Integrating eXplainable AI in Modern High-Energy Physics

Monica D’Onofrio, Cristiano Sebastiani, **Joseph Carmignani**

with contributions from Carl Gwilliam + Sam Valentine, Lennox Wood (MPhys at UoL)

University of Liverpool

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE
EuCAIFCon 2024

The MUCCA Project:
CHIST-ERA-19-XAI-009

Istituto Nazionale di Fisica Nucleare

with invaluable help and continuous support from our collaborators, S. Giagu (PI-MUCCA), computer scientists S. Scardapane, A. Devoto, D. Genovese – **La Sapienza**

j.carmignani@liverpool.ac.uk

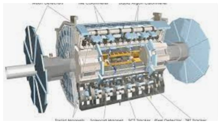
MUCCA
Multi-disciplinary Use Cases for Convergent new Approaches to AI explainability

Collaboration that brings together researchers from different fields: High Energy Physics, Medicine, Neuroscience and Computer science

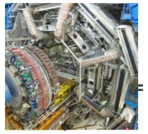
Goal to study xAI in heterogeneous cases *quantifying strengths* and solving weaknesses of new and state of the art methods on Deep Learning applications

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

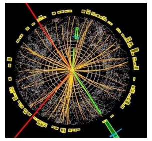
➢ A few of the tested XAI models:
Learning most important features for a given prediction -> Saliency maps
Estimating training data influence -> Gradient tracing , Datamodels , Trac-In



WP1: **HEP Physics**
Application of AI-methods to searches for New Physics at ATLAS @LHC. xAI to improve transparency and impact of systematics errors



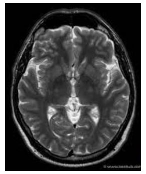
WP2: **HEP detectors**
Application of AI-methods to calorimeter detectors (PADME). xAI to improve performances and systematics comprehension



WP3: **HEP real time systems**
Develop AI-based real time selection algorithms for FPGAs at ATLAS. Use xAI methods to understand complex systems

WP7: **xAI tools**
Survey of xAI methods relevant for the use-cases, develop xAI usage pipelines: analysis of results

WP4: **Medical Imaging**
Develop xAI pipeline for segmentation of brain tumours in magnetic resonance imaging. Use publicly available databases for xAI developments, focusing on explainability of training strategy



WP6: **Neuroscience**
Test xAI techniques to uncover computational brain strategies and selection of dynamical neural models

WP5: **Functional imaging**
Test xAI methodology in respiratory systems. Analyse complex systems (passage of air and mucus) to derive model and test xAI

WP7 Publications: Hassija, V. *et al.* Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn Comput* **16**, 45–74 (2024)

XAI for high energy physics: outline

eXplainability (XAI) as *bridge* between the AI expert and scientists:

- How to select a **good** algorithm and a **valuable** XAI method?
- How to combine the explanations?

Let's find out how to eXplain the explanation!

Offline High Energy Physics applications useful as they offer a "fully" known pipeline: maximise signal efficiency and background rejection, understand events through features (WP1). Applications in **Real Time System** and **detector** developments equivalently relevant (WP2 and WP3).

This talk focus: Searches for new physics at the ATLAS experiment

- **DARK PHOTON**: light long-lived particles belonging to a new hidden sector not yet discovered because too feebly interacting with ordinary matter.
- **SUSY**: search for dark matter candidates resulting from the decay of new particles predicted by Supersymmetry.

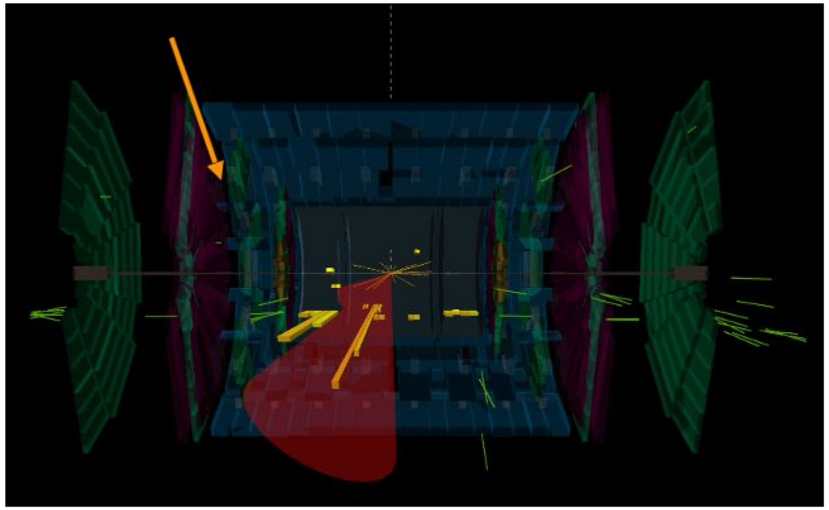
In the backup:

- **Real Time HEP systems – trigger at ATLAS, WP3 (flash talk and poster)**
- **Detector – development of PADME experiment Electromagnetic Calorimeter, WP2**

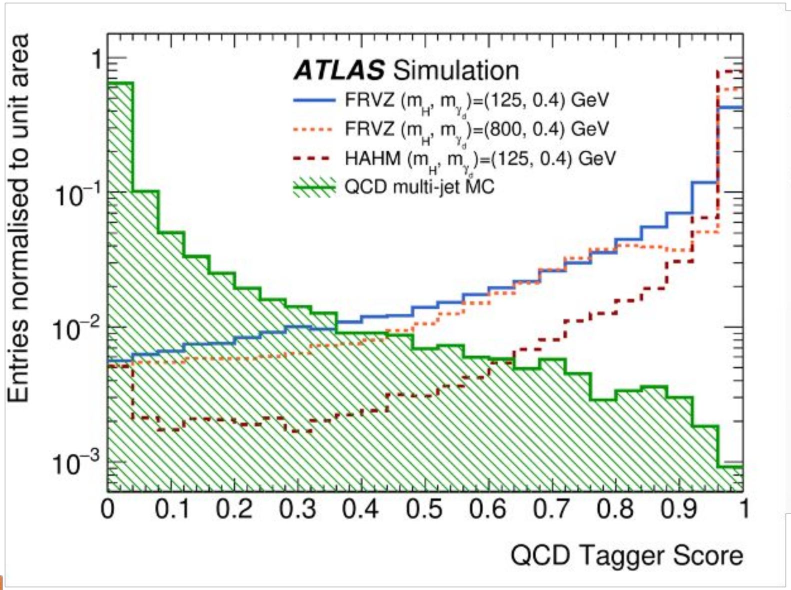
Search for dark-photon at ATLAS

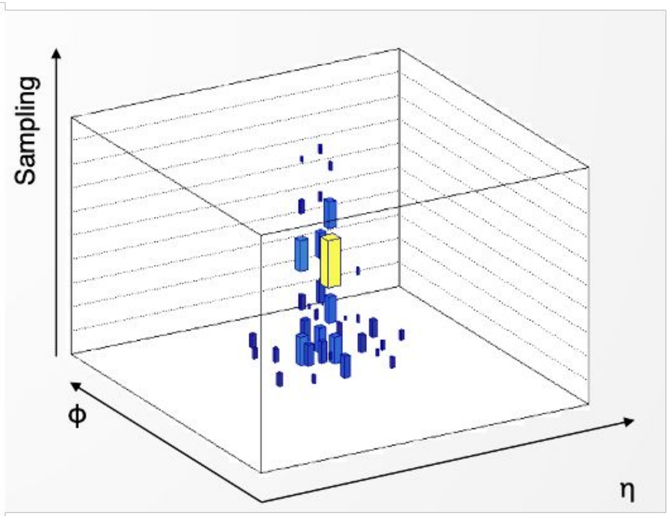
- Dark photons, foreseen in hidden sector models, are produced through SM-like Higgs decays, and decay in electrons, muons or pions → the "signature" is a collimated "jet" of leptons: displaced-lepton jet (DPJ)
- Standard object classification problem where a signal dark-photon leaves different signature in the detector wrt background. ML discriminator (3D-CNN) developed for the publication(s) → uses **image classification** trained to distinguish background processes from signal mapping clusters of particles jets in 3D coordinates

ATLAS calorimeter system



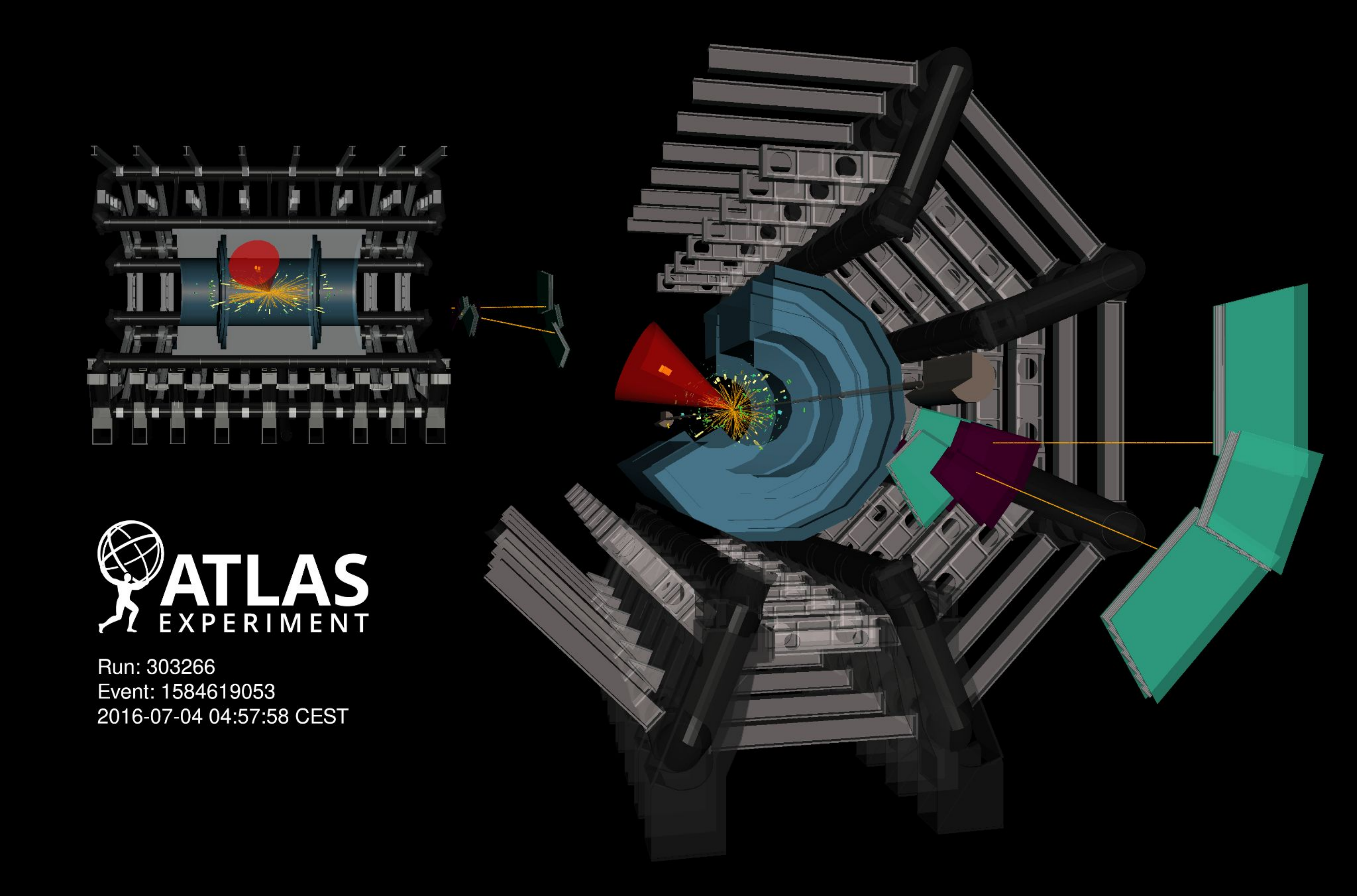
exploit the calorimeter granularity to parametrise the energy deposits: x, y, z, energy





3D jet images: Train a CNN
HOWEVER: Very sparse images -> sub-optimal

"Search for light long-lived neutral particles from Higgs boson decays via vector-boson-fusion production from pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector", submitted to EPJC (Inspire)
As well as following publication exploiting VBF signatures: https://arxiv.org/abs/2311.18298



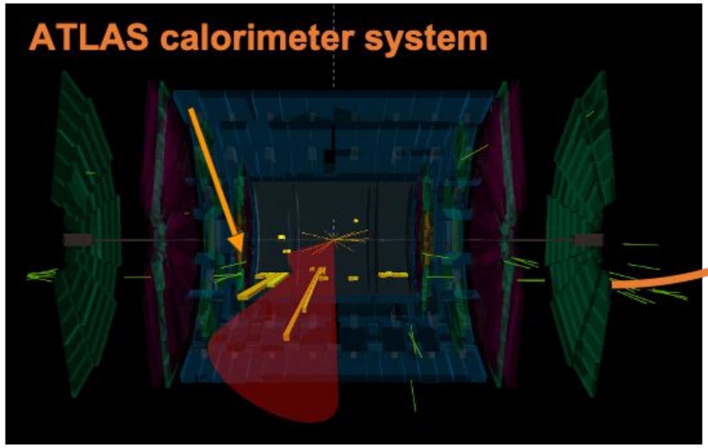
Event display for a data event in the ATLAS detector at CERN. A dark-photon candidate is shown here with this red cone: a highly energetic shower of particles originated far from the interaction point of the collisions.

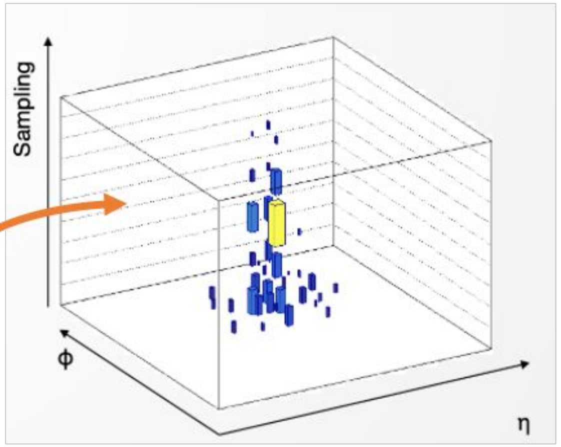
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2019-05/

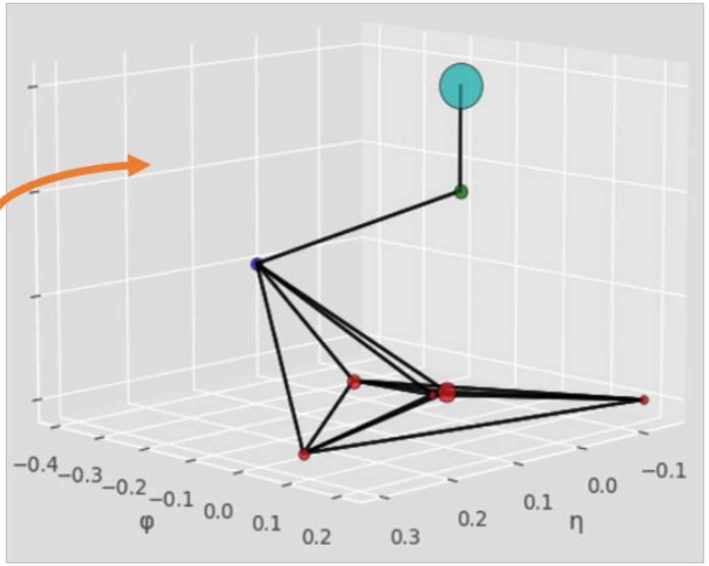
Dark-photon using GNN

Still use image classification trained to distinguish background processes from signal mapping clusters of particles jets in 3D coordinates

Use of additional higher level variables can be added as features to further improve the network performance, although the goal is to have them already ‘learned’ by the network by using only the low level inputs







Graphs: Train a fully optimised GNN
Small cloud space objects, Efficient and easy to manipulate

➢ A visual representation of Jet 3D images using node-by-node correspondence with an upgraded graph structure

Procedure:

- use ~500k images from signal (DPJ) and background (QCD jets) to build input dataset
- test impact of decisions taken a priori (3 models), implement eXplainability tools: PyG explainers (e.g., GNN Saliency Maps) and Captum’s data influence modelling (e.g., TracIn)

Graph Building and Performance

Dataset building:

- Node for every cluster in the calorimeter
- Normalized cluster energy and position as node attributes*
- Edge built if spatial covariant distance “**DR**” between two nodes is within an optimized distance parameter
- Covariant distance normalized as edge weight

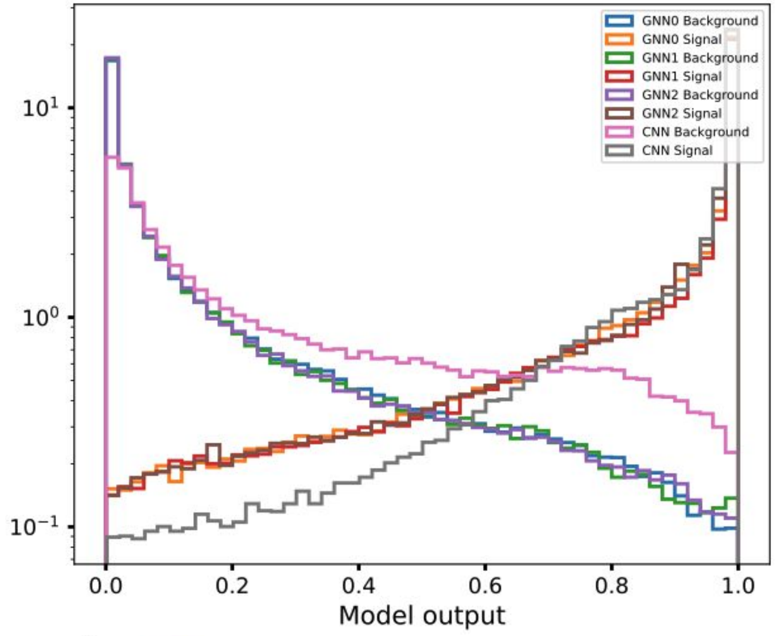
Graph Pre-processing:

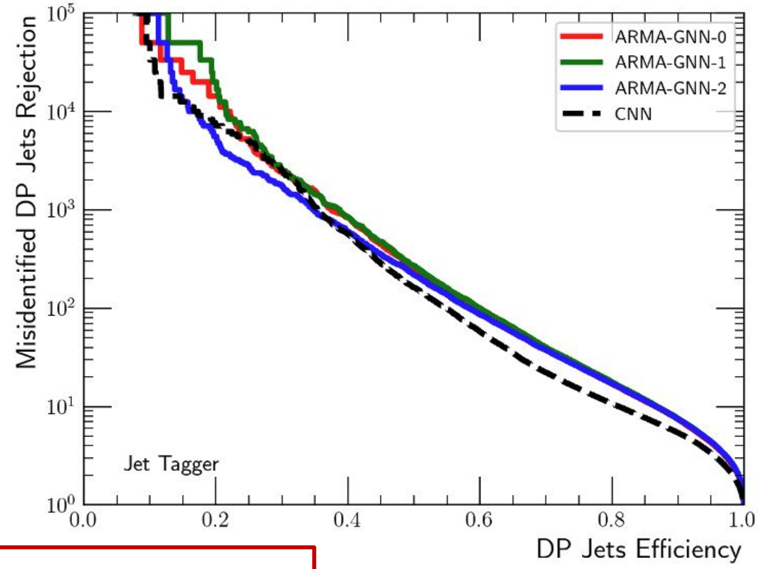
- Remove isolated and self-connected nodes
- Retain largest subgraph only to remove calorimeter noise

Model optimization and XAI implementation:

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

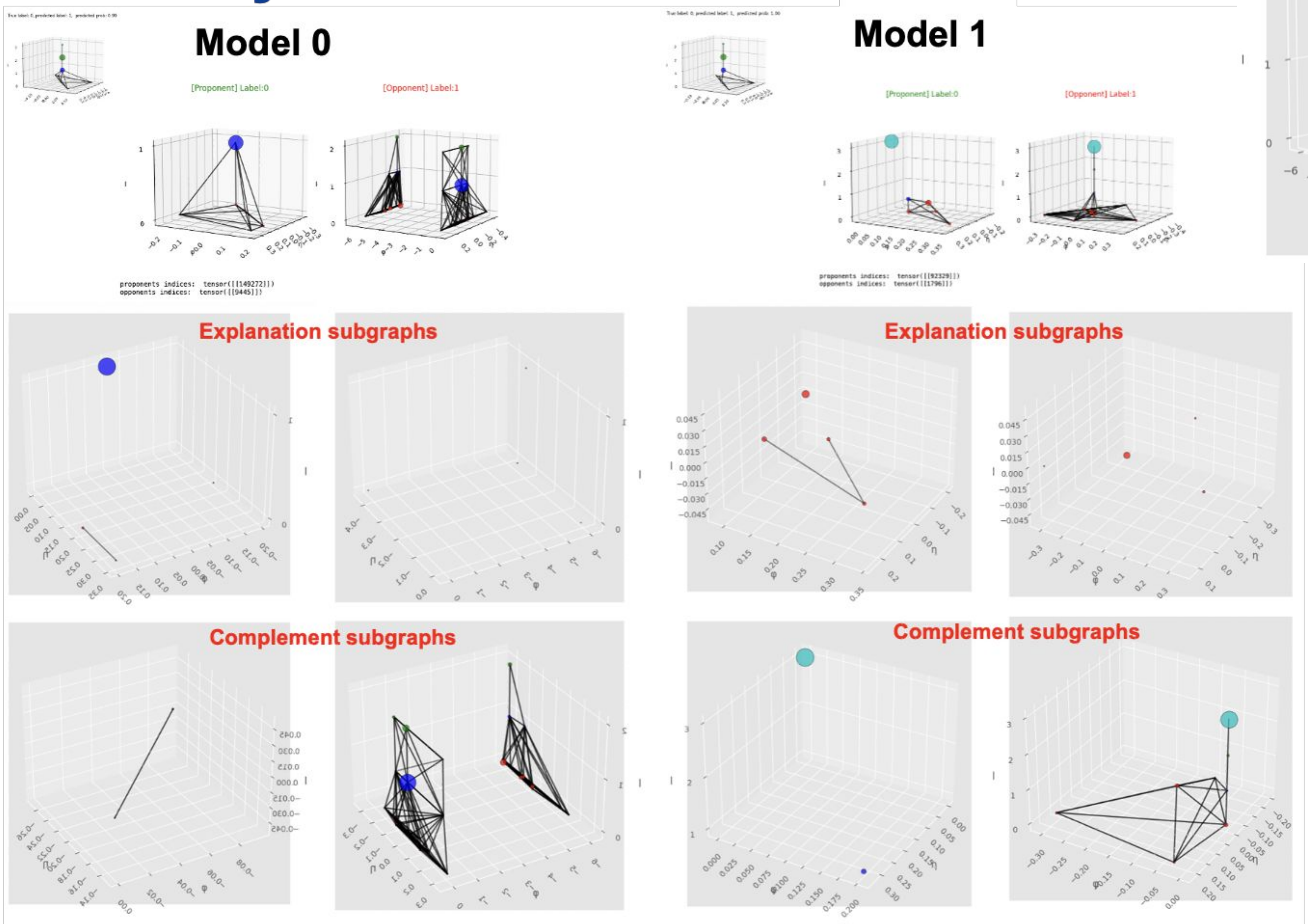
- Test multiple models
  - Model 0 No Preprocessing same as CNN Benchmark data (in graphs)
  - Model 1 Optimised DR = 0.6 within calo layers and 0.3 intra-calo layers (cuts based on performance metrics like accuracy and purity)
  - Model 2 Optimised number of nodes/edges/subgraphs: removing isolated nodes and disconnected sub-leading subgraphs made sense from a physics intuitive perspective and was not the best or at least similar with performance and eXplainability metrics
- Performance evaluation and comparison with 3D-CNN as Benchmark
- Main XAI layer (retrain): TRAC-IN* as data influence metric implemented producing proponents and opponents to any post-training data sampling (e.g., TP, FP & TN sets)
- Additional XAI layers (optional retraining): GNN and PYG Saliency Maps to **explain-the-explainer on the top-k nodes/edges level for any prop/opp sampling**

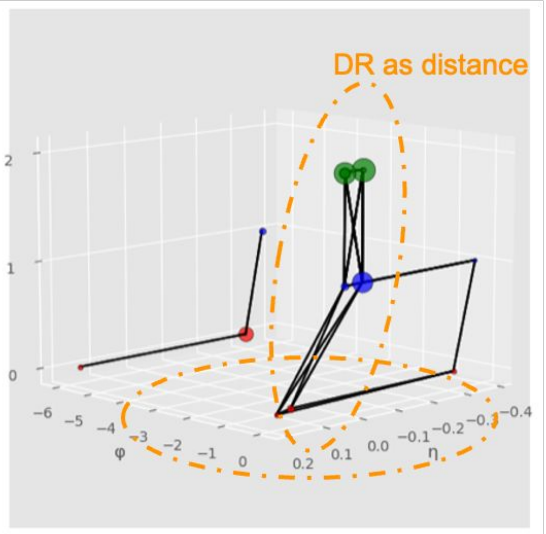




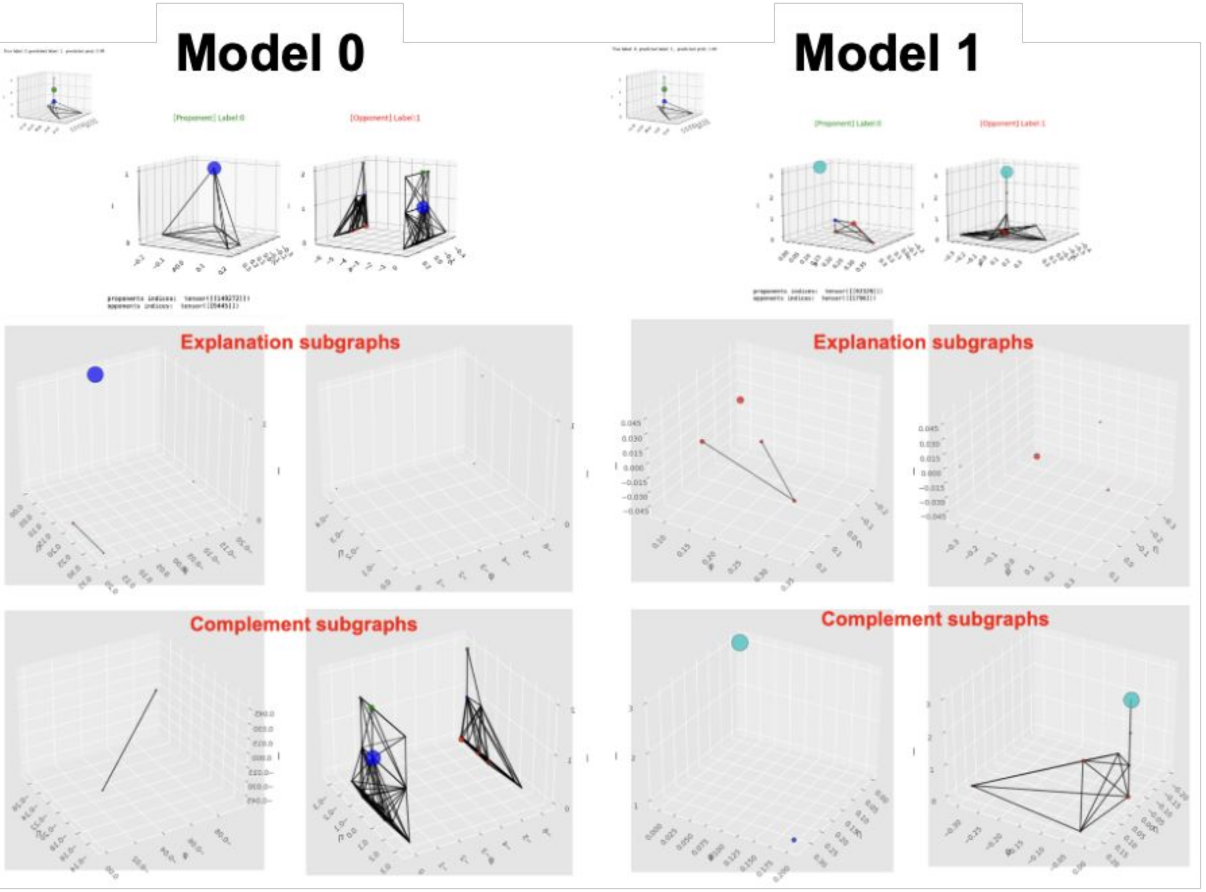
The GNN model out-performed the CNN model on all performance metrics tested at same signal efficiency score

XAI Data Analysis

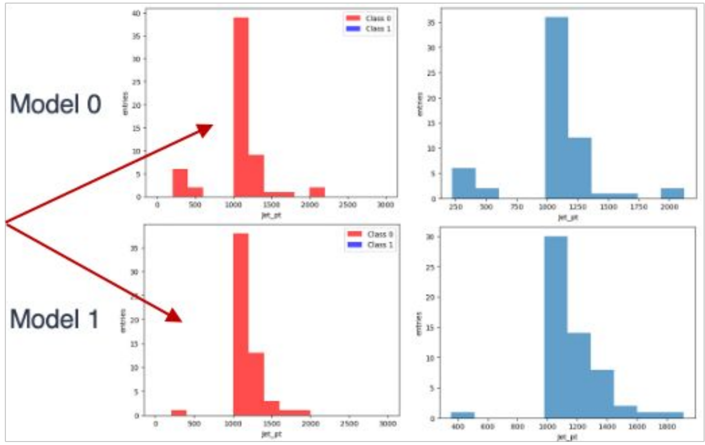




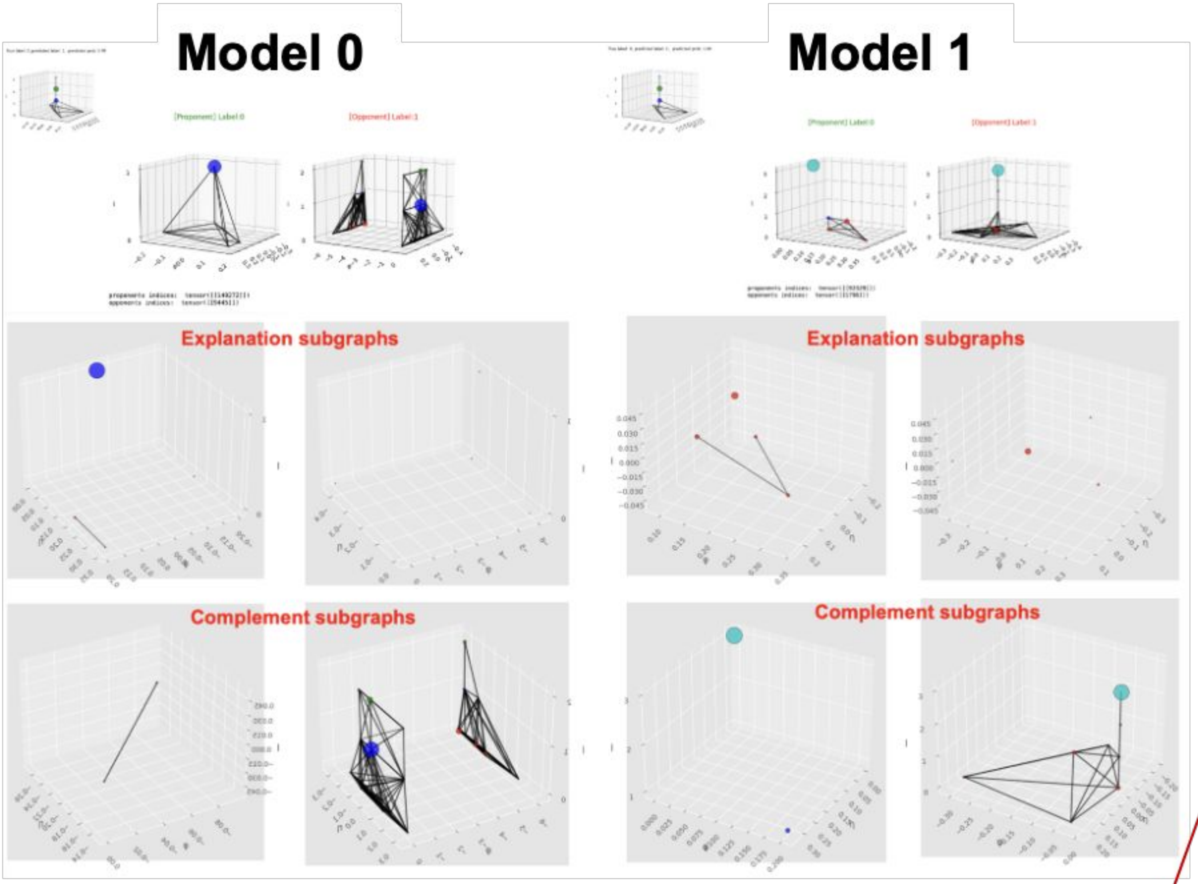
XAI Data Analysis



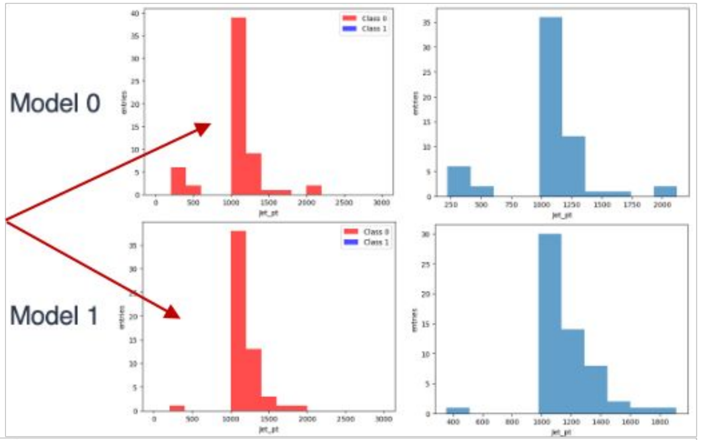
- FP proponents are entirely background as expected



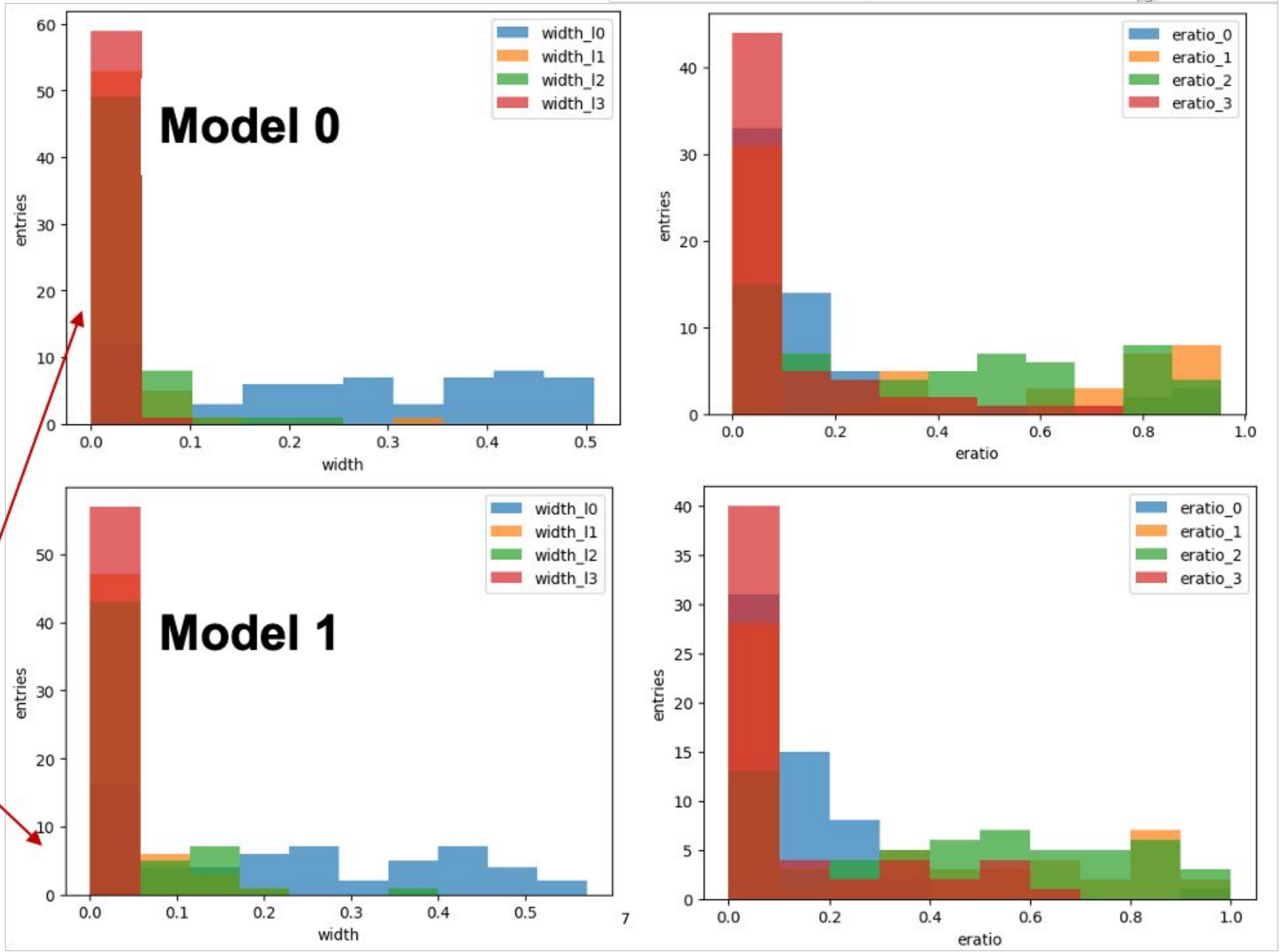
XAI Data Analysis



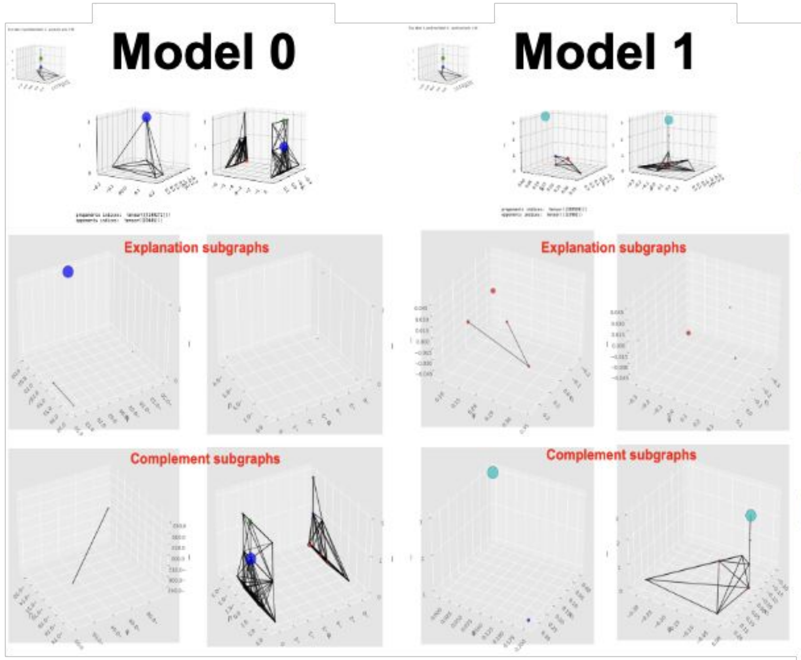
- FP proponents are entirely background as expected

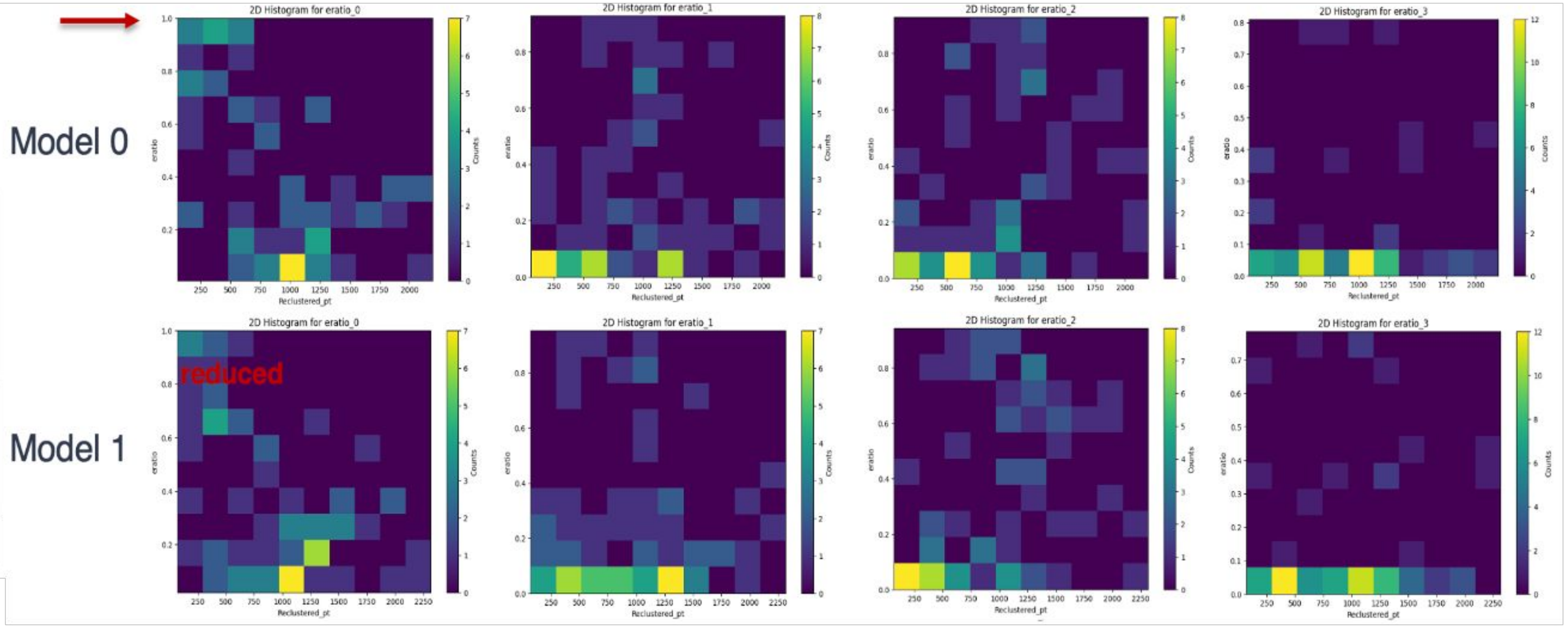


- Saliency Maps masks both nodes and edges show consistent selections majority of nodes and edges retained are **from EMB layer 0**



XAI Data Analysis





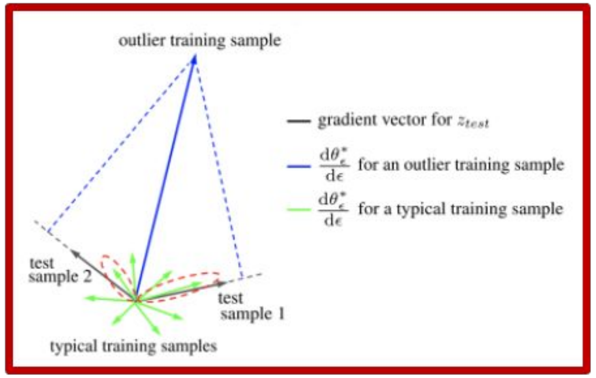
Saliency Maps outputs in 2D plots showing reconstructed JetpT against Energy ratio for each of the 4 layers

- ➢ Trac-In results show steady and consistent trustworthy results by reproducing nearly identical best scoring proponent-opponent **major pair** for all instances of **FPs set**.
- ➢ Trac-In proponents and opponents do not provide self contained explainability but gives more coherent outputs under **Model 1 criterium**.
- ➢ Saliency Maps are essential to explain Captum clear but open ended explainability, i.e., proponent/opponent minimal prototyping needed.
- ➢ **2D plots on the right and top histos from previous slide show reduced activity in layer 0 (low pT range) for FP *Proponents* as an instance**

XAI Recycling

★ Model 1 **retrained** **after removing Global Influencers/outliers "GIs" only**

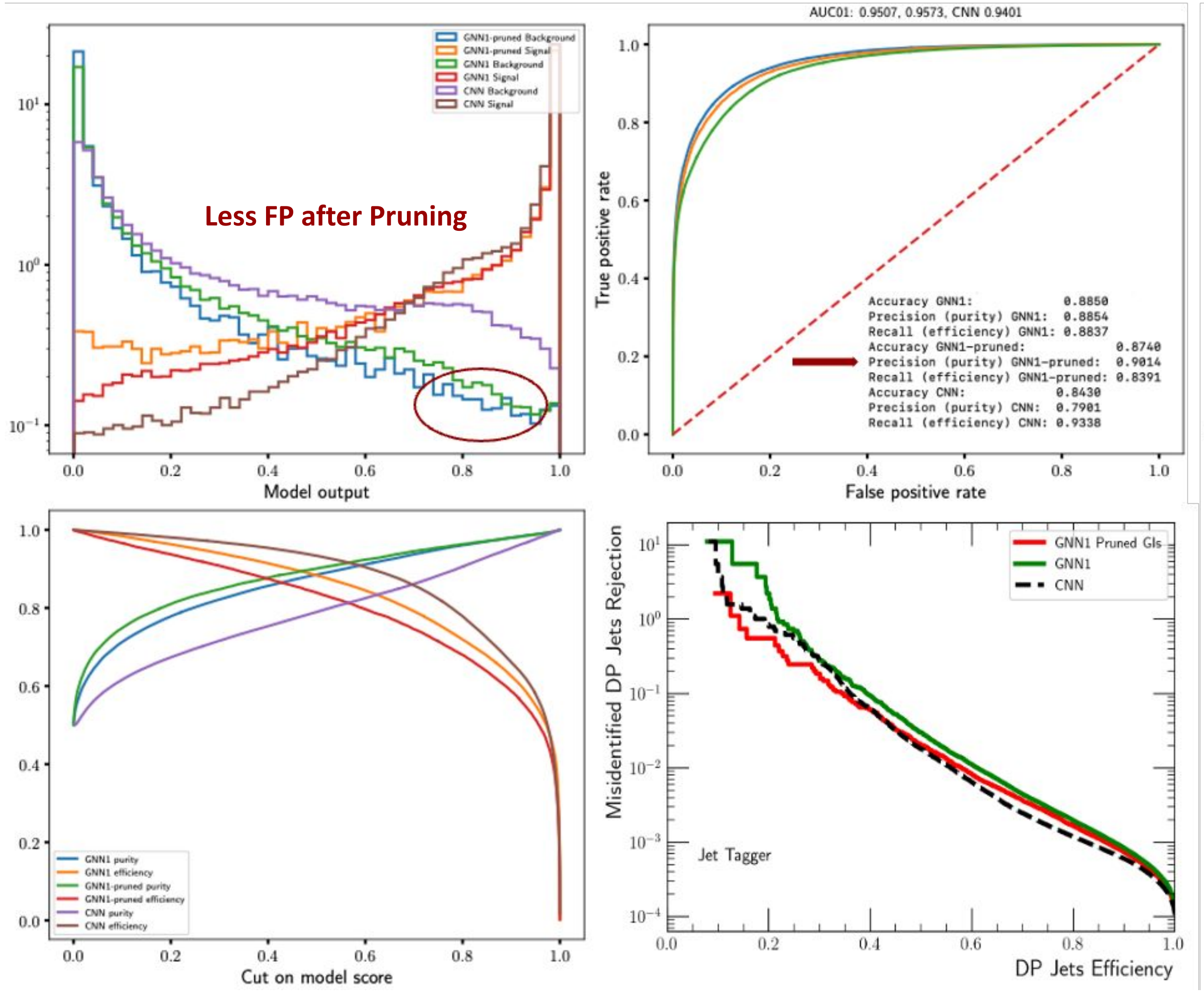
➢ **Leave-Some-Out Retraining Approach: 60%**

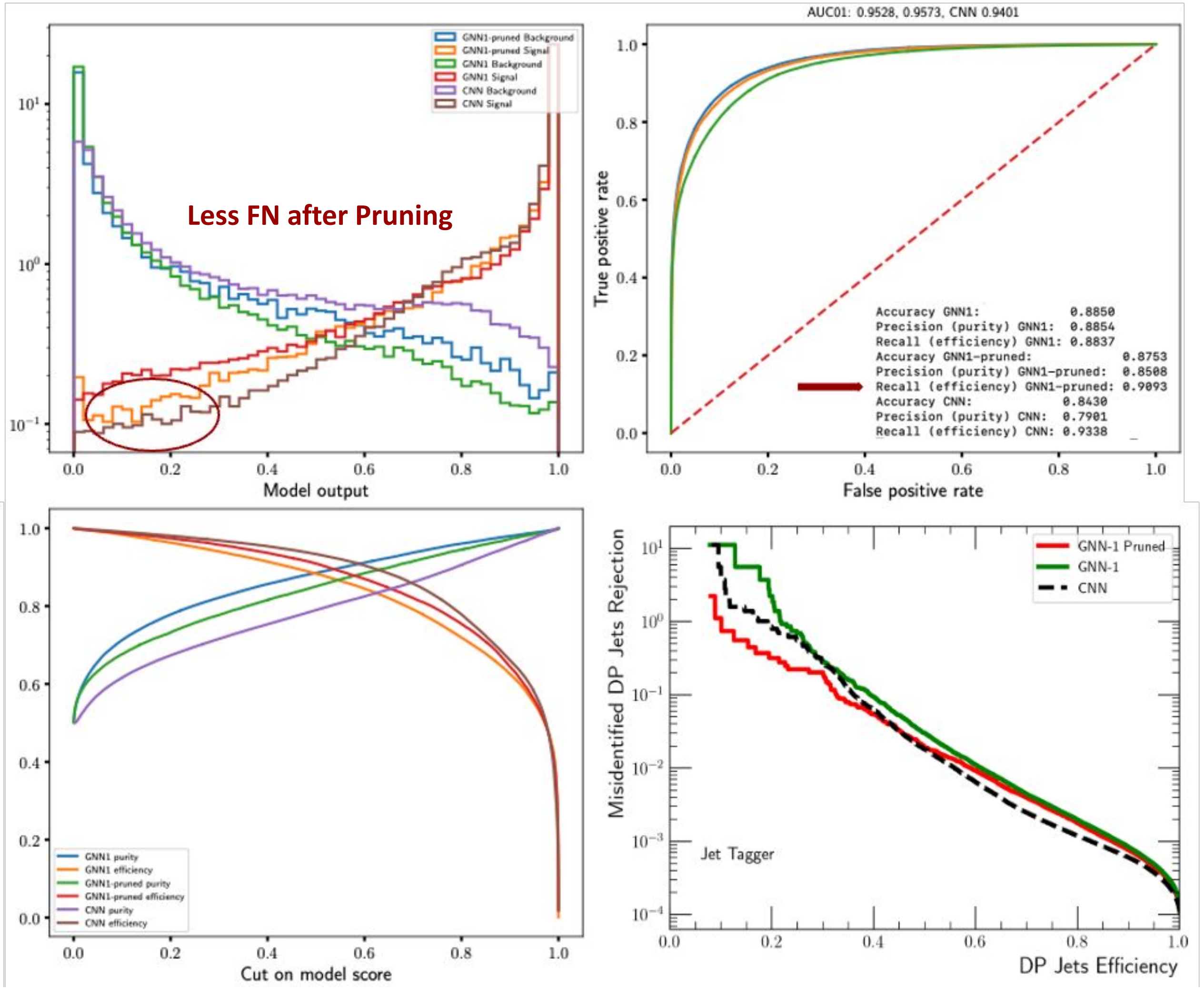


$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

★ Model 1 **retrained** **after removing GIs and proponents FP, GIs and opponents of TP and TN**





Preliminary Conclusions for Dark Photons

From Dark Photons pipeline we could draw some insights from eXplainability layers with the following:

1. Filtering out GIs did produce local data influencers that are relatable in physics terms: Shape of graphs are in coherence with what we expect from Signal vs Background Jet structures to be.

2. Proponents of TP and opponents of both TN and FP were almost totally Signal events, while the opposites were Backgrounds as expected.

3. On the analysis level, we could preliminarily deduce that extra calorimetric activity in Layer 0 is affecting the performance of the overall training. A fair comparison of the model, is made on all level of explainability and trends are especially important when persisting to Saliency Maps; being eXplainer of local data influence eXplainers.

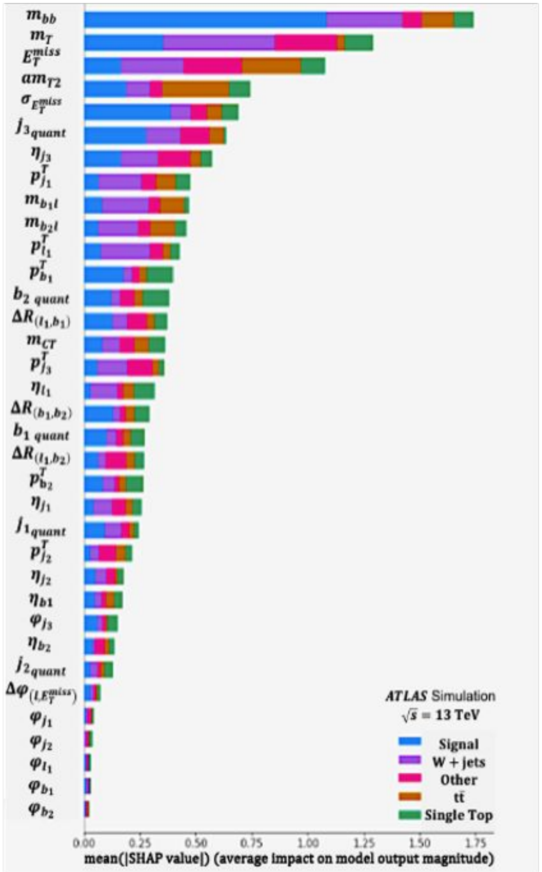
4. When filtering, resampling based on XAI results and retraining we notice slight improvements on some metrics but not strong enough to claim general supremacy.

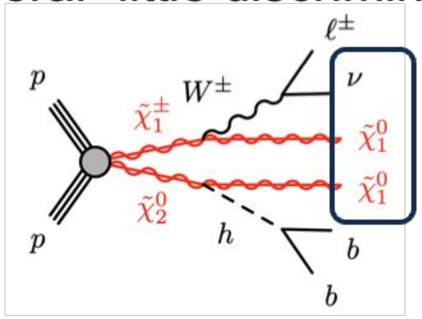
5. Based on 4, we see the need for more powerful self-inherently explainable models like Transformers and foundational models and we propose an example in the following test study with SUSY analysis.

★ More details of the study will be found in the paper (**out very soon**)… stay tuned!

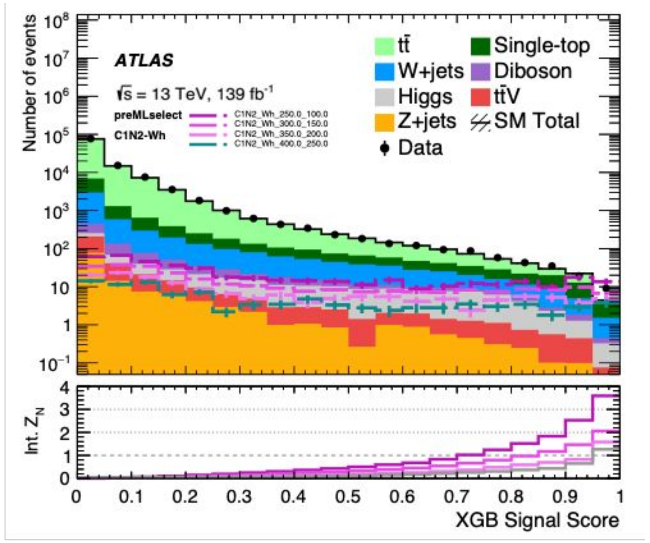
SUSY search at ATLAS: chargino-neutralino

- ”classical” case of rare signal over large SM background, with kinematic observables offering -in general- little discriminating power





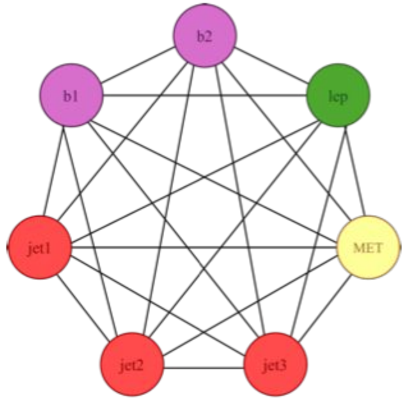
Missing ET (MET)



- For the data analysis, use a BDT (XGBoost) exploiting 30 variables (object based and complex)

- SHAP interpretability tool used to understand the relevance of the variables

- In our project, develop **GNN and Graph transformers**
→ Build graph for each event, one node for each particle and different features in nodes
→Test multiple models and evaluate performance

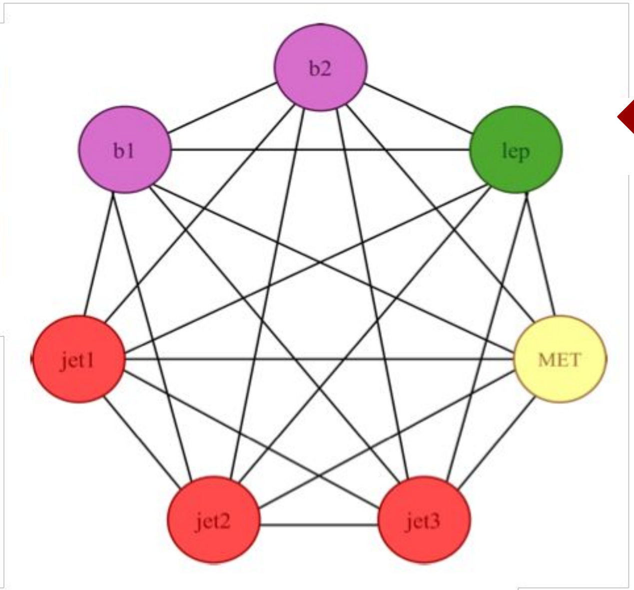


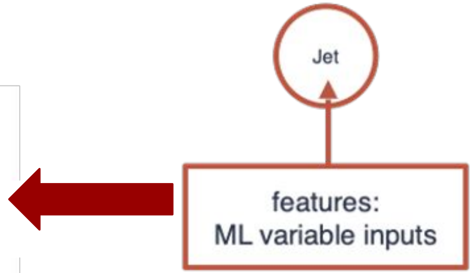
“Search for direct production of electroweakinos in final states with one lepton, jets and missing transverse momentum and in pp collisions at √s=13 TeV with the ATLAS detector”, JHEP 12 (2023) 167
Datasets and results published (CERN opendata and hepdata) and disseminated here

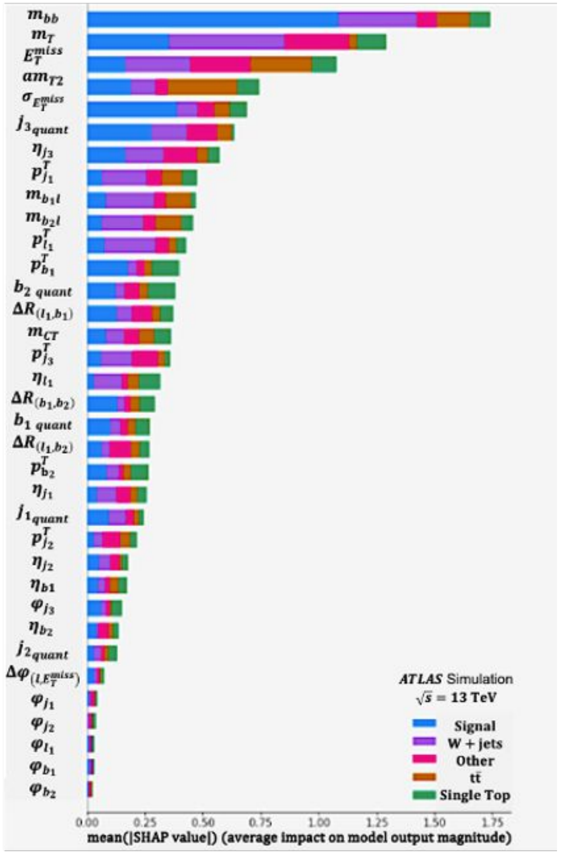
SUSY Pipeline: a closer look

Particle	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
jet1	'pTj1'	'etaj1'	'phij1'	'j1_quantile'	nan	nan
jet2	'pTj2'	'etaj2'	'phij2'	'j2_quantile'	nan	nan
jet3 (optional)	'pTj3'	'etaj3'	'phij3'	'j3_quantile'	nan	nan
b1	'pTb1'	'etab1'	'phib1'	'b1_quantile'	'b1m'	nan
b2	'pTb2'	'etab2'	'phib2'	'b2_quantile'	'b2m'	nan
lepton	'pTl1'	'etal1'	'phil1'	nan	nan	nan
energy	'ETMiss'	nan	'ETMissPhi'	nan	nan	'metsig_New'

Features of nodes







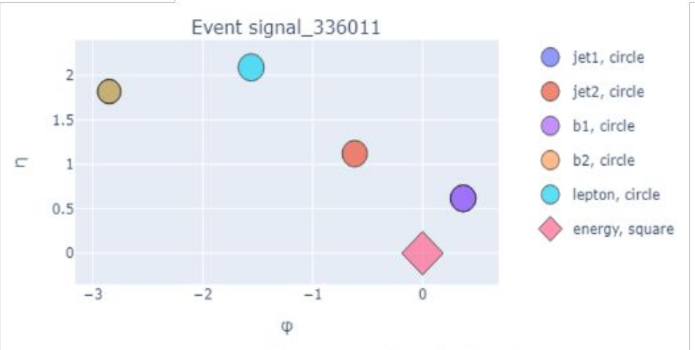
Dataset

- Each row of the dataset contains 1 graph of 6 or 7 nodes.
- Each graph is fully connected.
- Each graph has a maximum of 6 features.

Three types of graphs:

- **Signal:** SuSy Dark matter MC candidate events
- **Background1**: top-antitop quark pair decay Jets
- **Background2**: Single top quark decay Jets

➢ **Training a graph based model that performs binary classification (i.e. recognizes signal and background events)**

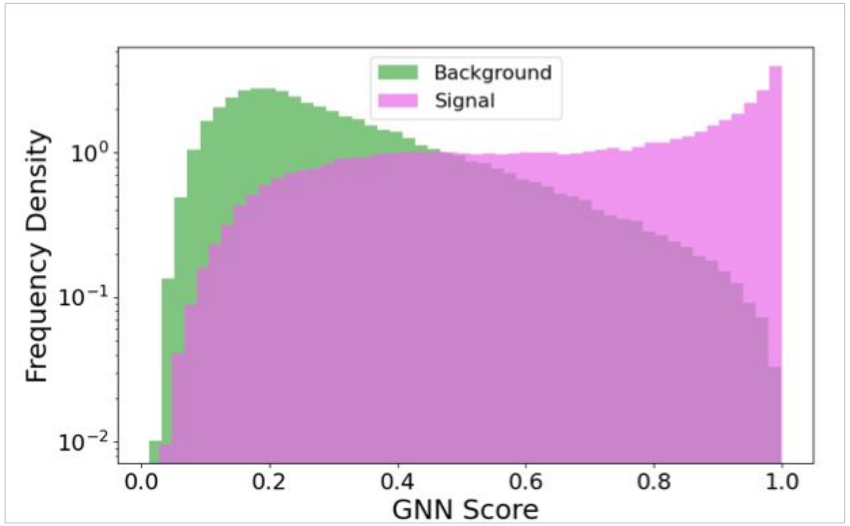


Representation of a Graph

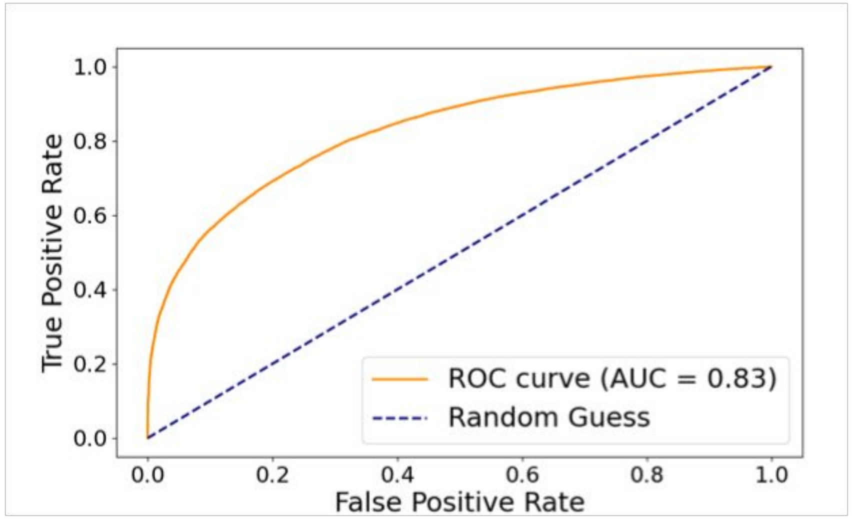
Signal events: 450k (same as BDT)

- Background1 events: 590k (BDT trained on 6m)
- Background2 events: 240k (BDT trained on 796k)

SUSY GNN Preliminary Results and Interpretation

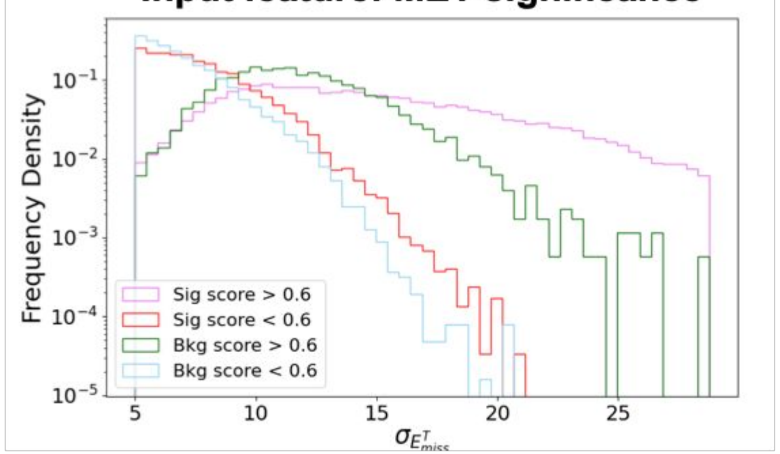


- Clearer distinction of signal vs background in score distribution wrt to XGBoost

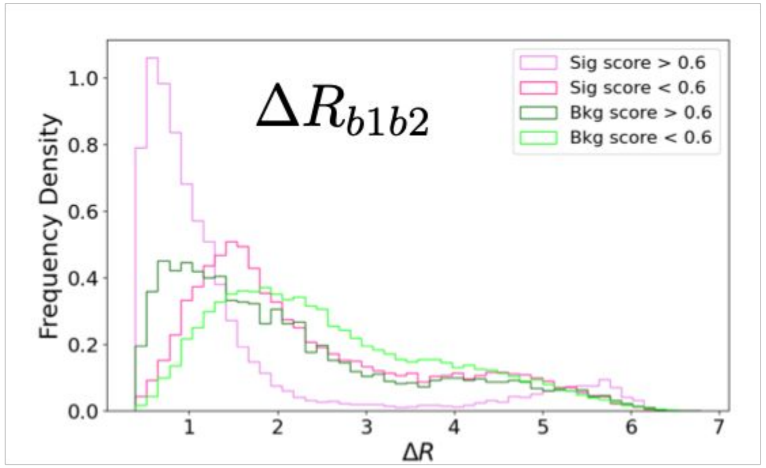


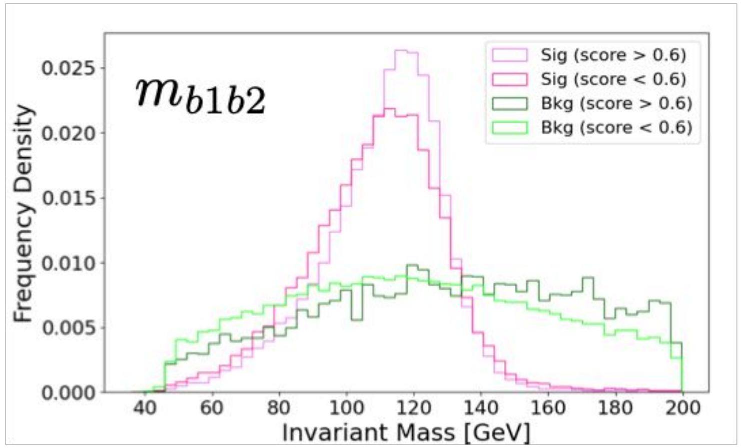
- The GNN learn from individual input features and ”transfer” knowledge to complex variables without need to use them for training like in XGBoost

Input feature: MET significance



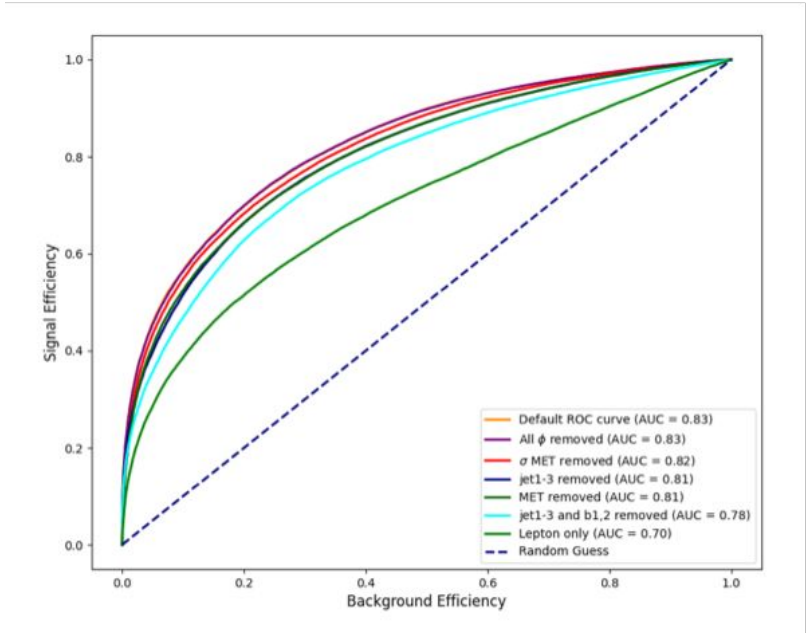
Output distributions





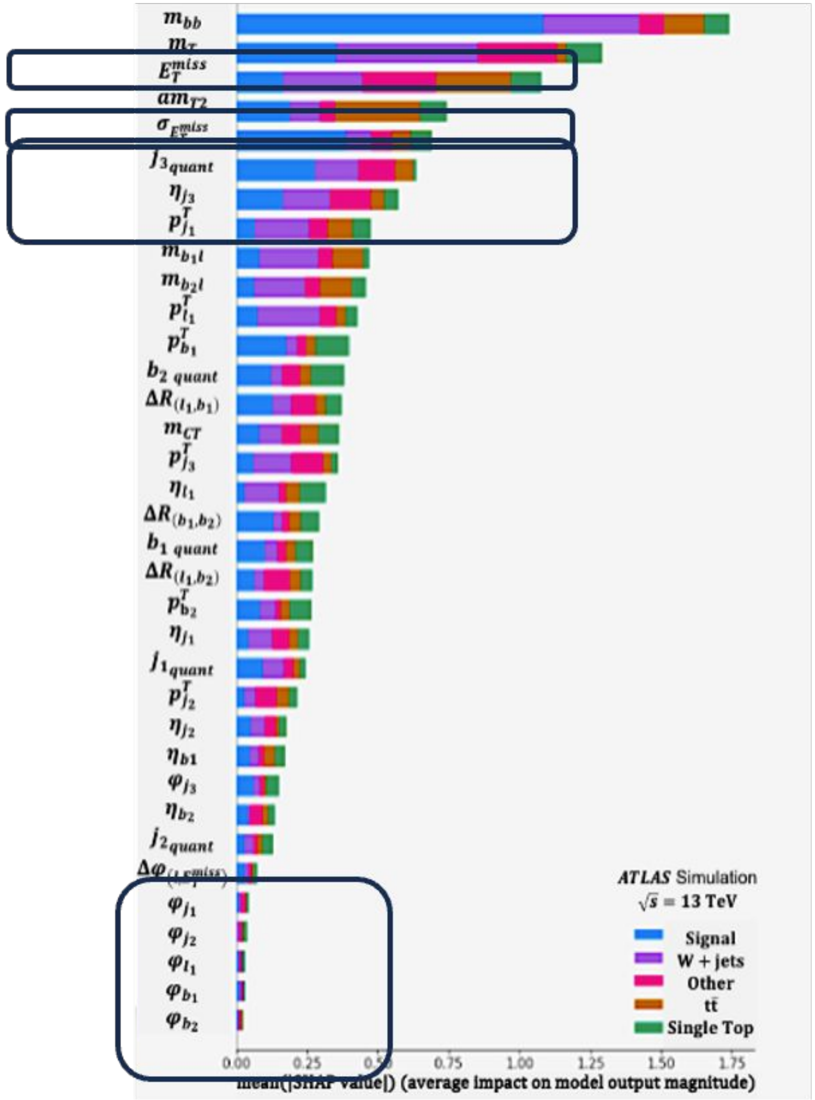
SUSY GNN results and interpretation

- Occlusion tests to understand how the network learns
- Hidden features progressive test:
  - All jet1-3 features,
  - The E_T^{miss} feature,
  - The feature for $\sigma_{E_T^{miss}}$,
  - All ϕ features, (not relevant according to SHAP)
  - All jet1-3, b1 and b2 features,
  - All features except lepton input features.



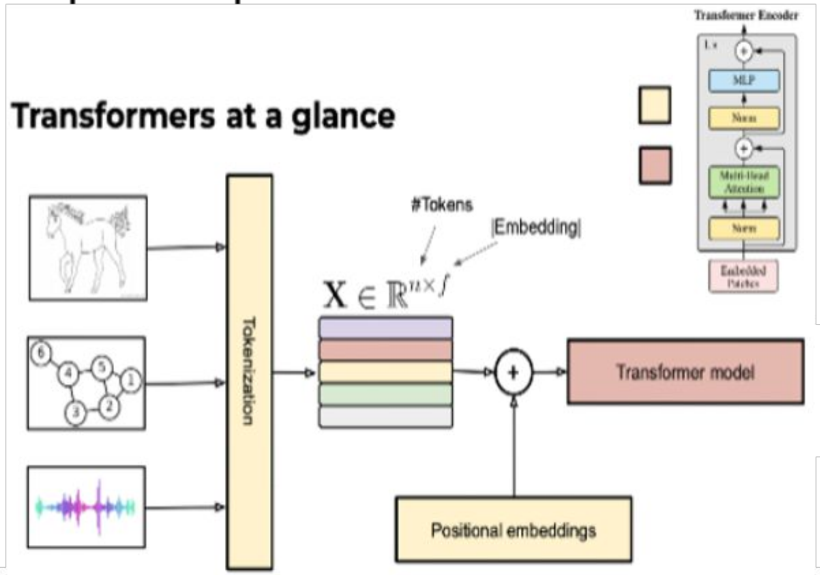
Occlusion	AUC Value
Default Value (No Occlusion)	**0.83**
All jet1-3 features	0.81
The E_T^{miss} feature	0.81
$\sigma_{E_T^{miss}}$ feature	0.82
All ϕ features	**0.83**
All jet1-3, b1 and b2 features	0.78
All nodes except lepton	0.70

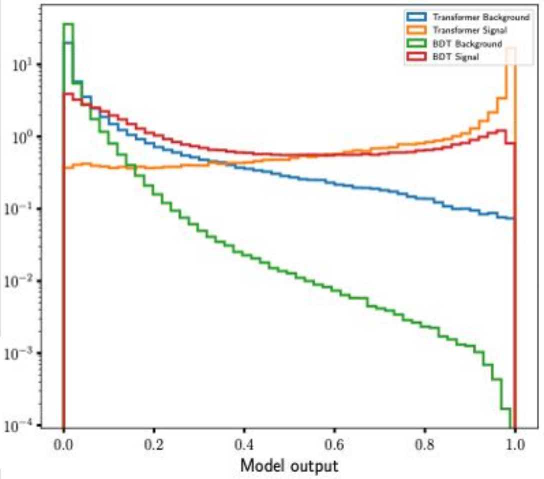
Preliminary conclusion: the GNN need less variables to understand the signal → hidden kinematic correlations easily exploited

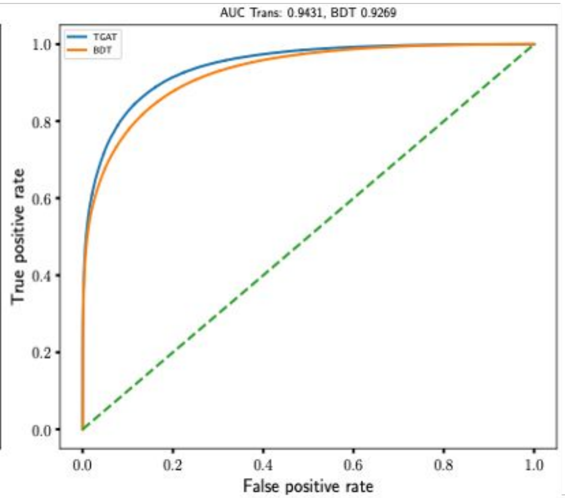


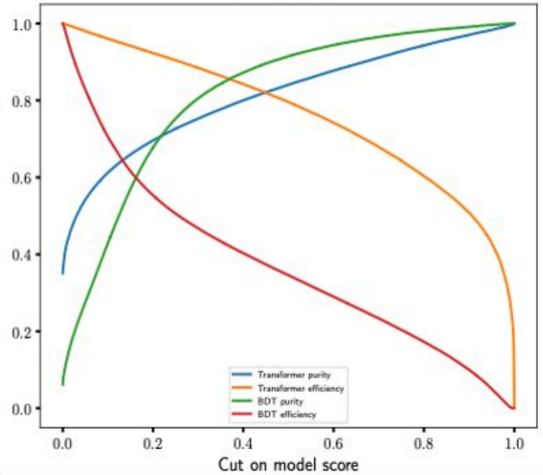
Moving to GNN Attention Transformers

➢ To better understand how to network learn and implement eXplainability directly, tested a couple of implementation of Attention Transformers.









T-GAT Model and Hyperparameters

- Tokenized inputs
- Graph Transformer with 2 GAT layers: 1 with 3 heads
- Features projected onto Query, Keys and Values.
- Adam optimizer with a cosine annealing LR 1e-3
- BCE with Logits loss, batch size of 512
- Signal events: 450k (same as BDT)
- Background1 events: 240/590k (BDT trained on 6m)
- Background2 events: 240k (BDT trained on 796k)

The two bkg are different in kinematics, currently studying dependency on samples size

➢ Transformer models performed better than GNNs shown before

➢ **AUC score and ROC curves** at the same level and slightly better than BDT multi-classifier optimized for the analysis

XAI Pipeline and preliminary analysis

Produce a «global» interpretation of the model. Tested with 3 and 4 Heads (shown here).

We first select these subsets of the dataset:

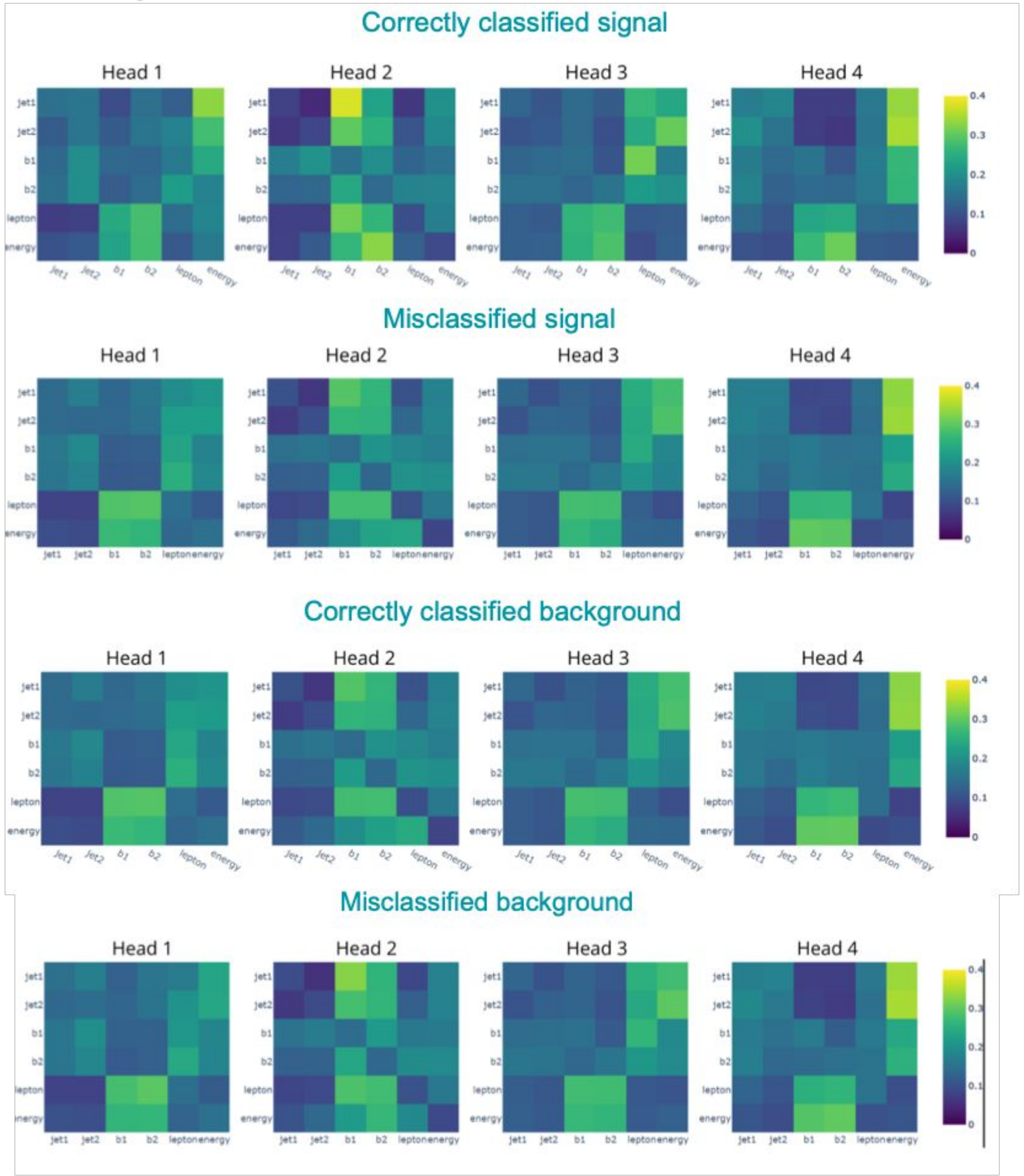
- The training set
- The test set
- **Signal correctly classified**
- **Signal misclassified**
- **Background correctly classified**
- **Background misclassified**

Then we print the attention matrix for each head (in average) and the attention for each node.

The idea is to use the attention scores to understand which connections of neurons are significative for the predictions.

Analysis of preliminary results show that attention scores reflect expected behaviour, i.e.:

- for correctly classified signal events, lepton pT and MET pay more attention to b1 and b2 (correlation expected).
- configuration of attention weights for misclassified signals is well overlapping with the configuration for correctly classified background



Conclusions and Prospects

★ AI techniques based on graphs are highly effective for classification application in High Energy Physics

★ Overall, some insights from eXplainability layers obtained using the two HEP benchmarks. Still, decoupled xAI techniques have limitations for easy-to-glimpse information for domain experts/scientists → **explainable-by-design** family of neural networks would be more useful in future.

In this talk, we have presented current investigations using the DARK PHOTON and SUSY searches as benchmark

★ DARK PHOTON:
- Reshaped and optimised DARK PHOTON ATLAS analysis training with GNN
- Explore XAI analysis options (Saliency Map and TracIn), as well as impact of global influencers - overall some expected features obtained, although less clear than expected

★ SUSY:
- Search XAI pipeline and data analysis in similar fashion to DARK PHOTON search: use of GNN show similar performance to BDT
- Optimised model with features and architectures that are explainable-by-design like Attention based Graph Transformers
- Replace graph convolution and spectral based models with Transformers showing most performing, while keeping same foundational selection on data and conditions in node/edge pruning → Attention scores show expected correlations for relevant features

Plans: wrap up results and publish, apply approaches to other cases to evaluate xAI and ability to understand NN

BACKUP

The Consortium

Sapienza University of Rome (IT)
Departments of Physics, Physiology, and Information Engineering

Funding agency: MUR

HEP: data-analysis, detectors, simulation AI: ML/DL methods in basic/applied research and industry, intelligent signal processing. Neurosciences: brain encoding of complex behaviours, ML in electrophysiology, multi-scale modelling approaches

Istituto Nazionale Fisica Nucleare (IT)
Rome group

Funding agency: INFN

Fundamental research with cutting edge technologies and instruments, applications in several fields (HEP, medicine imaging/diagnosis/prognosis/therapy)

Medlea S.r.l.s (IT)

MedLea

Funding agency: MUR

High tech startup, with an established track record in medical image analysis and high-performance simulation and capabilities of developing and deploying industry-standard software solutions

University of Sofia St.Kl.Ohridski (BG)
Faculty of Physics

Funding agency: BNSF

Extended expertise in detector development, firmware, experiment software in HEP

Polytechnic University of Bucharest (RO)
Department of Hydraulics, Hydraulic Equipment and Environmental Engineering

Funding agency: UEFIS CDI

Complex Fluids and Microfluidics expertise: mucus/saliva rheology, reconstruction and simulation of respiratory airways, AI applications for airflow predictions in respiratory conducts

University of Liverpool (UK)
Department of Physics

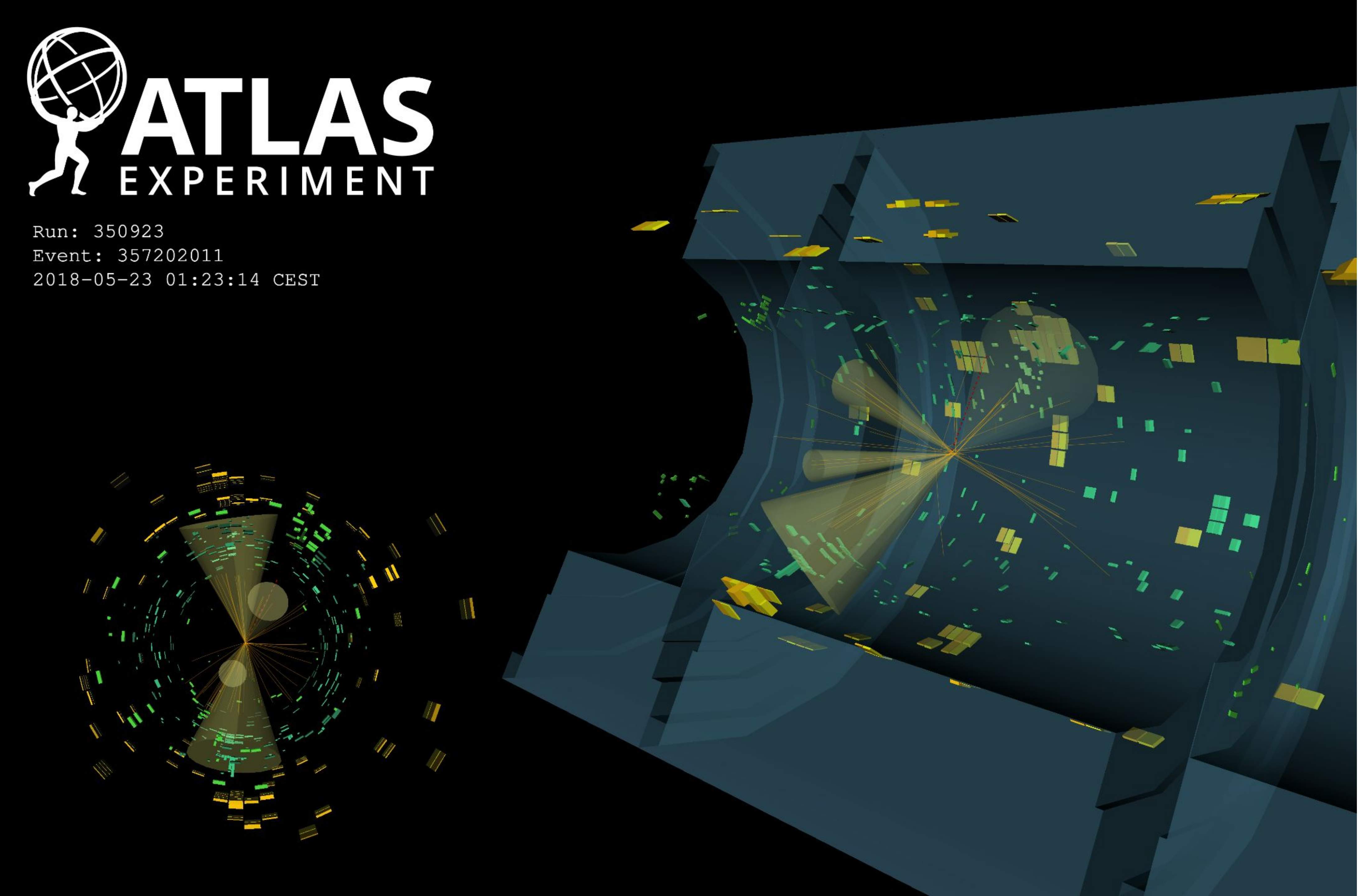
Funding agency: UKRI

Physics data analysis at hadron colliders experiments, simulation, ML and DL methods in HEP

Istituto Superiore di Sanità

Funding agency: INFN

Expertise in neural networks modeling, cortical network dynamics, theory inspired data analysis

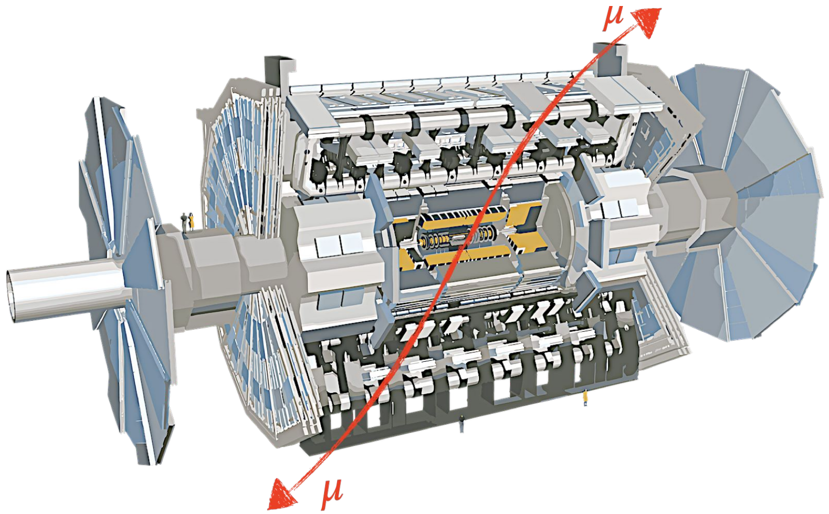


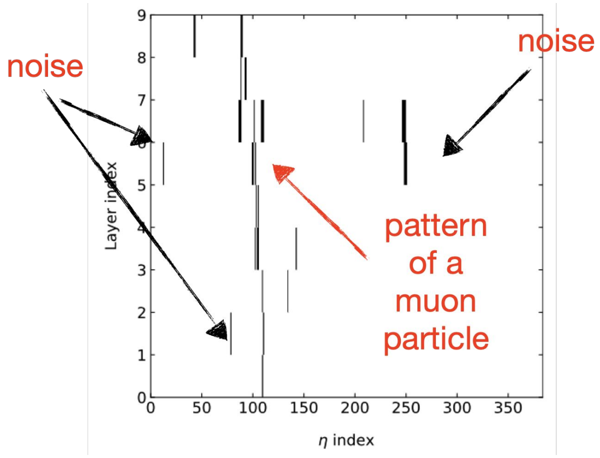
SUSY searches: Event display for a higgsino-like event in the low-mass channel of the multi-b search. Four jets (yellow cones) produced in the decay of the two Higgs boson candidates are observed, with low missing transverse momentum.

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2023-048/fig_13.png

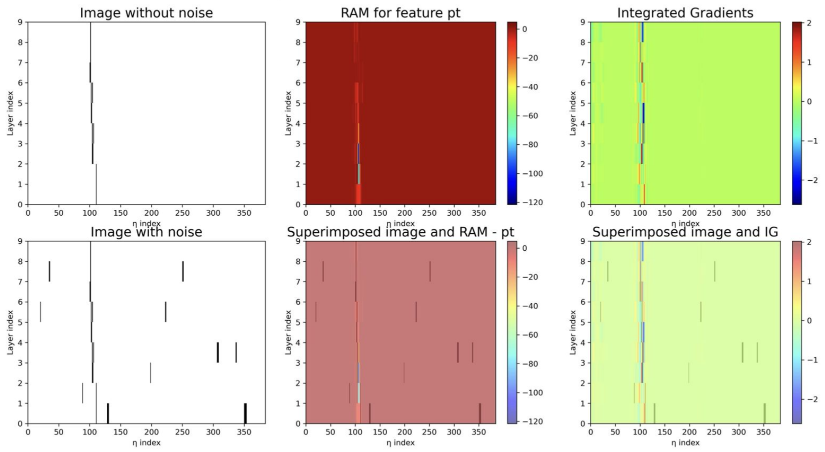
Real time HEP systems

Developed complete pipeline for a real-time AI based event selection. Explored an array of xAI (Attribution, Training influence) methods based for easy-to-understand explanations of models' predictions. Reported strengths and drawbacks in this particular scenario. Developed a novel explainability techniques based on Convolutional Soft Decision Trees





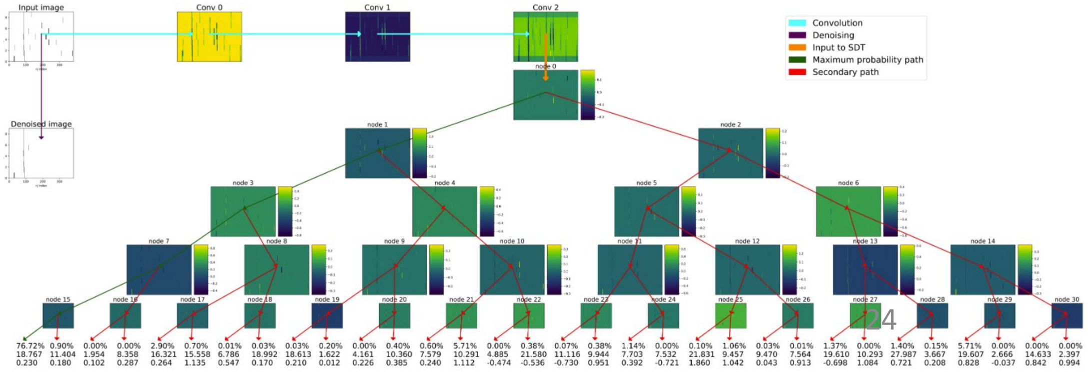
Attribution Algorithms:
Regression Activation Maps VS Integrated Gradients



- ultra-fast (<**400ns/inference**) DNN for identification of muonic particles in the muon spectrometer of the ATLAS detector at the LHC
- test xAI techniques over **extreme sparse data** and heavily compressed and quantised neural network models

➢ Flashtalk with Poster from yesterday

Distillation to Convolutional Soft Decision Trees

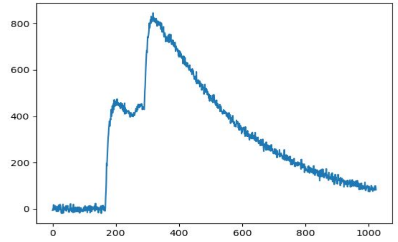


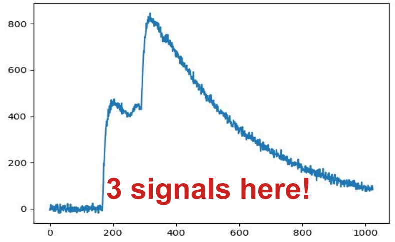
Eur. Phys. J. C 81, 969 (2021), Comput Softw Big Sci 7, 8 (2023)

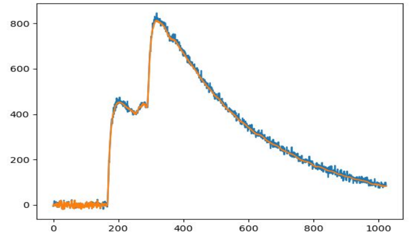
HEP detector

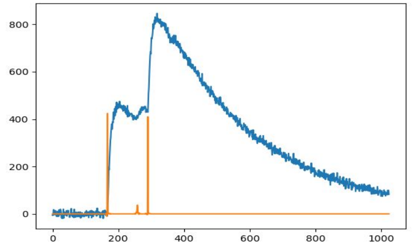
Development of a CNN autoencoder model for signal reconstruction. A time resolution better than 1 ns was achieved which is consistent with the needed performance of PADME Electromagnetic calorimeter

Signal simulation → Simple CNN for counting → Autoencoder → Modified autoencoder







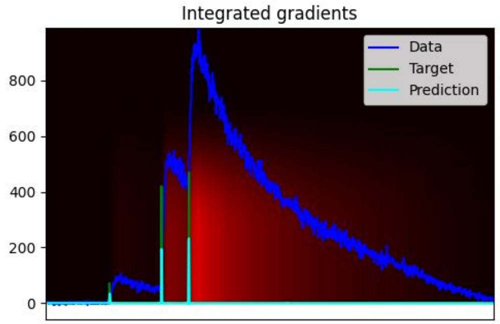


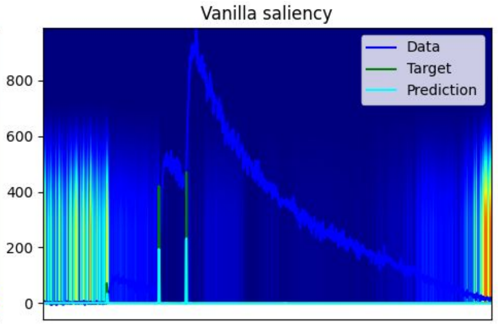
- Generation of noise + several waveforms similar to the expected real data from particle detectors

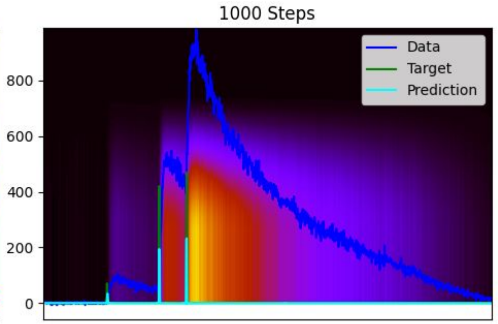
Classification task to identify the number of pulses in a waveform

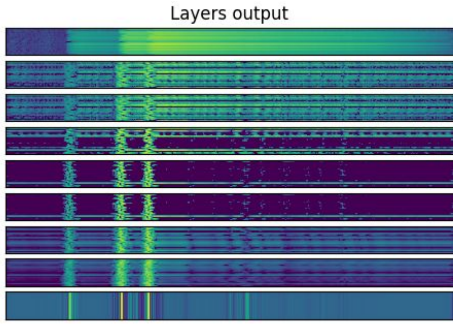
Convolutional autoencoder for signal and noise description

Desired output contains information about the time and amplitude







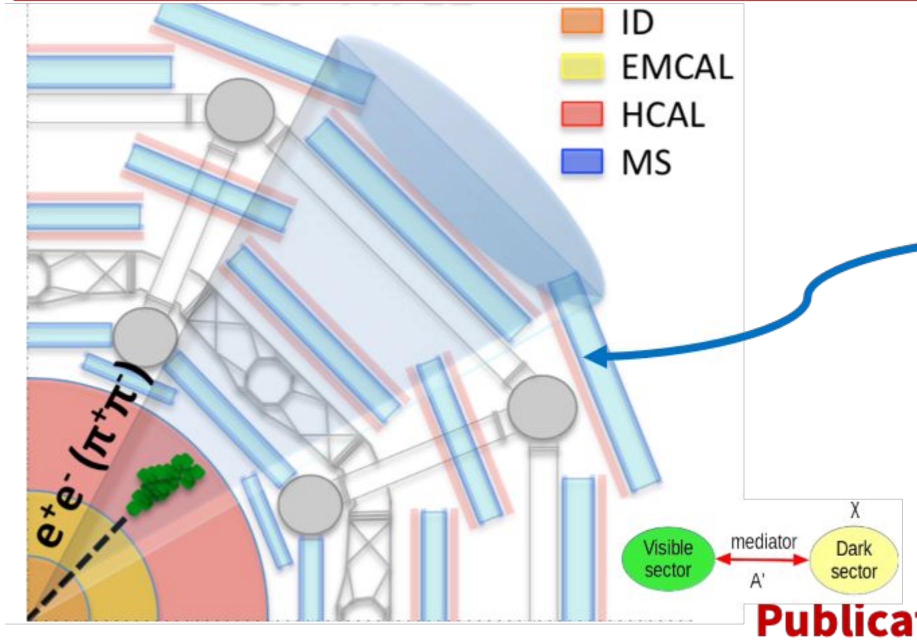


- All developed models were investigated with various explainability methods: integrated gradients, vanilla saliency, activations visualisation
- The best performing model is successfully introduced to the PADME and currently used in the analysis of real experiment data

K. Dimitrova on behalf of the PADME collaboration et al., Using Artificial Intelligence in the Reconstruction of Signals from the PADME Electromagnetic Calorimeter, Instruments 6 (2022) 4, 46
Valentin Buchakchiev et al., Pattern recognition and signal parameters extraction using machine learning methods, J. Phys.: Conf. Ser. 2668 (2024) 012001

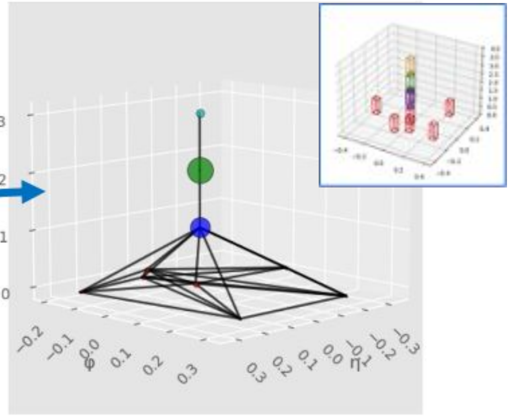
Search 2 – dark-photon

The ATLAS detector orthogonal view



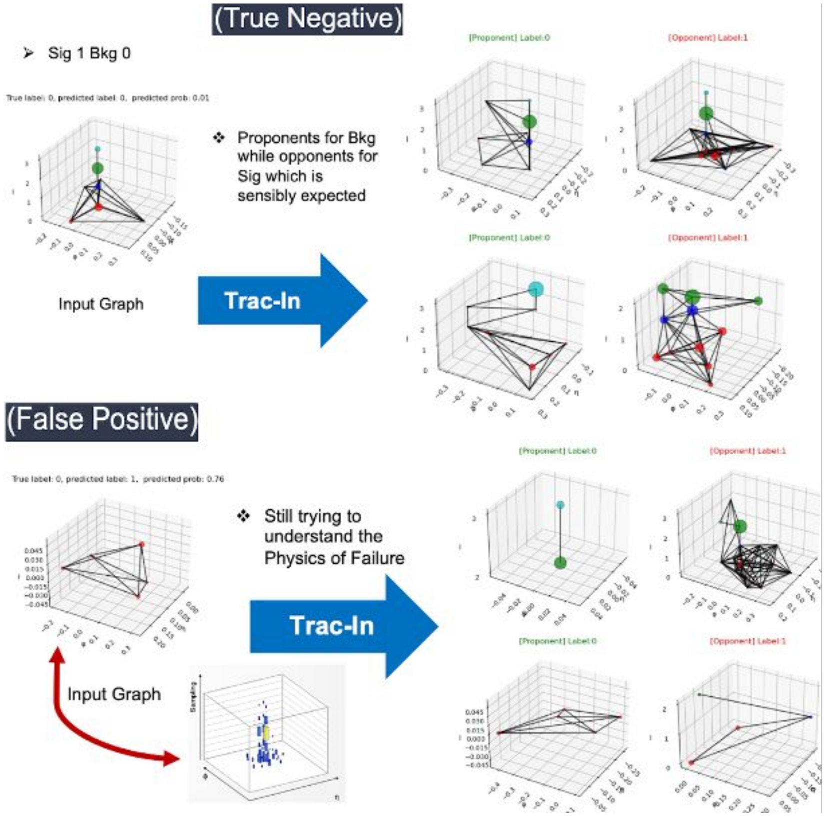
Explainability

- How does a signal look like?



- How does the NN understands true positive (signal), true negative or false positive ?

Publication ➡ **Work in progress!**



Search 1 - DARK – for “dark” photons, not yet discovered new particles. From images to graphs:

Dataset building:

- Build a graph for each ‘jet of particles’, one node for every energy cluster in the calorimeter (energy and position as node attributes), edges connected depending on spatial distance between nodes
- Model optimization and XAI implementation of TRAC-IN (data influence) and saliency maps.

“Search for light long-lived neutral particles from Higgs boson decays via vector-boson-fusion production from pp collisions at √s = 13 TeV with the ATLAS detector”, submitted to EPJC (Inspire)

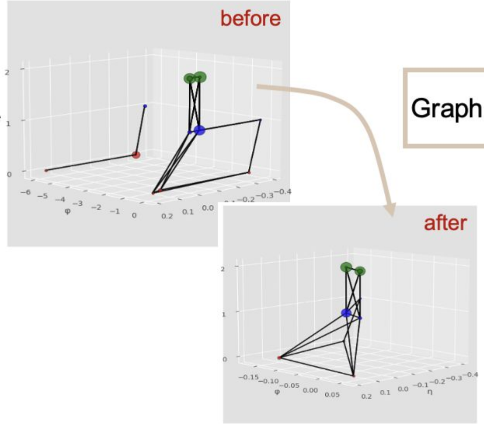
XAI Pipeline and model variants

Process **RAW** data information from **ATLAS calorimeter**: energy deposits relative position and energy distribution

Dataset building:
- Node for every cluster in the calorimeter
- Normalized cluster energy and position as node attributes*
- Edge built if spatial covariant distance between two nodes is within an optimized distance parameter
- Covariant distance normalized as edge weight

Graph Pre-processing:
- Remove isolated and self-connected nodes
- Retain largest subgraph only to remove calorimeter noise

Model optimization and XAI implementation:
- Test multiple models
- Performance evaluation and comparison with 3D-CNN
- Add XAI layers (TRAC-IN already implemented)



Graph pre-processing

Input dataset

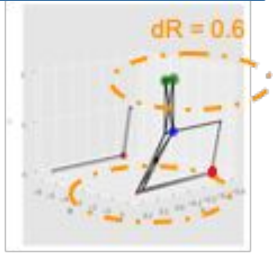
3x Graph Conv Layers

Output score

Model-0 • No preprocessing → *same as reference CNN from papaer*

Model-1 • Optimised delta-R layers → *we find DR = 0.6(within calo-layer), 0.3(intra calo-layer) is best based on metrics (accuracy and purity)*

Model-2 • Optimised number of nodes/subgraphs → *we find that remove isolated nodes (1 or 2) and subgraphs not connected to the core graph makes most sense from a physics perspective looking at proponents and opponents from Trac-In*



- Performance comparison between GNN model-0/1/2 and original CNN
- xAI TracIn and saliency maps available for each model to motivate the evolution of the preprocessing and the increase in performance