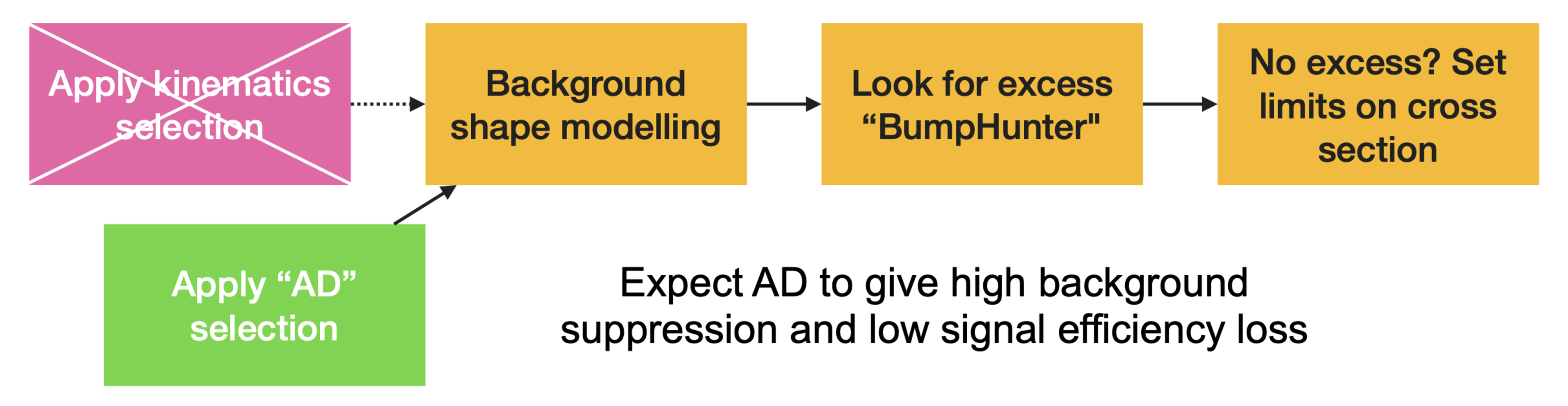
Search for new physics using unsupervised machine learning for anomaly detection with the ATLAS detector at the LHC

Chi Lung Cheng (University of Wisconsin-Madison)
On behalf of the ATLAS collaboration, ICNFP 2023

Introduction

- Traditional studies optimize signal regions using **BSM Monte Carlo**
- **Anomaly detection (AD)** is used as a new strategy that defines outlier events to look for new phenomena in two body invariant masses



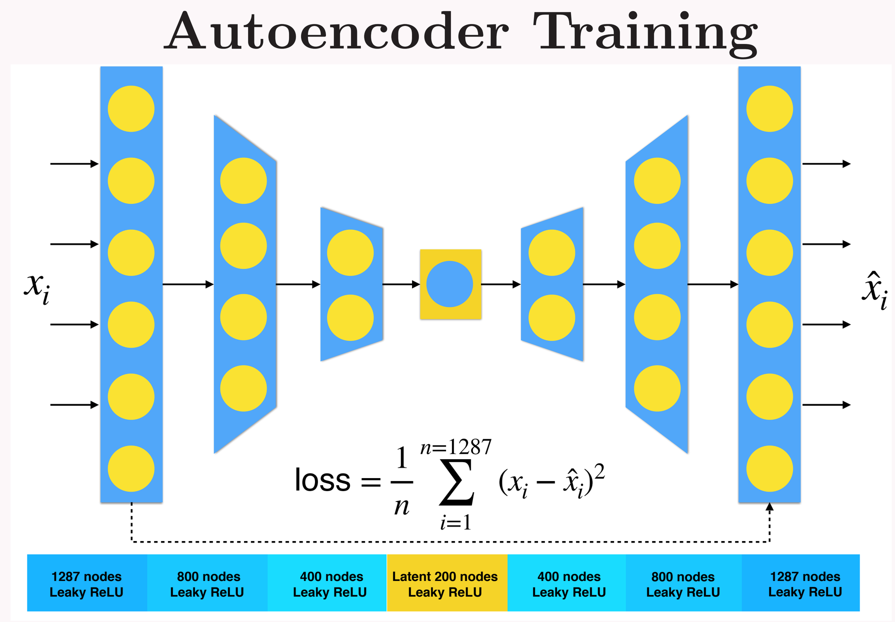
Advantage of using Anomaly Detection

- Not relying on specific signal hypothesis – Model independent
- Unsupervised training on data – no MC modelling required
- Full exploitation of event topologies on the standard reconstructed objects (jet, b-jet, e, μ, γ, met)

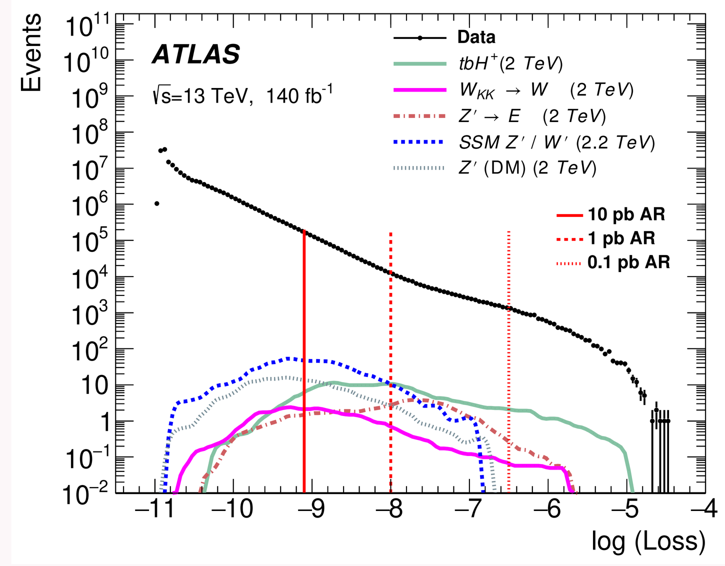
Analysis strategy

- One lepton trigger and pre-selection: $p_T^l > 60$ GeV, $p_T^{jet} > 30$ GeV
- Reconstruct **Rapidity Mass Matrix (RMM)** for each event
- Train **Autoencoder (AE)** using randomly selected 1% Run2 data
- Define **Anomaly Regions (AR)** using reconstruction loss from AE
- Likelihood fits on invariant mass spectrum and look for bumps

Autoencoder Training

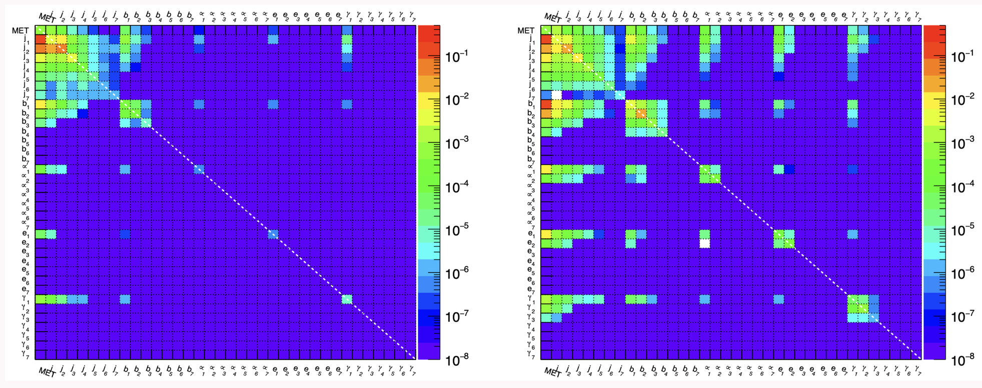


AR boundaries



Rapidity Mass Matrix (RMM)

- 2D matrix comprised of single- and double-particle characteristics of all reconstructed objects
- Different characteristics for different processes: e.g. **Multi-jets QCD** (left) VS **SM Higgs production with all Higgs decays allowed** (right)



Reference: 1810.06669

Background modelling

- Smoothing falling mass spectra
- Use SM MC and loose-lepton control regions to define bkg function:

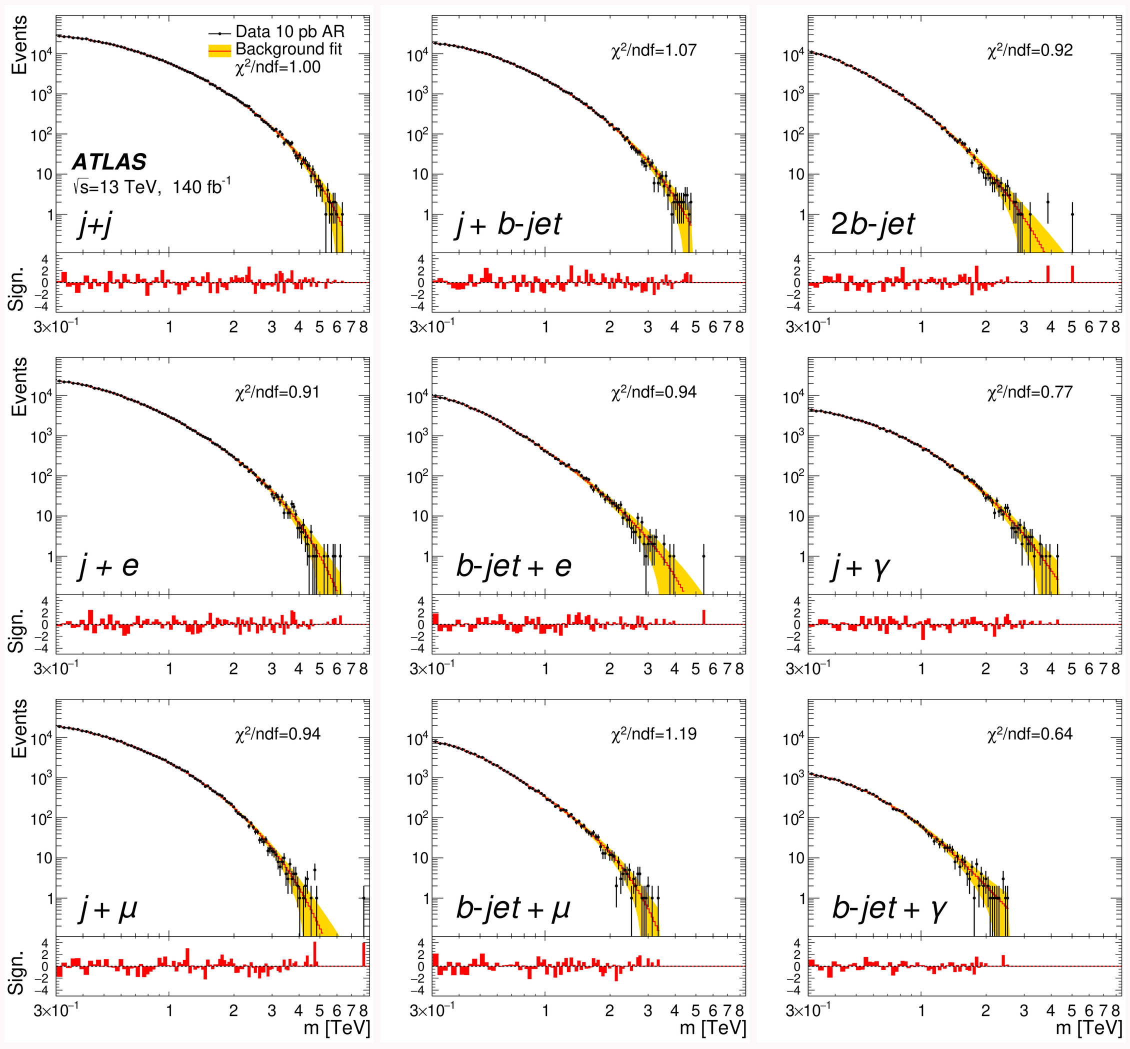
$$f(x) = p_1(1-x)^{p_2}x^{p_3+p_4\ln x+p_5\ln^2 x}$$

- Alternative bkg function for systematics:

$$f(x)_{\mathrm{alt}} = p_1(1-x)^{p_2}x^{p_3+p_4\ln x+p_5/\sqrt{x}}$$

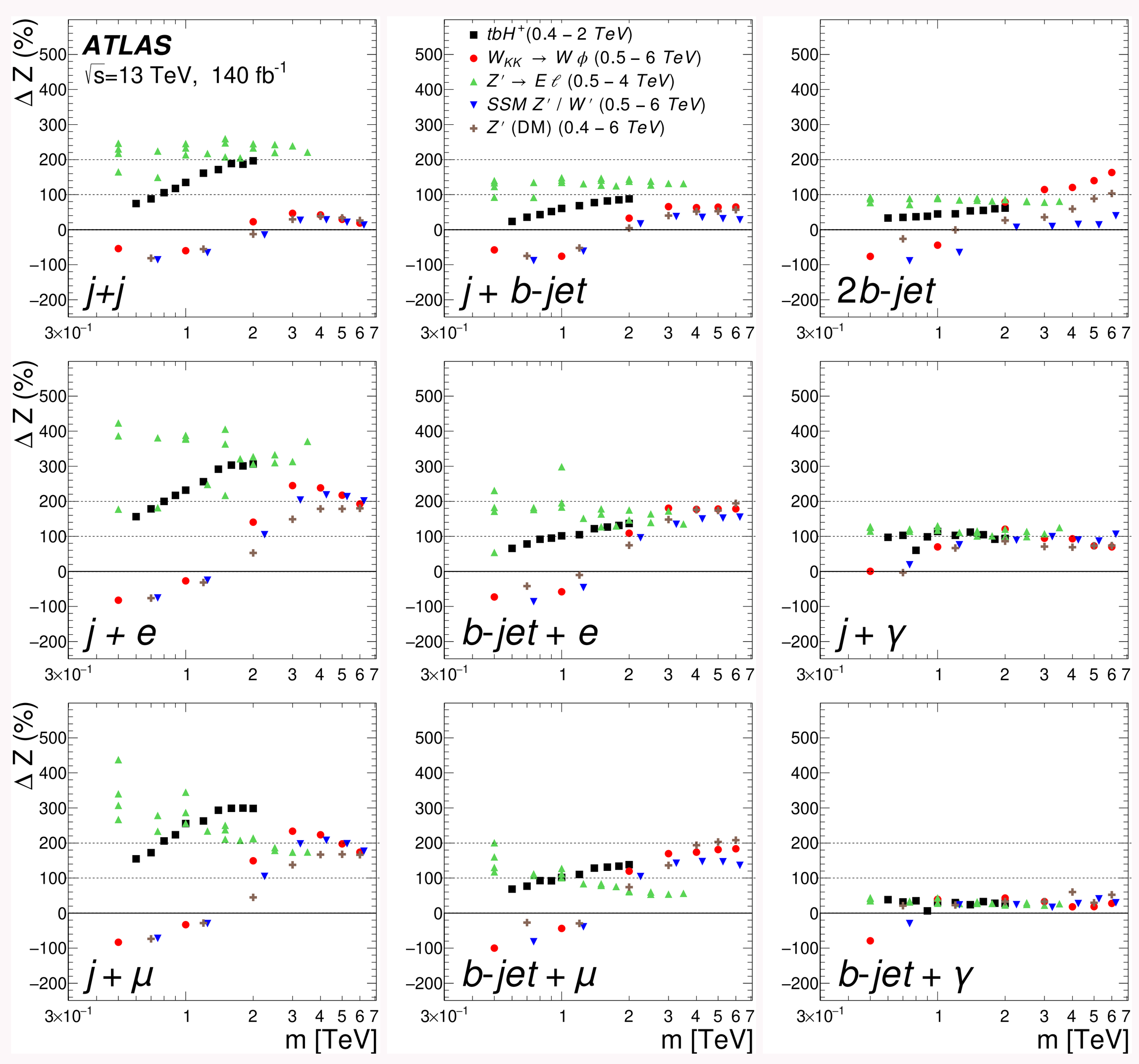
Results in 10 pb Anomaly Region

BumpHunter results



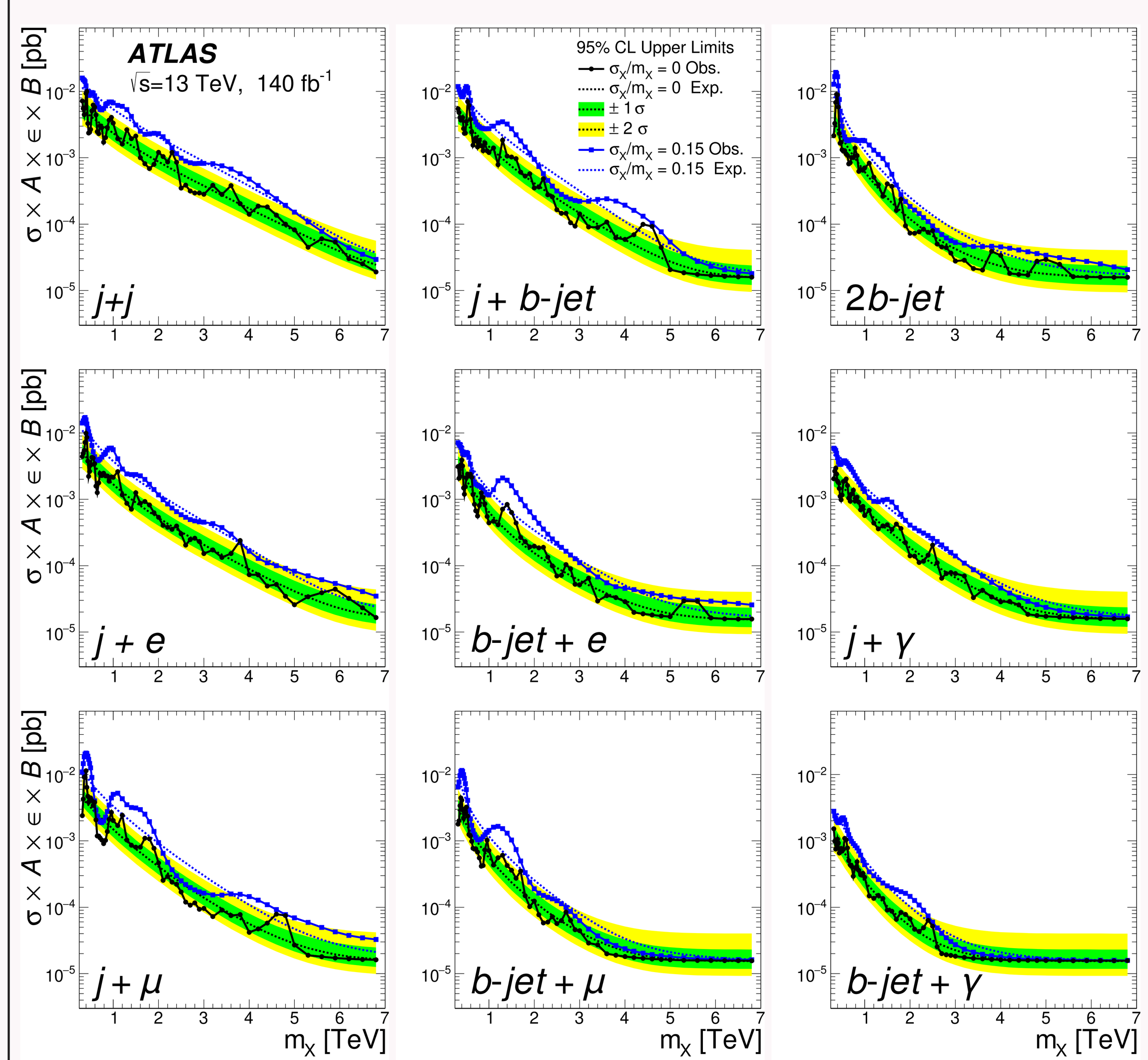
- Good agreement from mass spectrum fit
  - Normality tests on pulls passed for all masses
  - Bkg shape uncertainty shown in yellow
- Bin-by-bin significances in lower panel
  - Largest deviation is $m_{j\mu}$ at around 4.8 TeV

Sensitivity to BSM signals



- Calculate improvement in discovery significance
  - $Z = \sqrt{2\cdot[(S+B)\cdot\ln(1+S/B)-S]}$
- Sensitivity improvement quantified by ΔZ
  - $\Delta Z = 300-400\%$ for most BSM models
  - Directly translates into competitive limits

Upper limits of Gaussian signals



- 95% CL upper limits for Gaussian-shaped signals with two width hypotheses:
  - $\sigma = 0$ and $\sigma/m = 15\%$
- Largest deviation is consistent with a statistical fluctuation
  - Local 2.9 σ at $m_{j\mu} = 4.8$ TeV

Summary

- Successful application of **unsupervised machine learning** for anomaly detection using event level information
- **Model agnostic selection** based on data instead of BSM Monte Carlo
- 9 two-body invariant masses for jet+X (b-jet+X) analyzed in 3 outlier regions, no significant deviation observed
- In 10 pb Anomaly Region, the largest deviation (2.9 σ) for j+μ near 4.8 TeV is consistent with statistical fluctuation
- Analysis method shows improvement of sensitivity up to around 300% in model-independent limits of Gaussian-shaped signals

Reference: EXOT-2022-07