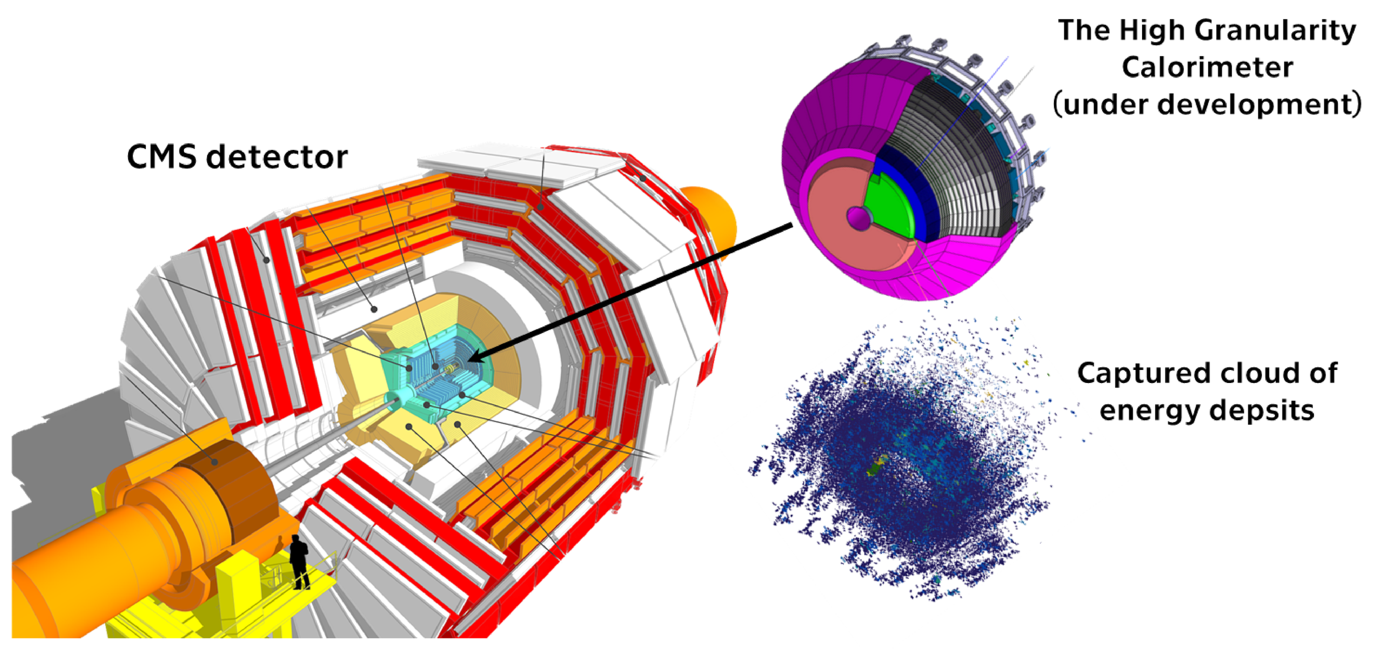
# Robust Edge Classification through Contrastive Learning on Graphs for Event Reconstruction


Christine Zeh [1, 2] Wahid Redjeb [2] Felice Pantaleo [2]

[1]Technical University Vienna [2]European Organization for Nuclear Research (CERN)


## Motivation

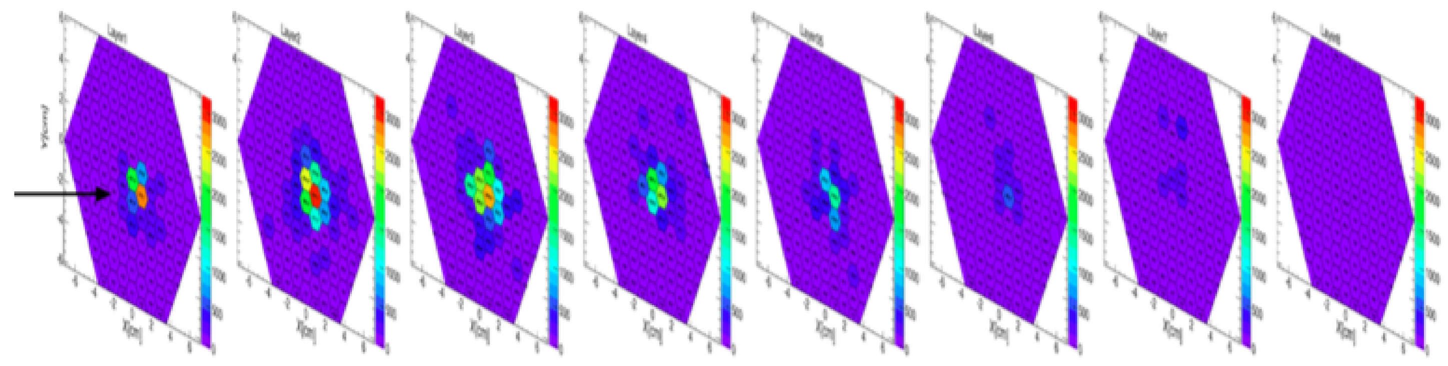
- Traditional deterministic clustering struggles with efficiency and scalability as event rates and complexity increase at the High-Luminosity LHC (HL-LHC). This is why a highly granular calorimeter is used to try to separate particles as much as possible, given that many more particles overlap with each other. This increases the size of events, as this results in many more energy deposits.
- **Goal:** Develop a Graph Neural Network (GNN) to cluster calorimeter energy deposits for real-time event reconstruction.




- **Challenges:**
  - Events contain up to 500k energy deposits (hits), which end up in $\approx 4000$ nodes per graph.
  - Only $\approx 2\%$ of edges represent true physical connections → strong class imbalance.
  - Need for robust and generalizable training.

## Event Reconstruction in the CMS HGCAL Endcap

- **The HGCAL Detector:**
  - The *High-Granularity Calorimeter (HGCAL)* is part of the *CMS experiment* for the CERN upgrade HL-LHC.
  - It is located in the *forward region*, close to the beam line, where particle densities are extremely high.
  - Composed of $\approx 6$ million silicon channels, it measures the *energy, position* and *time* of particles produced in collisions.
- **What "Event Reconstruction" Means:**
  - Each proton collision creates a "spray" of particles, and each of them leave energy deposits in the different sub-detectors of CMS.
  - *Event reconstruction* aims to transform the detector signals into meaningful patterns, building a structured description of what happened in the collision, identifying the types of particles that were produced and estimating their kinematic properties.

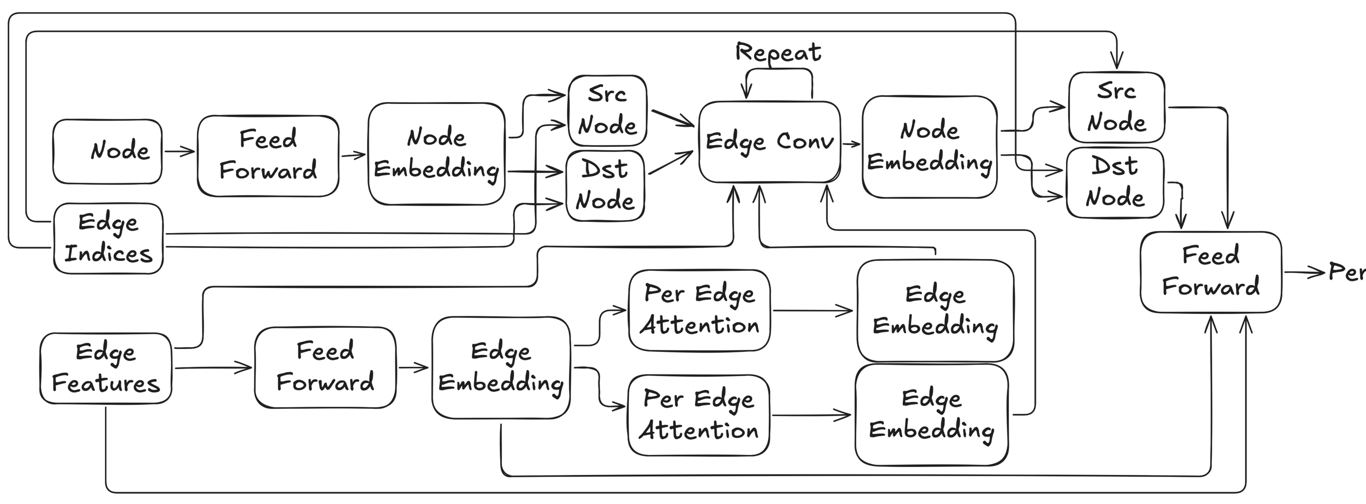


- **Why It's Challenging:**
  - Each event can produce *hundreds of thousands of hits* in the HGCAL.
  - Many hits overlap, especially during *HL-LHC* operation with up to 200 collisions occurring simultaneously.
  - The system must process all this information in *real time*, within milliseconds.
- **Current Approach:**
  - Uses *rule-based clustering algorithms* that group nearby hits into clusters, each likely representing one particle.
  - These methods work well but become *less efficient* and *scalable* as data rates and complexity increase.
- **New Direction:**
  - *Machine learning*, particularly *GNNs*, can learn to recognize particle patterns directly from data.
  - This approach enables *faster, more adaptive*, and *more accurate* reconstruction under extreme conditions.

## Graph Neural Network Architecture

A modified GNN for link prediction between CMS HGCAL 3D clusters, extending pairwise MLPs with neighborhood-aware message passing via attention-based EdgeConv [2] layers with skip connections.

- Node features, $X \in \mathbb{R}^{N_B \times 27}$ : 3D clusters containing multiple layer clusters
- Edge index, $I \in \mathbb{R}^{E_B \times 2}$ : Connections based on a $\eta - \phi$-cone, connecting physically possible clusters
- Edge features, $E \in \mathbb{R}^{E_B \times 8}$ : Relative feature information like spatial distance and difference in energy.




- *Latent Encoding:* Two-layer MLPs map features to latent spaces ($x_0 \in \mathbb{R}^{N_B \times 64}, e_0 \in \mathbb{R}^{E_B \times 32}$).
- *Edge Attention:* Static attention weights modulate neighbor contributions.
- *Neighborhood Aggregation:* Four EdgeConv layers with skip connections aggregate information up to four hops.
- *Link Prediction:* Edge embeddings passed through a two-layer MLP with sigmoid output for link scores.

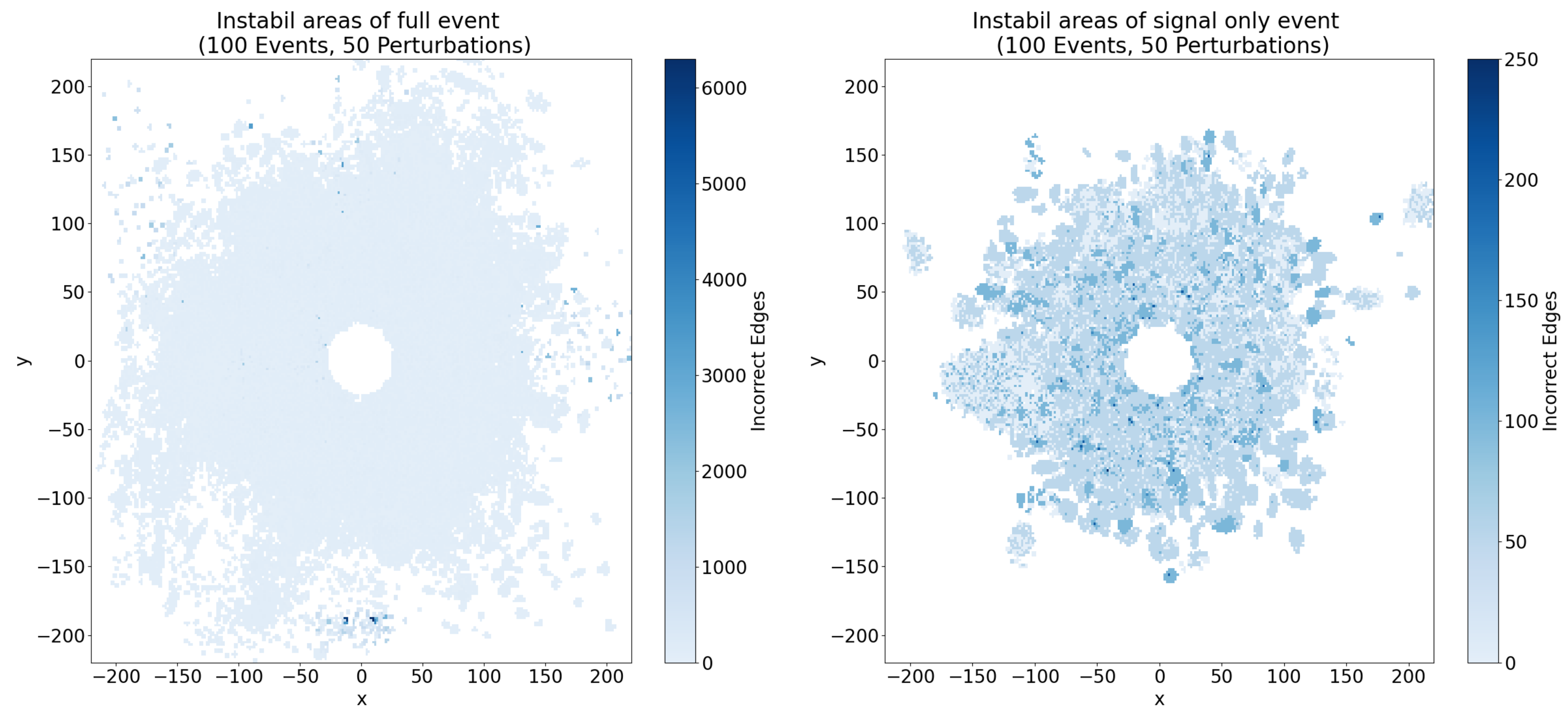
## Loss Function and Contrastive Learning

Combines correct data classification and reasonable representation to improve clustering robustness and generalization under high pile-up.

- **Focal loss:** Handles strong class imbalance ($\sim$2% true edges) by focusing on improbable training data.
- **Contrastive loss:** Learns robust embeddings by pulling similar samples closer and pushing dissimilar ones apart, improving generalization and stability [1].
- **Combined objective:**

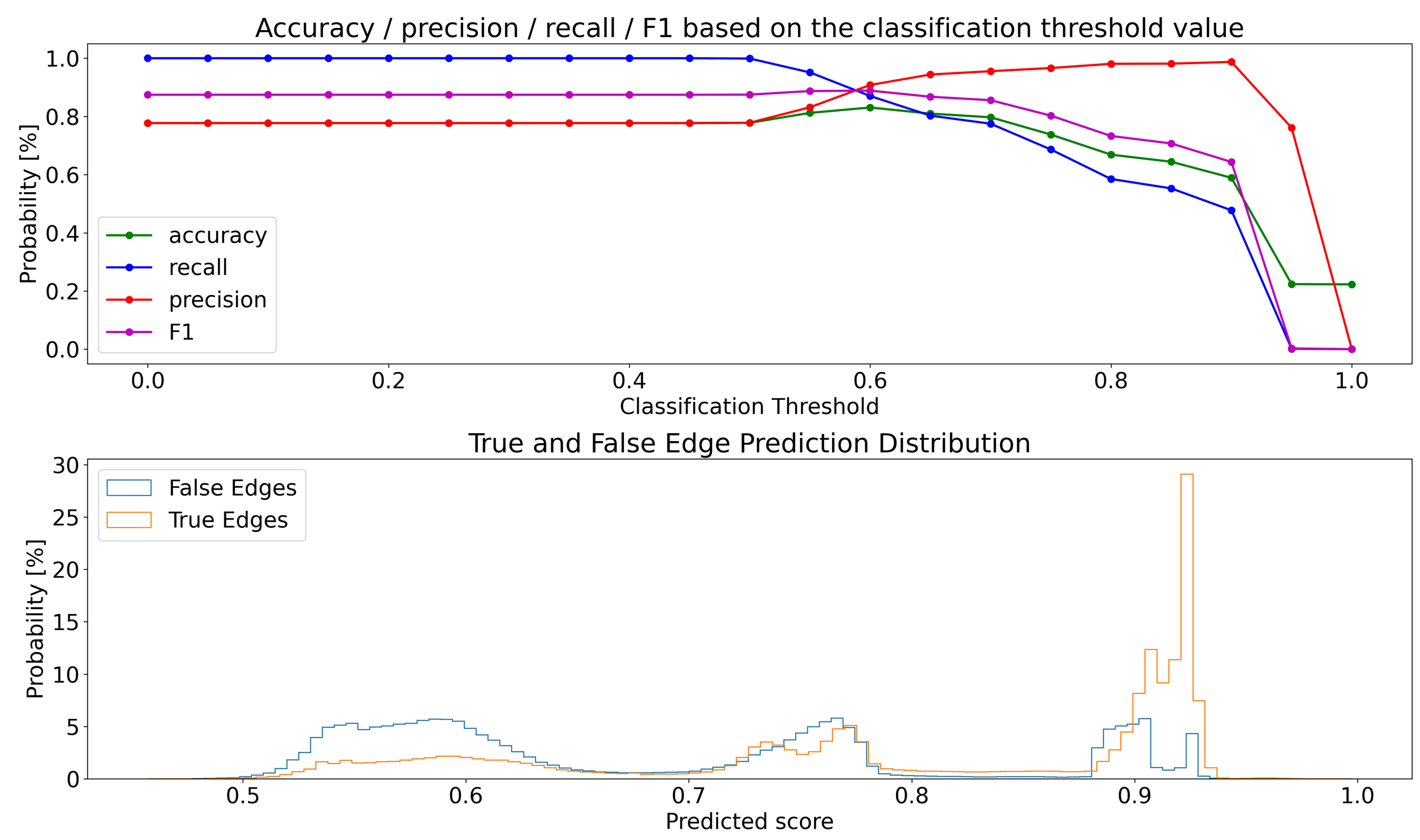
$$\mathcal{L} = \frac{2}{5}\mathcal{L}_{\text{focal}} + \frac{3}{5}\mathcal{L}_{\text{contrastive}}$$

- **PCA-based perturbation:** Positive samples are generated by shifting node embeddings along the first eigenvector, scaled by reconstruction error — capturing spatial uncertainty.




## Training Results

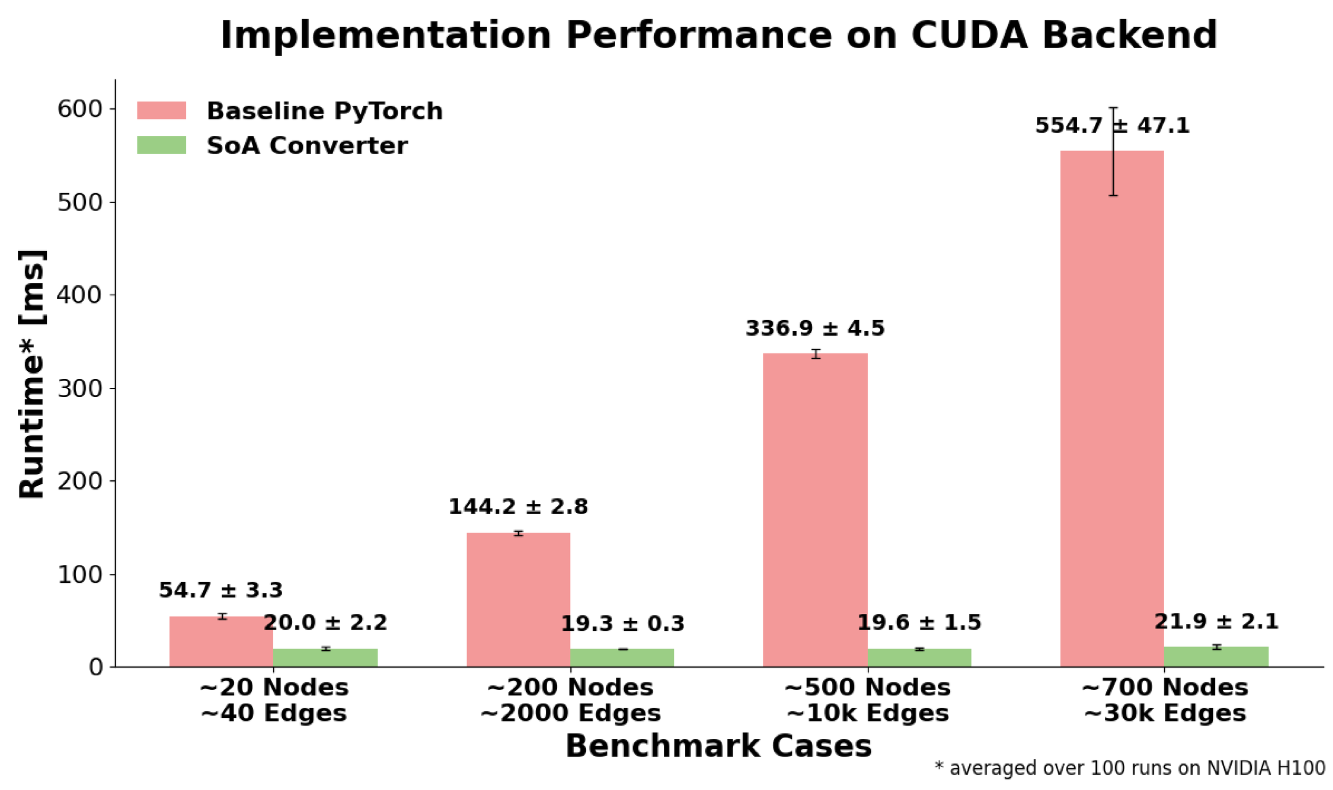
While validation accuracy drops compared to focal-only training, the model gains efficiency and handles more heterogeneous data.




**Next Steps:** Add attention to nodes to improve performance in high-$\eta$ regions, where cluster quality often degrades. Initial results show local improvements even without contrastive learning, but with reduced overall efficiency. The combined loss function remains unstable and requires further investigation.

## Real Time Reconstruction

Inference efficiency is achieved by minimizing copy operations and adopting a Structure of Arrays (SoA) layout for optimized memory access in the C++ software of CMS. This approach ensures that inference effort remains almost constant, even for large inputs.




## Partners & References


[1] Phuc H. Le-Khac et al. "Contrastive Representation Learning: A Framework and Review". In: *CoRR* abs/2010.05113 (2020). arXiv: `2010.05113`.

[2] Yue Wang et al. "Dynamic Graph CNN for Learning on Point Clouds". In: *CoRR* abs/1801.07829 (2018). arXiv: `1801.07829`.



Supported by the Eric & Wendy Schmidt Fund for Strategic Innovation (grant agreement SIF-2023-004)