detector resolution effects.

Revised this sentence in Section 4.4 to acknowledge that the observable Z width would be larger than 2.5 because of detector effects and that we picked a large buffer to avoid the natural width as well as these effects.

8. In your Data Preparation section, you define your event selection as requiring exactly two same-flavour leptons (dielectron or dimuon). You must explicitly state whether you required these leptons to have opposite electrical charges. Because the Z and Z' bosons are electrically neutral, conservation of charge dictates that they must decay into an oppositely charged pair (i.e., electron+positron or muon+ anti-muon). If you did not apply an opposite-sign (OS) requirement, your data likely contain unnecessary contamination from same-sign (SS) fake backgrounds and diboson processes. If you did apply this cut, be sure to add a sentence to explicitly mention it, as it is a fundamental requirement of the "golden channel."

Acknowledged that same-sign cuts were not made in section 3.1. Since correcting this error would require a lot of time to fix (because I would have to reprocess all of the data), I simply acknowledge this oversight in the limitations section and emphasize that it would be the first correction in a future analysis.

9. Your machine learning framework trains the models to specifically distinguish Z-peak events (signal) from low-mass Drell-Yan events (background). By training exclusively on Drell-Yan, your classifier is blind to the second most dominant background in high-mass dilepton searches: top-quark pair production. Because top quarks decay into W bosons and b-quarks, top+top-bar events produce two high-pT leptons, missing transverse energy (MET) from neutrinos, and multiple jets. Since you are already passing n_jets, jet kinematics, and MET into your classifiers, your model might be inadvertently scoring top+top-bar events as highly Z-like simply because they are highly energetic. Noting the lack of a b-jet veto (the standard method for removing top-quark contamination) in your Limitations section will demonstrate a mature understanding of background topologies.

Added a paragraph to limitations addressing the lack of a b-veto and acknowledging it as a future addition.

10. You note that your event selection aggregates both dielectron and dimuon events, and you apply the same 11 input features to both sets. In the ATLAS detector, electrons and muons have fundamentally different detection mechanisms (calorimeters vs. spectrometers), and their momentum resolutions scale differently at high energies. By combining them into one dataset without providing the classifier a feature to distinguish them (such as an is_muon boolean), the model is forced to average out channel-specific kinematic nuances. Consider adding a note in your limitations that future analyses should either train separate models for the electron-positron and muon-anti-muon channels or include lepton flavour as a designated input feature.

Added a sentence in the Limitations section to acknowledge this and note that it could be applicable to future analyses.

11. You creatively used the Standard Model Z boson peak (71.2-111.2 GeV) as your "signal" during training because you opted not to use simulated Z' events. While this brilliantly avoids relying on theoretical simulations, it inherently assumes that a 2000 GeV Z' looks

kinematically identical to a 91 GeV Z boson, just scaled up. In reality, the phase space, angular momentum conservation constraints, and parton distribution functions at the multi-TeV scale produce topologies that differ from a 91 GeV decay. Adding a sentence in your limitations acknowledging that a classifier optimized for the Z pole might be sub-optimal for capturing the exact topology of a multi-TeV Z' will demonstrate a highly sophisticated understanding of collider kinematics.

Improved explanation of the signal region in section 4.4 to emphasize the limitations of using Z-peak proxy signals as stand in for Z' kinematics.

12. You accurately state that your 2015 dataset has an integrated luminosity of 3.2 fb^{-1}, and you note that official full Run 2 analyses utilised 139 fb^{-1}. Because the full 139 fb^{-1} dataset has already ruled out a Z' in this mass range, finding a true, physical Z' in a 3.2 fb^{-1} subset of that exact same data is physically impossible; any localised excess would strictly be a statistical fluctuation. While you perfectly frame your paper as a "methodological demonstration", explicitly mentioning that the limited luminosity intrinsically precludes a novel discovery adds a mature, grounded perspective to your final discussion.

In the conclusion, added sentences explaining that the limited luminosity contributes to an expected absence of statistically significant results, but that the methodology remains successful nevertheless.

Reviewer #1

I enjoyed reading this revision. It is a considerably stronger and more thoughtful paper than the version submitted in the first round. The author has responded carefully to the reviewers and has understood the scientific reasons behind the comments. The manuscript now describes the work appropriately as a student demonstration of a machine-learning-assisted resonance-search workflow, rather than presenting it as a result comparable to an official ATLAS analysis. The distinction between a genuine Z′ signal and the Z-peak events used as a proxy training class is also much clearer. These changes give the paper a more accurate and mature foundation.

The additional work is genuinely impressive, particularly for a student researcher. The new score-versus-mass plots, category fractions, correlation measurements, revised background fits, and injection study show considerable initiative. I was especially impressed by the author's handling of the score-boundary error. Rather than minimizing the problem, the author identified its source, corrected the selection, repeated the downstream analysis, and reported the changed results openly. The revised χ^2/ndf of 1.38 and largest local fluctuation of 2.83σ at 1120 GeV now appear to be presented consistently. That degree of honesty and persistence is commendable at any career stage and exceptional in a student paper.

I have only a few requests for the final version. The manuscript should not suggest that the DNN has greater Z′ sensitivity simply because it has a higher AUC or produces the larger fluctuation at 1120 GeV. The AUC measures separation of the low-mass proxy classes, while the 2.83σ result is an observed fluctuation. The injection study should likewise be described carefully: as presented, it tests whether the fitted-spectrum procedure can recover an injected bump, but it does not directly measure the classifier's efficiency for genuine Z′ events.

I would also soften claims that the χ^2, correlation coefficients, and KS result "confirm" the absence of background sculpting. It would be more accurate to say that these checks find no strong evidence of harmful sculpting. If readily available, the KS p-value and sample sizes would be useful additions. The 111.2–250 GeV region should be described as the high-mass Drell–Yan continuum rather than simply a detector-smeared Z tail. Finally, the empty bullet, typographical mistakes, and remaining tracked-change artifacts should be removed from the publication copy.

These are modest corrections and do not alter my overall assessment. For a student journal, this is an ambitious, unusually transparent, and educationally valuable piece of work. The author should be proud of the revision, and I support its acceptance.

Reviewer #2

I recommend accepting this manuscript for publication because the author meticulously addressed all prior feedback, including correcting the DNN boundary-handling error, adding a signal-injection sensitivity study, and properly typesetting all mathematical equations. The revised paper presents a scientifically mature, reproducible machine-learning workflow for dilepton resonance searches that successfully demonstrates how to mitigate background sculpting using ATLAS Open Data. Furthermore, by deeply expanding the limitations section to address complex physical and statistical constraints, the author has elevated this methodology to fully meet the rigorous academic standards of an academic publication. The manuscript has high potential to be an insightful and motivating read for many in the Convergence Journal's target audience.

Thank you for the opportunity to review the revised manuscript, “Machine-Learning-Driven Search for a Z′ Boson in the Dilepton Channel Using ATLAS Open Data.” I have read the revised paper alongside the author’s point-by-point response. My recommendation is acceptance, with a few minor corrections before publication.

I enjoyed reading this revision. It is a considerably stronger and more thoughtful paper than the version submitted in the first round. The author has responded carefully to the reviewers and has understood the scientific reasons behind the comments. The manuscript now describes the work appropriately as a student demonstration of a machine-learning-assisted resonance-search workflow, rather than presenting it as a result comparable to an official ATLAS analysis. The distinction between a genuine Z′ signal and the Z-peak events used as a proxy training class is also much clearer. These changes give the paper a more accurate and mature foundation.

The additional work is genuinely impressive, particularly for a student researcher. The new score-versus-mass plots, category fractions, correlation measurements, revised background fits, and injection study show considerable initiative. I was especially impressed by the author’s handling of the score-boundary error. Rather than minimizing the problem, the author identified its source, corrected the selection, repeated the downstream analysis, and reported the changed results openly. The revised χ^2/ndf of 1.38 and largest local fluctuation of 2.83σ at 1120 GeV now appear to be presented consistently. That degree of honesty and persistence is commendable at any career stage and exceptional in a student paper.

I have only a few requests for the final version. The manuscript should not suggest that the DNN has greater Z′ sensitivity simply because it has a higher AUC or produces the larger fluctuation at 1120 GeV. The AUC measures separation of the low-mass proxy classes, while the 2.83σ result is an observed fluctuation. The injection study should likewise be described carefully: as presented, it tests whether the fitted-spectrum procedure can recover an injected bump, but it does not directly measure the classifier’s efficiency for genuine Z′ events.

I would also soften claims that the χ^2, correlation coefficients, and KS result “confirm” the absence of background sculpting. It would be more accurate to say that these checks find no strong evidence of harmful sculpting. If readily available, the KS p-value and sample sizes would be useful additions. The 111.2–250 GeV region should be described as the high-mass Drell–Yan continuum rather than simply a detector-smeared Z tail. Finally, the empty bullet, typographical mistakes, and remaining tracked-change artifacts should be removed from the publication copy.

These are modest corrections and do not alter my overall assessment. For a student journal, this is an ambitious, unusually transparent, and educationally valuable piece of work. The author should be proud of the revision, and I support its acceptance.