Machine Learning in Particle Physics

Benedetta Camaiani

Artificial Intelligence for Physical Sciences
January 29, 2025

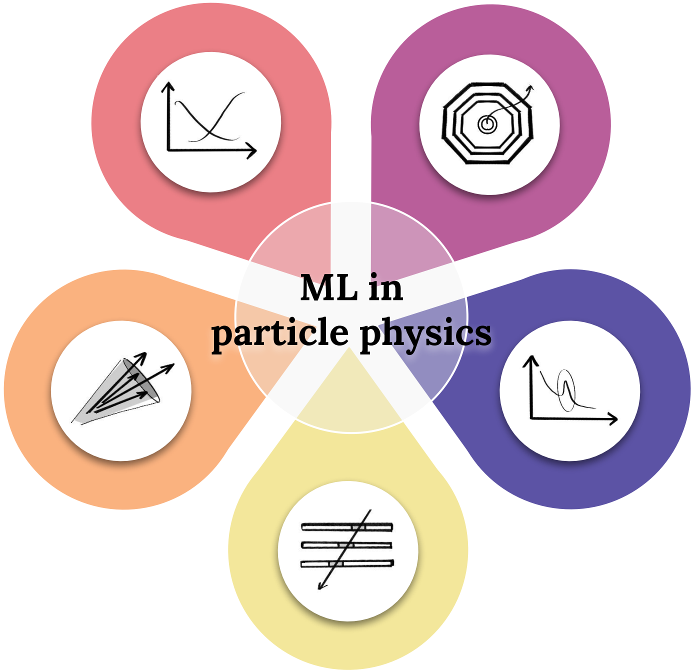
Introduction

- Over the last few decades, **machine learning** (ML) has become a powerful tool in particle physics, addressing challenges in handling high-dimensional and complex data
    - Early 2000s: boosted decision trees for event classification
    - Nearly a decade ago: deep learning for simulation and analysis workflows

- ML is nowadays used throughout the **entire workflow of particle physics experiments**, ranging from the detector level to the final analysis

In this talk we explore possible **applications of ML in the field of particle physics**, **highlighting the contributions from research groups of the Florentine area**

ML in particle physics



Event classification

Classification algorithm to separate rare signal events from backgrounds

Objects reconstruction

ML for the identification and reconstruction of physics objects within the detector

Event simulation

ML algorithm to accelerate the event simulation and generation

Anomaly detection

Unsupervised learning to search for unexpected or BSM phenomena

Detector applications

ML methods to improve detector modeling and response understanding

Event classification

- The most common ML task in particle physics is probably classification
- A key challenge is to distinguish rare signal events from overwhelming background processes
- ML is particularly suitable for this purpose as it captures correlations in high-dimensional data that are difficult to model analytically
    - Employ neural networks (NNs) in order to maximize the signal-to-background separation

Model agnostic classification

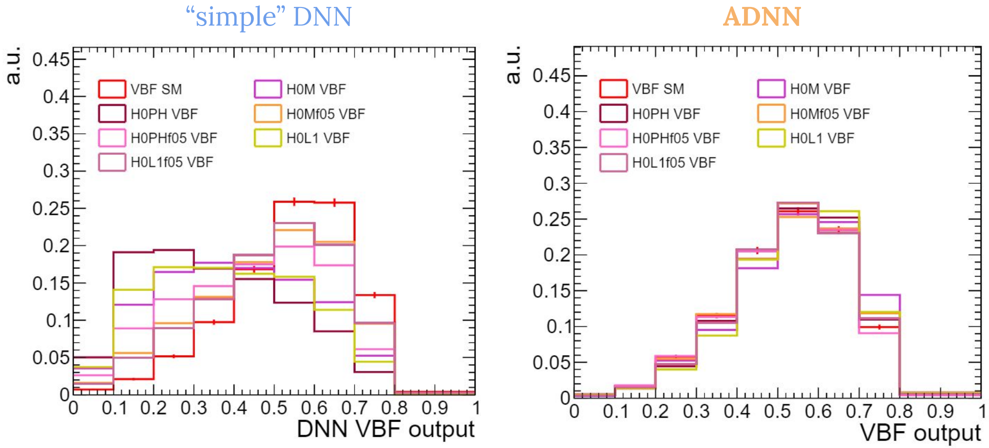
B. Camaiani, M. Lizzo, L. Viliani, P. Lenzi, V. Ciulli, L. Anderlini

Eur. Phys. J. C (2022) 82: 921

- The drawback of supervised NNs is that their output depends on the training data (MC samples)

- In the context of a Higgs boson analysis, the output depends on the physics hypothesis used to generate signal events
  - Vector boson fusion (VBF) under SM or BSM hypotheses

- Use **model-independent NN** to minimize the bias due to the theoretical modeling of the signal process
  - **Adversarial deep neural network** (ADNN) with a classifier and an adversary, *trained using multiple signal hypotheses*
  - The adversary forces the classifier to rely on domain-invariant features



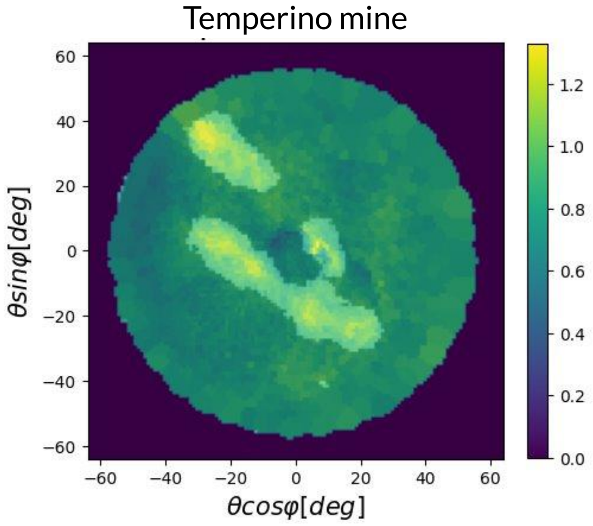
- Similar approach to define **systematic-aware NN** (S. Quinto)

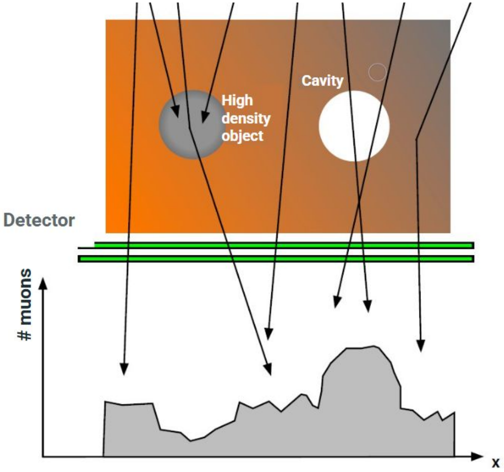
Image segmentation in muography

P. Paccagnella, C. Frosin, V. Ciulli, L. Bonechi, D. Borselli, R. D'Alessandro, R. Ciaranfi, S. Gonzi, T. Beni

- Muography uses cosmic muons to make radiography of very large targets, such as volcanoes and archaeological sites

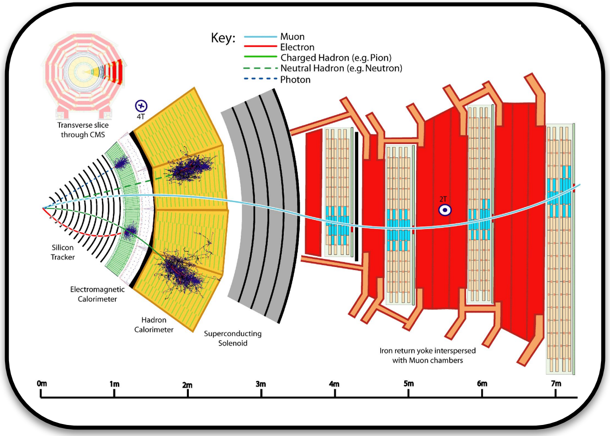
- Allows for the reconstruction of images that highlight variations in the density of the target





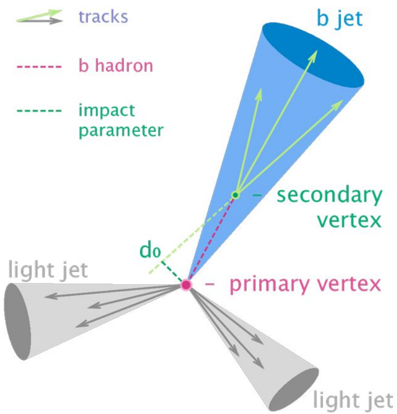
- **U-Net** employed to find cavities in the acquired images
  - CNN-based architecture with an encoder-decoder structure and skip connections that preserve spatial details

Object reconstruction



- "Objects" : any physical entity reconstructed from raw detector data (electrons, muons, jets, ...)
- Possible ML applications include **particle identification**, **energy regression**, **lepton reconstruction** from decay products, **jet reconstruction and flavour tagging** (e.g., b-tagging)

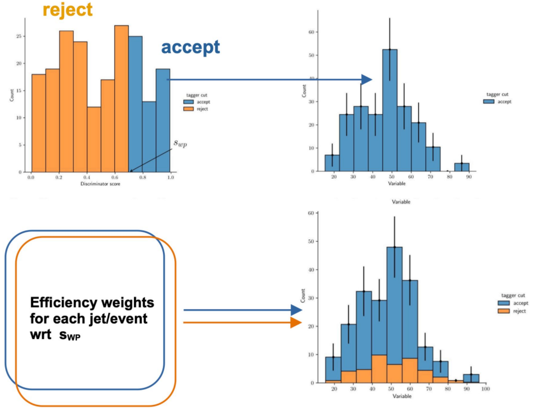
- B-tagging algorithms aim to identify jets originating from b-quarks
- They exploit primarily the relatively long lifetime of B-hadrons within the jets
  - secondary vertices
  - high impact parameters tracks
  - presence of charged leptons



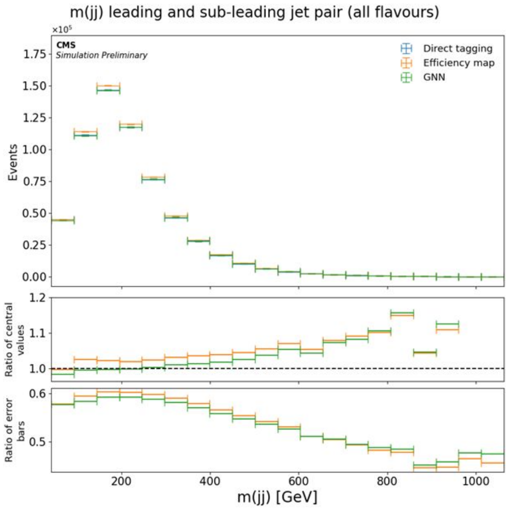
Efficiency parameterization

A. Calandri

CMS-DP-2022-051



- Several ways to estimate classifier efficiency
    - *direct tagging* ➡ limited simulation statistics due to a cut on the b-tagging discriminant to select a given phase-space enriched in b/c- jets
    - *2D efficiency maps* ➡ Limited in number of parameters and doesn't account for higher order correlations and effects from close-by jet activity



- **Multidimensional parameterization of efficiency weights using graph neural networks** (GNNs)
    - It takes full event as input and provides efficiency weights for each jet and for each working point of the classifier
    - **Higher order corrections** (p_T, η, φ, f_h) taken into account with respect to 2D efficiency maps
    - The radial separation (ΔR) between jets used as edge feature to capture **correlations among jets and environment-related effects**

Anomaly detection

- Anomaly detection algorithms are used to identify events that deviate from expected patterns, potentially revealing signs of new particles or interactions
    - Particularly suitable for **model agnostic searches**

- Beyond physics searches, anomaly detection is used to identify **detector malfunctions or calibration issues** by monitoring deviations in detector response

Data Quality Monitoring in CMS

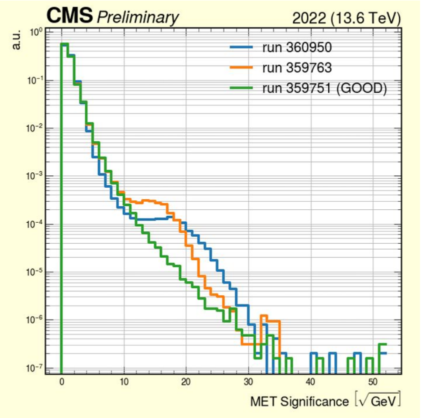
A. Papanastassiou, P. Lenzi

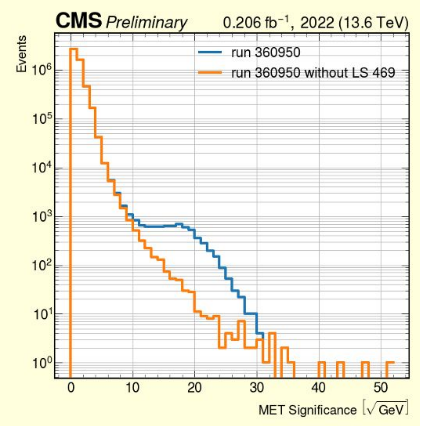
- In CMS, data is gathered in LumiSection (LS), corresponding to 23.31 s, and LSs are grouped in runs (typically a few hours long)

- During data certification, experts check several distributions of reconstructed variables to spot possible issue in each run

- An issue in some LSs would cause the entire run to be flagged as “BAD”

- Given the high number (~1000) of LSs in each run, **AutoEncoders** (AEs) are used to automate the data certification process
    - Trained on non anomalous data from “GOOD” runs
    - Tested on “BAD” runs to detect anomalous LSs
    - **Only the anomalous LSs are removed, rather than discarding the entire run**





Event simulation

- Traditional Monte Carlo simulations (e.g., using Pythia and Geant4) are computationally expensive
  - They require generating a large number of events and simulating detailed interactions and detector responses for each one

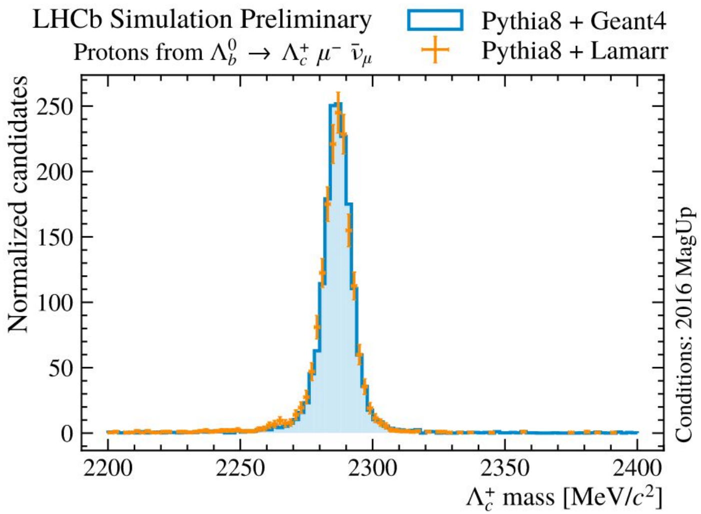
- ML-based simulations provide a way to drastically reduce the time and resources needed:
  - **Subdetectors simulation**, e.g., replicating the energy distributions in calorimeters
  - **Fast/flash simulation** of *full detector responses*, which reduces computational cost while maintaining high accuracy
  - **Event generation**, accelerating particle-level event production by learning and replicating physics process distributions

Lamarr

L. Anderlini, M. Barbetti

PoS ICHEP2022 (2022) 233, arXiv:2303.11428

- *Lamarr* is the novel flash simulation framework of LHCb
- It is conceived as a **pipeline of** (ML-based) **parameterizations** designed to replace both the simulation and reconstruction steps
- A modular architecture (∽20 different models) allows choosing the most suitable model for simulating each subdetector, ensuring high *flexibility*
- Two kinds of parameterization:
  - Acceptance/Reconstruction/Selection efficiencies → **DNNs**
  - Reconstructed features (e.g., smeared momenta or PID variables) → **GANs**
- A dedicated *python package* PIDGAN developed to simplify the use of GANs, offering ready-to-use implementations
- The parameterizations are transformed in C, compiled as shared objects, and linked directly to the simulation software (`scikinC` tool)



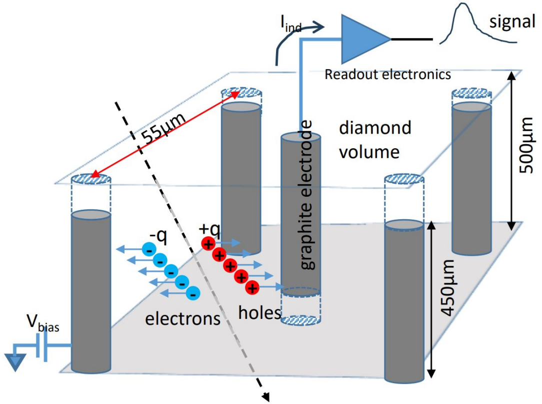
Detector applications

- ML algorithms can enhance the study and simulation of detector responses, providing faster and more accurate solutions compared to traditional methods

- Possible applications are related to:
    - **Study of the detector responses**
    - **Track and vertex reconstruction**
    - **Sensor calibration**

3D diamond detectors

A. Bombini, C. Buti, A. Anderlini, A. Rosa, R. Pietrini



- Simulation of 3D diamond pixel detectors for ionizing radiation
- Modeling detector responses traditionally involves solving complex differential equations based on the physics of particle interactions, often requiring detailed geometric modeling
- These methods rely on discretized grids (e.g., finite element methods) and are often computationally heavy
- **Physics Informed Neural Networks** (PINNs) are under investigation to compute **time-dependent potential maps**
- PINNs solve differential equations by incorporating physical laws directly into the learning process

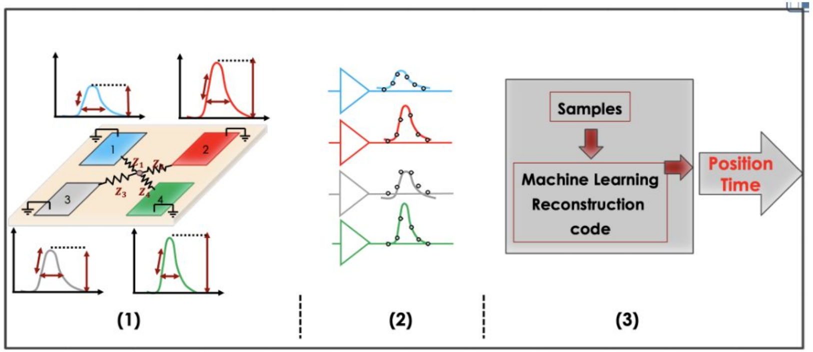
$$\mathcal{L} = \underbrace{\mathcal{L}_{PDE} + \mathcal{L}_{BC} + \mathcal{L}_{IC}}_{\textit{physics informed}} + \underbrace{\mathcal{L}_{data}}_{\textit{data informed}}$$

4D tracking

M. Lizzo, G. Bardelli, M. Bartolini, L. Viliani, G. Sguazzoni, A. Cassese

founded by PRIN 2022 - 2022KLK4LB

- DCRSD: DC-coupled resistive silicon detector
  - n-in-p sensor (LGAD)
  - 4 electrodes connected to each other via resistors (DC pads)
- Exploit charge sharing to reconstruct the position information
  - When a charged particle traverses the sensor, each pad sees a signal whose amplitude depends on its distance from the hit



- The goal of the project is to precisely determine the spatial and timing information of the charged particle
- Traditional method: center of charge → unable to provide timing information
- ML-based approach: **Recurrent Neural Network** (RNN) to retrieve the hit position and the time of arrival
  - Waveform modeled as time series

Conclusions

- We have seen a (non complete) overview of possible applications of ML in particle physics

- Many research groups in Florence actively involved in this field

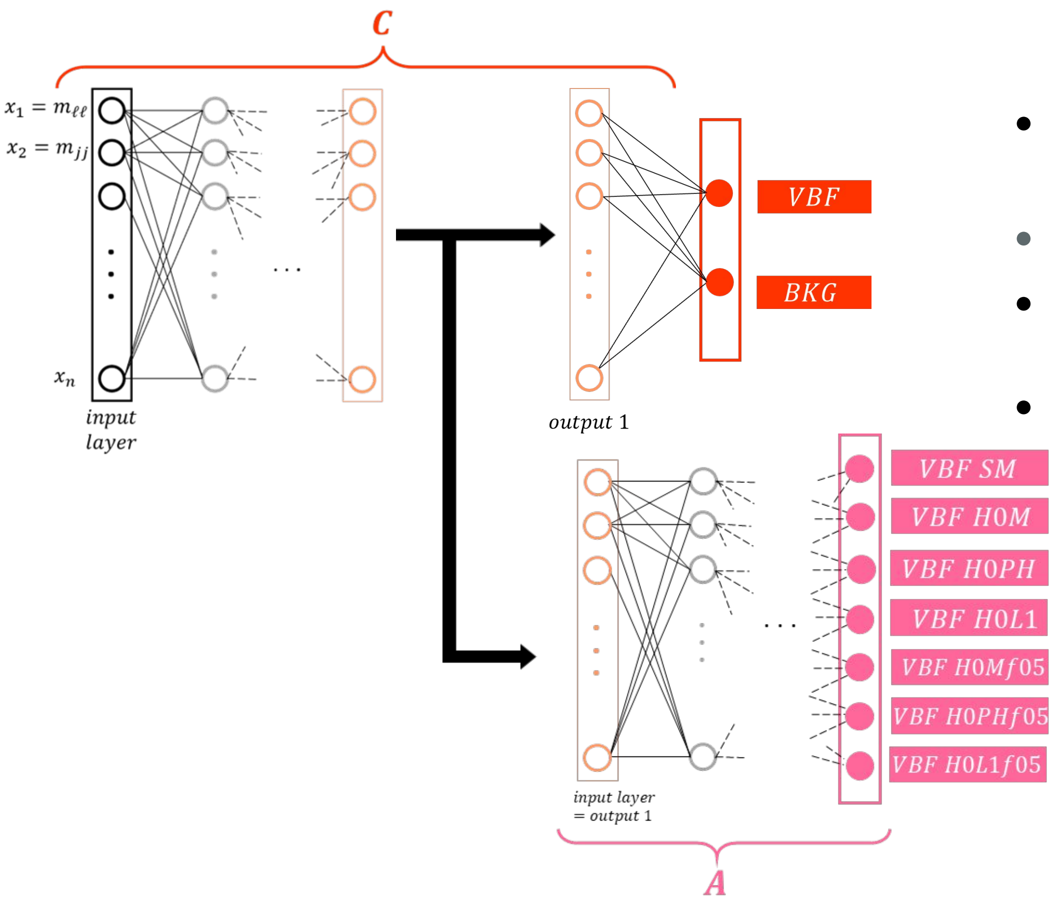
- If you are interested in any of the topics presented, do not hesitate to contact people involved!

Thanks for your attention

Backup slides

Adversarial deep neural network

Eur. Phys. J. C (2022) 82: 921



Classifier

- Takes as input the measurable kinematic variables of an event
- Trained on both **signal and background events**
- Aims to determine if the event is signal- or background-like
- Signal sample includes events coming from different *domains*, i.e. **different signal models**

Adversary

- ***Multiclass* neural network** trained **only on signal events** (SM + BSM hypotheses)
- **Tries to guess the physics model of signal events**, regressing the domain from the second-to-last layer of C

Competitive learning

- The classifier is penalized if its output contains too much information on the domain of origin of signal events

- If C manages to prevent A from identifying the signal model, then the **classification is independent of** the domains of origin of the events, i.e. **the physics model of signal events**

Compute first the gradient of $\mathcal{L}$ with respect to the C weights. A weights frozen in this step.

The parameter α regulates the interplay between A and C

Two-step training procedure

1. $\mathcal{L}oss = \mathcal{L}oss(C) - \underbrace{\alpha \cdot \mathcal{L}oss(A)}_{\text{Penalty term}}$

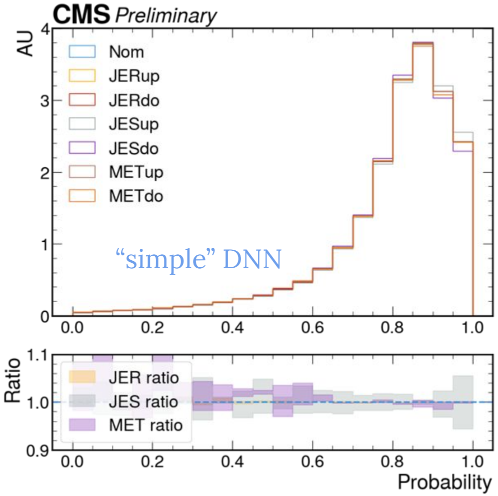
2. $\mathcal{L}oss(A)$

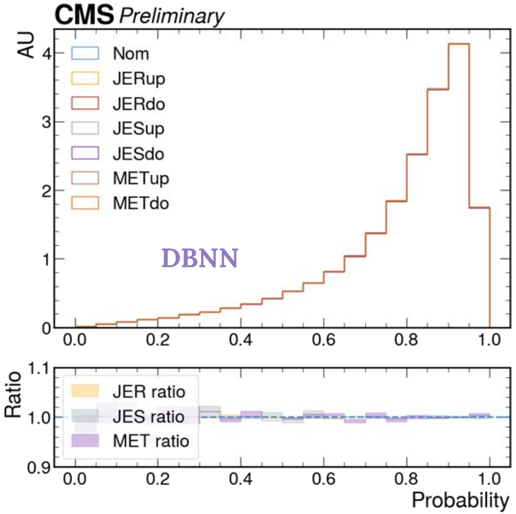
Compute the gradient of $\mathcal{L}(A)$ with respect to the A weights

Systematic-aware NN

S. Quinto, M. Lizzo, L. Viliani, P. Lenzi

- The goal of this analysis is to measure the Higgs boson production cross section in the WW boson decay channel using data collected during the LHC Run 3 → Run 2 measurement limited by systematic uncertainties
- Systematic uncertainties modify the shape of the binned observable used to extract the signal
- A NN has been implemented such that its output is sensitive as less as possible to systematic variations





- NN trained using a sample corresponding to the nominal value of the uncertainty
- Loss function includes an additional penalty term to minimize differences between nominal and varied histograms

$$\mathcal{L} = \mathcal{L}_c + \lambda\Lambda$$

Bagging technique to estimate GNN uncertainties

A. Calandri

- **<u>Bootstrap aggregation</u> performed to extract central values as well as uncertainty band of GNN efficiency prediction**
  - different trainings performed using training set samples with replacement
  - uncertainty on the histogram bin of result corresponds to spread of aggregations for each bin: **training uncertainty band around 5-10% for main kinematics features**