PROCEEDINGS OF SCIENCE

Automatizing the search for mass resonances using BumpNet

Jean-François Arguin,[1] **Émile Baril,**[1] **Ilan Bessudo,**[2] **Maryna Borysova,**[2] **Shikma Bressler,**[2] **Samuel Calvet,**[3] **Shalini Epari,**[1] **Ethan James Meszaros,**[1] **Elad Kliger,**[2] **Michael Kwok Lam Chu,**[2] **Hoang Dai Nghia Nguyen,**[1] **Julien Noce Donini,**[3] **Eva Mayer,**[3] **Bruna Pascual,**[3] **Amit Shkuri**[2] **and Muhammad Usman**[1]

[1]*Université de Montréal*

[2]*Weizmann Institute of Science*

[3]*Laboratoire de Physique de Clermont – UCA – CNRS/IN2P3*

E-mail: kwok.lam.chu@cern.ch

Modern experimental physics research, particularly in particle physics, requires extensive data analysis efforts to identify significant signals indicative of new physics. We present BumpNet, a novel Neural Network (NN) architecture designed to conduct model-independent searches for mass bumps arising from new physics phenomena. This model maps invariant mass histograms into statistical inference distributions to facilitate efficient signal detection. By focusing on experimental data without relying on simulations, BumpNet enables the identification of exclusive selections that significantly deviate from the Standard Model's known properties, marking them for further study. The NN minimizes resource-intensive tasks such as background estimation and systematic uncertainty evaluation, enabling rapid testing of multiple final states with only minor sensitivity loss compared to standard likelihood-based methods.

The model's performance is validated using training data from the Dark Machines dataset, with its predicted significance benchmarked against an ideal likelihood analysis. The results demonstrate negligible bias and variance below 1σ when tested on Gaussian-shaped signals. Furthermore, BumpNet's consistency is evaluated using data from the ATLAS Higgs discovery, reinforcing its reliability and applicability in real-world analyses.

Reference paper: JHEP02(2025)122

13th Large Hadron Collider Physics Conference
5 - 9 May 2025
National Taipei University, TaiWan

PoS(LHCP2025)175

1. Introduction

Despite the success of the Standard Model (SM) in describing elementary particles and their interactions, it is not yet a complete theory [1]. The search for signals beyond the Standard Model (BSM) is one of the main research directions at the Large Hadron Collider (LHC) at CERN. However, only a small fraction of possible final states has been systematically analyzed for new resonances [2]. Despite significant efforts, traditional analysis methods require substantial time and resources, which limits the number of final states and selections that can be tested.

The **Data Directed Paradigm (DDP)** [3] is a search strategy designed to efficiently identify regions of interest in the data. It leverages a theoretically well-established property of the SM combined with a tool specifically developed to detect deviations from this property. This enables rapid and systematic exploration of numerous final states and selection criteria in the search for BSM physics.

In this document, we present **BumpNet**, a neural network (NN) that extends the DDP. Instead of requiring explicit background modeling, BumpNet learns from synthetic and simulated data to predict, for each histogram bin, the statistical significance of a possible excess. This approach enables large-scale, model-independent resonance searches across thousands of histograms, offering both speed and sensitivity.

2. Methodology

BumpNet is designed to process invariant-mass histograms and predict the statistical significance of resonant signals as an excess of events (a bump) based on the likelihood ratio (LR) test. The NN is trained in a supervised manner using generated training and testing datasets. BumpNet outputs a z value for each bin, indicating where and how likely it is that the bin contains a bump.

2.1 Network Architecture

The architecture of the NN is illustrated in figure 1. The input to the network is the bin content of the histogram, normalized to the range [0,1], in the form of a tensor of dimensions (n_{bins}, 1), where n_{bins} is the variable number of bins in the invariant mass histogram. This input tensor is processed in parallel by four convolutional layers with kernel sizes of 3, 9, 15, and 25, enabling the model to capture features at different scales, from local relationships to broader smoothly falling background patterns. The outputs from the convolutional stacks are concatenated and combined with a skip connection from the raw input. The resulting representation of shape (n_{bins}, $4 \times 64 + 1$) is then fed into a multilayer perceptron (MLP) with three hidden layers (ReLU activations; 128, 64, and 32 neurons) and a final output layer providing a per-bin significance estimate.

The network is trained using the Adam optimizer [4] to minimize the mean squared error (MSE) loss function against the true significance derived from likelihood ratio tests.

2.2 Dataset preparation

The training and testing datasets are generated by injecting synthetic Gaussian signals into two types of background invariant-mass distributions:

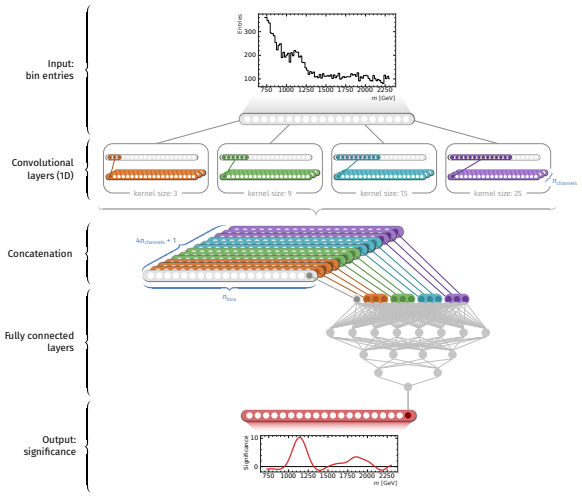


Figure 1: The BumpNet architecture.

- **Analytical functions**: eleven smoothly falling functions, shown below.

$$be^{-ax}, \quad ax+b, \quad \frac{1}{ax}+b, \quad \frac{1}{ax^2}+b, \quad \frac{1}{ax^3}+b,$$
$$\frac{1}{ax^4}+b, \quad a\left(x-x_2\right)^2+y_2, \quad -a\cdot\ln\left(x\right)+b,$$
$$\left(y_1-y_2\right)\cos\left(a\left(x-b\right)\right)+y_2, \quad \cosh\left(a\left(x-x_2\right)\right)+b.$$

  The parameters a and b are defined such that each curve decays between two points (x_1, y_1) and (x_2, y_2), which are randomly selected as the two endpoints of an invariant-mass histogram.

- **Simulation-based backgrounds**: histograms derived from the Dark Machines (DM) dataset [5], which models 10 $\text{fb}^{-1}$of LHC collisions at $\sqrt{s} = 13\,\text{TeV}$ across 26 Standard Model processes. Each process is weighted according to its cross-section, resulting in a realistic dataset that simulates data produced at the LHC.

To simulate realistic statistical fluctuations, both the signal and background histograms are Poisson-fluctuated bin by bin.

2.3 Histogram Processing

The DM dataset provides basic objects such as electrons, muons, photons, and jets. To approximate realistic LHC analyses, additional objects are constructed from these based on their kinematic properties. These include leptonic Z, boosted hadronic W/Z, boosted tops, and high-mass jets. Additional kinematic requirements on missing transverse energy ($E_{\text{T}}^{\text{miss}}$) and leading-object momentum are imposed, and events are split according to jet multiplicity.

An invariant-mass histogram is created for every combination of at least two objects within a defined subcategory. The invariant mass is calculated as the magnitude of the sum of the four-momenta of the selected objects.

Each histogram is then fitted with a 4th-order log-polynomial function, $\ln y = \sum_{n=0}^{4} P_n x^n$, to create a smooth background histogram, which is refered to the "mother" histogram in later context.

Rebinning procedures are applied to reflect detector resolution: the bin width scales with the quadratic sum of individual object resolutions, making signal peaks approximately one bin wide on the invariant-mass histogram. This ensures BumpNet's sensitivity to potential narrow resonances despite varying kinematics.

3. Results

The accuracy of BumpNet's predicted significance (z_{pred}) is compared to the true significance (z_{LR}), obtained from the likelihood ratio based on the known background shape of the "mother" histogram. The difference ΔZ_{max}, defined as the difference between the maximum values of z_{pred} ($Z^{\text{pred}}_{\text{max}}$) and z_{LR} ($Z^{\text{LR}}_{\text{max}}$) within a given histogram, is evaluated.

3.1 Performance over nominal histograms

BumpNet is first validated using the dataset prepared with the same procedures described in Sections 2.2 and 2.3. The validation histograms are generated using the same "mother" distributions but were not used during training.

On the synthetic test sets with injected Gaussian signals, the accuracy ΔZ_{max} is shown in Figure 2. For both background shapes generated from analytical functions and those from the DM dataset, BumpNet achieves unbiased predictions of maximum significance. The spread of ΔZ_{max} is about 0.6–0.7 σ, indicating high precision.

To evaluate BumpNet's ability to generalize to entirely new background shapes, additional tests are performed with histograms generated from background models not used during training. One such test uses the ATLAS dilepton resonance search [6]. The background model of this analysis is used to generate a BumpNet test dataset. Histograms with 98 and 28 bins are produced for the di-electron and di-muon distributions, respectively. Gaussian signals with a 1-bin width and statistical significance in the range of 1–10 σ are injected into the fluctuated distributions following the procedure detailed in Section 2.2. The results are evaluated using ΔZ_{max}, and the plot for the di-electron case is shown in Figure 2. The performance on the di-electron distributions is excellent, with no bias and small variance in ΔZ_{max}. The di-muon case performs slightly worse due to the smaller number of bins in its histograms.

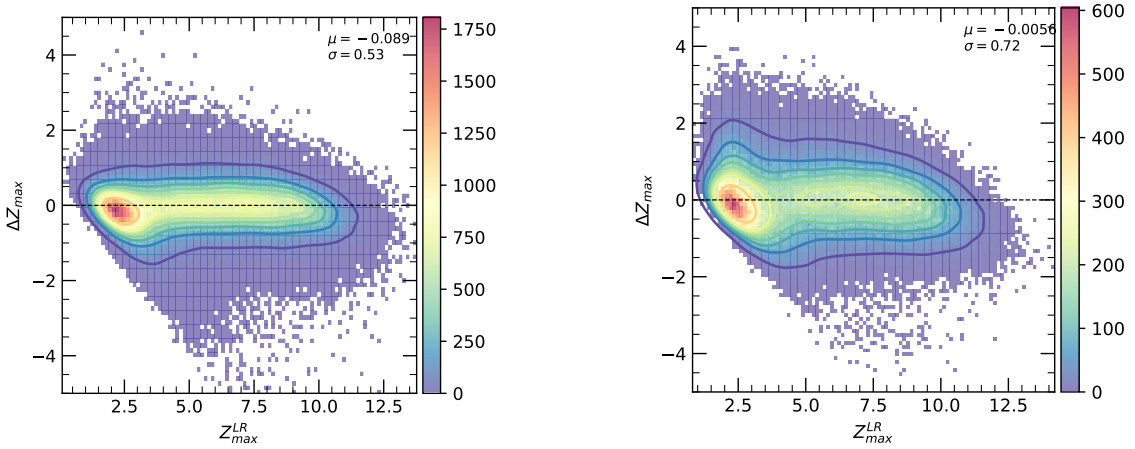
3.2 Performance over data and data-like signals

BumpNet has been applied both to realistic HEP data in the $H \to \gamma\gamma$ channel and to simulated BSM signals within the DM framework.

The data points from the $H \to \gamma\gamma$ histogram were extracted from the ATLAS Higgs discovery paper [7]. The performance of BumpNet's z_{pred} is shown in Figure 3, where the Higgs signal is clearly visible with a predicted significance of 4.5 σ, consistent with the published 4.2 σ.

For realistic BSM scenarios, a set of benchmark signals has been selected, including 600 GeV leptoquarks [8], a 50 GeV Z' boson [5], 1 TeV RPV stops [5], and a 1.5 TeV W' [9]. In all cases, BumpNet identifies significant bumps at the expected masses. The significance predictions for the first two signals are shown in Figure 3.

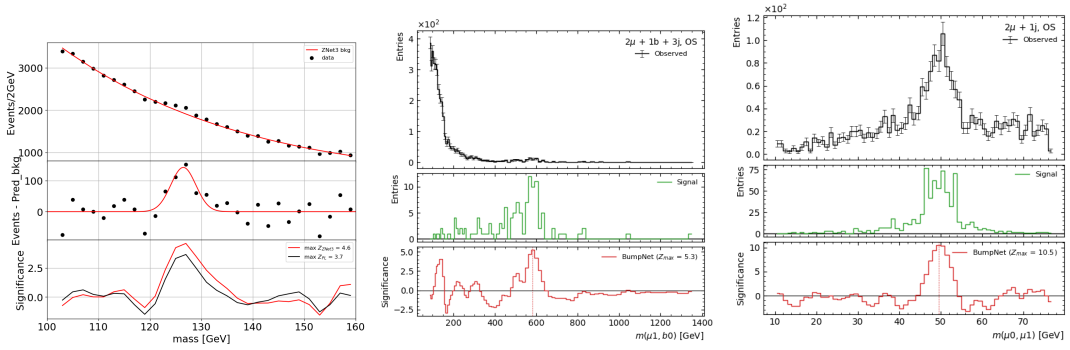
To mitigate the look-elsewhere effect, BumpNet outputs can be combined across statistically independent histograms. Observing correlated bumps at the same mass in multiple categories



(a) Function background shapes

(b) ATLAS di-electron background shapes

Figure 2: Distributions of ΔZ_{max} as a function of $Z^{\text{LR}}_{\text{max}}$ over testing data with backgrounds from (a) analytical functions and (b) ATLAS di-electron analysis.



(a) $H \to \gamma\gamma$

(b) Two $LQ \to b\mu$

(c) $Z' \to \mu\mu$

Figure 3: BumpNet's significance prediction over realistic signals.

strongly disfavors random fluctuations. A dedicated Global Analysis Algorithm (GAA) is under development to quantify global significance across thousands of histograms.

4. Conclusion

BumpNet provides a practical tool for resonance searches at the LHC. Its performance was benchmarked against that of an ideal analysis using the likelihood ratio test with perfect knowledge of signal and background shapes. The predictions show negligible to small biases and variance below 1 σ when tested on both Gaussian-shaped signals and more realistic signals from simulation. These results validate BumpNet's adaptability and robustness for challenging analysis scenarios, highlighting its promise for advancing signal detection in future high-energy physics applications.

References

[1] S. Weinberg, *Essay: Half a Century of the Standard Model*, Phys. Rev. Lett. **121** (22 2018) **page** 220001, URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.220001.

[2] ATLAS Collaboration, *ATLAS Public Results on Searches for New Phenomena*, [Online; accessed 19-December-2024], 2024, URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SearchesPublicResults.

[3] 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), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-022-10215-1.

[4] D. P. Kingma **and** J. Ba, *Adam: A Method for Stochastic Optimization*, 2017, arXiv: 1412.6980 [cs.LG], URL: https://arxiv.org/abs/1412.6980.

[5] T. Aarrestad **andothers**, *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, SciPost Physics **12** (2022), ISSN: 2542-4653, URL: http://dx.doi.org/10.21468/SciPostPhys.12.1.043.

[6] G. Aad **andothers**, *Search for high-mass dilepton resonances using 139 fb1 of pp collision data collected at s=13 TeV with the ATLAS detector*, Physics Letters B **796** (2019) **page** 68, ISSN: 0370-2693, URL: https://www.sciencedirect.com/science/article/pii/S0370269319304721.

[7] G. Aad **andothers**, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012) **page** 1, ISSN: 0370-2693, URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[8] I. Doršner **and** A. Greljo, *Leptoquark toolbox for precision collider studies*, Journal of High Energy Physics **2018** (2018), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP05(2018)126.

[9] J. de Favereau **andothers**, *DELPHES 3: a modular framework for fast simulation of a generic collider experiment*, Journal of High Energy Physics **2014** (2014), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP02(2014)057.

PoS(LHCP2025)175