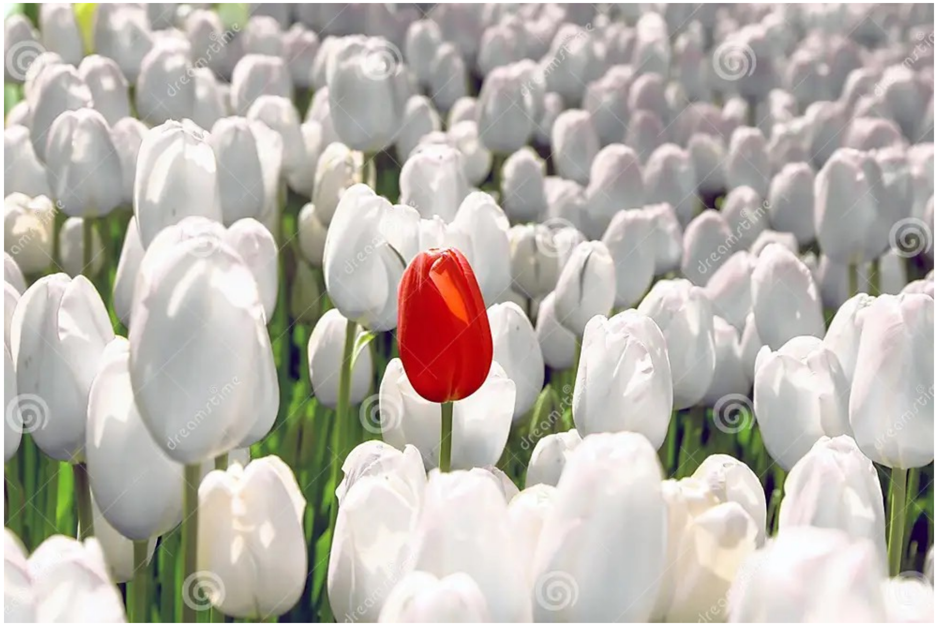
Search for new physics in two-body mass distributions using unsupervised ML for anomaly detection



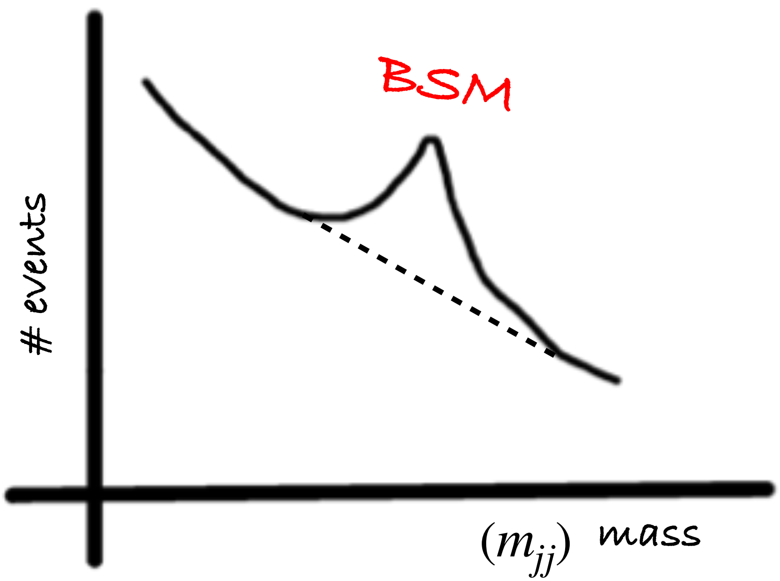
Wasikul Islam

University of Wisconsin-Madison, USA

On behalf of the ATLAS Collaboration

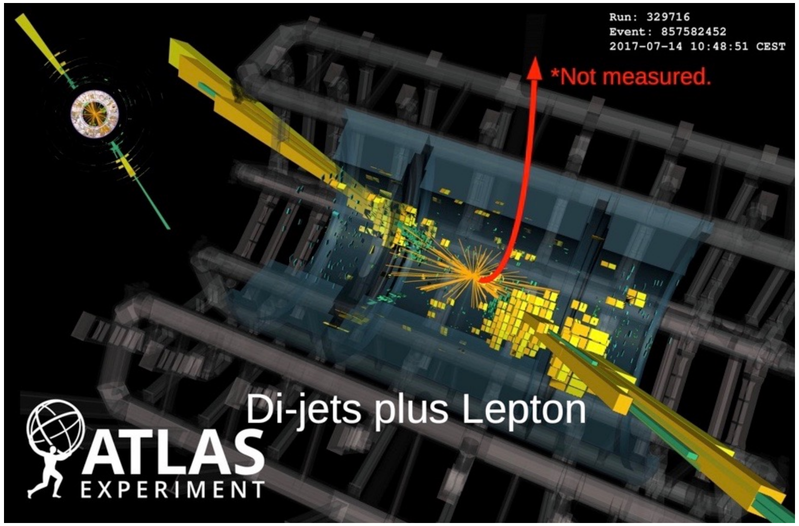
EAST BAY
BOOST 23
31 JUL - 4 AUG LAWRENCE BERKELEY NATIONAL LAB

Motivation



- Two body (& dijet) searches are very common in collider experiments.
- New phenomena produced in parton collision can decay to partons predicated by BSM & can manifest as narrow resonances above the observable spectra.
- Investigate a wide range of BSM theories
- Many studies have been performed on Dijet final states. (by UA1, UA2 Collaborations at the CERN SppS; CDF, D0 at the Tevatron & by both CMS and ATLAS experiments)

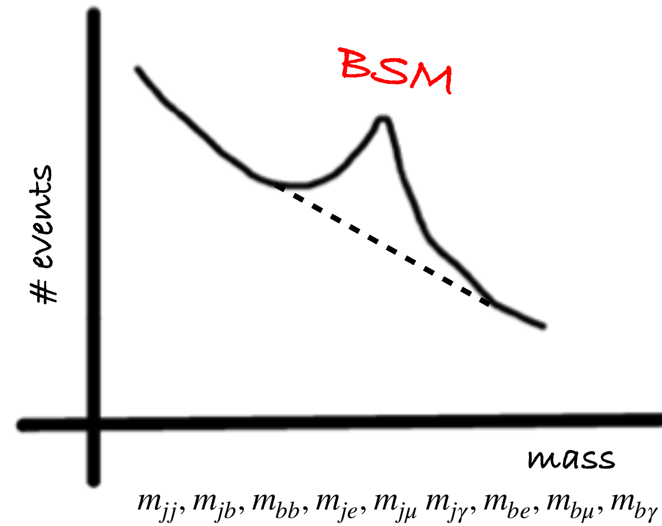
- Inclusive searches are typically restricted to $m_{jj} > 1.0$ TeV due to $p_{\mathrm{T}}^{\mathrm{jet}}$ trigger thresholds.
- Overcome trigger limitations by exploiting spectator objects, e.g. photons, leptons.
- We searched for dijet resonances using leptons as spectator objects/for triggering purpose.
- Requiring an additional lepton reduces the QCD multi-jet background rate.



JHEP07(2023)202

Motivation

- Model-independent searches in dijets are very common in ATLAS/CMS Experiments
- Example: Inclusive searches use “signal region” (y*<0.6). Ref : JHEP03(2020)145
- This y* cut cannot be effective when dijets are associated with other objects (leptons, photons..)
- Machine learning (ML) anomaly-detection methods open a new ways to study the collision events.
- Replace such cuts with a selection based on machine learning (ML) that defines “Anomaly region”, i.e. collision events that are most distinct from the Standard Model (pp events)



Our anomaly detection approach

~~Apply kinematics selection~~ → Background shape modeling → Look for excess “BumpHunter” Perform likelihood fits → No excess? Set limits on cross section

Apply “AD” selection Find “Anomaly region” → Background shape modeling

Advantages:

No need Monte Carlo simulations

Agnostic to BSM & unexpected signatures

Improves S/B sensitivity for new physics

What’s new ?

See Elham’s talk for other ML approaches in ATLAS. Ours is **Event based** anomaly detection using Unsupervised ML !

Analysis strategy

Reference : arxiv:2307.01612

Trigger (one lepton) and pre-selection (p_T^{lepton}*(trigger)*> 60 GeV , p_T^{jet} > 30 GeV)

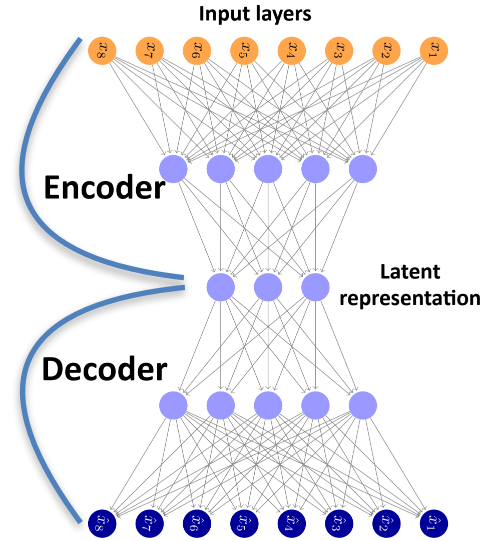
Reconstruct Rapidity Mass Matrix for each event

Rapidity-mass matrix

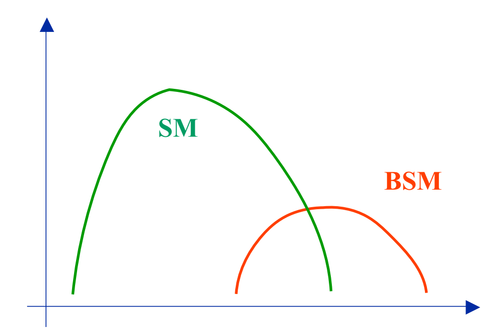
$$\begin{pmatrix} e_T^{miss} & m_T(j_1) & m_T(j_2) & \ldots m_T(j_N) & m_T(\mu_1) & \ldots m_T(\mu_N) \\ h_L(j_1) & e_T(j_1) & m(j_1,j_2) & \ldots m(j_1,j_N) & m(j_1,\mu_1) & ..m(j_1,\mu_N) \\ h_L(j_2) & h(j_1,j_2) & \delta e_T(j_2) & \ldots m(j_2,j_N) & m(j_2,\mu_1) & ..m(j_2,\mu_N) \\ \ldots & \ldots & \ldots & \ldots, & \ldots & \ldots \\ h_L(j_N) & h(j_1,j_N) & \ldots & \ldots \delta e_T(j_N) & m(j_N,\mu_1) & ..m(j_N,\mu_N) \\ h_L(\mu_1) & h(\mu_1,j_1) & h(\mu_1,j_2) & \ldots h(\mu_1,j_N) & e_T(\mu_1) & m(\mu_1,\mu_N) \\ h_L(\mu_2) & h(\mu_2,j_1) & h(\mu_1,j_2) & \ldots h(\mu_2,j_N) & h(\mu_1,\mu_2) & m(\mu_2,\mu_N) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \\ h_L(\mu_N) & h(\mu_N,j_1) & h(\mu_N,j_2) & \ldots h(\mu_N,j_N) & h(\mu_N,\mu_1) & \delta e_T(\mu_N) \end{pmatrix}$$

Train unsupervised ML model - autoencoder using 1% ATLAS Run 2 data

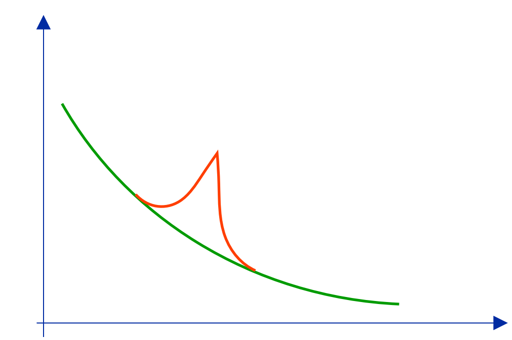
Autoencoder



Define signal region using reconstruction loss from autoencoder



Fit invariant mass spectrum, do statistical analysis, look for bumps in 9 different channels jet+X (b-jet+X) [X=e, μ, γ]



Rapidity mass matrix (RMM) for event representation

Missing momentum & Transverse masses

$$
\begin{pmatrix}
\mathbf{e}_{\mathbf{T}}^{\mathbf{miss}} & m_T(j_1) & m_T(j_2) & \ldots m_T(j_N) & m_T(\mu_1) & m_T(\mu_2) & \ldots m_T(\mu_N) \\
h_L(j_1) & \mathbf{e}_{\mathbf{T}}(\mathbf{j}_1) & m(j_1,j_2) & \ldots m(j_1,j_N) & m(j_1,\mu_1) & m(j_1,\mu_2) & \ldots m(j_1,\mu_N) \\
h_L(j_2) & h(j_1,j_2) & \delta\mathbf{e}_{\mathbf{T}}(\mathbf{j}_2) & \ldots m(j_2,j_N) & m(j_2,\mu_1) & m(j_2,\mu_2) & \ldots m(j_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots, & \ldots & \ldots & \ldots \\
h_L(j_N) & h(j_1,j_N) & \ldots & \ldots \delta\mathbf{e}_{\mathbf{T}}(\mathbf{j}_N) & m(j_N,\mu_1) & m(j_N,\mu_2) & \ldots m(j_N,\mu_N) \\
h_L(\mu_1) & h(\mu_1,j_1) & h(\mu_1,j_2) & \ldots h(\mu_1,j_N) & \mathbf{e}_{\mathbf{T}}(\mu_1) & m(\mu_1,\mu_2) & m(\mu_1,\mu_N) \\
h_L(\mu_2) & h(\mu_2,j_1) & h(\mu_1,j_2) & \ldots h(\mu_2,j_N) & h(\mu_1,\mu_2) & \delta\mathbf{e}_{\mathbf{T}}(\mu_2) & m(\mu_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
h_L(\mu_N) & h(\mu_N,j_1) & h(\mu_N,j_2) & \ldots h(\mu_N,j_N) & h(\mu_N,\mu_1) & h(\mu_N,\mu_2) & \delta\mathbf{e}_{\mathbf{T}}(\mu_N)
\end{pmatrix}
$$

Invariant masses m_{ij}/E_{CM}

E_T/E_{CM} scaled transverse energies & energy imbalances

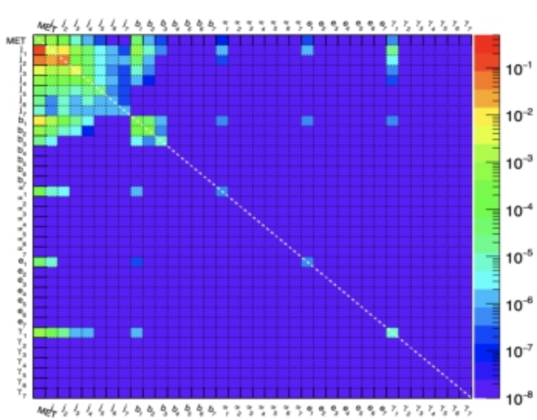
Lorentz factors

Rapidity difference h(i,j)~cosh(y_j - y_i)

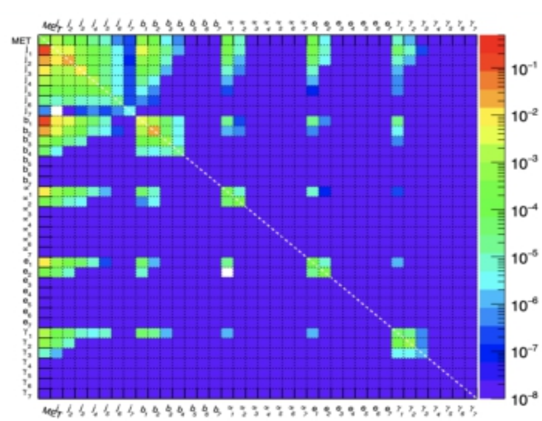
Reference : https://arxiv.org/abs/1805.11650

Examples :

Multi-jet QCD process



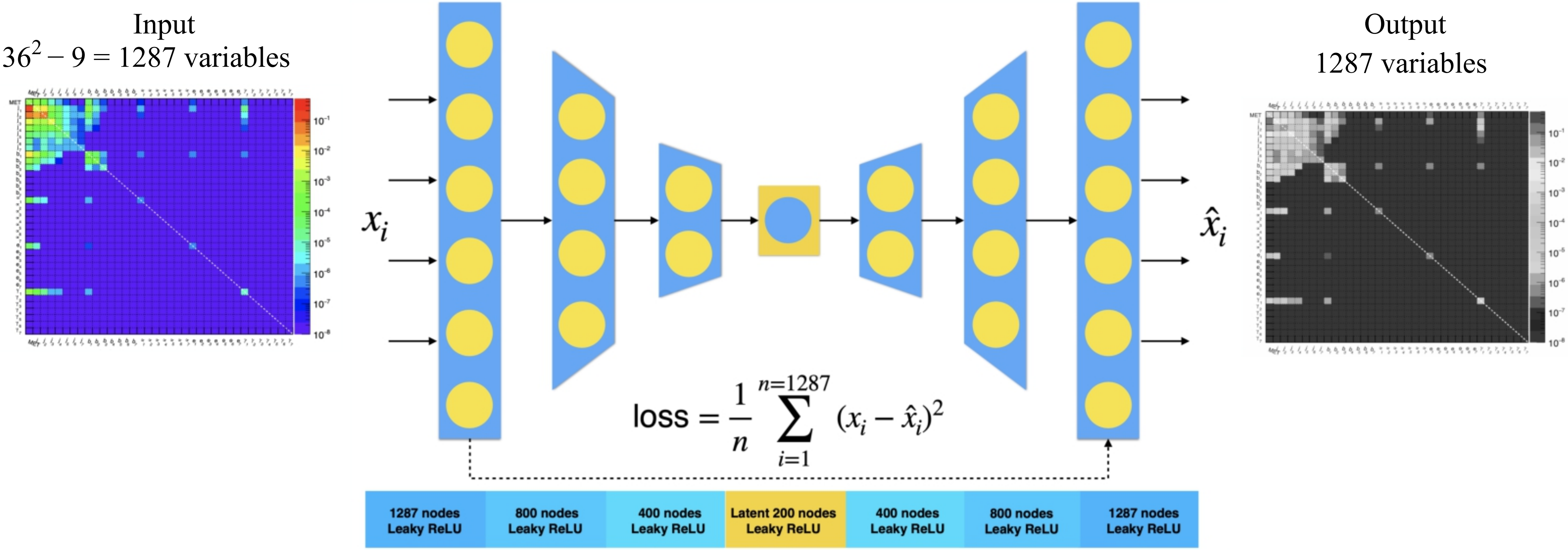
Higgs process



- Expected to have different characteristics for different processes
- RMM matrix is flattened to a 1-dimensional input vector before being fed into the AE
- Using event topologies on the standard reconstructed objects— object type & multiplicity, 4-momenta, two-body system information, all will contribute to the anomaly score.

Training with autoencoder
(using Unsupervised ML)

- Using randomly selected 1% collision data
  - · Sufficient statistics to train and well represent the full collision dataset
  - · Split to 7:3 for training and validation, monitor validation loss for early stopping
  - · 50 variations of random seed
- Tried other architectures such as Variational (convolutional) autoencoder and various sizes of the autoencoder. The selected one gives better performance
- TensorFlow is used to implement AE



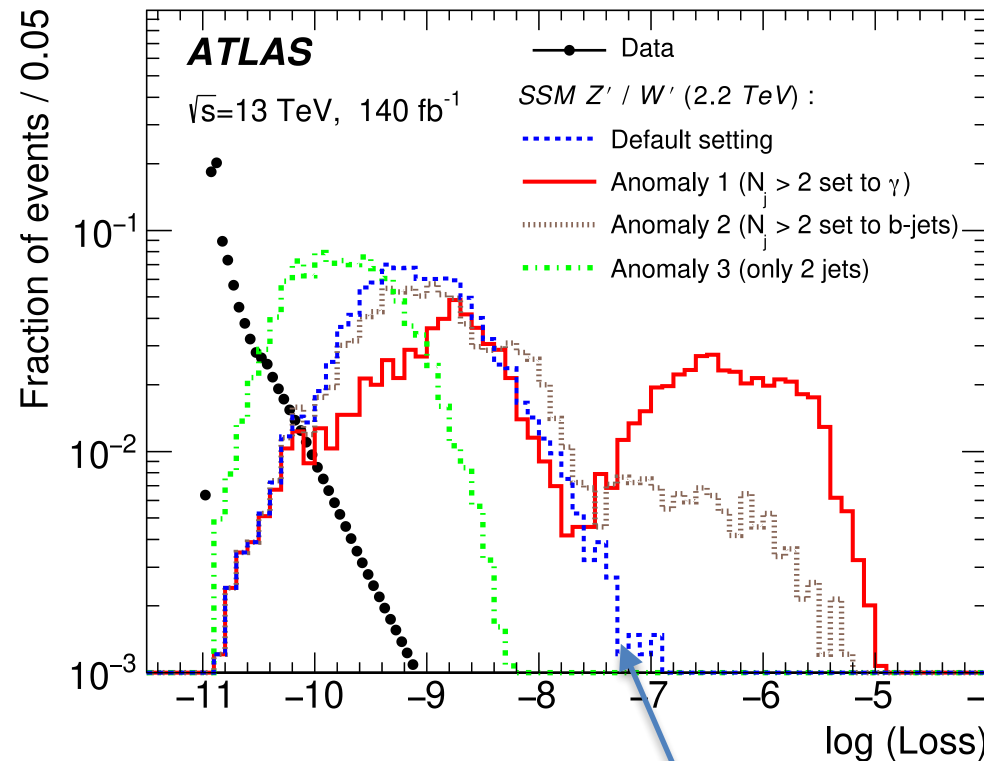
$$\text{loss} = \frac{1}{n}\sum_{i=1}^{n=1287}(x_i - \hat{x}_i)^2$$

Anomaly signal test

Can model detect anomalous events ?

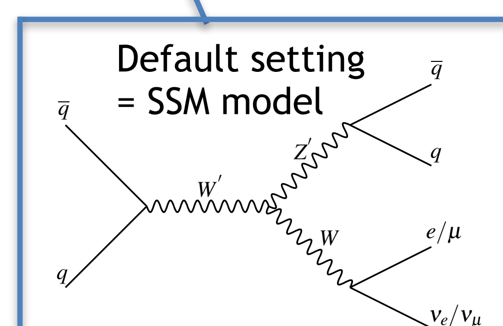
Make up anomaly events by shuffling objects and re-calculating RMM

- Anomaly 1: jets beyond 2 are set to photon
- Anomaly 2: jets beyond 2 are set to b-jets
- Anomaly 3: keep only 2 jets and 1 lepton
  - Less anomaly even than the original events



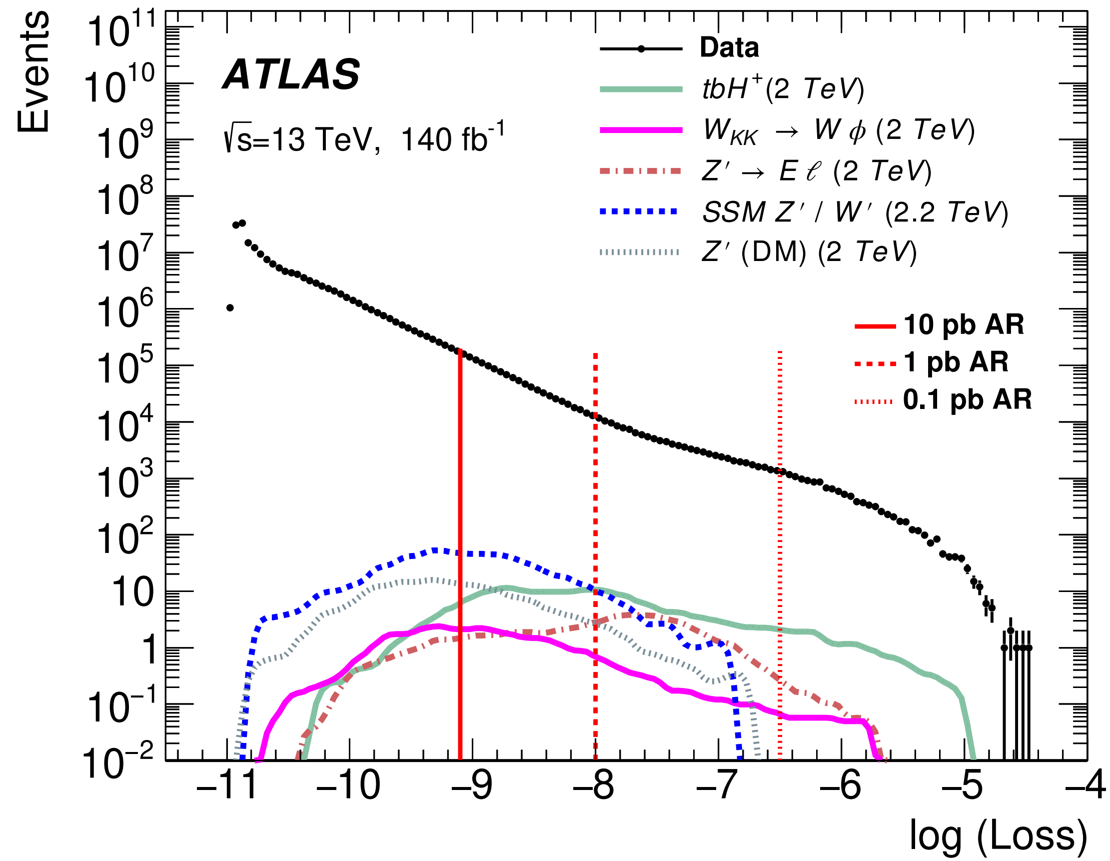
They all show up with large loss, in addition:

- More anomalous is seen when b-jet or photon multiplicity increases - expected
- Less anomalous is seen when multiplicity is low, anomaly may come from large p_{T}, $E_{\mathrm{T}}^{\mathrm{miss}}$, etc -expected



Anomaly region definitions

- Anomaly region should enhance BSM signal and suppress SM bkg
  - Need enough bkg for modelling
- We studied theoretical cross sections of various BSM models for two-body resonances for AR cuts.
- Select events corresponding to 10pb, 1pb, 0.1pb (× 140 fb -1) as 3 anomaly regions to cover different sensitivities.
- With Run 2 data, these correspond to1.4M, 140K, 14K events.
- 10pb AR is the main region to study in this analysis *(Corresponds to maximum value of cross sections for various BSM models considered)*



Loss distributions for BSM models

Background hypothesis testing

The mass spectra of background (SM) are expected to be smoothing falling.

Background hypothesis tested using:
1. Loose-electron control region (LE-CR)
2. LE-CR+MC control region
3. 10% of data (scaled, smoothed)

- For all the invariant masses, fits with different analytic functions were performed & the following p5 function was selected to describe background hypothesis.

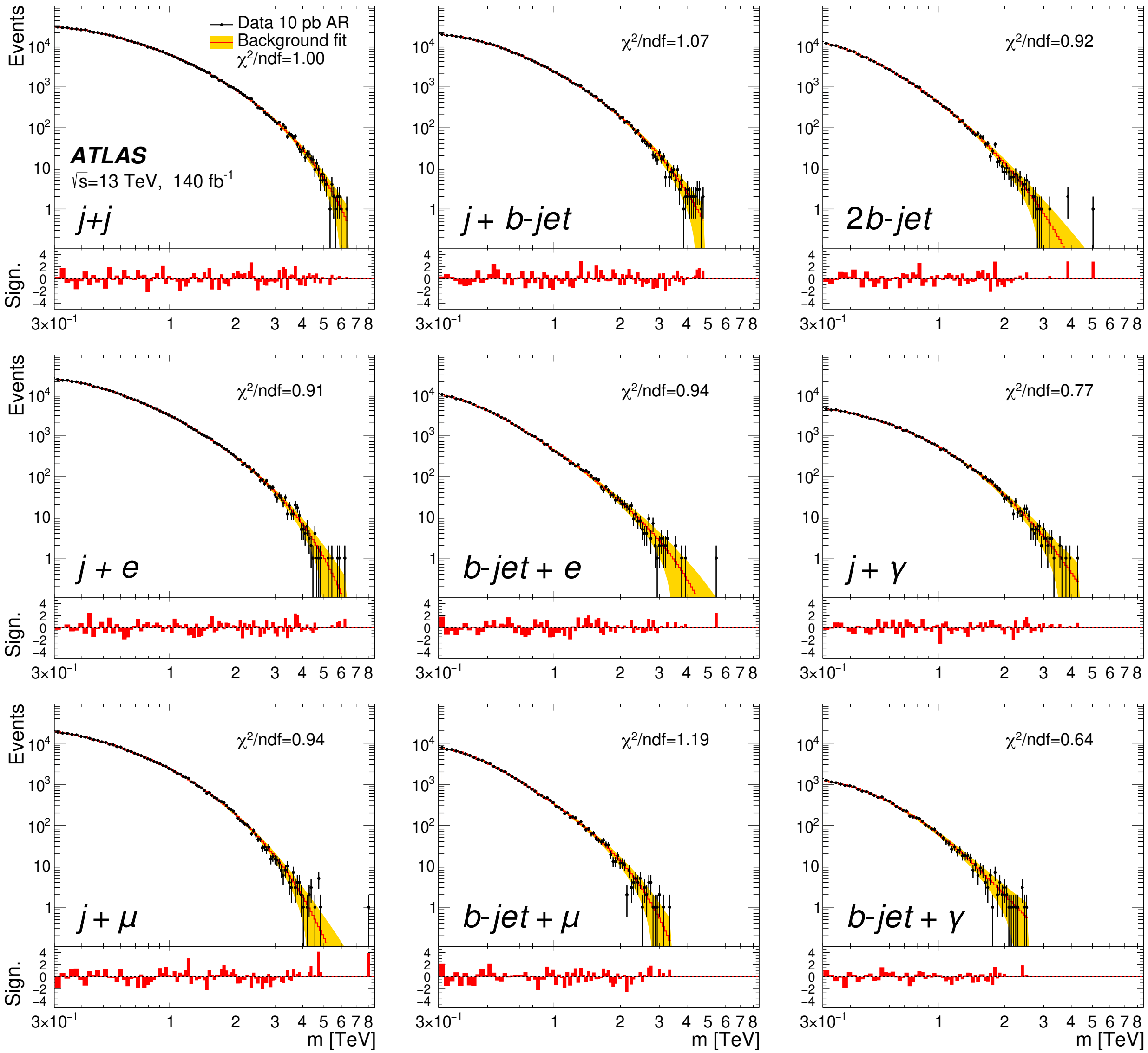
p5 function : $f(x) = p_1(1-x)^{p_2}x^{p_3+p_4 lnx+p_5 ln^2 x}$ Where x ≡ m_{jj} /√s ∈ [0,1]

Also used an alternative function form to estimate systematics

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

The background shapes are well described by the p5 after selecting events with AE (with high loss values) for all the 9 invariant masses.

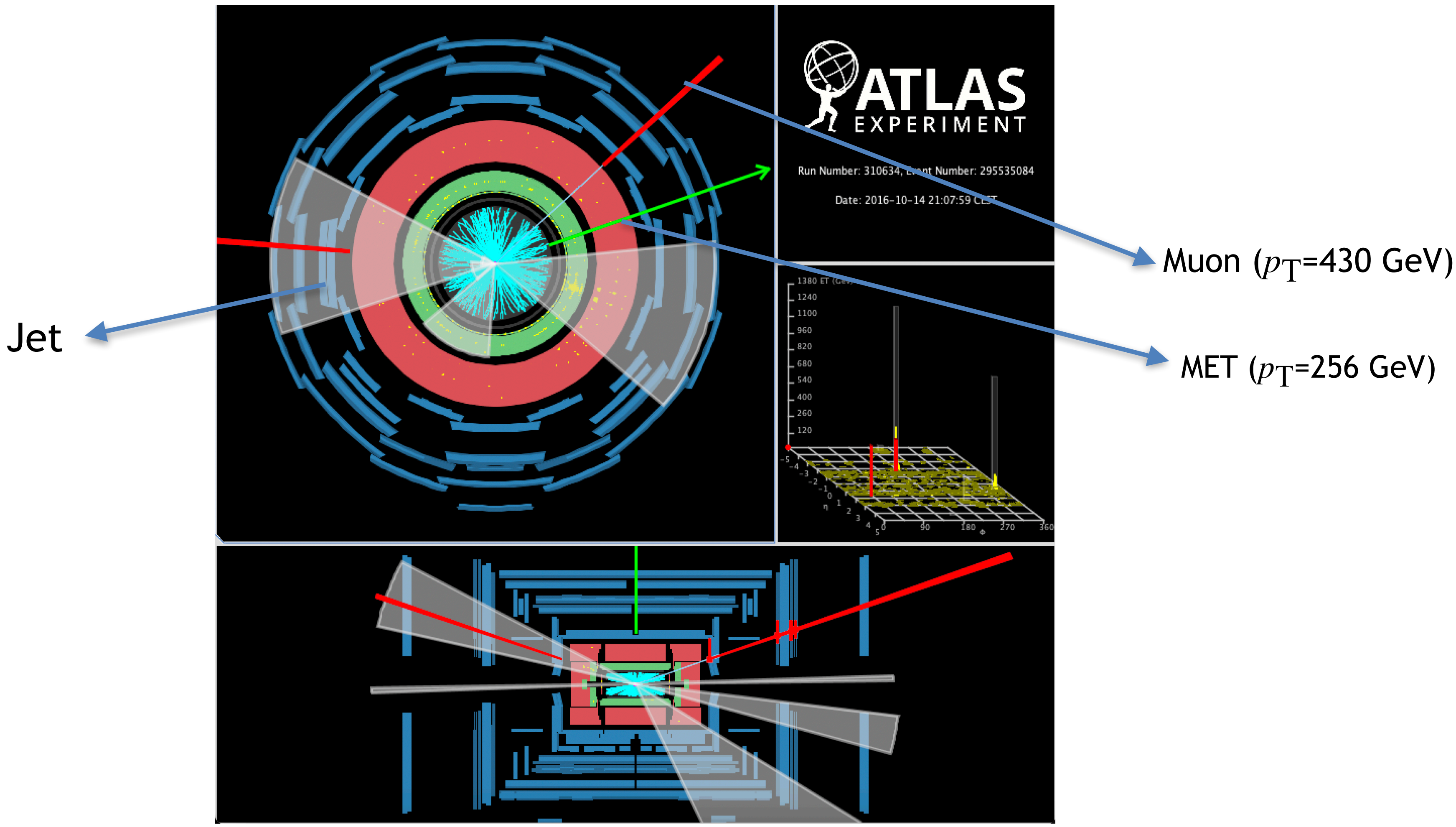
BumpHunter results for 10 pb AR



- BumpHunter results agree with p5 fit
- Tests of normality on pulls passed for all masses passed
- Background shape uncertainty using alternative function shown in yellow
- Largest deviation reported by BumpHunter is at $m_{j\mu} = \sim 4.8$ TeV
- The distributions are binned with increasing bin widths from 16 GeV to 150 GeV to reflect the jet energy resolution of detector

Event display in the j+μ channel

(An event passing 10 pb Anomaly Region with $m_{j\mu}$ = 4.72 TeV)



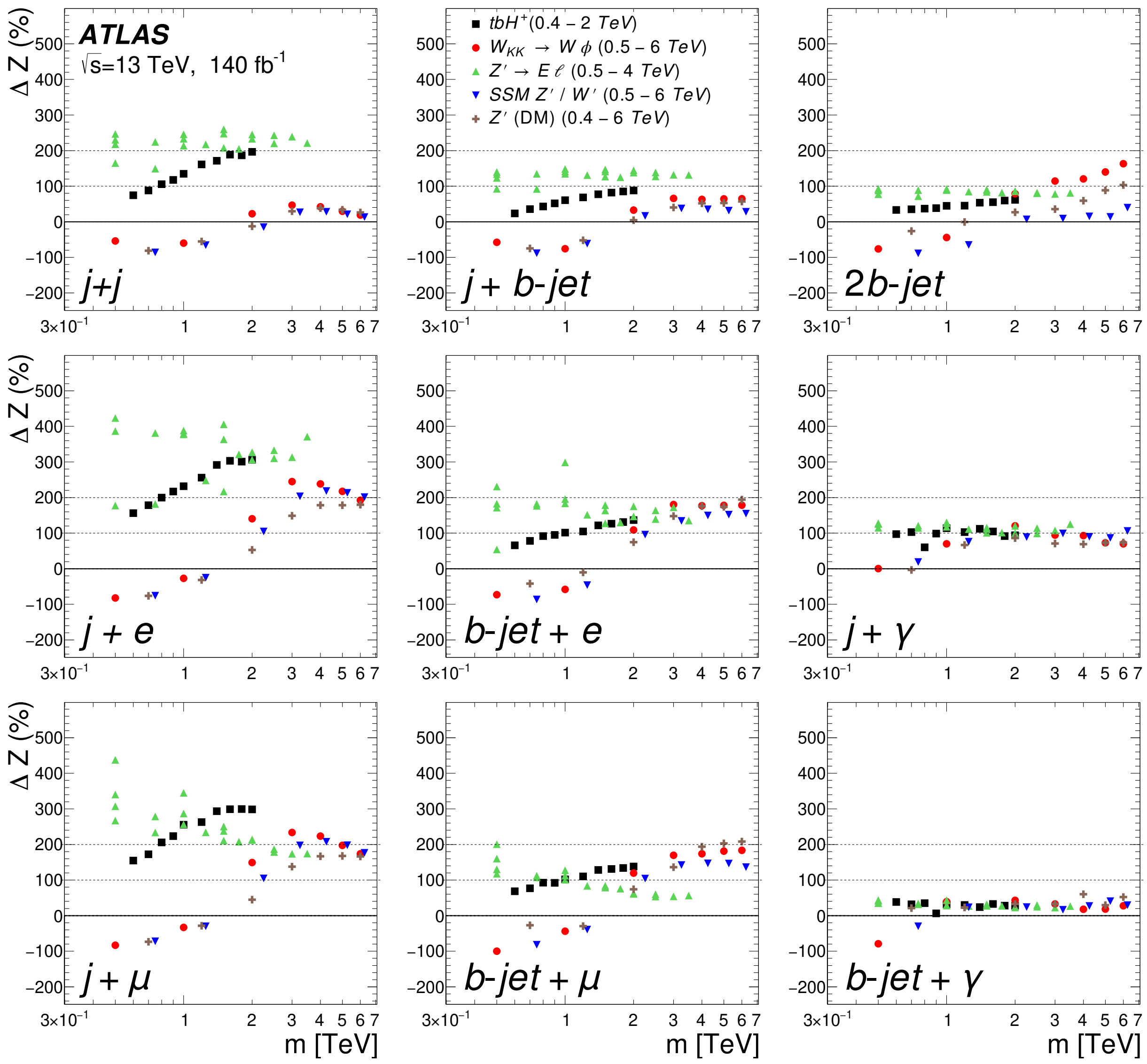
Discovery sensitivity of BSM models
(for 10 pb AR)

Sensitivity improvement quantified by ΔZ

$$Z = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

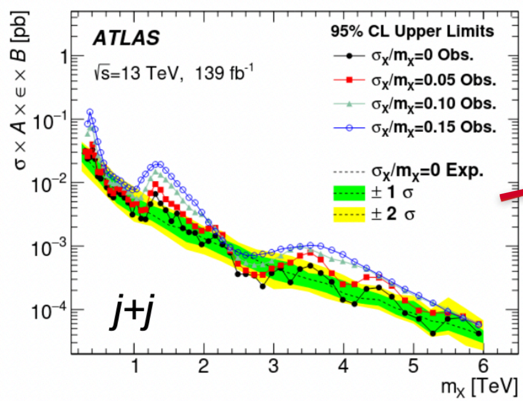
$$\Delta Z = ((Z_{AE}/Z)-1)\times 100\%$$

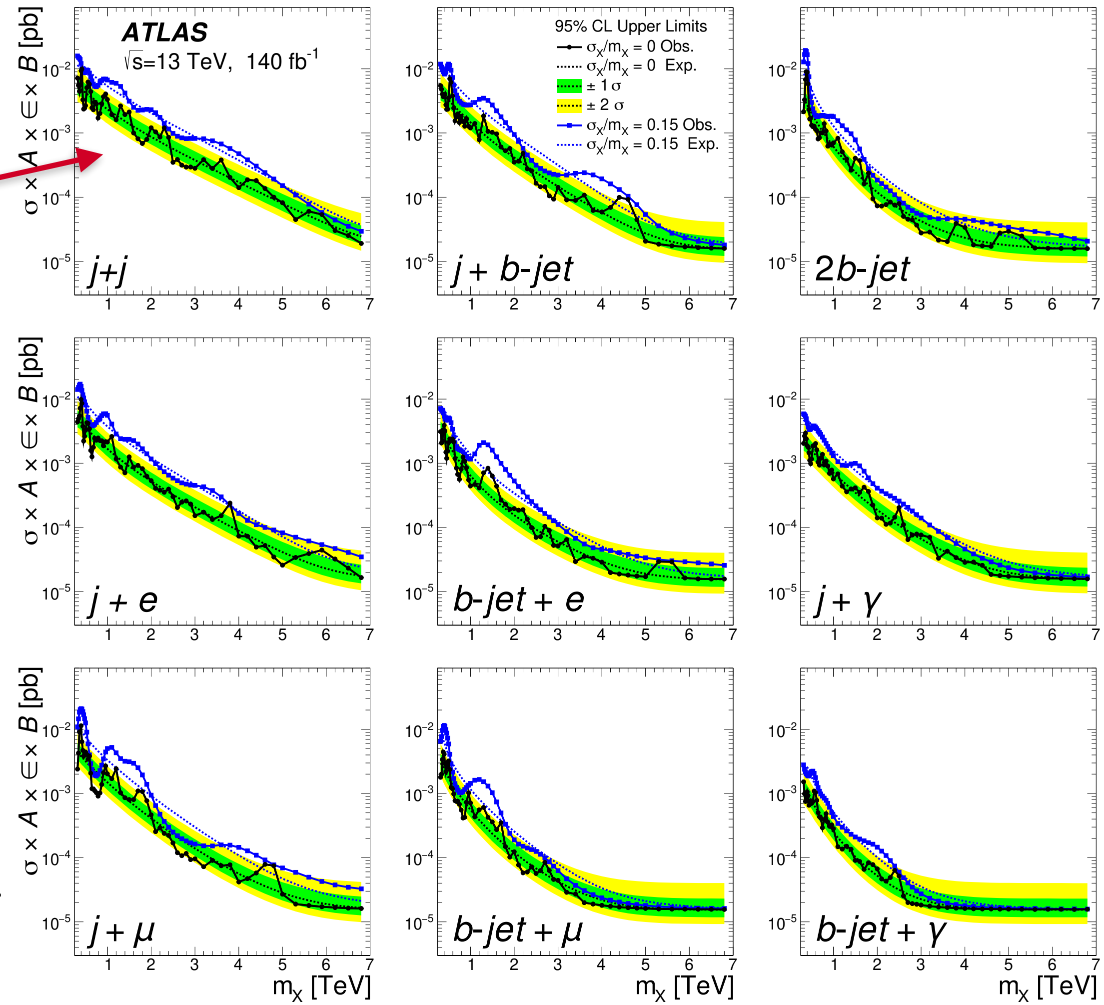
- Improvement tend to enhance with high mass as expected
- Higher pT, higher mass => more anomalous
- $\Delta Z > 0 \Rightarrow$ improvement
- Improvement directly translates into competitive limits



Upper limits of Gaussian signals

Published: JHEP 06 (2020) 151





- Signal width of $\sigma/m = 0\%$ and $\sigma/m = 15\%$ shown
- Mass points are spaced 5% apart from their preceding point starting from 0.3 TeV.
- Narrow signals have better limits as expected
- $\pm 1\sigma$, $\pm 2\sigma$ Error bands are from $\sigma/m = 0$ signals.
- Waves are similar, subject to local fluctuations
- Local significance 2.9 σ at $m_{j\mu} = 4.8$TeV, 2.8 σ at $m_{j\mu} = 1.2$TeV

- At masses < 1 TeV the limits are factor 2-3 better than for similar selection without autoencoder (JHEP 06 (2020) 151).

Summary & conclusions

- Searched for anomalies in 9 invariant masses in 3 Anomaly Regions in a mass range of 0.3 TeV - 8 TeV using full Run 2 data. Ref : arXiv:2307.01612 (Submitted to PRL).
- No statistically significant excess found in any channel.
- Largest deviation was found in j+μ channel near 4.8 TeV in 10 pb AR, with 2.9σ local excess.
- Limits have been set on generic gaussian signals of various widths and on different BSM physics scenarios.
- Model-independent limits have stronger potential to exclude generic heavy states with complex decays.
- Discovery sensitivity shows a large improvement after the anomaly region selection
- A successful application of unsupervised machine learning for anomaly detection using event level information has been demonstrated.

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

ATLAS
EXPERIMENT

Backup slides

Determination of Anomaly regions

We studied theoretical cross sections of various BSM models for two-body resonances (near 400 GeV mass scale). Some of the models we studied were :

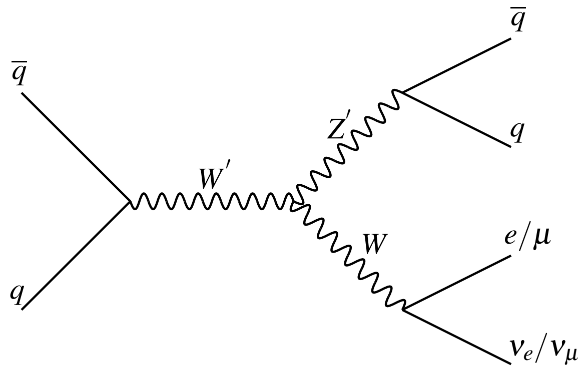
Sequential Standard Model
Charged Higgs (hMSSM, tan(β) = 1)
Charged Higgs (hMSSM, tan(β) = 0.5)
Simplified dark matter model
Composite lepton model (E=250 GeV)
Composite lepton model (E=500 GeV)
Technilocolr model
Radion model
Scalar Leptoquarks at B = 0.5
1st gen. scalar Leptoquarks
3rd gen. Leptoquarks
1st and 2nd gen. scalar Leptoquarks
SUSY with 1 lepton + jets
SUSY directRPVLQ (CC)
Heavy Higgs with lepton+jets

Anomaly region cuts of 10 pb, 1 pb & 0.1 pb could cover cross sections of most of the these models.

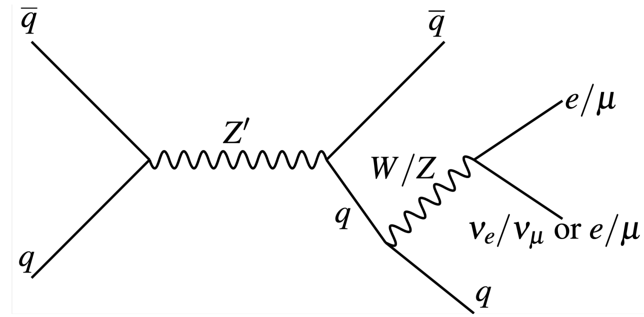
Three Anomaly Regions based on cuts on reconstruction loss from autoencoder :

1. 1st "BSM-region" (10 pb): Maximum value of the cross section for all BSM models
2. "2nd " region (1 pb)
3. 3rd "Data-region" (0.1 pb): Upper bound on published experimental limits for di-jet masses.

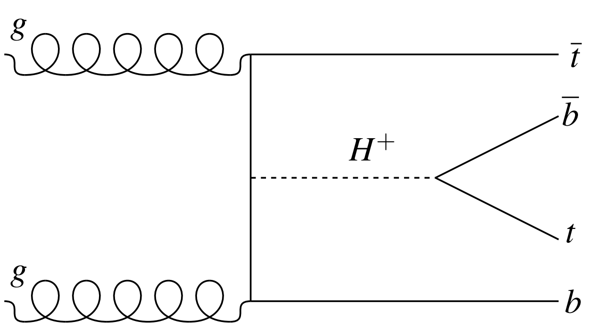
BSM models used



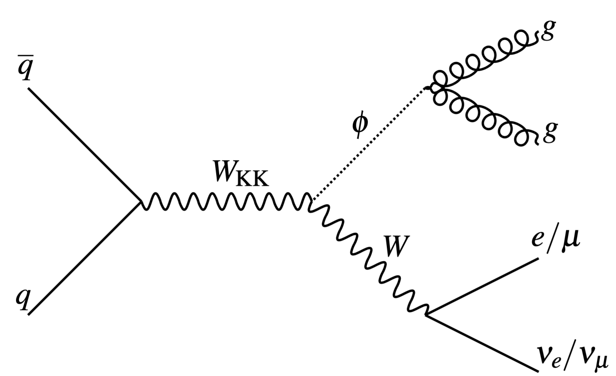
Sequential Standard Model



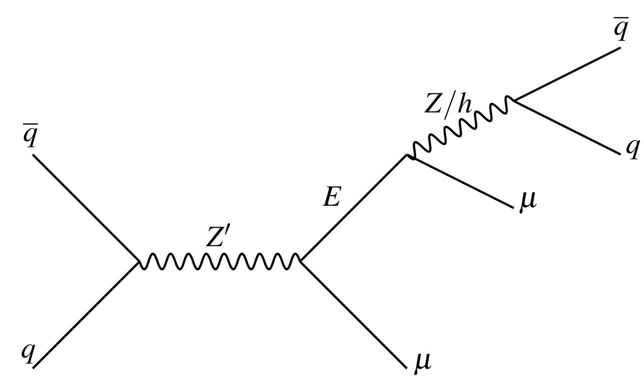
Simplified Dark Matter model



Charged Higgs model



Radion model



Composite lepton model

Usual LHC searches vs our anomaly-detection

Previous model-independent searches in dijet invariant masses (2009-Now)	Proposed strategy for searches in invariant masses using anomaly detection
1) Define “signal region” in terms of rapidity space y*=0.5*|y1-y2| or some other variable	1) Define “Anomaly region”, i.e. a region with large reconstruction loss of autoencoder trained on a fraction of data
2) Define background hypothesis for masses in signal region in the form of a monotonically decreasing function (or smoothing etc) for leading in pT jets	2) Define background hypothesis for masses in Anomaly region in the form of a monotonically decreasing function
3) Unblind: Run a likelihood fit on dijet invariant mass in the signal region and calculate local and global p-value for a largest deviation (“BumpHunter”)	3) Unblind: Run likelihood fits on invariant masses in the Anomaly region to calculate local and global p-value for largest deviation for all masses (modified “BumpHunter”)
4) Nothing found according to global p value? Run limits using Gaussian signal shapes or specific BSM models	4) Nothing found according to global p-value? Limits depending on availability of untested BSM models

Advantages:

No need Monte Carlo simulations

Agnostic to BSM & unexpected signatures

BumpHunter results for 10 pb WP

without Anomaly region cut