FlashSim: End-to-End simulation with flow matching

SCUOLA NORMALE SUPERIORE

INFN Istituto Nazionale di Fisica Nucleare

CMS

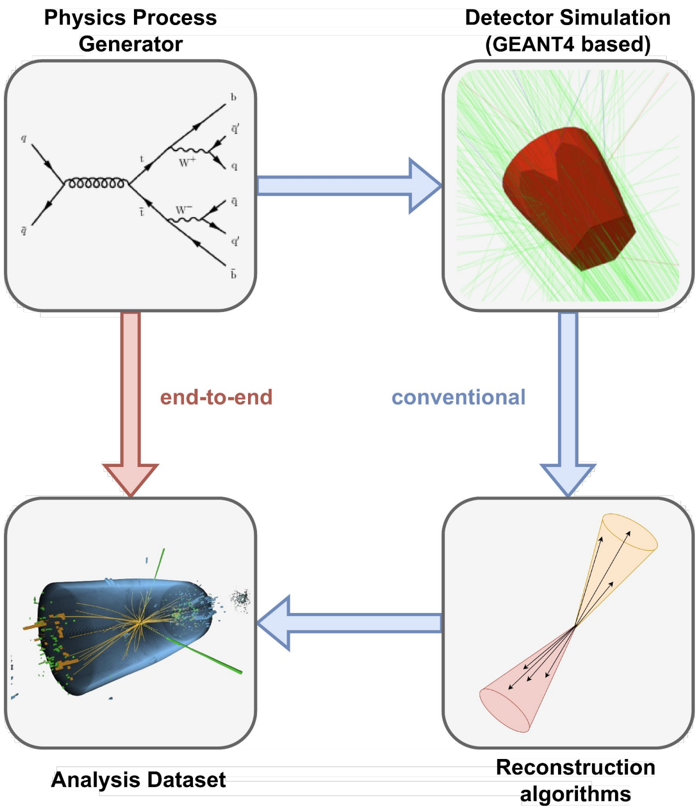
Francesco Vaselli on behalf of the CMS Collaboration

We propose an *end-to-end* approach for faster simulations

Main idea: going directly from the generator output objects to the high level analysis objects (jets, muons ...)!

We want something:

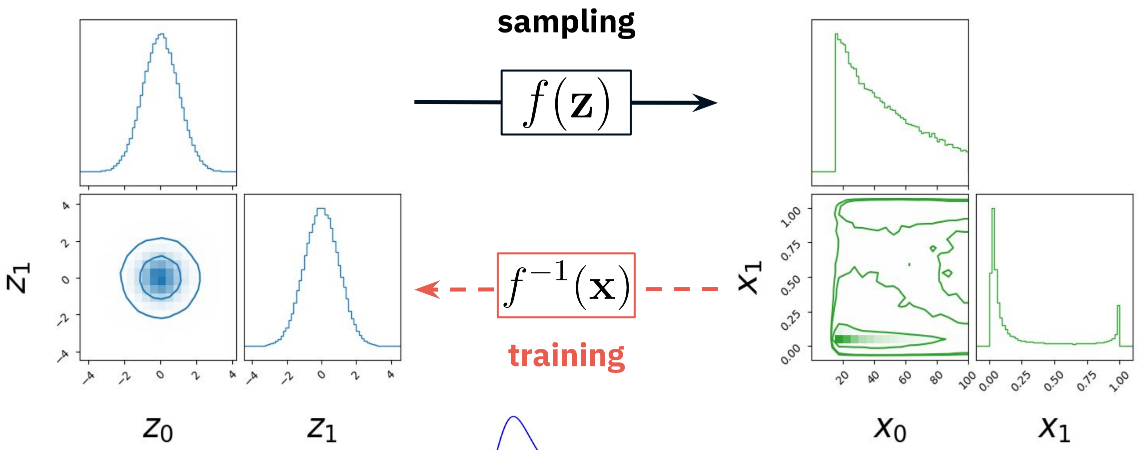
- Fast(er): reached ~kHz!
- Not analysis specific
- Depending on Gen (not just a generic event but the event)

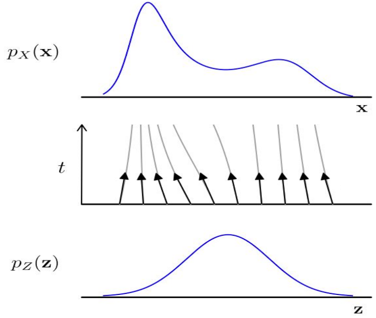


Continuous Normalizing Flows are the backbone of our approach!

We learn an invertible transformation, taking us from data *x* to noise *z*

Once *f* has been found we can invert it, start from noise and sample new data from the unknown PDF!





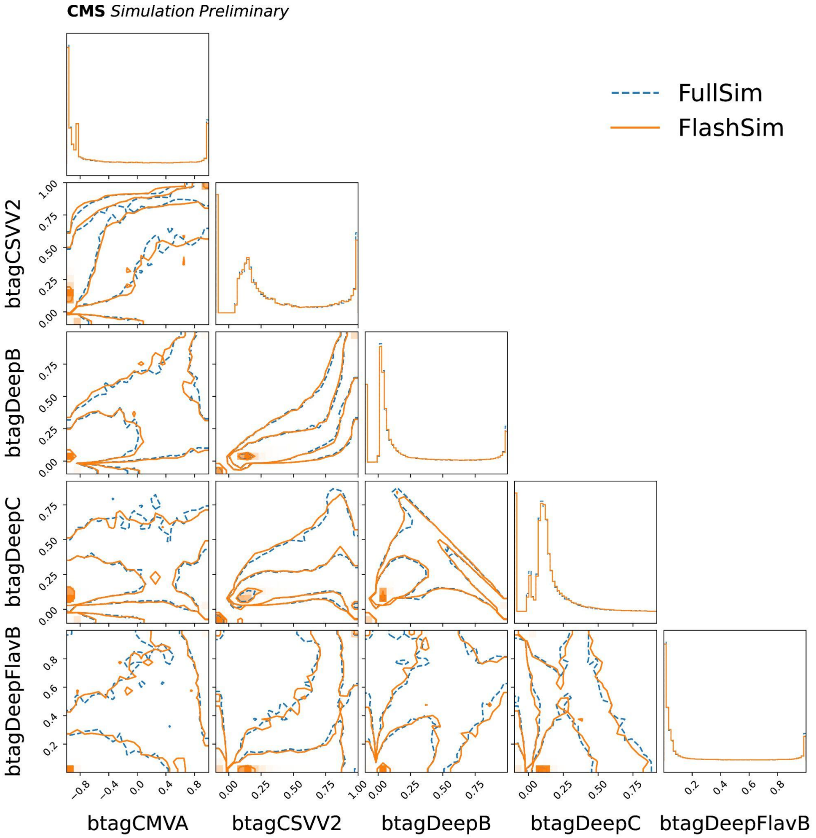
see https://arxiv.org/abs/2210.02747, andhttps://arxiv.org/abs/2302.00482, figure from https://ehoogeboom.github.io/post/en_flows/

Results are convincing

Simulation speed per object is around 10 kHz.
Our results accurately reproduce the Full Simulation data of the CMS Experiment, on both training and unseen processes, for:

- 1-d distributions;
- correlations between the variables;
- different physical processes;
- analysis-level plots.

For more:
Poster Session B, Thursday, Loc 70



refs: https://cds.cern.ch/record/2858890, and https://arxiv.org/abs/2402.13684

A Neural-Network-defined Gaussian Mixture Model for particle identification in LHCb

Giacomo Graziani (1), Lucio Anderlini (1), Saverio Mariani (1, 2, 3), **Edoardo Franzoso** (4), Luciano Libero Pappalardo (4,5), Pasquale di Nezza (6)

(1) INFN Sezione di Firenze, Florence, Italy

(2) Università degli studi di Firenze, Florence, Italy,

(3) European Organization for Nuclear Research (CERN), Geneva, Switzerland,

(4) INFN Sezione di Ferrara, Ferrara, Italy,

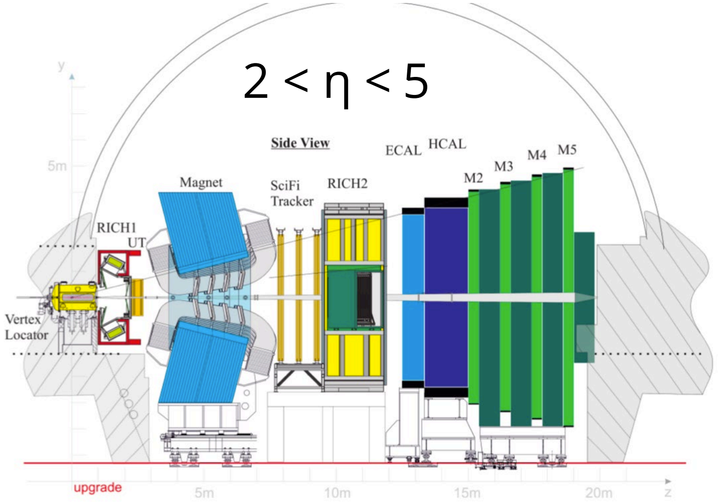
(5) Università degli studi di Ferrara, Ferrara, Italy,

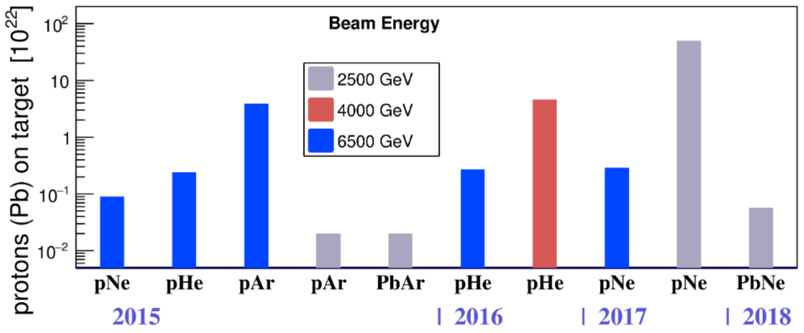
(6) INFN Laboratori Nazionali di Frascati, Frascati, Italy

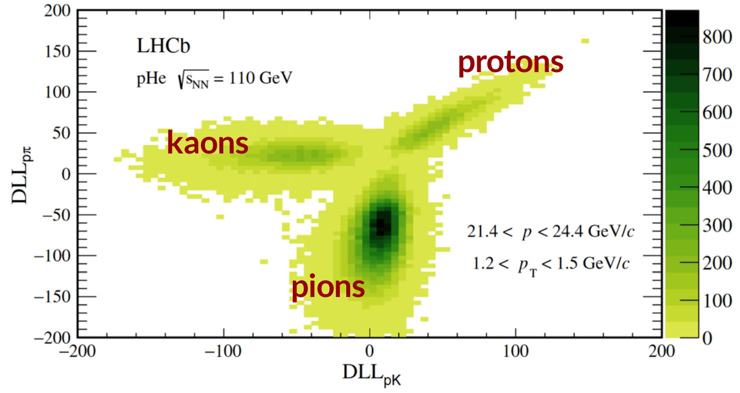
Motivation and idea

- **LHCb** relies on the Ring Imaging Cherenkov (**RICH**) detector system for the **charged hadron identification** (PID) in a wide momentum range (2 - 100 GeV/c)

- The PID was one of the dominant systematics int the measurement of cross-sections with fixed-target datasets in LHCb:







- This motivated the development of a **novel approach to the modelling of particle identification classifiers** using machine-learning techniques

  - **Explicitly model the marginal probability density function** (pdf) of the PID classifiers → must **depend on the experimental features** θ

  - Extract the marginal pdf using a **Gaussian Mixture Model**, whose parameters are predicted by Multi Layer Perceptrons trained on calibration data

  - **Model each hadron type** h (π, K and p) PID response independently using the appropriate control modes

Model and Validation

- In the **bidimensional PID space** the pdf x is defined as: $\underline{x}_h \sim \sum_{j=1}^{N_{g,h}} \alpha_{j,h}(\underline{\theta}) \frac{\exp(-\frac{1}{2}(\underline{x}_h - \underline{\mu}_{j,h}(\underline{\theta}))^T \Sigma_{j,h}^{-1}(\underline{\theta})(\underline{x}_h - \underline{\mu}_{j,h}(\underline{\theta})))}{2\pi\sqrt{\det(\Sigma_{j,h}(\underline{\theta}))}}$

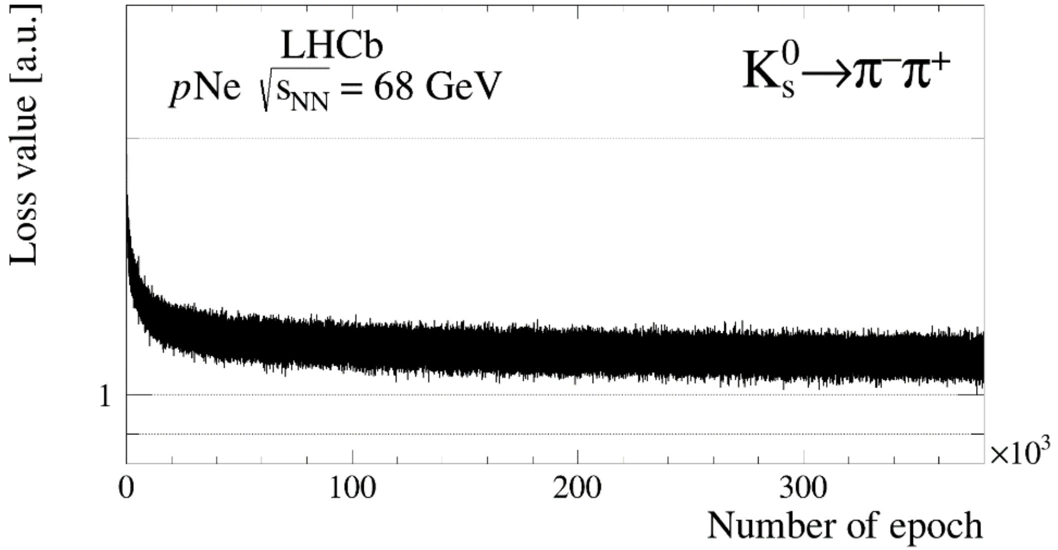
Training on proton-Neon sample

- The learning process of the networks relies on the **minimization** of the loss function, defined as the **negative log-likelihood**

$$\mathscr{L} = -\sum_{i=1}^{n_h} w_i \log\left[\sum_{j=1}^{N_{g,h}} \alpha_{j,h}(\underline{\theta}_i)\mathscr{G}(x_i, \mu_{j,h}(\underline{\theta}_i), \sigma_{j,h}(\underline{\theta}_i))\right]$$

- Training performed with **mini-batches gradient descent** with a user-defined number of Gaussians and NN structure

Input parameter	$K_S^0 \to \pi^-\pi^+$	$\overline{\Lambda} \to \overline{p}\pi^+$	$\phi \to K^-K^+$
Number of Gaussians	64	20	64
Number of NN nodes	128	128	128
Starting learning rate	10^{-3}	10^{-4}	$5 \cdot 10^{-6}$
Batch size	10000 events	10000 events	20000 events

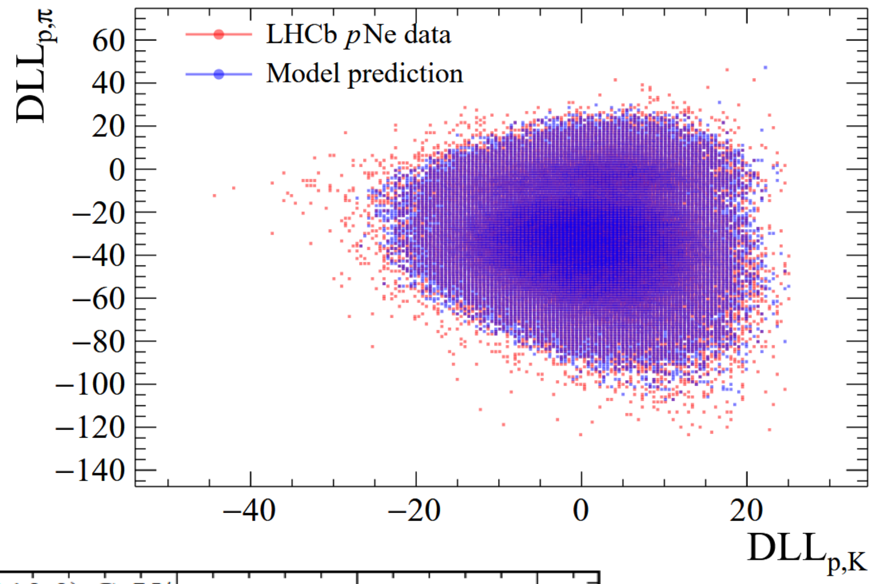


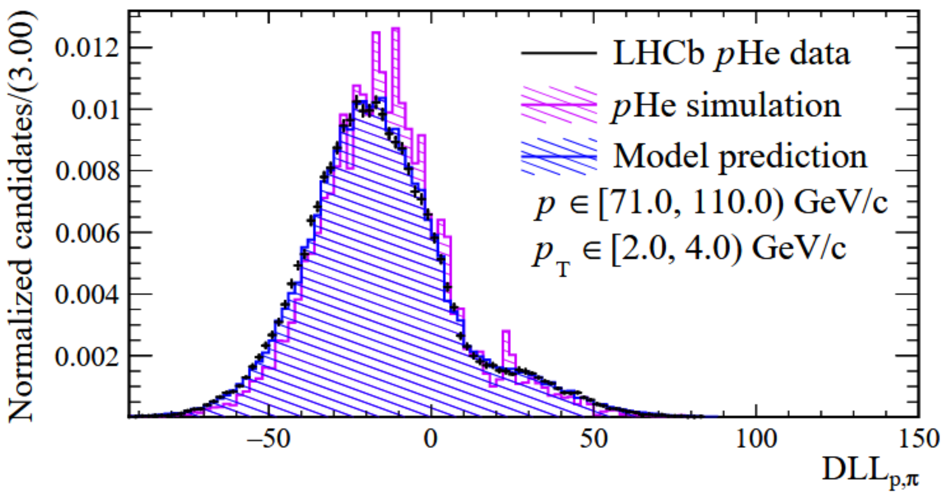
Paper : https://iopscience.iop.org/article/10.1088/1748-0221/17/02/P02018/meta

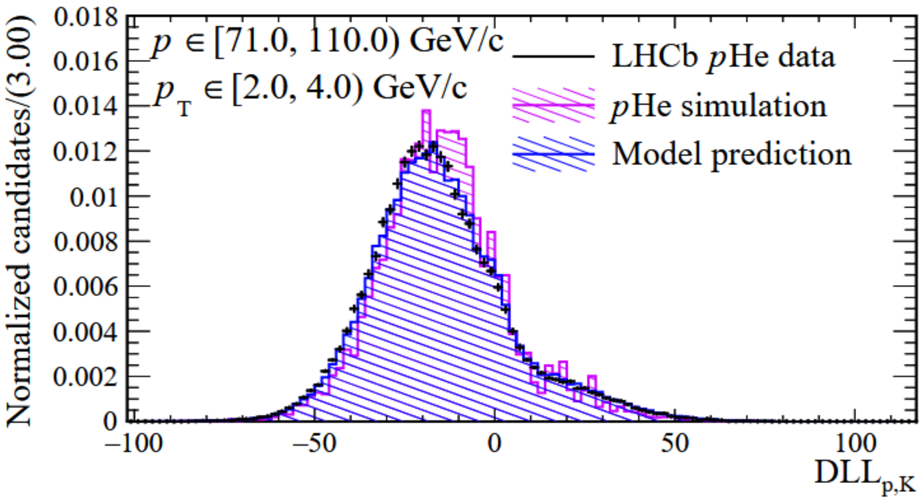
Model and Validation

Validation on <u>fixed-target</u> datasets

- Marginal-pdf prediction validated for the same training dataset and then **applied to two independent lower-statistics samples** of proton-Helium and proton-Argon collisions
    - **significant modifications of the kinematic distributions** of the produced particles
    - **different events multiplicity** and the detector occupancy
    - **identical detector and data-taking conditions** can be assumed
- **Data description quality improved by the model** compared to simulation
- The method is generic, not relying on a specific set of experimental feature variable







Paper : https://iopscience.iop.org/article/10.1088/1748-0221/17/02/P02018/meta

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

POLITECNICO MILANO 1863

CartCasLab

INFN Istituto Nazionale di Fisica Nucleare

Validating Explainable AI Techniques through High Energy Physics Data

Mariagrazia Monteleone[1], Francesca Camagni[1], Simone Gennai [2], Pietro Govoni [2,3], Chiara Paganelli[1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

[2] INFN - Sezione di Milano Bicocca, Italy

[3] Università degli Studi Milano Bicocca, Milano, Italy

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE

EuCAIFCon 2024

Background

Within the Compact Muon Solenoid (CMS) Collaboration, various Deep Neural Networks (DNNs) and Machine Learning (MLs) approaches have been employed to investigate the production of a new massive particle that undergoes decay into Higgs Boson pairs (HH) which further decay into a pair of b-quarks and a pair of tau leptons and discriminate the HH signal from the backgrounds.

However, the mentioned models which are employed are often **complex** and considered black boxes, making it challenging **to interpret** how the task was performed and the data analysis review process.

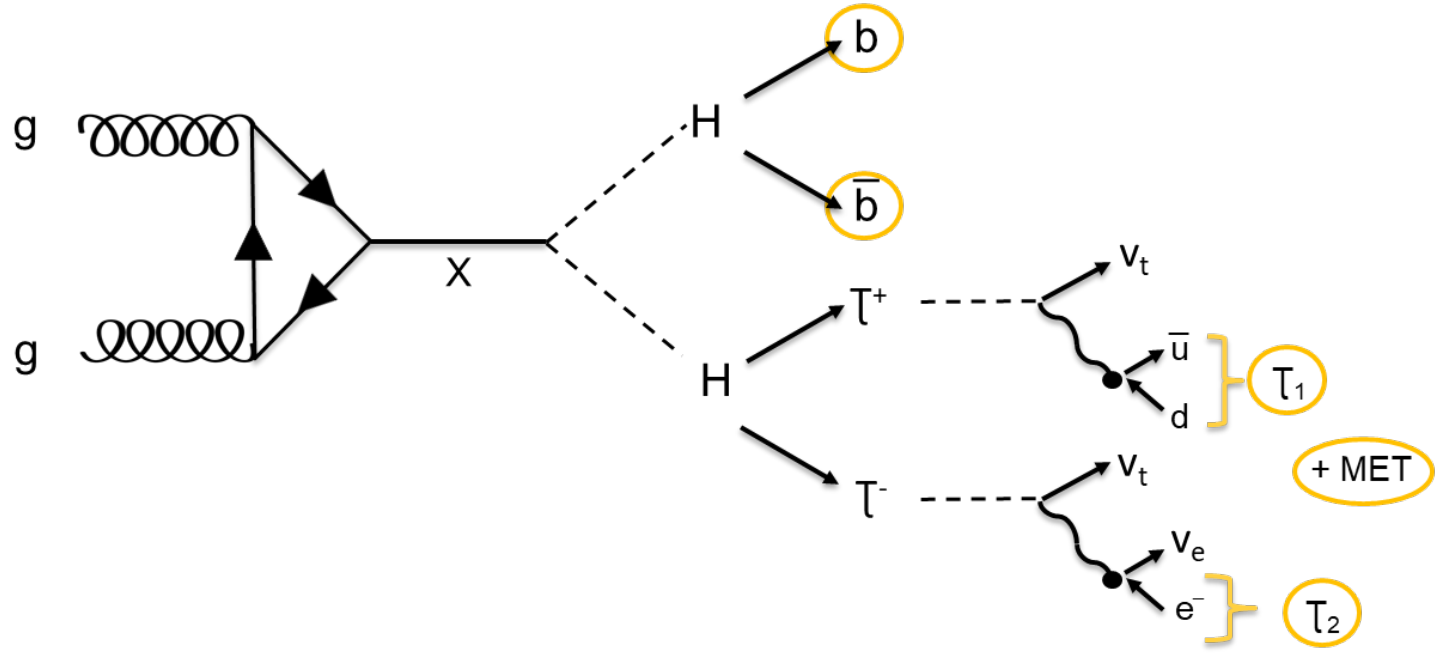


Fig. 1: Decay process outline

Aim of the work

This work aimed therefore to provide **a better understanding** of how the models work by validating an established **Explainable Artificial Intelligence (XAI)** technique such as SHapley Additive exPlanations (SHAP) [1], aiming for **more interpretable, trustworthy models** and predictions.

Workflow

A data pre-processing pipeline was established to select important features Recursive Feature Elimination based on SHAP values. This led to fine-tuning XGBoost for a classification task, whose features were compared with PCA results for validation and interpretation.

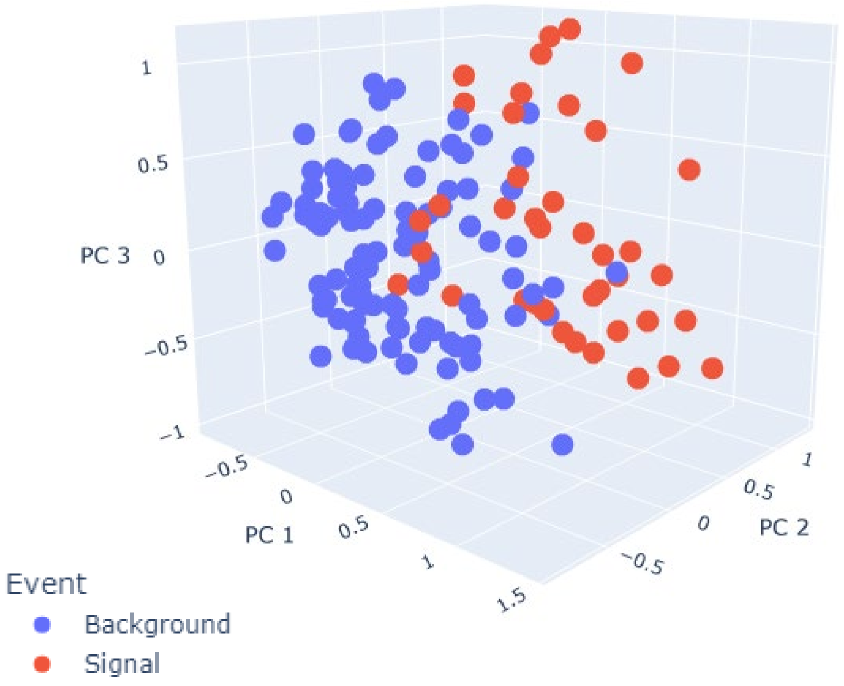


Fig. 2: PCA data visualization

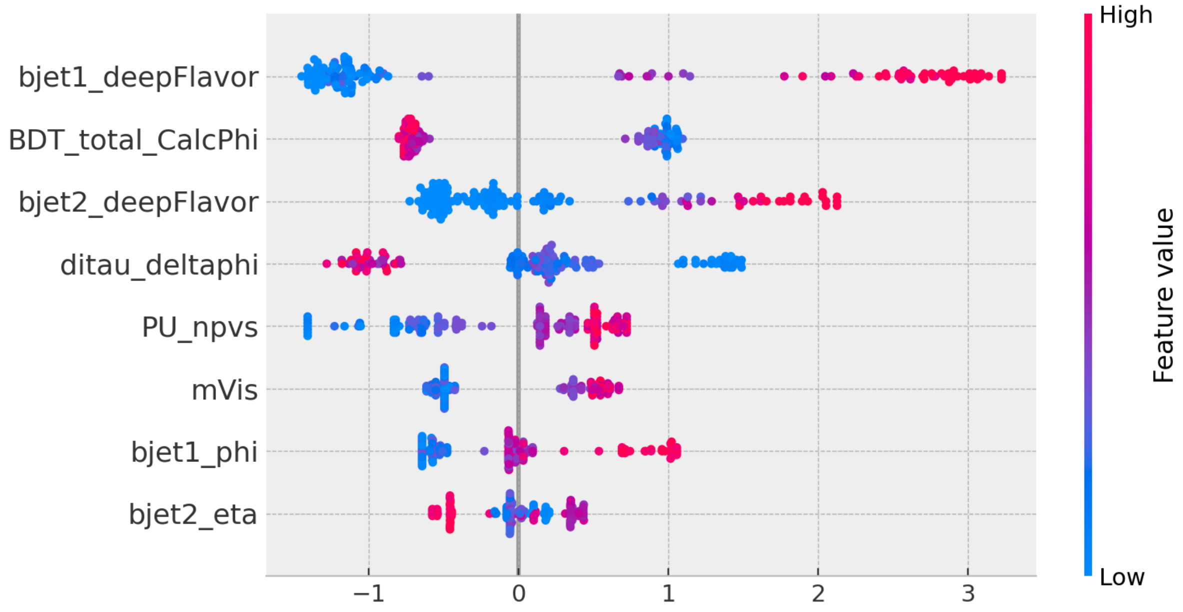


Fig. 3: SHAP value, impact on model output

Conclusions

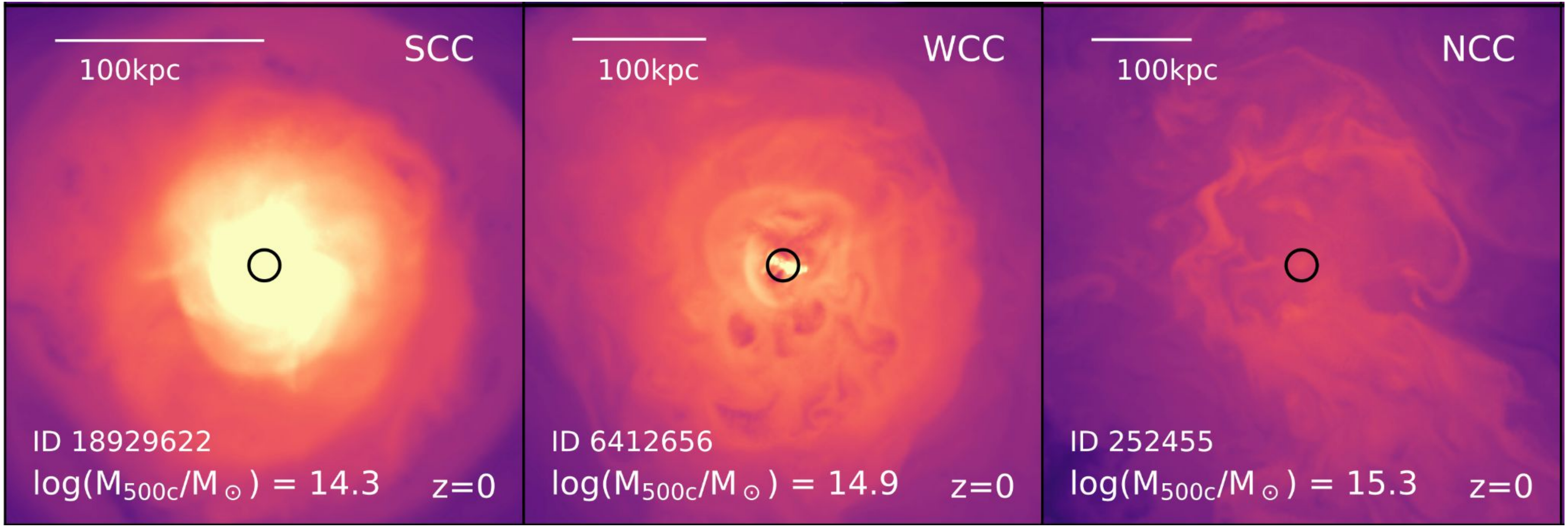
The results obtained with SHAP and PCA agreed on the importance of some of the features , the combination of the two techniques confirmed the reliability of SHAP as an established tool, but also the potential of High Energy Physics (HEP) domain as a **new technical validation tool.**

Understanding galaxy clusters with Contrastive Learning

Urmila Chadayammuri
EUCAIF 2024 | Amsterdam

Clusters are very diverse

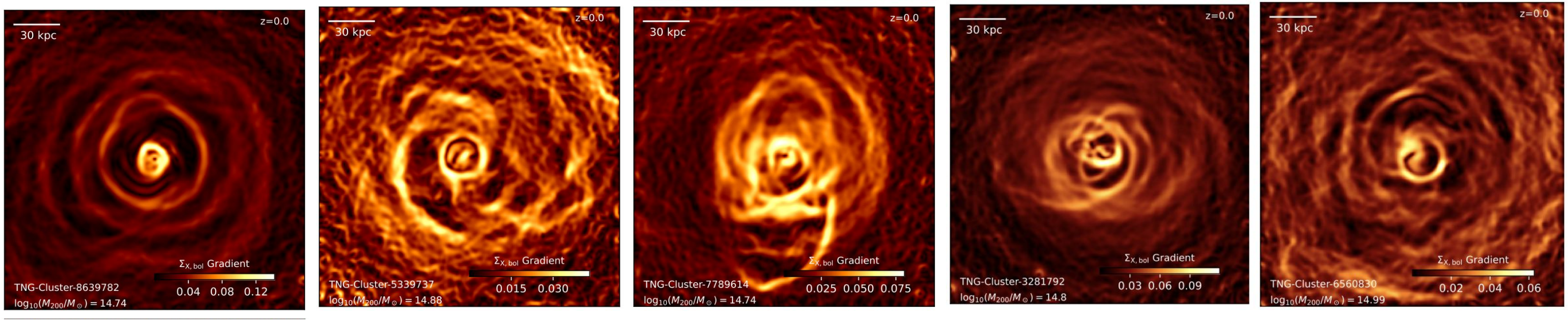
They have a wide variety of core thermodynamic profiles



Lehle+ 2024

Clusters are very diverse

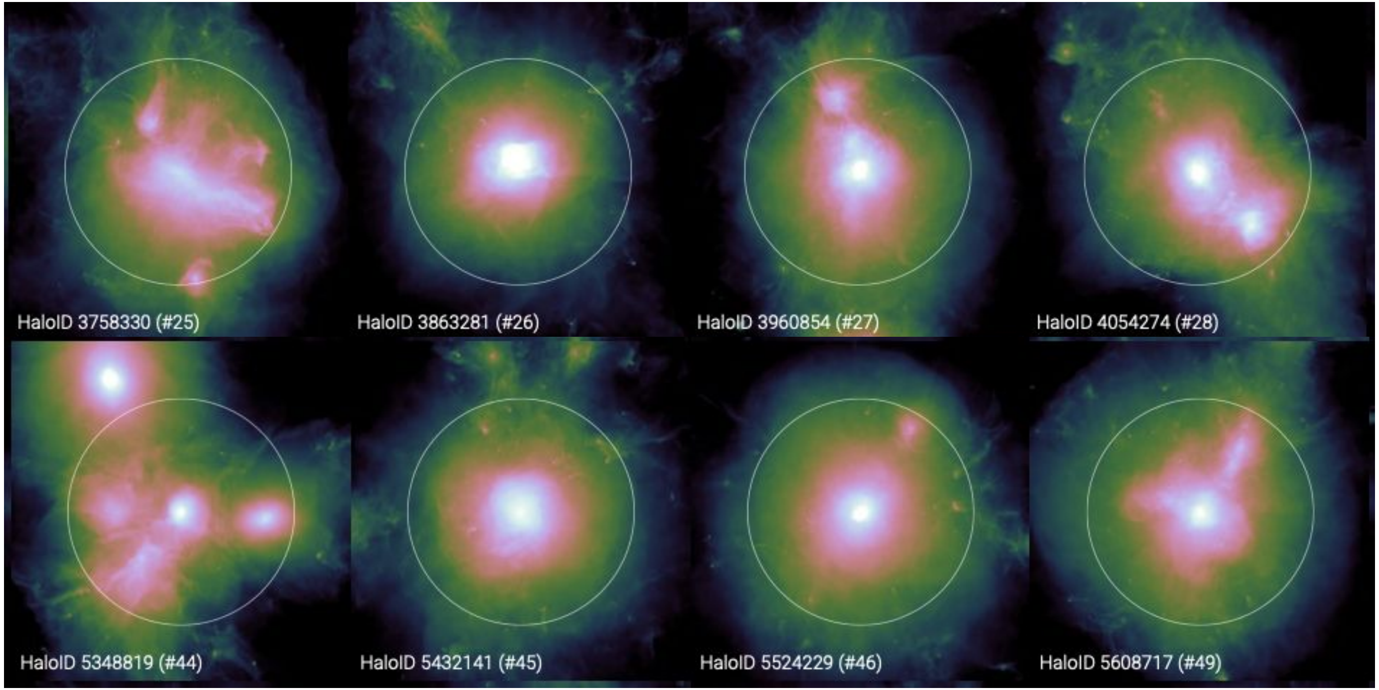
They have a wide variety of histories of AGN activity, in different phases



Truong+ 2024

Clusters are very diverse

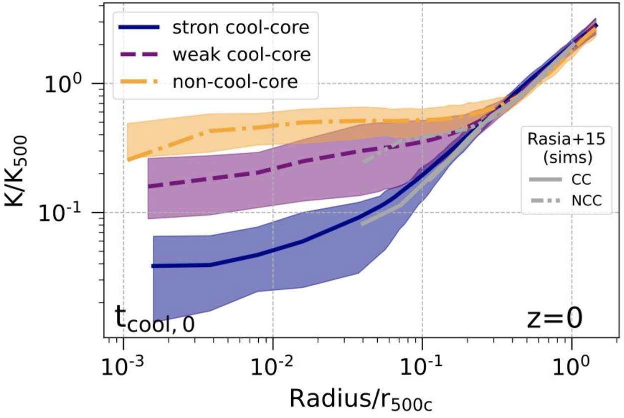
They can be in very different stages of merging



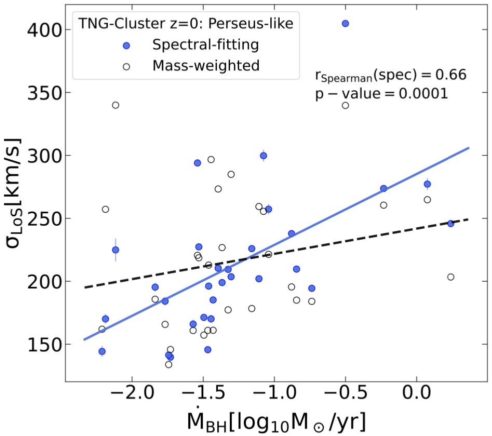
Nelson+ 2024

Clusters are very diverse

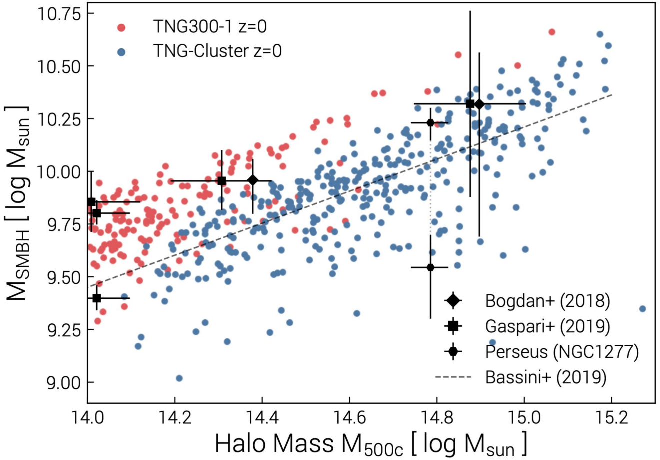
We try to characterise them in scalars or azimuthally averaged profiles



Lehle+ 2024



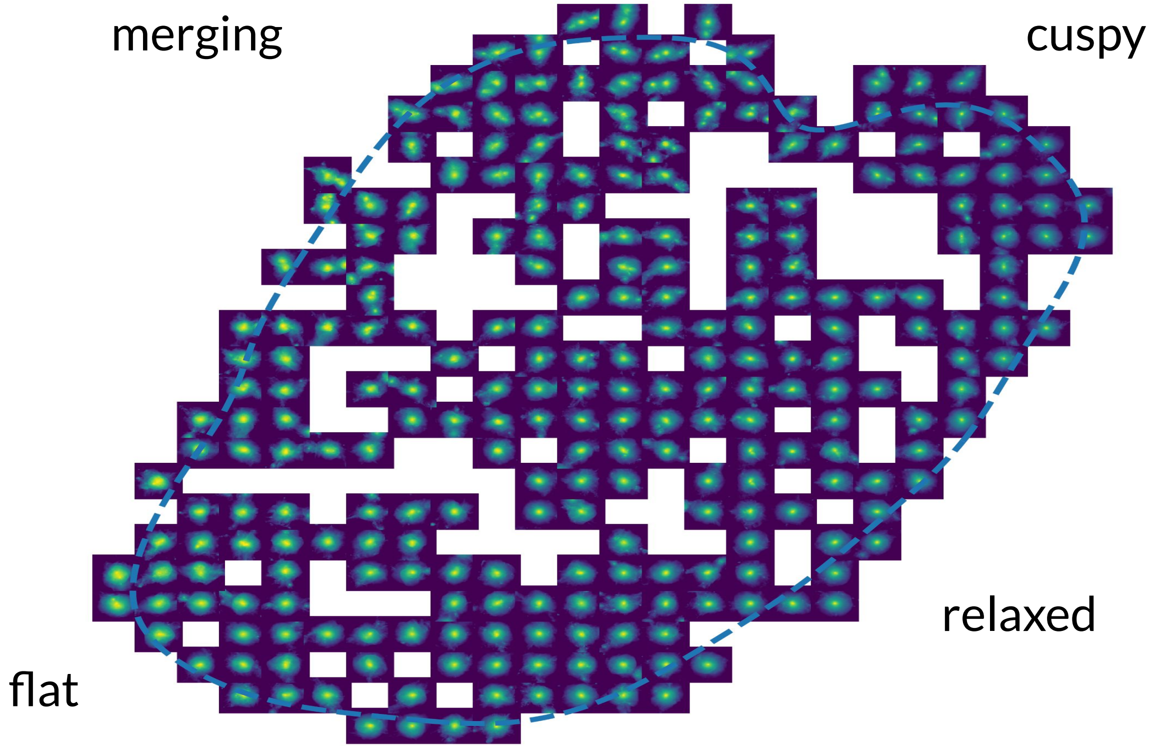
Truong+ 2024



Nelson+ 2024

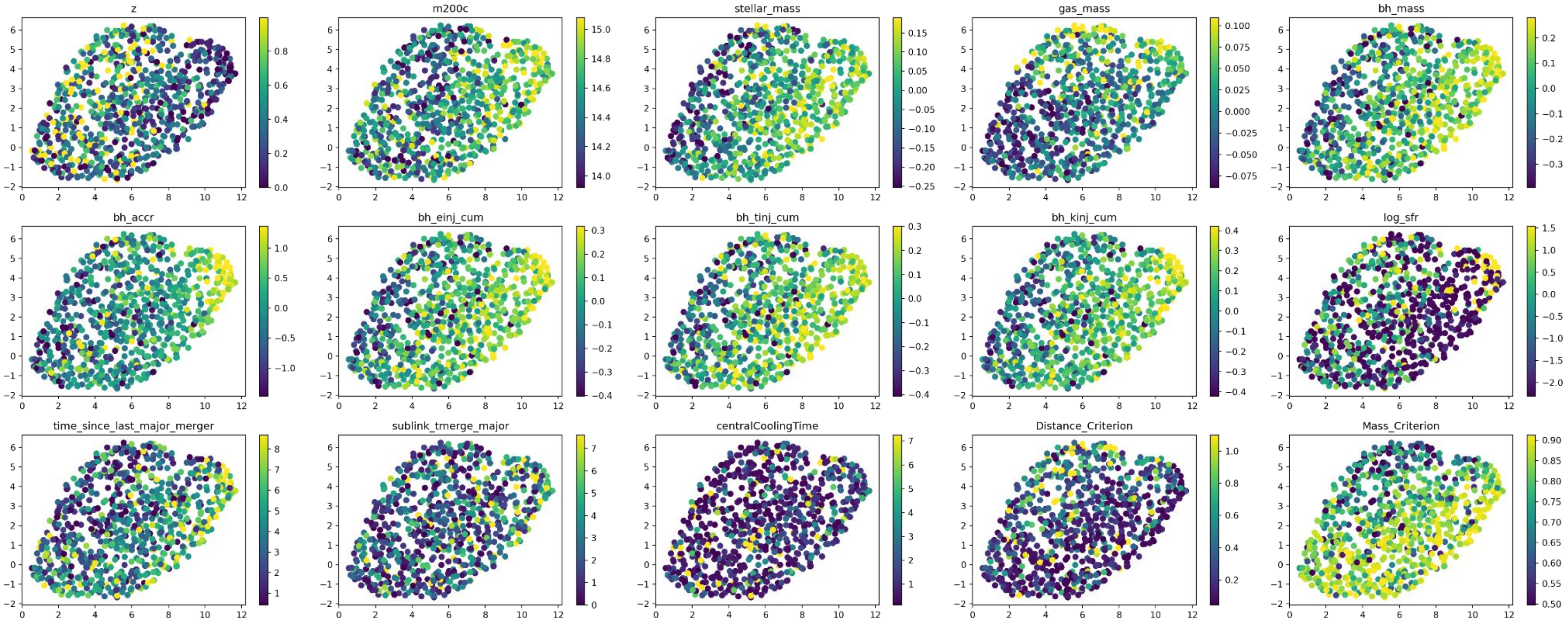
So what can we learn from cluster images?

2. We can sort the cluster images by similarity and identify cluster populations in image space



So what can we learn from cluster images?

2. The image-based sorting retains a lot of information about cluster mass, feedback and merger history - *even after correcting for halo mass*



Understanding correlations between feedback, mergers, and cluster cores

2. The image-based sorting retains a lot of information about cluster mass, feedback and merger history - <u>*even after correcting for halo mass*</u>

The median values of feedback, merger and cool-coreness metrics <u>are more correlated with each other</u> in similar regions of the representation space (bottom) than without image-based sorting (above)