量子場計測システム国際拠点
International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles
WPI research center at KEK

Louis Vaslin (QUP Postdoctoral fellow)

wpi KEK

Model-independent strategy for New Physics search at LHC using Anomaly Detection algorithms

Model Independent search strategy
1

– Limits of classical search strategy

Signal specific
=> Narrow field of search

Simulation dependent
=> Time and power consuming
=> Need better accuracy

– New strategy proposal

Signal agnostic
=> More generic search

Data-driven
=> Reduce simulation dependency

Use Anomaly Detection algorithms
=> Unsupervised Machine Learning
=> Model-independent bump hunt

Unsupervised Machine Learning algorithm
2

– GAN-AE[1]

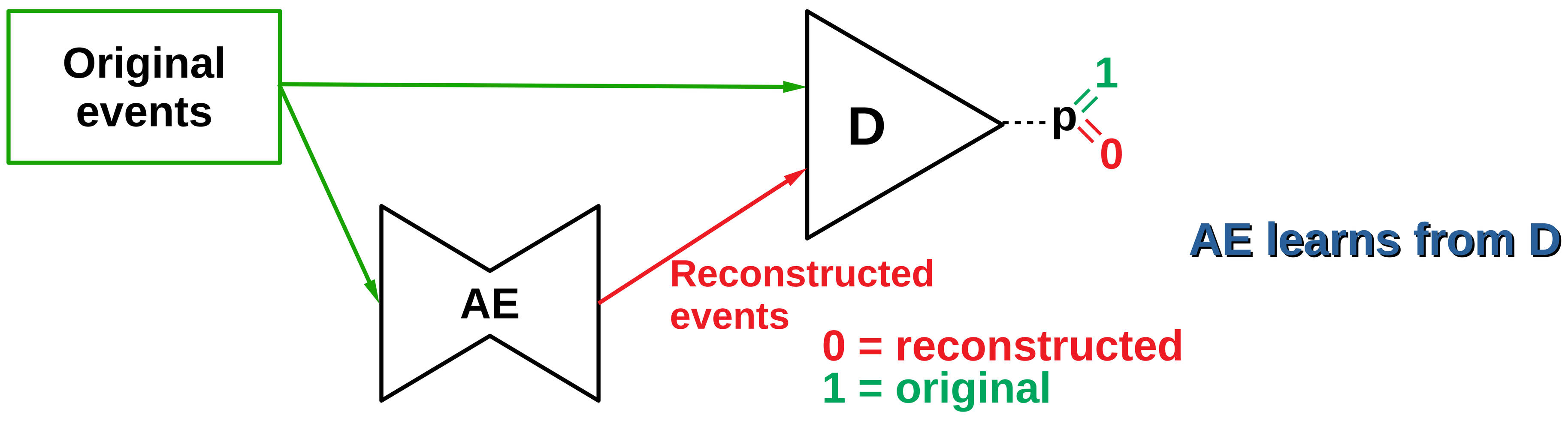
Combine a **Auto-Encoder** and a **Discriminant**

Objectives

AE : Reconstruct events as accurately as possible
D : Discriminate reconstructed and original events

Loss

D : Binary cross-entropy (classification)
AE : **Combine loss**
reconstruction error **+** discriminant information



Schematic of the GAN-AE architecture

Training and application
3

– Loss and anomaly score

D loss : $bc = -(\underbrace{y}_{\text{target}} \log(\underbrace{p}_{\text{output}}) + (1-y)\log(1-p))$

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

Information from D
Use D loss with '*switched labels*'

Reconstruction error
Used as anomaly score

– Application

Training **directly on data**

Use only AE for application
D is used as a *training proxy*

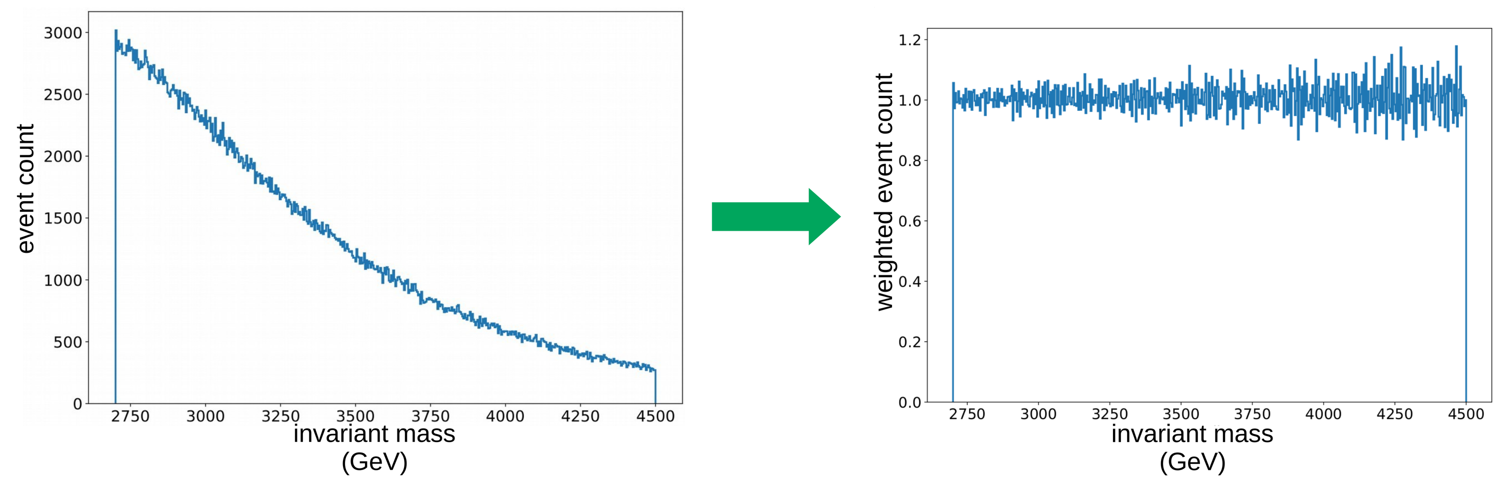
Select event with **high anomaly score**
rare New Physics signal => anomalous events

Compare anomalous data with a *reference background*
Need a proper model

Mass sculpting mitigation
4

– Mass-based event reweighing

Flatten the mass distribution to reduce bias



– Distance Corelation regularization (DisCo)

Decorrelate anomaly score and invariant mass distributions

Mass distribution **invariant** when applying selection on anomaly score

Allow for data-driven background modeling

Model independent bump-hunting
5

– pyBumpHunter[2]

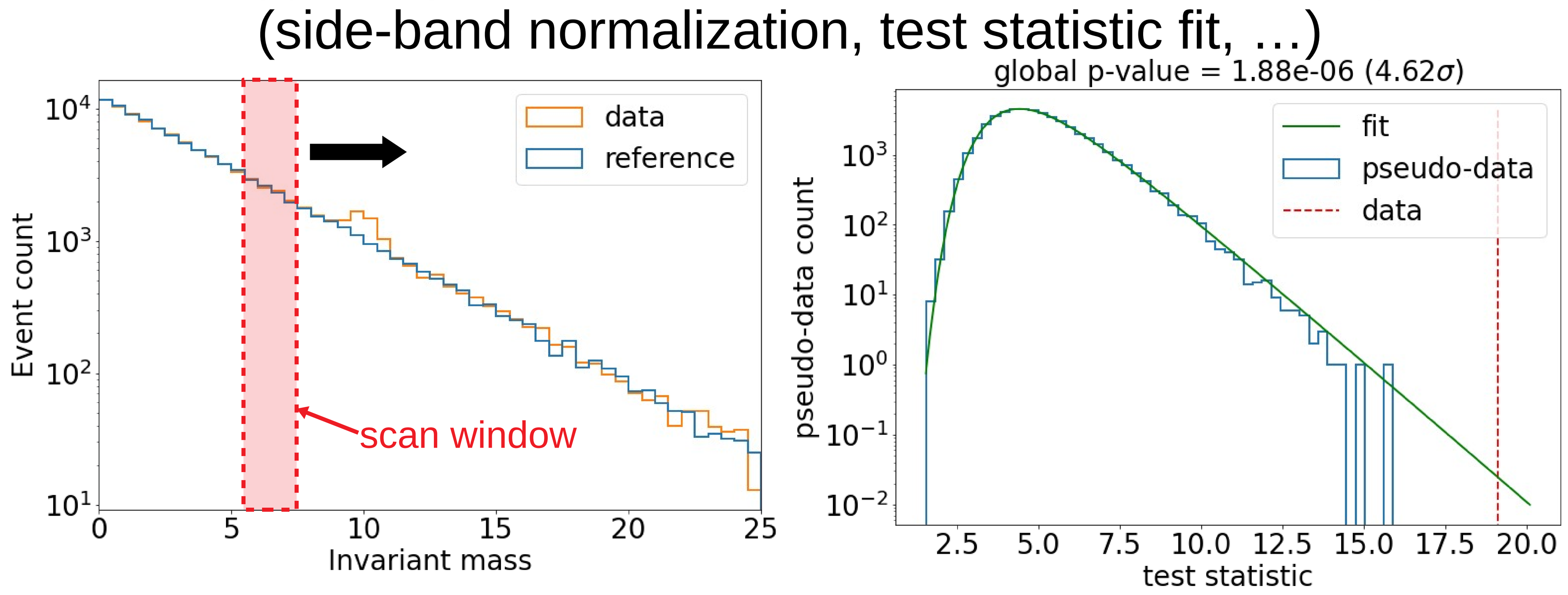
New implementation of the BumpHunter algorithm in python

Compare data histogram with a reference background

Look for the interval with lowest local p-value

Compute **global p-value** using background-only pseudo-data histograms

pyBumpHunter includes ***new features*** to the algorithm (side-band normalization, test statistic fit, …)



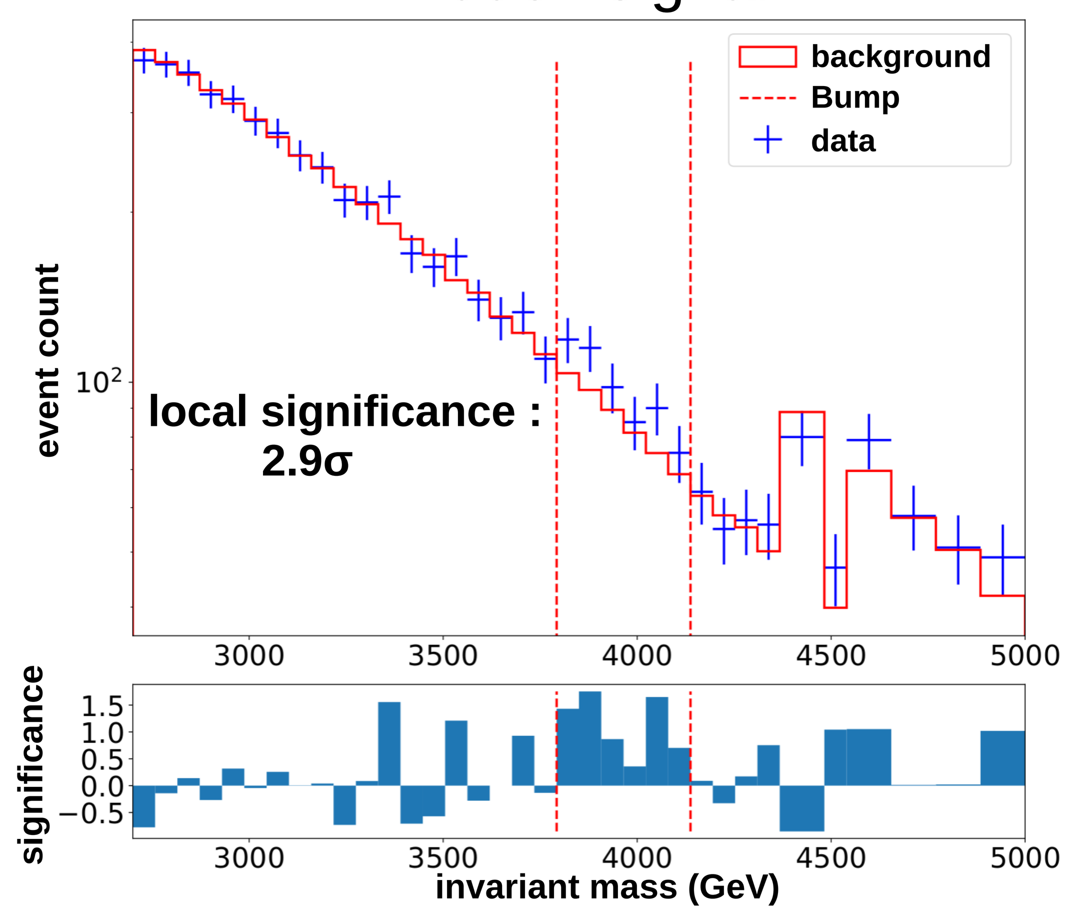
Results
6

– LHC Olympics 2020 dataset[3]

Community challenge to promote **Anomaly Detection algorithms**

Black box dataset :

Background dominated (multijet)
Hidden signal : Z' → XY → (qq)(qq) → 2 large jets



Model trained directly on unknown data

Deviation identified with almost 3σ local significance

True signal mass is 3.8 TeV
=> Correctly identified

No major deviation *outside of selected mass interval*
=> Good background modeling

References

[1] Vaslin, L., Barra, V. & Donini, J. GAN-AE: an anomaly detection algorithm for New Physics search in LHC data. Eur. Phys. J. C 83, 1008 (2023)
[2] Vaslin L. & Calvet S. & Barra V. & Donini J. pyBumpHunter: A model independent bump hunting tool in Python for high energy physics analyses. SciPost Phys. Codebases 15 (2023)
[3] Kasieczka G. & Nachman B. & Shih D. & al. The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. Rep. Prog. Phys. 84 124201 (2021)