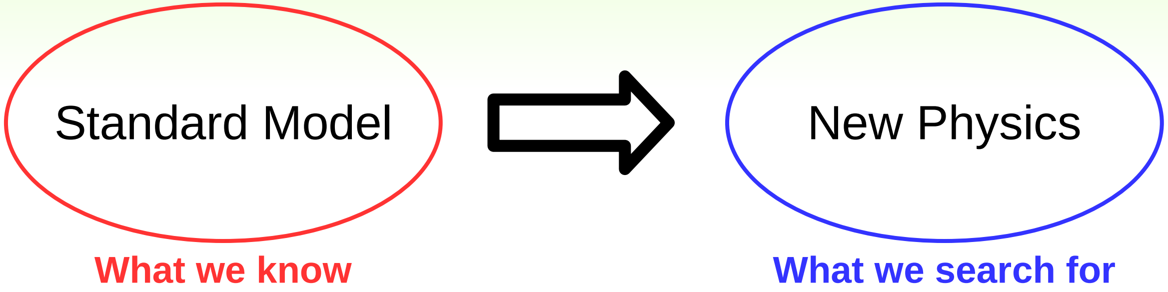
Anomaly Detection in LHC data with the GAN-AE algorithm

Presenter :

VASLIN Louis – Postdoctoral fellow at QUP

Context

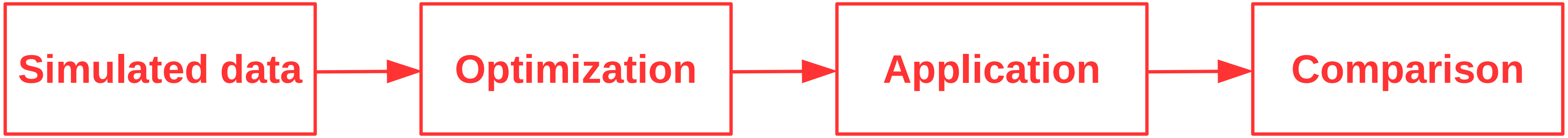
- Search for New Physics



<u>Looking for the unknown</u>

- Classical strategy

Heavy resonance search



Simulated data:
- <u>Signal hypothesis</u>
- *Computationally expensive*

Optimization:
- <u>Supervised</u> Machine Learning

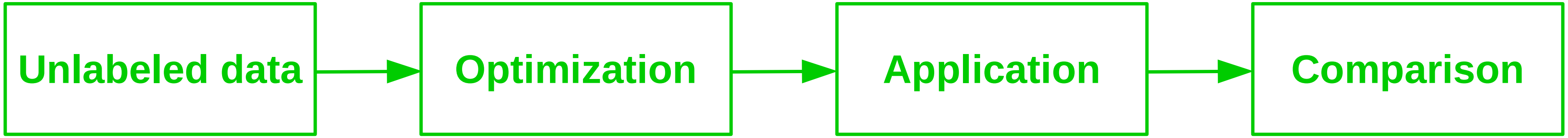
Application:
- Real data
- Background rejection

Comparison:
- <u>Simulated</u> reference background
- Look for deviations

New Approach

- New Analysis strategy

  Objectives :
  - No signal hypothesis
  - No simulation (data-driven background)



Unlabeled data
- Can be real data
- Signal is rare

Optimization
- *Unsupervised* Machine Learning
- **Anomaly detection**

Application
- Select anomalous events

anomaly
=
signal-like

Comparison
- Data-driven reference
- Model independent bump hunting

Anomaly detection algorithms

- Auto-Encoder

Objective :

learn **alternative representation** of data
for best **reconstruction**

Loss Function :

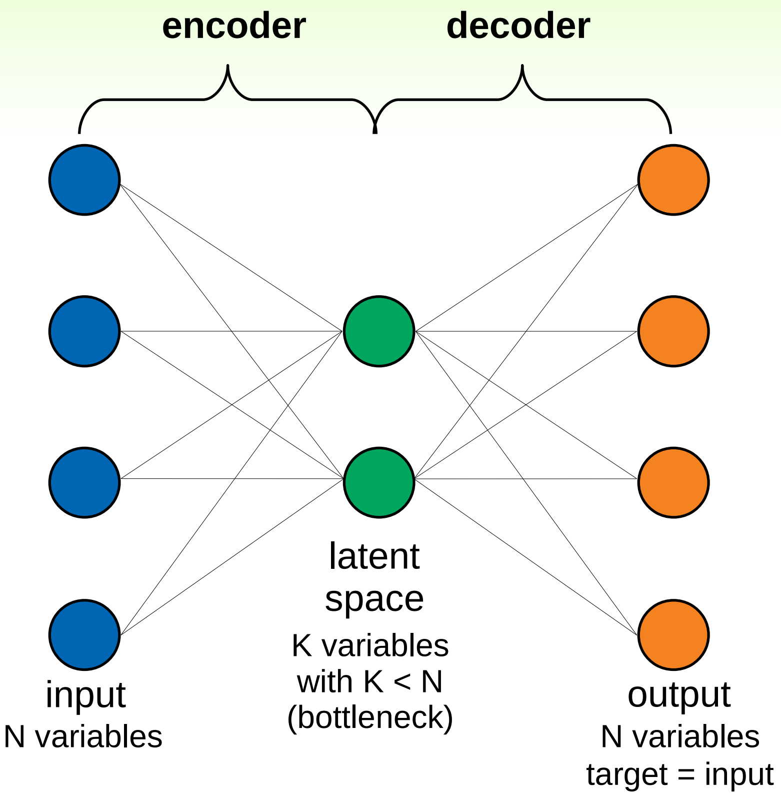
Reconstruction error (distance)

Application to anomaly detection :

common events => **good** reconstruction
rare event (anomaly) => **bad** reconstruction

=> Anomaly score

New Physics = anomaly



Anomaly detection algorithms

- GAN-AE

Adversarial model inspired by GANs

Objective :

Use the <u>discriminant output</u>
as an <u>additional constraint</u>

Two loss functions :

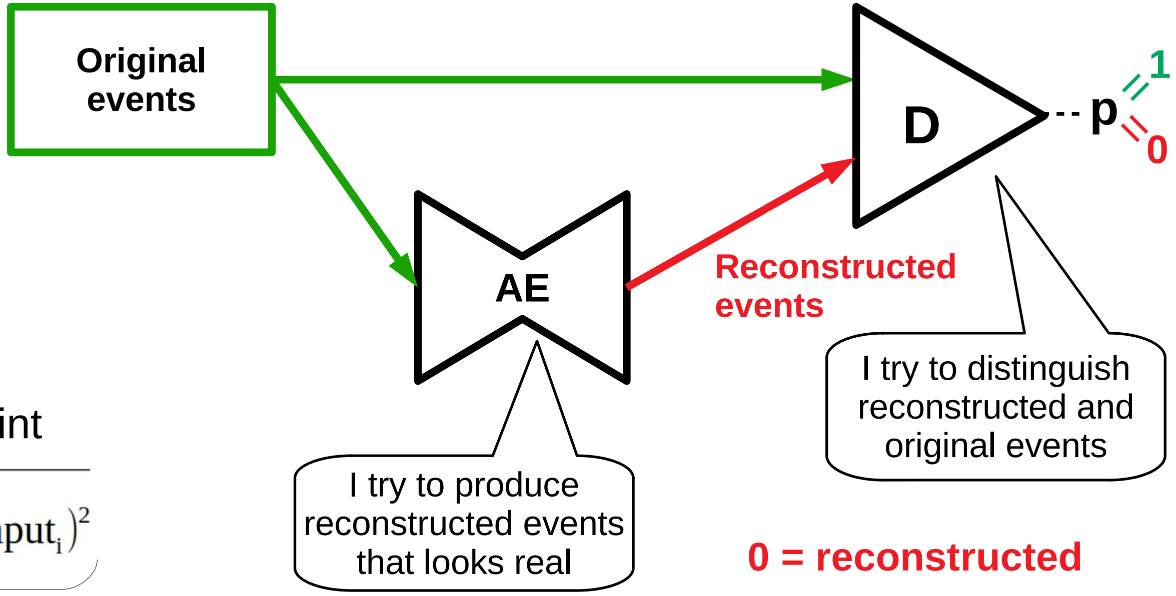
D loss = binary cross-entropy

$$bc = -\left(y \log(p) + (1-y)\log(1-p)\right)$$

AE Loss = reco error + D constraint

$$\text{loss} = \underbrace{bc_{(y=1)}}_{\text{D constraint}} + \underbrace{\varepsilon \times \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{output}_i - \text{input}_i)^2}}_{\text{reco error}}$$

reco error
=
<u>anomaly score</u>



0 = reconstructed

1 = original

Bacground modeling

- What reference background ?

Model independent search
=> **Data driven reference**

New Physics is rare
=> Data is mostly background

Before selection on anomaly score
=> Signal is invisible

Use data before selection ?

Bacground modeling

- What reference background ?

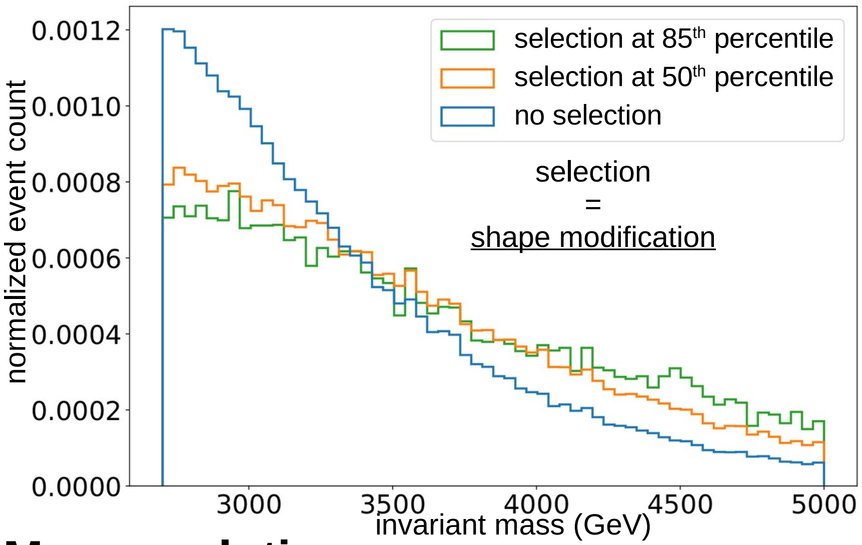
Model independent search
=> **Data driven reference**

New Physics is rare
=> Data is mostly background

Before selection on anomaly score
=> Signal is invisible

Use data before selection ?

Not straight forward …



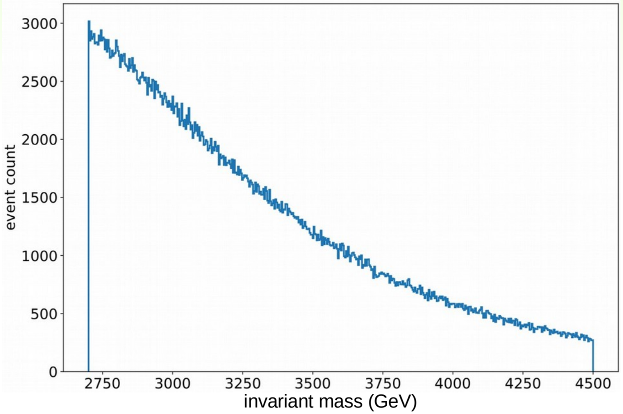
Mass sculpting
Must be avoided

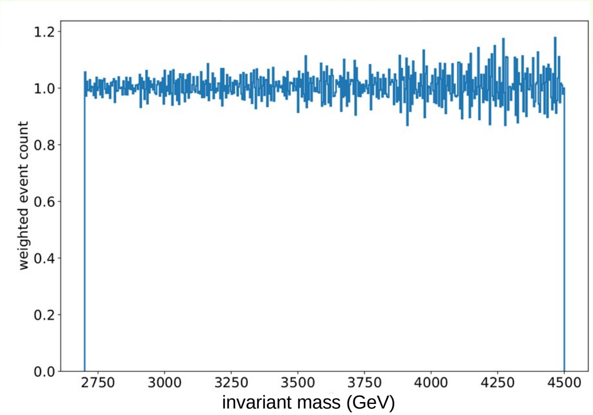
Mass sculpting mitigation

- Event reweighing method

  Compute **event weights** based on invariant mass

  Uniform distribution

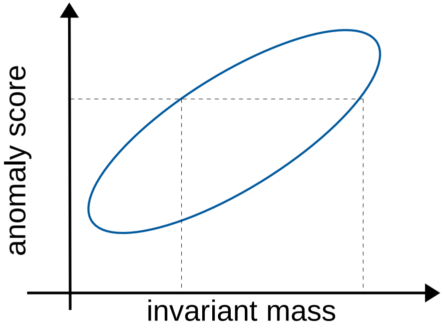


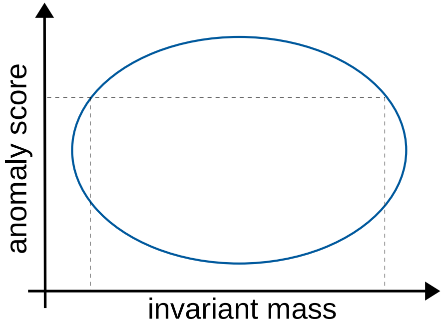


- Distance Correlation (DisCo) regularization

  Decorrelate invariant mass and anomaly score distributions

  Need resampling (batch computation)





Mass sculpting mitigation

- Modified loss function

<u>Apply weights to AE loss</u>

<u>Compute DisCo regularization on batch</u>

New AE loss expression :

$$\mathrm{loss} = \sum_{j=1}^{N_b} w_j \left(bc_{(\mathbf{y=1}),j} + \varepsilon \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{output}_{j,i} - \mathrm{input}_{j,i}\right)^2} \right) + \alpha \times \mathrm{DisCo}(X, Y)$$

number of events in batch

event weight

D constraint

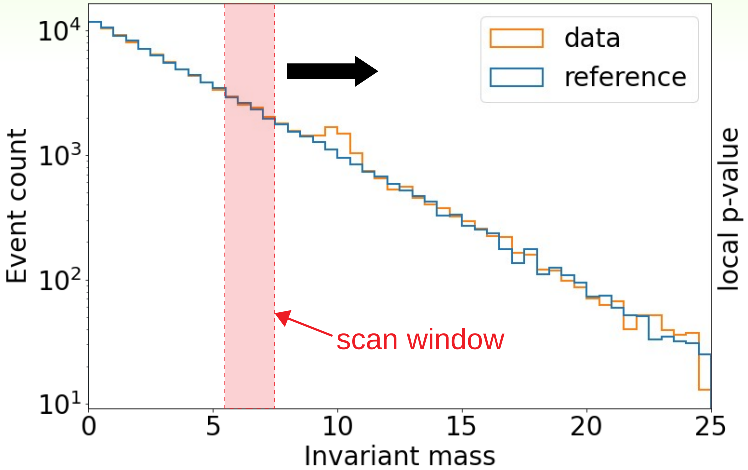
reco error

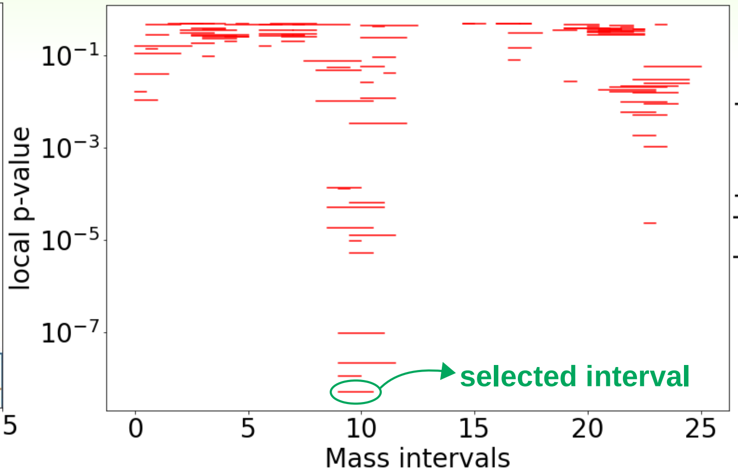
DisCo regularization

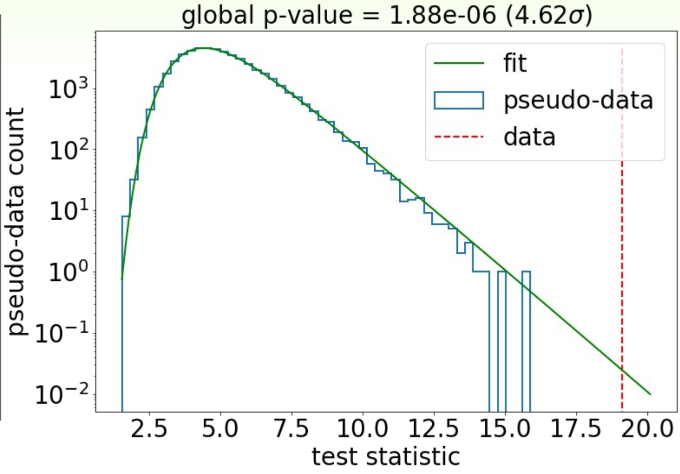
Allow for *<u>both</u>* **anomaly detection** and **background modeling**

BumpHunter

- Principle







Compare **data histogram** with **reference background**

Test deferent intervals of different width (**variable scan window**)

For each tested interval compute **local p-value**

Select **most significant deviation** (smallest local p-value)

Repeat process on **backgroung-only pseudo-data**

test statistic distribution

Compute **global significance**

BumpHunter

- pyBumpHunter

  New implementation in **python**

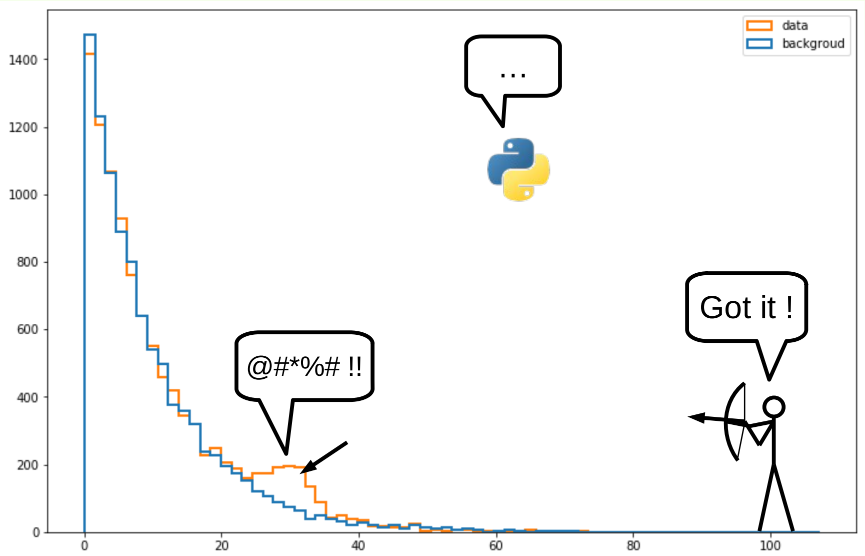
  Public release on [GitHub]() and [PyPI]()

  Included in [Scikit-HEP]() environment

  New features ([arXiv:2101.08320]() [arXiv:2211.07446]())

  - Automatic fit of test statistic distribution
  - Improved multi-channel combination
  - BumpHunter with 2D histograms
  - Side-band normalization
  - Sensitivity test by signal injection

  Ongoing development (more will come)



Application

- LHC Olympics 2020 challenge

Objective :

Develop <u>anomaly detection</u> algorithm for *New Physics* search

Contribution to community paper (arXiv:2101.08320)

Open dataset :

Simulated data with ATLAS-like settings

RnD data

dijet/trijet data

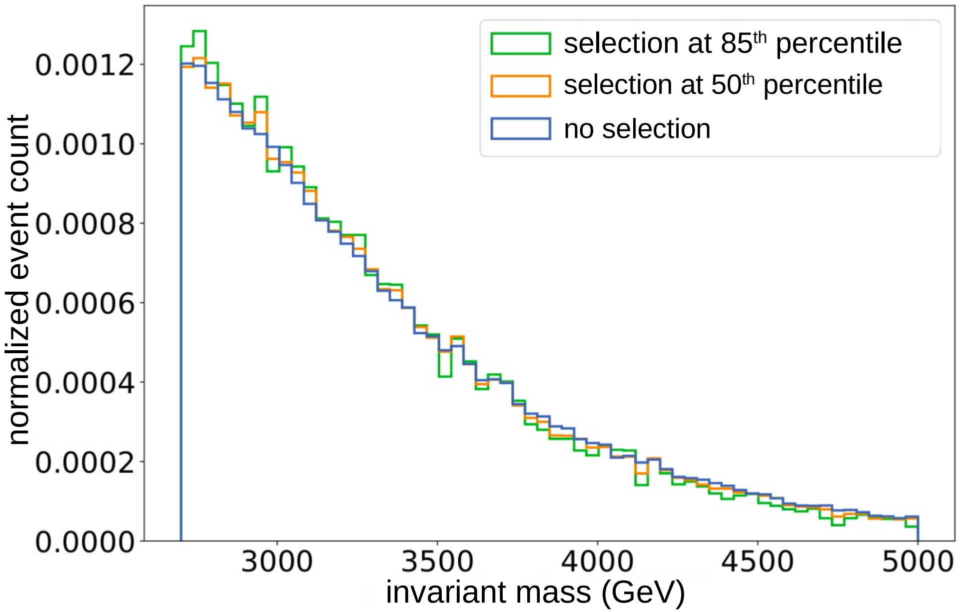
1 background + 2 signal samples

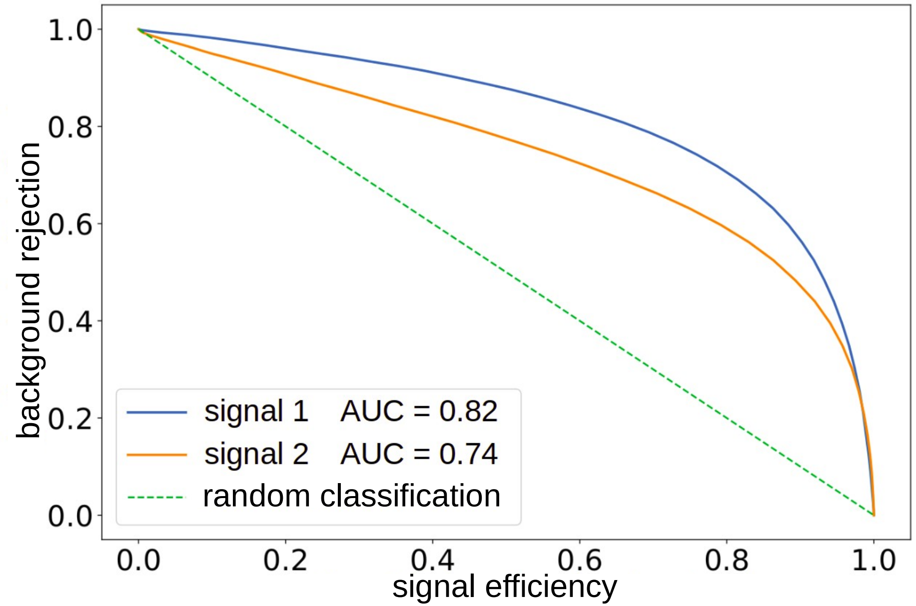
Black Box data

3 unknown sample with jet events (no label available)

Results

- Results on RnD data





Test on background-only sample

Apply selection at different threshold

Mass sculpting is negligible

Test on background/signal mixture

Test different selection threshold

Signal and background separation

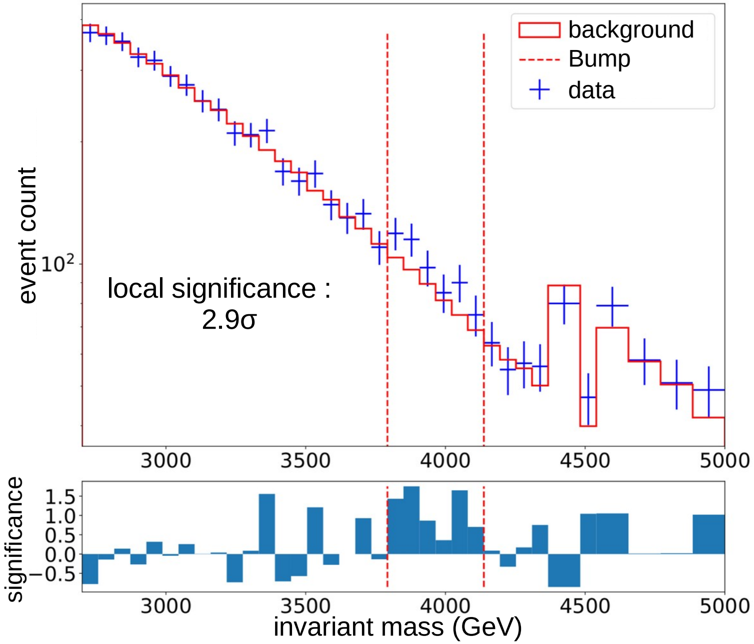
Results

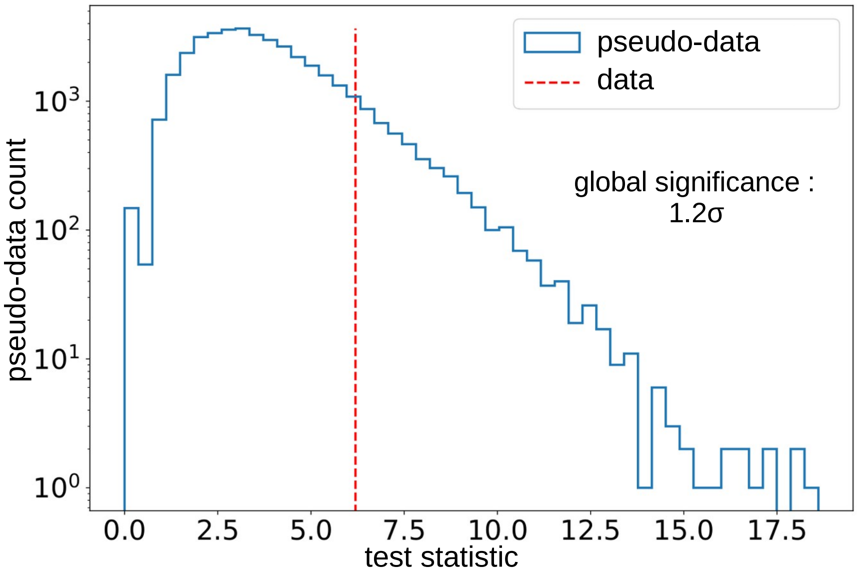
- Results on 1st Black-Box data

Training on 100k events

Application to <u>all black box events</u>

Selection at 99th percentile of anomaly score





Bump identified

mass = <u>3.97 TeV</u>

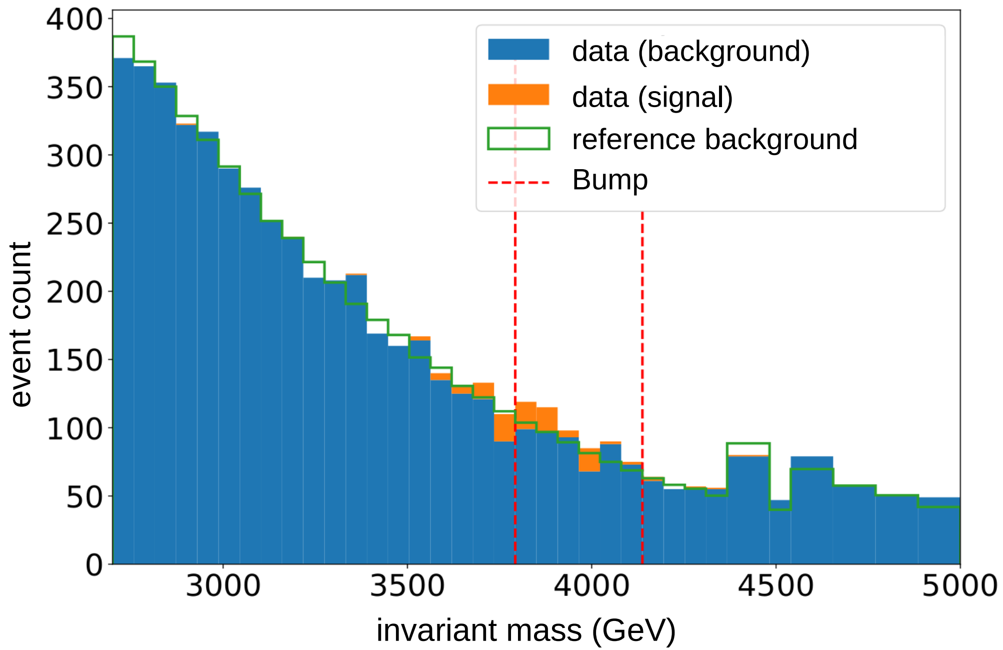
<u>Good background modeling</u>

Almost <u>3σ local</u>

=> signal hint

Results

- Results on 1st Black-Box data



Check results

Real labels revealed *after the challenge*

Solution

There were a signal

Same topology than RnD signal 1

mass = 3.8 TeV

Conclusion

Signal identified with **good mass precision**

Signal efficiency > 15%   S/B ratio **x 20**

No mass sculpting

Summary

- New analysis strategy

  Use **anomaly detection** based on unsupervised Machine Learning

  Data-driven background modeling with mass sculpting mitigation techniques

  Model independent bump hunt with improved version of BumpHunter

- Application to LHC Olympics 2020 dataset

  Good background modeling

  Improvement of signal significance

  Complete strategy

Thank you !

ありがとうございます！

BACKUP

GAN-AE training

- 1st training step

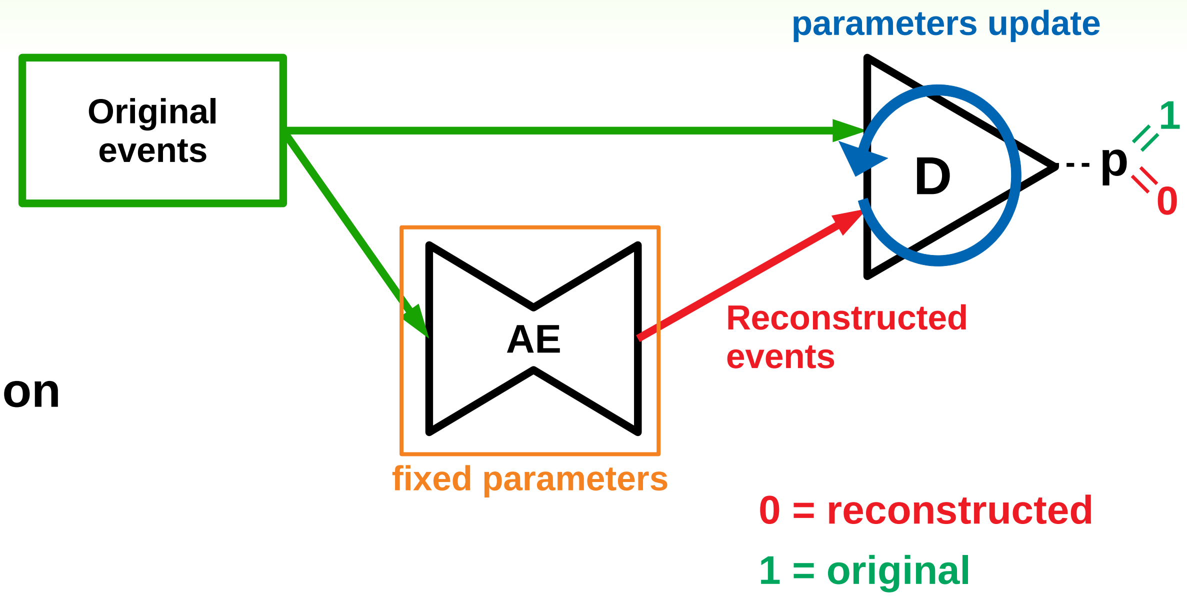
  Train parameters of discriminant

  Keep parameters of AE fixed

  Mix of original and reconstructed events

  => Usual binary classification

  Repeat for a few epochs



GAN-AE training

- 2^{nd} training step

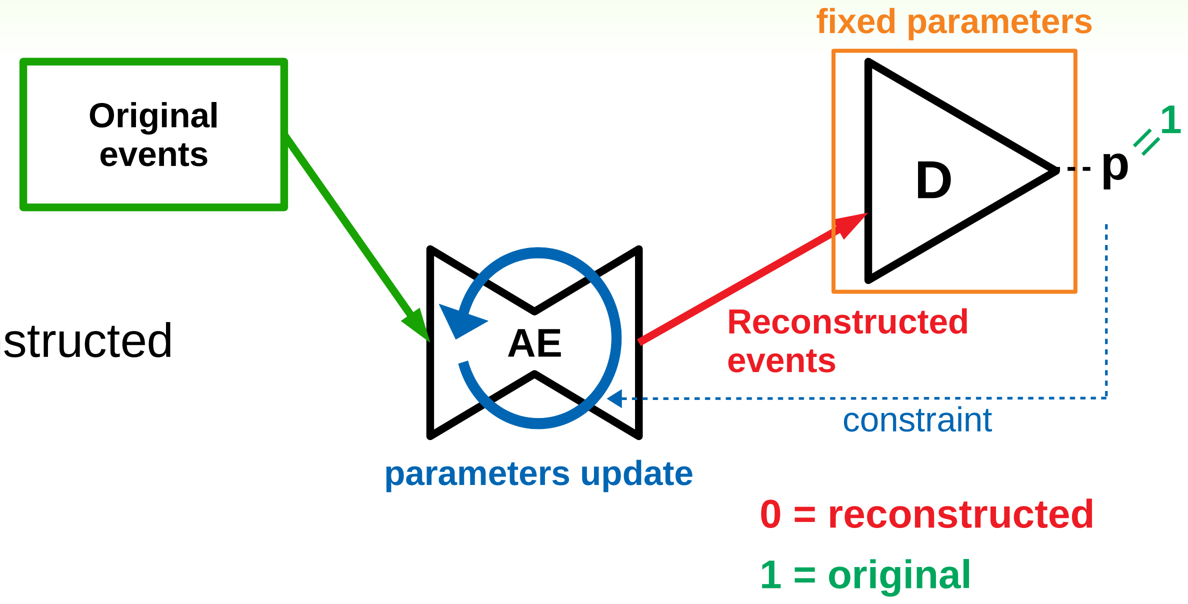
  Train parameters of AE

  Keep parameters of D fixed

  Only original events

  => Try to make D label reconstructed events as original events

  Repeat for a few epochs



DisCo regularization

Condition of independence

Distribution of X and Y are independent => $f_{XY} = f_X f_Y$

Distance covariance

Measure of independence between X and Y (arXiv:0803.4101)

$$dCov^2 = \| f_{XY}(s,t) - f_X(t) f_Y(s) \|^2 = \int \left| f_{XY}(s,t) - f_X(s) f_Y(t)^2 \right| w(s,t)\, dt\, ds$$

Empirical form (arXiv:2001.05310)

$$dCov^2(X,Y) = \langle |X-X'||Y-Y'| \rangle + \langle |X-X'| \rangle \langle |Y-Y'| \rangle - 2 \langle |X-X'||Y-Y''| \rangle$$

<.> = mean |.| = Euclidean norm

Distance Correlation (DisCo)

$$DisCo(X,Y) = \frac{dCov^2(X,Y)}{dCov(X,X)\, dCov(Y,Y)}$$

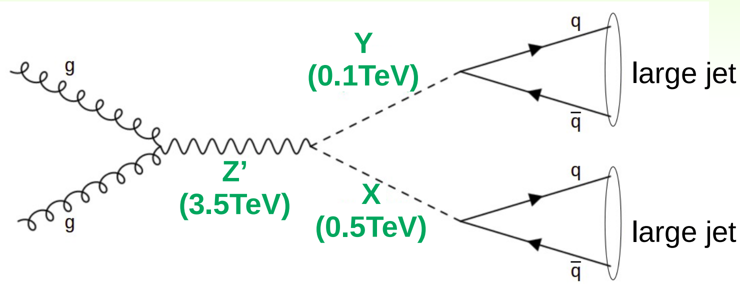
LHC Olympics 2020 data

- RnD data

  Background : QCD (multijet)

  Signal 1 : Z' → XY → (qq)(qq)

  Signal 2 : Z' → XY → 3-jet-like

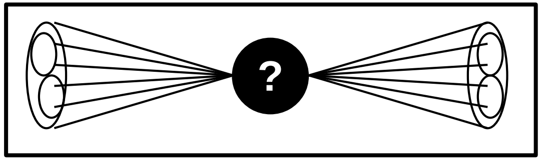


- Black Box data

  3 sample with unlabeled data

  Different simulation settings (*background is not same as RnD*)

  Objective : Find if there is a singal hidden in Black Box samples



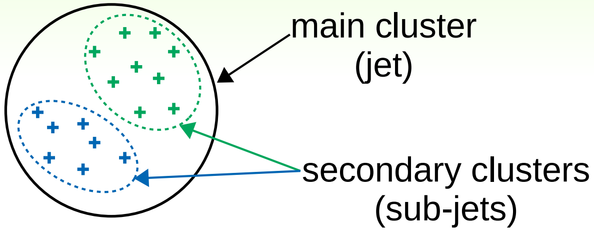
Variables

- Clustering

  Up to 700 raw jet constituents

  Clustering with anti-Kt algorithm (FastJet)

  2 step clustering (2 mains jet + subjets)



- General features

  Jet 4-vectors (E, p_T, η, φ)

  Jet mass

  Number of jet constituent

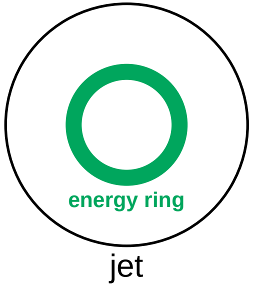
  dijet/trijet invariant mass

- Substructure variables

  Number of sub-jets

  N-subjetiness (3 per jets + 2 ratios)

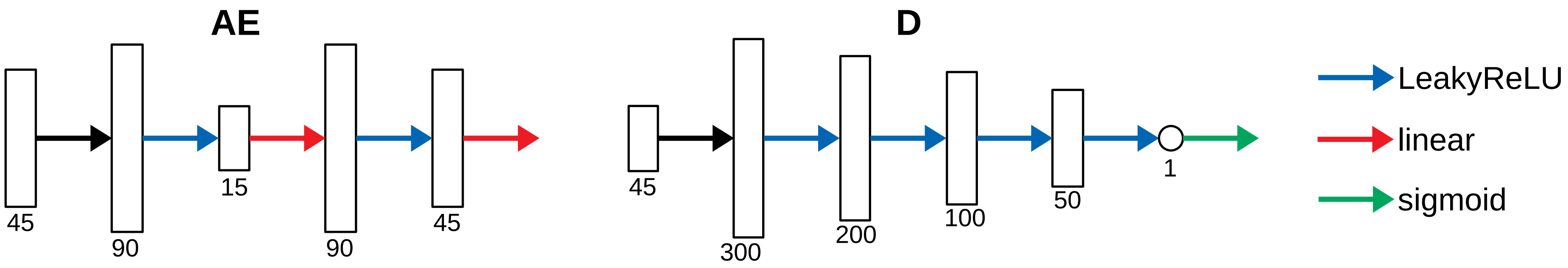
  Energy rings



Model hyperparameters

Preparing the GAN-AE

Network architectures (for **dijet** clustering)



Training hyperparameters

Number of cycles : 100

D epochs per cycle : 7

AE epochs per cycle : 5

Event per batch : 2048

Pretrain AE for 5 epochs

Dropout on hidden layers : 20%

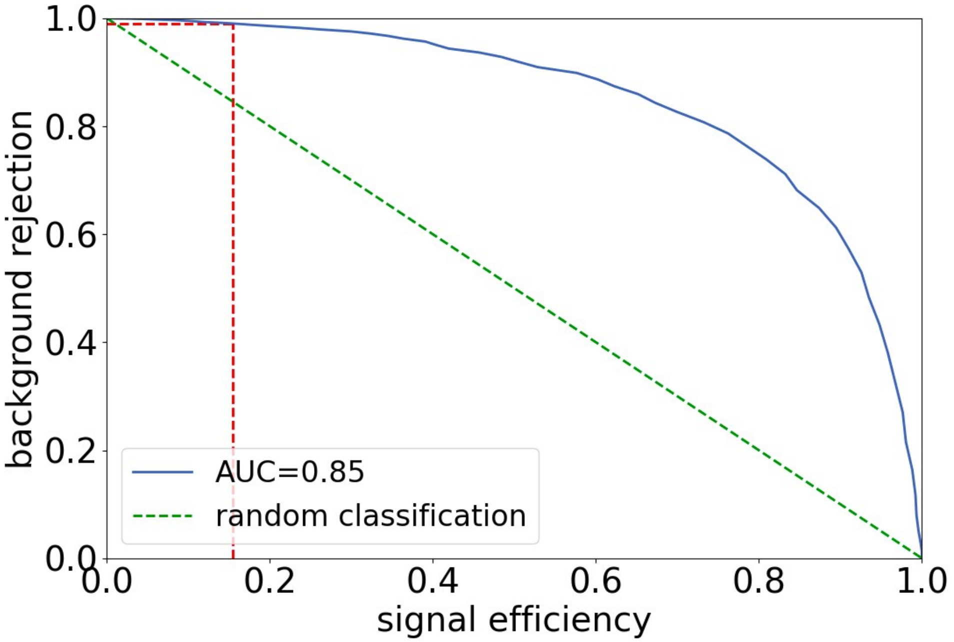
Loss hyperparameters

Reconstruction error (ε) : 6.0

DisCo regularization (α) : 65.0

Complementary results

- Black Box 1 ROC curve



Red lines correspond to working point (99th percentile selection threshold)

Better AUC than what we obtained on RnD

Complementary results

- Summary of all Black Box samples

	Bump mass	Local significance	Global significance	True mass
Black-Box 1	3.97 TeV	2.9σ	1.2σ	3.8 TeV
Black-Box 2	3.31 TeV	1.5σ	–0.43σ	X
Black-Box 3	3.77 TeV	1.5σ	–0.94σ	4.2 TeV

Black-Box 2

<u>No signal</u>

Global significance **very low**

=> <u>compatible with fluctuations</u>

No fake signal

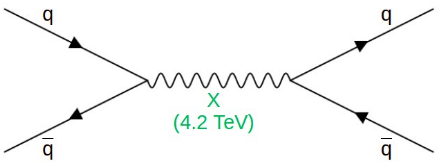
Black-Box 3

<u>Complex signal</u>

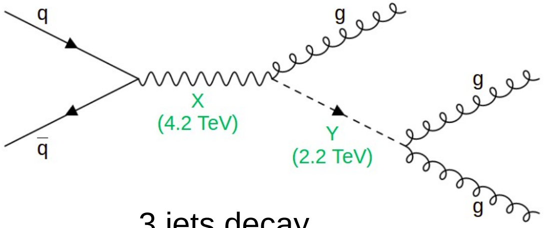
=> particle with 2 **decay modes**

Fail to identify the signal

Dijet search not optimal



2 jets decay
37.5%



3 jets decay
63.5%