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

Stefano Giagu
AI@INFN Bologna, 2-3 Maggio 2022

THE MUCCA PROJECT

- CHIST-ERA IV xAI H2020 EU grant 2.2021-7.2024
- **Ultimate goal:** quantifying strengths and solving weaknesses of new and state of the art xAI methods
- **Strategy:** study explainability techniques in different use-cases intentionally chosen to be heterogeneous with respect to the types of data, learning tasks, scientific questions
- **Multidisciplinary Collaboration** that brings together researchers from different fields:
  - high energy physics
  - applied physics in medicine
  - neuroscience
  - computer science

Three phases:

I - apply xAI techniques

II - identify possibile shortcomings of the techniques and metric to evaluate explainability & interpretability

III - combine methods and knowledge to develop general procedures and engineering pipelines for explainable AI

MUCCA CONSORTIUM

project overarches multiple disciplines, from fundamental science to medical clinic and neuroscience, putting together world-experts from the respective fields

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)

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

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

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

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

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 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

MUCCA's PEOPLE

- Sapienza Univ.: S. Ferraina, **S.G.**, L. Rambelli, S. Scardapane, A. Uncini + students
- INFN: G. Bardella, A. Ciardiello, T. Torda, **C. Voena**
- ISS: P. Del Giudice†, G. Gigante, M. Mattia
- MedLea srls: **S. Melchionna**, M. Pratim Borthakur
- Liverpool Univ.: J. Carmignani, **M. D'Onofrio**, C. Sebastiani + students
- Sophia Univ.: **V. Kozhuharov**, G. Georgiev + students
- Bucharest Poli.: **C. Balan**, D. Broboana, E. Chiriac, E. Magos, C. Patrascu, N. Tanase + students

MUCCA WORK PLAN

AI EXPLAINABILITY

- xAI is a broad field of research in AI concerning development of tools to increase trust and understanding of a model’s predictions

- **Main issues with xAI:**
  - strong trade-off between interpretability and representation power of ML models
    - intrinsically interpretable models (linear regression, decision trees, …) orthogonal to models with strong representational power (Deep NN)
  - most xAI methods are oriented towards practitioners of ML (e.g. help experts in making better models), much less toward end-users (e.g. radiologists in the case of AI applied on medical imaging)
  - different xAI methods may disagree on the “explanation”, they may not be always accurate, and they lack principled evaluation metrics

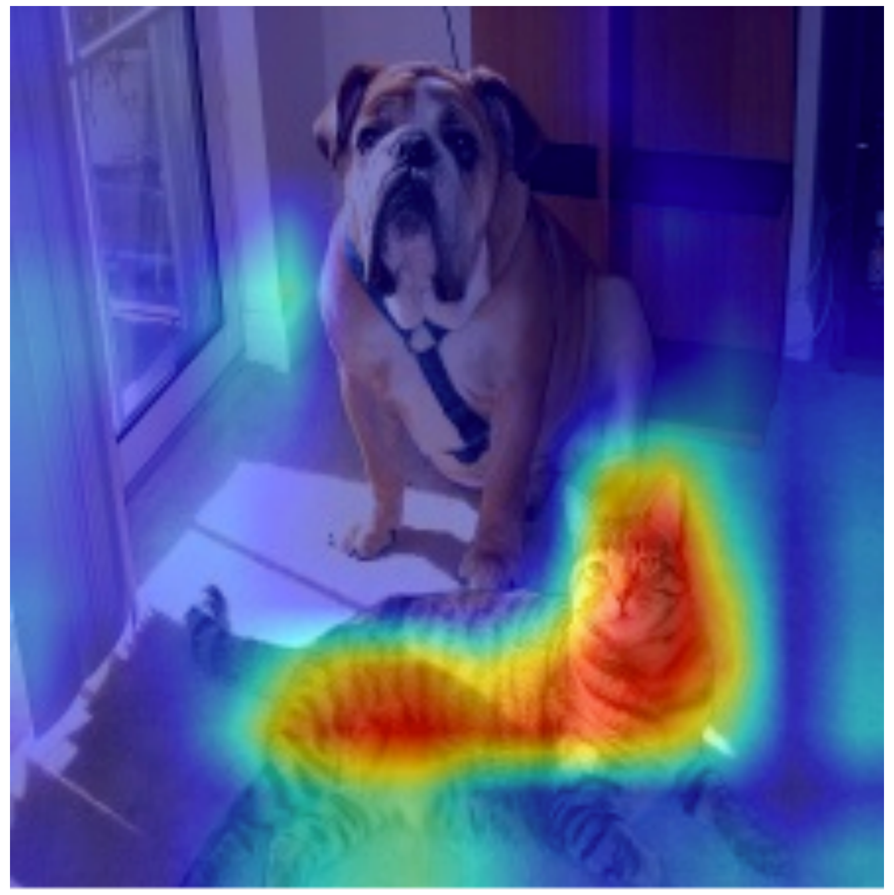
EXPLAINABILITY METHODS

- can be categorised wether they provide global or local explanations and what type of information they provide in output:
  - Visualisation methods: help to understand the correlations between output and input by highlighting the characteristics of the DNN input (or intermediate stages) that most strongly influence the associated output
  - Methods based on data influence: 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
  - Synthetic methods: a separate model of ML is developed, a sort of “white box” trained to mimic the input-output behavior of the DNN. The white box model is more easily explained and / or has the purpose of identifying the decision rules or input characteristics that influence the network outputs
  - Intrinsic methods: DNNs created specifically to provide an explanation of the reason for the output together with the output. Intrinsically explainable DNNs simultaneously optimize both model performance and a certain quality of the explanations produced

A BACK-PROP BASED METHOD: GRAD-CAM HEAT-MAPS

- display the relevance of features based on the magnitude of the gradients flowing through the network layers during training
  - starts with the output feature map of one of the convolutional layers produced by a given input image
  - each channel of the input feature map is weighted with the gradient of the class with respect to the channel, the weights are then propagated to the pixels of the input image
  - useful to measure how much each pixel/region of the input image activate the category predicted by the network

predicted class:
indian elephant



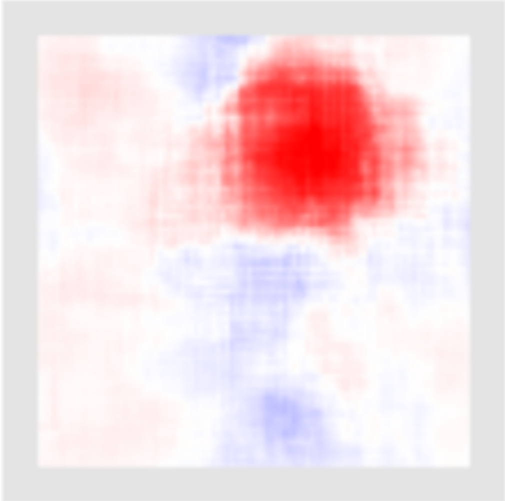
predicted class:
cat

Selvaraju et al, 2017

A PERTURBATION-BASED METHOD: OCCLUSION SENSITIVITY

- display the relevance of the features by comparing the network output for a certain input and for a suitably modified copy of the input
  - underlining hypothesis: performances of a model significantly change when influent elements of the input are masked off (techniques often used in physics to understand transfer function of black box systems)
  - a gray patch is placed in different regions of the input image in order to occlude the overlapping pixels, for each region is checked how much the output prediction of the model changes
  - saliency maps built by weighting each pixel (or group of pixels) by the output prediction variation



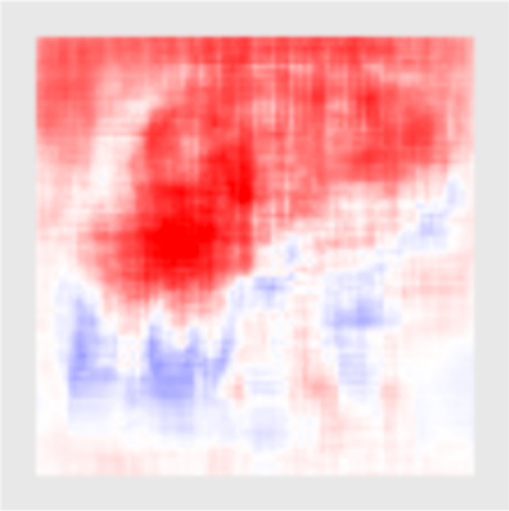


class
dog

original
image

occlusion
mask
32x32

alexnet
stride 2



class
elephant

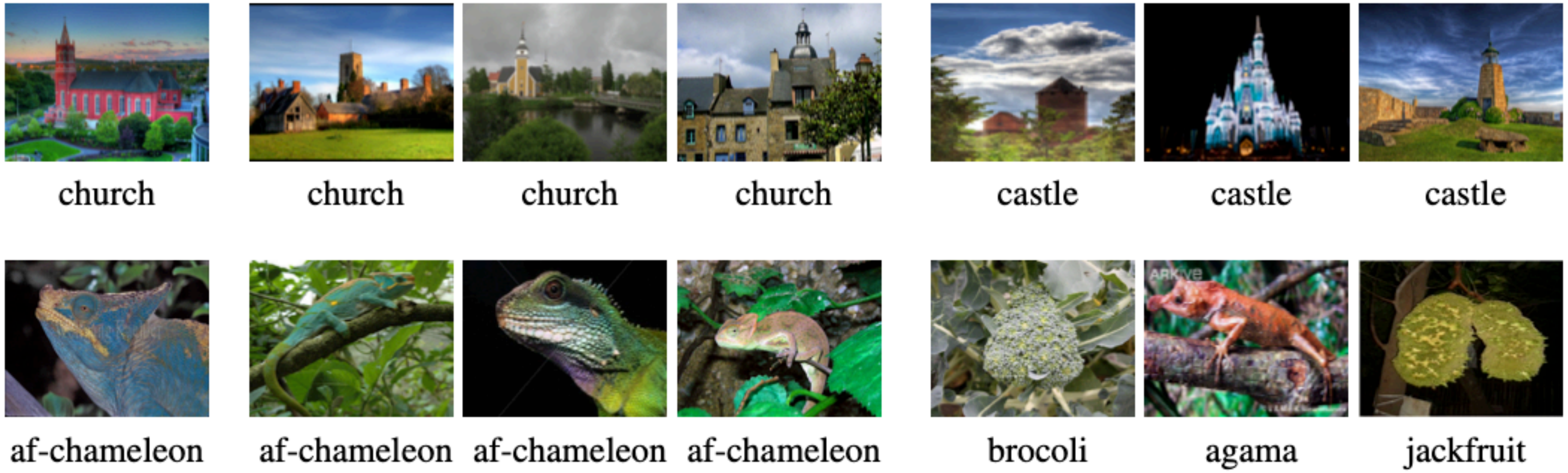
Zhou et al, 2014

A DATA INFLUENCE METHOD: GRADIENT TRACING

- 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
- approximate the ideal influence of a point z on the point z' by storing k checkpoints during the training of the model and computing:

$$\text{Influence}(z,z') \simeq \sum_{i=1}^{k} \eta \nabla l(w_i, z) \cdot \nabla l(w_i, z')$$

test image | proponents (reduce loss) | opponents (increase loss)

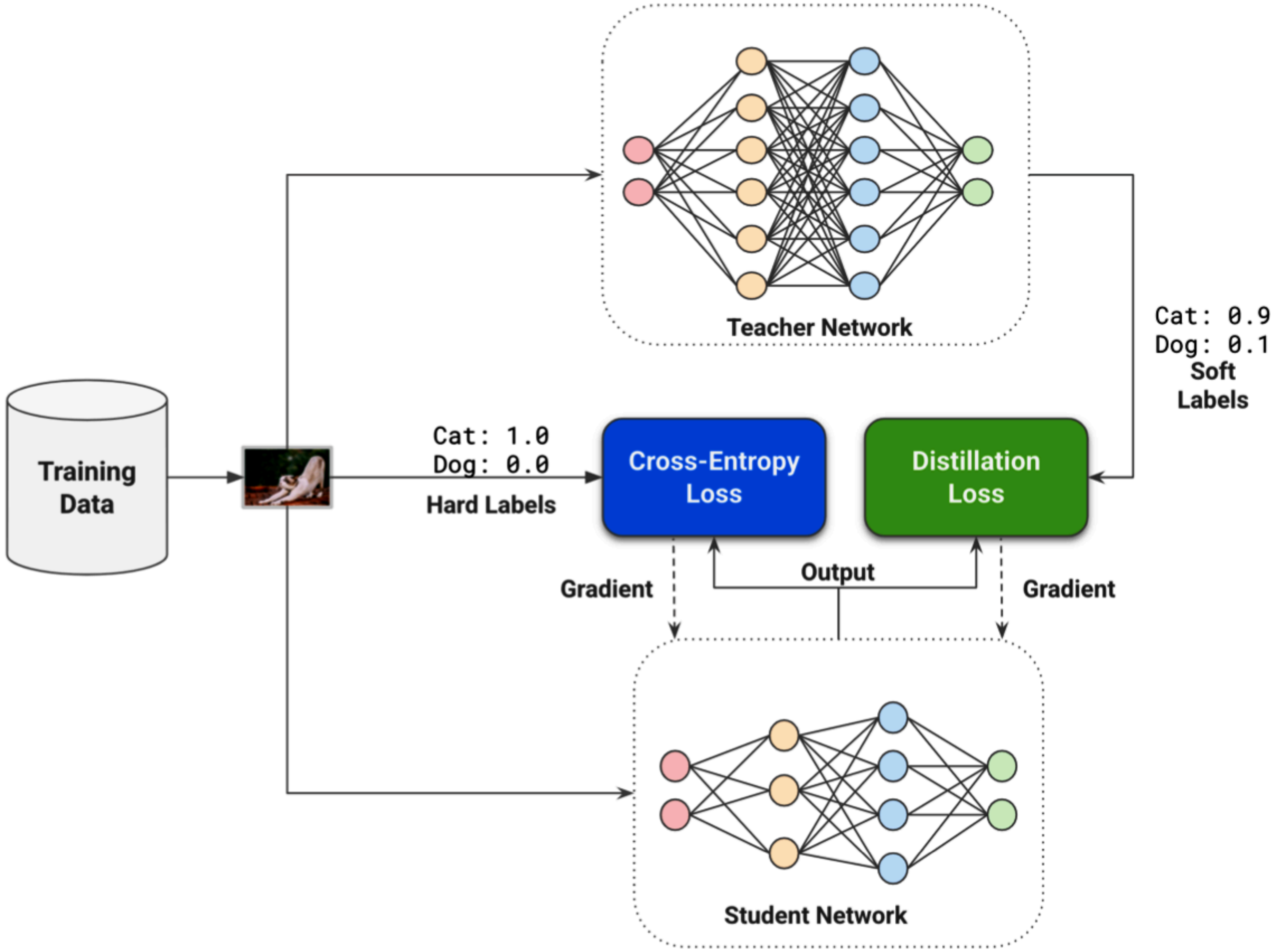


Pruthi et al, 2020

A SYNTHETIC METHOD: KNOWLEDGE TRANSFER BY DISTILLATION

transfer knowledge learned by a larger neural network pre-trained for the same task to a smaller, simpler and more explainable model

- the teacher is used to generate soft labels that replace the values of the ground truth labels with the probabilities estimated by the teacher that the input belongs to each class
- during the training the student model can learn both from the hard (ground truth) and from the soft labels produced by the teacher



$$L = \alpha L_{xent} + \beta L_{dist}$$

$$L_{dist} = \text{x-entropy}(\hat{y}, y^{soft}; \mathbf{w})$$

$$y_i^{soft} = \frac{\exp \frac{Z_i^{teacher}}{T}}{\sum_j \exp \frac{Z_j^{teacher}}{T}}$$

T: temperature parameter which acts as a smooting for the distribution of soft labels

distillation facilitate student's training by allowing to capture relationships between classes that are not represented in the hard labels of the training dataset

Hinton et al, 2014

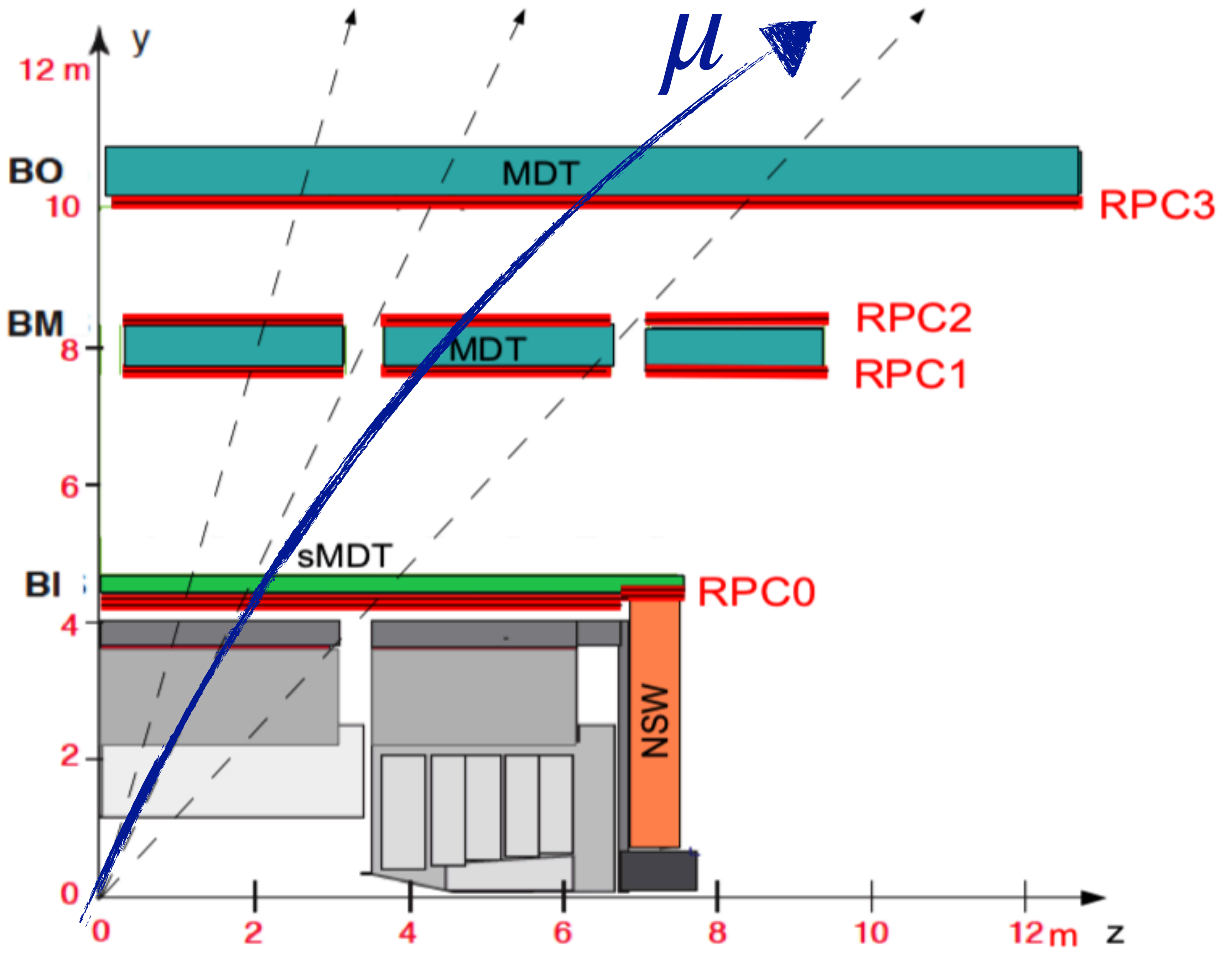
A MUCCA USE-CASE: xAI ON DNN FOR REAL-TIME TRIGGERS IN HEP

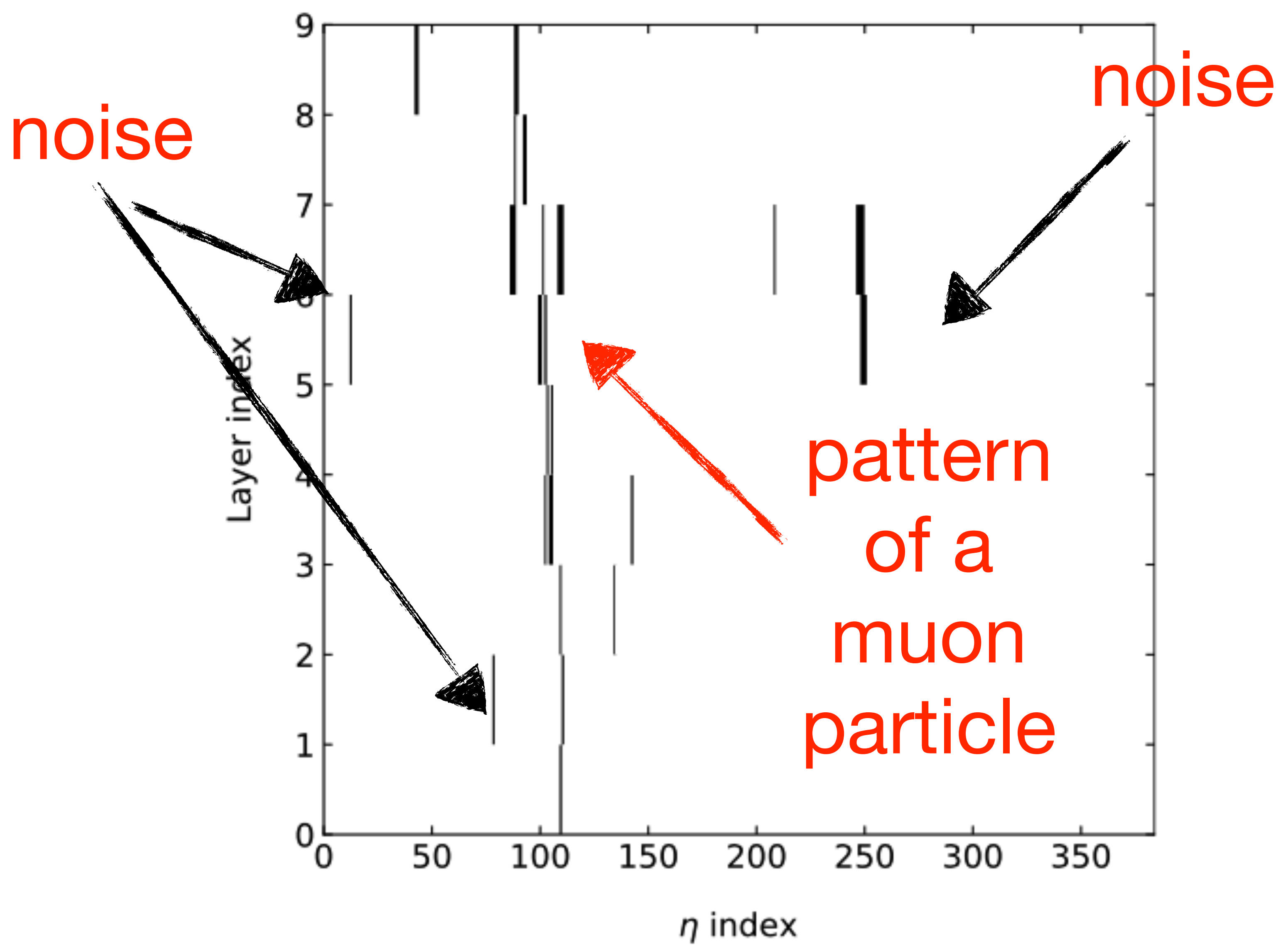
Goal: accurately reconstruct the momentum and angle of the muon track from the RPC detector hit information **in less than 400ns** (3 orders of magnitude faster than fastest AI models on CPUs and GPUs)

Latency and FPGA resource occupancy are in a trade-off relationship, while AI model performance strongly depends on the neural network scale

Strategy: multi-stage **AI model compression** and simplification based on **aggressive quantisation** and **knowledge transfer techniques** to avoid degradation of physics performances

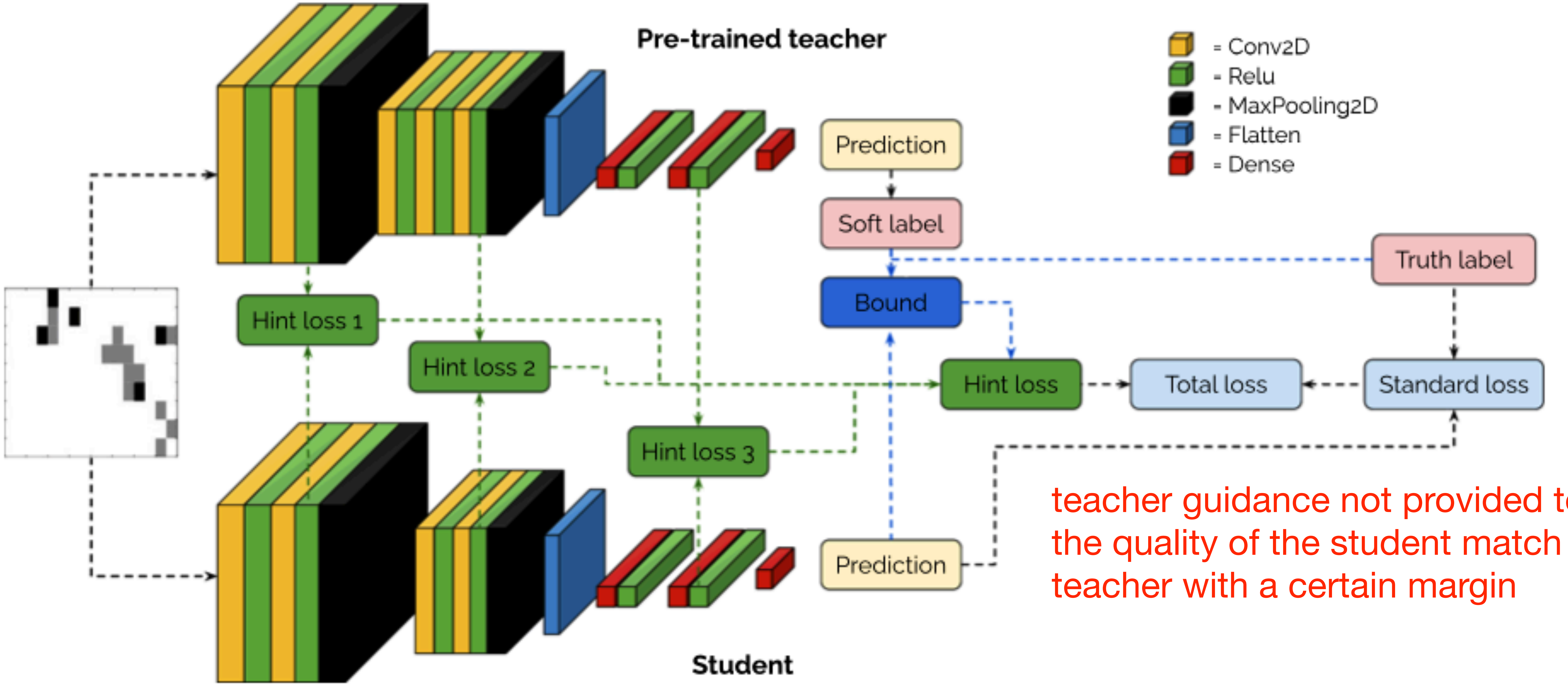
xAI: lightweight models obtained using distillation easier to explain, but extreme sparsity on data and model quantization may challenge xAI methods





KNOWLEDGE TRANSFER FOR CNN MODEL COMPRESSION

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



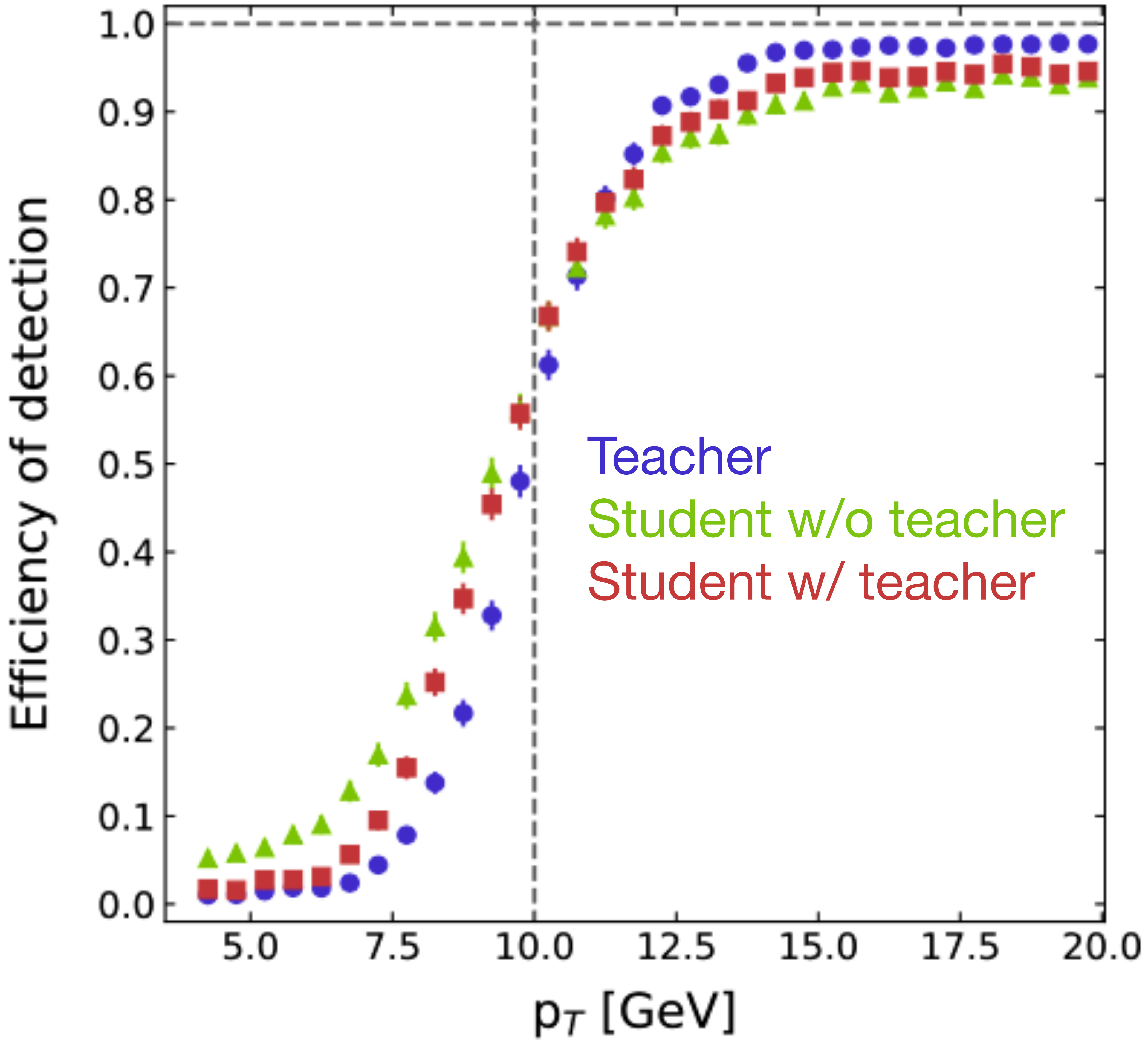
teacher guidance not provided to the student once the quality of the student match or surpass that of the teacher with a certain margin

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

S. Francescato, S.Giagu, F. Riti, G.Russo, L.Sabetta, F.Tortonesi, Eur. Phys. J. C (2021) 81:969

PRELIMINARY PERFORMANCES

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

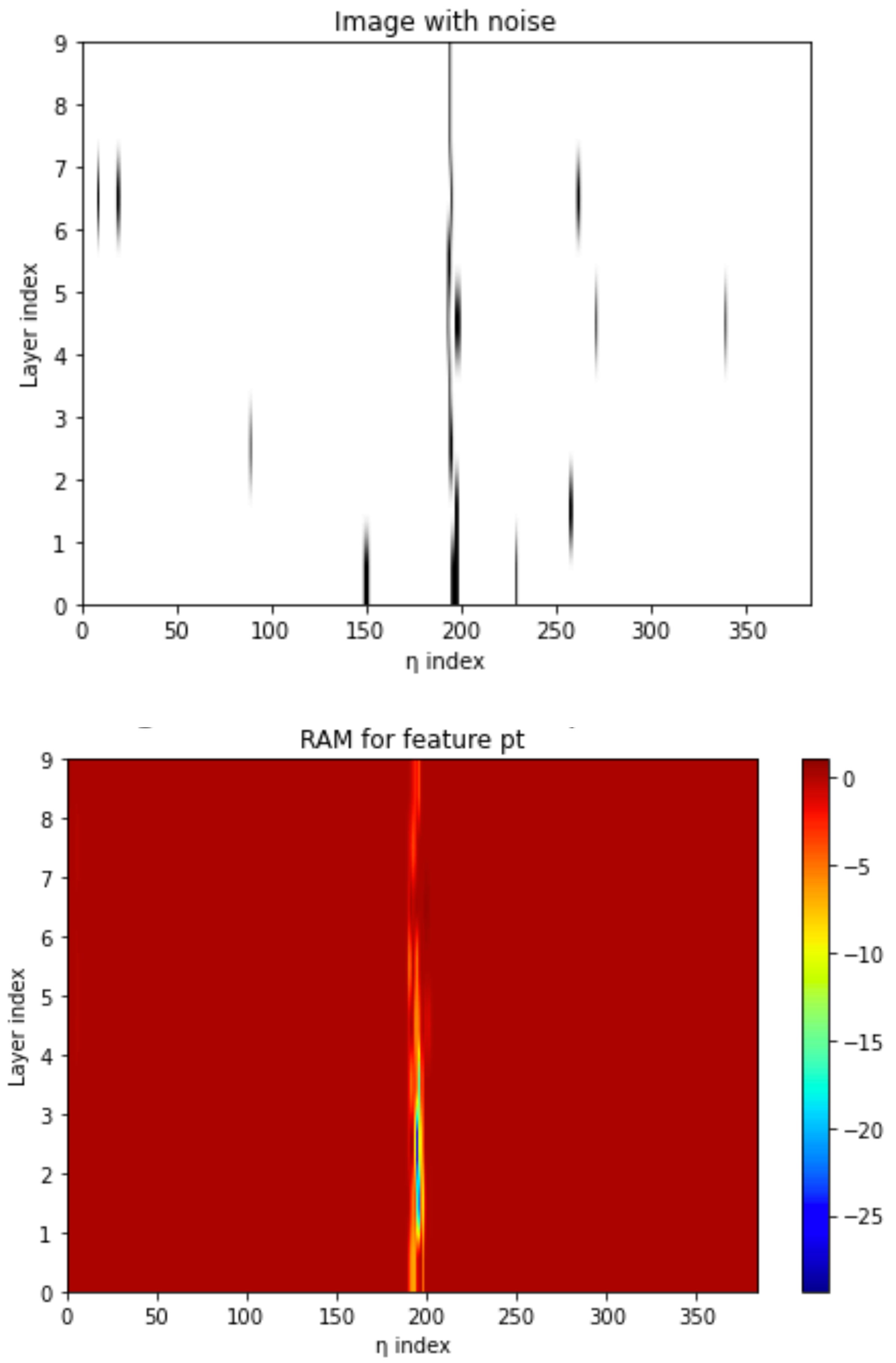
- Teacher fp32: 5 ms (Tesla V100 GPU)
- Student 4 bit: 438 ns (hls4ml)
- Student 4 bit: 84 ns (our VHDL implementation)

S. Francescato, S.Giagu, F. Riti, G.Russo, L.Sabetta, F.Tortonesi, Eur. Phys. J. C (2021) 81:969

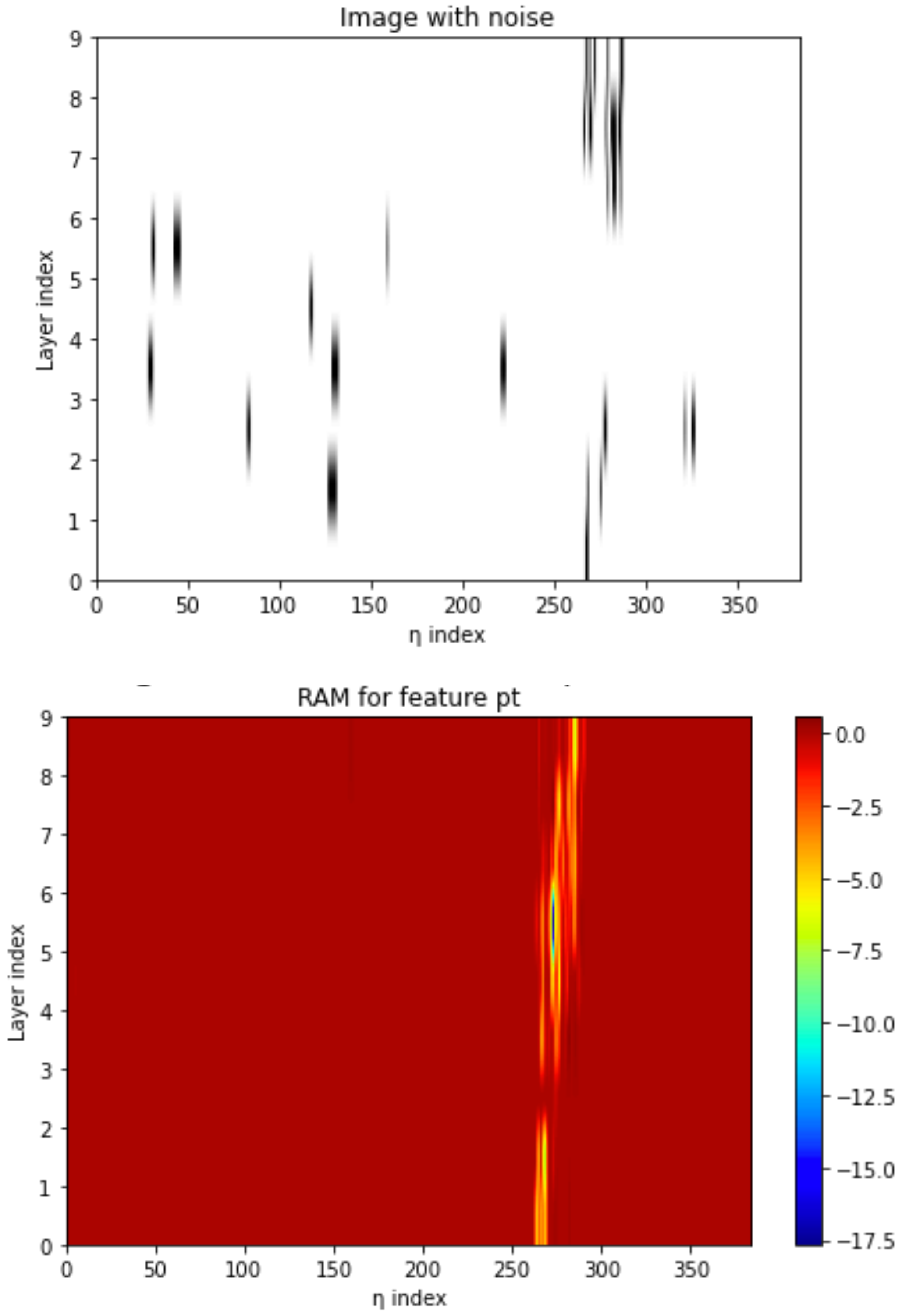
xAI VIA HEAT MAPS

- visualize pixels that have contributed the most to the track reconstruction
- heat maps obtained with the RAM technique (regression activation maps (generalise grad-CAM for regression tasks))

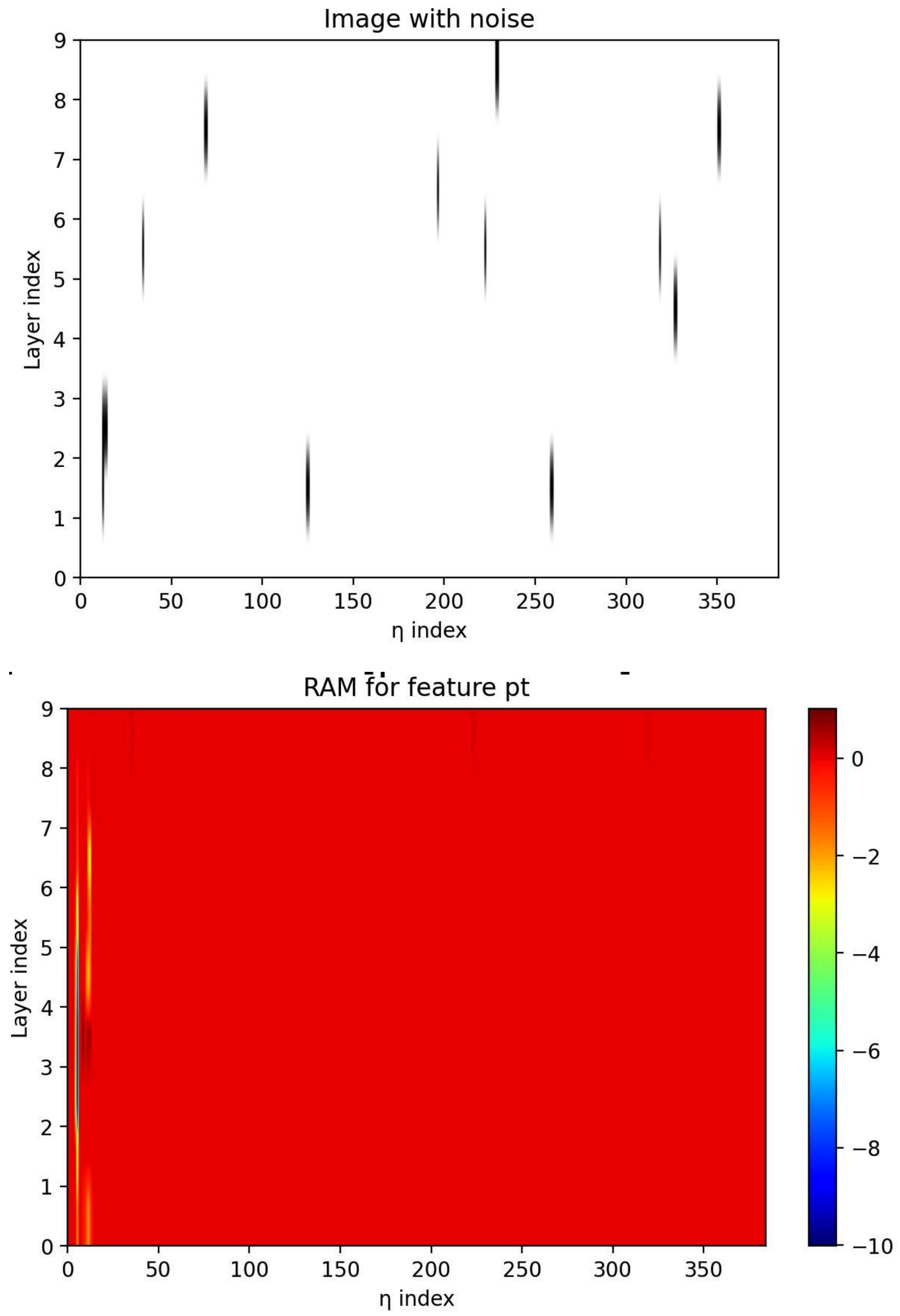
true positive case



false positive case



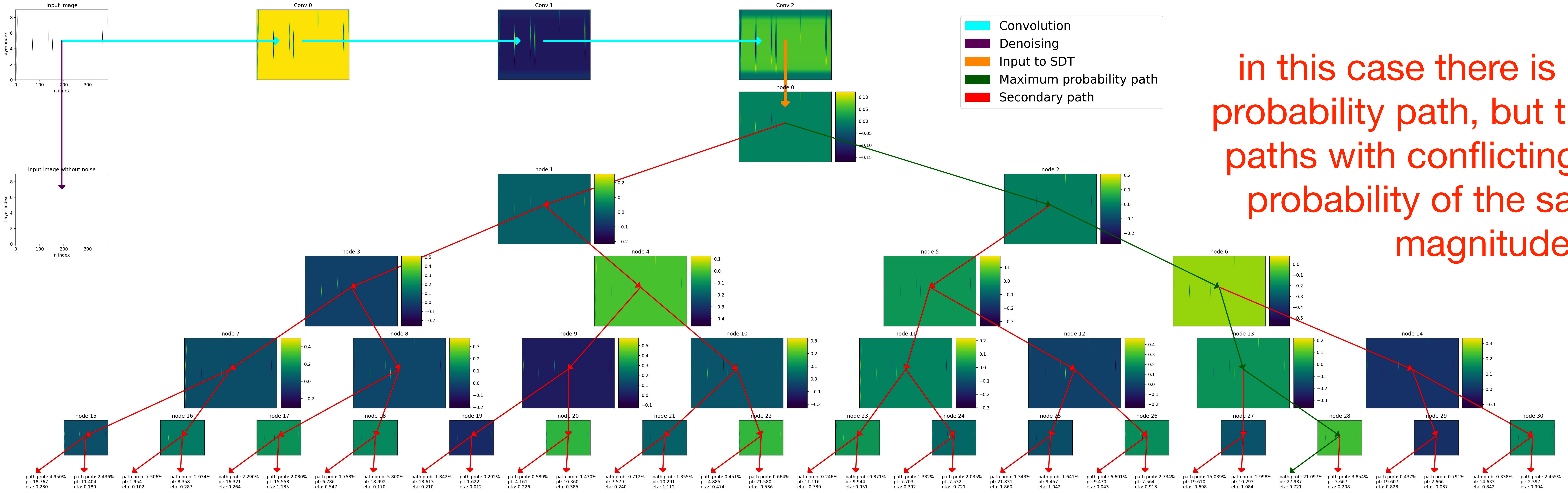
noise-only FP case



xAI VIA DISTILLATION TO CONVOLUTIONAL SOFT DECISION TREES

- teacher distilled to a intrinsically explainable student, as an example a decision tree (Convolutional Soft Decision Tree)
- Soft Decision Trees (SDTs) are capable to consider each output leaf node with a specific probability that will contribute to the final outcome of the model
- Convolution SDTs are an improvement of this idea with Convolutional layers on top to provide a latent representation of the input data to be passed to the hierarchical mixture of the trees

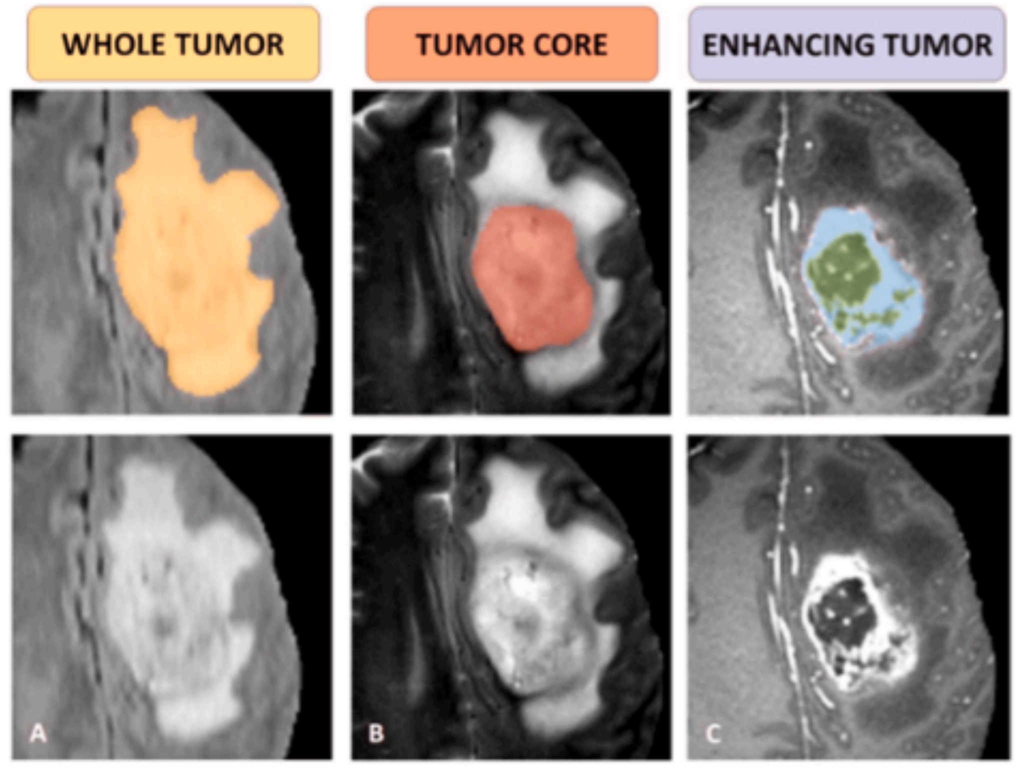
Real = [pt: 0.0, eta: 0.0000]   Predicted = [pt: 15.4892, eta: 0.2855]



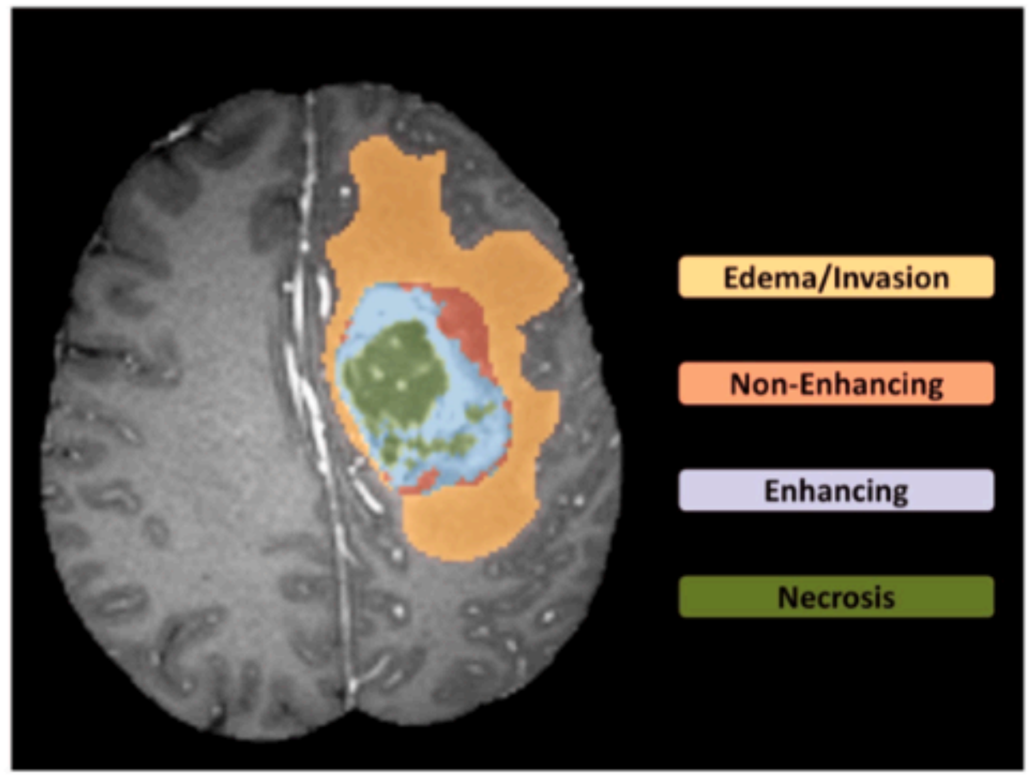
in this case there is no dominant probability path, but there are many paths with conflicting outputs and probability of the same order of magnitude …

A MUCCA USE-CASE: xAI IN MEDICAL IMAGING

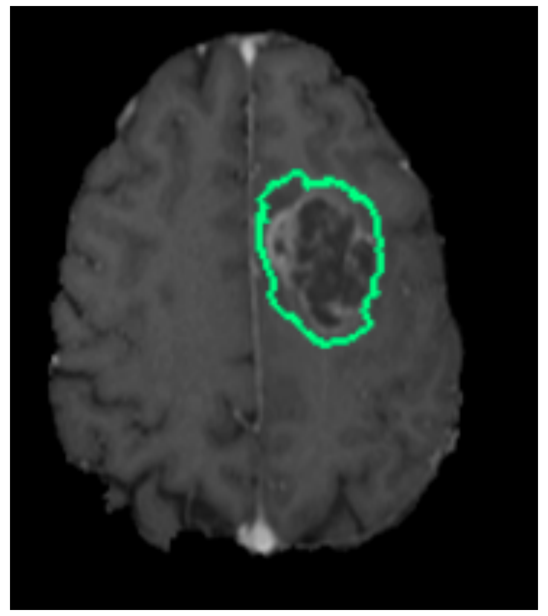
- use open MRI images databases to train DNN for segmentation tasks of both anatomical brain structures and healthy/pathological tissue
- apply state of the art xAI algorithms and test their ability to produce consensus on final users quantifying it by appropriate metric
- study stability of the metric vs different datasets, training strategies, architecture constraints, data augmentation, …



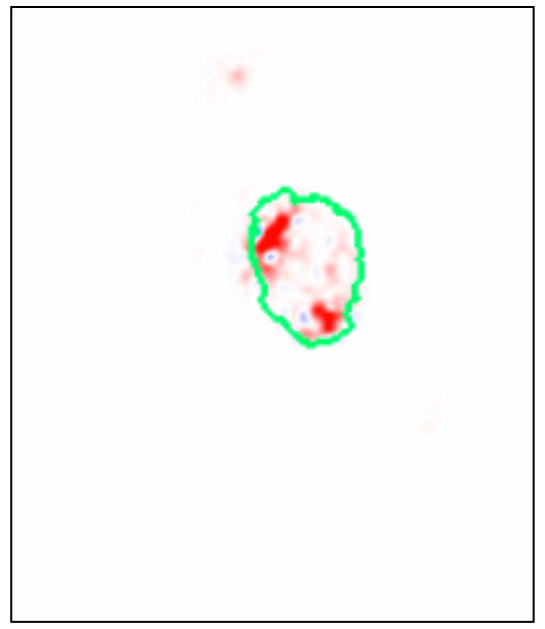
BraTs17

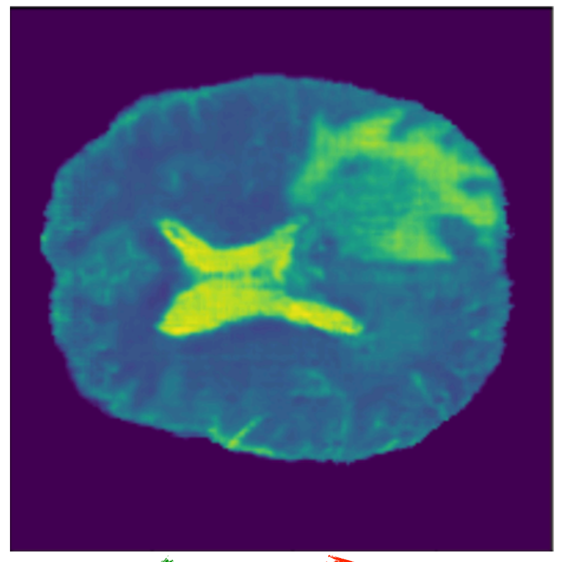


UNet , DeepLabV3+, ResNet3D



saliency maps

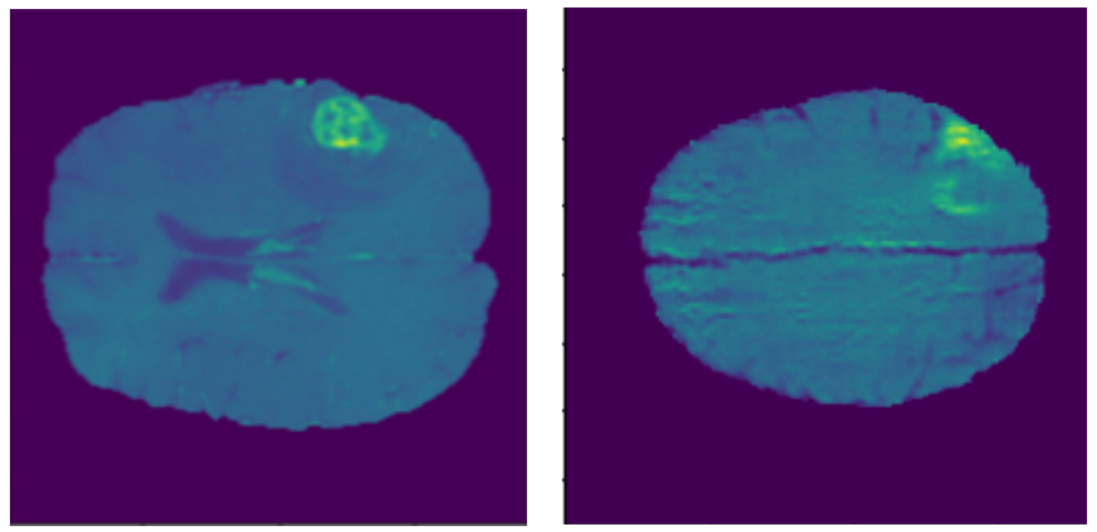


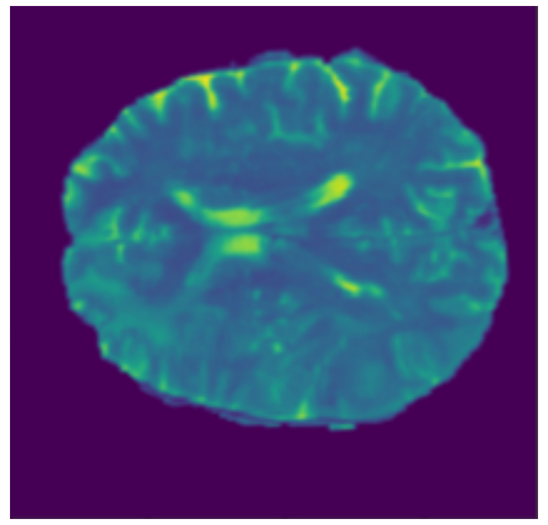


gradient tracing

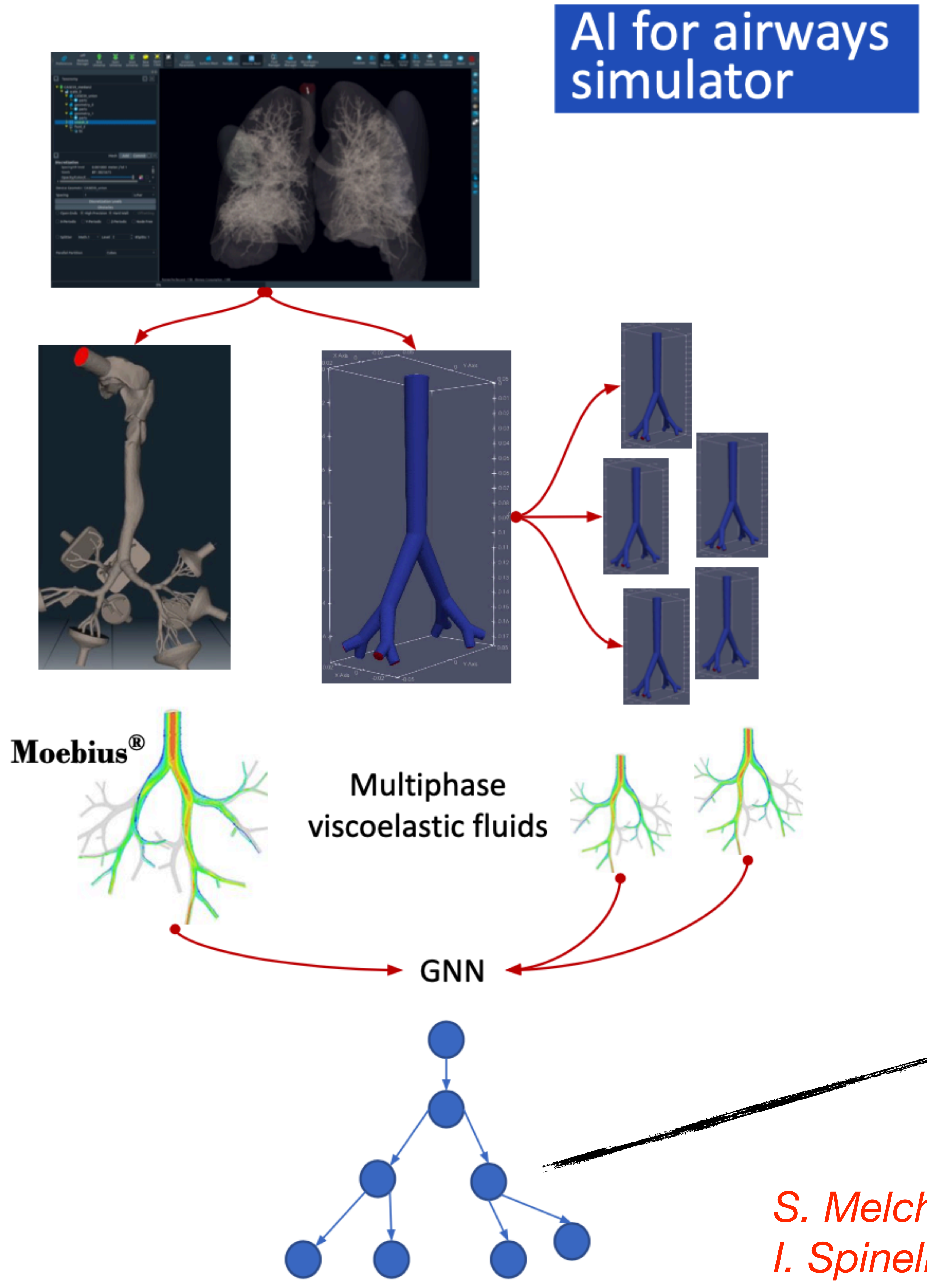
proponents

opponent





A MUCCA USE-CASE: xAI IN FUNCTIONAL IMAGING



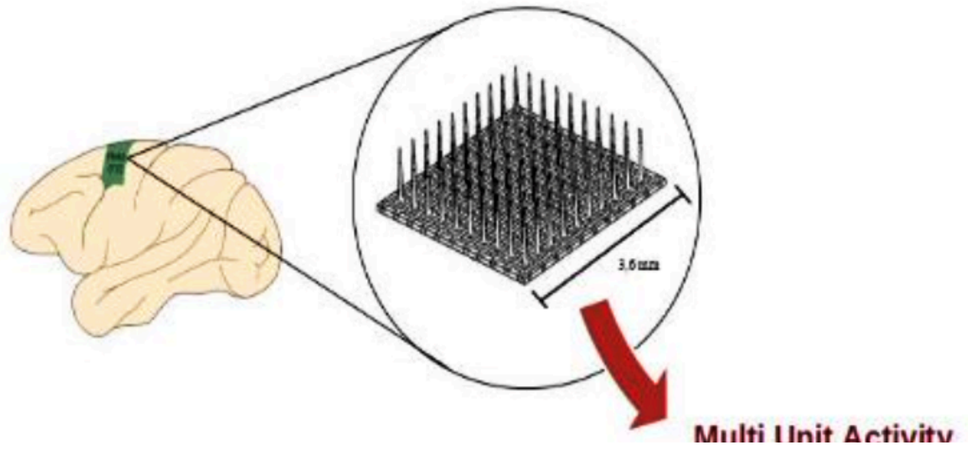
- Develop an integrated approach for 3D reconstruction from medical images to perform fluid dynamics simulation & experiments on respiratory tracts (airways)
- Assess airflow and air+mucus dynamics in respiratory tracts: Newtonian and non-Newtonian rheology
- Reach a high level of automation to handle several geometries (patients)

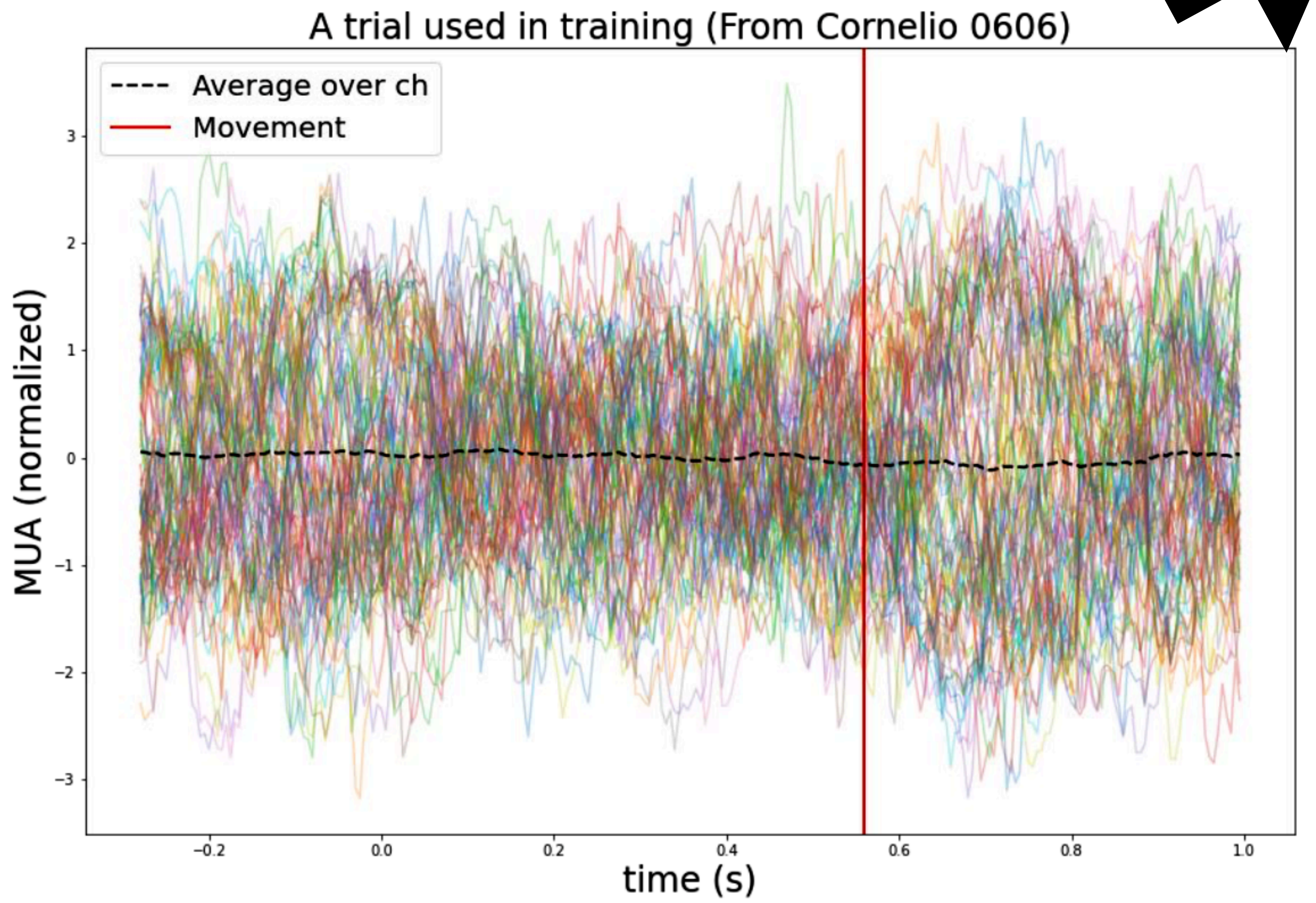
- Graph Neural Network based fluid dynamic simulation
- explainability via meta-learning

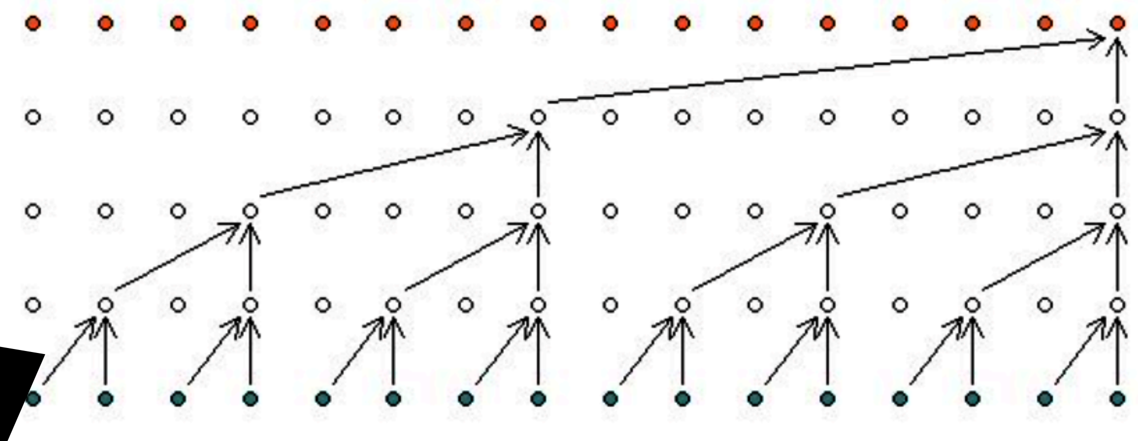
S. Melchionna, Moebius fluid dynamics simulation in complex geometries, 2020
I. Spinelli, S. Scardapane, A. Uncini, A meta-learning approach to train graph neural networks 2021

A MUCCA USE-CASE: xAI IN NEURO SCIENCE

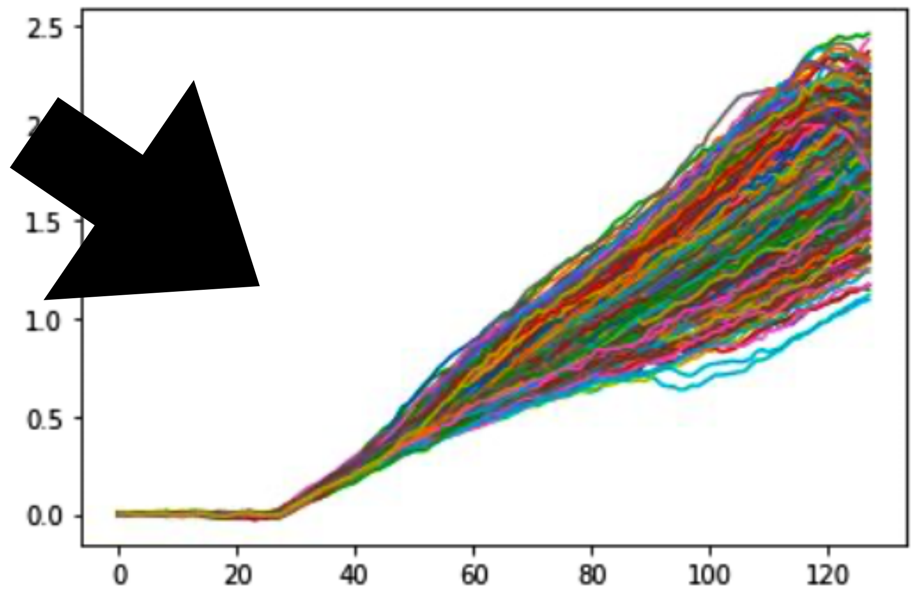
- goal: uncover computational brain strategies while non human primates perform tasks requiring the inhibition of planned movements







horse race model

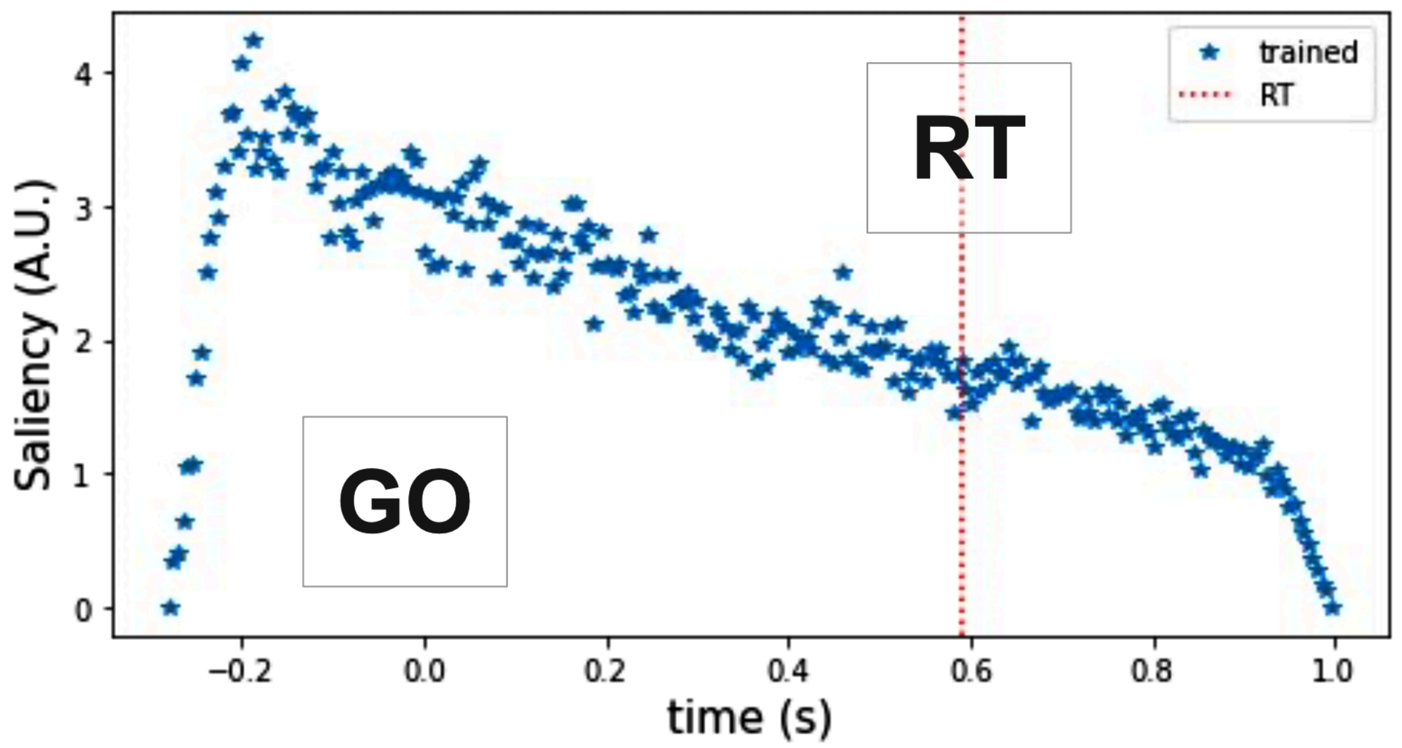


DNN mapping the complex multidimensional sequence to a simpler one (a linear ramp) preserving causality: wavenet with dilated causal convolutions

functional explainability: where is, in time, the information used to build the ramp?

$$x(t) + \delta x(t) \rightarrow f(t) + \delta f(t)$$

$$df(t) \rightarrow \text{DNN}^{-1} \rightarrow \delta x(t)$$



SUMMARY AND EXPECTED IMPACT

- **Status of the project:** some delay wrt the original plans due to Covid19 restrictions and delay in obtaining funding from one of the funding agency, nevertheless:
  - successfully implemented appropriate AI algorithms for all the use cases
  - performed an extensive survey and analysis of state-of-the art xAI methods and developed new ones, identified the most suitable ones to be used for the next phase of the project
- **Expected Results:** knowledge base and xAI tools (documentation and procedures/engineering pipelines)

Multiple level impact:

1. enable users to better understand AI models and diagnosis limitation using xAI
2. systematic understanding of which xAI methods better adapts to specific applications
3. skill development and training for young researcher