Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Amit Shkuri

מאת
עמית שקורי

פרדיגמה מוכוונת נתונים לחיפושי תהודה מעבר למודל הסטנדרטי

A data-directed paradigm for BSM resonance searches

Advisors:
Shikma Bressler
Eilam Gross

מנחים:
שקמה ברסלר
עילם גרוס

May 2024

ניסן ה׳תשפ״ד

Acknowledgments

My heartfelt gratitude is extended to my advisors: Shikma, for her guidance, mentorship, and insightful feedback throughout my research, and Eilam, whose innovative ideas and unique perspective on tackling challenging goals have been crucial. Their input was invaluable.

Thanks to my DDP group colleagues, Nilotpal and Etienne, for their feedback and support. Our collaboration was not only effective but also enjoyable.

To everyone in the DDP group—Ali, Ilya, Samuel, Julien, Joséphine, Radoslav, Fannie, Bruna, and Jean-François—working alongside such a diverse and talented group of scientists has been a privilege.

My friends at the Weizmann Institute have been a pillar of support during these years; I am deeply grateful for your friendship.

Most importantly, I thank my wife, parents, and brothers for their constant encouragement and belief in me. None of this would have been possible without your support.

Abstract

The bump-hunt Data-Directed Paradigm (DDP) was proposed as a method to efficiently search for resonant Beyond the Standard Model (BSM) physics. It uses a Neural Network (NN) to rapidly map an invariant mass distribution into a distribution of statistical significance for excesses of events.

This thesis presents advances made to the original proposed model. In particular, the NN architecture was improved, and the input and target training datasets were enhanced. All modifications are discussed in detail. They are followed by a comprehensive study of several NNs performances on different test datasets.

A single NN is shown to perform well on a large variety of mass distributions with varying number of bins, signal widths, and dynamic ranges. These results pave the way towards employing the bump-hunt DDP on real experimental data.

List of Abbreviations

BSM Beyond the Standard Model

VEV Vacuum Expectation Value

SSB Spontaneous Symmetry Breaking

QCD Quantum Chromodynamics

MLE Maximum Likelihood Estimation

PL Profile Likelihood

PLR Profile Likelihood Ratio

PLRT Profile Likelihood Ratio Test

DDP Data-Directed Paradigm

NN Neural Network

MC Monte Carlo

DR Dynamic Range

DM Dark Machines

SM Standard Model

PDF Parton Distribution Function

LHC Large Hadron Collider

ML Machine Learning

CWola Classification WithOut Labels

ANODE Anomaly Detection with Density Estimation

CATHODE Classifying Anomalies Through Outer Density Estimation

CNN Convolutional Neural Network

MSE Mean Squared Error

AE Auto Encoder

VAE Variational Auto Encoder

pdf probability density function

HEP High Energy Physics

Contents

1 Introduction

The Standard Model (SM) of particle physics is a cornerstone of our understanding of the universe, providing a comprehensive framework to describe the elementary particles and their interactions. Despite its profound success in predicting experimental results, including the last landmark discovery of the Higgs boson, the SM does not explain various phenomena. Among others, it doesn't explain the origin of dark matter, the baryon asymmetry or gravity. Thus, the SM can be thought of as an effective low-energy theory embedded within a broader Beyond the Standard Model (BSM) framework(s).

Some proposed BSM theories suggest the existence of new particles and forces yet to be discovered, and many predict the existence of new resonances at the TeV scale and below. Thus, the search for new resonances, either theoretically-predicted or model-agnostic, is a core strategy for discovery in experimental high energy physics, including at the LHC (e.g recently [1],[2],[3])

With almost no exception, all BSM searches have been conducted following the blind analysis paradigm, in which an enormous amount of time and effort is invested before looking at the data, i.e., on background modeling and systematic uncertainty estimation. These resource-intensive tasks have allowed only a limited region in the space spanned by all observables to be explored to date. Indeed, searches for resonances typically focus on inclusive final states – di-lepton, di-photon, di-jet, etc. – ignoring all other observables and avoiding exclusive selections such as di-lepton + jets, di-jets + missing transverse momentum, di-photon within a $t\bar{t}$ topology, etc. Moreover, event selection is usually optimized relative to predefined signal models within the studied final states, making these studies sensitive to very specific signal scenarios. So far, no significant indication of BSM physics has been found.

Complimentary to the blind analysis paradigm, the recently proposed Data-Directed Paradigm (DDP) approach leverages computational advances to accelerate the search for BSM phenomena by directly analyzing experimental data. It avoids the need to use time and resource consuming tasks such as Monte Carlo (MC) simulation, background modeling, etc..

Given the large number of plausible signals that could manifest in an infinite number of exclusive regions, and moreover, the limited time, manpower, and resources at hand, approaches like the DDP might provide our best chance for discovering BSM physics.

A DDP search is based on two key ingredients:

1. A theoretically well-established property of the SM based on which deviations can be searched for – here we exploit the fact that within the SM, in the absence of resonances, almost any invariant mass distribution is smoothly falling.

2. An efficient algorithm to scan the observable space in search for these deviations – here we train a deep Neural Network (NN) to map any invariant mass distribution into a distribution of statistical significance for excesses of events ("bumps"). The latter is known as a "z" distribution and is based on the Profile Likelihood (PL) test for positive signals [4].

The challenge of bump-hunting is an excellent showcase for a search in the DDP; even a simple implementation was shown to achieve good accuracy [5]. As long as the underlying background distribution is smoothly falling, a single trained NN, as described in this thesis, can quickly perform statistical inference from many selections of observed data.

Our research improves over the original DDP work by developing a more robust deep NN designed for detailed analysis of mass distributions. The improvement is mainly three-fold. First, a novel Convolutional Neural Network (CNN) architecture was developed. It is tailored with multi-dimensional layers and variable-size kernels making the NN insensitive to the binning of the mass

distributions. Second, the training dataset was enriched with more functional forms and data-like distributions. The Dynamic Range (DR) was broadened as well. Third, various signal scenarios are considered. Finally, we developed a set of tests to evaluate and compare the performance of different NNs. It provides a step towards quantifying the systematic uncertainties associated with the NN prediction.

This thesis is organized as follows: in Section 2 we provide the scientific background relevant to this work. The DDP concepts are detailed in Section 4. The methodologies developed in this work are described in Section 5 followed by a comprehensive set of results in Section 6. Finally, conclusions and possible future directions are given in Section 7.

2 Scientific background

2.1 The Standard Model of particle physics

The SM accounts for the strong, weak, electromagnetic, and Yukawa interactions of the elementary particles. It is defined (e.g., in [6]) by the symmetry group:

$$SU(3)_C \times SU(2)_L \times U(1)_Y. \tag{1}$$

There are three fermion generations, each consisting of five different representations:

$$Q_{L_i}(3,2)_{+1/6}, \quad U_{R_i}(3,1)_{+2/3}, \quad D_{R_i}(3,1)_{-1/3}, \quad L_{L_i}(1,2)_{-1/2}, \quad E_{R_i}(1,1)_{-1}, \quad i = 1,2,3. \tag{2}$$

There is a single scalar multiplet:

$$\phi(1,2)_{+1/2}. \tag{3}$$

We use the notation $(A, B)_Y$ where A is the representation under $SU(3)_C$, B is the representation under $SU(2)_L$, and Y is the hypercharge. The fermions that transform as triplets of $SU(3)_C$ are called quarks, while those that transform as singlets of $SU(3)_C$ are called leptons.

The masses of fermions and quarks are generated through the Yukawa interactions:

$$-\mathcal{L}_{\text{Yuk}} = Y^u_{ij}\overline{Q}_{Li}U_{Rj}\tilde{\phi} + Y^d_{ij}\overline{Q}_{Li}D_{Rj}\phi + Y^e_{ij}\overline{L}_{Li}E_{Rj}\phi + \text{h.c.}. \tag{4}$$

According to the Goldstone theorem, once the scalar field acquires a Vacuum Expectation Value (VEV), $|\langle\phi\rangle| = \frac{v}{\sqrt{2}}$ and since it is an $SU(3)_C$ singlet, we obtain the following Spontaneous Symmetry Breaking (SSB)

$$SU(3)_C \times SU(2)_L \times U(1)_Y \to SU(3)_C \times U(1)_{EM}. \tag{5}$$

This process assigns mass to particles and establishes their electromagnetic charges $Q = T_3 + Y$, where T_3 is the weak isospin and Y the hypercharge. The mass terms arise when replacing ϕ with its VEV and after diagonalizing the Yukawa matrices.

While the neutrinos (ν_e, ν_μ, ν_τ) are massless, the charged lepton (e, μ, τ) masses are given by

$$-\mathcal{L}_{m_l} = \frac{y_e v}{\sqrt{2}}\overline{e_L}e_R + \frac{y_\mu v}{\sqrt{2}}\overline{\mu_L}\mu_R + \frac{y_\tau v}{\sqrt{2}}\overline{\tau_L}\tau_R + \text{h.c.}. \tag{6}$$

The up-type (u_R, c_R, t_R) and down-type (d_R, s_R, b_R) quarks accumulate mass through

$$\begin{aligned} -\mathcal{L}_{m_u} &= \frac{y_u v}{\sqrt{2}}\overline{u_L}u_R + \frac{y_c v}{\sqrt{2}}\overline{c_L}c_R + \frac{y_t v}{\sqrt{2}}\overline{t_L}t_R + \text{h.c.} \\ -\mathcal{L}_{M_d} &= \frac{y_d v}{\sqrt{2}}\overline{d_L}d_R + \frac{y_s v}{\sqrt{2}}\overline{s_L}s_R + \frac{y_b v}{\sqrt{2}}\overline{b_L}b_R + \text{h.c.}. \end{aligned} \tag{7}$$

Finally, the mass of the weak gauge bosons originate through

$$\mathcal{L}_{m_V} = -\frac{1}{4}g^2v^2W^{+\mu}W^-_\mu - \frac{1}{8}(g^2+g'^2)v^2Z^\mu Z_\mu. \tag{8}$$

The defining characteristics of the SM particles are summarized in Table 1.

Table 1: The SM particles

particle	color	Q	mass $[v]$
$W^\pm$	(1)	±1	$\frac{1}{2}g$
Z^0	(1)	0	$\frac{1}{2}\sqrt{g^2+g'^2}$
γ	(1)	0	0
g	(8)	0	0
h	(1)	0	$\sqrt{2\lambda}$
e,μ,τ	(1)	-1	$y_{e,\mu,\tau}/\sqrt{2}$
ν_e,ν_μ,ν_τ	(1)	0	0
u,c,t	(3)	$+2/3$	$y_{u,c,t}/\sqrt{2}$
d,s,b	(3)	$-1/3$	$y_{d,s,b}/\sqrt{2}$

2.2 Bumps in invariant mass distributions

Sharp structures in invariant mass spectra are not common in background processes, which tend to produce smooth distributions. Hence, bump-hunting - the method of discovering new particles through peaks in the invariant mass distribution of their decay products - is commonly used in particle physics. It confirmed, e.g., the existence of particles such as the W [7], Z [8], and Higgs bosons [9, 10], and it is a key method in searching for BSM particles.

In the following, we briefly discuss the theoretical origin of resonances in two particle collisions, e.g., $p_Ap_B \to \{p_f\}$ scattering where p_A and p_B are the incoming particles and $\{p_f\} = p_1, p_2, ...$ are the outgoing particles. The differential cross-section is given by [11]

$$d\sigma = d\Pi_n \frac{\mathcal{M}|(p_Ap_B \to \{p_f\})|^2}{4E_AE_B|v_A - v_B|} \tag{9}$$

where $\mathcal{M}$ is the invariant matrix element, E_A (E_B) and v_A (v_B) are the energy and velocity of particle A (B), respectively, and $\int d\Pi_n$ is the Lorentz invariant phase space integral over the outgoing particles

$$\int d\Pi_n = \prod_f \int \frac{d^3p_f}{(2\pi)^3}\frac{1}{2E_f}(2\pi)^4\delta^{(4)}\left(\sum_f p_f - p_A - p_B\right). \tag{10}$$

For two outgoing particles, it is convenient to carry out the phase space integral in the Center of Mass (COM) frame, where $(\vec{p_1})_{\rm CM} = -(\vec{p_2})_{\rm CM}$

$$\int d\Pi_2 = \int \frac{d^3p_1}{(2\pi)^3}\frac{1}{2E_12E_2}(2\pi)\delta(E_{\rm CM} - E_1 - E_2) = \int d\Omega_{\rm CM}\frac{|\vec{p}_1|_{\rm CM}}{16\pi^2E_{\rm CM}} \tag{11}$$

and E_{CM} is the energy of the system in the CM frame

$$E_{\text{CM}} = \sum_i p_i^0|_{\text{CM frame}} = \sqrt{\left|\left(\sum_i p_i^0, \vec{0}\right)\right|^2} = \sqrt{\left|\sum_i p_i\right|^2} \equiv \sqrt{s}. \tag{12}$$

The differential cross-section simplifies to

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{CM}} = \frac{|\mathcal{M}(p_A p_B \to \{p_f\})|^2}{4E_A E_B |v_A - v_B|} \frac{|p_1|}{16\pi^2 E_{\text{CM}}} \tag{13}$$

and in the special case where all the masses of the incoming and outgoing particles are equal

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{CM}} = \frac{|\mathcal{M}(p_A p_B \to \{p_f\})|^2}{64\pi^2 E_{\text{CM}}^2}. \tag{14}$$

$\mathcal{M}$ depends on the type of interactions and exact particles involved. Considering a scattering process mediated by the exchange of some χ particle.

We can compute the $\chi\chi \to \chi\chi$ cross-section. The amplitude is

$$\mathcal{M} = g^2 \left[\frac{i}{(p_1+p_2)^2 - m^2} + \frac{i}{(p_1-p_3)^2 - m^2} + \frac{i}{(p_1-p_4)^2 - m^2}\right]. \tag{15}$$

We have $(p_1+p_2)^2 = p_1^2 + p_2^2 + 2p_1 \cdot p_2 = 4m_\chi^2 + 4\vec{p_1}^2$, so when $\vec{p_1}^2 = \frac{1}{4}(m^2 - 4m_\chi^2)$ the first element of the amplitude diverges. However, the cross-section is a physical quantity and must not diverge. Neglecting the two non-diverging terms, the divergence of the first term can be regulated with the following addition in the denominator:

$$\left(\frac{d\sigma}{d\Omega}\right)_{CM} = \frac{1}{64\pi^2 E_{CM}^2} \left|\frac{1}{(p_1+p_2)^2 - m^2}\right|^2 \to \frac{1}{64\pi^2 E_{CM}^2} \frac{1}{|(p_1+p_2)^2 - m^2|^2 + m^2\Gamma^2}. \tag{16}$$

This is obtained when employing propagators of the form

$$\frac{i}{p^2 - m^2 + im\Gamma} \tag{17}$$

known as the relativistic invariant Breit-Wigner formula. where Γ is the decay width of a particle with mass m.

In conclusion, an unstable particle states appear in scattering experiments as a resonance. Near the resonance energy, the scattering amplitude is given by the Breit-Wigner formula and the differential cross-section peaks. The width of the resonance peak is equal to the decay rate of the unstable state.

2.3 Smoothly falling backgrounds

Following [12], we define:

- $f_i(x_i)$ - the Parton Distribution Function (PDF), i.e., the probability that f_i acquires fraction x_i of the total proton energy
- s is the total center-of-mass energy squared in the pp system

- $\hat{s} = x_1 x_2 s$ - the event energy
- $\hat{\sigma}_{ij}(\hat{s})$ - the partonic cross-section at energy $\hat{s}$ that depends on the specific matrix elements of the process

The differential cross-section of the process $pp \to$ final state as a function of the event energy is given by

$$\frac{d\sigma(pp \to f)}{d\hat{s}} = \sum_{i,j} \hat{\sigma}_{ij}(\hat{s}) \int_0^1 \int_0^1 dx_i dx_j f_i(x_i) f_j(x_j) \delta(\hat{s} - x_i x_j s) \tag{18}$$

While the PDF of the valence quarks peak at about $x = 0.2$, the PDFs of the other partons are steeply falling with x. Thus, the probability of getting an event of a given $\hat{s}$ energy decreases with increasing x_i and x_j. It can be shown that for most processes, this falling behavior does not change for precise calculations of $\hat{\sigma}_{ij}(\hat{s})$. Considering for example Large Hadron Collider (LHC) top pair production where the partonic cross-section is

$$\hat{\sigma}_{gg \to t\bar{t}} = \frac{\pi \alpha_s^2 \beta}{48 \hat{s}} \left(31\beta + \left(\frac{33}{\beta} - 18\beta + \beta^3 \right) \ln \left[\frac{1+\beta}{1-\beta} \right] - 59 \right), \tag{19}$$

where $\beta = \sqrt{1 - \frac{4m_t^2}{\hat{s}}}$. Apart from the $t\bar{t}$ production at a typical $\tau = \frac{\hat{s}}{s}$ at LHC14: $(\frac{2m_t}{14TeV})^2 \sim 6 \times 10^{-4}$, the cross-section falls fast with the fraction of the total energy used in the production.

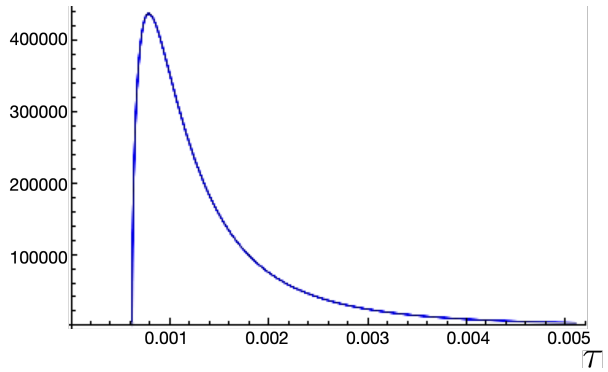

Figure 1: $gg \to t\bar{t}$ coss-section as a function of the ratio between the event-energy and pp collision energy.

2.4 Statistical Inference

Statistical analysis is carried out to assess the agreement between the predictions of any proposed model and observed data. Let $\vec{x}$ be an outcome of a measurement. A hypothesis H is a statement about the probability of observing data $\vec{x}$. For continuous variables, it specifies a probability density function (pdf). The null hypothesis, H_0, states that the data consists of background-only events, while the alternative hypothesis, H_1, assumes that the data consists of background and signal events. $P(\vec{x}|H)$ denotes the probability of finding data $\vec{x}$ under hypothesis H.

To evaluate the null hypothesis against an alternative hypothesis, we define test statistics. It is used to specify a subset of the data, e.g., the critical region (w), chosen such that under H_0 the probability of observing data within w is less than or equal to a significance level (α)

$$P(x \in w|H_0) \leq \alpha. \tag{20}$$

The critical region is defined such that observing x within the region w suggests that the event's occurrence is less likely under the null hypothesis H_0 and more likely under the alternative H_1. Therefore, H_0 is dismissed in favor of H_1.

The likelihood function, $\mathcal{L}(\vec{\theta})$, measures the probability of observing the data under a model defined by the parameters $\vec{\theta}$. We drop the $\vec{\cdot}$ hereafter. Maximum Likelihood Estimation (MLE) seeks the parameter values, $\hat{\theta}$, that maximize the likelihood function, thereby identifying the model parameters most likely to have generated the observed data. To evaluate hypotheses, the likelihood ratio contrasts the likelihoods under H_0 and H_1, calculated as:

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} \tag{21}$$

where θ_0 represents parameters under H_0 and $\hat{\theta}$ are the MLE parameters under the alternative hypothesis H_1. The log-likelihood ratio, a refinement of this concept, is defined as:

$$\lambda = -2\ln(\Lambda) = -2\ln\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right), \tag{22}$$

facilitating easier computations and interpretation, especially in complex models. The distribution of λ under H_0 approximates a χ^2 distribution when the number of degrees of freedom n is the number of free parameters μ in the hypothesis [13], allowing for convenient significance testing.

While the likelihood ratio is a powerful comparative tool, its direct application can be complicated by the presence of nuisance parameters — parameters that are not of primary interest but must be accounted for. The PL refines the likelihood analysis by focusing on a parameter of interest, μ, while accounting for nuisance parameters, $\vec{\theta}$, through maximization. This approach simplifies statistical inference by isolating the effect of μ. The Profile Likelihood Ratio (PLR) for a parameter μ is then defined similarly to the likelihood ratio, but within the context of the PL, as:

$$\mathrm{PLR}(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \tag{23}$$

where $\mathcal{L}(\mu, \hat{\hat{\theta}})$ represents the PL when μ is fixed, and $\hat{\hat{\theta}}$ are the nuisance parameters optimized for this fixed μ. $\mathcal{L}(\hat{\mu}, \hat{\theta})$ denotes the maximum PL, where both μ and θ are optimized. This ratio assesses the strength of evidence for μ relative to the best-fit value $\hat{\mu}$, facilitating a focused analysis on the parameter of interest.

The application of the likelihood-ratio test stands out as a pivotal method in particle physics. This test contrasts the observed data with predictions under the null hypothesis to derive a p-value. The agreement, or lack thereof, between the observed data and the null hypothesis, is further quantified using the significance level, Z, calculated as

$$Z = \Phi^{-1}(1-p), \tag{24}$$

with Φ^{-1} representing the inverse of the cumulative distribution function of a standard normal distribution. A high Z value, indicating the number of standard deviations the observed result is from the expected under the null hypothesis, suggests significant evidence against the null hypothesis, potentially heralding the discovery of new physical phenomena.

Following [4], the Profile Likelihood Ratio Test (PLRT) is used to estimate z. Denote the expectation value of the observed data n_i in a generic bin i by

$$E(n_i) = \mu s_i + b_i. \tag{25}$$

Here, b_i and s_i are the background and signal contributions, respectively, and μ is the signal strength parameter. In this framework, the null hypothesis (background only), H_0, is given by $\mu = 0$.
We can construct the likelihood function as a bin-by-bin product of Poisson distributions:

$$L(\mu) = \prod_{i=1}^{N} \frac{E(n_i)^{n_i}}{n_i!} e^{-E(n_i)} = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}. \tag{26}$$

The likelihood function term for bin i represents the probability to observe the recorded n_i events in bin i, within the mass interval $[m_i, m_{i+1}]$.
The profile likelihood ratio is given by

$$\Lambda(\mu = 0) = \frac{\mathcal{L}(\mu = 0)}{\mathcal{L}(\hat{\mu})} \tag{27}$$

and the test statistic q_0 becomes

$$q_0 = -2\ln[L(\mu = 0)] + 2\ln[L(\hat{\mu})]. \tag{28}$$

In the case where μ is the only free parameter, The bin-by-bin significance is given by $z^i_{PL} = \sqrt{q_0}$.

2.5 Machine Learning and Neural Networks

Machine Learning (ML) encompasses algorithms that enable performing specific tasks without using explicit instructions. At its core, ML is about constructing and studying systems that can learn from data to make predictions or decisions. It is broadly categorized into three learning methods:

- Supervised Learning: Involves learning a function that maps an input to an output based on example input-output pairs. Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, the learning algorithm optimizes a loss function $L(f(x_i), y_i)$ to find the best approximation of the function mapping inputs to outputs. For example, in regression problems, the Mean Squared Error (MSE) loss is commonly minimized:
$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2. \tag{29}$$

- Unsupervised Learning: Involves learning patterns from an unlabeled dataset $\mathcal{D} = \{x_i\}_{i=1}^N$. Most of the tasks include clustering and dimensionality reduction, where the aim is to discover the inherent structure of the data.

- Self-supervised Learning: Uses the input data to generate its own labels, which are then used to train the model.

Deep Learning algorithms NNs consist of neurons organized in layers. A neuron computes the weighted sum of its inputs and applies an activation function. The output of a neuron in a fully connected layer can be described as:

$$y = \sigma\left(\sum_j w_j x_j + b\right), \tag{30}$$

where σ is the activation function, w_j are the weights, x_j are the inputs, and b is the bias.

CNNs are specialized NNs for processing data with a grid-like topology featuring layers that perform convolutions. The convolution operation for a single filter k on an input image X is defined as:

$$S(i,j) = (k * X)(i,j) = \sum_m \sum_n X_{i+m,j+n} k_{m,n}, \tag{31}$$

where S is the feature map resulting from applying the filter k over X. CNNs learn to recognize spatial hierarchies in data by applying multiple filters that each learns to recognize different features, from simple edges in early layers to complex objects in deeper layers. For instance, in analyzing particle physics data represented as histograms, filters in a CNN can learn to identify specific shapes, peaks, or bumps indicative of underlying physical processes or particle interactions. The architecture typically comprises several convolutional layers, possibly embedded with pooling layers to reduce dimensionality, followed by fully connected layers that culminate in an output layer tailored for the specific task (e.g., classification).

3 ML-based anomaly detection methods

The complexity and magnitude of data make traditional methods for detecting anomalies insufficient, calling for novel and efficient, e.g., ML-based techniques. A detailed description of anomaly detection methods using ML can be found in [14] and reference therein. It distinguishes between four types of approaches:

- Supervised learning in which the input data, signal, and background are perfectly labeled.
- Semi-supervised learning in which either the signal or the background is labeled.
- Weakly-supervised learning in which the labels are imperfect.
- Unsupervised learning in which the input data is not labeled at all.

In one of the first attempts to employ ML for anomaly detection in a supervised manner [15], MC was used to train a classifier to discriminate between standard Quantum Chromodynamics (QCD) jets and a spectrum of potential signal jets, aiming to enhance the detection of new physical events. The method relies heavily on the accuracy of the simulations and the choice of signals for training. Thus, discrepancies between simulations and real data compromise the classifier's performance.

In [16], a classifier was trained to distinguish between impure samples mixing quark-jets and gluon-jets. An assumption was made according to which the proportion of events in either class is known. Even though per-instance labels were not available, a supervised task was established, leveraging the class proportions as the target.

In [17], the Classification WithOut Labels (CWola) framework was developed. It was shown that, for large enough samples, weakly-supervised training is as efficient as supervised one, provided that the fraction of signal and background events in the two datasets is different. The method was first used in an ATLAS search for di-jet resonances [18]. Two samples were created: signal-like, consisting of all events in the di-jet mass of interest, and background-like, consisting of events in the sideband around the mass of interest. The classifier was then trained to classify signal events and background events based on the mass of the two jets (which was assumed to be independent of the di-jet invariant mass). once signal events were selected, a standard side-band fit was performed to model the SM background and for the statistical inference. The method outperformed traditional inclusive searches while dedicated searches for, e.g., massive resonances decaying to two fat jets had a better sensitivity. In general, the method was shown to be effective mainly for narrow resonances,

and its performance drops for broader ones where signals could appear in both datasets. Its main limitation arises from possible correlations between the discriminating variables and the variable of interest.

The Anomaly Detection with Density Estimation (ANODE) [19] and Classifying Anomalies Through Outer Density Estimation (CATHODE) [20] attempt to address this limitation. In ANODE, a NN designed for density estimation is applied to interpolate probability densities from the sidebands to the signal region. This technique compares the interpolated probability density to the actual observed density in the signal region to construct a likelihood ratio for identifying potential signals of new physics. CATHODE, on the other hand, doesn't try to estimate the density of the signal directly. Instead, it uses ANODE for background estimation to generate synthetic background samples that are then mixed into the real data. Then, it uses the CWola classifier in order to distinguish between the synthetic background and the potential signal. Essentially, CATHODE skips a step compared to ANODE; it doesn't try to predict the background in the signal region directly but uses a machine learning model to learn what the signal might look like based on artificial samples.

Unsupervised (a.k.a self-supervised) method relies on dimensionality reduction or data compression to capture the essence of the data rather than the exact details. It can be useful for highlighting anomalies that stand out from the norm. Multiple studies in High Energy Physics (HEP) have employed Auto Encoders (AEs) to identify potential new physics by spotting outliers within the data [21–26]. For instance, in [27], this methodology has been applied to probe for new physics phenomena within the invariant mass spectrum of di-jet or jet-lepton pairs. The output from an AE is used to discriminate possible signals from the SM background, thus improving the signal-to-background ratio.

4 Bump Hunt DDP

The Bump hunt DDP proposed in [5] uses ML to rapidly and accurately map invariant mass distributions into statistical significance (z) distributions. It allows to infer efficiently the existence of excess of events in a given mass window, i.e., identify bumps. The NN is trained in a supervised manner for which pairs of input mass distribution (input data) and significance distributions (target) are generated. The input data aimed at representing realistic invariant mass distributions[1] and was randomly chosen from a variety of ten smoothly falling functions

$$\begin{aligned} &be^{-ax};\quad ax+b;\quad \frac{1}{ax}+b;\quad \frac{1}{ax^2}+b;\quad \frac{1}{ax^3}+b \\ &\frac{1}{ax^4}+b;\quad \frac{1}{ax^n}+b;\quad a(x-x^2)^2+y^2;\quad -a\ln(x)+b \\ &(y_1-y_2)\cos(a(x-b))+y_2;\quad \cosh(a(x-x^2))+b. \end{aligned} \tag{32}$$

The target, z_{PL}, is calculated using the PLRT as discussed in Section 2.4 and its maximal value denoted as z_{PL}^{max} used for performance studies. The trained NN is validated to confirm that its predictions are consistent and that its loss value converges. Finally, its predictions are evaluated on the test set, and its performance is discussed.

Dataset generation: The dataset generation procedure used in [5] is shown schematically in Figure 2. It consists of the following steps:

[1] In contrast to individual events, as in the [18–20]

- **Background Generation:** For each randomly selected function, the parameters a and b are defined such that each curve decays between two points, (x_1, y_1) and (x_2, y_2), where $x_1 < x_2$ are the centers of the extreme bins and $y_1 > y_2$ are randomized from the interval $[100, 10000]$. Here, x_1 and x_2 represent the minimum and maximum mass values (or bin number), and y_1 and x_2 stand for the minimum and maximum number of entries in each bin, i.e., DR.
- **Adding Fluctuations:** Poisson fluctuations are introduced to the background data to reflect fluctuations in real data
- **Signal Injection:** To the fluctuated background, Gaussian-shaped signals with a fixed width of three mass bins are added, emulating potential bumps. To improve the desired feature detection, the NN is trained with a data set containing signals with significance in the range [1,20] σ.

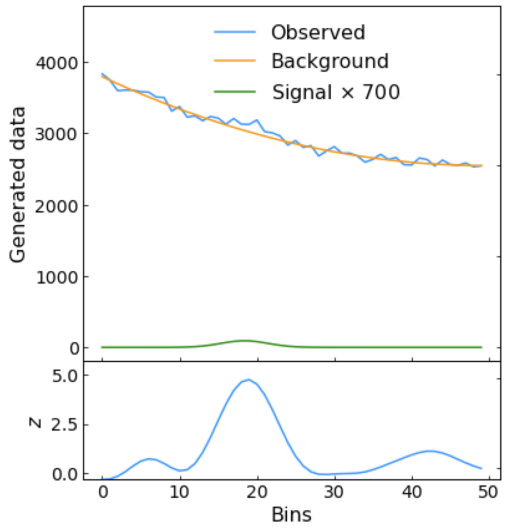

Figure 2: Illustration of the sample generation procedure. (top) A smoothly decaying background curve (orange) is generated over 100 bins. Each bin is assigned a Poisson fluctuation. A signal with significance relative to the fluctuated background (green) is added to it, producing the observed data (blue). (bottom) The corresponding significance distribution, z, is calculated analytically.

The resulting data is globally scaled to the interval $[0, 1]$ under a linear transformation and used to train a *Znet* architecture (Figure 3). It is based on a dense layer followed by six 1-dimensional convolutional layers. The latter is intended for feature detection, while the former is useful in suppressing position-dependent biases. A "rectified linear unit" activation function is used. The "Adam" optimizer is used to minimize the MSE loss function over 200 epochs at a learning rate of 0.0003 with a batch size of 100. The training sample comprises 600,000 histograms, 20% of which are used for validation, and 150,000 testing samples.

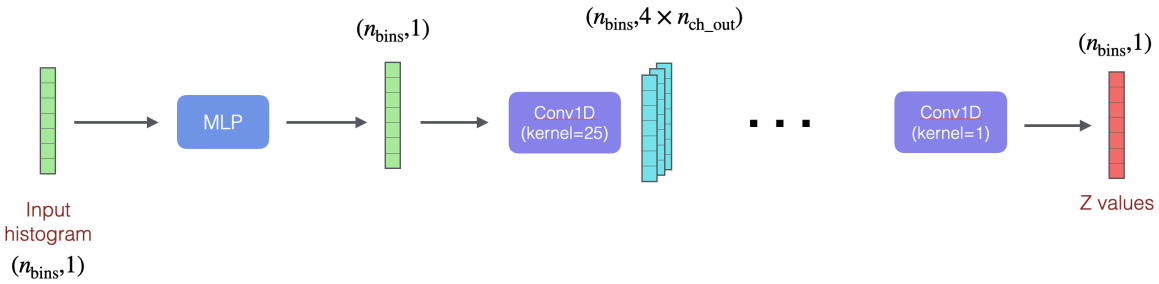

Figure 3: The *Znet* architecture: the input histogram is processed through a fully connected layer followed by multiple 1D convolutional layers. The output bin-by-bin histogram indicates the significance values.

Finally, the performance of the NN is determined on a testing data set by evaluating its ability to identify bumps with a significance of 3σ – the common definition of a "hint" for BSM physics.

5 Methods

The bump-hunt DDP proposed in [5] uses NN to rapidly map invariant mass distributions into statistical inference. Thus significantly reducing the time it takes to identify bumps in the data and allowing it to rapidly scan a large number of final states and selections. To prove the concept, it was shown that a simple, fully connected NN is capable of predicting accurately the bin-by-bin significance of bumps over smoothly falling backgrounds. However, the method was demonstrated only with synthetic (non-realistic) data and had several limitations. First, it only worked for distributions of a fixed number of mass bins. Second, it was used with a relatively narrow DR (100-10000 entries in each bin). Third, the background pdfs was given by analytical smoothly falling functions (all concave), with a Gaussian-shaped signal with a fixed 3-bin width, thus limited only for signal scenarios that fit these conditions.

In this work, we further generalize the bump-hunt DDP with a more advanced NN architecture and richer training dataset. We show that a single network can give accurate significance predictions for histograms with a varying number of bins, mass widths, and DRs. For background generated from smoothly falling analytical functions as well as simulation of real physics process and HEP data. The advanced NN architecture and training dataset are detailed below.

5.1 Znet3: A Rigorous Bump Hunter Architecture

In our recent study, we designed a uniquely structured CNN rather than a basic one to keep the information in every single input bin and its neighbors. It is also insensitive to properties (number of bins, DR, etc.) of the input mass distributions.

The key difference between the architecture of the NN (i.e. *Znet*, Figure 3) used in [5] and the one developed along this work (i.e. *Znet*3, Figure 4), is the order of the layers. In *Znet*, the first layer is fully connected, and the other layers are convolutional. Thus, the bin-by-bin information is lost at an early stage, and all the histograms must have the same number of bins. *Znet*3 begins with convolutional layers. Hence, the number of bins in the histogram can vary. Each convolutional layer has a different kernel size, allowing the identification of different data features - a small kernel to identify the bumps and a large kernel to address the smoothly falling background.

The outputs from the convolutional layers are concatenated before the connected layer, which maps them into a significant distribution. The decision to use skip connections is made to retain essential information from the input data throughout all the layers and ensure that no input characteristics are lost during the stages. Unlike *Znet*, *Znet*3 also uses the smooth background shape to provide an auxiliary prediction of the underlying background.

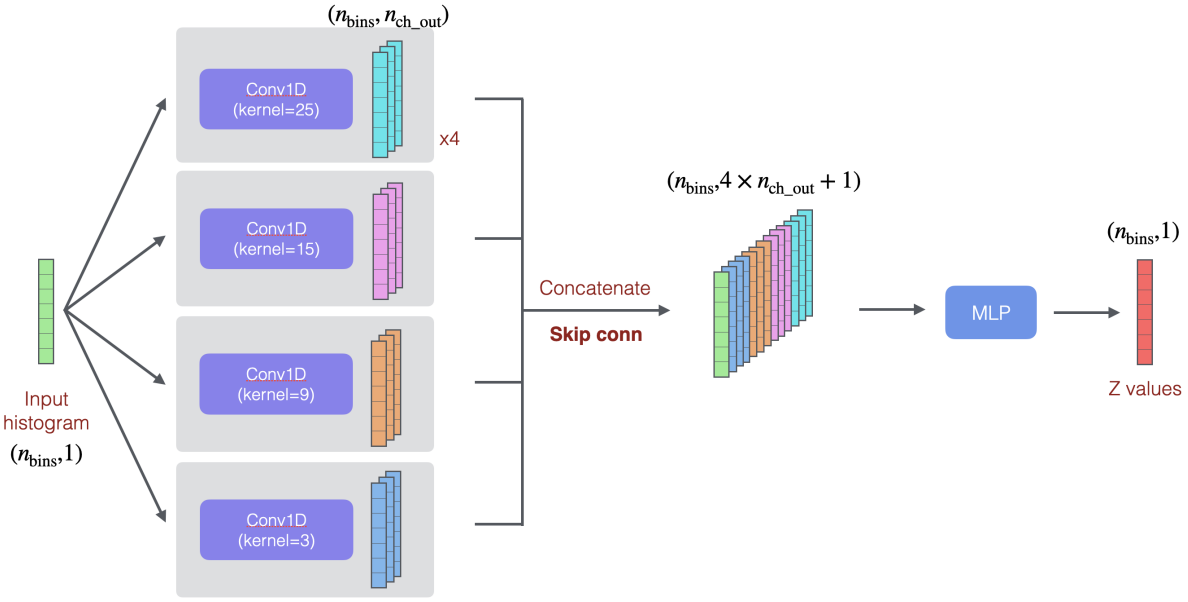

Figure 4: *Znet*3 architecture. This diagram illustrates the input histogram being processed through convolutional layers with varying kernel sizes. Each layer's output is concatenated with the input via skip connections, resulting in a unified distribution that outputs the final Z values.

An example Z_{pred} significance prediction of *Znet*3 overlaid on z_{PL} target is shown in Figure 5 (bottom). The input data distribution is shown as well (top) along with the smooth background and signal shapes used to generate it.

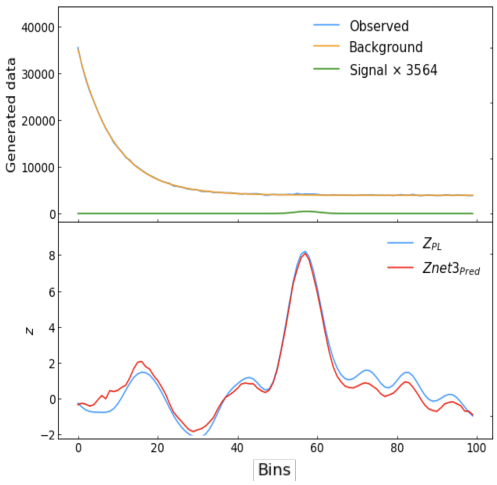

Figure 5: (top) The smooth background (orange) and signal (green) giving rise to a fluctuated data distribution (blue). (bottom) The NN significance prediction (red) and the PLRT significance (blue).

5.2 Training datasets

We generate the inputs of the NN as histograms of observed events, $N = B + S$. The number of bins in the histogram was in the range between 30 - 100 (the bin width reflecting a given detector resolution), and the number of entries in each been was in the range of 10 - 100,000. For all data types, the generation process followed the one described in Section 4.

5.2.1 Background shapes

Extended functions: As illustrated in Figure 6, in the original work, each configuration of number of bins and DR, gave rise to exactly 10 functional forms. Thus, many potential background shapes were left uncovered.

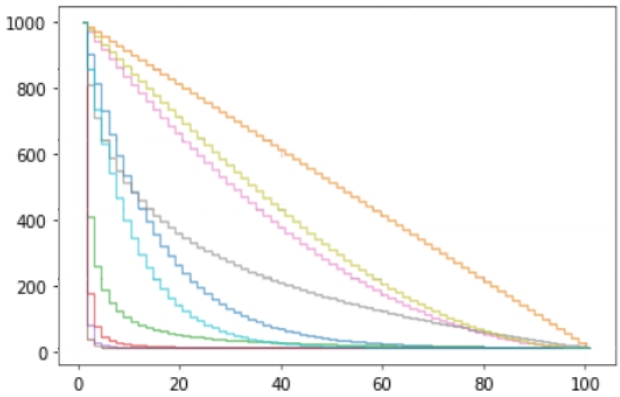

Figure 6: All ten possible functional forms for a given DR and number of histogram bins.

To enhance the training dataset, once the functional form is fixed, the function is cut at a random intermediate x value. The function is then stretched to fit the original range. The process is demonstrated in Figure 7. It results in a large variety of shapes, offering a more comprehensive training dataset, as seen in Figure 8.

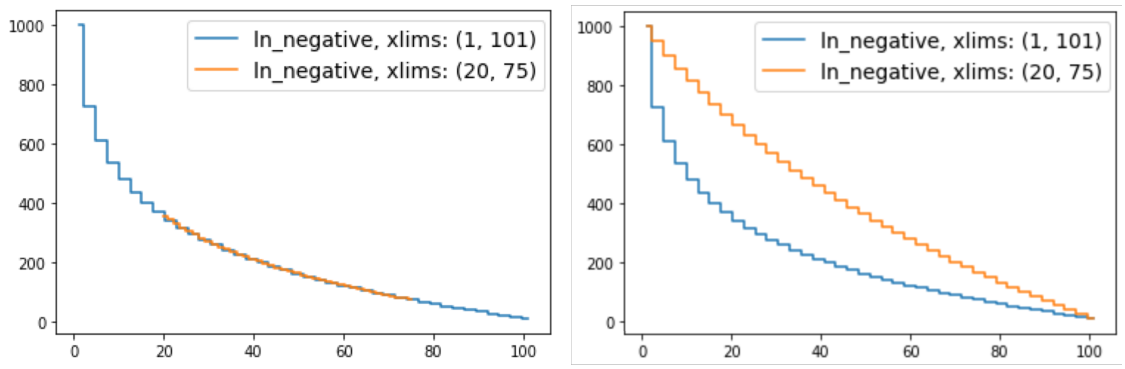

Figure 7: Left: a single functional form (blue) and the partial part used in the stretching procedure (orange). Right: The stretched (orange) and the original functional form (blue).

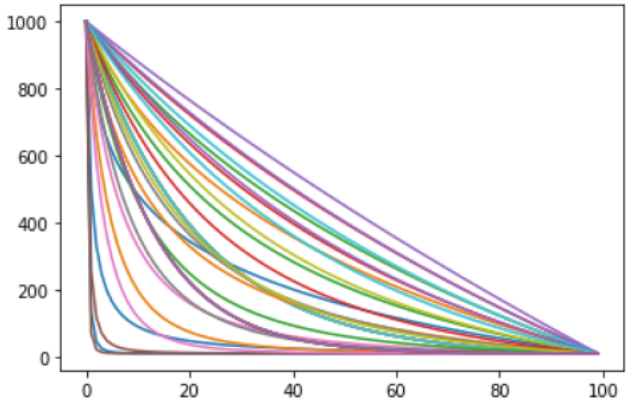

Figure 8: Example set of functional forms obtained for a given DR and number of histogram bins after the stretching procedure.

Dark Machine: Real data distributions are emulated using simulated data produced for a data challenge [28] in the context of the Dark Machines initiative (https://www.darkmachines.org/). The Dark Machines (DM) samples provide SM backgrounds for 10 fb^{-1} of pp collisions at 13 TeV, together with 15 signal benchmarks providing dark matter candidates. The simulated events are passed through a detector simulation using DELPHES3 [29], and the following objects are reconstructed if they satisfy some LHC-experiment standard kinematic selections: electrons (pT $>$ 15 GeV, $|\eta| < 2.5$), muons (pT $>$ 15 GeV, $|\eta| < 2.7$), jet (pT $>$ 20 GeV, $|\eta| < 2.8$) and photons (pT $>$ 20 GeV, $|\eta| < 2.37$). The jets can be b-tagged, meaning they are likely to originate from a b-parton.

DM samples are split into four channels, corresponding to four non-exclusive pre-selections:

- *channel 1*: HT $>$ 600 GeV, MET $>$ 200 GeV, MET/HT $>$ 0.2, $\geq$ 4 jets (pT $>$ 50) and $\geq$ 1 jet (pT $>$ 200GeV)
- *channel 2a*: MET $>$ 50 GeV, $\geq$ 3 leptons (electron or muon)
- *channel 2b*: HT $>$ 50 GeV, MET $>$ 50 GeV, $\geq$ 2 leptons (electron or muon)
- *channel 3*: HT $>$ 600 GeV, MET $>$ 100 GeV

In this work, we generated DM-based data distributions specifically from *channel 3*. The object definitions are refined to be closer to what can be performed on real data when one wants to maximize the use of available information. Two same-flavor opposite charge leptons with a mass in the range $m_Z = 91.18 \pm 15$ GeV are replaced by a Z-boson candidate[2]

Jets with a mass between 60 and 110 GeV are considered as hadronic V-boson, Vhad candidates (i.e., either a boosted W- or Z-boson decaying hadronically), while the ones with a mass between 110 and 200 GeV are tagged as top-quark candidates. Jets with mass larger than 200 GeV are particularly interesting in the context of anomaly detection as they could originate from boosted High Mass particles. They consequently have a dedicated object definition named HM jet.

Each pre-selected DM sample is divided into thousands of exclusive categories according to the numbers of electrons, muons, gammas, Z, Vhad, top, and HM jets. The categories having exactly two charged leptons are divided into two exclusive categories, depending on whether the leptons have the same sign of electric charge or not. The number of light ($m < 60$GeV) jets, b-tagged or not, is not used at this point. About 63000 categories are extracted from the DM dataset.

the data obtained in each selection is handled to avoid humps at the low mass region of the invariant mass distribution and to unify the expected width of narrow mass resonances. Considering the momentum resolution of each object included in the invariant mass calculation, the histograms are defined with varying bin sizes such that narrow mass resonances are described well by Gaussian of 1 bin width. The smoothly falling background distribution is obtained by fitting the histogram to the function $lna + bx + cx^2 + dx^2$.

HEP data:

- Di-Lepton: Dataset corresponding to real data was generated using the di-lepton mass background template provided in [2]. The functional form is given by:

$$f_{\ell\ell}(m_{\ell\ell}) = f_{BW,Z}(m_{\ell\ell}) \cdot (1 - x^c)^b \cdot x^{\sum_{i=0}^{3} p_i \log(x)^i}, \tag{33}$$

  where $x = m_{\ell\ell}/\sqrt{s}$, and the parameters b and p_i with $i = 0, ..3$ are left free in the fit to data and are independent for the di-electron and di-muon channels. The parameter c is set to 1 for the di-electron channel and 1/3 for the di-muon channel. The function $f_{BW,Z}(m_{\ell\ell})$ is a non-relativistic Breit–Wigner function characterized by $m_Z = 91.1876$ GeV.

- $H \to \gamma\gamma$: Higgs $\to \gamma\gamma$ data was extracted manually from the plot in the ATLAS discovery paper [9].

5.2.2 Targets

Signal Width: In the original work, a fixed width of 3 mass bins was assumed. In reality, the width of detected signals can vary due to detector resolution. Calibration efforts can reduce some of this variability, improving the uniformity and accuracy of signal detection. In this work, we have tested two possible solutions. First, train different NN for each signal width hypothesis: 1,2,3 and 4 mass bins. Second, train a NN with targets calculated to maximize the likelihood also with respect to the signal width. Concretely, for a given invariant mass histogram, the PLRT test was calculated with different signal widths, and the largest value was selected as a target in each mass bin.

[2]There can be several Z candidates in one event, as long as they do not share a lepton. In case one lepton can be assigned to two Z-boson candidates, the Z with the mass closest to m_Z is kept.

Dynamic Ranges: In the original work, the DR was limited to the range between 100 and 10,000 entries per bin. Considering real-data invariant mass distributions, e.g., [1, 2], this was broadened to the range between 10 and 100,000.

Number of mass bins: The dataset generation process involves creating histograms with varying numbers of bins, ranging uniformly from 30 to 100. This procedure is repeated multiple times to compile a comprehensive training dataset encompassing a wide range of histogram sizes.

5.3 Performance criteria

The NN performance was quantified by comparing Z_{pred} – the maximal predicted significance in each distribution, and Z_{pl} – the maximal significance calculated using the PLRT.

For each trained NN, we looked at the spread of $Z_{\text{pred}} - Z_{\text{pl}}$ as a function of different variables e.g., number of mass bins in the histogram, DR, etc. The performance was quantified by looking at the biases of these distributions and by their spread (Figure 9). We further looked at the 68% and 95% confidence interval obtained by the NN prediction and compared it to the one obtained with the PLRT.

6 Results

The list of different *Znet*3 NNs tested are summarised in Table 2. They differ by the mixture of training datasets and targets. All networks were trained with histograms with a number of bins ranging from 30 to 100 and with DR of 10 - 100,000. The training datasets consisted of ∼6.5M functions-based histograms and ∼6.5M DM-based histograms for a total of either ∼6.5M or ∼13M training datasets for the functions only and functions & DM NNs, respectively. For both cases, 90% of the datasets were used for the training and 10% for validation.

Table 2: Overview of Trained Neural Networks

NN Label	**background source**	**signal width**
F1w	functions	1
F2w	functions	2
F3w	functions	3
F4w	functions	4
Fmw	functions	mix 1-4
DMF1w	DM & functions (50/50)	1
DMF2w	DM & functions (50/50)	2
DMF3w	DM & functions (50/50)	3
DMF4w	DM & functions (50/50)	4
DMFmw	DM & functions (50/50)	mix 1-4

The NNs were tested on datasets generated from functions and DM samples that were or weren't used for the training and on samples generated from the HEP distributions. Several such datasets were generated for the specific tests described below.

A detailed evaluation of the DMFmw NN performance is given in Section 6.1. A comparison between the performance of NNs trained with different input datasets and targets is given in

Sections 6.2.1-6.2.2. A comparison between NNs performance on different test datasets is given in Sections 6.3-6.4.

6.1 DMFmw - performance

We first present the performance of DMFmw on a mixture of ∼1M function-based and ∼1M DM-based datasets. Out of the ∼1M DM-based datasets, half were generated from distributions (fitted functions) that were used in the training, and half were generated from distributions not used in the training. Signals were injected at all widths (1-4) and at significance in the range 0-10σ. As can be seen in Figure 9 (left), almost no bias is present in the NN prediction relative to the truth injected significance (μ=-0.02), and the spread is at the level of 1σ. This is also reflected in the confidence intervals (Figure 9, right), e.g., 5σ significance is expected to be detected by the NN at 65% in the interval between 4-6σ and at 95% in the interval between 3-7σ.

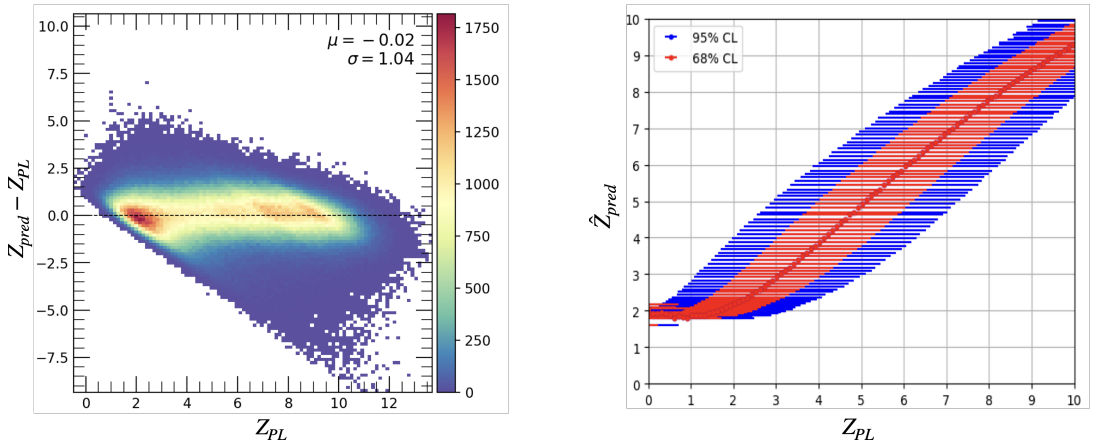

Figure 9: Left: The difference between $Z_{\rm pred}$ and $Z_{\rm pl}$ as a function of $Z_{\rm pl}$. Dense regions are shown in red and yellow, while sparse regions are shown in blue. Right: Relative to $Z_{\rm pl}$, the red and blue bars represent 68% and 95% confidence intervals, respectively.

Performance at different dynamic ranges: Three function-based test datasets were generated at different DRs: 10-1000, 10-10000, and 10-100000. Each of ∼1M distributions. Some degradation in performance is observed for the narrower dynamic ranges, as seen in Figure 10, suggesting that the NN performance could be further tuned, e.g., by enhancing the training dataset with samples of narrower dynamic range. This is left for future work.

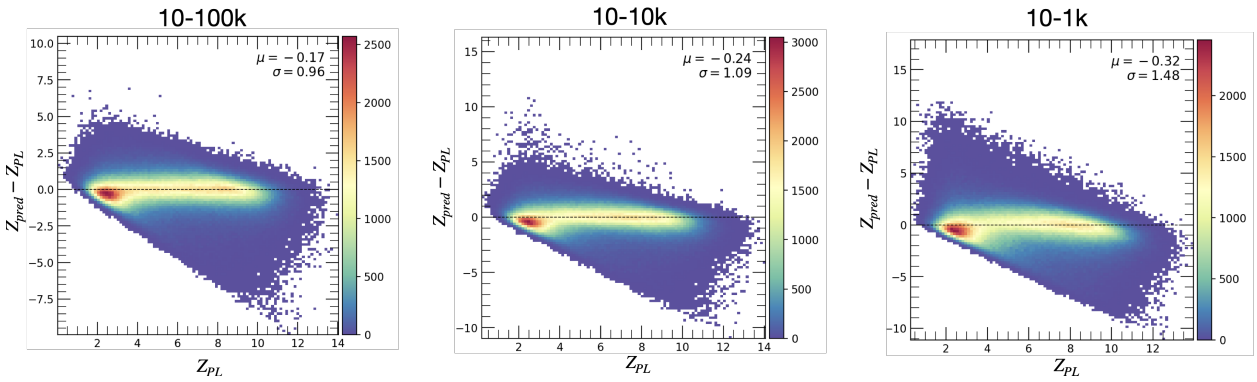

Figure 10: DMFmw performance on data with varying DRs. Left: 10-100000. Middle: 10-10000. Right: 10-1000.

Performance at different number of bins: Three function-based test datasets were generated with a different number of bins: 30 bins, 60 bins, and 100 bins. The NN performance is shown in Figure 11. Somewhat poorer performance in terms of both bias and spread is recorded for datasets with a small number of bins. This is attributed to the fact that the smaller the number of bins is, the broader the width of the signal relative to the smoothly falling distribution, making it harder for the NN to identify a bump over the background.

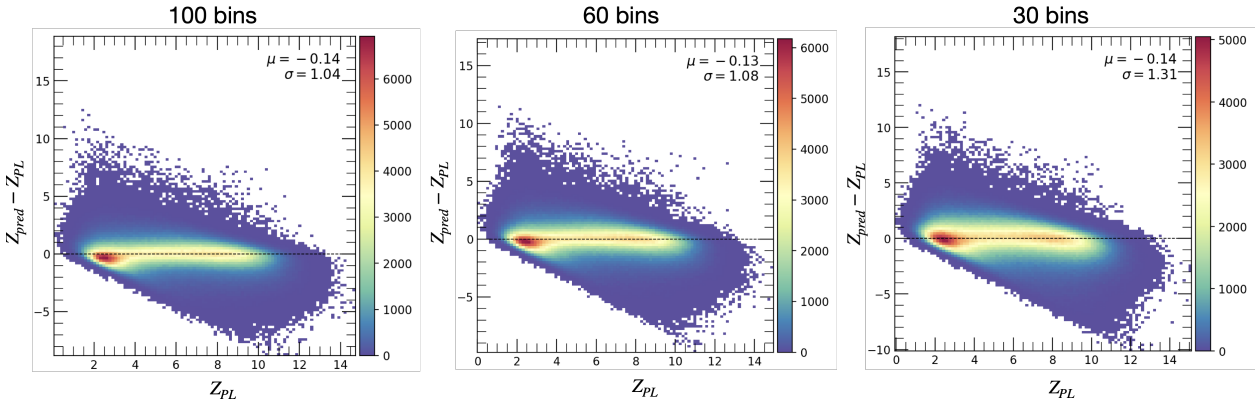

Figure 11: DMFmw performance on histograms with varying number of bins. Left: 100 bins. Middle: 60 bins. Right: 30 bin.

Performance at different signal width: ∼1M function-based test datasets were generated for each signal width (1-4) and for histograms with each number of bins (30,40,...,100). The performance of DMFmw in predicting the significance in all these datasets is summarized in Figure 12.

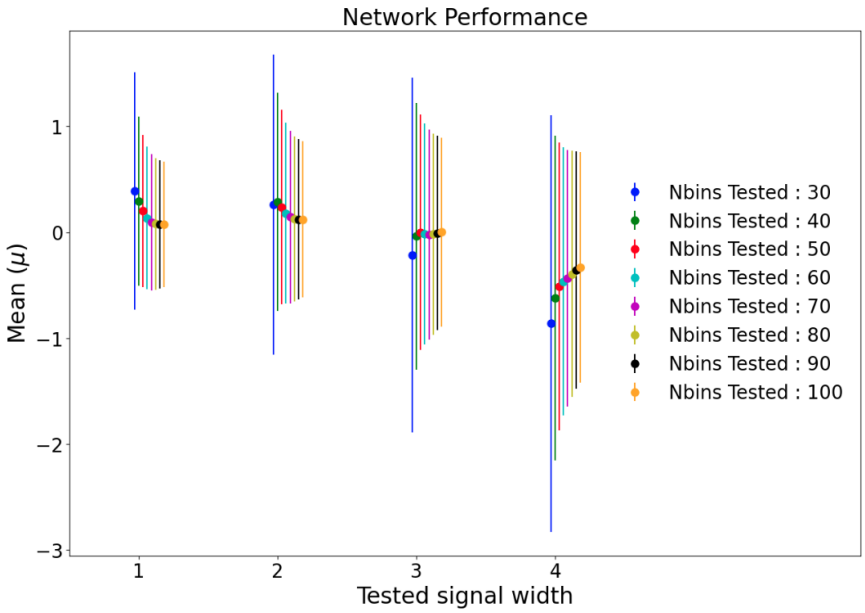

Figure 12: DMFmw performance on function-based dataset relative to the injected signal width and number of bins in the data histogram.

The largest biases are recorded for wide signals (4 bin width), indicating the challenging task of identifying wide signals. As shown before, somewhat poorer performance is recorded for datasets with a small number of bins.

Performance on DM data: Figure 13 summarizes the performance of DMFmw measured on datasets generated from DM samples with multiple injected signal widths. As seen in Figure 14, similar performance is recorded when measured on samples generated from distributions that were used in the training and from samples that weren't used in the training. These results pave the way towards constructing a NN whose predictions are trust-able also for distributions it was never trained on.

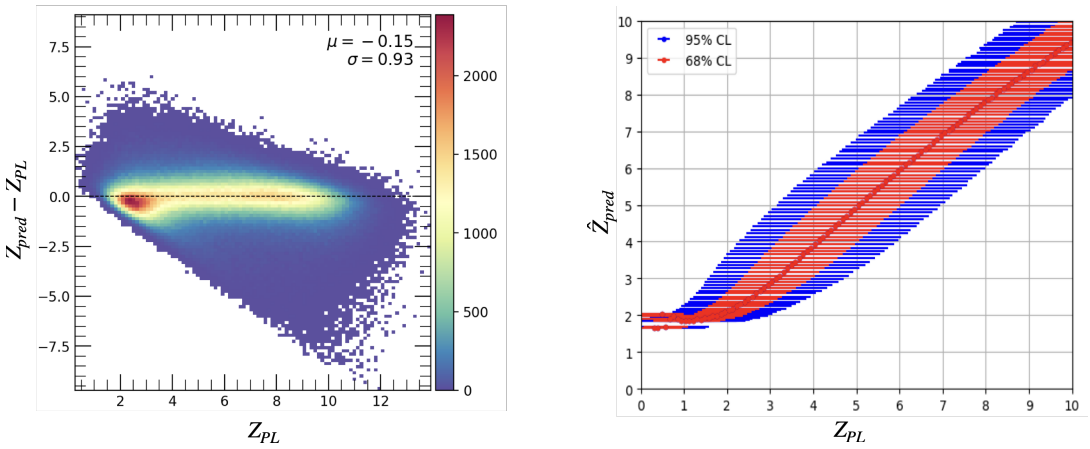

Figure 13: DMFmw performance on DM-based dataset.

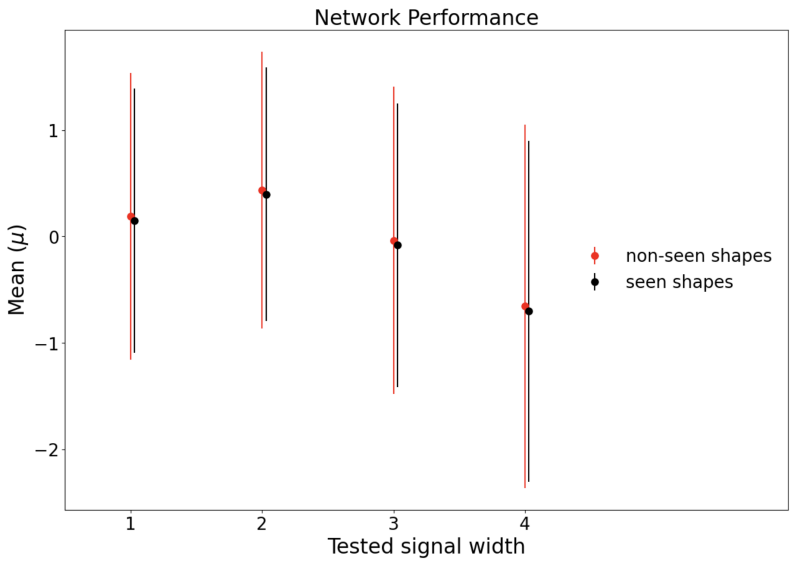

Figure 14: DMFmw performance on DM based datasets that were and weren't used for the training relative to different injected signal widths.

6.2 Comparing NNs performance on functions-based dataset

6.2.1 NNs trained with different input datasets

The performance of DMFmw is compared to that of Fmw in Figure 15. As can be seen, including DM data into the training process did not degrade the performance of DMFmw measured solely on the original 'functions' dataset. This suggests that the training dataset could be enriched without degrading the NN, albeit a larger training dataset and perhaps larger NN might be needed.

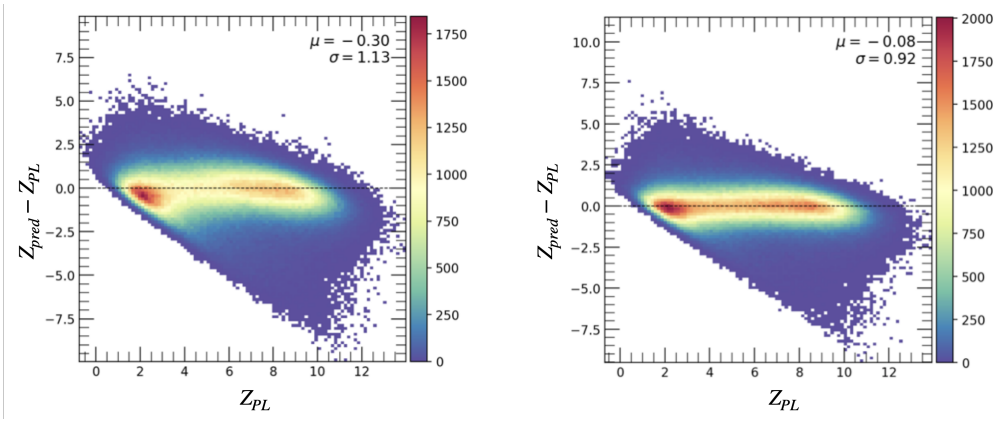

Figure 15: (left) Fmw and (right) DMFmw performance on functions-based dataset

6.2.2 NNs trained with different targets

The performance of DMFmw is compared in Figure 16 to that of DMF1w-DMF4w (of Table 2). As discussed in Section 5.2.2, the former was trained with a target calculated with the signal width as a free parameter in the likelihood fit and the latter with fixed signal widths.

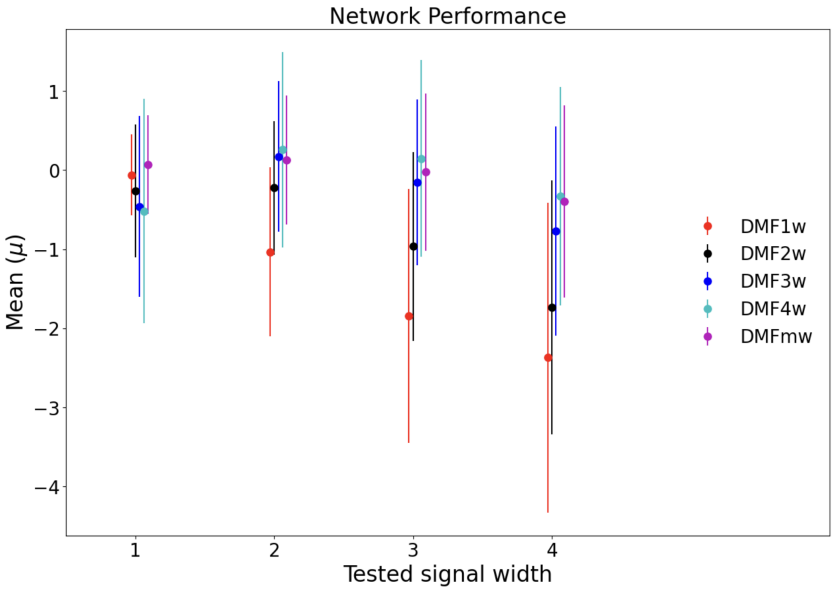

Figure 16: DMFmw, DMF1w-DMF4w performance on functions-based dataset relative to different injected signal width.

In all signal widths, DMFmw performs almost or exactly as well as the NN trained with the specific width. Moreover, it outperforms the fixed width NNs when tested with signal at other widths. For instance, when assessing data with a 3 signal width, DMFmw performs better than DMF1w. This trend is consistent across untrained regions.

6.3 Comparing NNs performance on DM-based data

Figure 17 demonstrates the necessity of training the NN also with DM data. As can be seen, Fmw prediction on data generated from DM samples is very poor. This suggests that a mixture of MC based and function based datasets will be needed in the future to accurately predict also in real data.

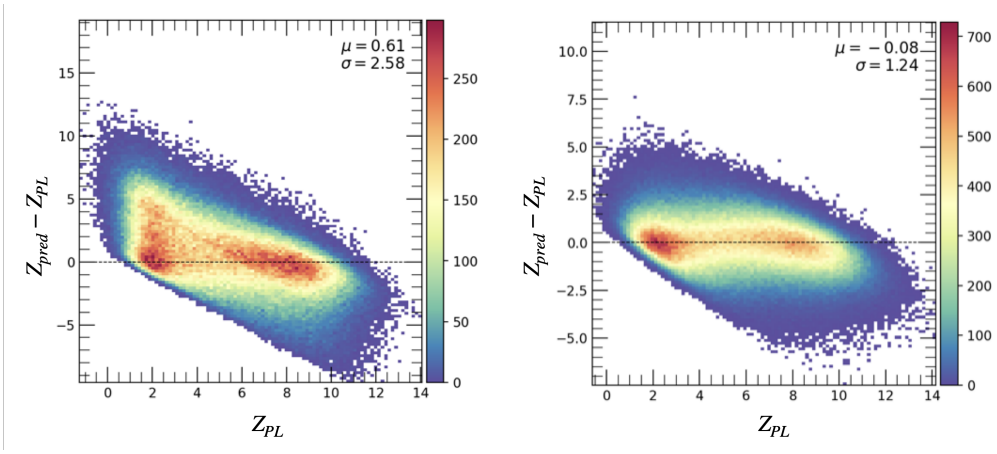

Figure 17: (left) Fmw and (right) DMFmw performance on DM-based dataset.

Figure 18 summarizes the comparison between networks that were trained with different inputs

of signal width and tested with DM test data. It shows a trend similar to the one observed with functions-based dataset (Figure 16).

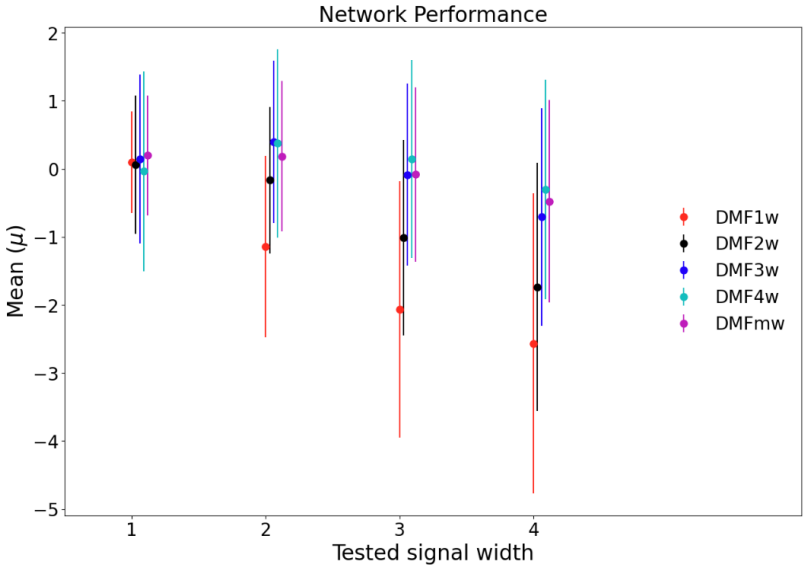

Figure 18: DMFmw, DMF1w-DMF4w performance on DM-based dataset relative to different injected signal width.

6.4 Comparing NNs performance on HEP data

6.4.1 $H \to \gamma\gamma$

The $H \to \gamma\gamma$ distribution was extracted from the plot in the Higgs discovery paper of ATLAS [9]

Figure 19 (lowest pad) shows the predicted significance of the Higgs signal from DMF1w-DMF3w and DMFmw NNs. To guide the eye, the background prediction (auxiliary material) is overlaid on the data points (top pad) and a Gaussian fit of the data after background subtraction is shown in the middle.

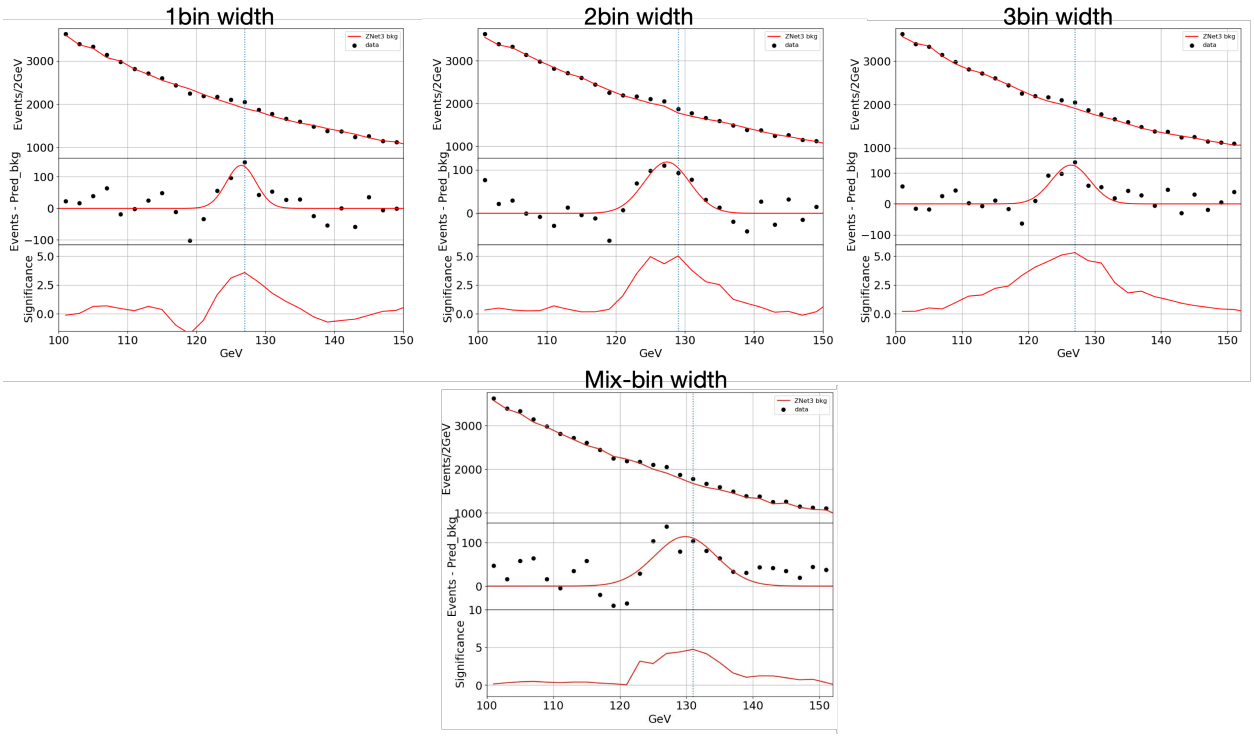

Figure 19: The $H \to \gamma\gamma$ data points overlaid with the background prediction (top panels) as well as the background subtracted data points and the Gaussian fit (middle panels) and Z_{pred} (bottom panels). All shown for DMF1w-DMF3w (top row) and DMFmw (bottom row).

Figure 20 shows the prediction of DMF1w (right) along with the official results of ATLAS (left) [9]. The NN predicts the Higgs with a significance of $Z_{Pred} = 4.6$, while its target $Z_{PL} = 3.7$. The latter was calculated relative to the background shape extracted from the plot of ATLAS. We stress that DMFmw predicted $Z_{Pred} = 4.9$, slightly worse than DMF1w. This could be expected given that the signal width assumed in the ATLAS paper was of 1 bin.

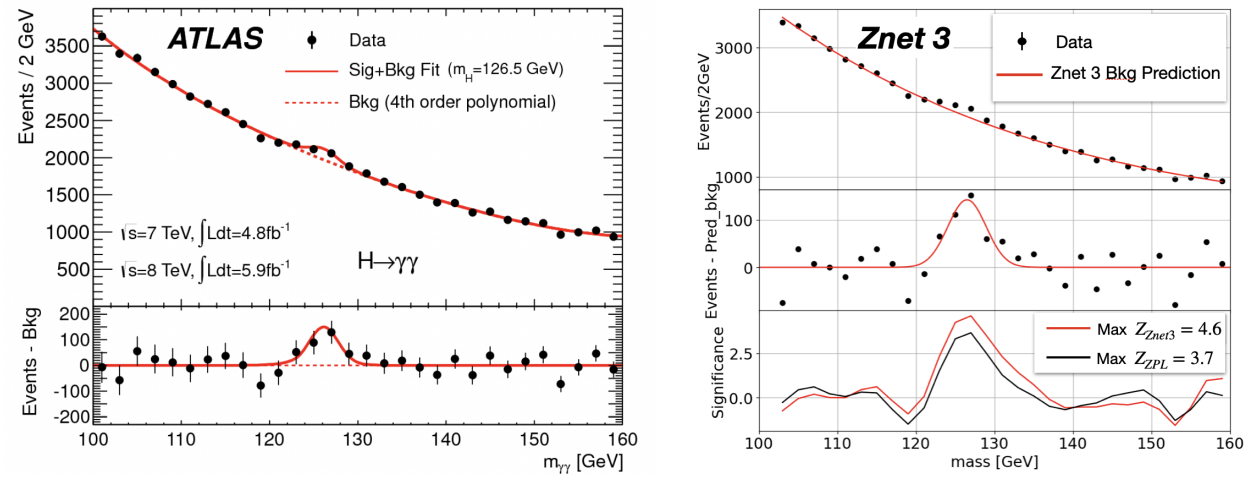

Figure 20: $H \to \gamma\gamma$: ATLAS results (Left) and DMF1w predictions (Right). ATLAS results are taken from [9].

6.4.2 Di-Lepton

We generated 320K di-lepton histograms with varying a number of bins from 30 to 100, and DR between 10 - 100,000 as described in 5.2.1. Signals were injected at all widths (1-4) at significance in the range 0-10 σ.

Figure 21 details the performance of DMFmw on this dataset. Relative to the performance on both function- and DM-based dataset, the performance of DMFmw on the di-lepton dataset is broader and show a larger bias.

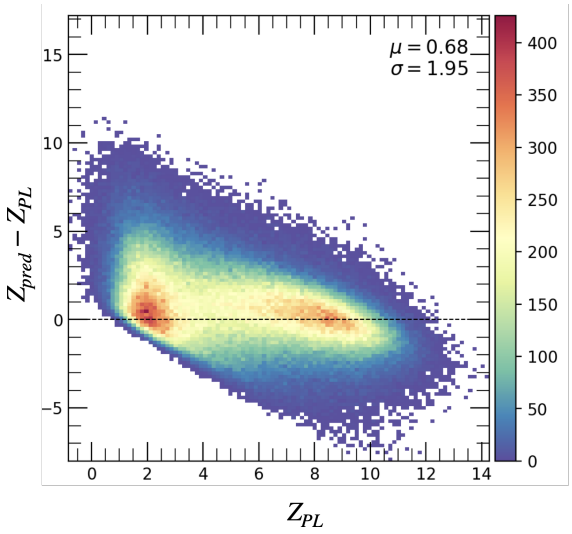

Figure 21: DMFmw performance on di-lepton-based dataset.

While this result is not yet clear (and contradicts our observation in 6.2.1), we have noticed that the addition of DM-based distributions into the training dataset degrades the NN performance on the di-lepton dataset. This is illustrated in Figure 22 where the relative fraction of DM-based distributions in the training dataset varies. Further investigations are left for future work.

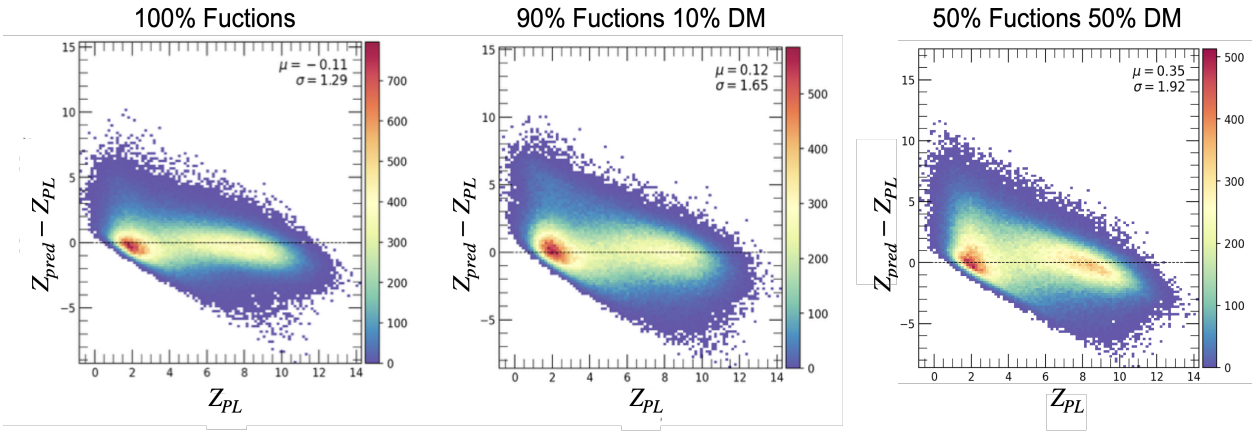

Figure 22: The performance on di-lepton-based samples of NNs trained with different fraction of DM samples in the training dataset. Left: 0%. Middle: 10%. Right: 50%.

7 Conclusions & Future Directions

This thesis presents the progress made in the development of bump-hunt DDP. An improved NN architecture as well as enhanced training dataset were combined to form a NN that provides a robust significant prediction over a broad range of invariant mass distributions. This should allow scanning efficiently large portion of the data in search for bumps and identify regions in which the data itself exhibits deviation from the SM prediction.

We have shown that a NN trained with a mixture of input distributions generated from a set of smoothly falling functions and DM samples provides precise prediction when tested with a dataset generated from smoothly falling functions, DM samples that were and weren't used in the training itself. When the targets were calculated with a mixed width assumption, the performance of the NN had only minor dependency on the width of the injected signal (Figure 16). Poorer performance was measured with distributions generated from HEP data (Figure 21), suggesting that the training dataset should be further enriched.

As expected,NN that was trained only with smoothly falling functions performed well on these functions but worse on data generated from DM distributions. (Figure 17). NN trained with a fixed signal performed very slightly better on distributions containing signals at the same width and worse on datasets generated with signals of other widths (Figure 16).

Future work should focus on refining the input training dataset. In particular, we consider using Variational Auto Encoder (VAE) to enhance the variety of functional shapes and use simulation to generate more examples of data-like distributions. The DDP search for bumps in the ATLAS run-II data is at a preliminary stage. We plan to complete it and reveal as many potential signals in the data as possible. A key ingredient is developing a robust method to evaluate the systematic uncertainty associated with the NN prediction.

References

[1] G. Aad et al., *Search for new resonances in mass distributions of jet pairs using 139 fb^-1 of pp collisions at $\sqrt{s}$ = 13 tev with the atlas detector*, *Journal of High Energy Physics* **2020** (2020) 145.

[2] G. Aad et al., *Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* **796** (2019) 68 [1903.06248].

[3] G. Aad and others (ATLAS Collaboration), *Search for heavy higgs bosons decaying into two tau leptons with the atlas detector using pp collisions at $\sqrt{s} = 13$ tev*, *Phys. Rev. Lett.* **125** (2020) 051801.

[4] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *The European Physical Journal C* **71** (2011) .

[5] S. Volkovich, F. De Vito Halevy and S. Bressler, *A data-directed paradigm for bsm searches: the bump-hunting example*, *The European Physical Journal C* **82** (2022) .

[6] Y. Grossman and Y. Nir, *The Standard Model: From Fundamental Symmetries to Experimental Tests*. Princeton University Press, 10, 2023.

[7] UA1 collaboration, *Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV*, *Phys. Lett. B* **122** (1983) 103.

[8] P. M. Watkins, *DISCOVERY OF THE W AND Z BOSONS*, *Contemp. Phys.* **27** (1986) 291.

[9] ATLAS Collaboration collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys.Lett.* **B716** (2012) 1 [1207.7214].

[10] S. Chatrchyan et al., *Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc*, *Phys. Lett. B* **716** (2012) 30.

[11] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*. Addison-Wesley, Reading, USA, 1995.

[12] G. Perez, "Lhc notes." https://indico.cern.ch/event/341178/contributions/795892/attachments/669829/920774/LHC_Gilad_Perez_final.pdf, 2014.

[13] S. S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, *The Annals of Mathematical Statistics* **9** (1938) 60 .

[14] V. Belis, P. Odagiu and T. K. Aarrestad, *Machine learning for anomaly detection in particle physics*, *Reviews in Physics* **12** (2024) 100091.

[15] J. A. Aguilar-Saavedra, J. Collins and R. K. Mishra, *A generic anti-qcd jet tagger*, *Journal of High Energy Physics* **2017** (2017) .

[16] L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly supervised classification in high energy physics*, *Journal of High Energy Physics* **2017** (2017) .

[17] E. Metodiev, B. Nachman and J. Thaler, *Classification without labels: learning from mixed samples in high energy physics*, *Journal of High Energy Physics* **2017** (2017) 174.

[18] J. H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038.

[19] B. Nachman and D. Shih, *Anomaly detection with density estimation*, *Physical Review D* **101** (2020) .

[20] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel et al., *Classifying anomalies through outer density estimation*, *Phys. Rev. D* **106** (2022) 055006.

[21] T. Cheng, J.-F. m. c. Arguin, J. Leissner-Martin, J. Pilette and T. Golling, *Variational autoencoders for anomalous jet tagging*, *Phys. Rev. D* **107** (2023) 016002.

[22] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, *Phys. Rev. D* **101** (2020) 075021.

[23] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty detection meets collider physics*, *Phys. Rev. D* **101** (2020) 076015.

[24] T. Heimel, G. Kasieczka, T. Plehn and J. M. Thompson, *QCD or what?*, *SciPost Phys.* **6** (2019) 030.

[25] A. Blance, M. Spannowsky and P. Waite, *Adversarially-trained autoencoders for robust unsupervised new physics searches*, *Journal of High Energy Physics* **2019** (2019) .

[26] T. S. Roy and A. H. Vijay, *A robust anomaly finder based on autoencoders*, 2020.

[27] A. Collaboration, *Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ tev with the atlas detector*, *Phys. Rev. Lett.* **132** (2024) 081801 [2307.01612].

[28] T. Aarrestad, M. van Beekveld, M. Bona, A. Boveia, S. Caron, J. Davies et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, *SciPost Phys.* **12** (2022) 043.

[29] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [1307.6346].