Integrating Explainable AI in Data Analyses of the ATLAS Experiment at CERN

University of Liverpool
Phystat Statistics meet ML 24

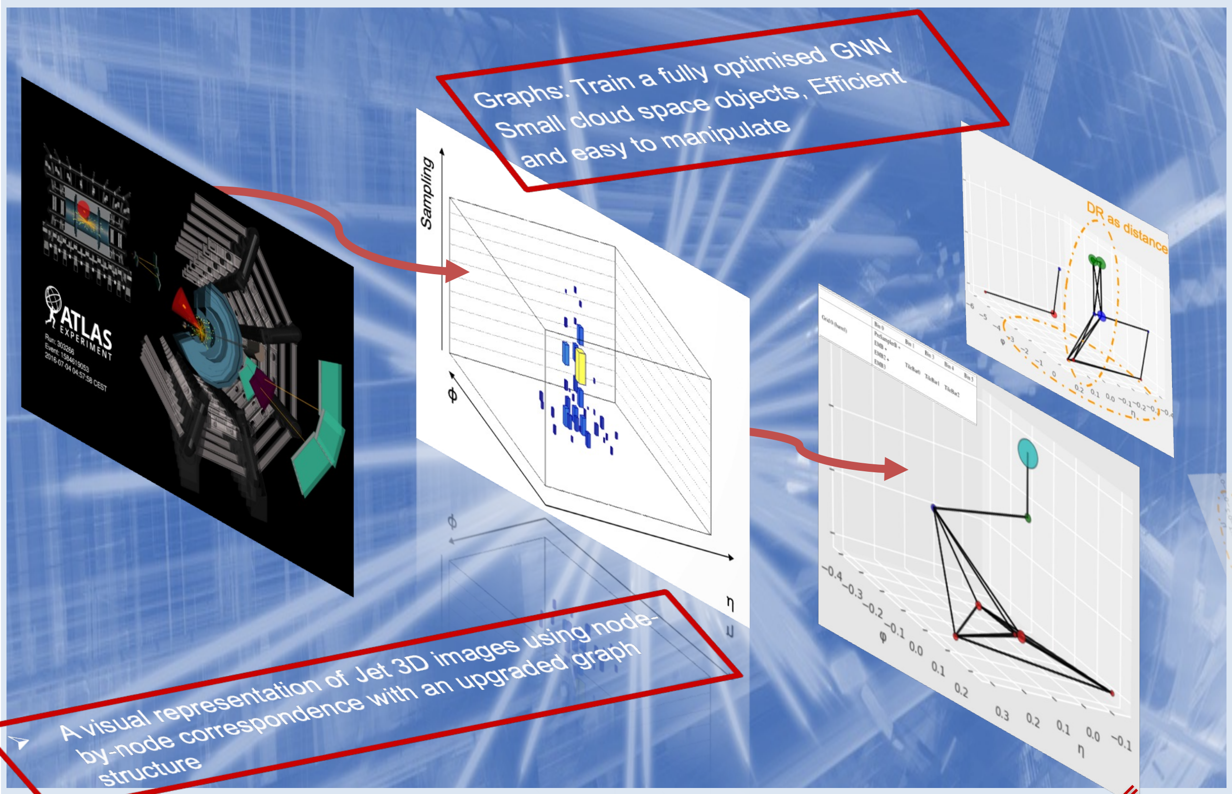
The MUCCA project: **Multi-disciplinary Use Cases for Convergent new Approaches to AI explainability**

Machine Learning has been used in HEP data analysis for over a decade but remains a "black-box" in decision-making; in this work, we apply Graph Neural Networks and XAI tools to analyse ATLAS data.

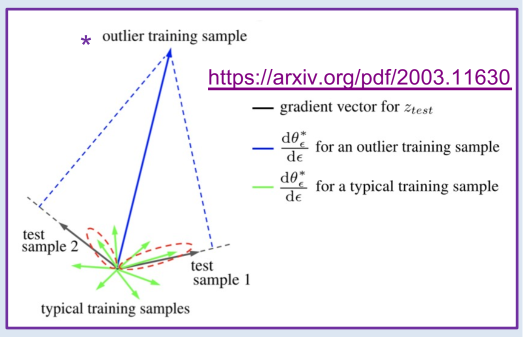
Two benchmark analyses are considered in the project, one on Supersymmetry [JHEP 12 (2023) 167] and one on searches for dark photons [JHEP 06 (2023) 153] which is considered here.

- ➢ "Dark" photons search, light long lived particle belonging to a new hidden sector:
- Signal leaves different signature in the detector than background
- 500K signal (dark photon jets) and background (QCD jets) data published with the paper
- Discriminate background processes from signal
- Mapping clusters of hadrons (jets) in 3D coordinates:
  - →3D objects/Jets (Very Low-Level)
  - →eta, phi, sampling layer and energy

Three phases:
1. Apply XAI-NPUT techniques
2. Identify shortcomings and metrics
3. Get new transparent algorithms

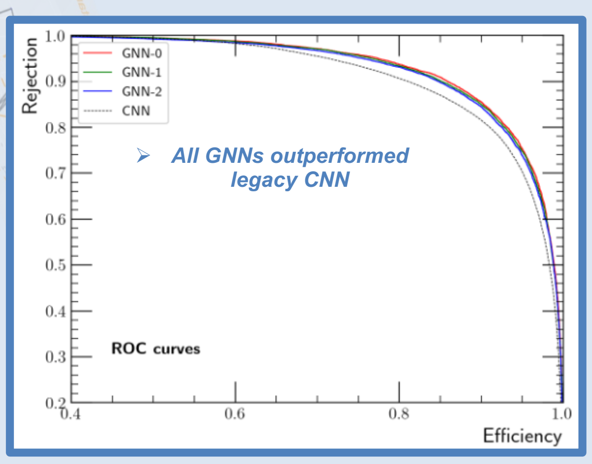


$$^{*}\ \tilde{r}(g,g') = \sum_{t\in\mathcal{C}} \left[\nabla_{w_t} l(w_t,g)\right]^{\top} \nabla_{w_t} l(w_t,g')$$



DarkJetGraphs (Git @Carmigna):
- •Node for every cluster in the calorimeter
- •Normalized cluster energy and position
- •Edge built with spatial covariant distance "**DR**"

Graph Pre-processing:
- •Remove isolated and self-connected nodes (Baseline GNN-0)
- •Retain largest subgraph as calorimeter noise (**GNN-2**)
- •Exclusive selection on cone distance condition (Best **GNN-1**)

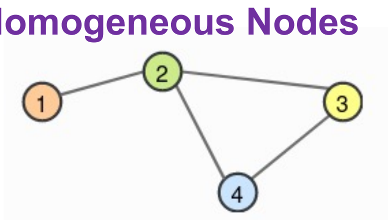


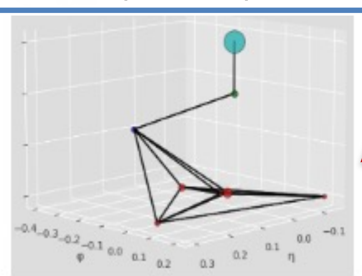
➢ **Model optimization and *Data Informed* XAI sampling:**
- Benchmarking legacy 3D-CNN
- LSOR Procedure (*Leave-Some-Out-Retrain*)

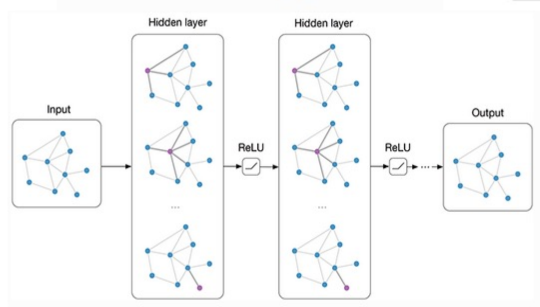
➢ **Kappa* pruning cuts over Gradient Vectors with differentiable Programming**

DarkJetGraphs sample Model

3D Graph Visualizations
Homogeneous Nodes





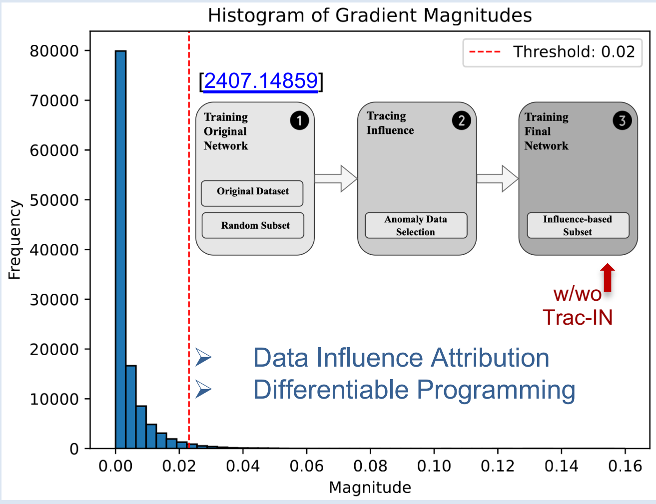


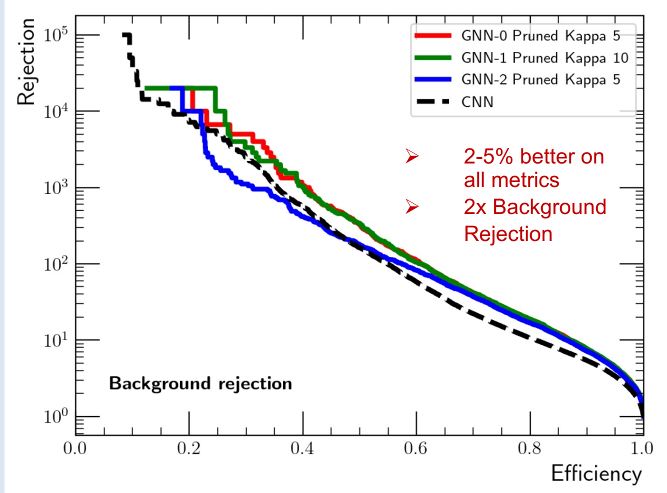
BatchNorm Layer
ARMA Conv Layers
Global Pooling Layers
FC Linear Layer
Output score

- ➢ 4 ARMA Conv Graph Spectral Blocks
- ➢ Or TGAT Transformers
- ➢ Multi-Classifier

- ➢ **Powerful representations for hetero features**
  - ➢ **MPLs using edge attribute ||distance||**

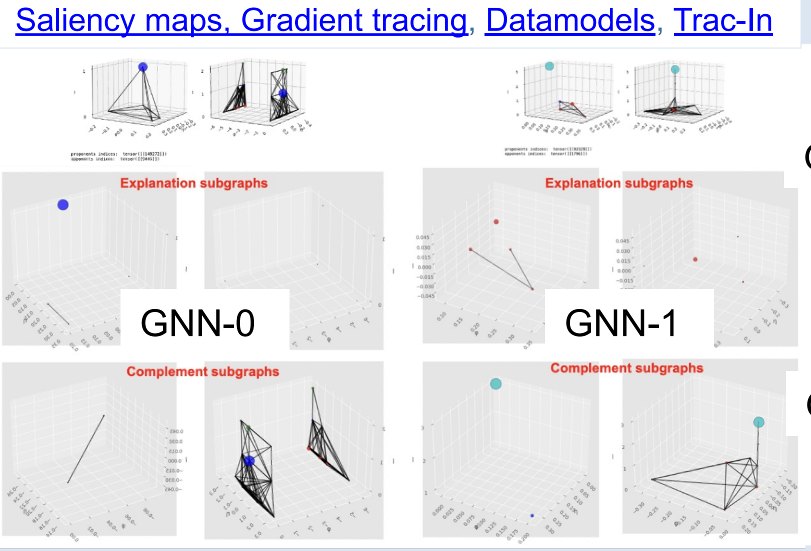
LSOR with X-tra layers:
- ❖ TRAC-IN* as data influence from sampling (TP, FP, FN & TN sets) and relevance *r(g,g')*
- ❖ GNN-Explainer or PYG Saliency Maps to **explain-the-explainer on the top-k nodes/edges.**



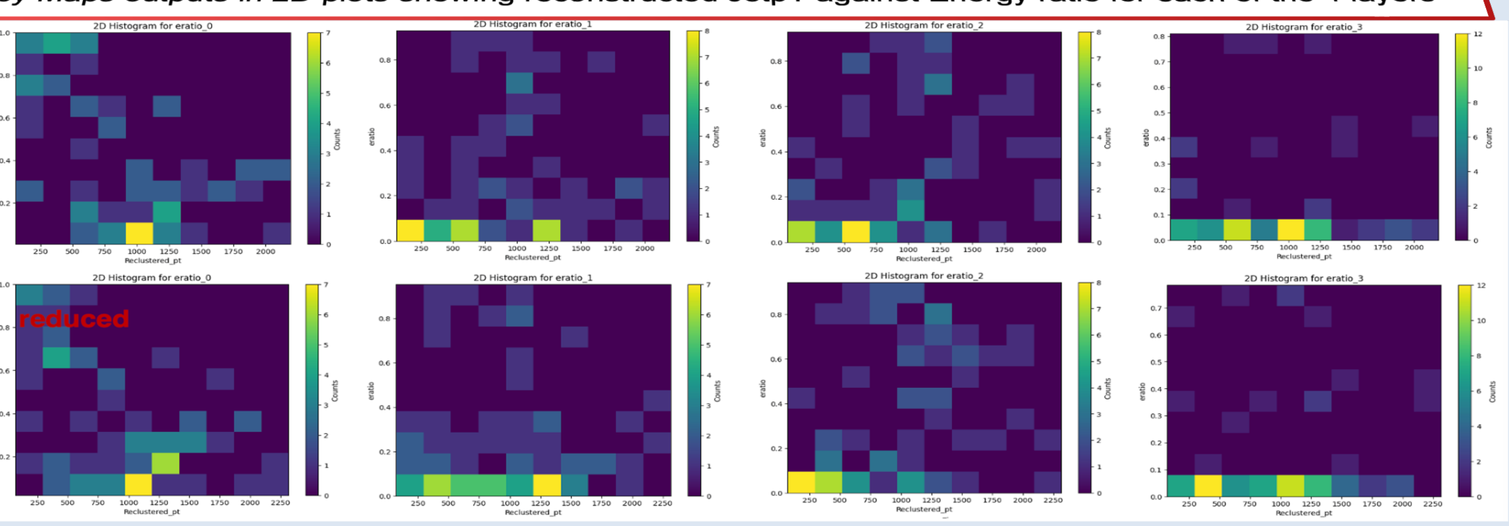


- ➢ Saliency Maps are essential to explain Captum clear but open ended explainability, i.e., proponent/opponent minimal prototyping needed.
- ➢ **2D plots below show reduced activity in layer 0 (low pT range) for FP *Proponents* as an instance.**

Saliency maps, Gradient tracing, Datamodels, Trac-In



Saliency Maps outputs in 2D plots showing reconstructed JetpT against Energy ratio for each of the 4 layers



- ➢ GNN-based DarkJetGraphs models consistently outperformed legacy CNN, showing a 2-5% improvement across all metrics and achieving a 2x better QCD jets rejection.
- ➢ Explainable AI (XAI) methods such as Saliency Maps and Trac-IN provided enhanced interpretability of model outputs, offering critical insights into model behaviours.
- ➢ Kappa pruning technique with differentiable programming to interpret the data proven valid to enhance performance further.
- ➢ Paper in progress.

Dr. Joseph Carmignani, Dr. Cristiano Sebastiani, Prof. Monica D'Onofrio (University of Liverpool)
Alessio Devoto, Dr. Simone Scardapane, Prof. Stefano Giagu « PI MUCCA » (La Sapienza)