Advancing Explainable AI

Testing and Enhancing Techniques Across Multidisciplinary Use-Cases

Presenter: Simone Scardapane

Sapienza Università di Roma

Introduction

The **MUCCA** project

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

CHIST-ERA IV xAI H2020 EU grant 2.2021-7.2024

ISTITVTO SVPERIORE DI SANITÀ

UNIVERSITY OF LIVERPOOL

POLITEHNICA UNIVERSITATEA DIN BUCUREȘTI 1818

СВ. КЛИМЕНТ ОХРИДСКИ

MedLea

INFN Istituto Nazionale di Fisica Nucleare

STVDIVM VRBIS

The MUCCA consortium

Istituto Nazionale Fisica Nucleare (IT)
Rome group

Fundamental research with cutting edge technologies and instruments, applications (HEP, medicine)

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.

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

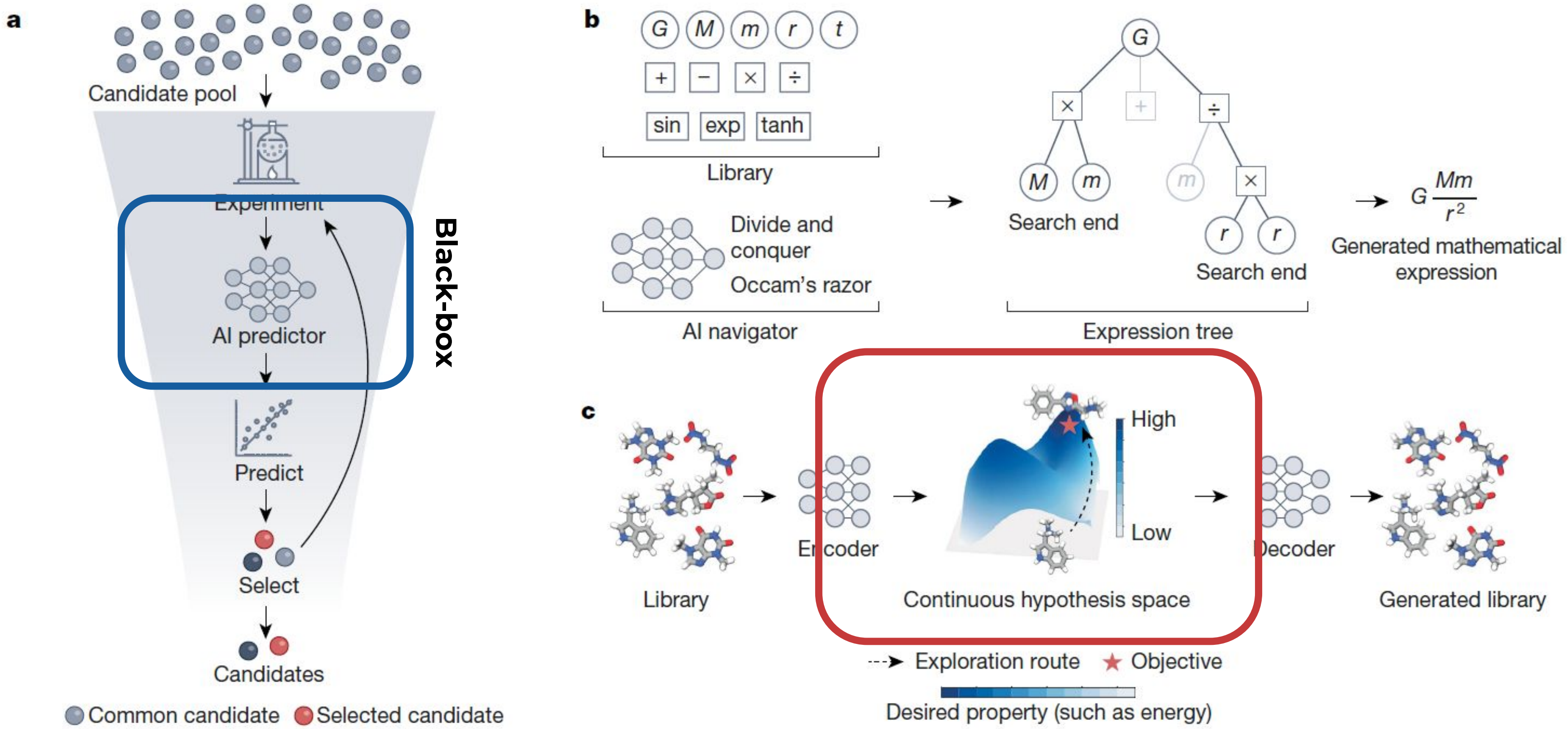
physics data analysis at hadron colliders experiments, simulation, ML and DL methods in HEP

Istituto Superiore di Sanità (IT)

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

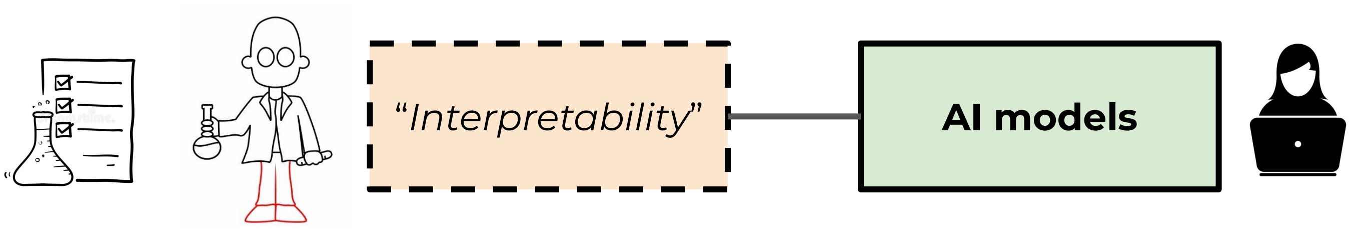
AI for scientific discovery

[Scientific discovery in the age of artificial intelligence | Nature]



Contents

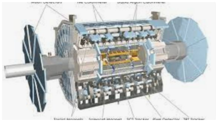
Explainability (xAI) as the potential "*bridge*" between the AI expert and the scientist.



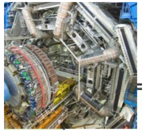
Research questions:

1. How to select a **"good" xAI algorithm**? Which method among hundreds (saliency map, data attribution, ...)?
2. How to combine multiple, potentially contradictory explanations (**convergent explanation**)?
3. How do we "**explain the explanation**"?

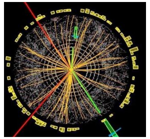
The Use Cases



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



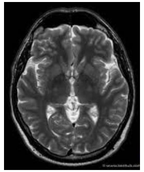
WP3: **HEP real time systems**
Develop AI-based real time selection algorithms for FPGAs at ATLAS. Use xAI methods to understand complex systems

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

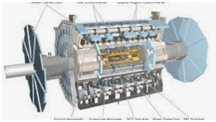
WP4: **Medical Imaging**
Develop xAI pipeline for segmentation of brain tumours in magnetic resonance imaging. Use publicy available databases for xAI developments, focusing on explainability of training strategy

WP6: **Neuro-science**
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

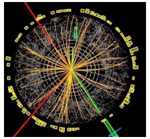


The Use Cases



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 relvant 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 publicy available databases for xAI developments, focusing on explainability of training strategy

WP6: Neuro-science
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

MUCCA use cases

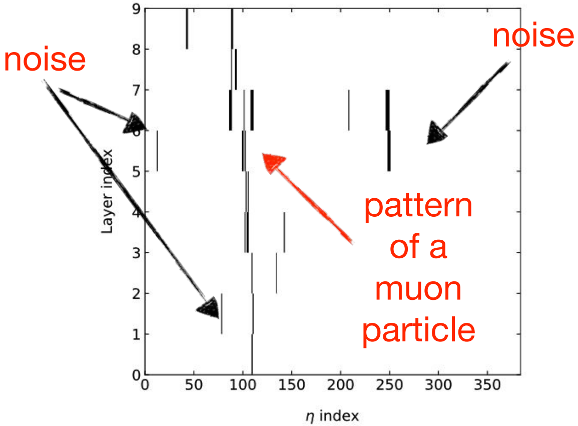
Real-time HEP triggers

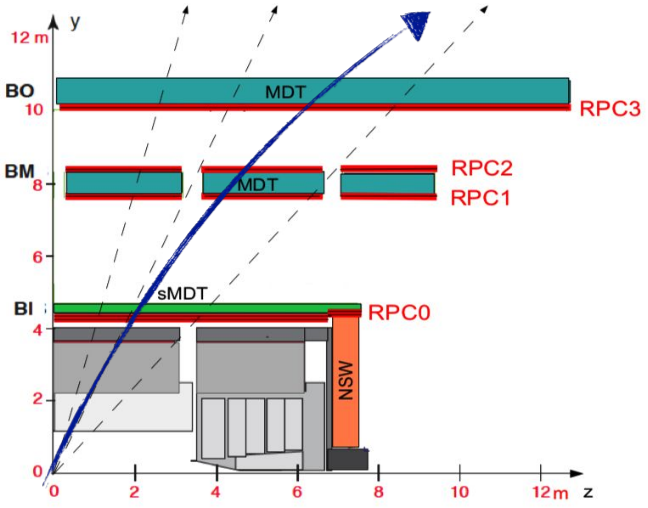
Real-time Triggers in HEP

[Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP | The European Physical Journal C]

Goal: reconstruct momentum and angle of a muon track from the RPC detector hit information **in less than 400ns**.

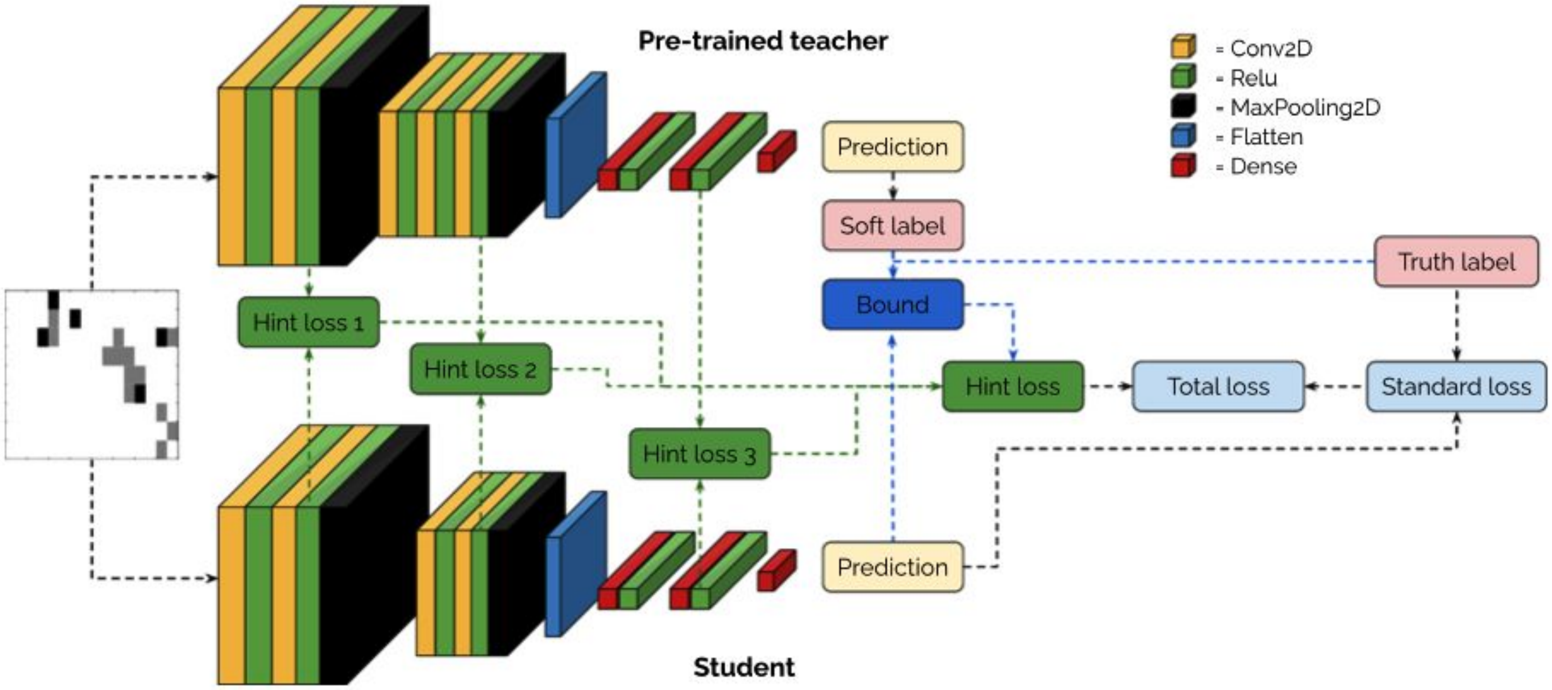
Strategy: multi-stage **AI model compression** based on **quantisation** and **knowledge transfer**.





The model

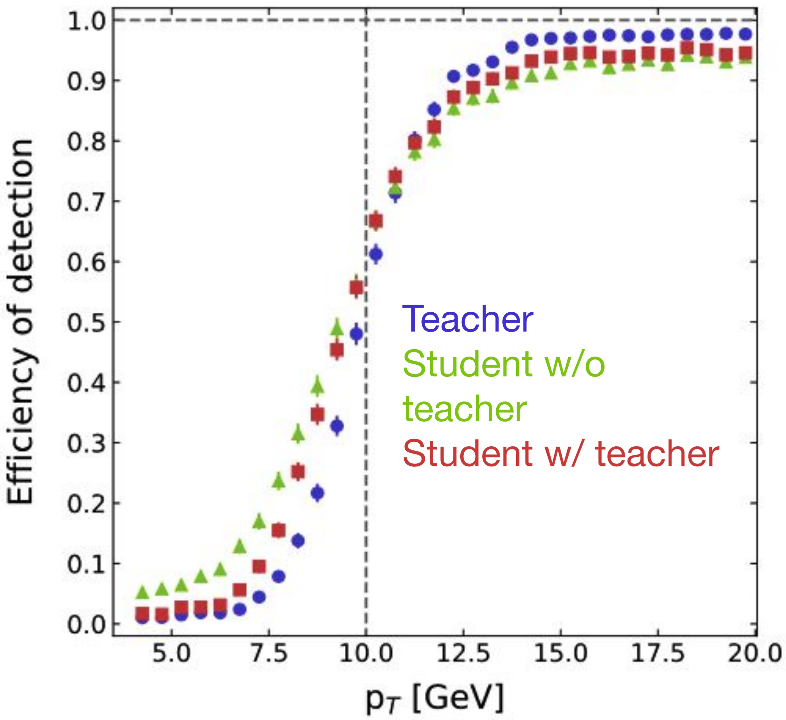
[Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP | The European Physical Journal C]



Performance

[Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP | The European Physical Journal C]

Single muon trigger efficiency curve for a nominal threshold of 10 GeV



FPGA resource occupation

Table 3 Percentage occupancy relative to the total FPGA available resources (model xcvu13p-fhga2104-2L-e [14])

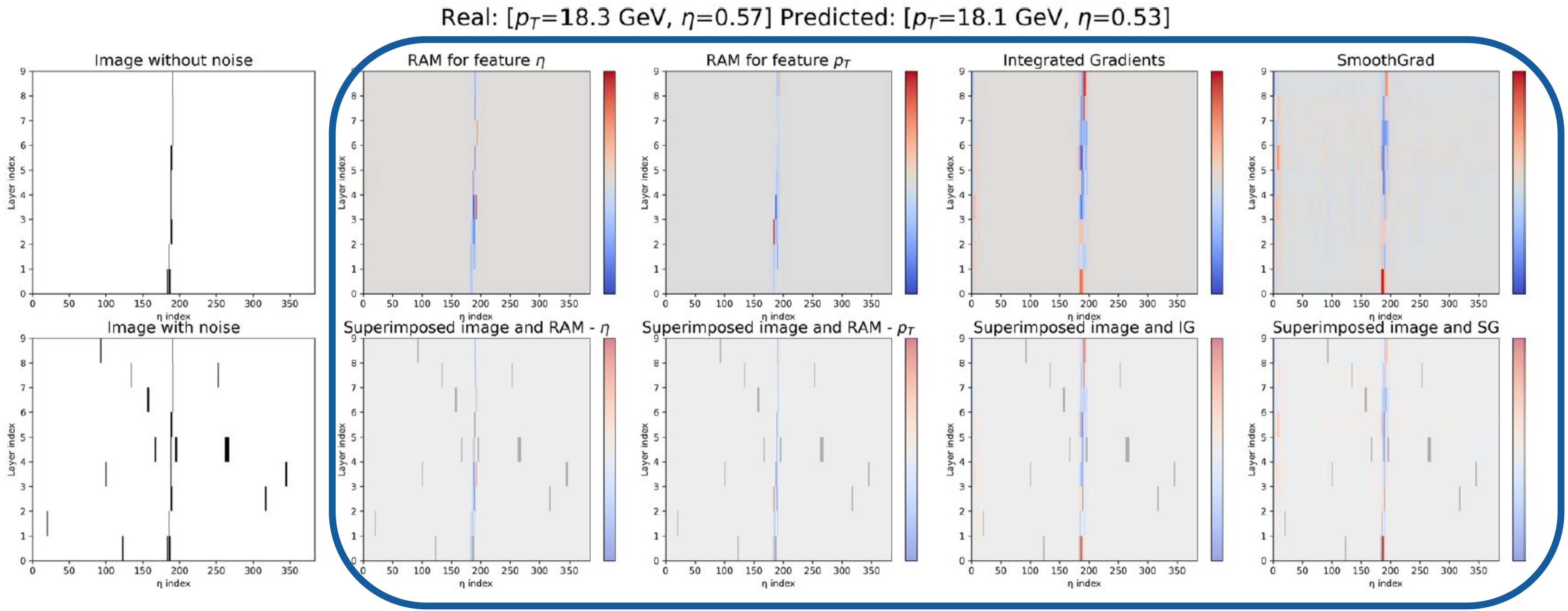
Model (9 × 16)	BRAM	DSPs	FF	LUT
Teacher (%)	20.9	258.0	69.4	15.3
Student 32 bit (%)	3.2	31.0	8.4	2.7
QStudent 4 bit (%)	0.2	0.05	0.4	1.7

Inference time per event on FPGA Xilinx Ultrascale+ XCV13P

- Teacher fp32: 5 ms (Tesla V100 GPU)
- Student 4 bit: 438 ns (hls4ml)
- Student 4 bit: 84 ns (our VHDL implementation)

Strategy 1: saliency maps

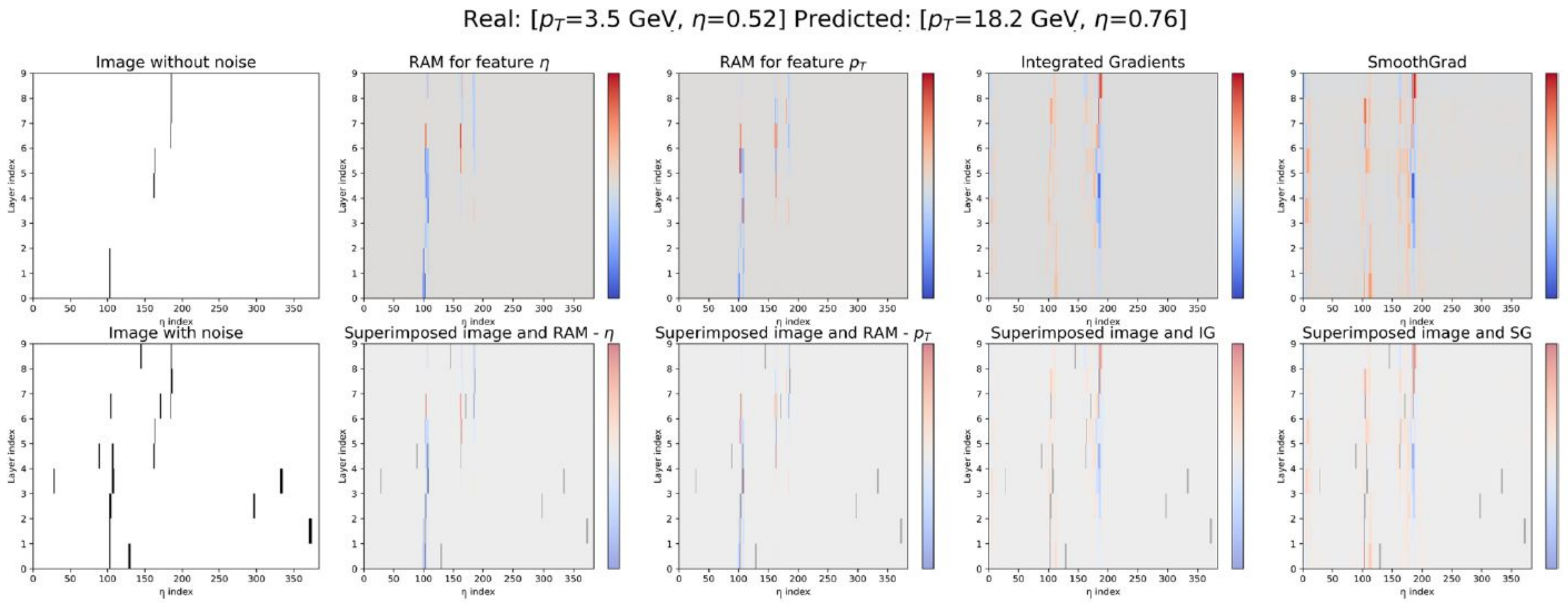
[Convergent Approaches to AI Explainability for HEP Muonic Particles Pattern Recognition | Computing and Software for Big Science]



Overabundance of (potentially conflicting) explanations!

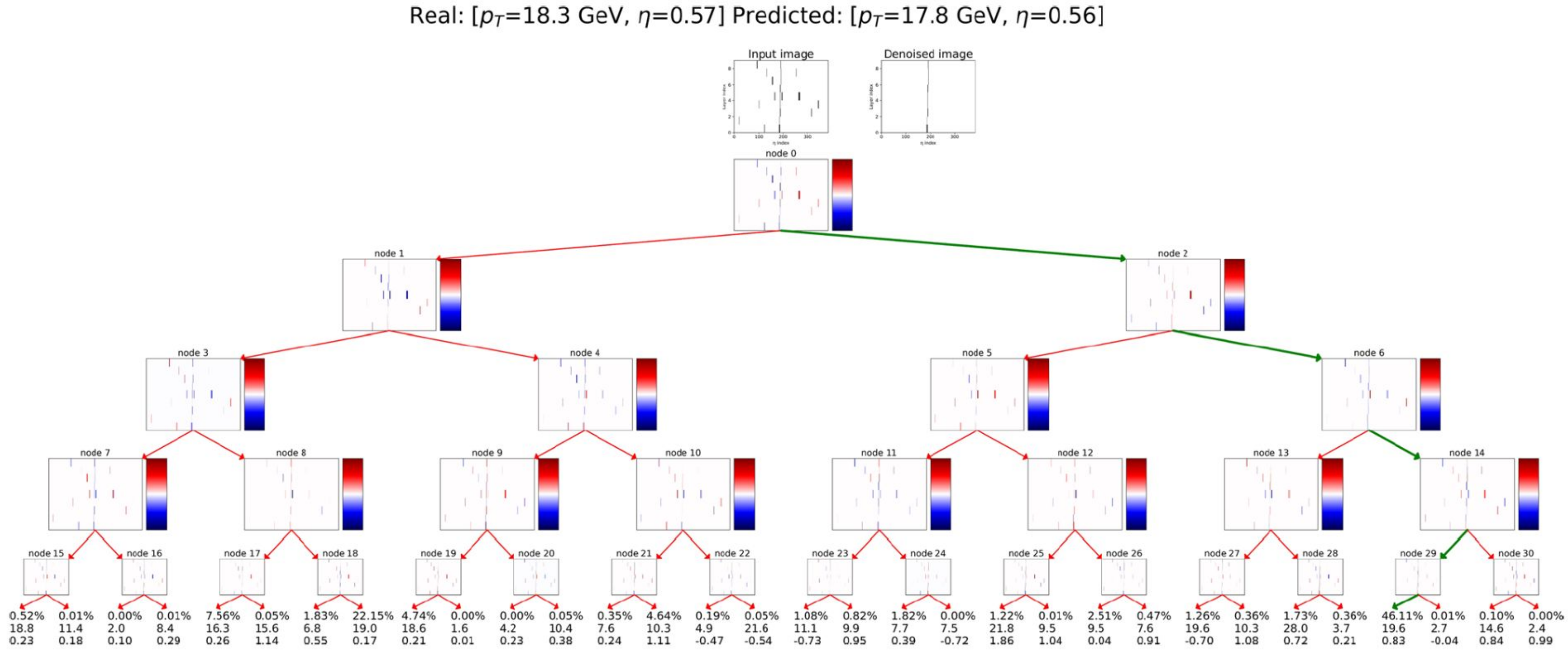
Strategy 1: saliency maps

[Convergent Approaches to AI Explainability for HEP Muonic Particles Pattern Recognition | Computing and Software for Big Science]



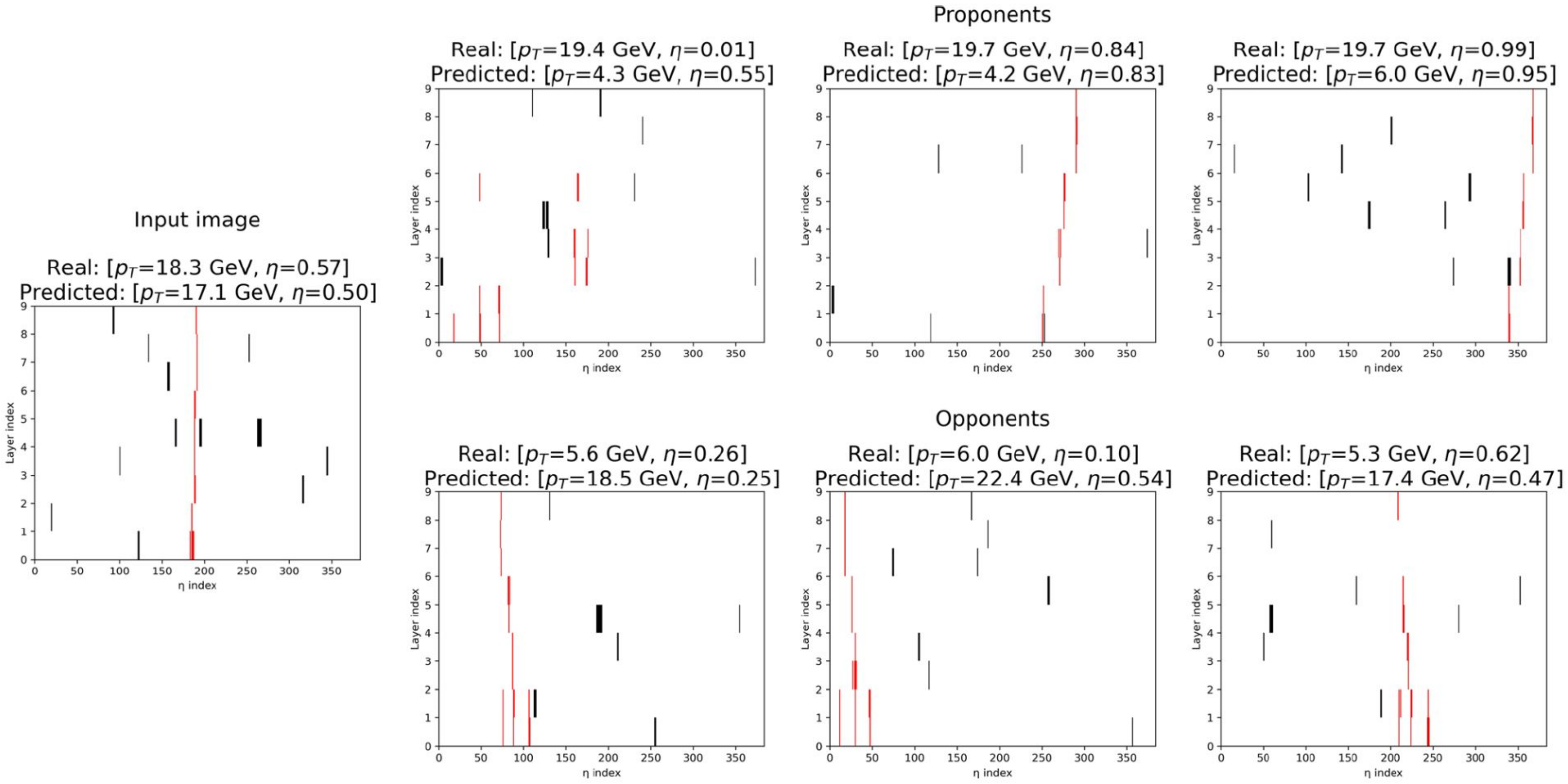
Strategy 2: soft decision trees

[Convergent Approaches to AI Explainability for HEP Muonic Particles Pattern Recognition | Computing and Software for Big Science]



Strategy 3: data attribution

[Convergent Approaches to AI Explainability for HEP Muonic Particles Pattern Recognition | Computing and Software for Big Science]



MUCCA use cases

Search for **new physics** at ATLAS

Introduction

Goal: use two searches for new physics at ATLAS Collaboration at CERN as demonstrators of employability of ML techniques and testbed for xAI.

Search 1 - SUSY: for dark matter candidates resulting from the decay of new particles predicted by Supersymmetry.

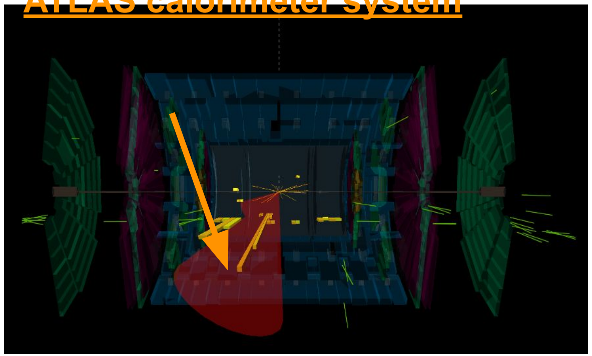
Search 2 - DARK: for “dark” photons, light particles belonging to a new hidden sector not yet discovered because too feebly interacting with ordinary matter.

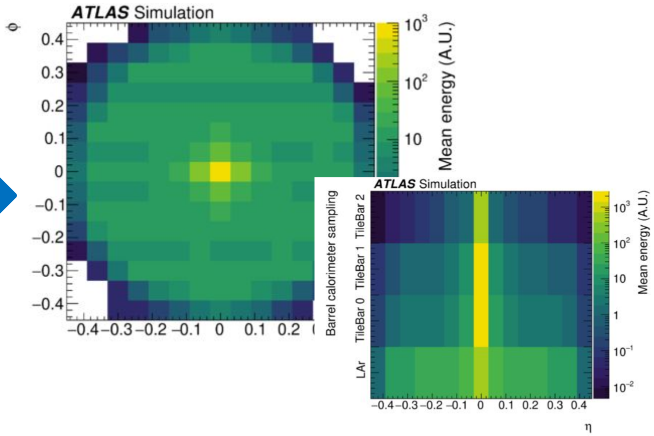
DARK- Dark photons search

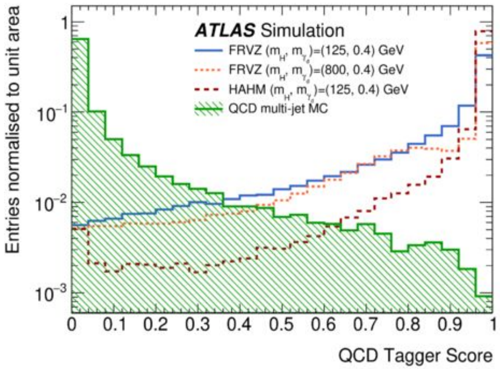
[2206.12181] Search for light long-lived neutral particles that decay to collimated pairs of leptons or light hadrons in pp collisions at $\sqrt{s}=13$ TeV with the ATLAS detector

Signal leaves different signature in the detector wrt background (signal signature is effectively an unknown). ML discriminator (3D-CNN) uses image classification trained to distinguish background processes from signal mapping clusters of hadrons (jets) in 3D coordinates.

ATLAS calorimeter system

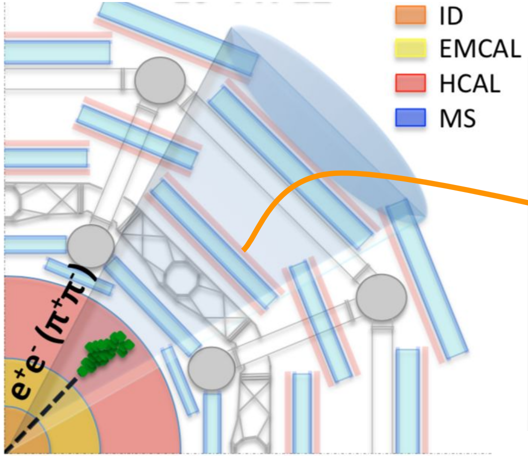




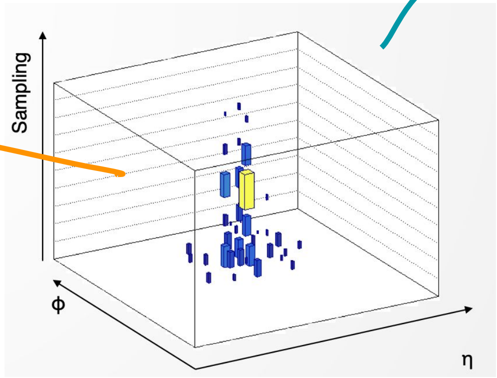


The full pipeline

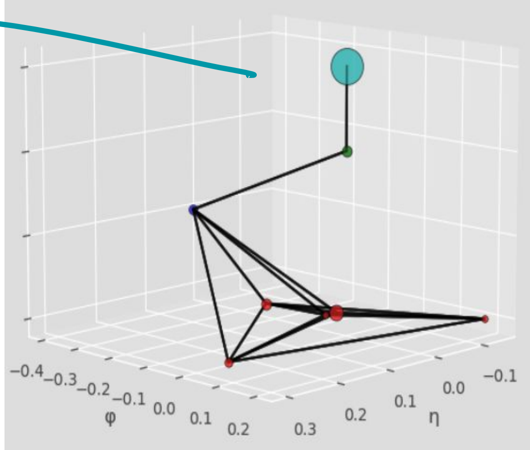
The ATLAS detector
orthogonal view



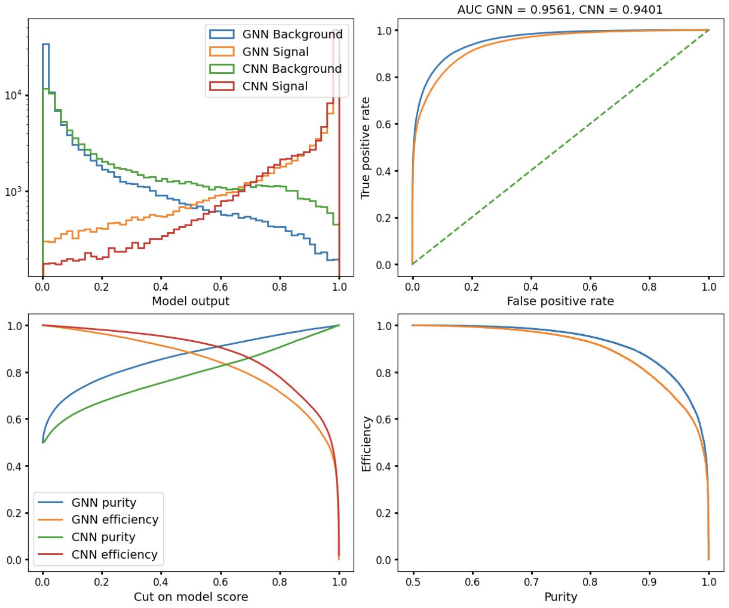
3D image
(sparse)

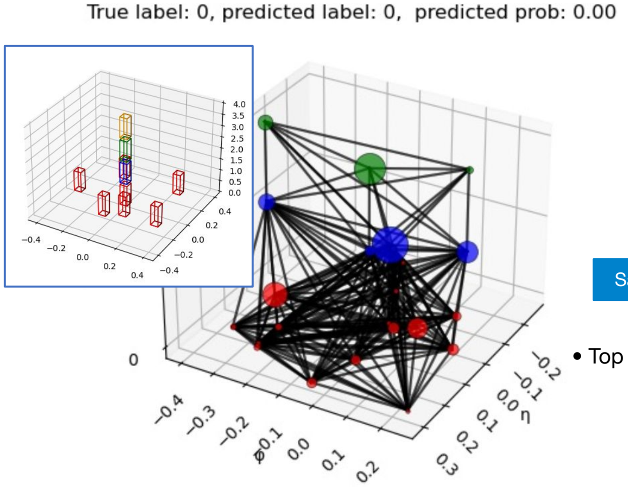


Graph representation
(sparse)

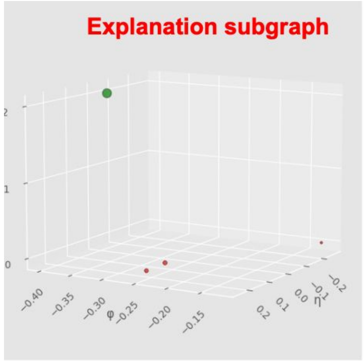


Ongoing research (unpublished)

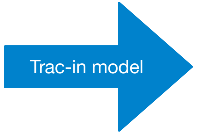


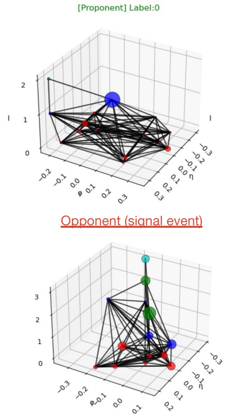


- Top 4 influential nodes



- Top influential data from training dataset





Takeaways

1. Novel AI techniques are highly effective (especially graph neural networks and compression algorithms).

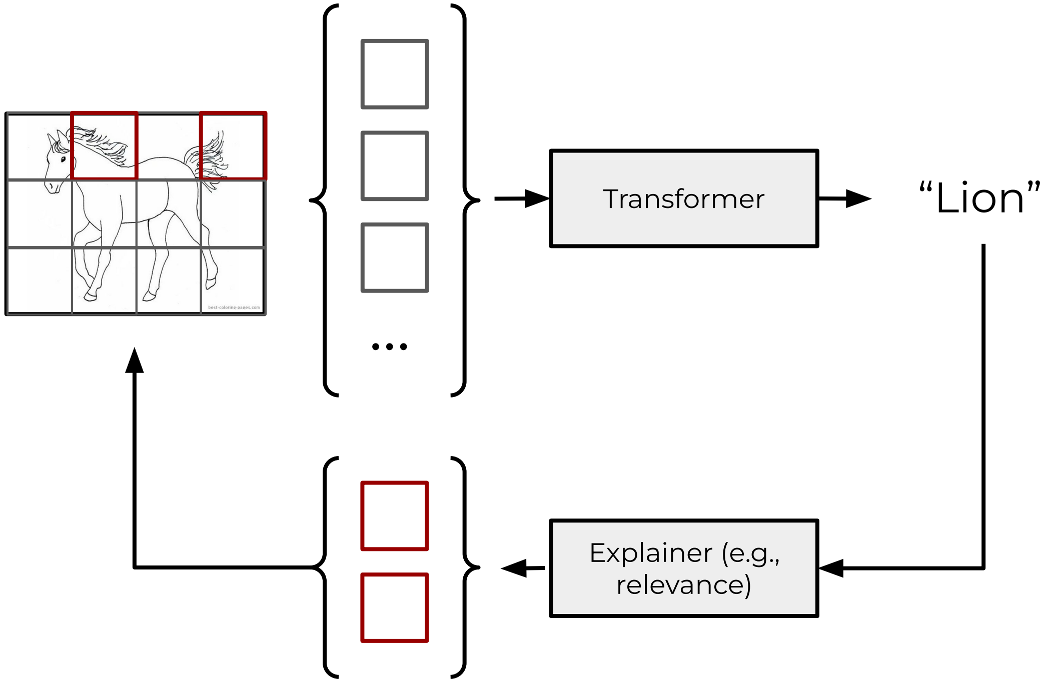
2. Too many, incompatible xAI techniques are inadequate to provide an easy-to-glimpse information to the scientists. Even for an AI expert, combining them is non trivial.

3. In the future, we will probably need a novel, **explainable-by-design** family of neural networks.

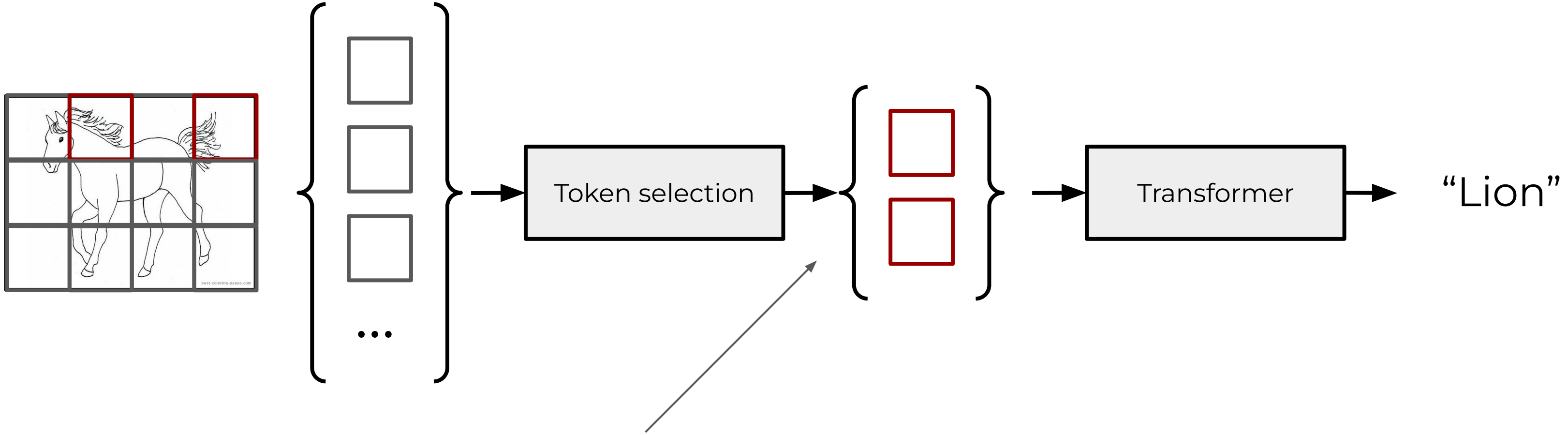
Conclusion

A new generation of **xAI**?

Post-hoc explainability



“Intrinsic” intepretability



Discrete selection!

A practical example

[[2112.07658] AdaViT: Adaptive Tokens for Efficient Vision Transformer]

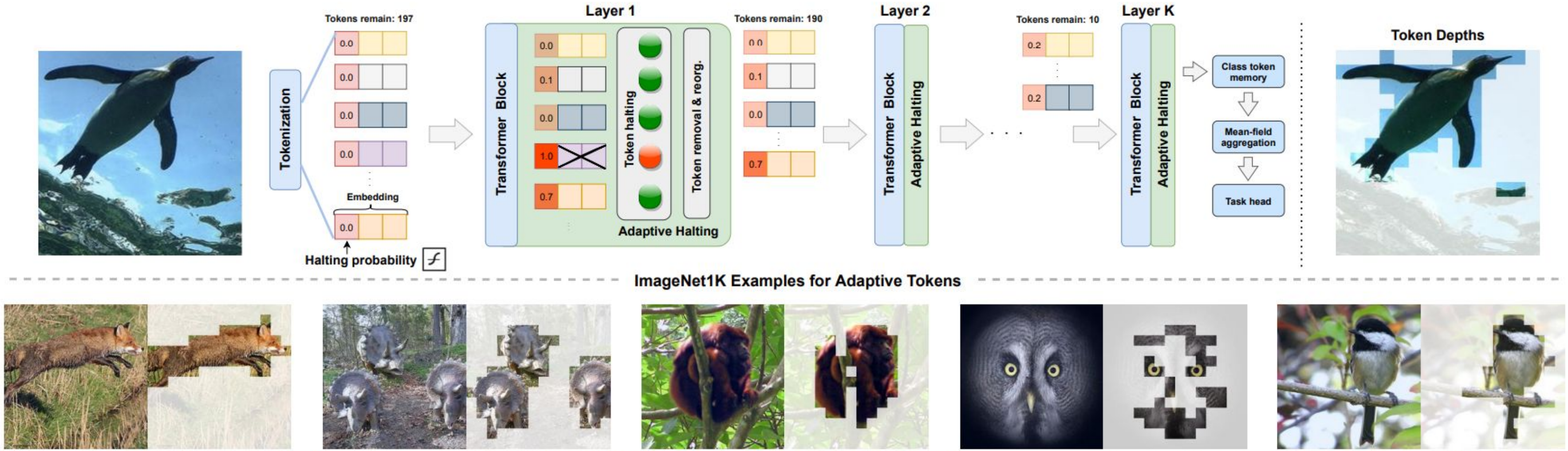
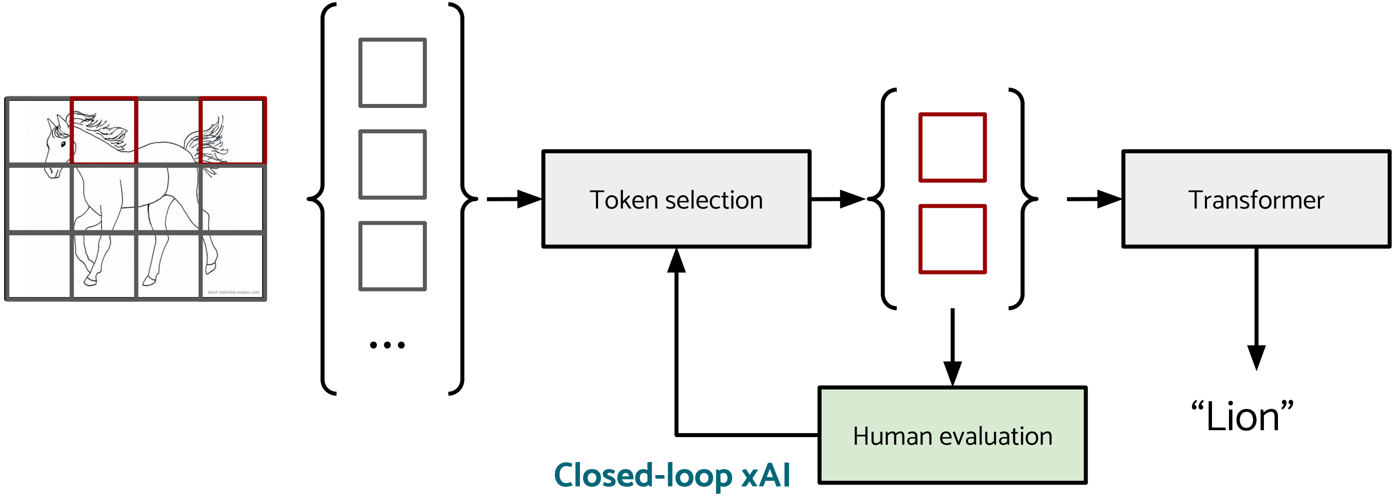


Figure 1. We introduce A-ViT, a method to enable ***adaptive token*** computation for vision transformers. We augment the vision transformer block with adaptive halting module that computes a halting probability per token. The module reuses the parameters of existing blocks and it borrows a single neuron from the last dense layer in each block to compute the halting probability, imposing no extra parameters or computations. A token is discarded once reaching the halting condition. Via adaptively halting tokens, we perform dense compute only on the active tokens deemed informative for the task. As a result, successive blocks in vision transformers gradually receive less tokens, leading to faster inference. Learnt token halting vary across images, yet align ***surprisingly well*** with image semantics (see examples above and more in Fig. 3). This results in immediate, out-of-the-box inference speedup on off-the-shelf computational platform.

In-the-loop explainability (controllability)



Thanks for listening

Simone Scardapane
Assistant Professor

https://www.sscardapane.it/

https://twitter.com/s_scardapane