Anomalous Jet Identification via Sequence Modeling

Alan Kahn, Julia Gonski, Inês Ochoa, Daniel Williams, Gustaaf Brooijmans

14 July 2021

DPF Annual Meeting

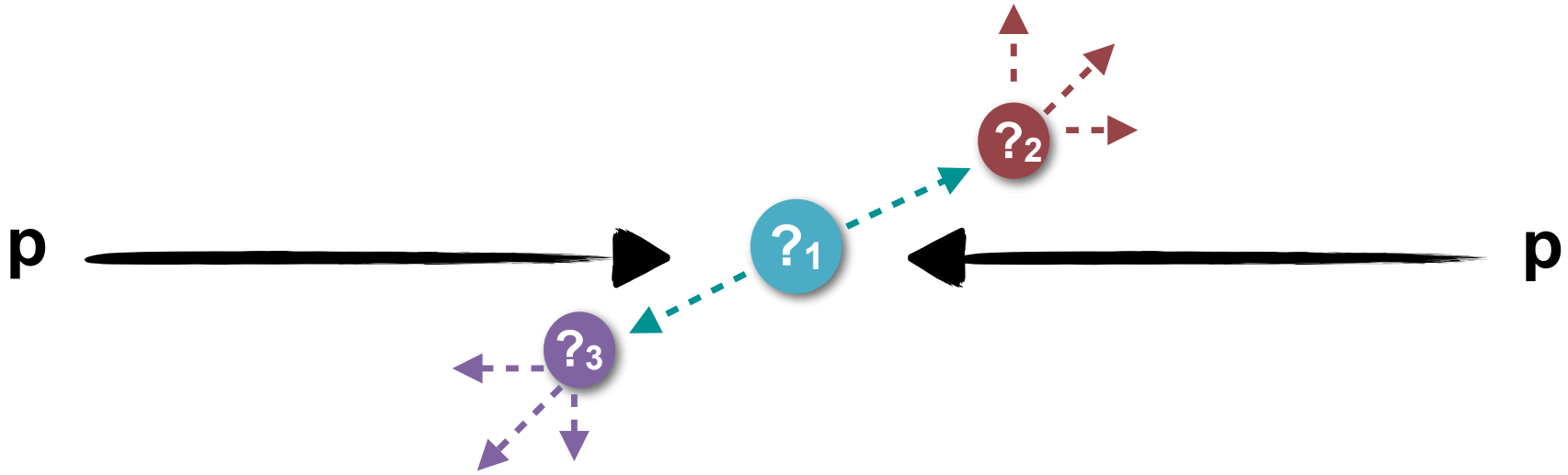
Outline

- Motivation: anomaly detection in HEP
- VRNN Architecture
- Search for generic new hadronic resonance:
  - Dataset
  - Preprocessing & Training
  - Analysis Results

Anomaly Detection in HEP

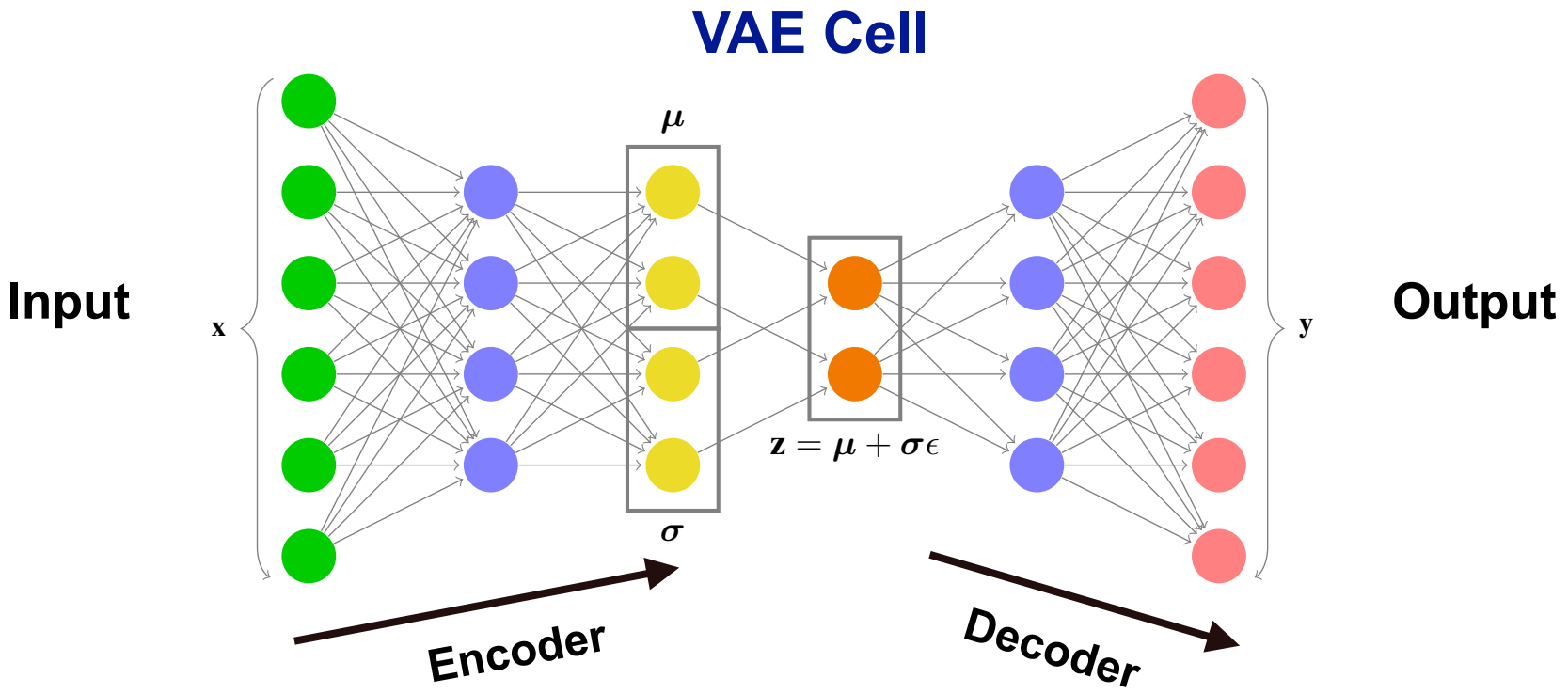
- **Anomaly detection (AD)** = identify features of the data that are inconsistent with a background-only model
  - "Unsupervised" learning = train on data; no signal hypothesis
- At the **Large Hadron Collider**: no recent new physics + many exclusion results → develop strong model independent search program

➡Focus here on **resonant new physics** in **hadronic final states**



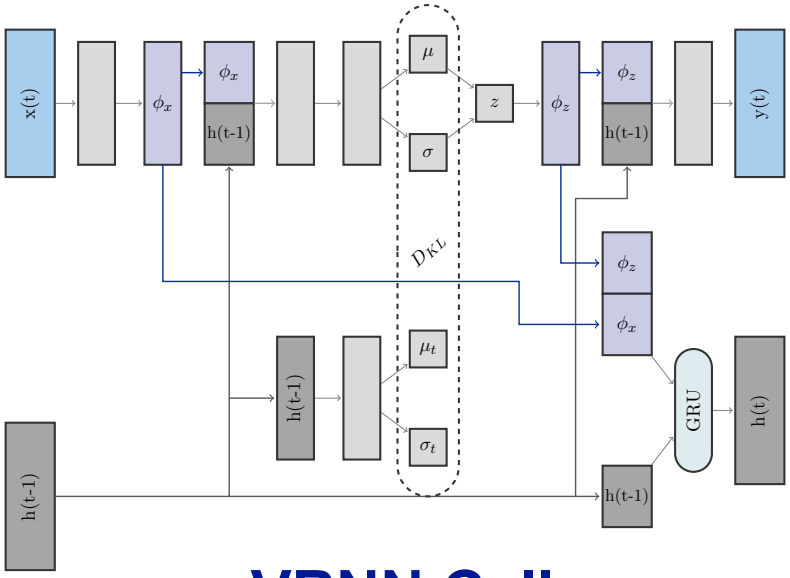
Autoencoders for AD

- **Autoencoder:** generative model that *encodes* input in lower-dimensional latent space, *decodes* from latent space, and checks reconstruction error
- **Variational autoencoder:** perform Bayesian inference by sampling from a multivariate Gaussian latent space



Why VRNN?

- **Variational RNN:** recurrent neural network (RNN) that updates a VAE latent space at each time step; accommodates variable-length input sequences



VRNN Cell

Loss

$$\mathcal{L}(t) = |\mathbf{y}(t) - \mathbf{x}(t)|^2 + \lambda D_{KL}(z||z_t)$$

Mean-squared reconstruction error

Kullback-Leibler Divergence

➡Train over data using large-R jet constituent 4-vectors as inputs

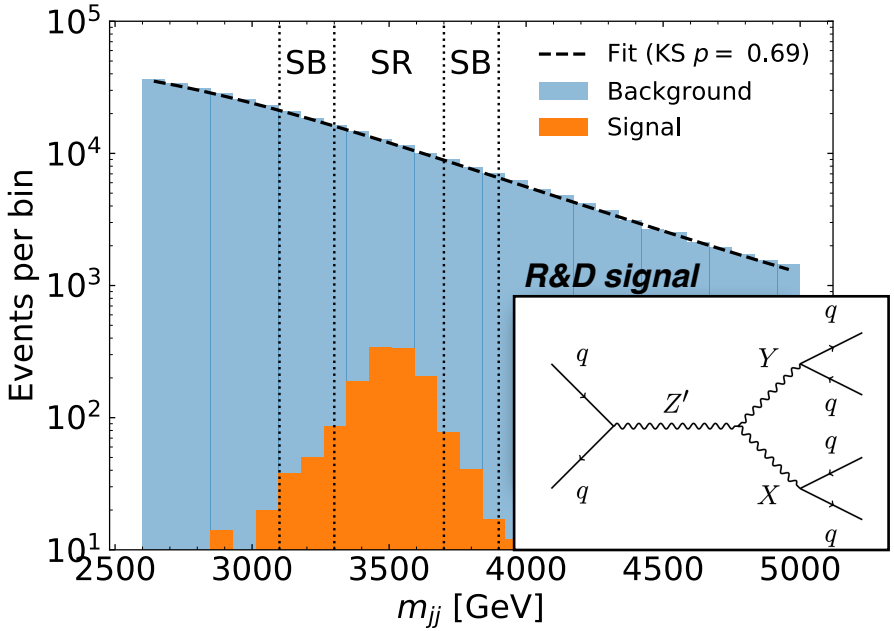
Dataset

- LHC Olympics dataset: Pythia generated + Delphes detector simulation (no pileup)

➡arXiv:2101.08320

- **Signal:** 3.5 TeV Z' → 500 GeV X + 100 GeV Y
  - Two substructure hypotheses: 2-pronged and 3-pronged X/Y decays

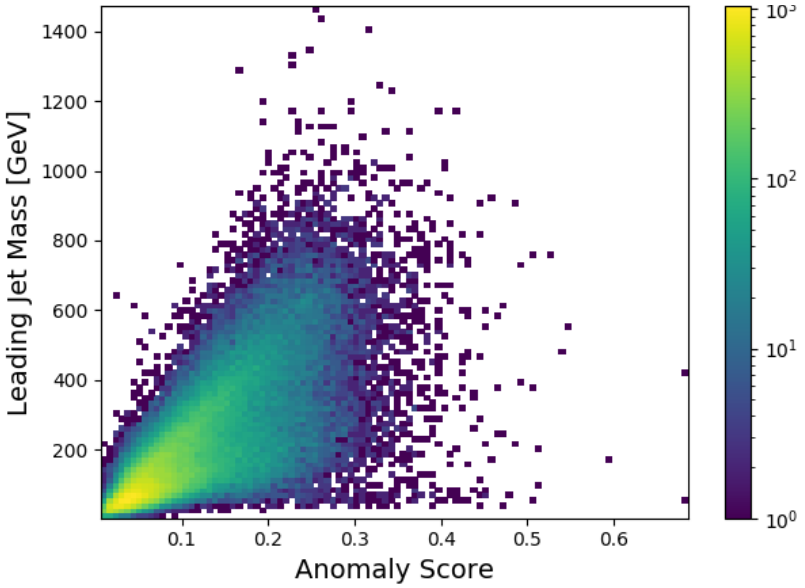
- Reconstruction = two large-radius (R=1.0) jets
  - Trigger: 1 large-R jet with p_T > 1.2 TeV



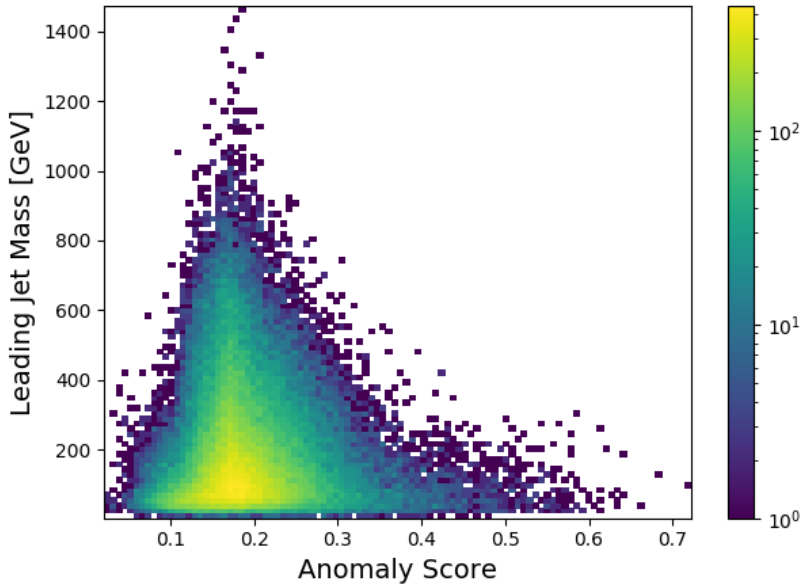
Alignment

- **Goal**: remove mass and pT information from input jets to avoid tagging on kinematics alone
- **Procedure**:
  1. Rescale each jet to the same mass
  2. Boost each jet to the same energy
  3. Rotate each jet to the same η/Φ orientation
- **Result**: anomaly score far less correlated with mass in background jets

No Alignment



With Alignment



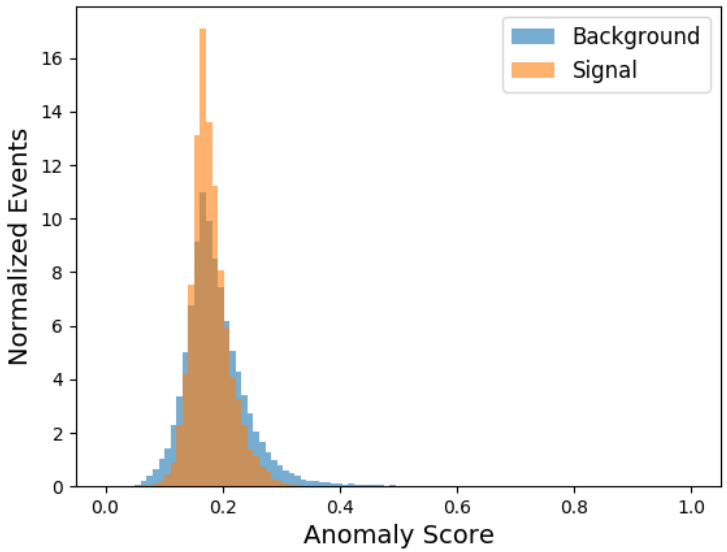
Sequence Ordering

- In a recurrent architecture, apt sequence modeling of jets (eg. order of constituents) can highlight importance sequence features & boost performance
- Select **k_t-distance** ordering to highlight substructure: n^{th} constituent has highest k_t-distance relative to previous, starting with highest pT constituent

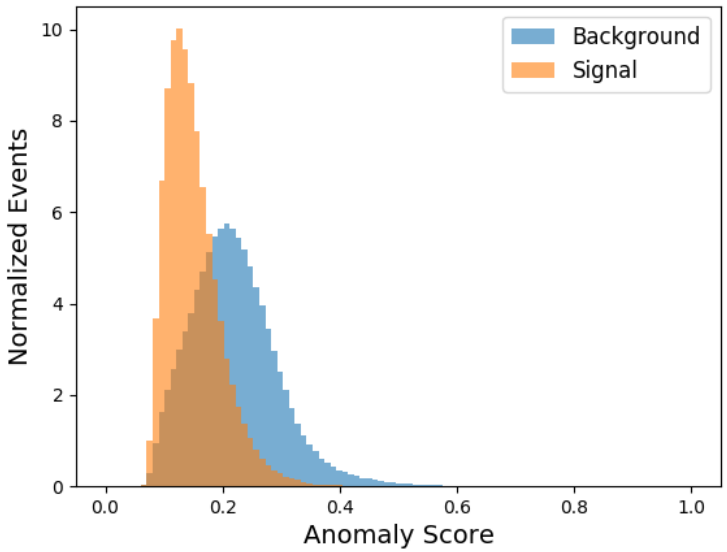
$$c_n = max(p_{Tn}\Delta R_{n,n-1})$$

- **Result:** better separation of two-prong signal from diffuse QCD background than pT-sorting

pT-sorted



k_t-sorted



Analysis Application

- Compute anomaly score for each jet
  - Higher KL divergence = higher loss = lower anomaly score
  - ➡ Transform such that higher AS corresponds to more anomalous jets

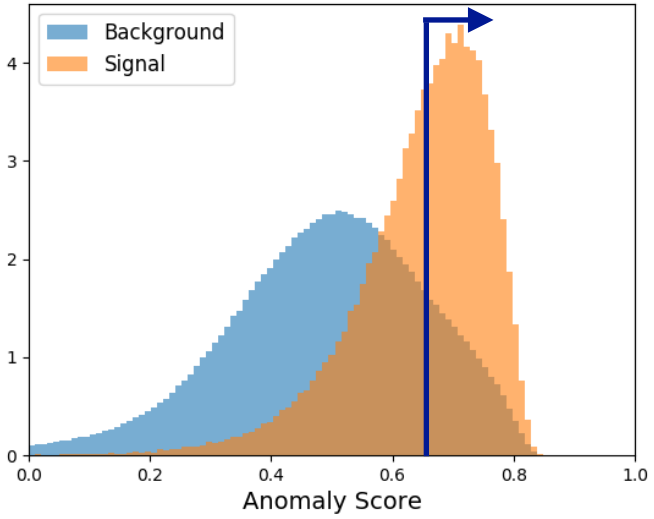
Jet Anomaly Score

$$\rho = 1 - e^{-\overline{D_{KL}}}$$

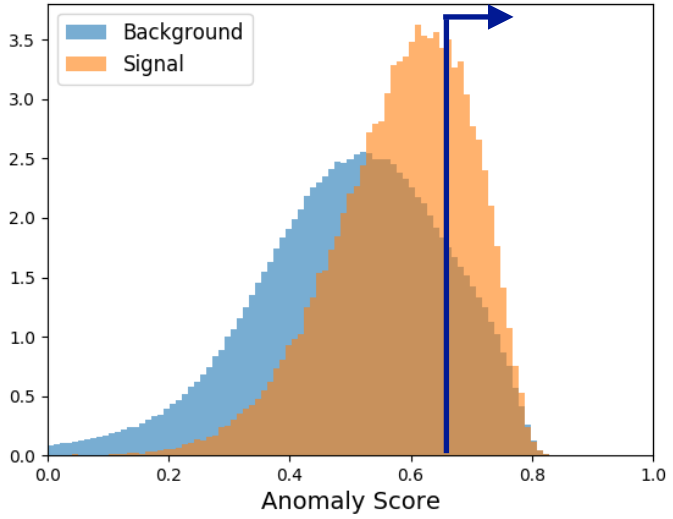
$$\rho' = 1 - \left(\frac{\rho}{2\overline{\rho}}\right)$$

- **Analysis strategy:** cut-and-count on $\rho' > 0.65$ as sole signal region selection & test signal significance in bins of mJJ

2-prong



3-prong

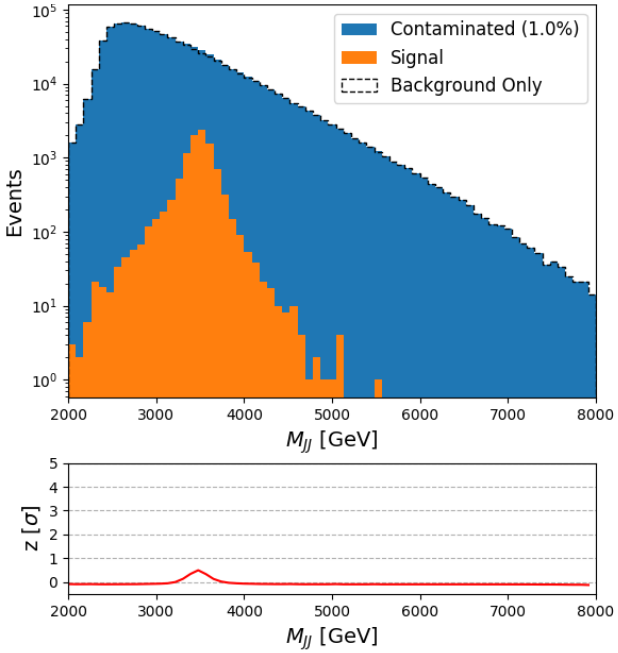


Results: 2-Prong Signal

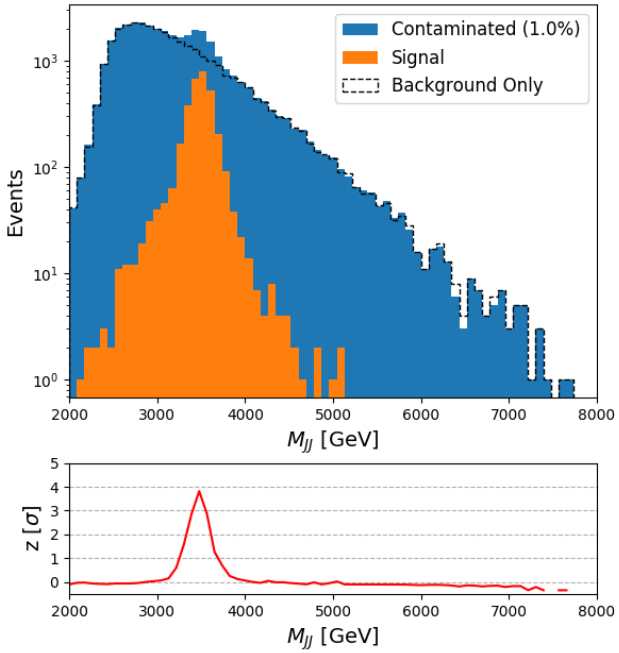
- Perform bump hunt on m_{JJ} with selection on *Event Score* = max of two leading jet Anomaly Scores
- Dataset = background + 1% signal contamination

➡ Enhance a **0.5σ two-prong signal excess to 4.0σ** solely from an Event Score cut at 0.65

No cut



Event Score > 0.65



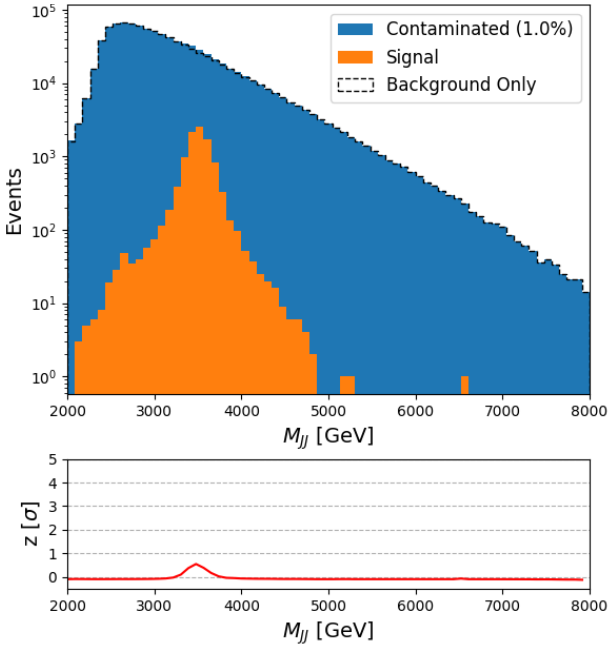
Results: 3-Prong Signal

- Perform bump hunt on m_{JJ} with selection on *Event Score* = max of two leading jet Anomaly Scores
- Dataset = background + 1% signal contamination

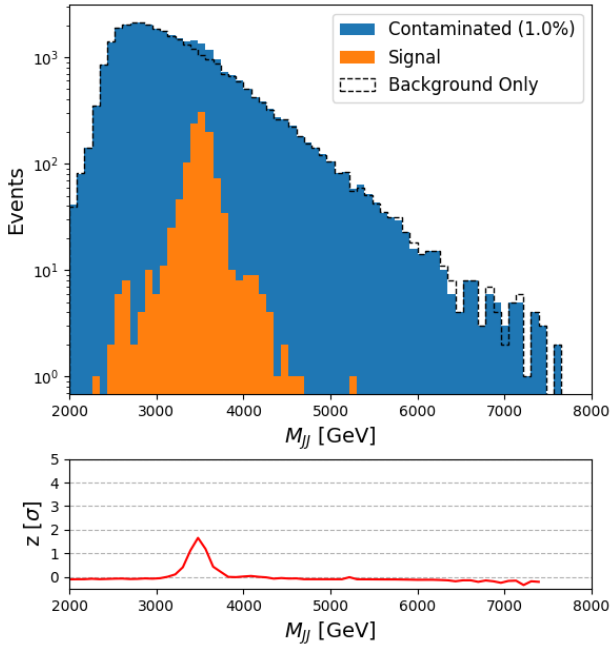
➡ Enhance a **0.5σ two-prong signal excess to 4.0σ** solely from an Event Score cut at 0.65

➡ Enhance a **0.5σ three-prong excess to 1.5σ** using the same score

No cut



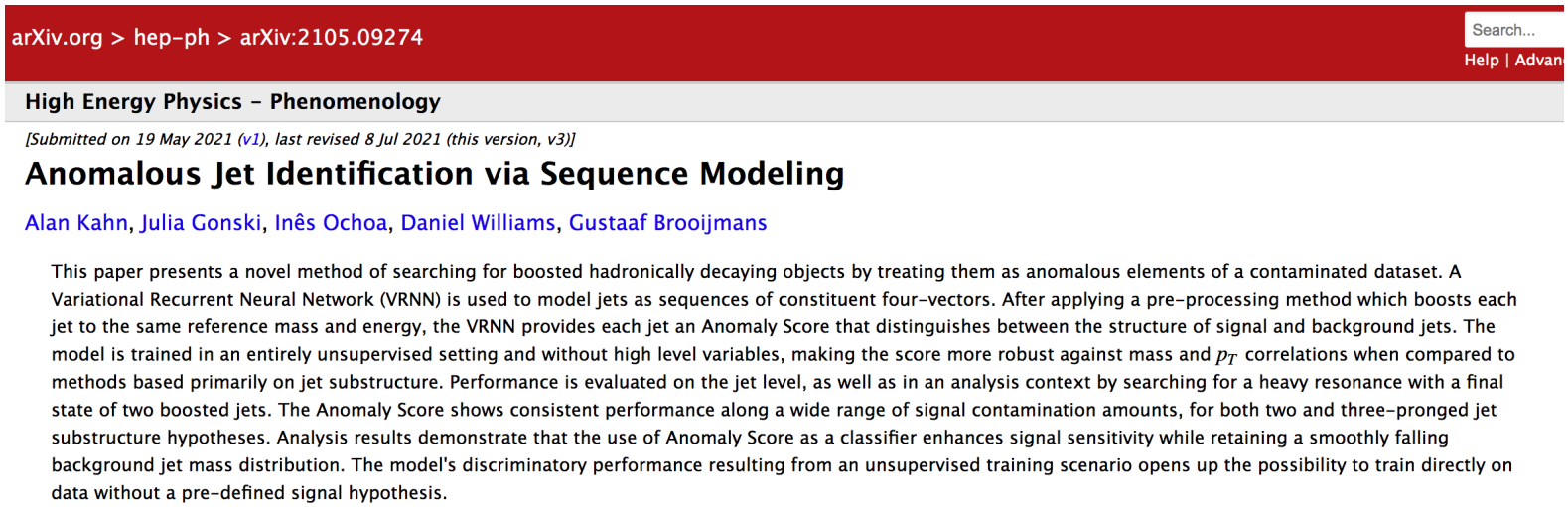
Event Score > 0.65



Conclusions

- We demonstrate an application of the VRNN architecture to data-driven unsupervised learning on *pp* collisions
  - Many thanks to the LHC Olympics organizers for the dataset & guidance!
- Resulting Anomaly Score is able to enhance both two- and three-prong substructure hypotheses over background from multijet processes
- Assessing application towards ATLAS data/physics analysis

➡ arXiv:2105.09274 & accepted to JINST!



arXiv.org > hep-ph > arXiv:2105.09274

Search...

Help | Advan

High Energy Physics - Phenomenology

[Submitted on 19 May 2021 (v1), last revised 8 Jul 2021 (this version, v3)]

Anomalous Jet Identification via Sequence Modeling

Alan Kahn, Julia Gonski, Inês Ochoa, Daniel Williams, Gustaaf Brooijmans

This paper presents a novel method of searching for boosted hadronically decaying objects by treating them as anomalous elements of a contaminated dataset. A Variational Recurrent Neural Network (VRNN) is used to model jets as sequences of constituent four-vectors. After applying a pre-processing method which boosts each jet to the same reference mass and energy, the VRNN provides each jet an Anomaly Score that distinguishes between the structure of signal and background jets. The model is trained in an entirely unsupervised setting and without high level variables, making the score more robust against mass and p_T correlations when compared to methods based primarily on jet substructure. Performance is evaluated on the jet level, as well as in an analysis context by searching for a heavy resonance with a final state of two boosted jets. The Anomaly Score shows consistent performance along a wide range of signal contamination amounts, for both two and three-pronged jet substructure hypotheses. Analysis results demonstrate that the use of Anomaly Score as a classifier enhances signal sensitivity while retaining a smoothly falling background jet mass distribution. The model's discriminatory performance resulting from an unsupervised training scenario opens up the possibility to train directly on data without a pre-defined signal hypothesis.

Backup

Technical Details

- Dataset processing: amended pyjet with FastJet + fjcontrib functionality
- PyTorch library
- Optimizer: Adam, initial learning rate 10^{-5}
- 500 epochs
- Regularization: gradient clipping with clip value 10

Alignment Algorithm

Start

Boost jet in z direction until $\eta_{Jet} = 0$

Rotate jet about z axis until $\phi_{Jet} = 0$

Rescale jet four-vector such that m_{Jet} =0.25 GeV

Boost jet along its axis until E_{Jet} =1 GeV

Rotate jet about x axis until hardest constituent has $\eta_1 = 0, \phi_1 > 0$

if *Any constituents have* $\Delta R > 1$[a] **then**

- Remove all constituents with $\Delta R > 1$
- Rebuild jet with remaining constituents
- Repeat from start

else

- continue

end

if *Number of constituents* > 20 **then**

- Keep up-to the first 20 constituents, ordered in p_T
- Rebuild jet with remaining constituents
- Repeat from start

else

- continue

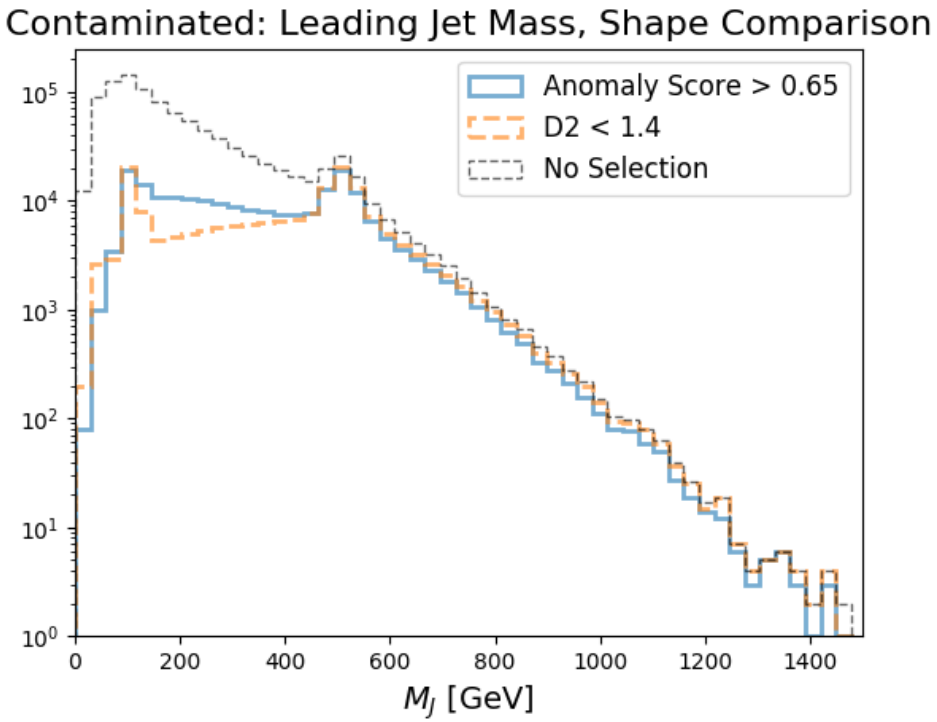
end

Reflect constituents about ϕ axis such that the second hardest constituent has $\eta_2 > 0$

[a] ΔR is computed as $\sqrt{\eta^2 + \phi^2}$ for each constituent, where η and ϕ are measured relative to the x axis.

Comparison to D2

- Dataset = 2-prong X% contaminated
- Selections: D2 < 1.4 / AS > 0.65 (equivalent background rejection)



Performance vs. Contamination

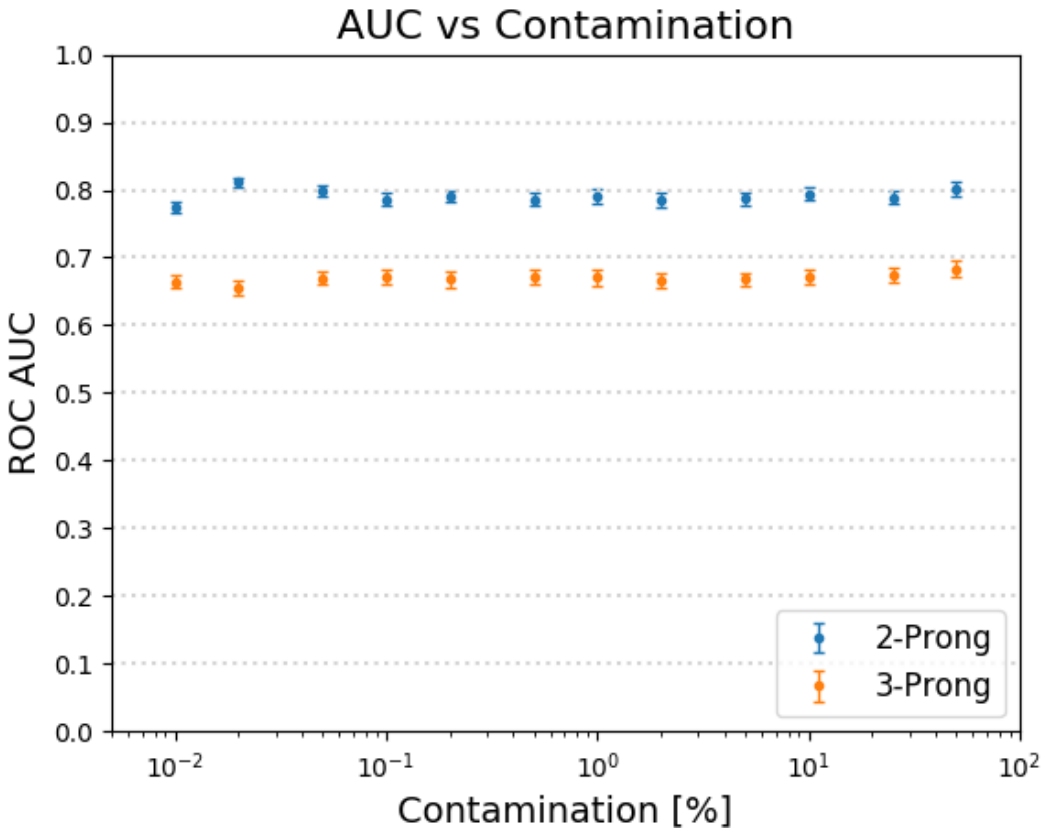


Figure 11. ROC AUC vs. percent signal contamination in training datasets. The performance of the Anomaly Score is consistent across a wide range of contamination levels.

Performance vs. Training Time

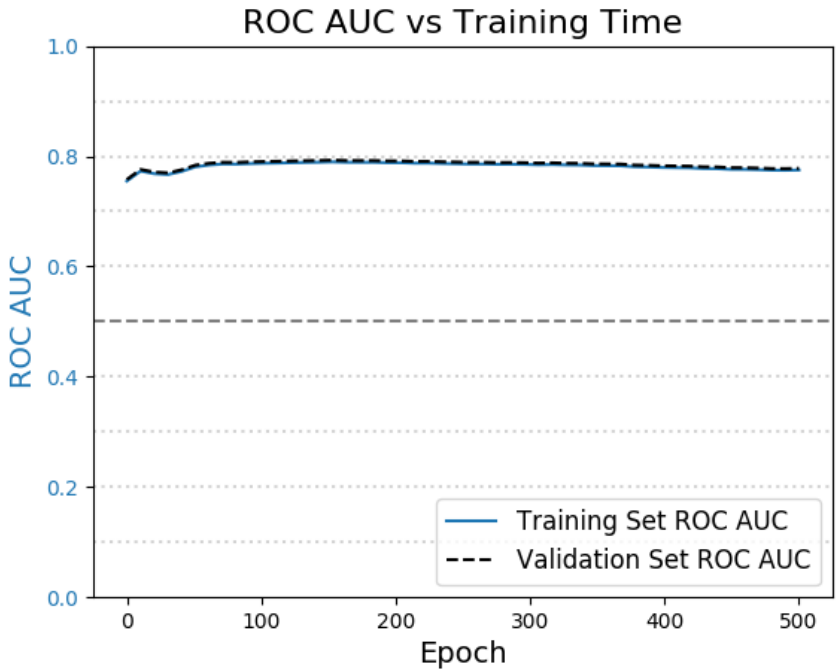


Figure 6. Area Under the Curve (ROC AUC) vs. training time in epochs on a 1% signal-contaminated dataset. The VRNN reaches an optimal performance quickly, and retains this performance over a long training period. The difference in performance between the training and validation sets is a result of the former containing elements of signal.

A Word on Jets

- **Jets** = sprays of hadronic particles reconstructed with clustering algorithms into a cone
- Higher mass exclusions for new particles + high energy collisions = high momentum outputs
  - **Constituents:** individual hadrons in jet
  - **Boosting:** collimation of constituents due to high momentum parent
  - **Substructure**: synthesizing correlations between jet constituents to determine particle content in large radius jet