MUCCA

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

Joseph Carmignani, Monica D'Onofrio, Cristiano Sebastiani

The MUCCA Project

Ultimate Goal: *quantifying strengths* and solving weaknesses of new and state of the art xAI methods

different data, learning tasks, scientific questions

Strategy: study xAI in *heterogeneous use cases* from High Energy Physics (HEP), medical imaging, diagnosis of pulmonary, tracheal and nasal disease, neuroscience

Collaboration: that brings together researchers from different fields

- High Energy Physics
- Applied Physics to Medicine
- Neuroscience
- Computer science

Multidisciplinarity

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

CHIST-ERA 2022 1st Prize winning Video:
http://widgixeu-responseuploads.s3.amazonaws.com/fileuploads/90010018/90420529/140-2f7727dc2c9debd53715877981b60bce_MUCCA_videO_720p.mp4

The Consortium

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

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

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

Medlea S.r.l.s (IT)

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

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

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

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

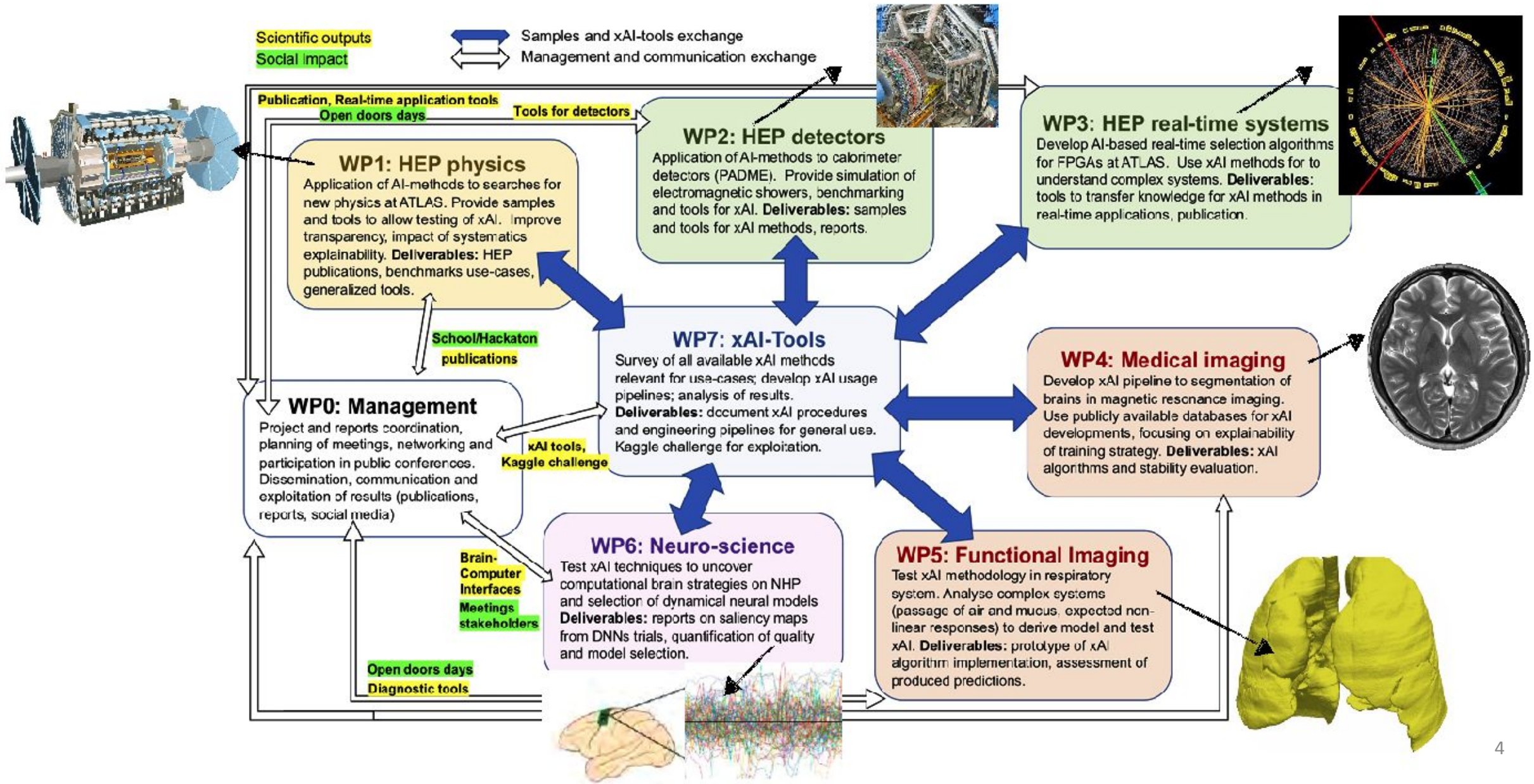
University of Liverpool (UK)
Department of Physics

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

Istituto Superiore di Sanità (IT)

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

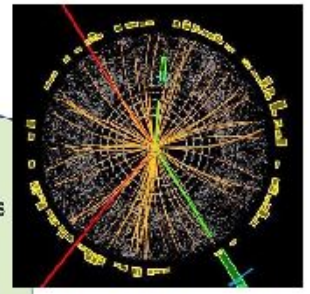
The Work Plan



High-Energy physics work-packages

Scientific outputs
Social impact

Samples and xAI-tools exchange
Management and communication exchange



Publication, Real-time application tools
Open doors days

Tools for detectors

WP1: HEP physics
Application of AI-methods to searches for new physics at ATLAS. Provide samples and tools to allow testing of xAI. Improve transparency, impact of systematics explainability. **Deliverables:** HEP publications, benchmarks use-cases, generalized tools.

WP2: HEP detectors
Application of AI-methods to calorimeter detectors (PADME). Provide simulation of electromagnetic showers, benchmarking and tools for xAI. **Deliverables:** samples and tools for xAI methods, reports.

WP3: HEP real-time systems
Develop AI-based real-time selection algorithms for FPGAs at ATLAS. Use xAI methods for to understand complex systems. **Deliverables:** tools to transfer knowledge for xAI methods in real-time applications, publication.

School/Hackaton
publications

WP7: xAI-Tools
Survey of all available xAI methods relevant for use-cases; develop xAI usage pipelines; analysis of results.
Deliverables: document xAI procedures and engineering pipelines for general use. Kaggle challenge for exploitation.

WP4: Medical imaging
Develop xAI pipeline to segmentation of brains in magnetic resonance imaging. Use publicly available databases for xAI developments, focusing on explainability of training strategy. **Deliverables:** xAI algorithms and stability evaluation.

WP0: Management
Project and reports coordination, planning of meetings, networking and participation in public conferences. Dissemination, communication and exploitation of results (publications, reports, social media)

xAI tools,
Kaggle challenge

Brain-Computer Interfaces
Meetings stakeholders

WP6: Neuro-science
Test xAI techniques to uncover computational brain strategies on NHP and selection of dynamical neural models
Deliverables: reports on saliency maps from DNNs trials, quantification of quality and model selection.

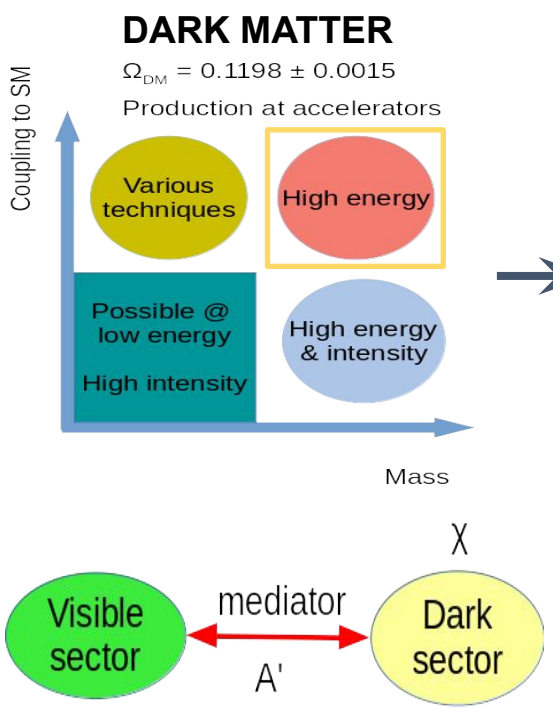
WP5: Functional Imaging
Test xAI methodology in respiratory system. Analyse complex systems (passage of air and mucus, expected non-linear responses) to derive model and test xAI. **Deliverables:** prototype of xAI algorithm implementation, assessment of produced predictions.

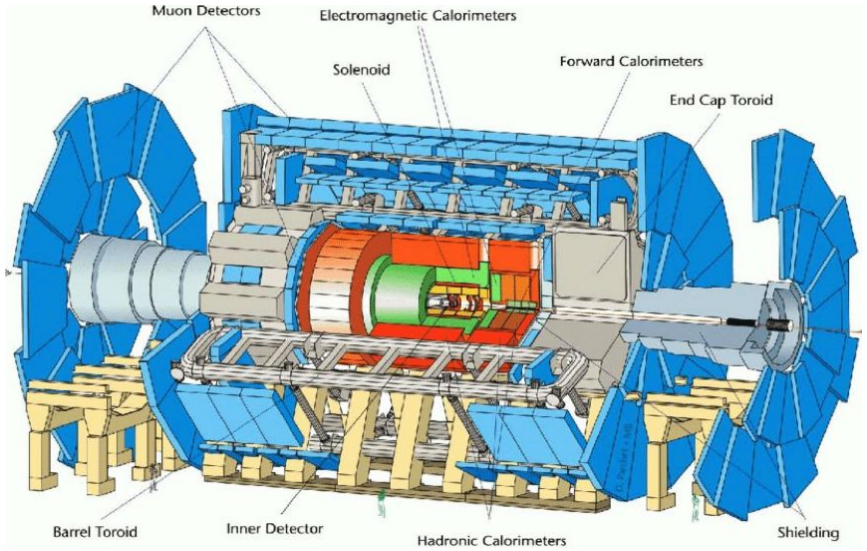
Open doors days
Diagnostic tools

HEP Uses Case 1, 2: offline data analysis

WP1 (HEP Case 1) M. D'Onofrio (PI), J. Carmignani, C. Sebastiani:
application and development of AI algorithms (CNN, Graph NN), targeted to event classification and process discrimination, for new physics (dark sector) and dark matter searches at the ATLAS experiment at the CERN LHC

Weakly interactive massive particles or dark sectors (> 100 MeV dark matter)

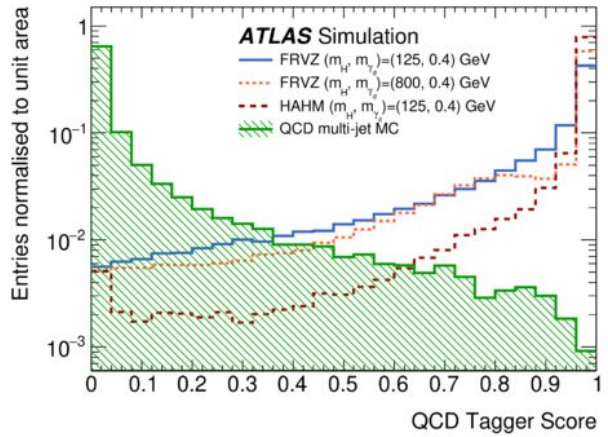




The ATLAS experiment

Information from detector used to build NN and extract potential signals

https://inspirehep.net/literature/2100410



Liverpool leads WP1, more details in the following of this talk

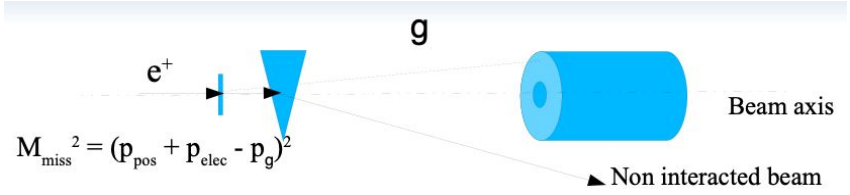
HEP Uses Case 1, 2: offline data analysis

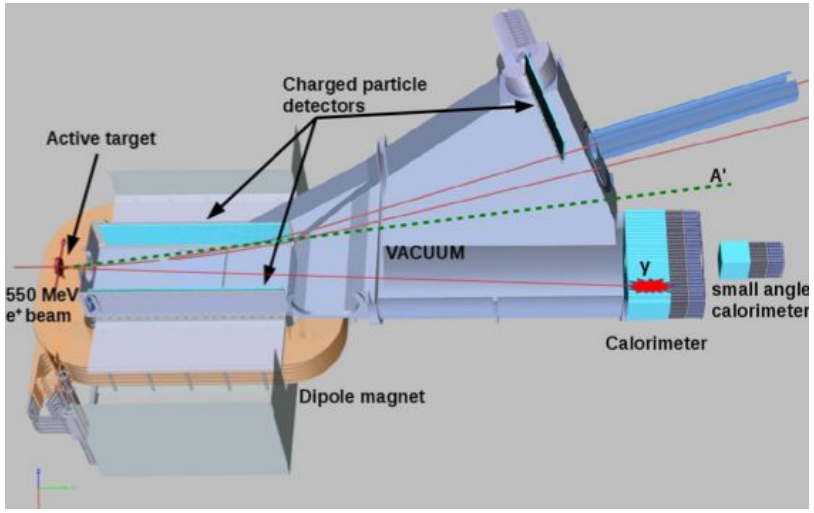
WP2 (HEP Case 2): *AI algorithms* (CNN, autoencoder) for identification of pulses and other specific parameters of detector responses, in particular of the calorimeter (ECAL)

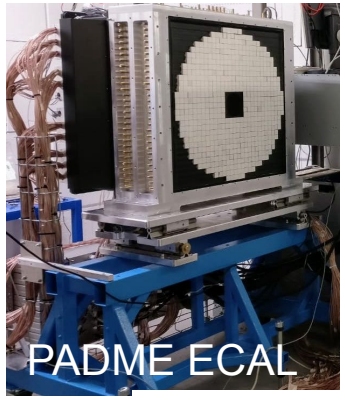
Use PADME apparatus: PADME is a fixed target experiment located at the Beam Test Facility at the Laboratori Nazionali di Frascati designed to search for a massive dark photon in the process , using a positron beam of energy up to 550 MeV.

Dark matter targeted: dark sector (low masses)

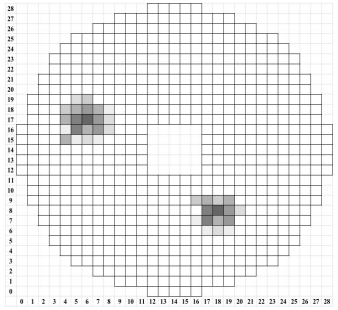
Positron **A**nnihilation into **D**ark **M**atter **E**xperiment





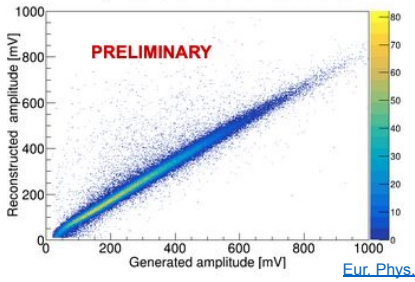


Two photon showers in the ECAL (bkg)

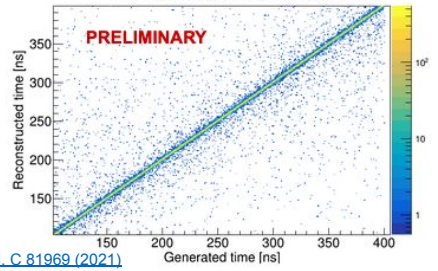


AI algorithms successfully developed and applied to identify pulses, determine amplitude and time of arrival for signal (1 g) and bkg (2 g) events in simulated data of the ECAL

Amplitude reconstruction

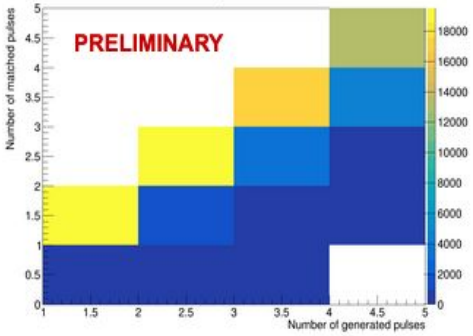


Time reconstruction



Eur. Phys. J. C 81969 (2021)

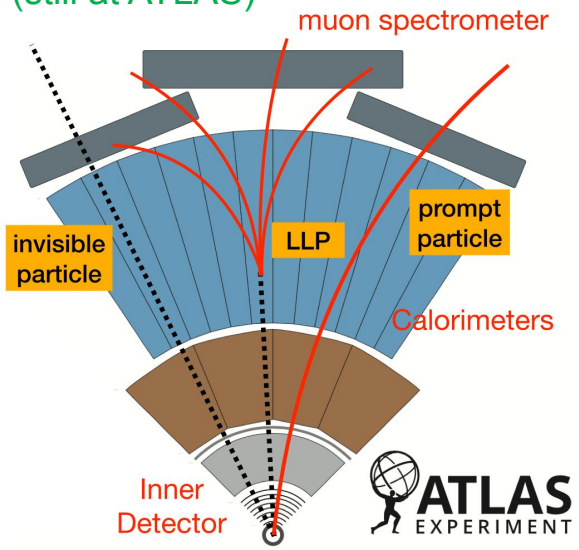
Pulse separation



HEP Uses Case 3: real time analysis

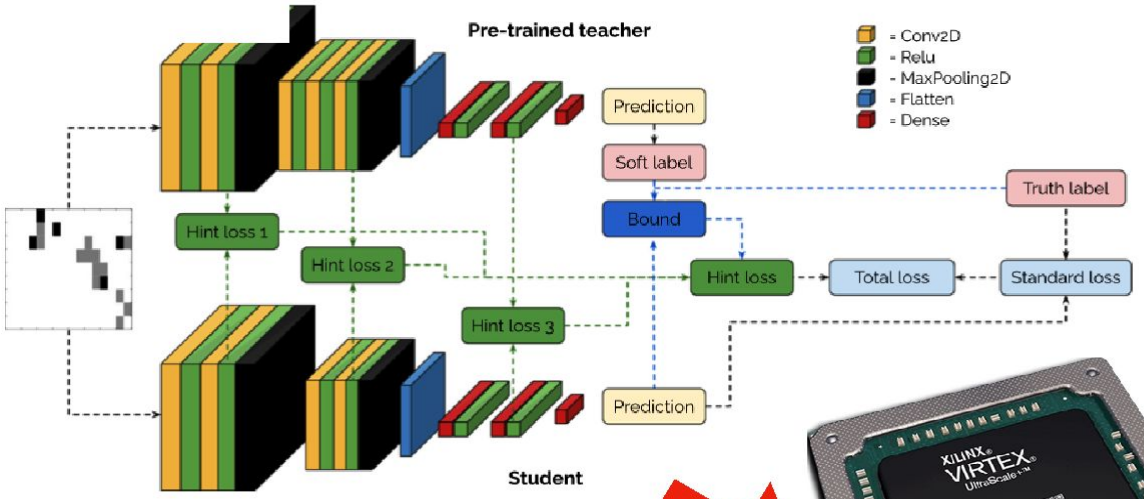
WP3 (HEP Case 3): focus on the ability of reconstructing in **real time** unconventional signatures e.g. from dark sector particles that traverse the detector before decaying (long-lived particle or LLP) as those searched for in WP1. *Status: developed complete pipeline for an AI based event selection* algorithm. CNN model with compression and simplification strategies to make easier to interpret, and faster to execute the AI model, for the conversion and implementation in the firmware of FPGA accelerators. Obtained CNN inference in 80/150ns/image

Signature of interest
(still at ATLAS)



Must suppress not-interesting events that is several orders of magnitude larger than the LLP signal

One of the outcomes: transfer knowledge learned by a larger neural network pre-trained for the same task to a smaller and quantised (4-bits per activations and weights) model



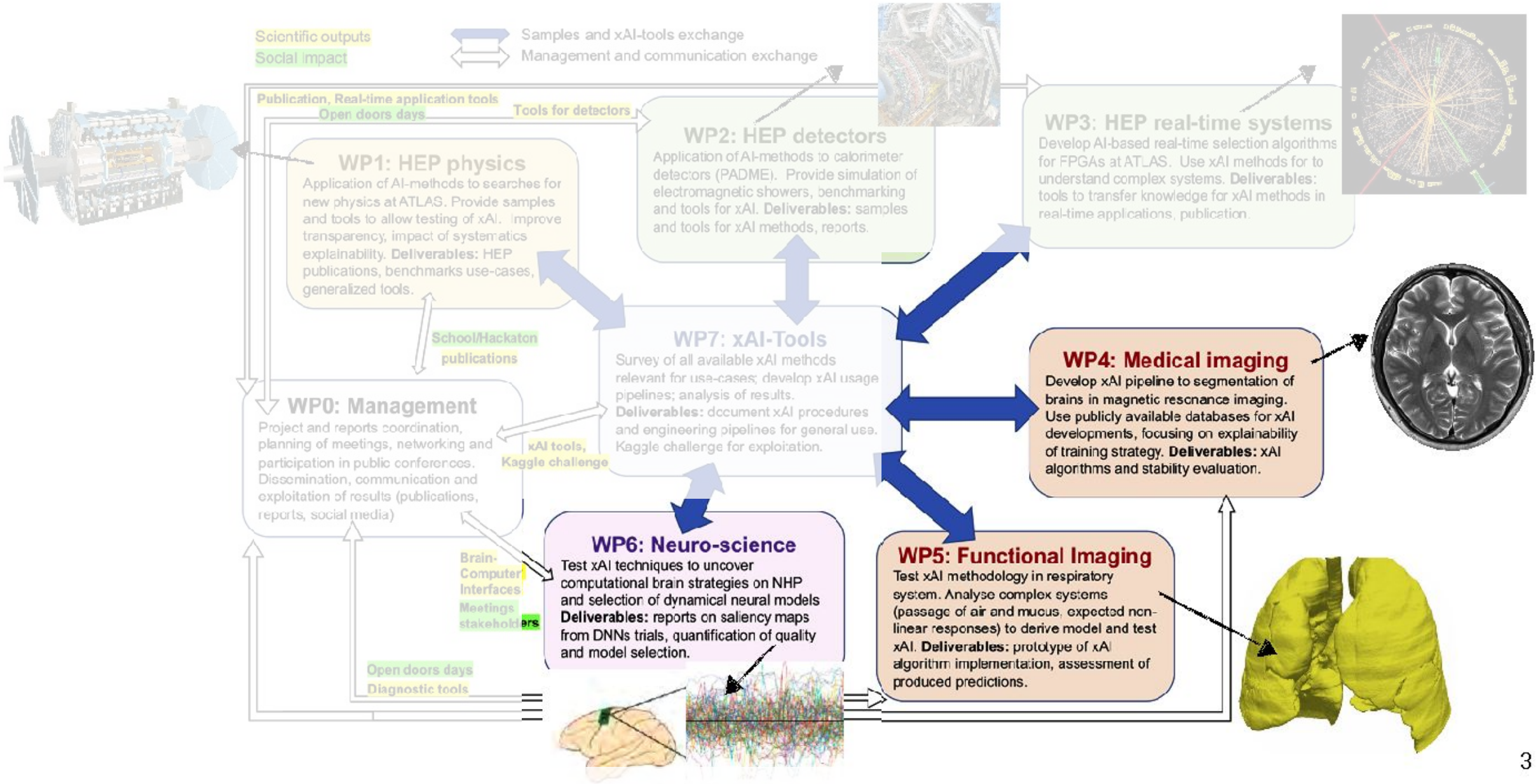
model compression

FPGA

obtained a reduction on size of the model of a factor 100 with only a limited reduction in performance

Eur. Phys. J. C 81969 (2021)

Health-care (medical-based) work packages



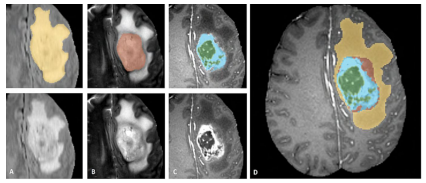
Brain in magnetic resonance imaging

WP4 (MED-1): aims to develop a pipeline to provide xAI tools suitable for medical applications, proven to work in a specific task, the segmentation of the brain in magnetic resonance imaging (MRI).

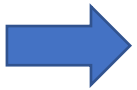
Status: Implemented *AI models for the brain lesion segmentation* in the Brats17 MRI dataset (Unet2D, Resnet 3D). Data augmentation techniques to enhance performances tested.
Selected state-of the art xAI algorithms, several under implementation

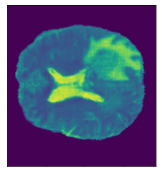
UNet pipeline

Brats17



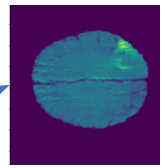
state-of-the-art xAI

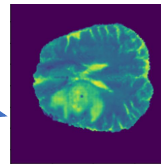




training influence (gradient tracing)

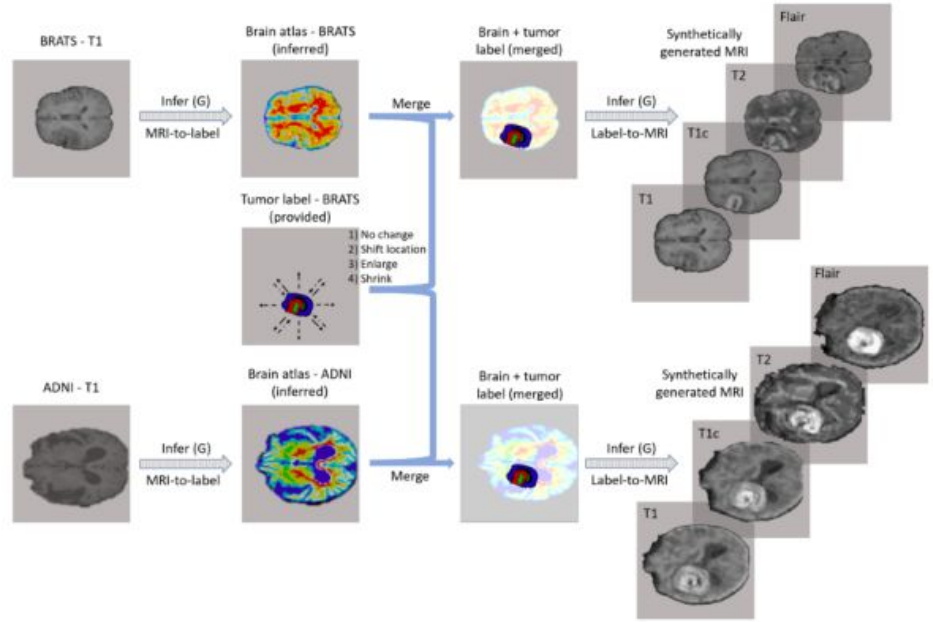
"far" from





saliency maps

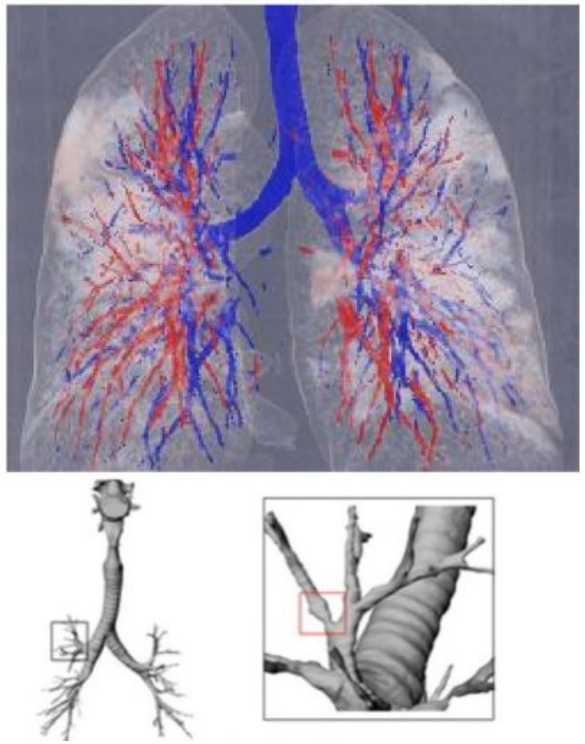
cGAN pipeline

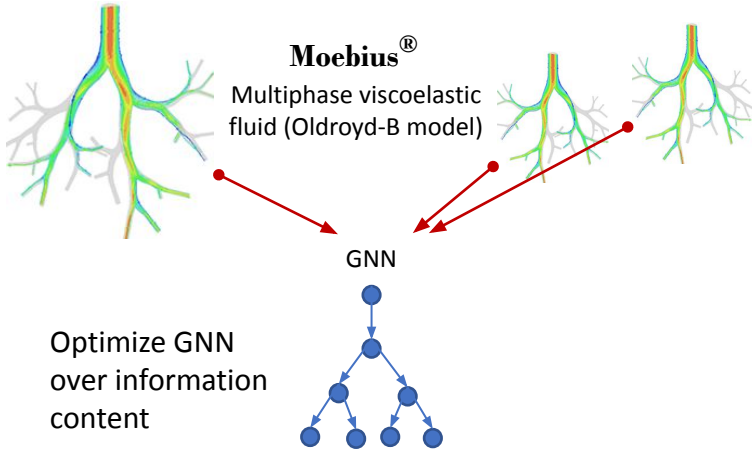


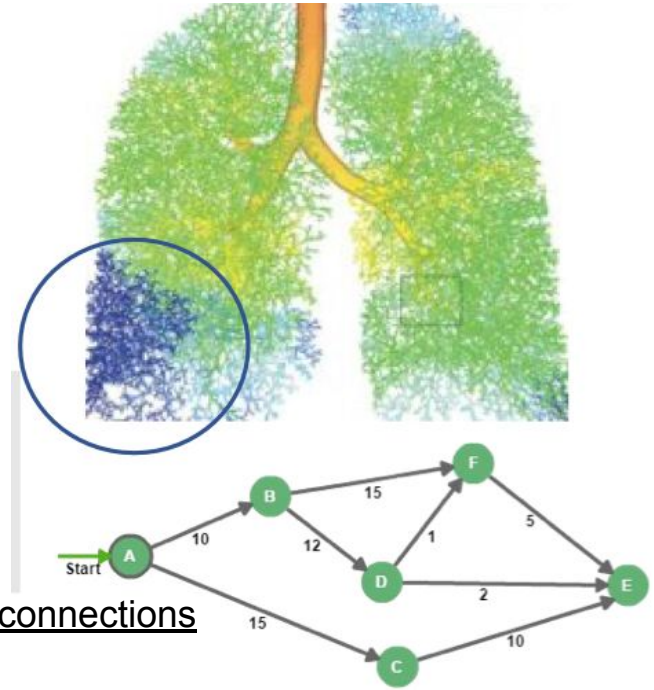
Functional imaging for respiratory system

WP5 (MED-2): Aim to develop an integrated approach for 3D reconstruction from medical images to perform simulation & experiments on respiratory tracts (airways) and assess airflow and air+mucus dynamics in respiratory tracts. Validation of simulation results with idealized and real data from patients, reaching a high level of automation to handle several geometries (patients)

Status: procedure for the realization of the prototypes of the trachea bifurcation (reconstruction of the geometry from the CT scan, numerical code) completed. Study of the GNN model for the simulation of the air-flow and explainability steps in progress







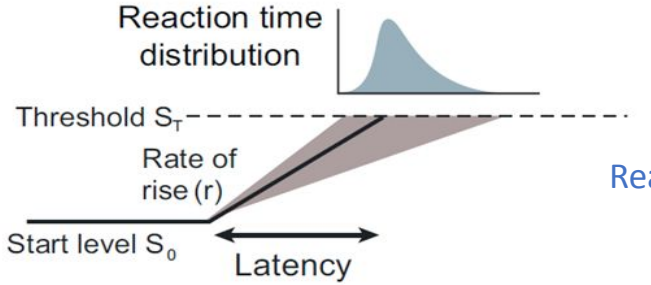
NN connections

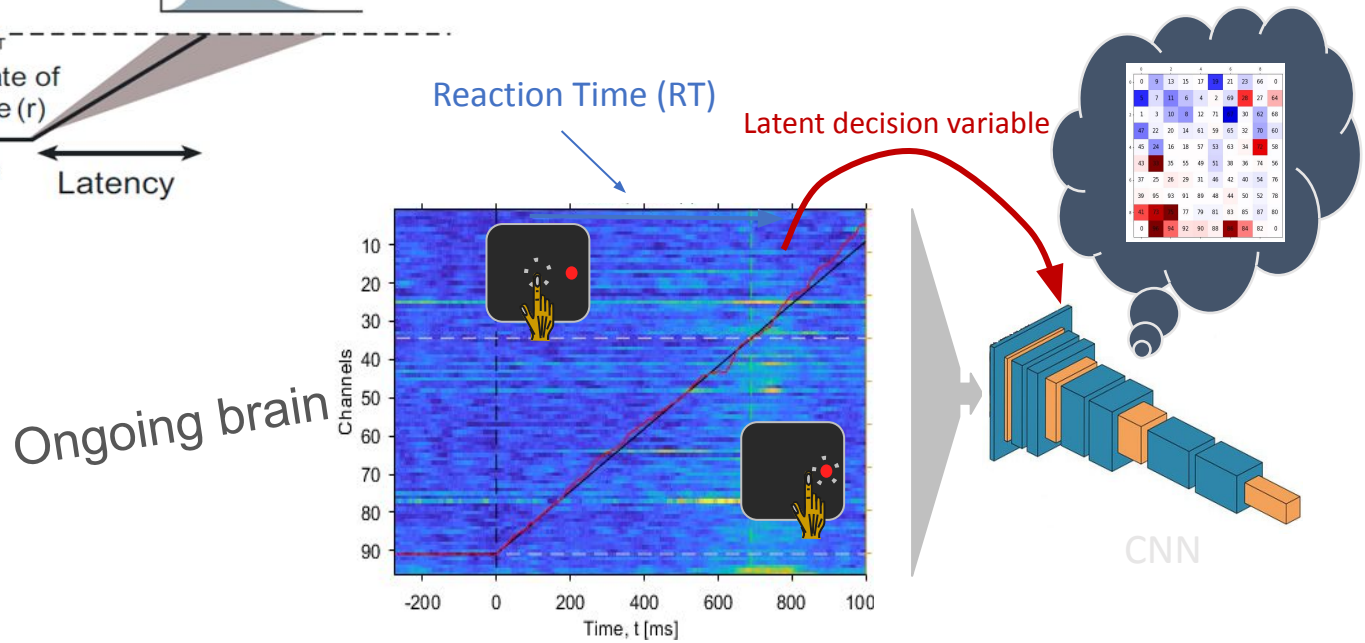
xAI: Understand system outputs simply and quickly (e.g. why certain regions are not ventilated). Give confidence in the AI diagnosis by providing clarity of why.

Neuroscience use-case

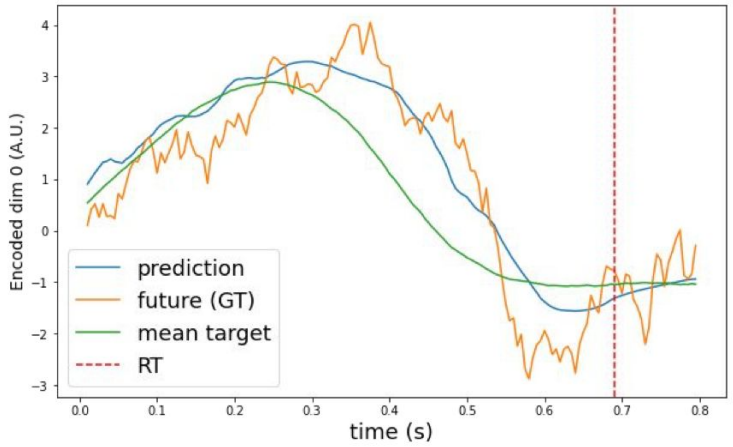
WP6 (MED-3, NS1): A twofold exploitation of xAI techniques, to: 1) uncover computational brain strategies while non human primates (NHP) perform tasks requiring the inhibition of planned movements; 2) optimally select dynamical neural models that will be developed to explain the observed task-related cortical dynamics.

Status: designed and realized a specific CNN (fed by electrophysiological signals) based on a ResNet to uncover an inner decision value increasing in time as a linear ramp eventually allowing to predict at single-trial level the onset timing of overt movements. Test of various xAI algorithms underway

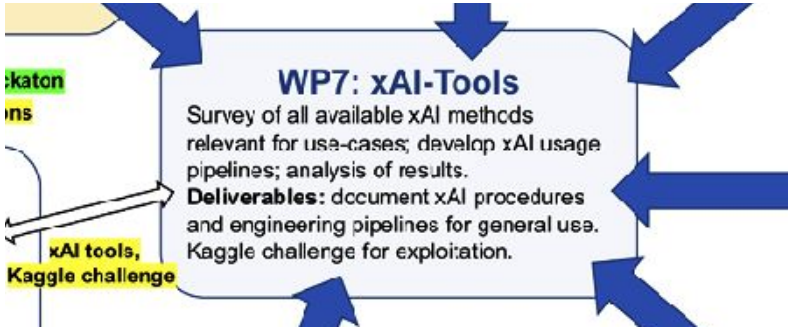




Good prediction and good understanding of source of errors and how the NN learns..



WP7: xAI Tools



WP7 liaises with all other work-packages in iterative mode:

1. **Survey of xAI methods** (done)

- XAI tools can be categorized depending on whether they provide **global** or **local** explanations.
- Some methods are **model-agnostic** (they only need the outputs of the models), other are **model-specific**.
- Finally, methods can be categorized depending on what type of information they provide in output.

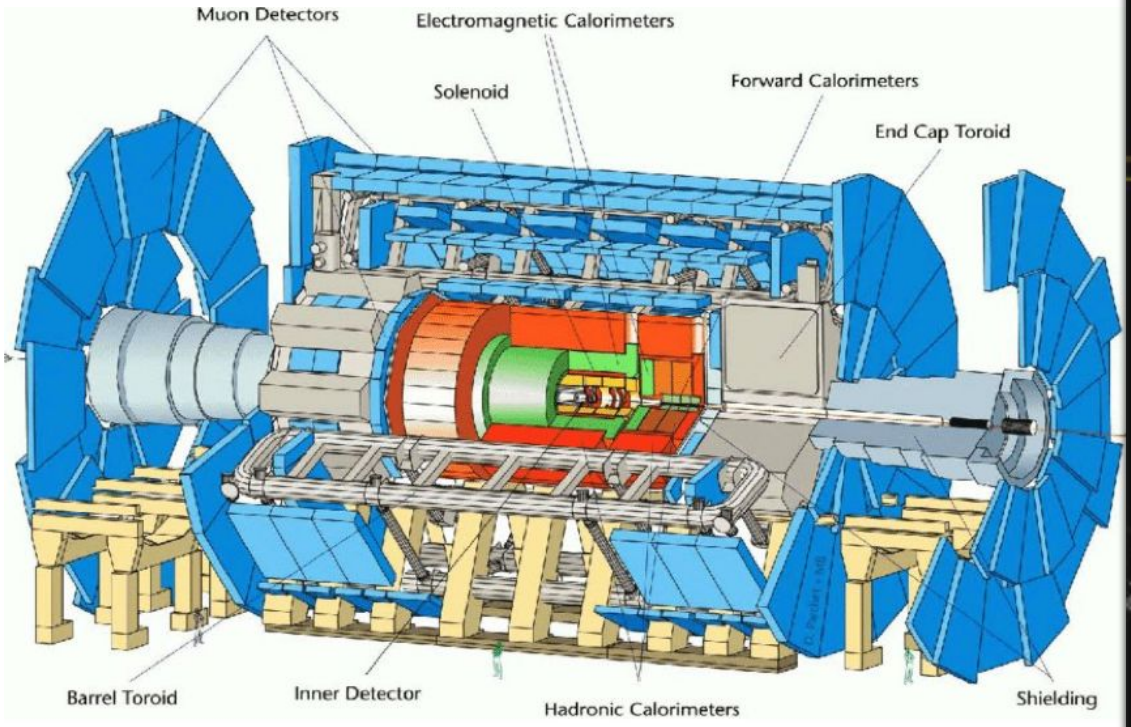
2. **Tools delivery to use-cases** (in progress)

E.g. for WP1, iterating on usage of Tools such as Captum to understand the way NN learns from high-level and low-level inputs in data analysis, but also hands—on tutorial and school/hackathon to be organised in 2023.

3. **Engineering pipelines for general xAI applications**
ultimate goal of the consortium for all use cases

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

A closer look to WP1



Two benchmark analyses:
Searches for Supersymmetry
Searches for dark photons

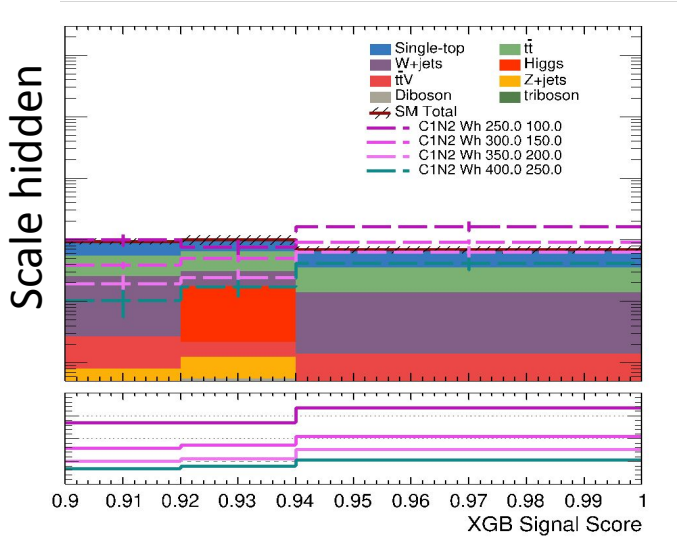
Searches for Supersymmetric particles

Benchmark-1: Search for dark matter candidates resulting from the decay of new particles predicted by Supersymmetry – the typical HEP case:

- Extract small signal of interest from large SM background
- Subtle/complex differences in variable correlations distinguish signal from background
- Complex numerical instance data, well-defined categories (underlying physics processes)

→This is the classic use-case for ML classification.

Build ML discriminator to distinguish backgrounds from SUSY signals, trained on simulated Monte Carlo samples, use classifier output score as discriminant variable for hypothesis testing



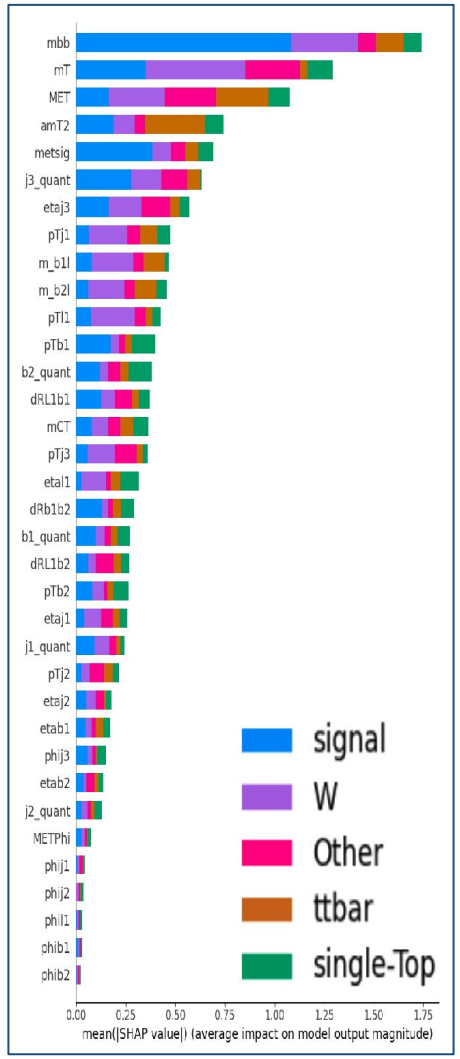
Tested multiple ML classifiers: BDT, NN. Use BDT (**XGBoost**) for reduced complexity, constructed from regression tree functions, using multi-classification with output scores containing the predicted probability of an event being in each class.

So far, used SHAP (SHapley Adaptive exPlanations, 2017) to identify variables with largest impact for signal

In progress:

- finalisation of the data analysis
- Building eXplainable Graph data, Use GNNs and New SOTA Metrics

Goal: Reduce dependencies on modelling from input variables



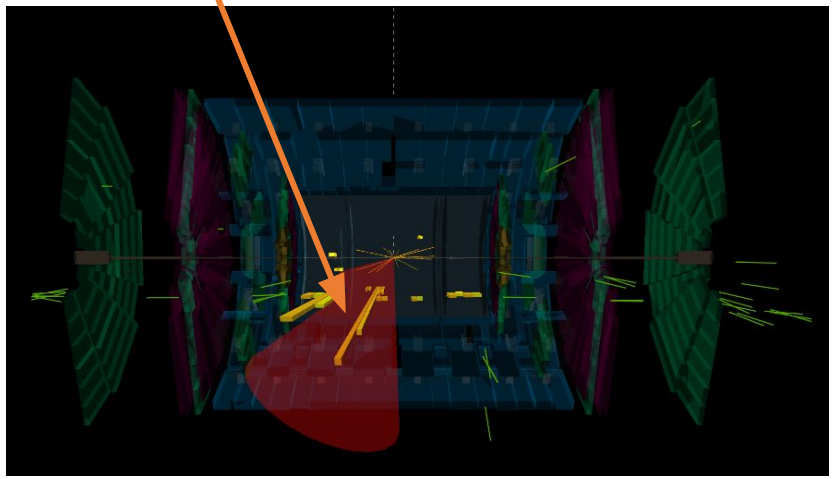
Searches for dark photons

Benchmark-2: Search for “dark” photons, light particles belonging to a new hidden sector not yet discovered because too feebly interacting with ordinary matter:

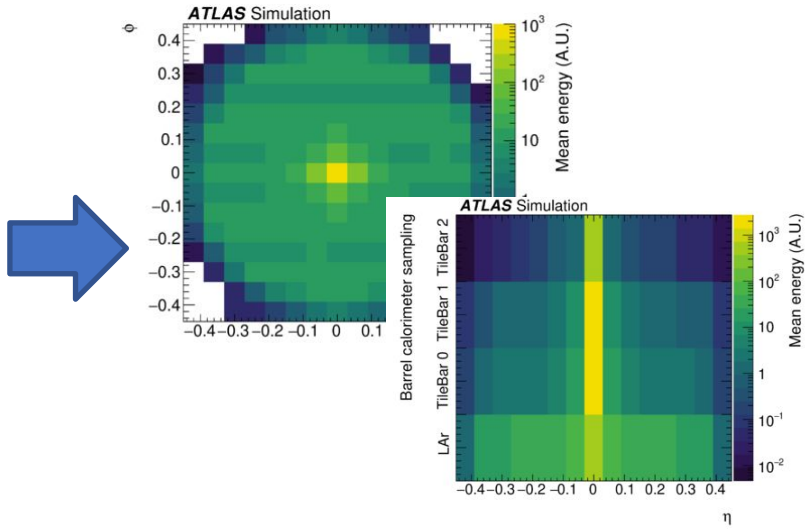
- In this case, signal leaves different signature in the detector wrt background
  - signal signature is effectively an unknown – study of systematics on the signal is non-trivial
- ML discriminator use image classification trained to distinguish background processes from signal mapping clusters of hadrons (jets) in 3D coordinates

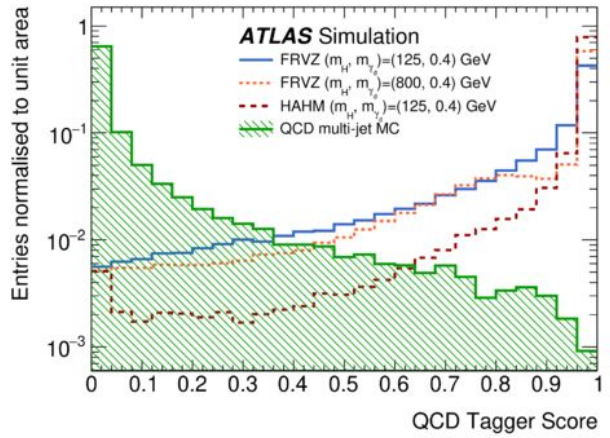
More advanced, also because publication has been finalised

ATLAS calorimeter system



In the ATLAS data-analysis:
Build a map of jet energy deposits in ATLAS detector from: calorimeter cell positions (eta, phi, sampling layer) and energy

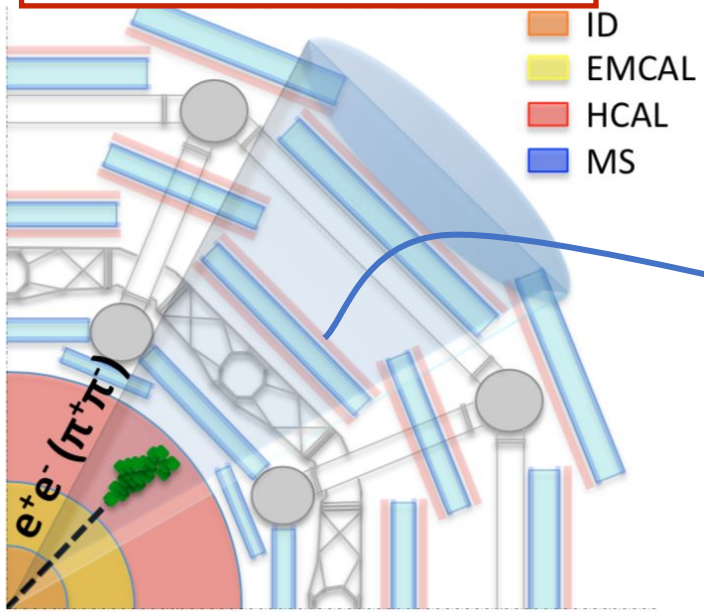


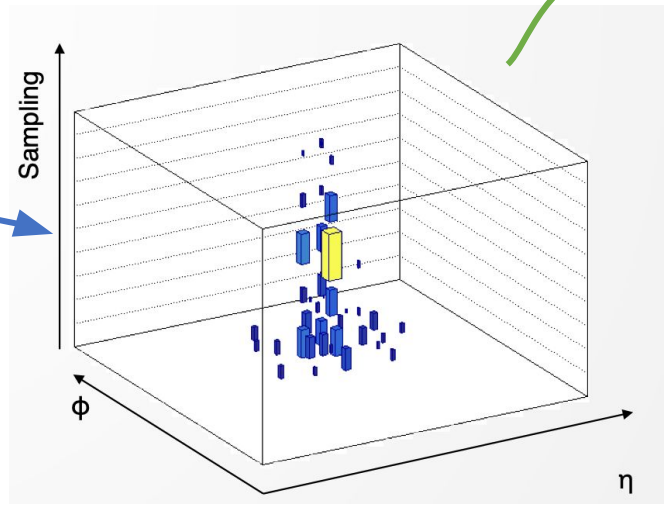


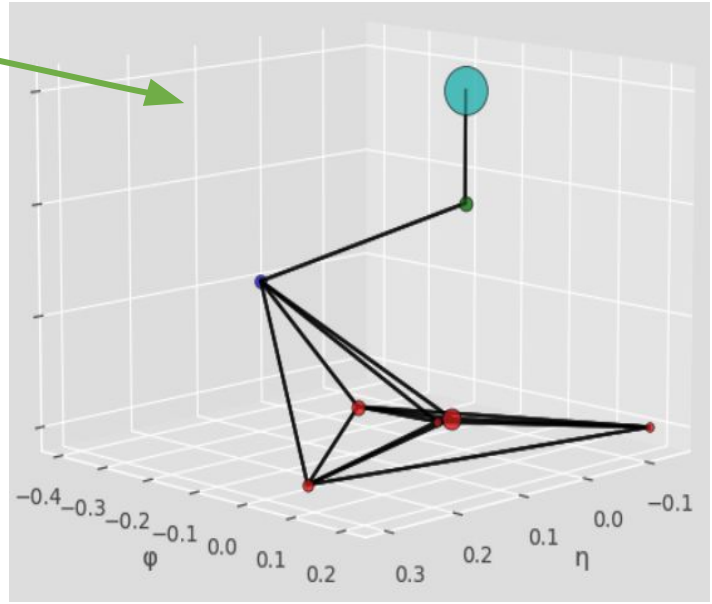
Low level inputs for jet discrimination

Extract low level information from the ATLAS calorimeter from a single jet in either 3D images or graphs

The ATLAS detector orthogonal view







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

3D jet images:
- Train a CNN used as reference for the study
- Very sparse images -> sub-optimal

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

Additional higher level variable can 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

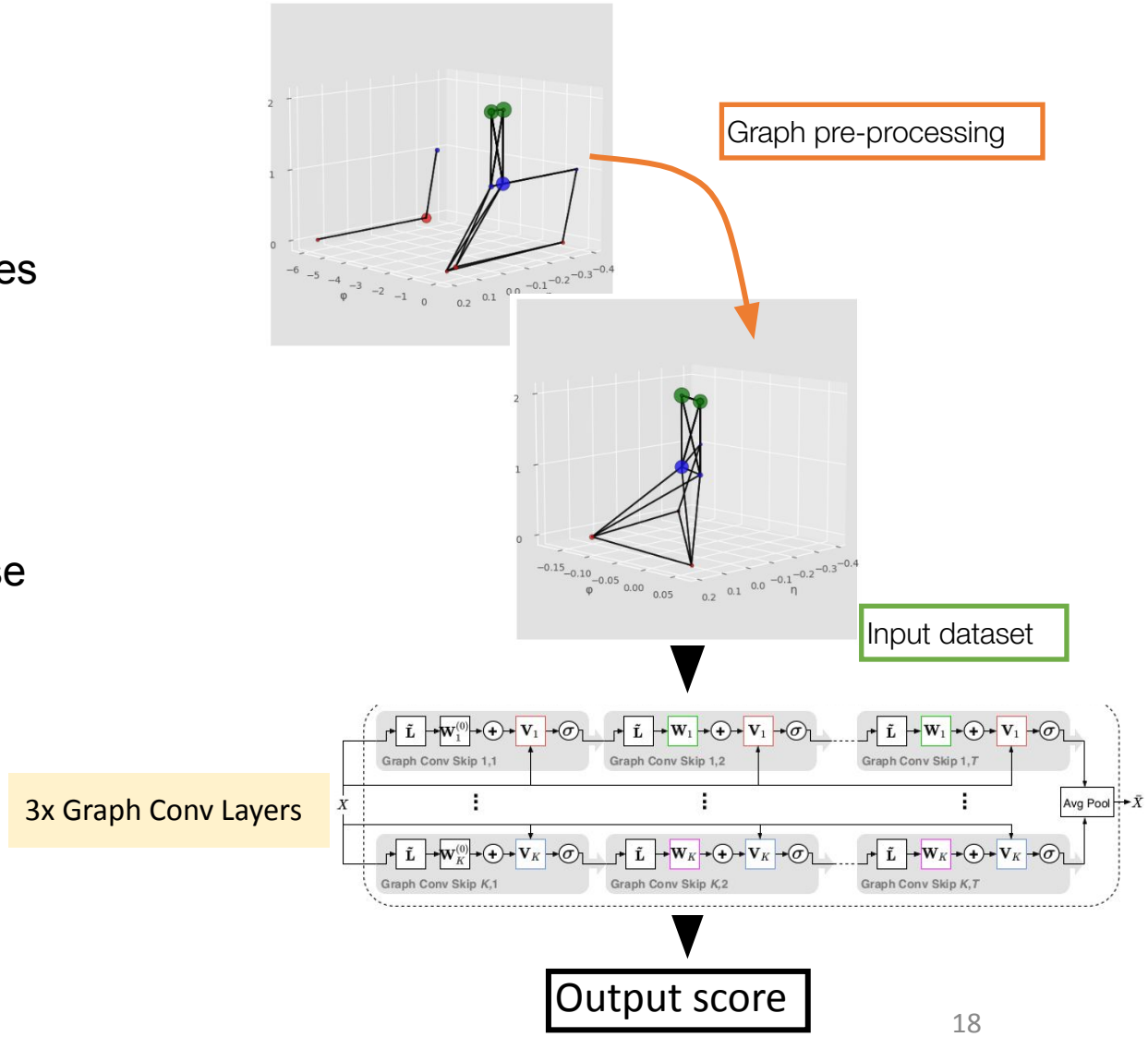
Optimisation and training

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

Dataset building:
- Node for every cluster in the calorimeter
- Normalised cluster energy as node attribute
- Edge built if spatial covariant distance between two nodes is within an optimised distance parameter
- Covariant distance as node weight

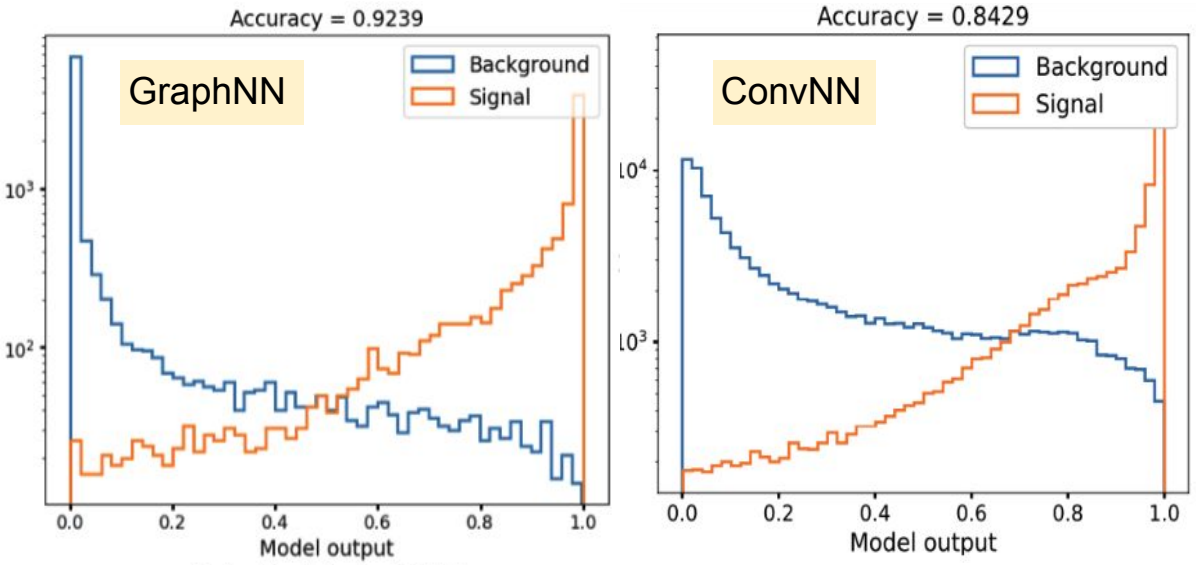
Pre-processing:
- Remove isolated and self-connected nodes
- Retain largest subgraph only to remove calorimeter noise

Model optimisation and xAI:
- Test multiple models with Pytorch Geometric libraries
- Performance evaluation and comparison with CNN
- Add xAI layers



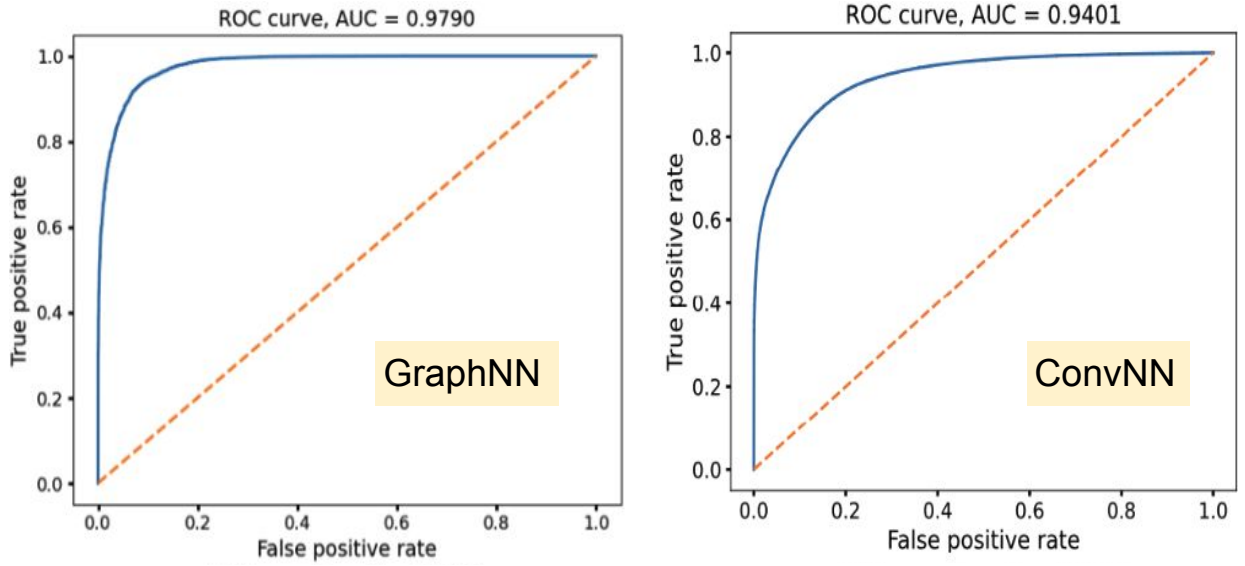
A quick model comparison…

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



Classification accuracy

Area Under ROC Curve

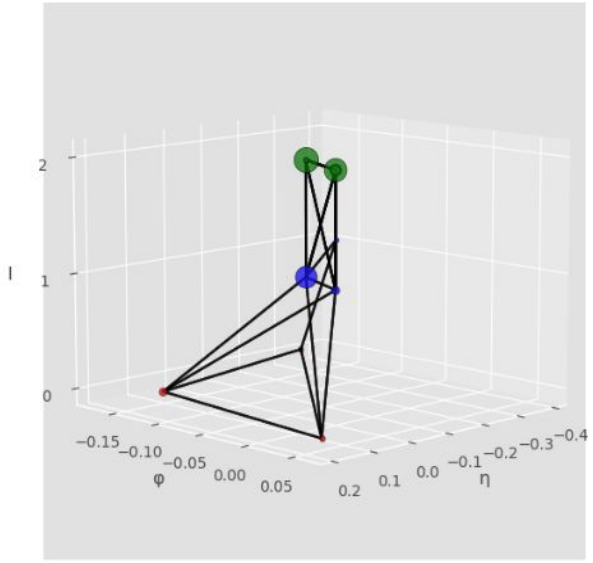


As expected graph dataset are proven very effective for classification of sparse image of HEP calorimeter detectors!

the X in X-AI

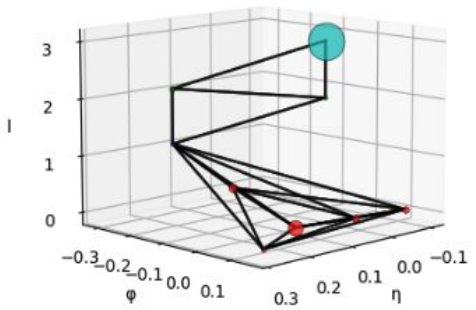
Exploit **explainer** layers to better understand the network prediction, to be implemented by WP7

Let's try estimating training data influence by tracing gradient descent with TracIn!
The model yields dynamic results producing best scoring Proponents/Opponents from training to explain predictions

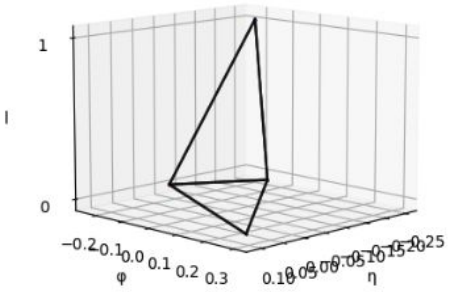


This graph is correctly predicted as signal

Training dataset most influential graphs for this prediction:



signal



background

This class of methods explore the influence of single **data points** on the prediction, e.g., how much training on a certain point has influenced the prediction on a separate point (computed across the entire training run)

Summary

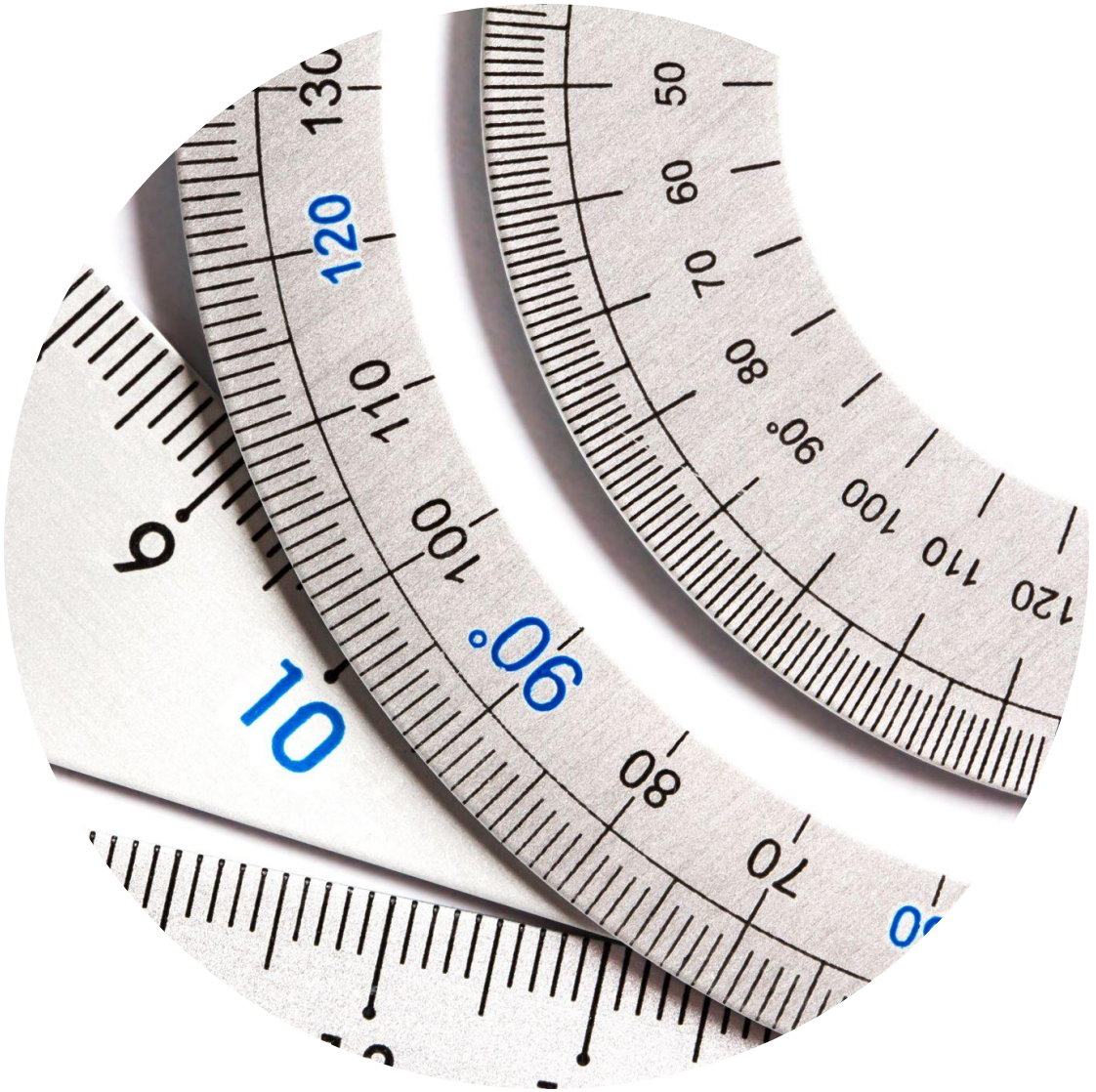
The MUCCA consortium aims to study xAI in *heterogeneous use cases* from High Energy Physics, medical imaging, diagnosis of pulmonary, tracheal and nasal disease and neuroscience, with the ultimate goal to *quantifying strengths* and solving weaknesses of new state of the art methods.

Status so far:

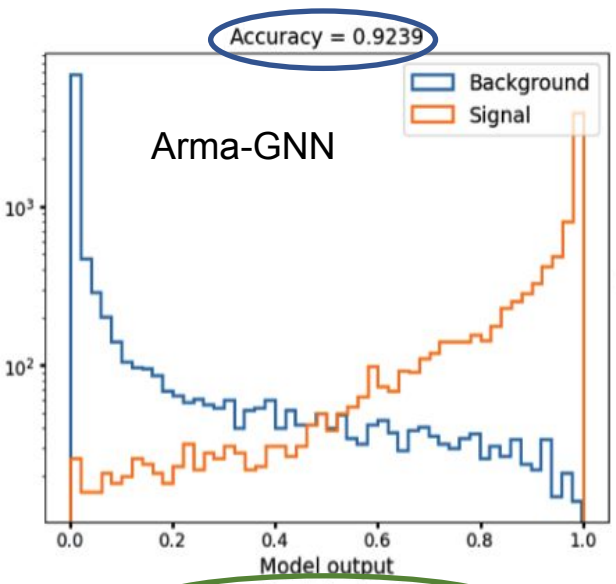
- successfully implement appropriate AI algorithms for all the use cases
  - for **WP1** the ATLAS analyses have been either published or are being finalised. For the dark-photon (published) a GNN that outperforms the current setup has been developed, xAI tools have been used to identify first shortcomings and ways to improve. We are now finalising this study aiming for publication in 2023
- perform an extensive survey and analysis of state-of-the art xAI methods by **WP7**
- identify suitable xAI algorithms for the next phase, that now are under implementation

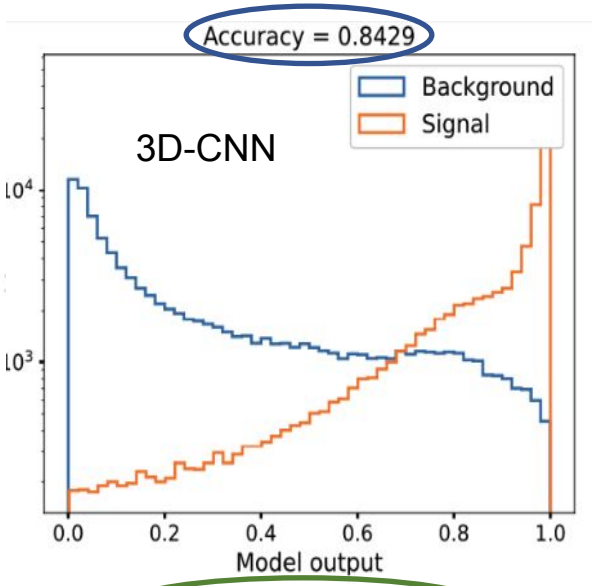
Expected results will allow a systematic understanding of which xAI methods better adapt to specific applications as well as skill development and training for young researchers.

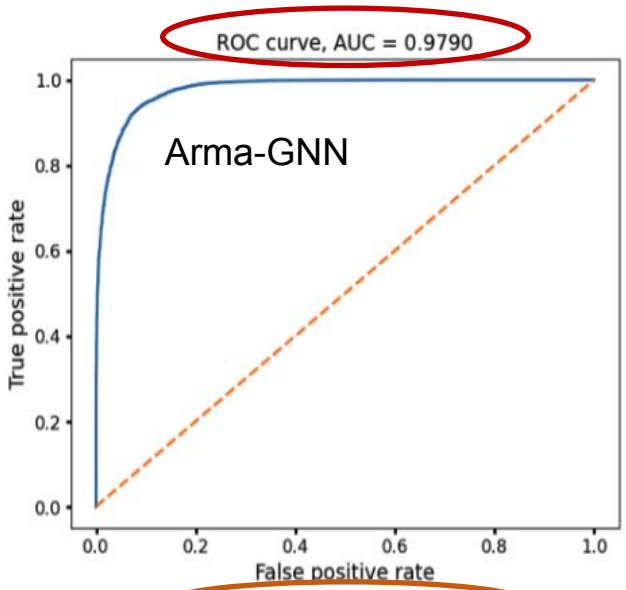
Additional slides

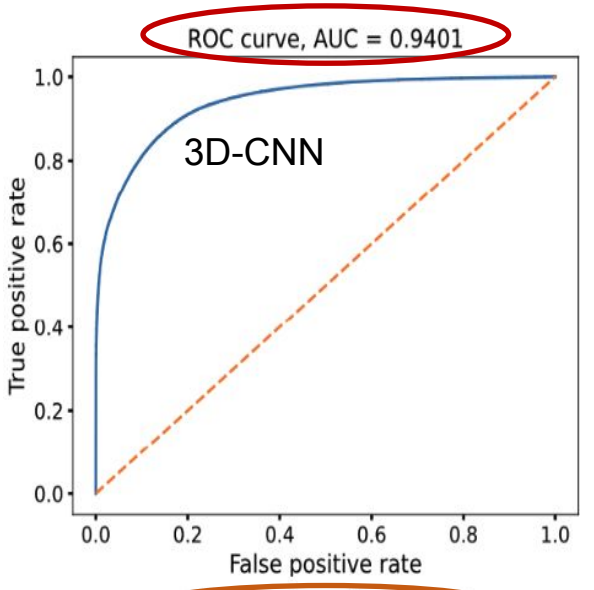


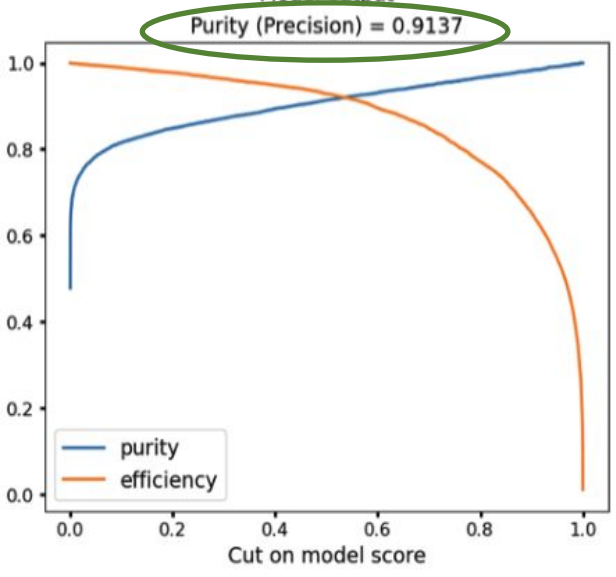
Performance checks

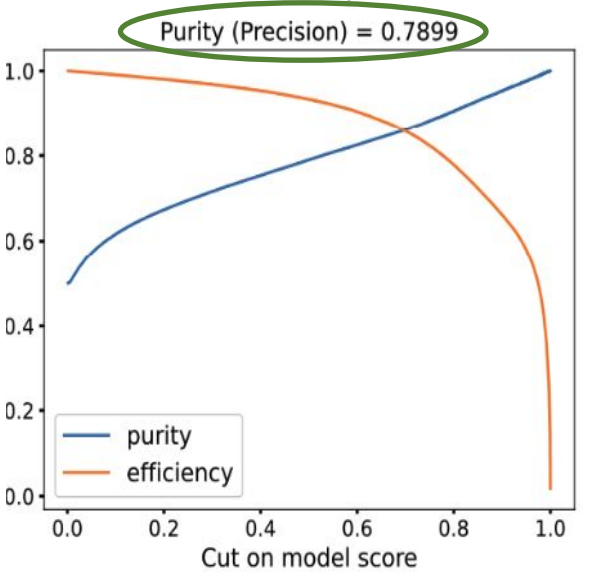


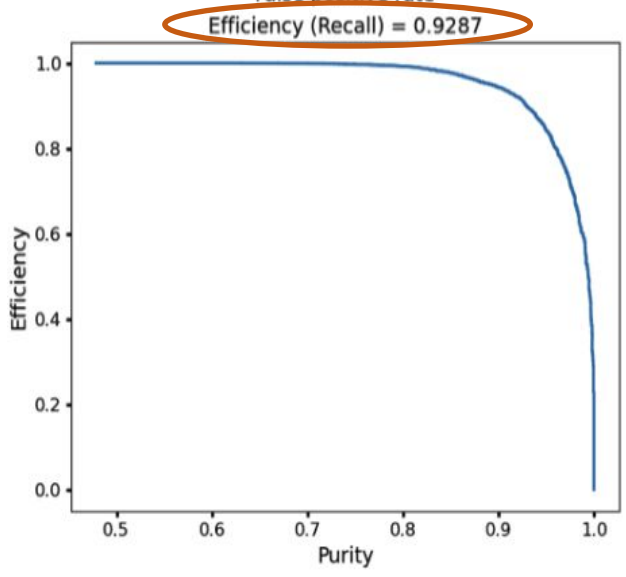


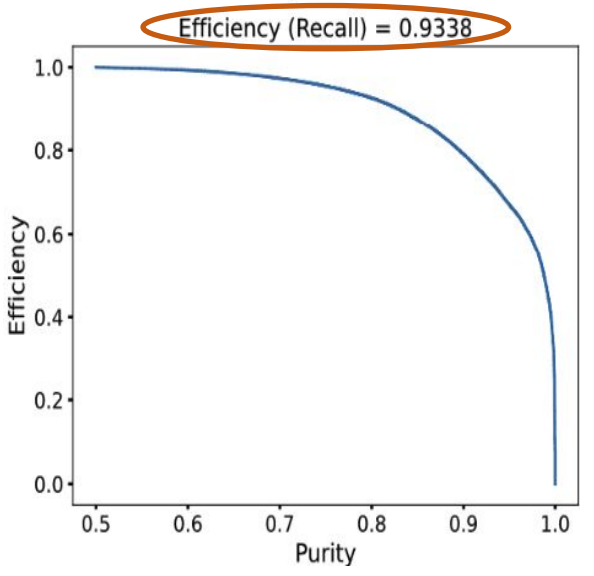










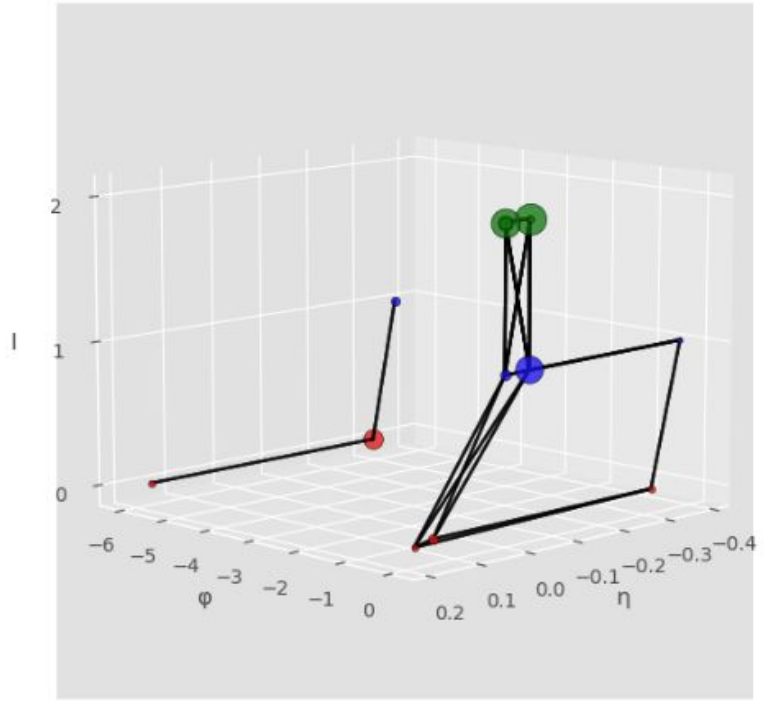


➢ Arma-GNN model using Jet selective Graph Dataset out-performed on all most common **performance metrics** at same Efficiency level


Some Captum Results

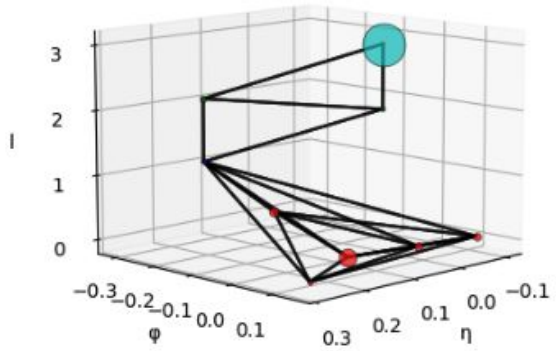
ArmaGNN Jet tagger trained on 0.6_0.6 Full dataset

```
Test example:Data(edge_index=[2, 146], x=[14, 4], edge_attr=[146, 1], y=[1])
True label: 0
Predicted label: 0
Predicted prob: 1.6132928521983558e-06
```



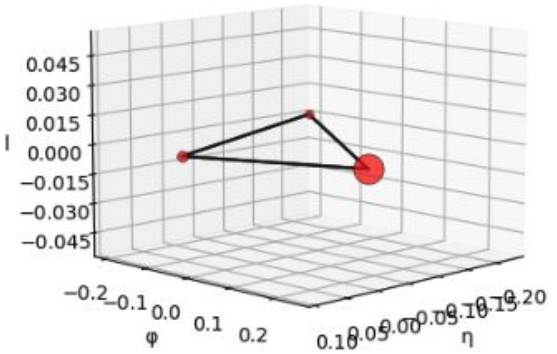
Proponents:

Label:0



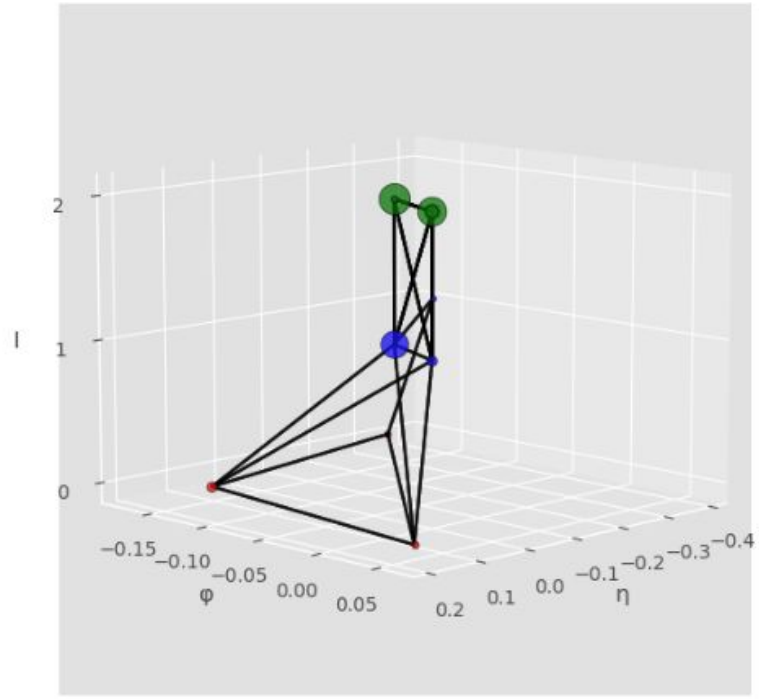
Opponents:

Label:1



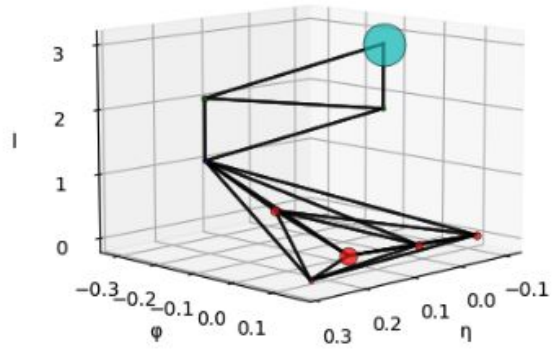
ArmaGNN Jet tagger trained on 0.6_0.6 Main Subgraphs

```
Test example:Data(edge_index=[2, 135], x=[11, 4], edge_attr=[135, 1], y=[1])
True label: 0
Predicted label: 0
Predicted prob: 5.9570318455826055e-08
```



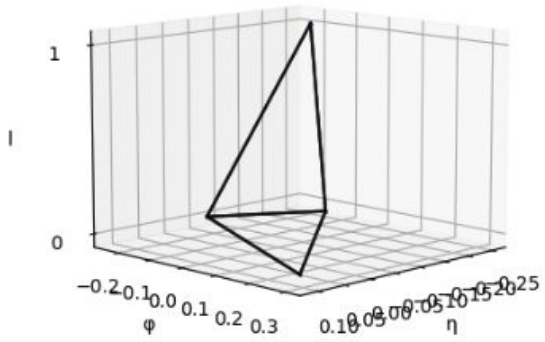
Proponents:

Label:0



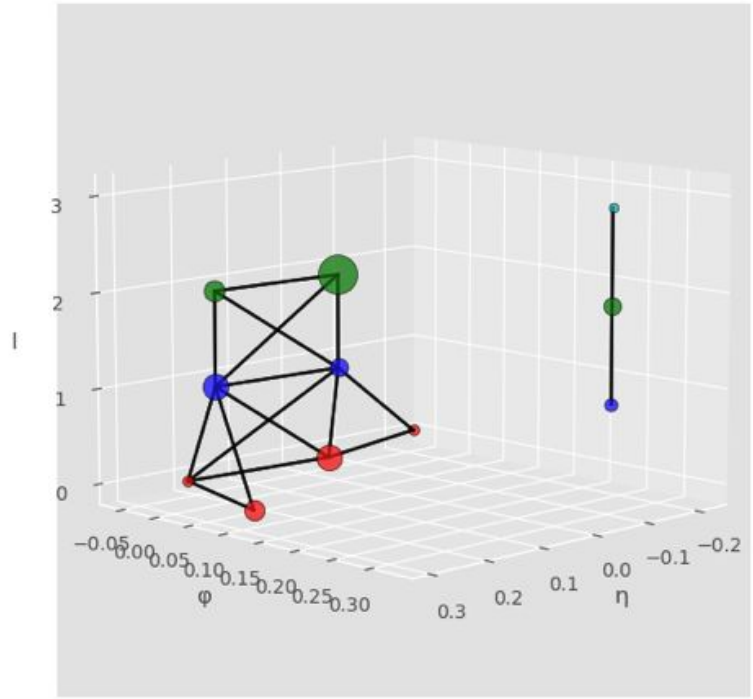
Opponents:

Label:1



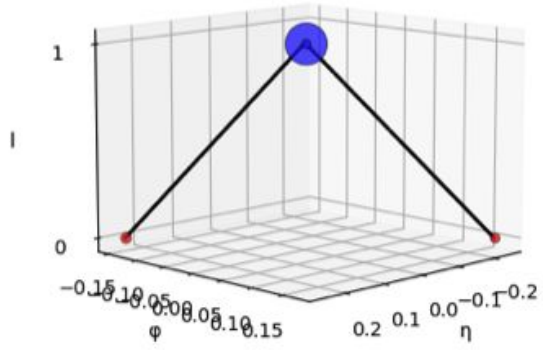
ArmaGNN Jet tagger trained on 0.3_0.3 Full dataset

```
Test example: Data(edge_index=[2, 79], x=[11, 4], edge_attr=[79, 1], y=[1])
True label: 0
Predicted label: 0
Predicted prob: 0.48466432094573975
```



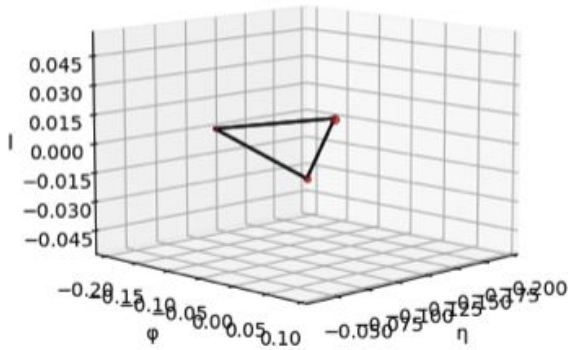
Proponents:

Label:0



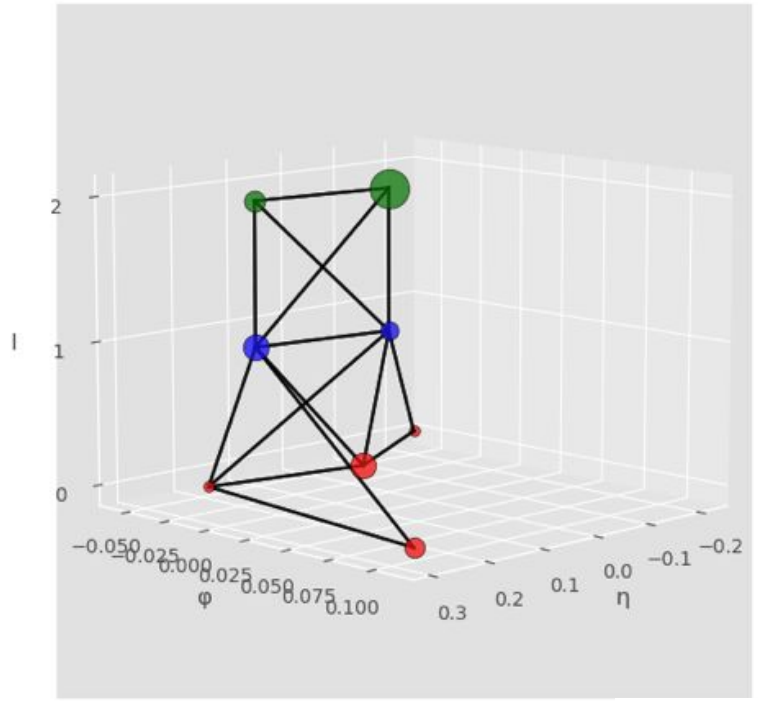
Opponents:

Label:1



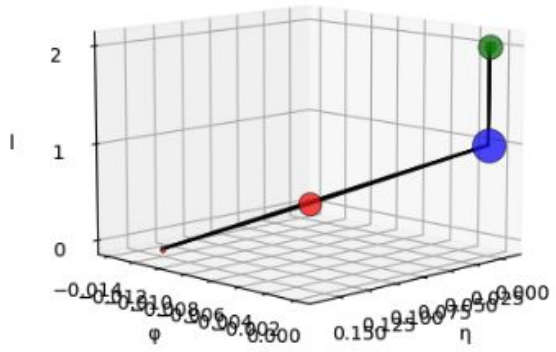
ArmaGNN Jet tagger trained on 0.3_0.3 Main Subgraphs

```
Test example: Data(edge_index=[2, 68], x=[8, 4], edge_attr=[68, 1], y=[1])
True label: 0
Predicted label: 0
Predicted prob: 0.2386256456375122
```



Proponents:

Label:0



Opponents:

Label:1