Détection d’anomalies en physique des particules

Groupe IA du programme Data - 19 Octobre 2023

Julien Donini - UCA / LPC

UNIVERSITÉ Clermont Auvergne & associés

cnrs IN2P3 Les deux infinis

LPC Particules pLasmas Univers applicationS

Plan : détection d’anomalies au LHC

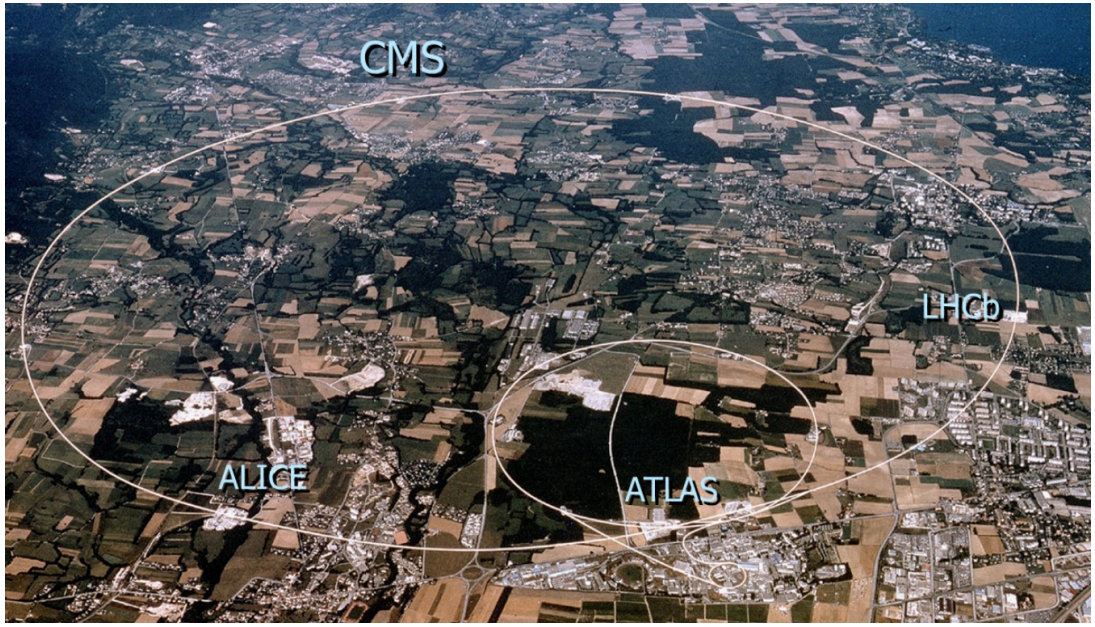
Contexte en physique des particules

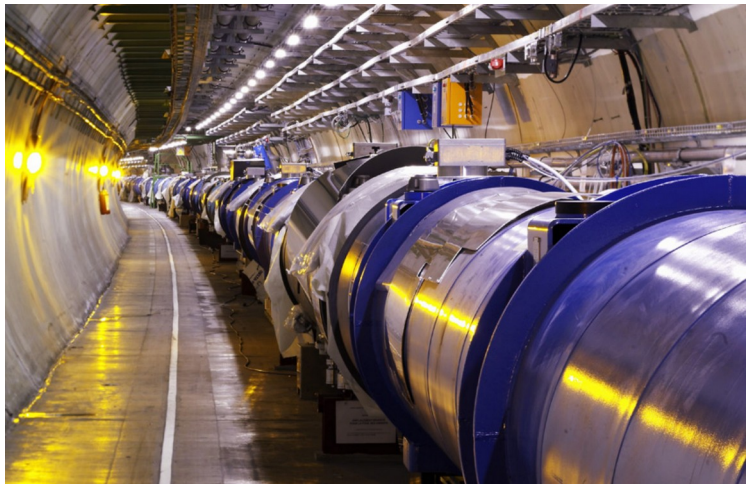
Approches développées localement

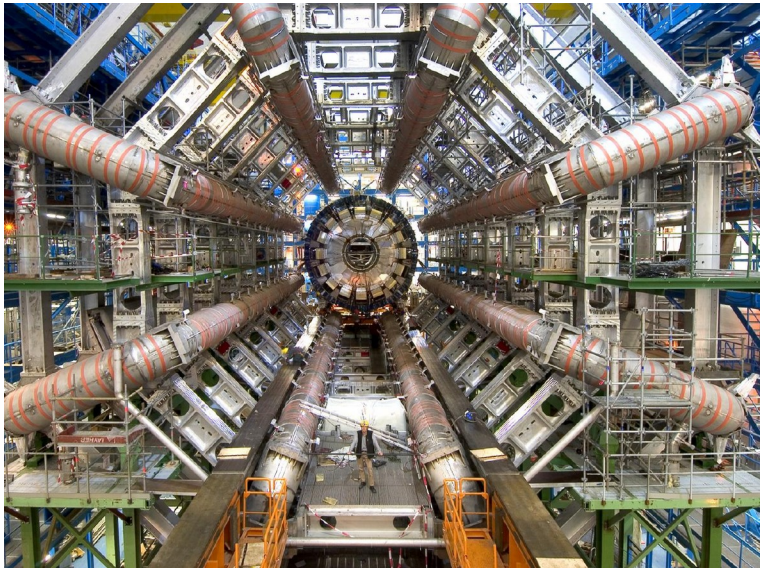
- GAN-AE
- DDP

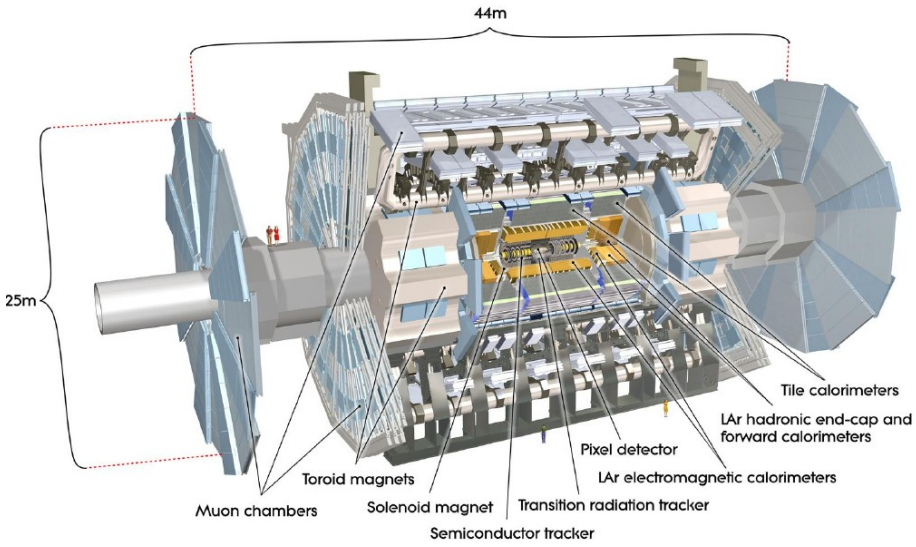
Perspectives

Le LHC en 1 slide

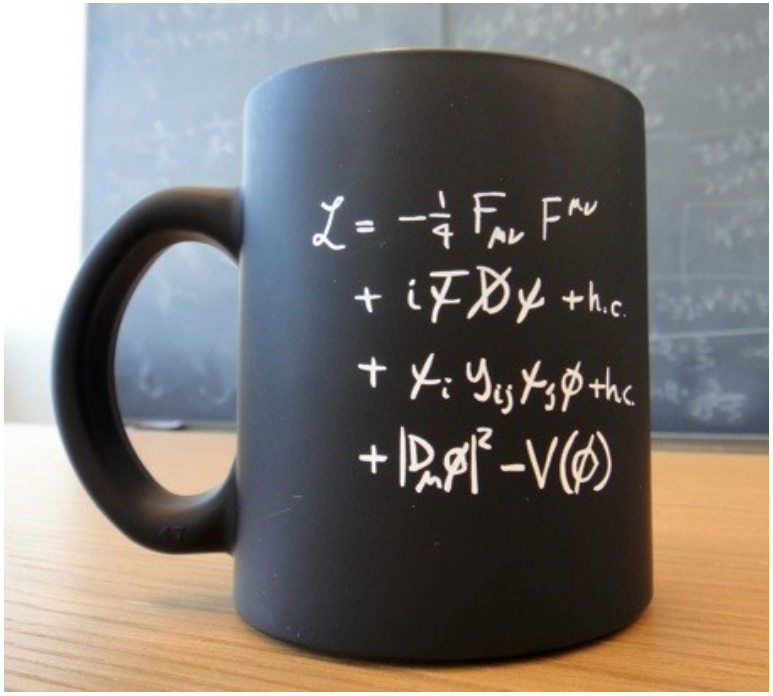


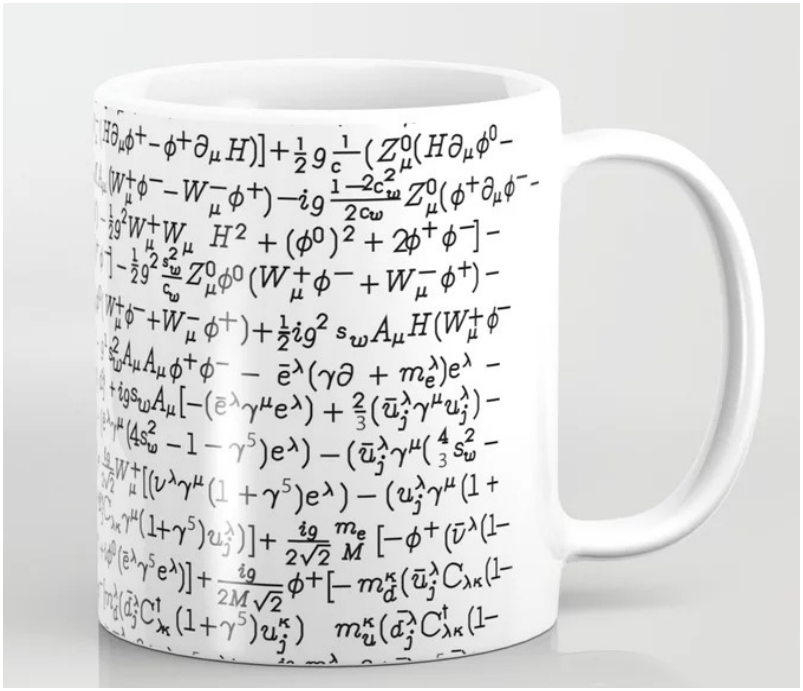






La physique des particules en 2 mugs



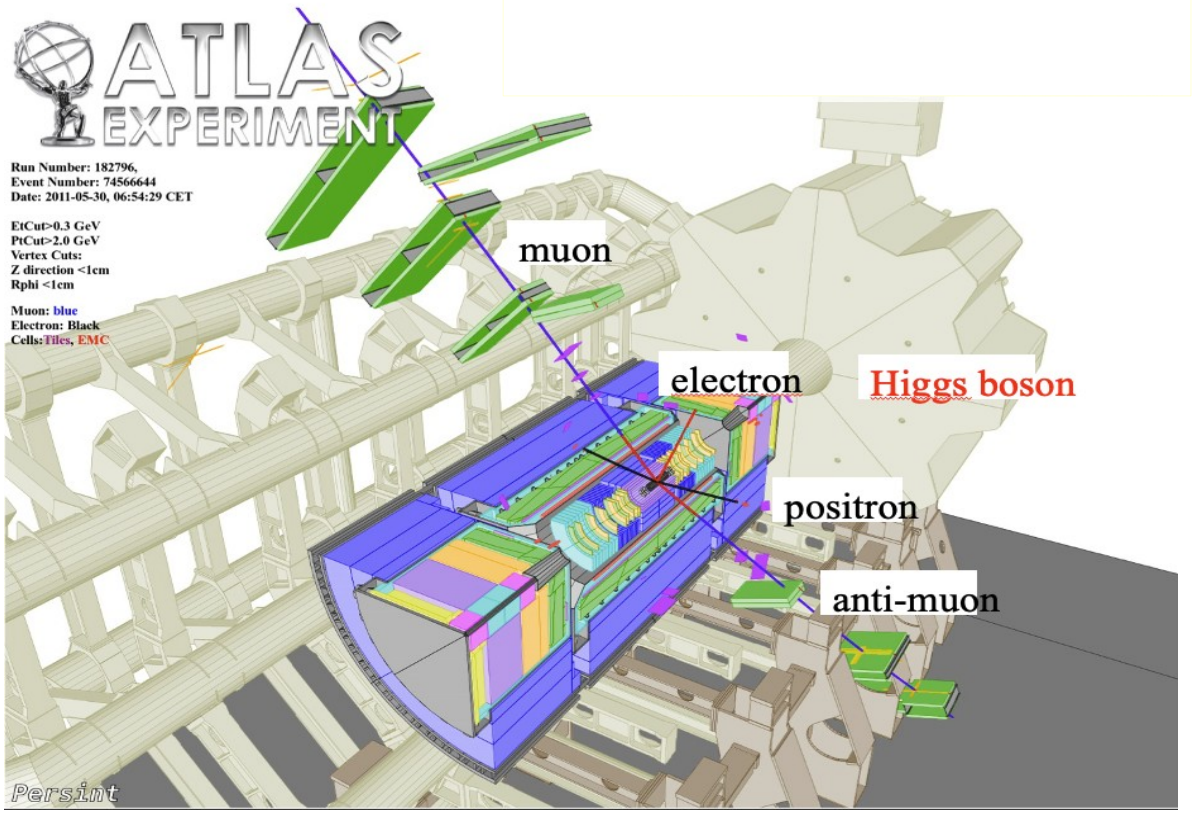


+ de 50 ans de **succès théoriques et expérimentaux**

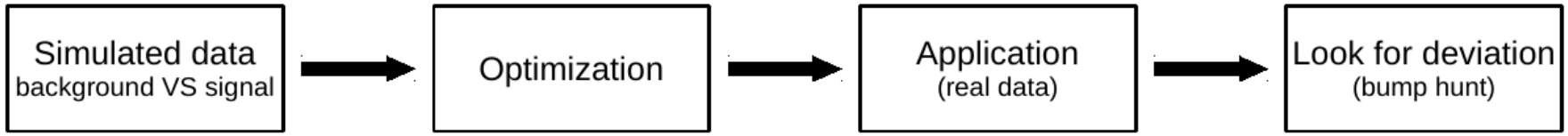
… mais **théorie incomplète** : « nouvelle physique » ?

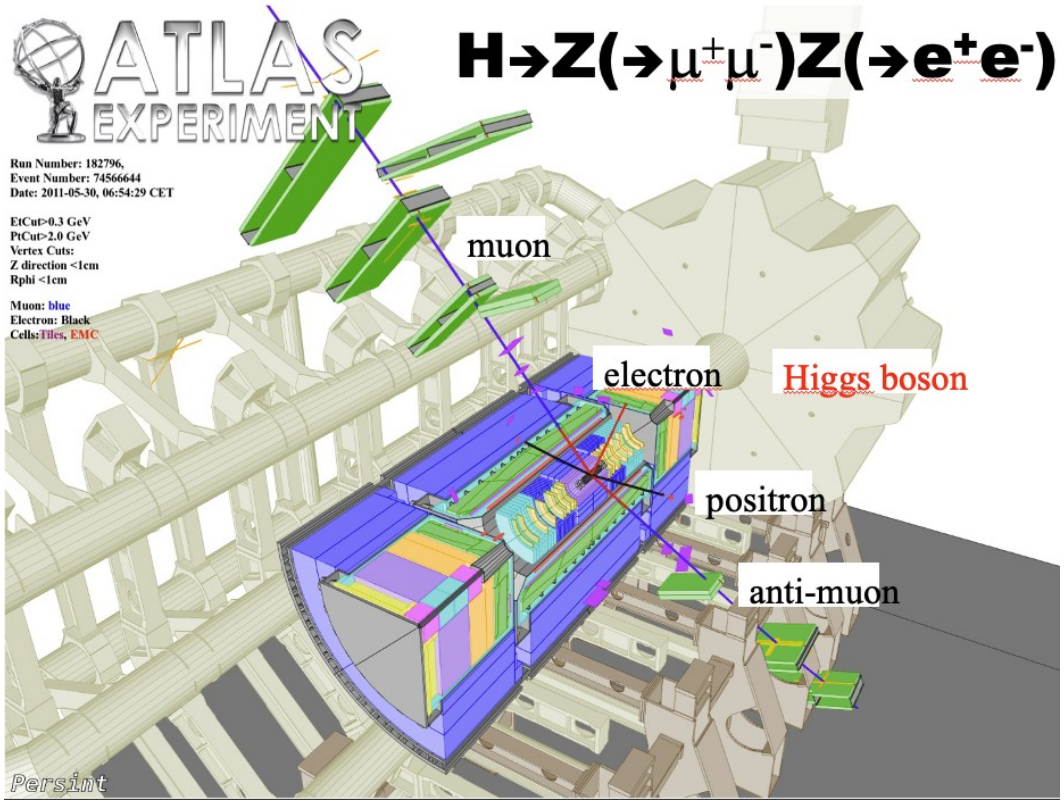
Approche “classique”

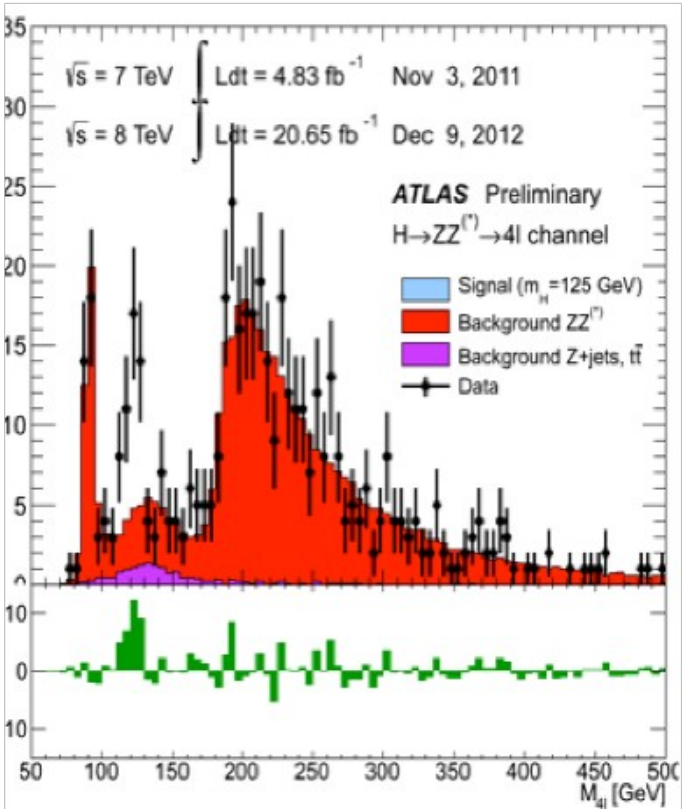
Exploration d’un **grand nombre de signatures** pour rechercher dans un espace à **grande dimensionnalité** un signal par rapport à un bruit de fond



Approche “classique”



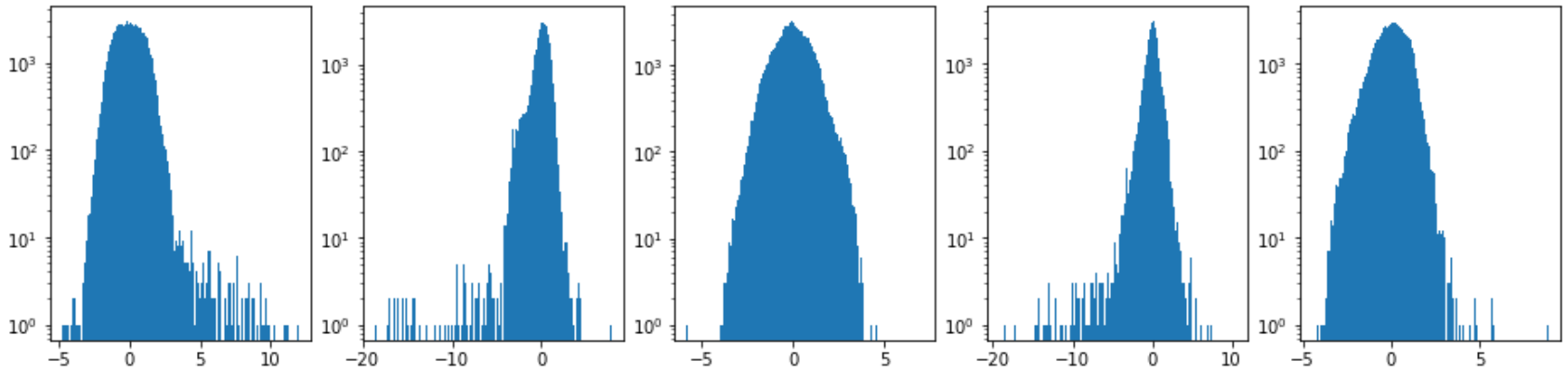




Anomalies au LHC

Tout nouveau phénomène physiques **non prévu** par la théorie !

- Signal **rare** et bruit de fond complexe
- Nécessite d'explorer des **données multidimensionnelles**
- **Approches** Machine Learning indispensables



Présence d'anomalies ?

Détection d'anomalies au LHC

HEPML-LivingReview

Anomaly detection.

- Learning New Physics from a Machine [DOI]
- Anomaly Detection for Resonant New Physics with Machine Learning [DOI]
- Extending the search for new resonances with machine learning [DOI]
- Learning Multivariate New Physics [DOI]
- Searching for New Physics with Deep Autoencoders [DOI]
- QCD or What? [DOI]
- A robust anomaly finder based on autoencoder
- Variational Autoencoders for New Physics Mining at the Large Hadron Collider [DOI]
- Adversarially-trained autoencoders for robust unsupervised new physics searches [DOI]
- Novelty Detection Meets Collider Physics [DOI]
- Guiding New Physics Searches with Unsupervised Learning [DOI]
- Does SUSY have friends? A new approach for LHC event analysis [DOI]
- Nonparametric semisupervised classification for signal detection in high energy physics
- Uncovering latent jet substructure [DOI]
- Simulation Assisted Likelihood-free Anomaly Detection [DOI]
- Anomaly Detection with Density Estimation [DOI]
- A generic anti-QCD jet tagger [DOI]
- Transferability of Deep Learning Models in Searches for New Physics at Colliders [DOI]
- Use of a Generalized Energy Mover's Distance in the Search for Rare Phenomena at Colliders [DOI]
- Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark [DOI]
- Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector [DOI]
- Learning the latent structure of collider events [DOI]
- Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders [DOI]
- Tag N' Train: A Technique to Train Improved Classifiers on Unlabeled Data [DOI]
- Variational Autoencoders for Anomalous Jet Tagging
- Anomaly Awareness
- Unsupervised Outlier Detection in Heavy-Ion Collisions
- Decoding Dark Matter Substructure without Supervision
- Mass Unspecific Supervised Tagging (MUST) for boosted jets [DOI]
- Simulation-Assisted Decorrelation for Resonant Anomaly Detection
- Anomaly Detection With Conditional Variational Autoencoders
- Unsupervised clustering for collider physics
- Combining outlier analysis algorithms to identify new physics at the LHC
- Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge
- Uncovering hidden patterns in collider events with Bayesian probabilistic models

+ ...

100 articles et pre-publications

- Autoencodeurs de toutes sortes
- Estimation de densité
- Modèles génératifs
- Transformers et attention
- Approches Bayesiennes
- Classificateurs multi-classe
- Réseaux récurrents, etc
- Détection en temps réel
- ...

Preuves de **concept** surtout

Quelque applications aux **données** LHC

Détection d’anomalies au LPC

4 thèses sur la détection d’anomalies au LHC (1 en cours)

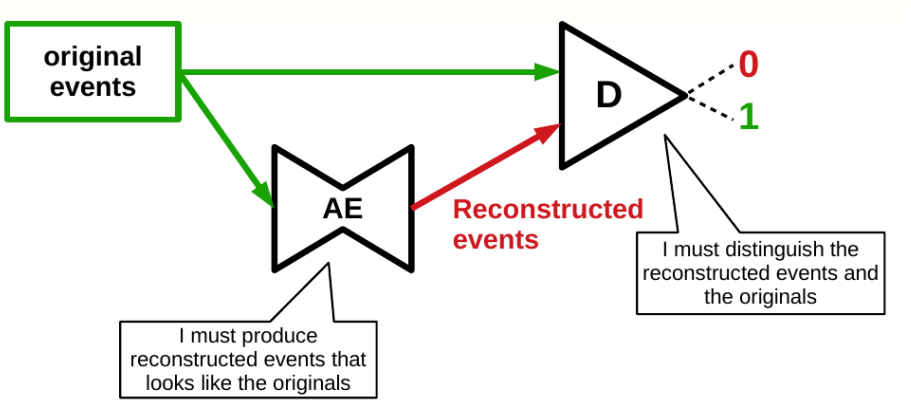
- 2016-2019 : **Fabricio Jimenez -** Processus gaussiens
- 2019-2022 : **Louis Vaslin** (co-direction V. Barra) - GAN-AE
- 2019-2022 : **Ioan Dinu** (co-tutelle C. Alexa) - Autoencodeur probabiliste
- 2023-2026 : **Eva Mayer** (co-direction S. Calvet et M. Michel) - DDP

+ ...

Détection d’anomalies : méthode GAN-AE

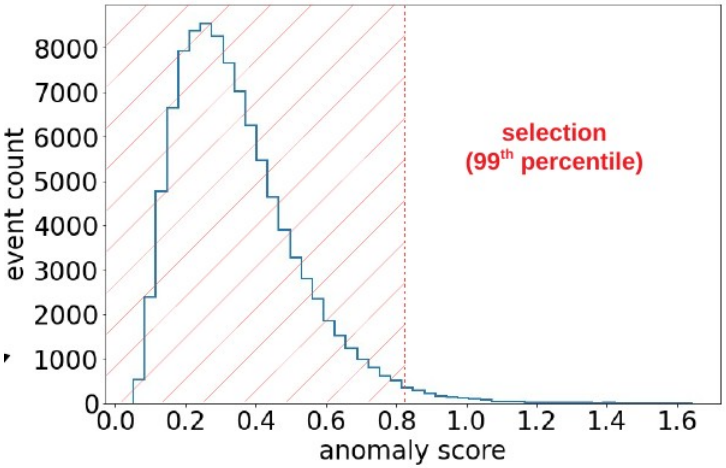
Thèse **Louis Vaslin** (LPC) 2019-2022, co-encadrée par Vincent Barra et J.D.

Détection d'anomalies : méthode GAN-AE



Apprentissage non-supervisé: **GAN + Auto-encodeur** (à la place du générateur).

Calcul d'un **score d'anomalies** à partir des données :

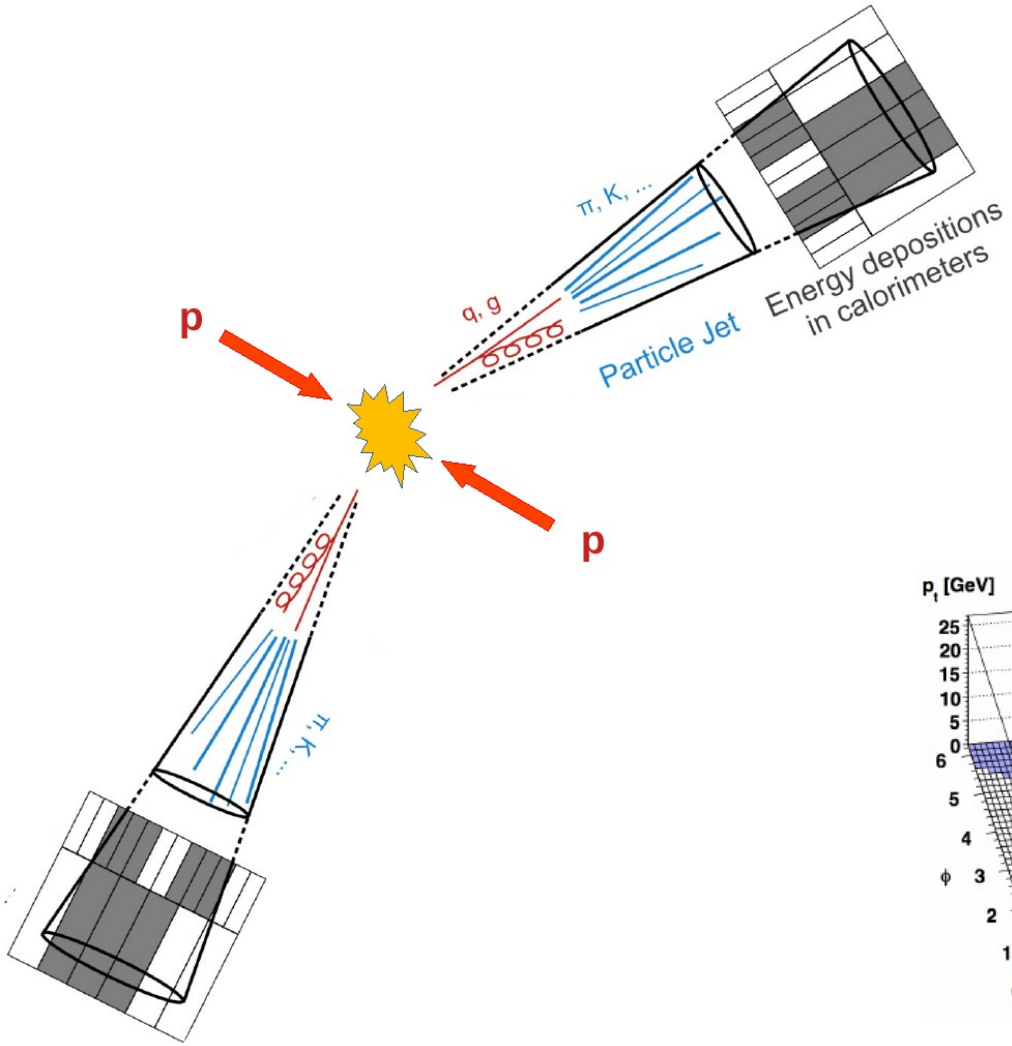


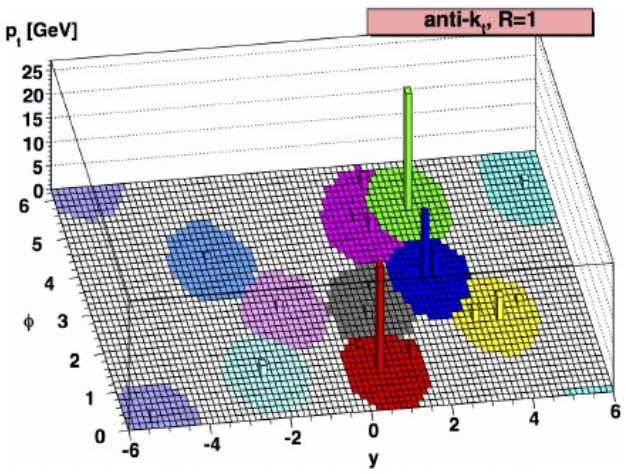
Sélection de données "anormales" potentiellement riches en signal

Données : challenge LHC Olympics

Recherche d'un **signal inconnu** dans les données

Black-boxes : bruit de fond + signal rare

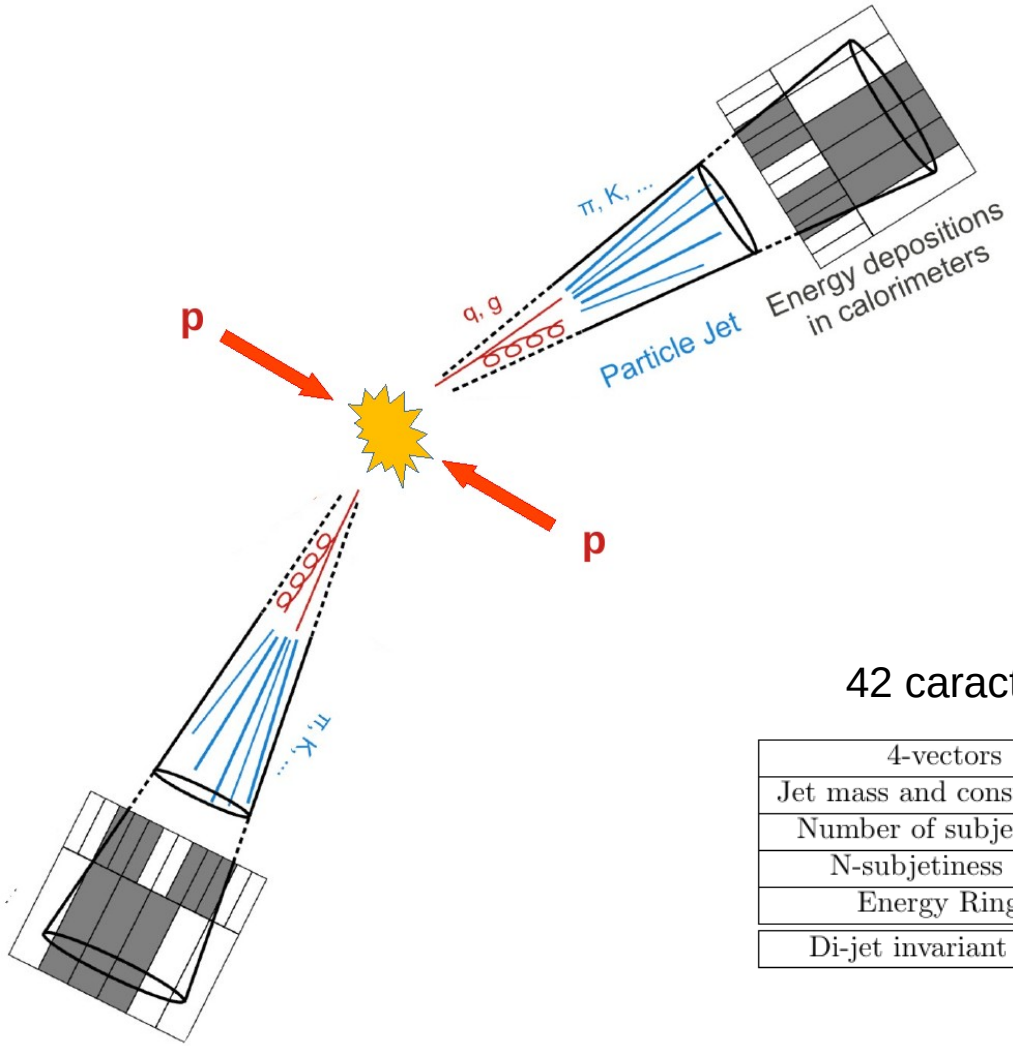




Données : challenge LHC Olympics

Recherche d'un **signal inconnu** dans les données

Black-boxes : bruit de fond + signal rare



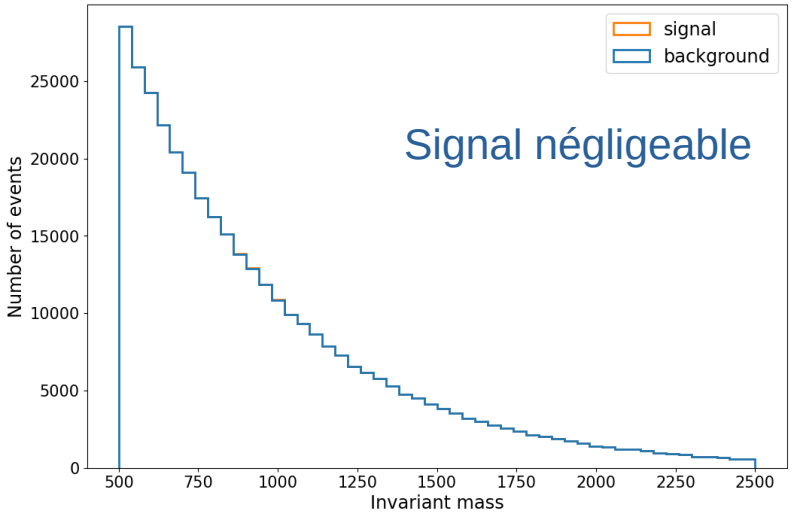
42 caractéristiques en entrée

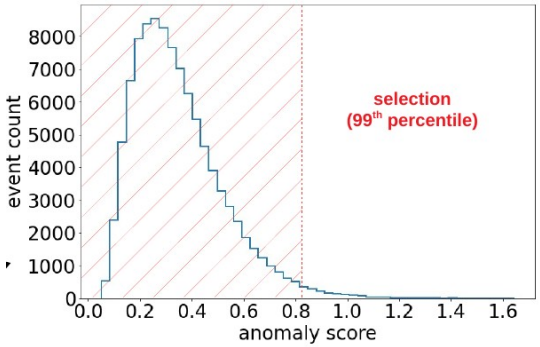
4-vectors	p_T, η, ϕ, E
Jet mass and constituents	m_{jet}, n_c
Number of subjets [20]	N_{incl}, N_{excl}
N-subjetiness [21]	τ_1, τ_2, τ_3, τ_{21}, τ_{31}
Energy Rings	$E_{ring,1}$, $E_{ring,2}$, ... , $E_{ring,10}$
Di-jet invariant mass	mjj

L'algorithme GAN-AE

L'algorithme doit à la fois

- **Isoler** une région des données potentiellement **riche en signal**
- **Prédire** la **forme** du bruit de fond



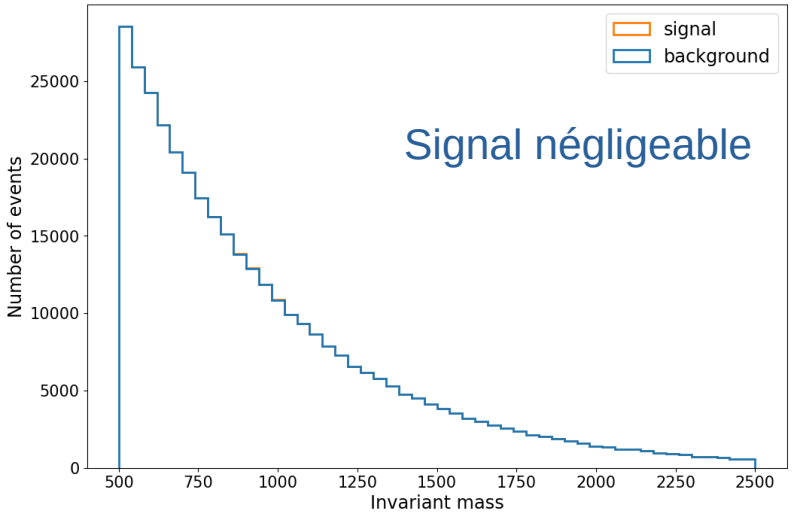


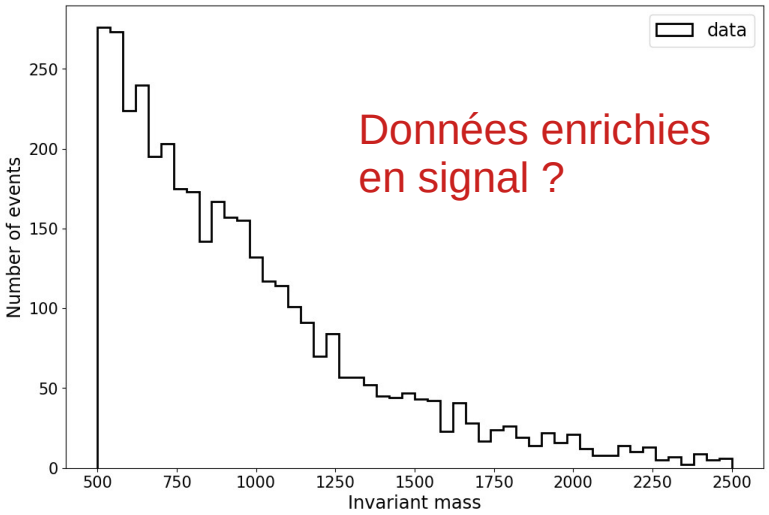
Sélection score anomalie

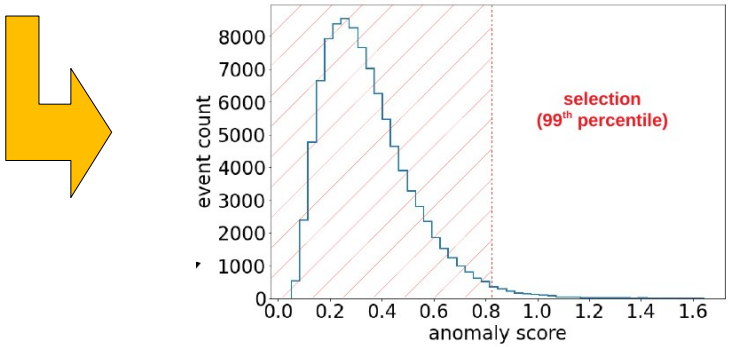
L'algorithme GAN-AE

L'algorithme doit à la fois

- **Isoler** une région des données potentiellement **riche en signal**
- **Prédire** la **forme** du bruit de fond





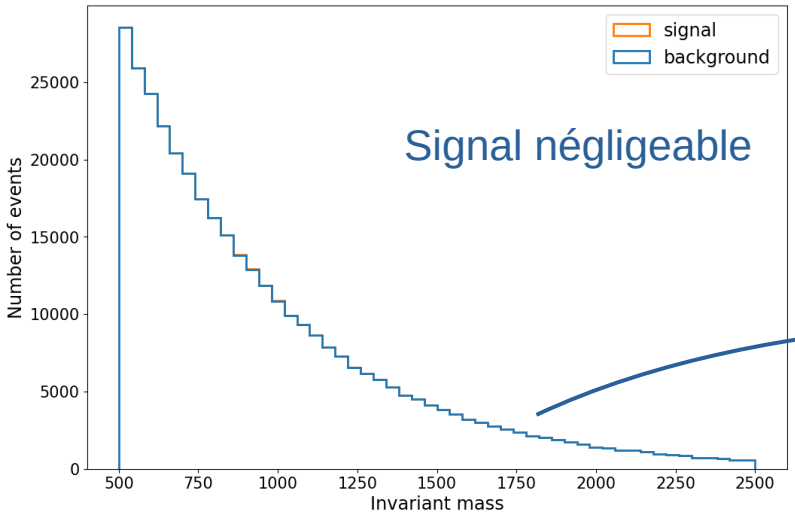


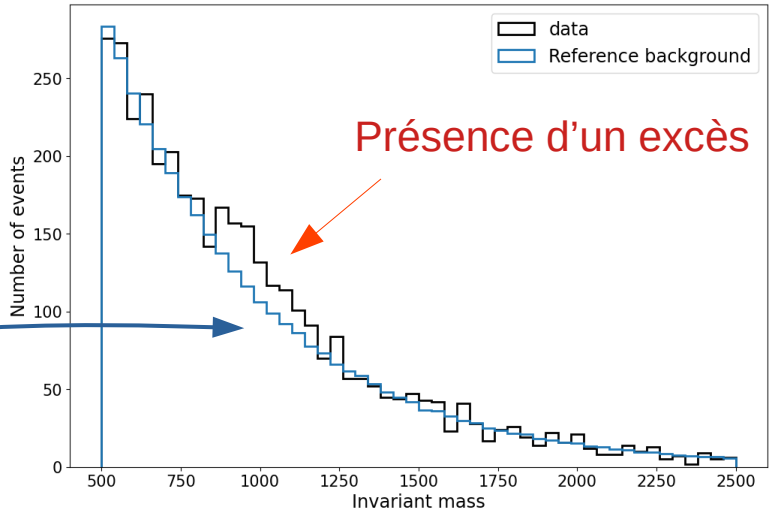
Sélection score anomalie

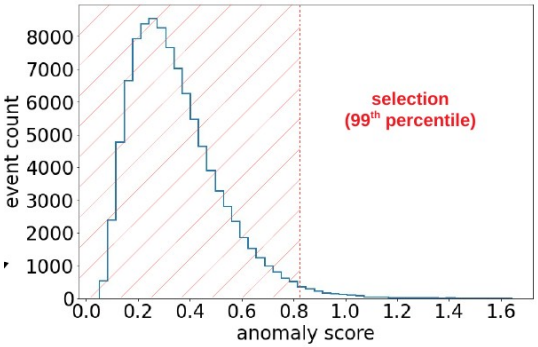
L'algorithme GAN-AE

L'algorithme doit à la fois

- **Isoler** une région des données potentiellement **riche en signal**
- **Prédire** la **forme** du bruit de fond



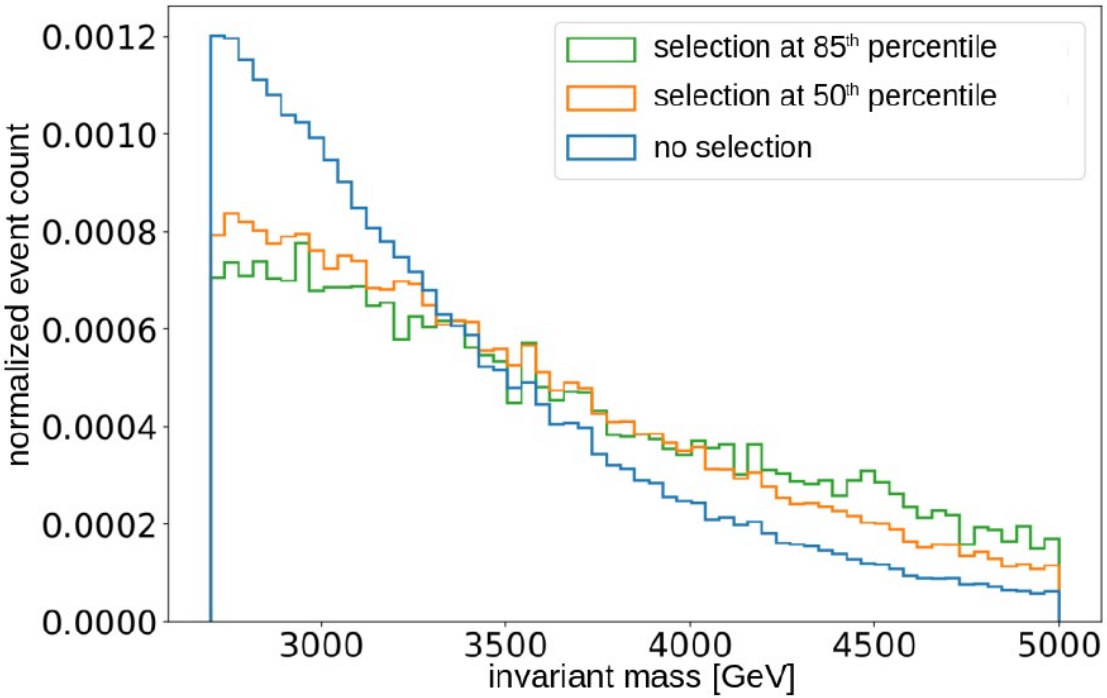




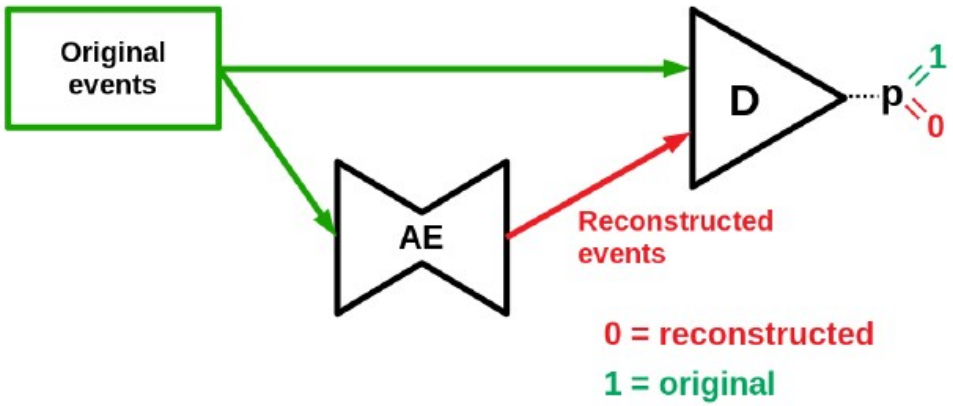
Sélection score anomalie

L'algorithme GAN-AE

Problème : une sélection sur le score d'anomalies déforme la prédiction



Entraînement du GAN-AE



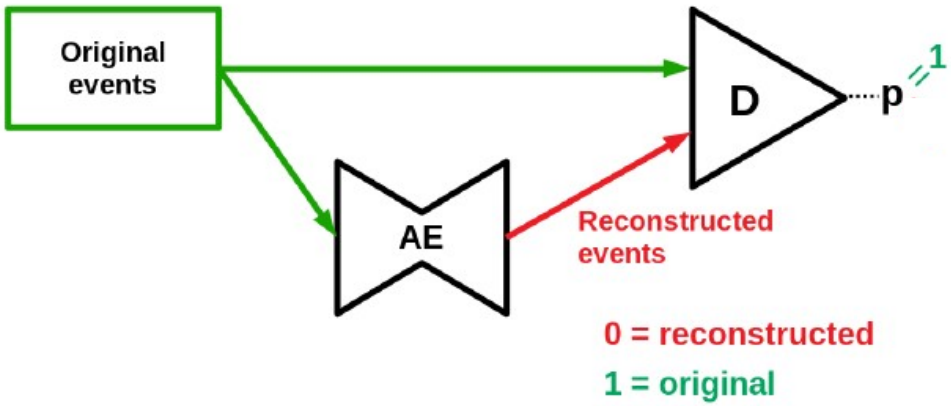
Discriminant

- Classification des échantillons originaux et reconstruits
- Loss : binary cross-entropy

$$\mathrm{bc}\left(y^{(d)}, y^{(l)}\right) = -\left[y^{(l)} \log\left(y^{(d)}\right) + \left(1 - y^{(l)}\right) \log\left(1 - y^{(d)}\right)\right],$$

Output — Label (0 ou 1)

Entraînement du GAN-AE



Auto-encodeur

- Reconstruction des événements pour tromper D
- Loss : erreur de reconstruction + binary cross-entropy (contrainte)

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

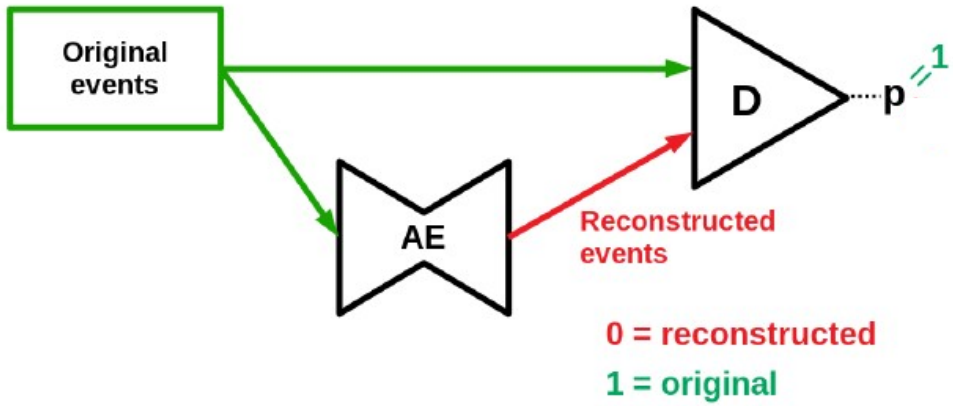
binary cross-entropy (with *switched* label)

hyperparameter

input size

Problème: déformation de la prédiction du bruit de fond

Entraînement du GAN-AE



Auto-encodeur

- Pondération des événements
- Décorrélation du score anomalie et de la masse (DisCo)

$$\text{loss} = \sum_{j=1}^{N_b} \boxed{w_j} \left(\underbrace{bc_{(y=1),j}}_{\text{D information}} + \varepsilon \underbrace{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\text{output}_{j,i} - \text{input}_{j,i})^2}}_{\text{reco error}} \right) + \alpha \times \boxed{\text{DisCo}(X, Y)}$$

number of events in batch

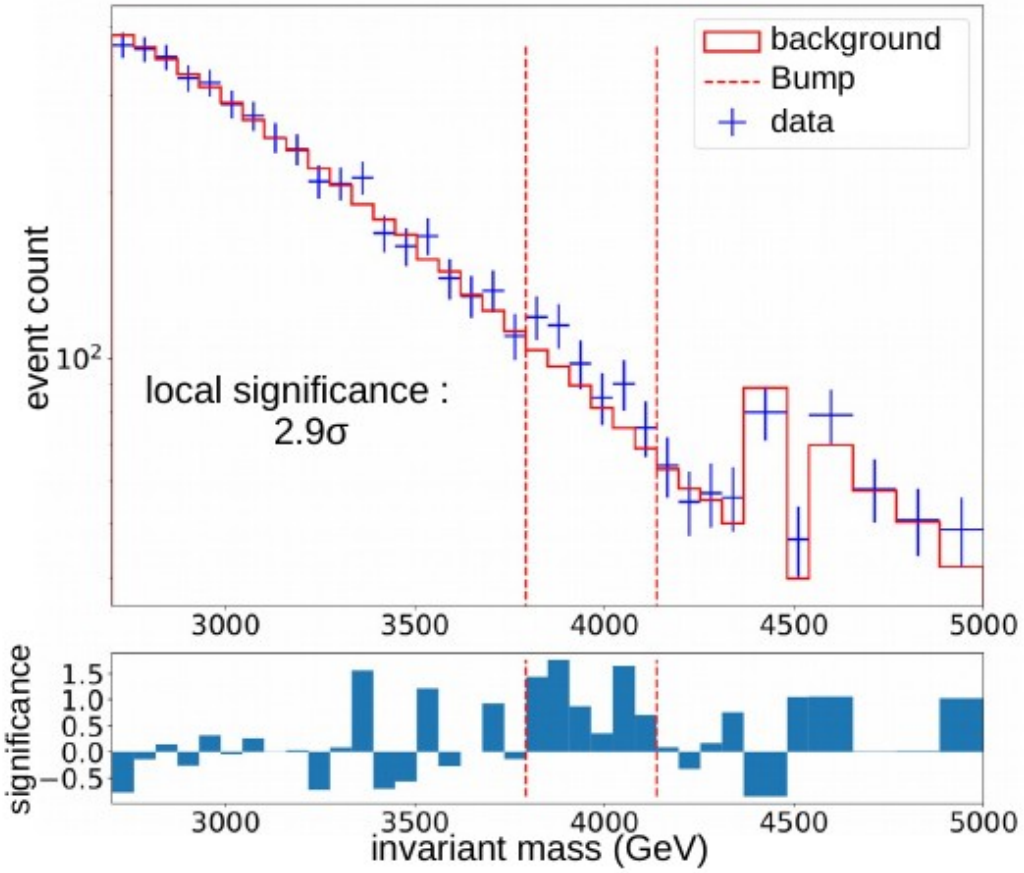
new hyperparameter

event weight

DisCo regularization

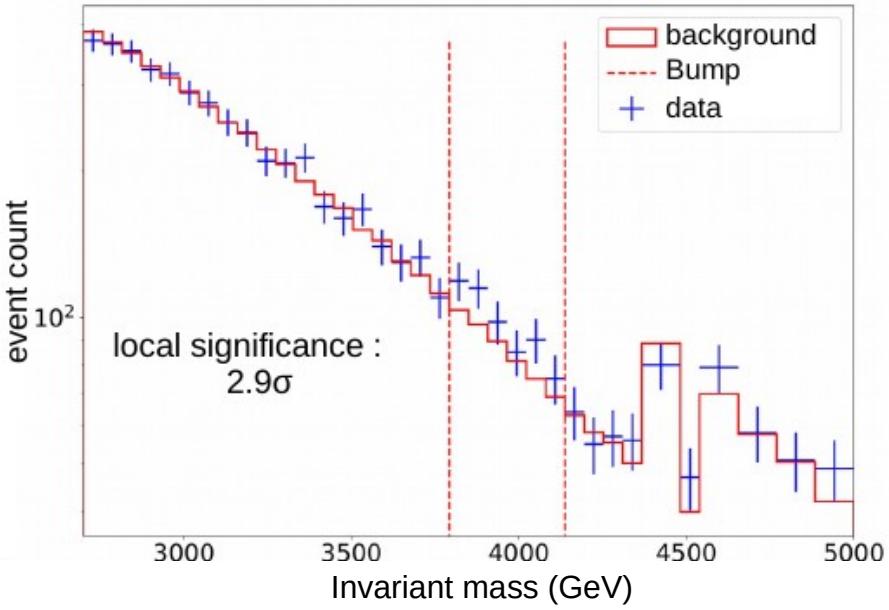
Détection d'anomalies en physique des particules

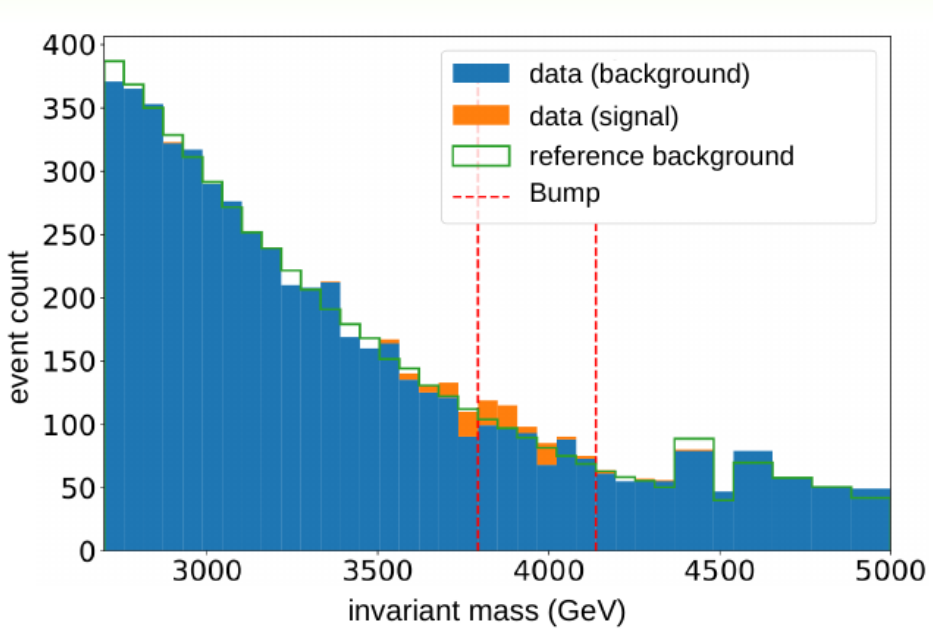
- GAN-AE : sélection d'un sous-ensemble des données
- Recherche d'un « bump » : algorithme pyBumpHunter



Détection d'anomalies en physique des particules

- GAN-AE : sélection d'un sous-ensemble des données
- Recherche d'un « bump » : algorithme pyBumpHunter





L'excès observé correspond bien au signal inconnu injecté
(Rapport S/B x 20 après sélection sur le score d'anomalie)

Publication en cours de revue par **European Physics Journal C**
3 Pre-prints: arXiv:2208.14760, arXiv:2211.07446, arXiv:2305.15179

Data Directed Paradigm

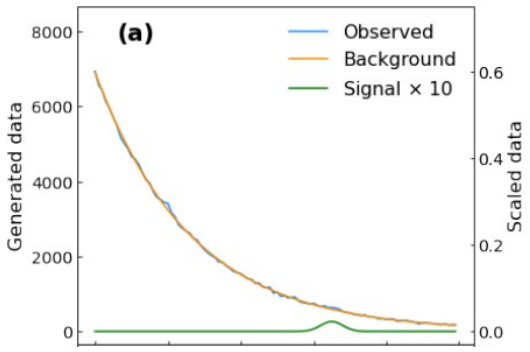
LPC (S. Calvet, J. Donini) - Tel Aviv – Montreal
+ Eva Mayer (co-supervision S.C, J.D et Manon Michel)

Recherche multi-signatures

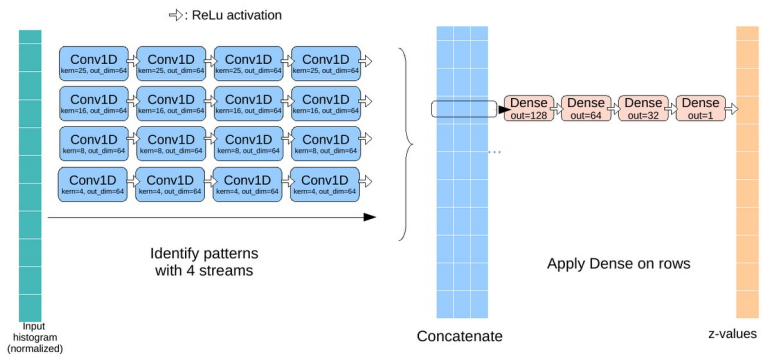
Recherche de “**bumps**” statistiquement significatifs dans les données mesurées

Entraînement :

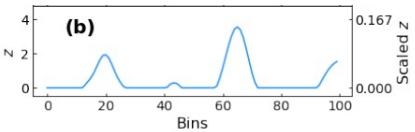
Distribution



Réseau de neurone (supervisé)



Sortie: excès du signal (significance)

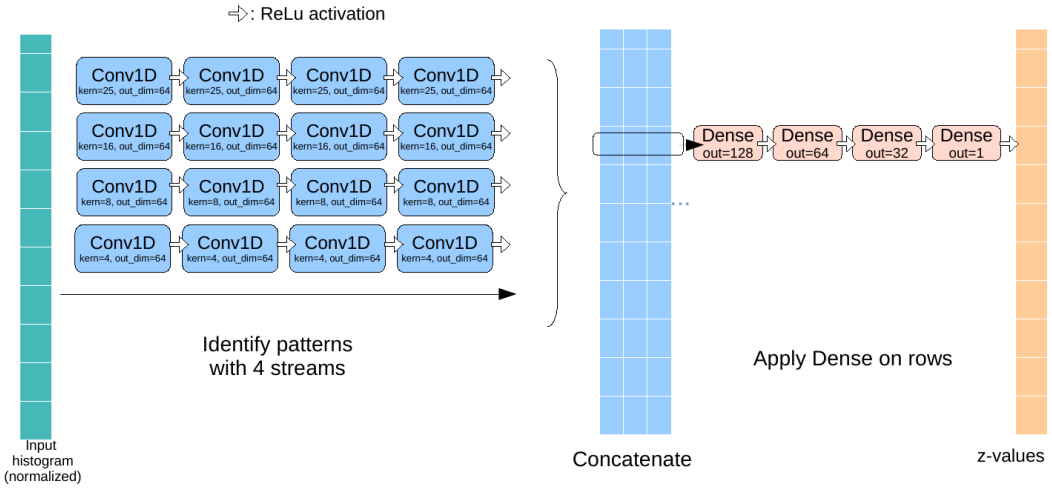


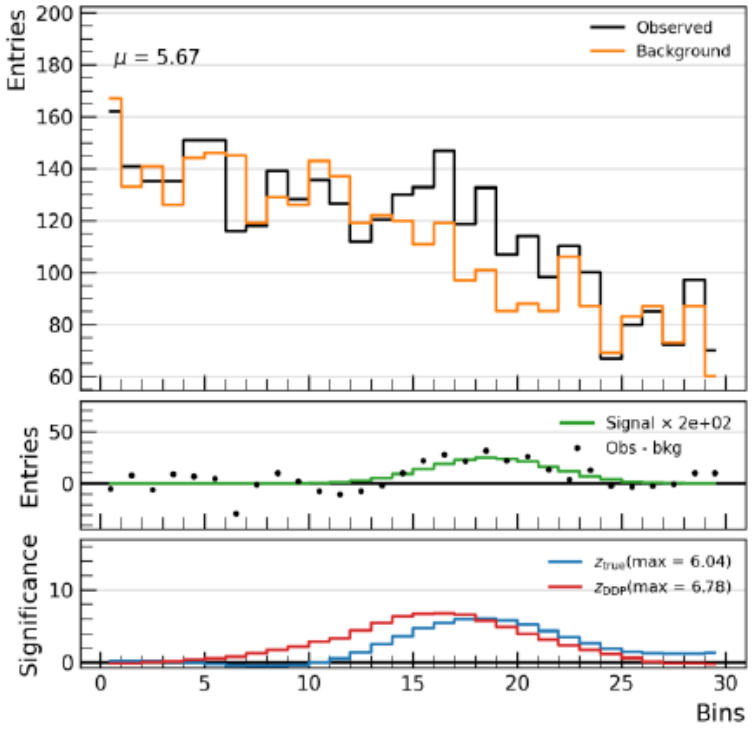
Preuve de **concept** → Préparation à l’analyse des données du détecteur ATLAS

Recherche multi-signatures

Recherche de “**bumps**” statistiquement significatifs dans les données mesurées

Test :

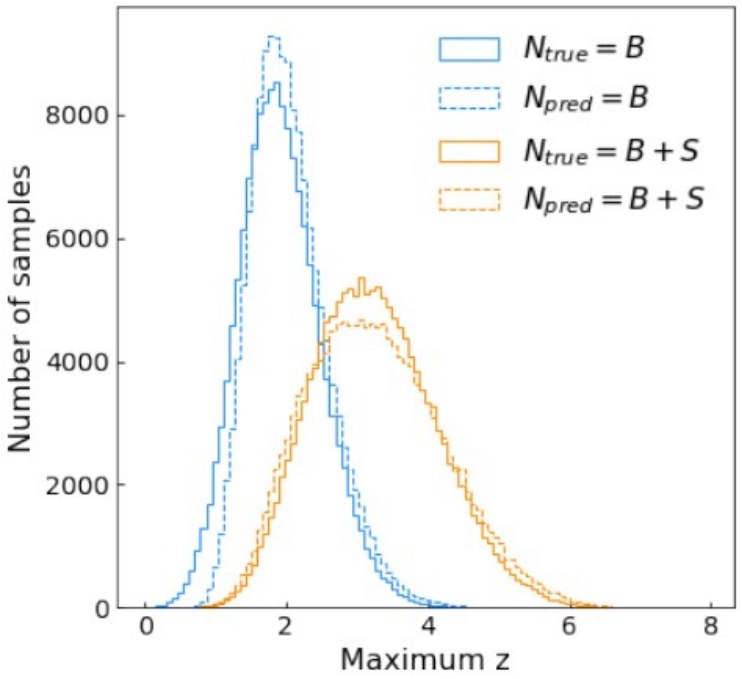


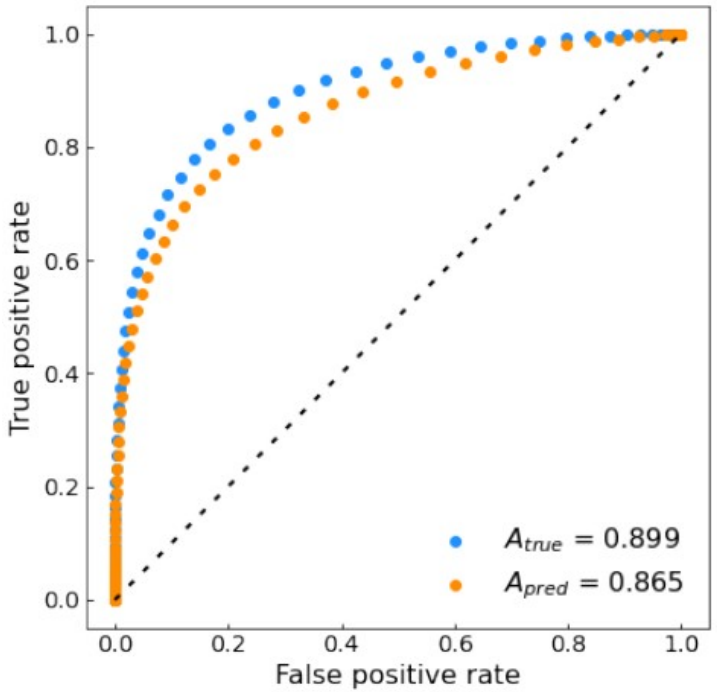


Preuve de **concept** → Préparation à l’analyse des données du détecteur ATLAS

Recherche multi-signatures

Preuve de **concept** (arXiv:2107.11573)





Préparation à l'analyse des données du détecteur **ATLAS**

But : scan rapide de **milliers** de distributions pour rechercher un **signal**

Workshop Anomaly Detection : 4-7 mars 2024

Site web: https://indico.in2p3.fr/event/30272/

6 novembre : ouverture inscription + soumission abstract

Local organizing committee:

Samuel Calvet (LPC)
Alexandre Claude (LPC)
Julien Donini (LPC)
Cyril Galpier (LPC)
Marine Hebert (LPC)
Emille Ishida (LPC)
Maria Pruzhinskaya (LPC)

Scientific organizing committee:

Emille Ishida (LPC, chair)
Vincent Barra (LIMOS)
Anja Butter (LPNHE)
Julien Donini (LPC)
Tommaso Dorigo (INFN)
Adnan Ghribi (GANIL)
Francois Lanusse (CEA)
Carole Lartizien (CREATIS)
Louis Lyons (Oxford)
Paula Sanchez (ESO)
Pietro Vischia (UniOvi and ICTEA)

Perspectives

La détection d'anomalies est un **sujet très actif** au LHC

- Recherche signaux, détection malfonction détecteurs, systèmes de déclenchement, etc …
- Foisonnement de méthodes développées

Localement : LPC + collaborations LIMOS et LMBP

Sujet transverse **fédérateur** (AAP IA Cluster, Prime...)

Backup material

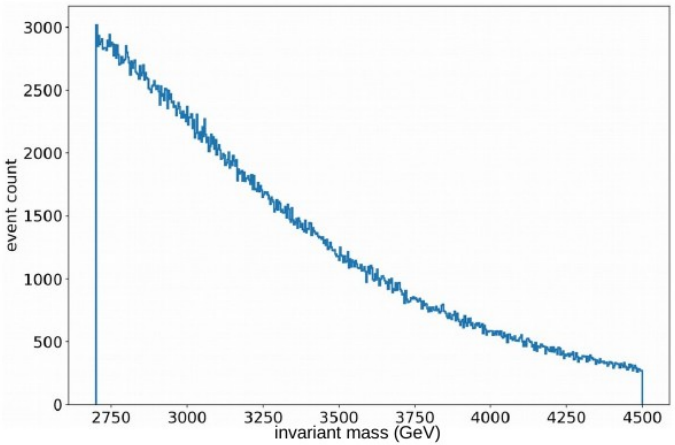
Repondération des événements

Slide Louis Vaslin

Distribution is **not uniform**

low mass => more common

high mass => more rare

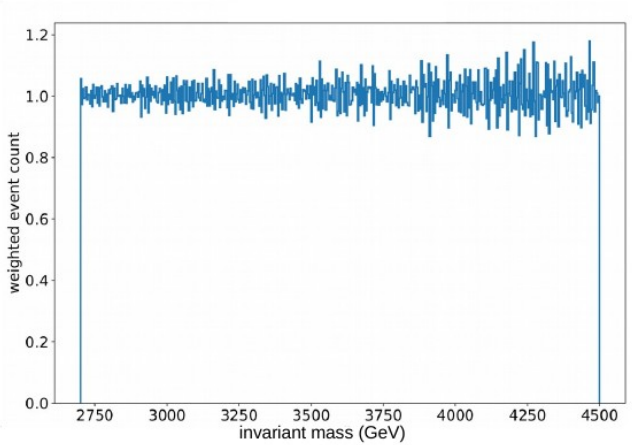


Event reweighing

compute **weights** based on invariant mass

=> *Make the distribution uniform*

Used when computing AE loss

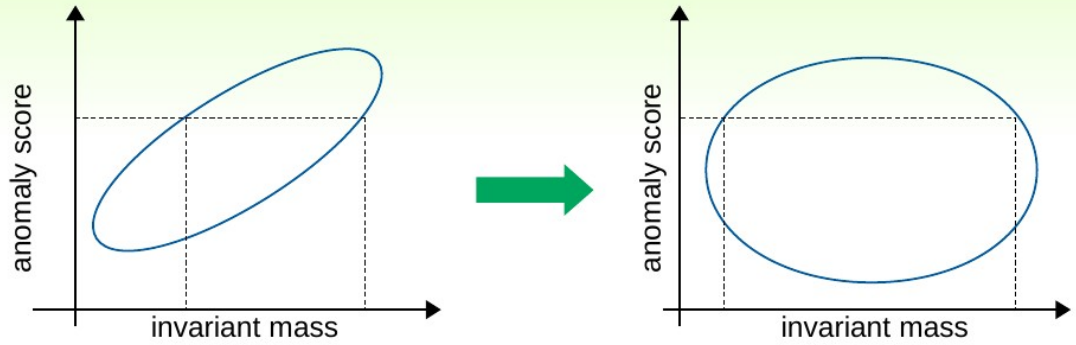


Décorrélation score anomalie vs masse

Slide Louis Vaslin

Enforcing decorrelation

Decorelate **anomaly score** and **invariant mass**



The **Distance Correlation** term (DisCo)

Distance covariance :

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

invariant mass **anomaly score**

Sampling of X and Y

=> Need a **batch** of events

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

Optimization des hyperparamètres

Hyperparameter	Value
Latent space dimension	14
AE hidden dimension	84
D hidden dimensions	{300, 200, 100, 50}
ε (reconstruction term)	6.0
α (DisCo term)	65.0
Dropout rate	20%
Number of training cycles	100
AE epochs per cycle	5
D epochs per cycle	7
Pre-training of AE	True

pyBumpHunter

BumpHunter

- **Model agnostic bump hunting algorithm**
- Evaluate both **local** and **global** p-value

Public implementation

- **Several extensions** of the base algorithm
- **Integrated** in Scikit-HEP, available on GitHub
- **Documented** in 2208.14760, 2211.07446
- Publié dans SciPost Physics