Université Paris Cité

MACHINE LEARNING FOR ASTROPHYSICS

3RD EDITION VALLETTA, 31 AUG - 4 SEP, 2026

APC

Machine Learning and Deep Learning in Astroparticle Physics:

From Cosmic Rays to Neutrino and Gamma-Ray Astronomy

Yvonne Becherini, Laboratoire Astroparticule et Cosmologie,
Université Paris Cité

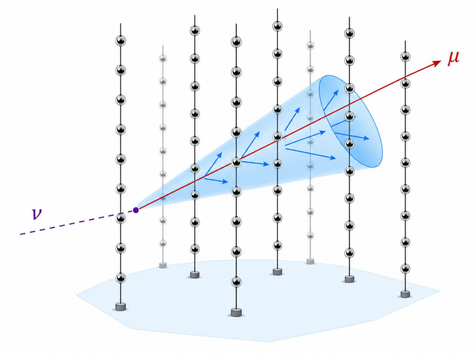
Astroparticle physics: ML gains start at the detector level

Extreme backgrounds ➔ Focus on the low-level reconstruction and rejection part

ASTROPARTICLE PHYSICS

Often S/B ≪ 1 already at event level

γ/hadron • ν/μ • rare showers



→ **Detector-level observables** → **ML / DL reco + rejection** → **Clean(er) physics sample** → **Source / particle parameters**

pixels/hits/ waveforms

Sensitivity Gain: 1) reduce residual background while retaining signal ⇒ S ≍ Nsig / √Nbkg
2) Angular and energy resolution

OBSERVATIONAL ASTROPHYSICS

Often starts closer to astronomical observables

images • spectra • light curves • catalogues

Observable → Physical inference

Astroparticle ML historically enters earlier in the analysis chain: at reconstruction, classification and event selection.

Astroparticle physics is an inverse problem

Observed

pixels/light/hits

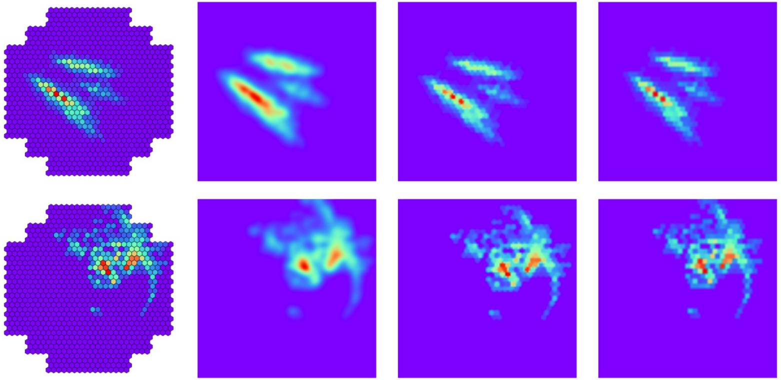
→

Wanted

physical quantities

Particle type
Energy
Direction

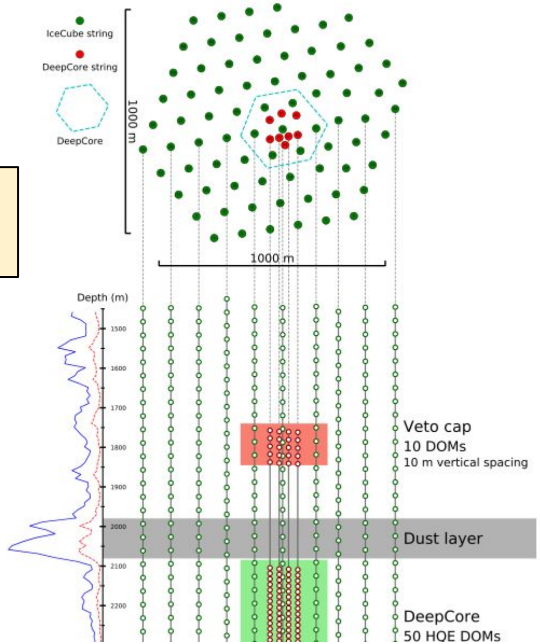
IACT camera
image-like shower view



The detector determines the useful ML representation

Shilon et al. 2019, arXiv:1803.10698;
Abbasi et al. 2022, arXiv:2209.03042.

neutrino telescope
sparse 3D hits



CNNs are natural for camera images; GNNs and transformers become natural when the detector is irregular, sparse or time-resolved.

Evolution of the ML role: deeper information extraction

cuts / likelihoods	RF / BDT	CNN / GNN / transformers	posteriors / SBI
engineered observables	classical multivariate classification/regression	detector-level representation learning	calibrated uncertainty for physics inference

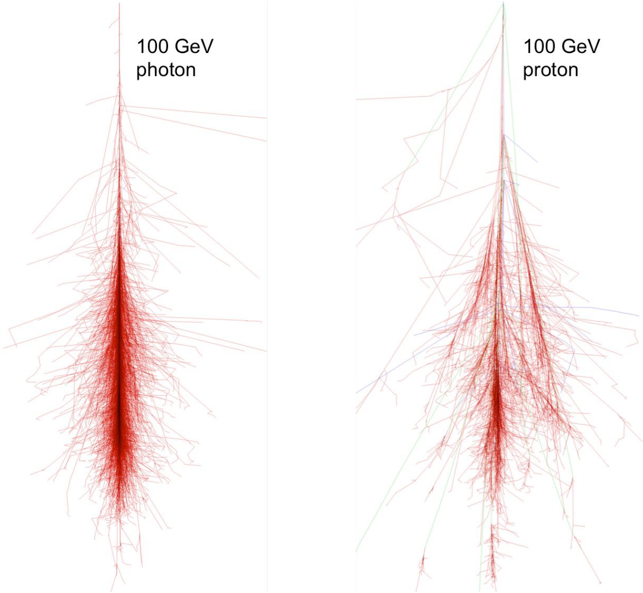
ML evolves from “classify events” to “extract latent physical information with uncertainties”.

What becomes possible?

sensitivity gains
combine observables;
reduce background;
faint-source reach

hidden physical information
Shower maximum;
images/graphs learn
morphology and timing

Shower development: Gammas vs Protons & Observational Techniques

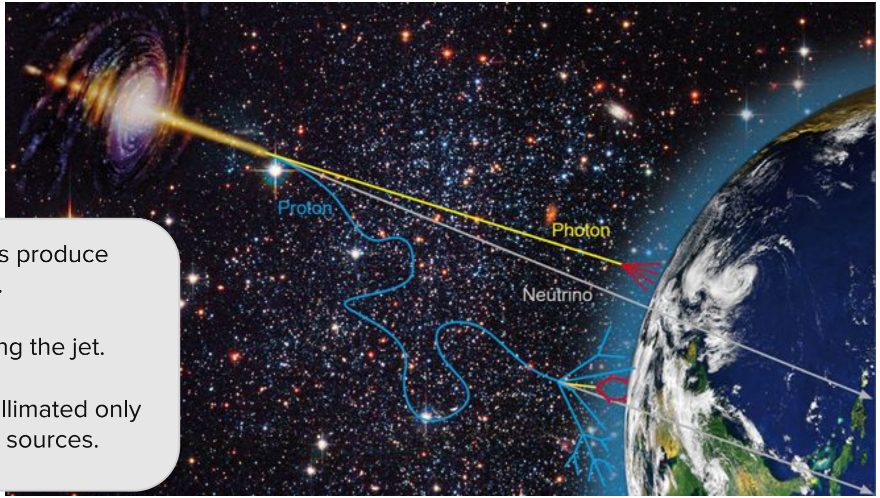


Electromagnetic Showers (pair production in the Coulomb field of an atmospheric nucleus and bremsstrahlung)

- Highly beamed, compact, symmetrical, and extremely muon-poor.

Hadronic Showers (Hadronic pN interactions)

- Complex multi-particle networks.
- Wider lateral footprint with significant muon, pion, and hadronic core components.



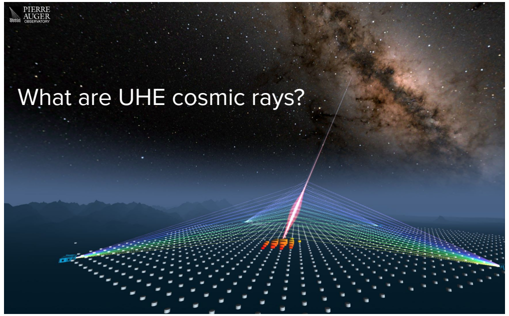
Protons accelerated at cosmic-ray sites produce secondary ν and γ-rays via pion decay.

Neutrinos and γ-rays stay beamed along the jet.

Protons are magnetically deflected, collimated only at the highest energies and for nearby sources.

Research infrastructures: detector response and physics goals

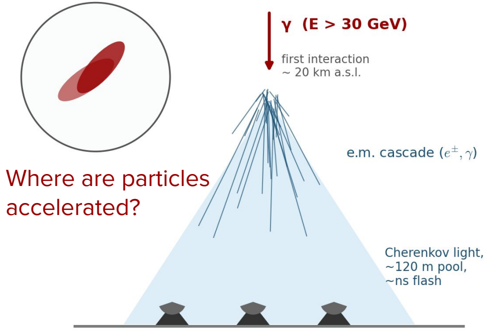
Pierre Auger Observatory



UHECR air showers: hybrid surface + fluorescence detector

hybrid cosmic-ray air-shower observatory

Imaging Atmospheric Cherenkov Telescopes



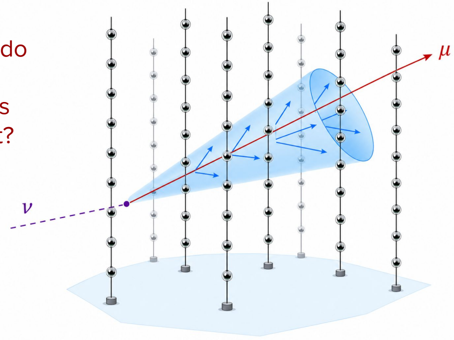
stereo IACT array — image the shower, reconstruct direction & energy

TeV γ rays: atmosphere as calorimeter, cameras as fast imagers

H.E.S.S. / MAGIC / VERITAS → CTAO

Neutrino telescopes

Where do cosmic hadrons interact?



cosmic ν: sparse 3D arrays in ice or deep sea water

IceCube / KM3NeT / Baikal-GVD and next generation

What is measured

- ground footprint + ns traces
- fluorescence profile
- Xmax, energy, muon/EM content

- ns Cherenkov images
- stereo → direction, impact, energy
- morphology → hadron rejection

- Cherenkov light in ice / sea
- hits: position, time, charge
- track vs cascade topology

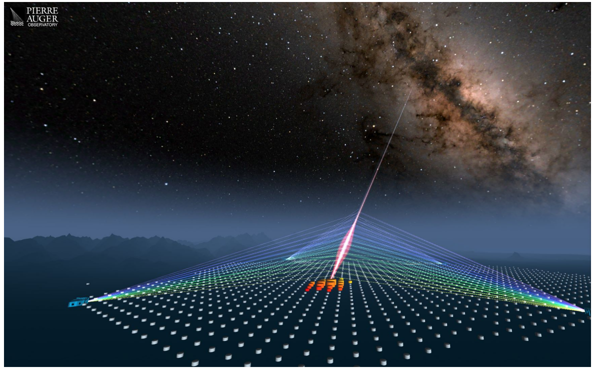
Physics goals

- composition above 10^{18} eV
- Galactic → extragalactic
- UHE photons and neutrinos

- TeV–PeV accelerators, PeVatrons
- leptonic vs hadronic emission
- AGN, SNRs, PWNe, EBL, DM

- astrophysical ν sources
- hadronic signatures of CRs
- γ–ν–CR connection

Cosmic rays: learning the shower from the ground



Surface arrays sample the shower at ground level:

- near-100% duty cycle, huge exposure
- indirect access to **mass composition**
- timing traces carry shower-development information

What Auger/TA-like arrays measure

- lateral shower footprint
- time traces in each station
- signal risetime and curvature
- geometry + energy proxies

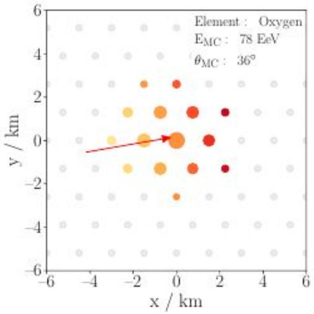
What are ultra-high-energy cosmic rays made of, and how does their composition evolve with energy?

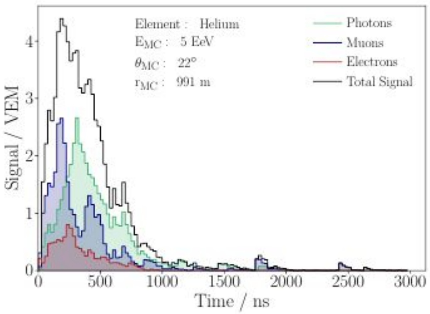
- The key observable is the atmospheric depth at which the shower reaches maximum development (X_max).
- X_max is strongly composition-sensitive
- X_max is measured with the fluorescence detector, but this has only about 15% duty cycle, while the surface detector operates nearly continuously.

Case-study question: can surface-detector traces recover information normally associated with fluorescence Xmax?

Figure: Erdmann, Glombitza & Walz, Astropart. Phys. 97 (2018), arXiv:1708.00647. Context: Pierre Auger Coll. 2021, arXiv:2101.02946.

Auger: reconstruction of shower maximum (Xmax) from surface-detector traces





LSTM

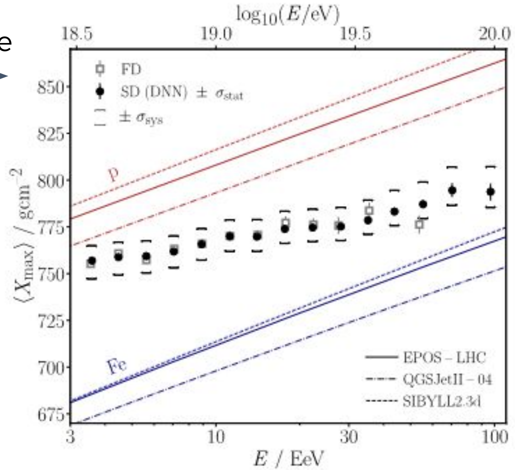
learns time-trace features per WCD station

hexagonal CNN

uses triangular/hexagonal array symmetry

Pierre Auger Coll., JINST 16 (2021) P07019; Phys. Rev. Lett. 134 (2025) 021001.

close to the ground →



- LSTM → converts each station waveform into learned feature channels; encodes the temporal structure of each station trace
- hexagonal CNN → learns the spatial correlations of those features across the shower footprint.

≂10×

more statistics than FD Xmax sample

100 EeV

composition reach at extreme energies

Trained on simulations (MC ≠ data), calibrated on hybrid Fluorescence Detector data

Gamma-ray astronomy: from shower images to faint-source sensitivity

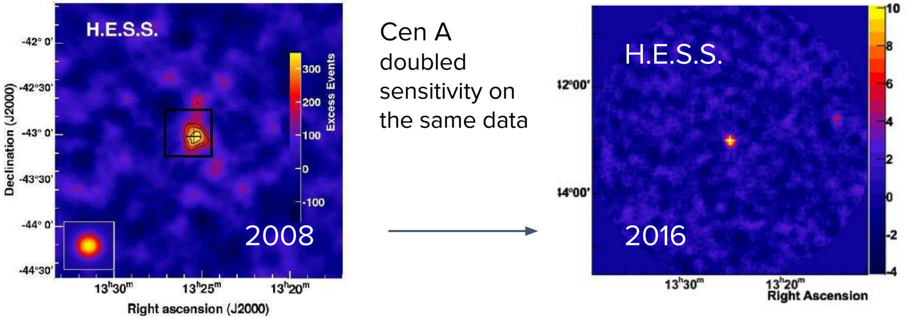
From a set of telescope images, we wish to infer the particle type, its energy and its direction

The dominant challenge is γ ≪ hadronic cosmic-ray background.

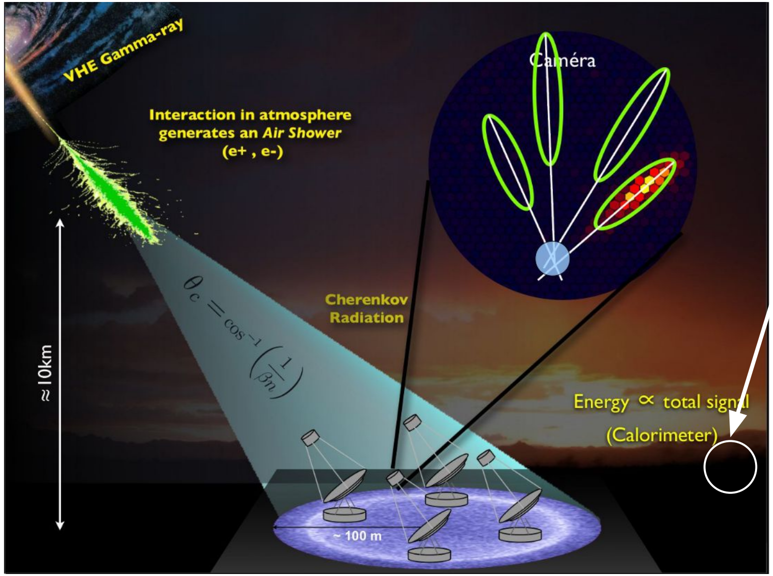
Classical approach

- Reduce each image to hand-engineered quantities:
    - width, length, intensity
    - impact-related parameters
    - stereo quantities
- Then use **Random Forests/BDTs/multivariate classifiers**.

Better gamma/hadron separation → better sensitivity to faint sources.



Cen A doubled sensitivity on the same data



Becherini et al, Astroparticle Physics, Volume 34, Issue 12, July 2011, Pages 858–870

What DL tries to exploit

camera pixels → **CNN** → γ/h E, dir

Beyond engineered features:

pixel-level morphology and timing

multi-telescope information fusion

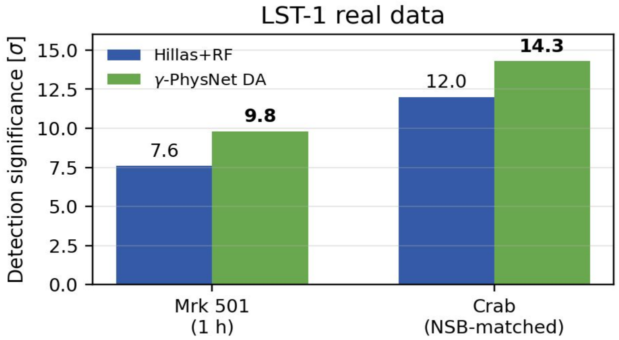
multi-task reconstruction: γ/h, energy, direction

The high bar:

DL competes not only with Hillas+RF, but with mature likelihood/template chains

So the relevant metric is not only AUC on MC, but stable gain on real data. (Domain shift)

Clear real-data example: LST-1 CTAO



Tangible observational gain, but depends on matching MC to data.

LST-1 real data: Vuillaume et al. 2021, PoS ICRC2021 703, arXiv:2108.04130 (Mrk 501); Dellaiera et al. 2024, arXiv:2403.13633 (Crab, NSB tuning).

Most supervised ML chains in astroparticle physics are trained on Monte Carlo simulations.

p_MC(x) ≠ **p_data(x)**

Sources of domain shift

- night-sky background and atmosphere
- detector ageing / calibration
- hadronic interaction models
- trigger and selection effects

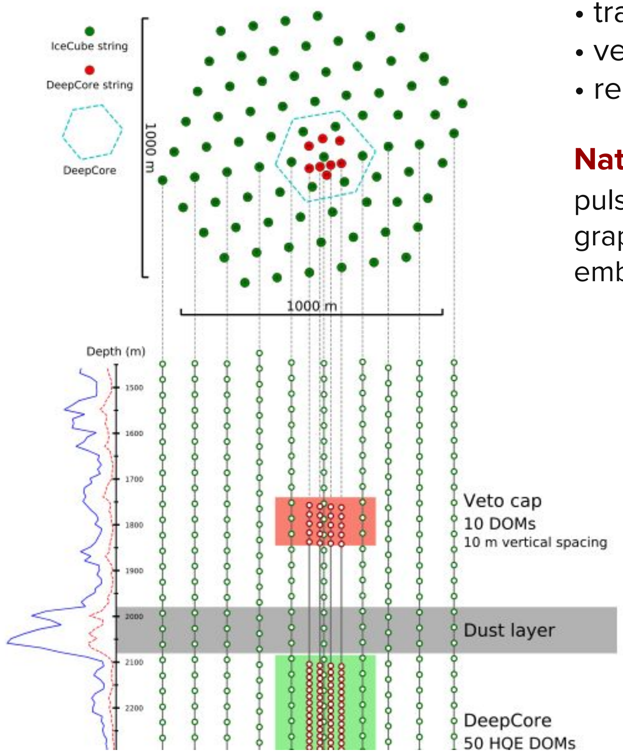
Methods entering the field

- data/MC validation on standard candles
- domain adaptation
- systematic-aware training
- uncertainty calibration

A major barrier to adoption in IACT analyses:
mature reconstruction chains are robust on real data, while DL gains can degrade under domain shift.

Neutrino telescopes: sparse 3D space-time patterns

ICECUBE NEUTRINO OBSERVATORY



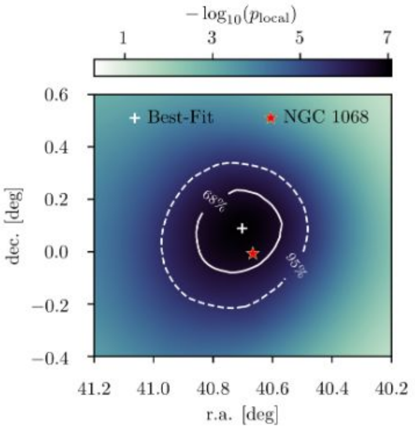
Different detector language, but same goals

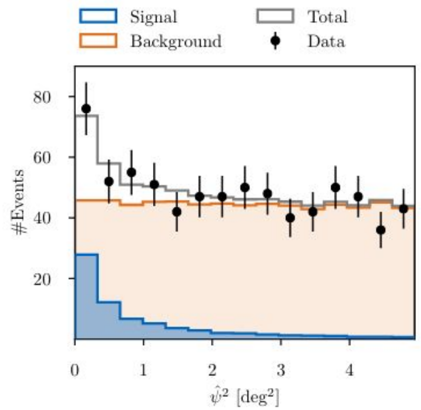
- atmospheric background rejection
- track /cascade classification
- vertex, direction and energy reconstruction
- real-time event selection for follow-up

Natural representations

pulses ➜
graph nodes/point cloud/time-aware embeddings

- Gradient boosted decision tree for angular errors (so standard ML)
- CNN for energy reconstruction (DL)





“Evidence for neutrino emission from the nearby active galaxy NGC 1068” Science, 2022

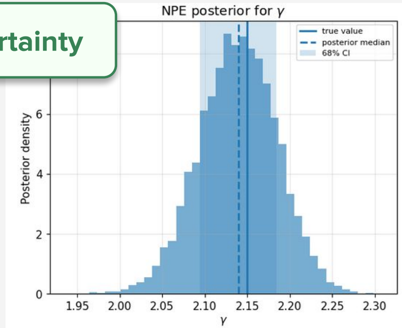
New frontiers

From point estimates to posteriors

classical output
E, direction, class

physics output
p(θ|detector data)

usable uncertainty



Pre-trained/foundation models for Data Filtering

- TabPFN for tabular data
- Fine-Tuned DINOv2 on H.E.S.S. images

Simulation-Based Inference

Representation Learning

raw detector hits

learned event embedding

classification

energy / direction

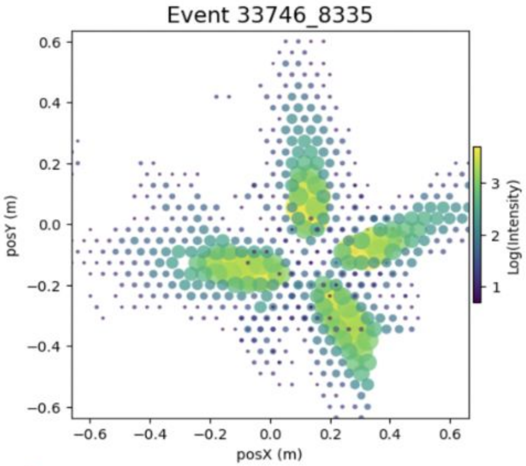
anomaly search

source association

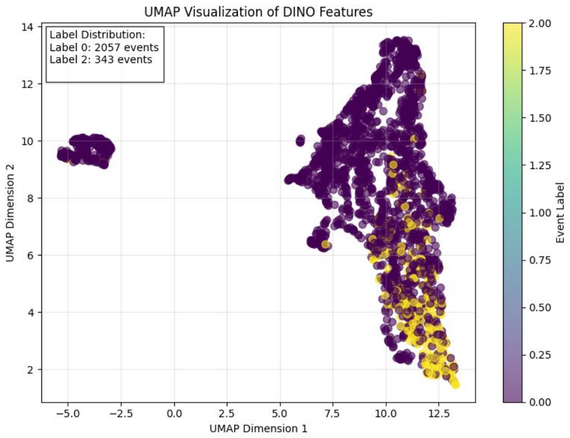
- Learned embeddings can be reused across analyses
- Promising for low-statistics and transient searches
- Still needs physics-aware validation

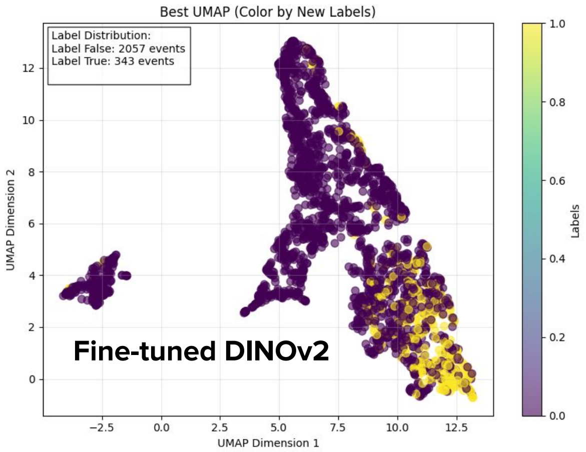
Applications of image-based foundation models to image classification in H.E.S.S.

DINOv2 on real H.E.S.S. data to see if we can distinguish gamma rays from hadronic showers with no supervision



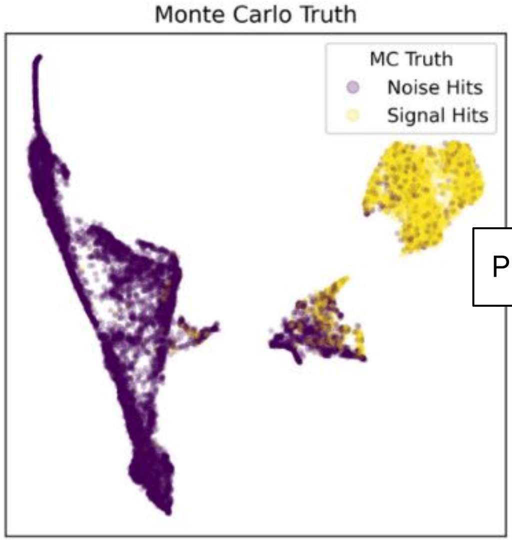
Frozen DINOv2 features



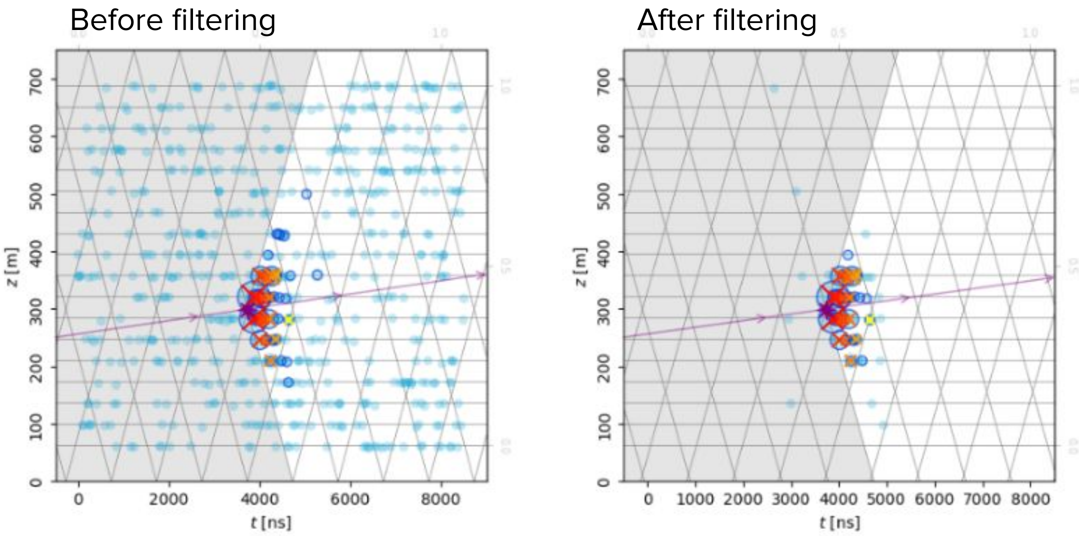


G. Aractingi, G. Jeannin
Master internships

- GNN to filter out K40 background from data to accelerate the track reconstruction procedure
- PMT hits are nodes (position, time, charge); physically motivated space-time compatibility defines the edges. **Sparse + irregular geometry → GNN.**
- Exploit unlabelled data/semi-supervised learning → reduce simulation dependence
- Binary classifier sees physics events and pure background events → the network learns the representation of the hits



Preliminary



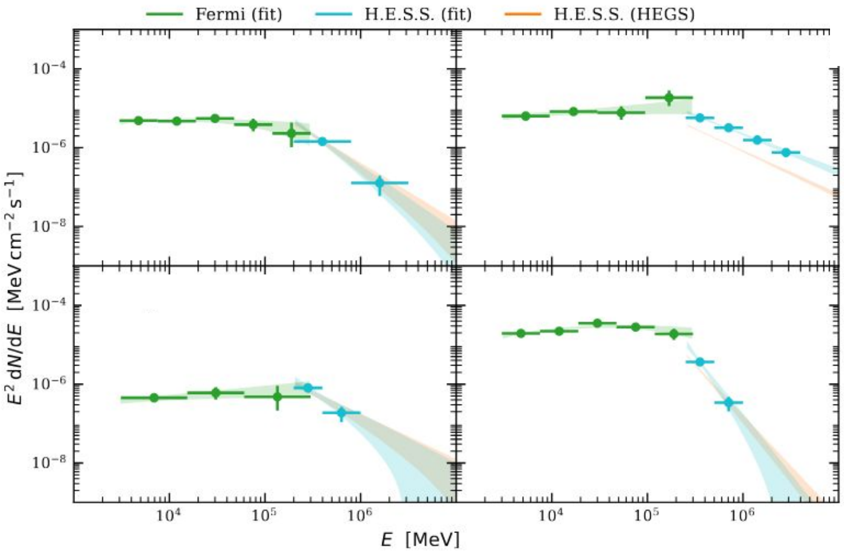
Max Eff, PhD Thesis

BL Lac Redshift Constraints from Inverse Compton Emission with Supervised ML

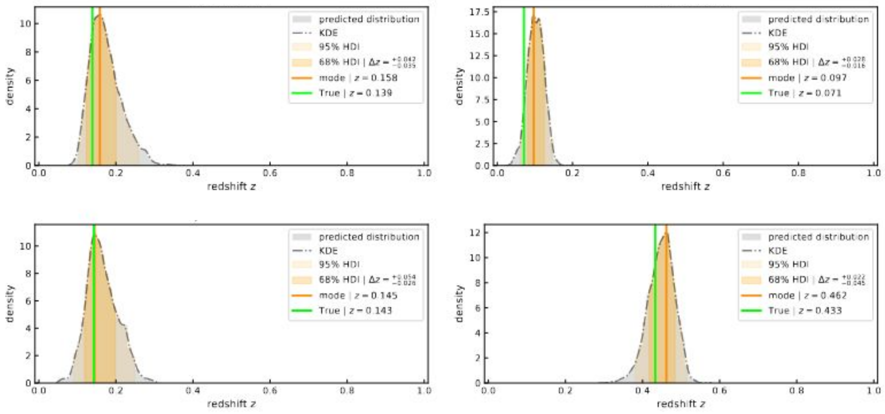
Goal: Estimate BL Lac redshifts from **EBL-attenuated inverse-Compton SEDs**, using supervised ML trained on a large set of physically motivated synthetic spectra.

Max Eff, PhD Thesis

Same inverse-problem logic, now at source level



Preliminary



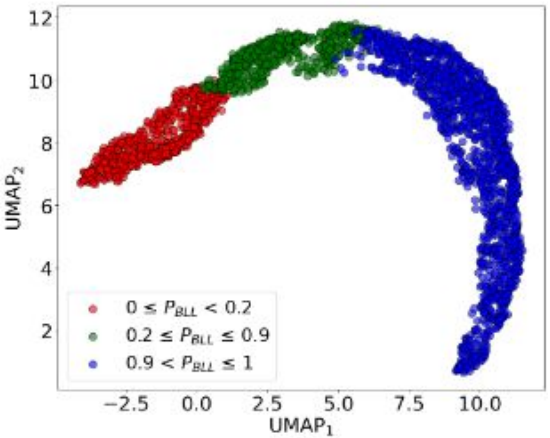
Simulation-based inference: the inverse mapping is learned from physically simulated SEDs, while the redshift posterior is obtained by Monte-Carlo propagation of the measured spectral uncertainties.

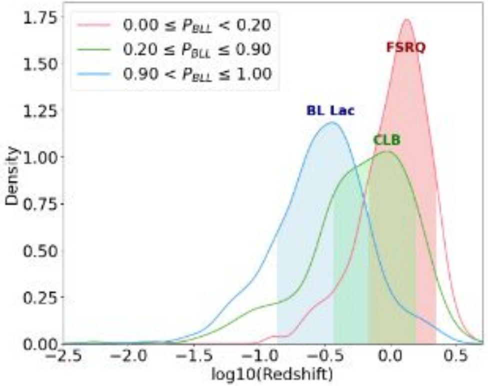
Next step: Neural SBI

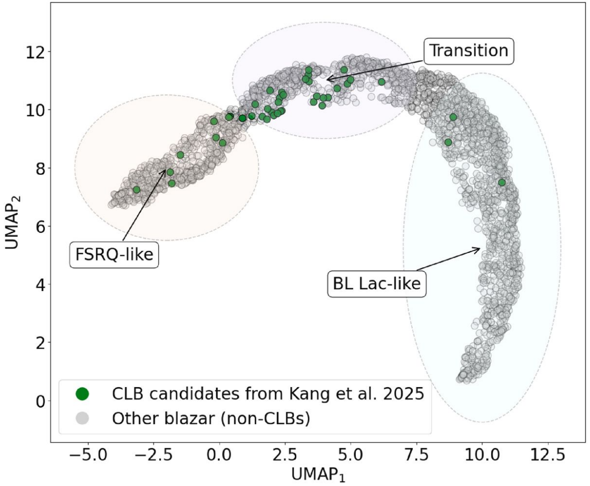
Blazar evolution beyond discrete classes

- Classification of blazars from the Fermi 4LAC-DR3 catalogue, focusing on blazars of uncertain type.
- TabPFN as a foundation model for tabular data.
- The result suggests a continuum between FSRQs and BL Lacs, with Changing-Look Blazars as possible transitional objects.
- Here the ML output is not only a label.
  A probability score plus latent-space structure becomes a way to discuss physical evolution.

Goal: Study blazar evolution through latent space investigation







Oukacha & Becherini, Towards a unified scheme of blazar evolution, A&A 2025

Detector level: ML as reconstruction

- **Auger:** DL recovers Xmax from surface traces, 10x the fluorescence statistics and composition reach to 100 EeV.
- **IACTs:** measurable gains on real data at low energies, but the baseline is mature likelihood chains and gains are sensitive to observing conditions.
- **Neutrino telescopes:** graphs and transformers match sparse 3D space-time data

Source level: the same inverse problem, one step up

- SEDs → redshift posteriors: simulation-based inference gives a distribution, not a point estimate.
- Catalogue features → latent space: classification becomes a way to discuss physical evolution, not just a label.
- Images → pre-trained foundation models: DINOv2 usable off the shelf where labels are scarce.

The common bottleneck is not architecture, it is the simulation gap

Every supervised chain here is trained on Monte Carlo. The interesting direction is reducing that dependence: domain adaptation, weak supervision from event-level labels, self-supervised pretraining on unlabelled data.

Thank you