# AI-Assisted Analysis to Enhance Discovery Potential in High Energy Physics

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

*Asrith Krishna Radhakrishnan, Maximiliano Sioli, Lorenzo Rinaldi*
*Physics and astronomy department, University of Bologna*

## Motivation

Traditional LHC searches follow a supervised script: test a known signal against background. This model-dependent approach scored big wins, like the Higgs, but has yet to uncover new physics.

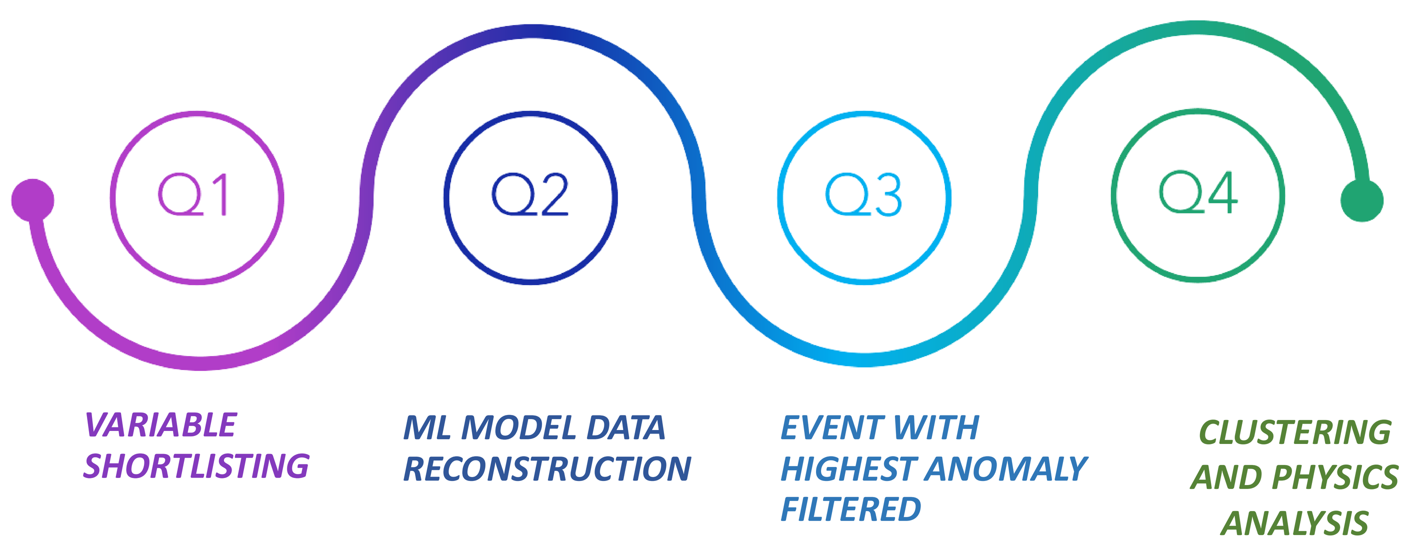
**The challenge**? The BSM landscape is vast, and fixed triggers or cuts can bias what gets seen. A Novel approach: **unsupervised, model-agnostic strategies**. Powered by **generative deep learning**, these methods learn the patterns from data and flag the outliers. We aim to catch **the rare**, **the BSM event**.

## Novelty

We present an **unsupervised**, **transformer-based** reconstruction model trained solely on Run 2 2015 proton-proton collision data from the ATLAS experiment.

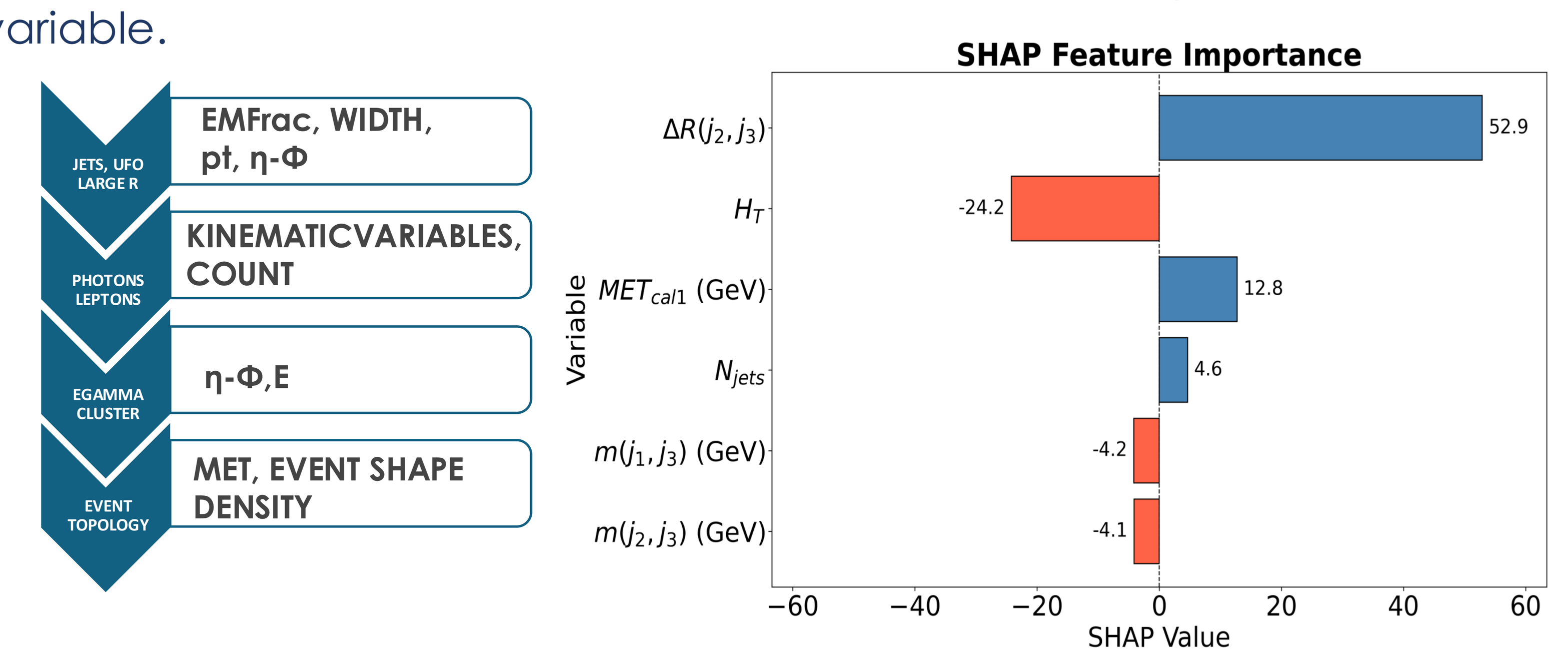
We developed a **Masked Tab-Transformer [1]** that handles **high-dimensional**, **sparse** tabular detector data, aiming to reconstruct events with large fractions of **missing inputs**. Our model reconstructs all detector-level features: jets, leptons, photons, missing energy, calorimeter deposits, and event topology.

## Model Development Workflow



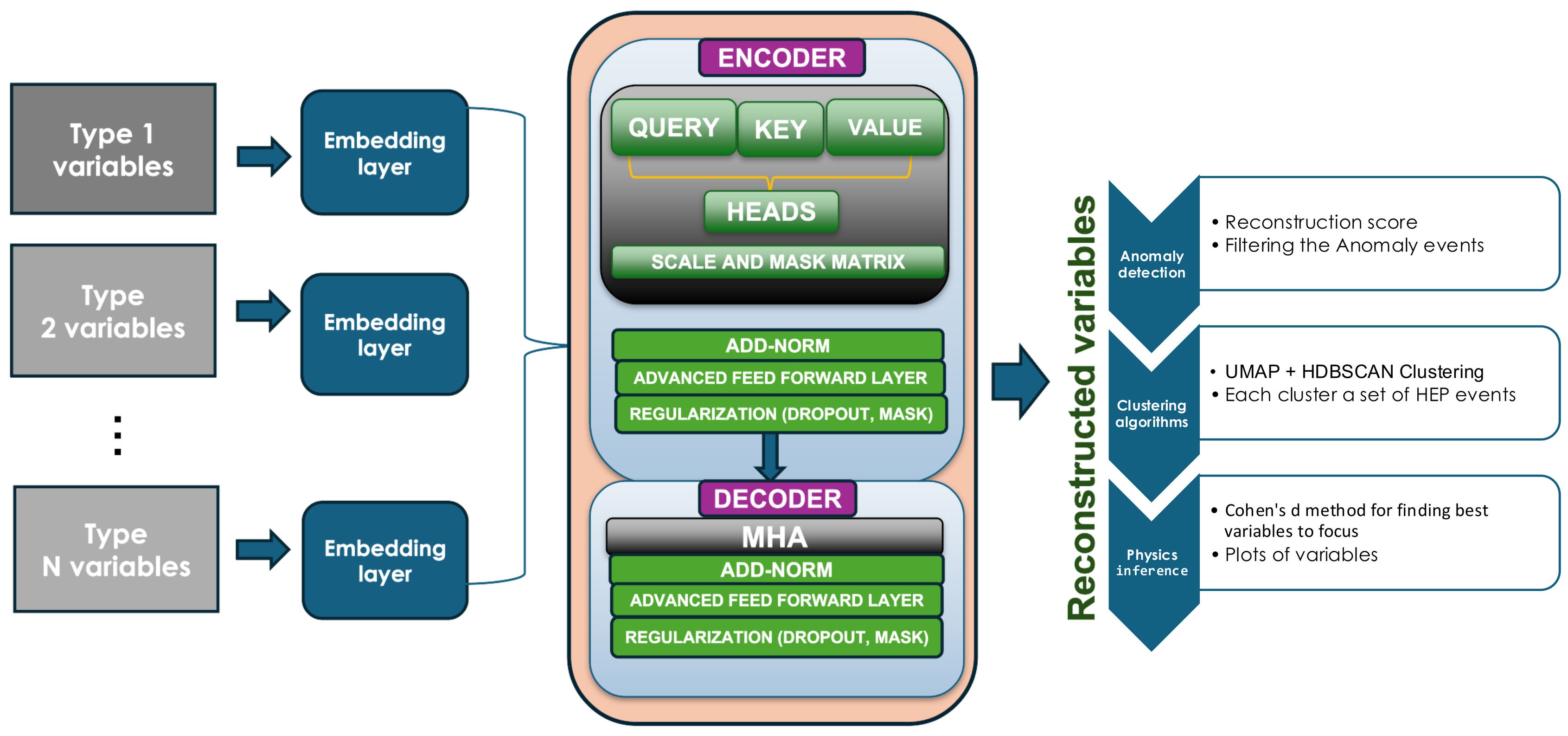

We analyze the **ATLAS Run-2 open dataset** (13 TeV, 2015) [2] a with ~5 million proton–proton collisions and full detector readouts. From each event, we extract low-level feature jets, leptons, photons, MET, calorimeter deposits, and event shape density.

- ❑ Preprocessing and scaling using **ROOT RDataFrame** on **CERN's SWAN** interfaced with **Dask** and **HTCondor.**
- ❑ Streaming output **.root** files directly as TensorFlow via **TMVA TF batch generator** [5].
- ❑ **SHAP**-based feature selection to reduce redundancy and remove noisy variable.




**Tab-Transformer autoencoder with masked self-attention and custom masked loss (weighted loss).**

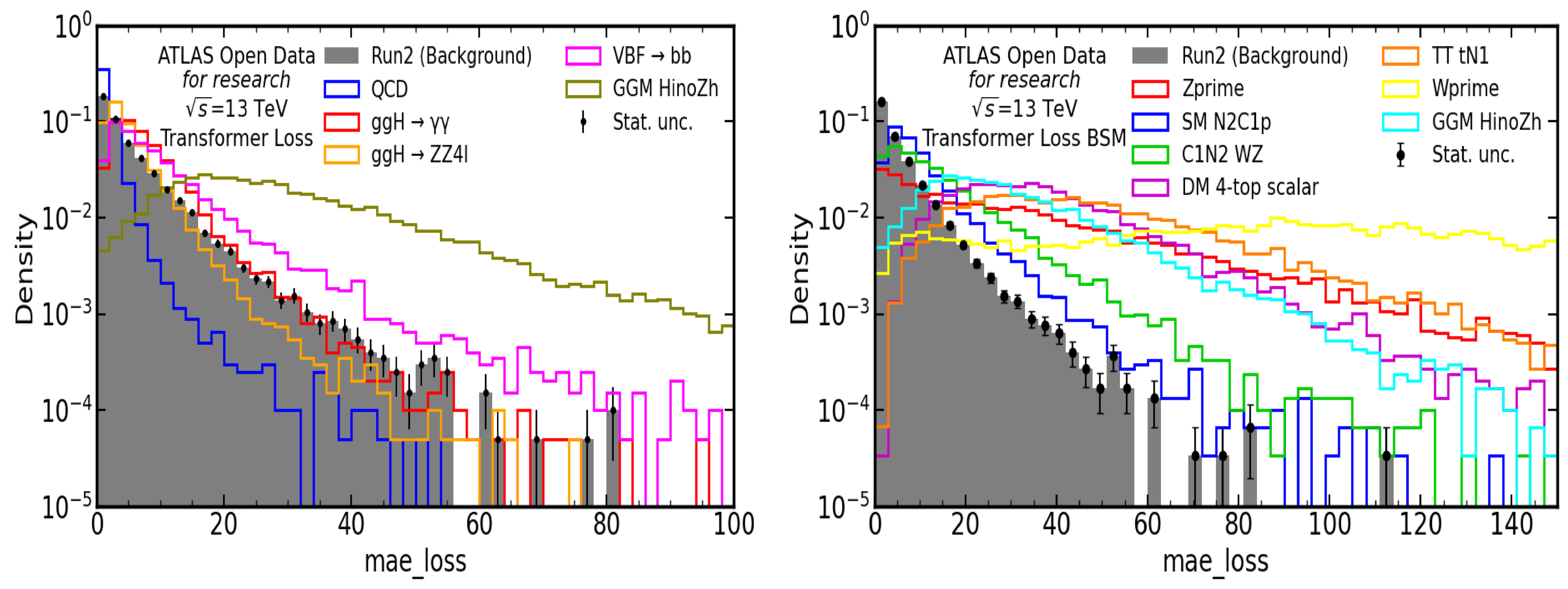
- ❑ Outliers are organized with **UMAP[3] + HDBSCAN[4]**, revealing structured patterns (e.g., high-pT leptons, jet-rich events).
- ❑ **Cohen's d** quantifies statistical significance, highlighting clusters that truly deviate from SM pattern




## Results

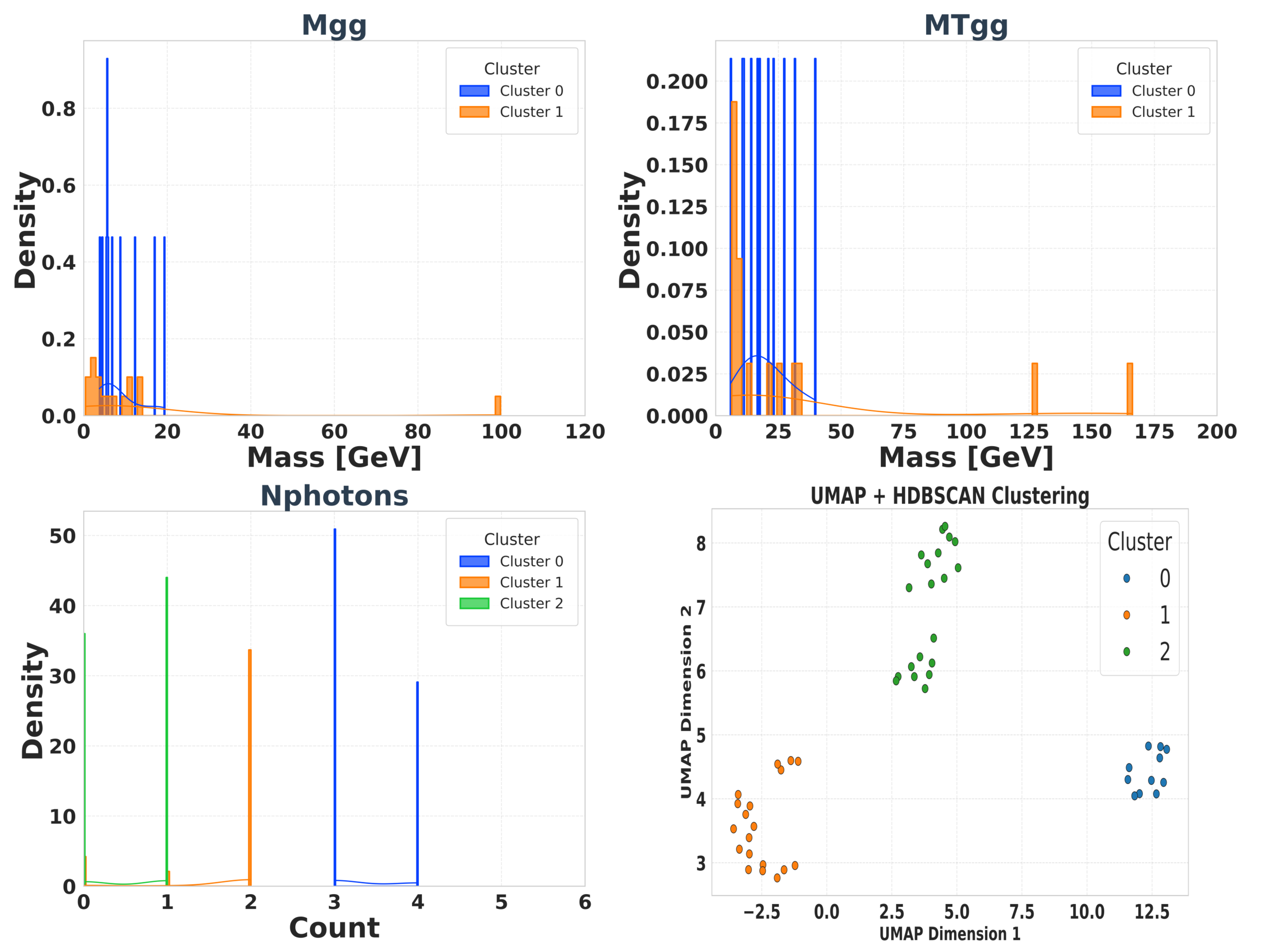
**Testing on simulated BSM samples** (SUSY, exotic resonances, rare Higgs/top decays) treated as anomalies.

- ❑ **SM backgrounds** reconstructed with high fidelity (low loss).
- ❑ **BSM events** show systematically higher reconstruction loss.
- ❑ Examples: leptonic SUSY cascades and heavy resonance decays produce clear high-loss tails, well-separated from SM validation samples.




- ❑ Performed clustering analysis to identify the most important variables driving the separation.
- ❑ Based on the Cohen's d value plotted important variable in clusters.

### Distributions by Cluster




## Unbiased Search

**Fully unsupervised pipeline** trained on raw detector-level data, with no assumptions on final states.

- ❑ **Transformer autoencoder** learns the Standard Model features from training events, flags outliers via reconstruction loss acting as a data-driven trigger for rare or unknown phenomena.
- ❑ **UMAP + HDBSCAN** cluster anomalies, revealing structured event classes.
- ❑ **Scalable & interpretable**, expanding discovery reach in underexplored phase space corners.
- ❑ **Future: larger datasets**, Improve network architecture, calibration, **richer inputs (tracking, timing), real-time trigger-level deployment** for truly model-independent new physics searches

## Reference

[1] Arik & Pfister, *TabTransformer*, AAAI 2021, arXiv:2012.06678.
[2] ATLAS Open Data, *pp collisions at √s = 13 TeV*, Run 2 (2015–2016), doi:10.7483/OPENDATA.ATLAS.9HK7.P5SI
[3] L. McInnes et al., arXiv:1802.03426 [stat.ML]
[4] R.J.G.B. Campello et al., *Adv. Knowl. Discov. Data Min.* (2013), Springer, ISBN 978-3-642-37456-2
[5] ROOT TMVA, RBatchGenerator TF (2024), root.cern.ch

## Acknowledgement

- ❑ I would like to thank Dr. Giovanni Guerrieri for his valuable guidance in the effective utilization of resources and infrastructure for this work.
- ❑ I would also like to thank the University of Bologna and Huawei for providing financial support for this project.