Unsupervised tagging of semivisible jets with energy-based autoencoders in CMS

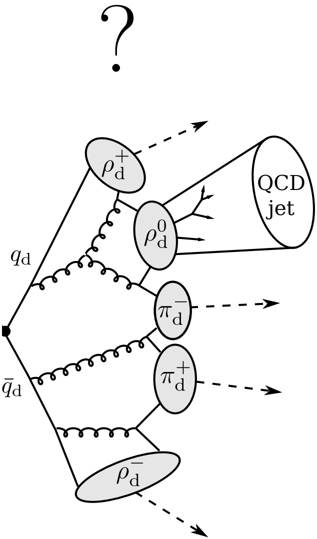
Florian Eble, on behalf of the CMS collaboration

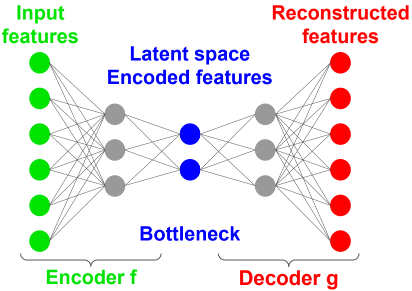
ETH zürich

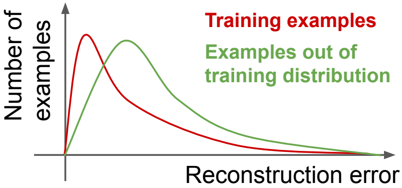
01/05/2024

EuCAIFCon 2024

Anomaly detection to search for semivisible jets (SVJs)



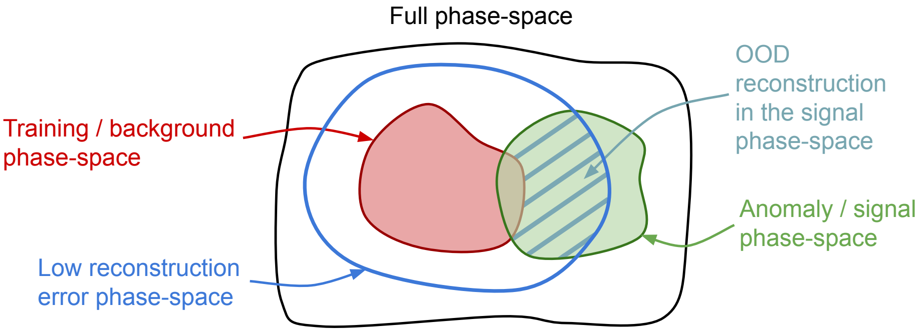




The problem of out-of-distribution (OOD) reconstruction

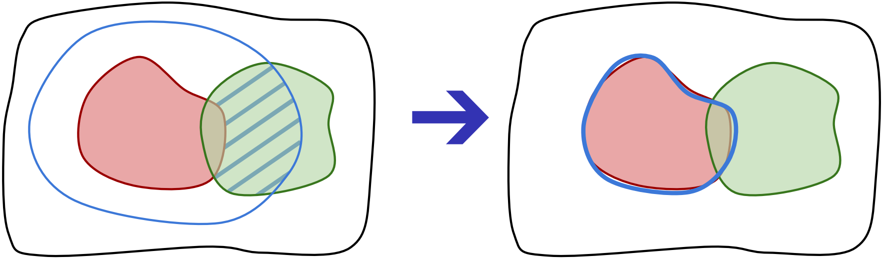
- **AEs are free to minimize reco error outside the background phase space!** including the unknown signal phase space...

➔ This is the problem of **OOD reconstruction**:



Normalized Autoencoders (NAE)

- NAE features a mechanism to suppress OOD reconstruction
- First introduced in arXiv:2105.05735 and used in HEP in arXiv:2206.14225



- NAE paradigm:
  - Define a probability distribution p_θ so that high probability regions have low reco error
  - Sample from p_θ via a MCMC
  - Minimize the distance between the background and p_θ probability distributions
- We propose a different metric to measure this distance, using the Earth Mover's Distance (a.k.a Wasserstein distance) and train NAEs in a fully signal-agnostic fashion

Multi-class classification of gamma-ray sources and the nature of excess of GeV gamma rays near the Galactic center

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Dmitry Malyshev

EuCAIFcon, Amsterdam, 30.04 – 03.05 2024
Poster session B, location 116

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

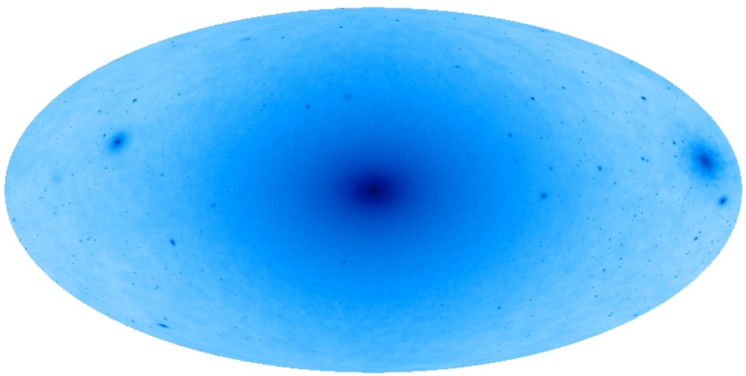
Deutsche Forschungsgemeinschaft
German Research Foundation

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Dark matter in the Galactic center

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

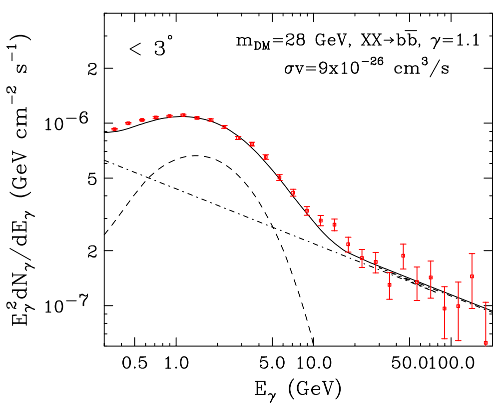
- Galactic center (GC) is the strongest possible source of dark matter (DM) annihilation signal



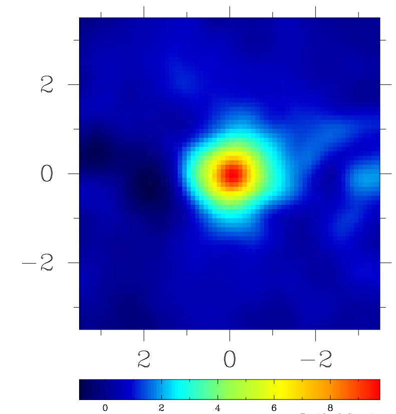
Via Lactea II simulation of a Milky-way-like galaxy. Kuhlen et al. (2009)

- Excess consistent with DM annihilation was detected in Fermi-LAT gamma-ray data two months after the data became public

Goodenough & Hooper (2009), Vitale & Morselli (2009), Hooper & Linden (2011), Abazajian & Kaplinghat (2012), Hooper & Slatyer (2013), Gordon & Macias (2013), Calore et al. (2015), Daylan et al. (2016), Ajello et al. (2016), Ackermann et al. (2017) etc



Goodenough & Hooper (2009)



Abazajian & Kaplinghat (2012)

Astrophysical explanation

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- A population of millisecond pulsars (MSPs) near the GC can explain the Galactic center excess (GCE)

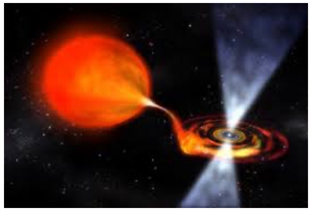
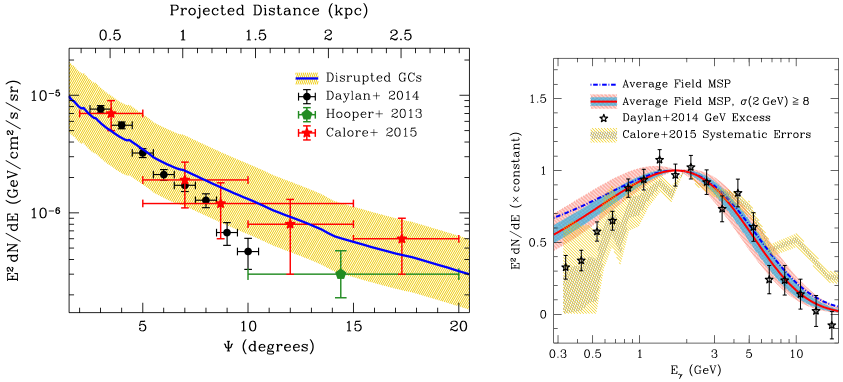
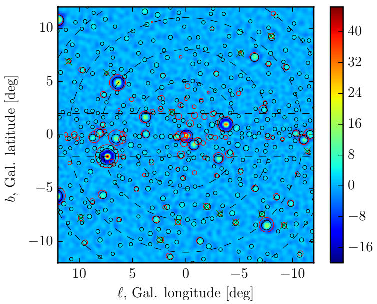


Image credit: NASA



Brandt & Kocsis (2015)



Bartels et al. (2016)

- Statistical studies

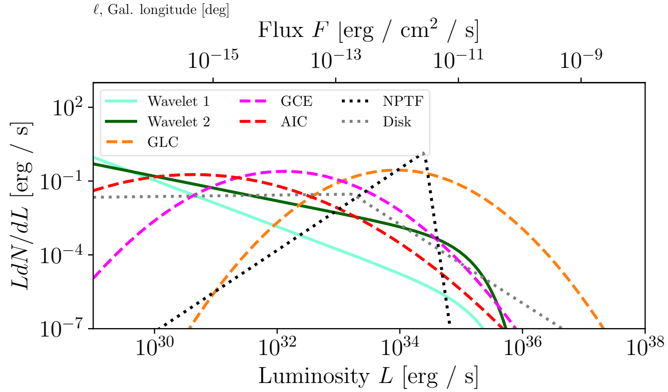
  Lee et al. (2015, 2016), Bartels et al. (2016), Leane & Slatyer (2019, 2020), Zhong et al. (2020), List et al. (2020), Calore et al. (2021), Mishra-Sharma & Cranmer (2022), Caron et al. (2023), Manconi et al. (2024) etc.

  - Based on statistical properties of the Gamma-ray data

- Population studies

  Brandt & Kocsis (2015), Hooper & Linden (2016), Bartels et al. (2018), Ploeg et al. (2020), Dinsmore & Slatyer (2022) etc.

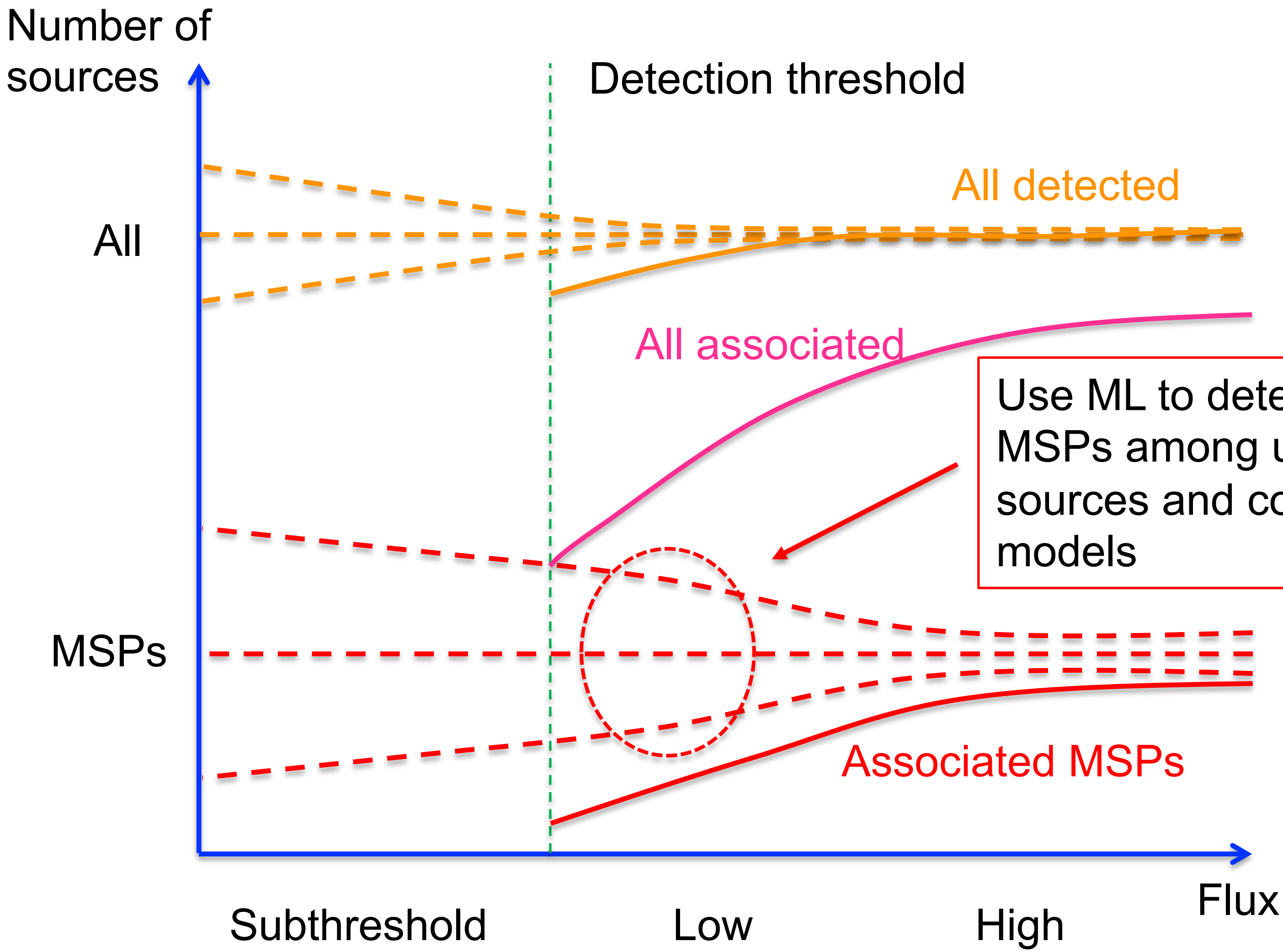
  - Associated (bright) MSPs are used to constrain the models

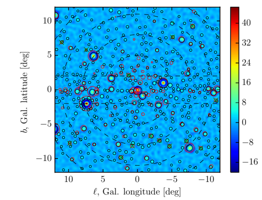


Dinsmore & Slatyer (2022)

How can machine learning help?

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS



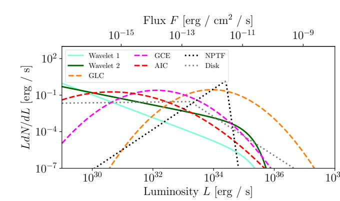


Bartels et al. (2016)

Statistical studies

Use ML to determine MSPs among unassociated sources and constrain MSP models

Population studies



Dinsmore & Slatyer (2022)

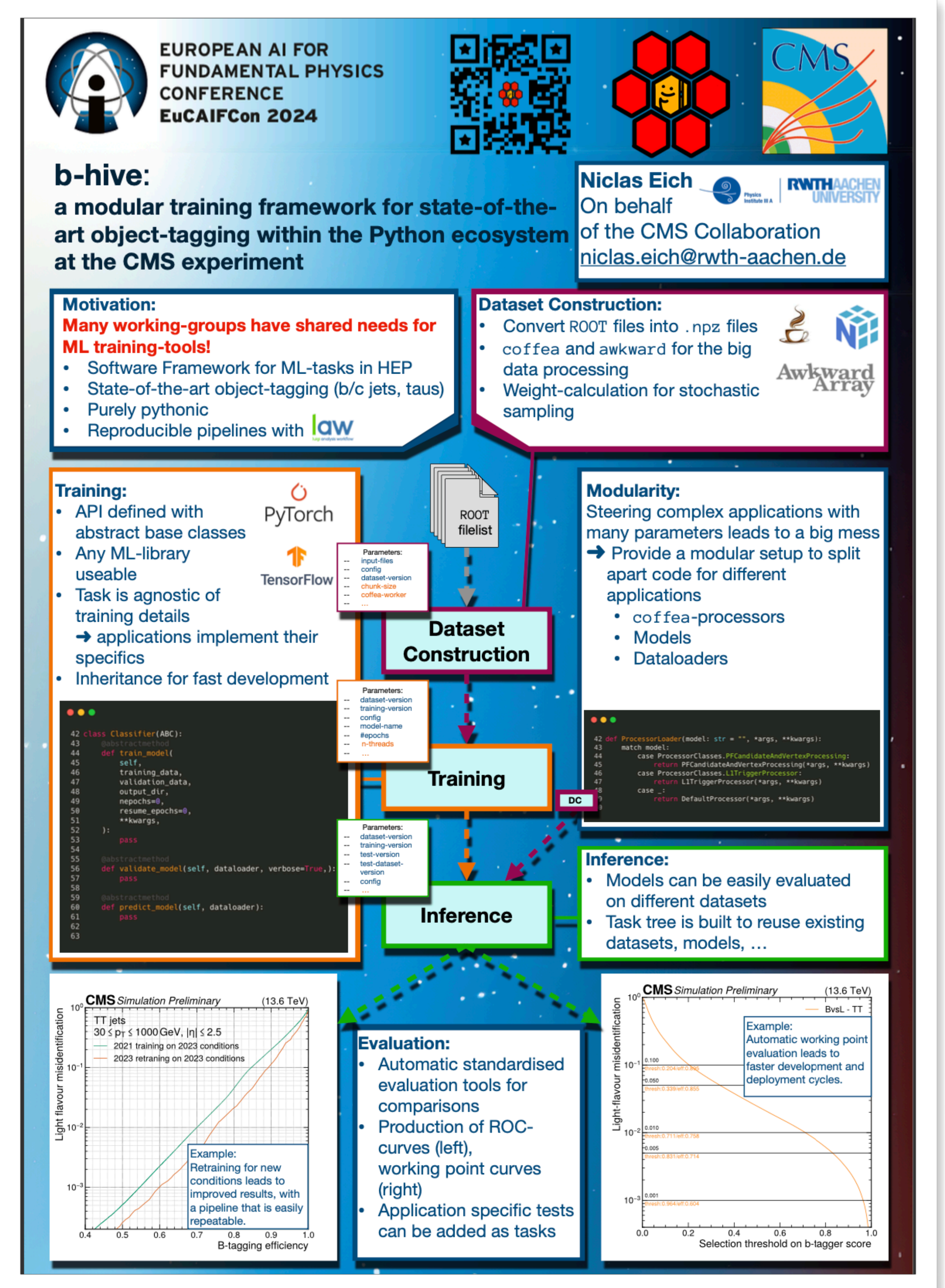
Details and results: poster session B, location 116 & arXiv:2401.04565

b-hive:
a modular training framework for state-of-the-art object-tagging within the Python ecosystem at the CMS experiment

Niclas Eich
On behalf of
the CMS Collaboration
III. Physikalisches Institut A
RWTH AACHEN UNIVERSITY

Motivation:

- Everybody wants to do Machine-Learning trainings
- Full end-to-end pipeline is way harder than an example Notebook
  - Big data processing (`ROOT` files)
  - Conversion into a ML-friendly format (`.npy`/`.npz`)
  - Deploy state-of-the-art models
    - ParticleNet (graph-convolutions)
    - Transformer models



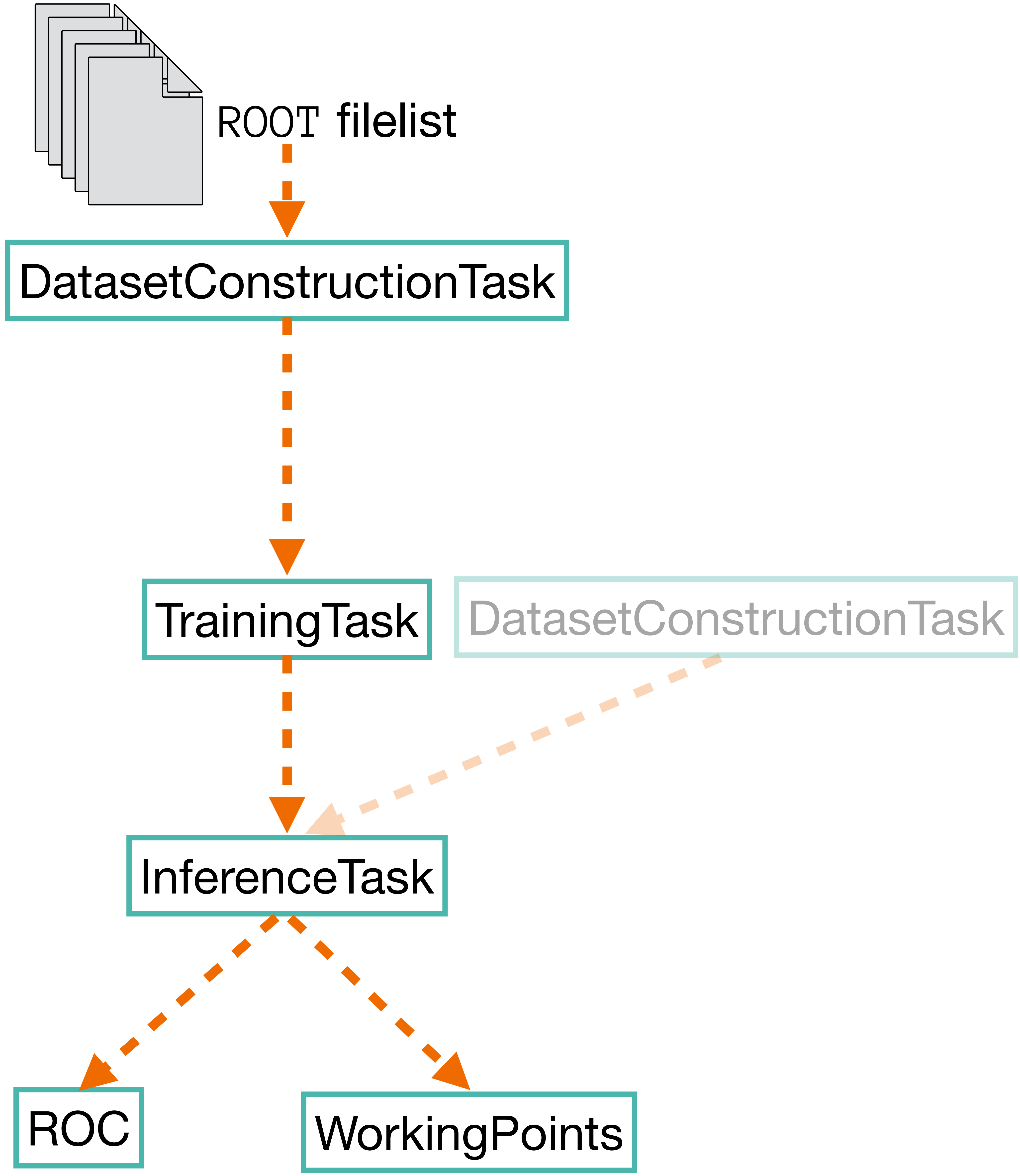
Also:

- Have clearly defined workflows (not your 7 bash scripts!)
- Make trainings repeatable
- Standardized evaluation tools

b-hive attacks these problems

- Pythonic training framework
- Workflow management with `law`
- `coffea`, `awkward`, `numpy` for the heavy data lifting
- No ML-framework lock in
  - `TensorFlow` and `PyTorch` can be used
  - ➜ Modular Setup

```yaml
bins_pt:
  - 30
  - 100
  - 500
  - 2000

bins_eta:
  - -2.5
  - -2.0
  - -0.5
  - 0.5
  - 2.0
  - 2.5

treename:
  "producer/custom_tree"

global_features:
  - "jet_pt"
  - "jet_eta"
  - "n_Cpfcand"
  - "n_Npfcand"
  - "nsv"
  - "npv"

pf_candidate_features:
  - "pfcand_pt"
  - "pfcand_eta"
  - "pfcand_trackPt"
  - "pfcand_trackEta"
  - "pfcand_trackDeltaR"

truths:
  - "isB"
  - "isUDS"
  - "isG"
```



- Easy configuration for different working-groups
- New applications are embedded in the pipeline

➜ Knowledge-sharing by code-sharing

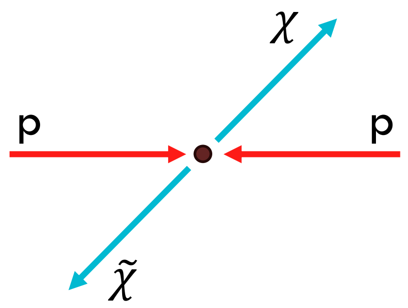
Have a look: CERN-CMS-DP-2024-020

ENERGY-BASED GRAPH AUTOENCODERS FOR SEMIVISIBLE JET TAGGING IN THE LUND REPRESENTATION

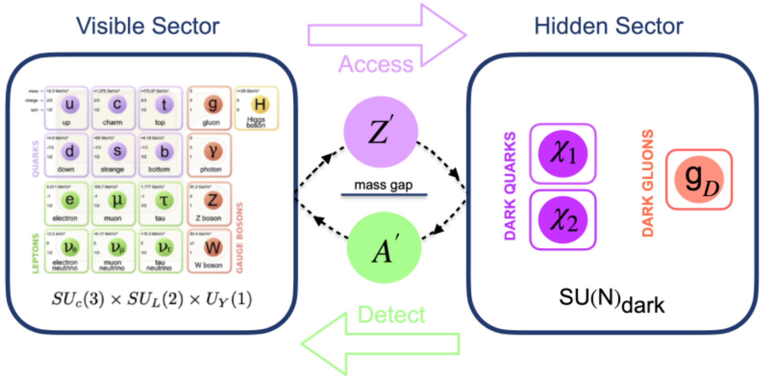
Annapaola De Cosa, Roberto Seidita, Florian Eble, Christoph Ribbe[1]

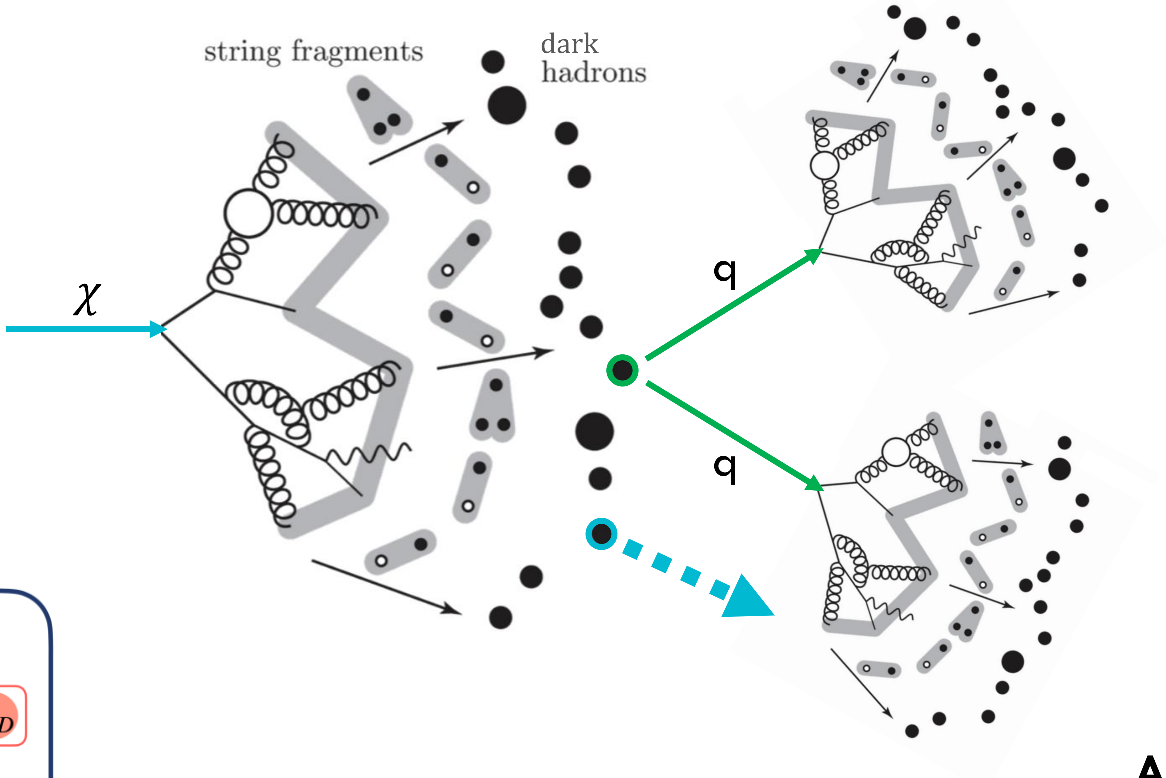
[1]Did most of the work but could not join

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE EuCAIFCon 2024

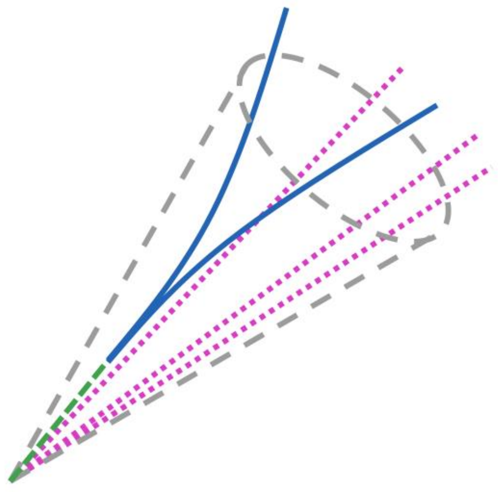


Dark quarks from a proton-proton collision





Shower of **stable (invisible)** and **unstable (visible)** dark particles



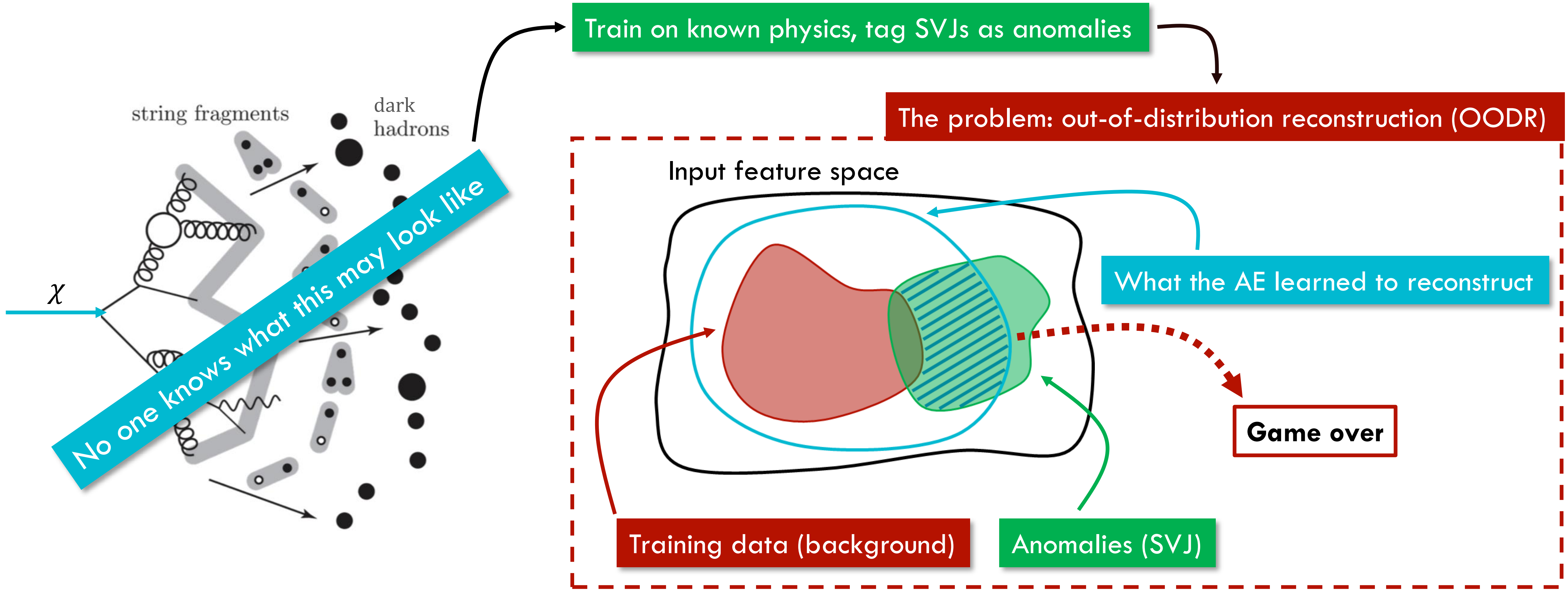
A semivisible jet (SVJ): a spray of particles with missing stuff between them

WHY UNSUPERVISED?

ETH zürich

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE EuCAIFCon 2024

Annapaola De Cosa, Roberto Seidita,
Florian Eble, Christoph Ribbe



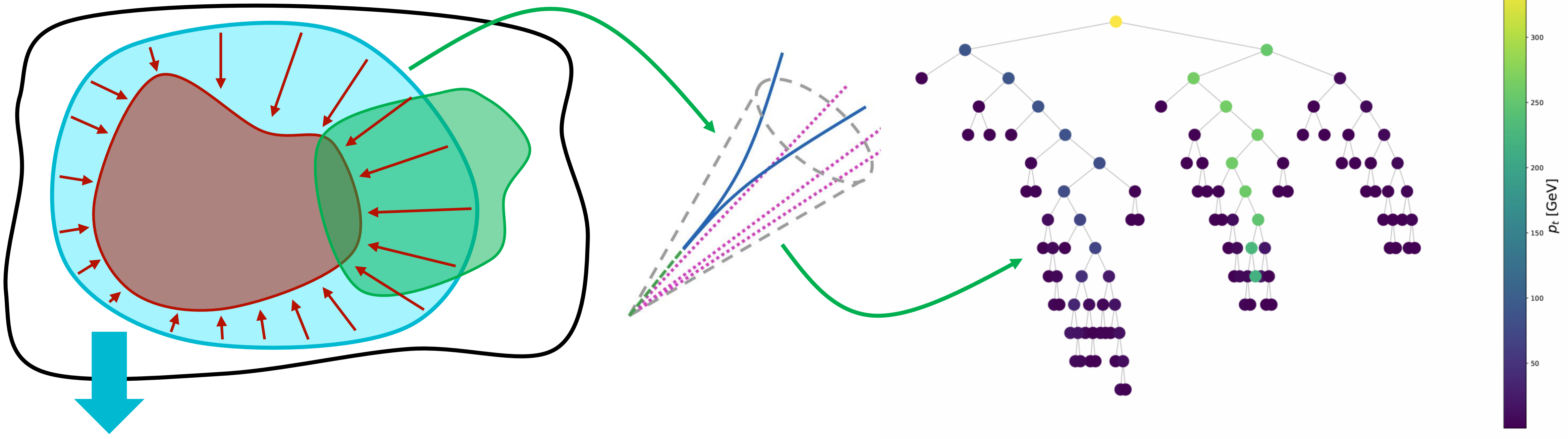
JETS AS LUND GRAPHS

ETH zürich

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE
EuCAIFCon 2024

Annapaola De Cosa, Roberto Seidita,
Florian Eble, Christoph Ribbe

Input feature space



Sample here and enforce $p_{model} = p_{data}$

Normalized autoencoder

How do you sample from graph space? Come to the session tomorrow!