The 2nd AI+HEP in East Asia
Auditorium room, Building #3, KEK, Japan

Foundation Model for Decay Event Classification at BESIII

Jingde Chen, Zijie Shang, Tong Liu, Ke Li

Institute of High Energy Physics, Chinese Academy of Sciences,
Experimental Physics Center, China

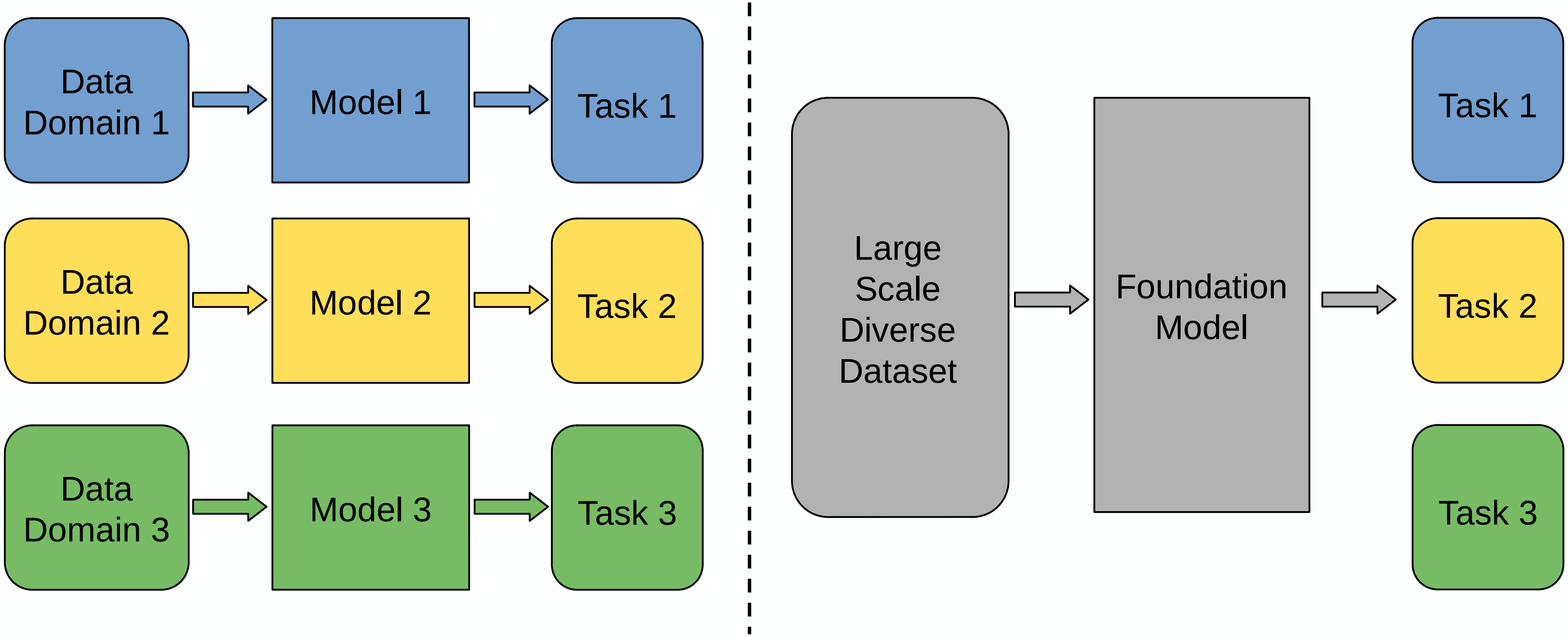
Jan 19th–23th, 2026

Part 1: Introduction

Introduction

What is a Foundation Model:

Any model that is trained on broad data (generally using **self-supervision** at scale) that can be adapted (e.g., **fine-tuned**) to a wide range of **downstream tasks**.



Why Foundation Model:

■ Traditionally, machine learning models are trained separately for **each specific task**.

■ A Foundation Model is pre-trained once and serves as a shared foundation for **many downstream tasks**.

Introduction

How to design Foundation Model

★ Foundation Model ≈ Pretext Tasks + Loss Functions ★

(1). Pretext Tasks: The task being solved is not of genuine interest, but is solved only for the true purpose of learning a good data representation.

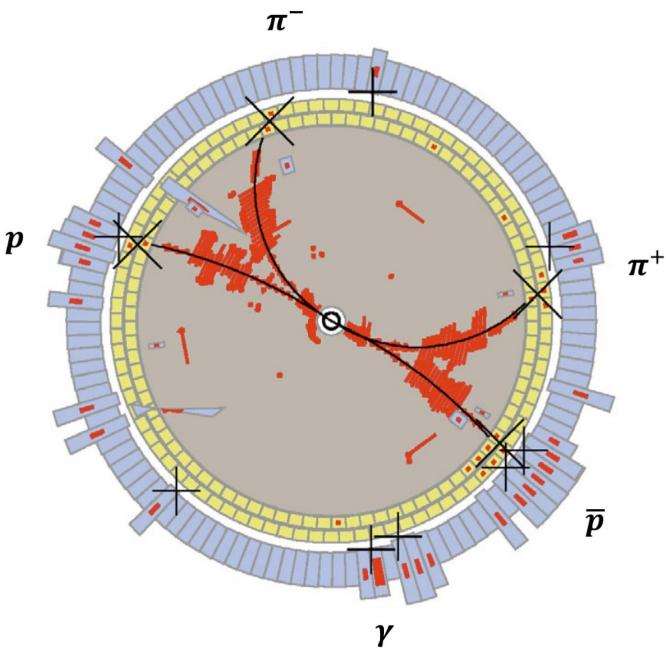
(2). Loss Functions: Loss functions can often be investigated independently of pretext tasks.

(3). Pretext tasks ≈Contrastive learning and masking learning.

Foundation Model in BESIII

■ Purpose: To cluster **all BESIII decay channels (>3,000 categories)**; our foundation model instead performs clustering at the collision-event level, potentially involving **tens to hundreds of thousands of clusters**.

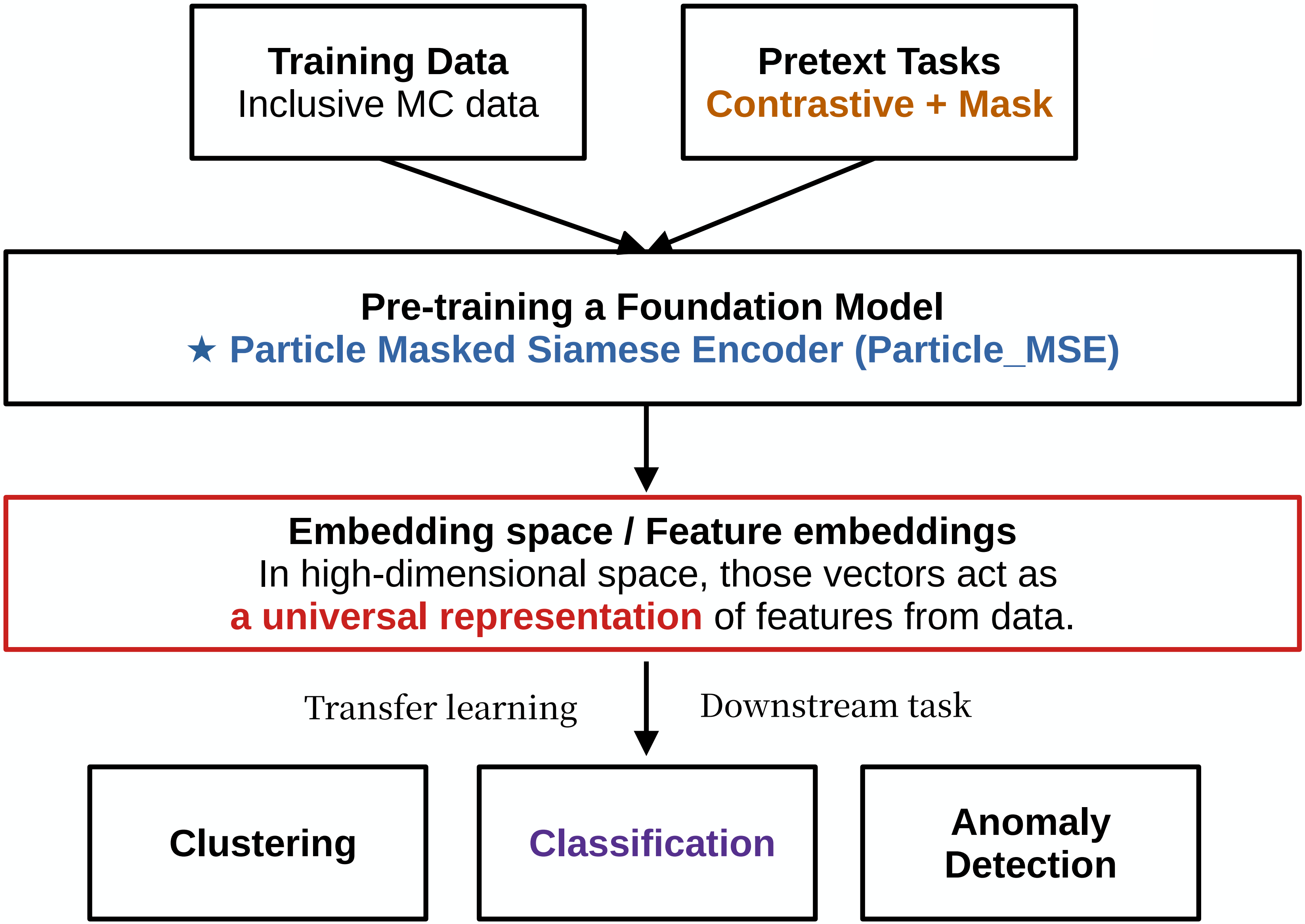
★ Research Highlight: **An extreme multi-class classification problem.**



Part 2: Research Method

Research Method

Research Procedure:



Research Method

Training Data:

◎ Decay event: **J/ψ** decay (Inclusive MC data)

◎ Event classes: **More than 1000 decay classes** in J/ψ decay. **Only the top 99** most frequent **classes** are **used for model training**.

◎ Training dataset: 99 classes of J/ψ decay events (= **3M samples**; expected to scale to over **100M in future versions**)

◎ Input features:

⇒ **Kinematic**:
['q','mdc_p3_px', 'mdc_p3_py', 'mdc_p3_pz', 'vz0', 'vr0']

⇒ **Particle identification information**:
['pid_prob_type']

⇒ **Electromagnetic calorimeter** :
['emc_numHits', 'emc_e3x3', 'emc_e5x5', 'emc_energy', 'emc_x', 'emc_y', 'emc_z', 'emc_time', 'EoP', 'emc_secmom', 'emc_latmom']

⇒ **Muon Counter** :
['muc_dpt', 'muc_numLayer', 'muc_maxHitsInlater', 'muc_chisq', 'muc_dof', 'muc_numHits']

Downstream Tasks (Excluded from Training Data):

◉ Task 1 – Clustering

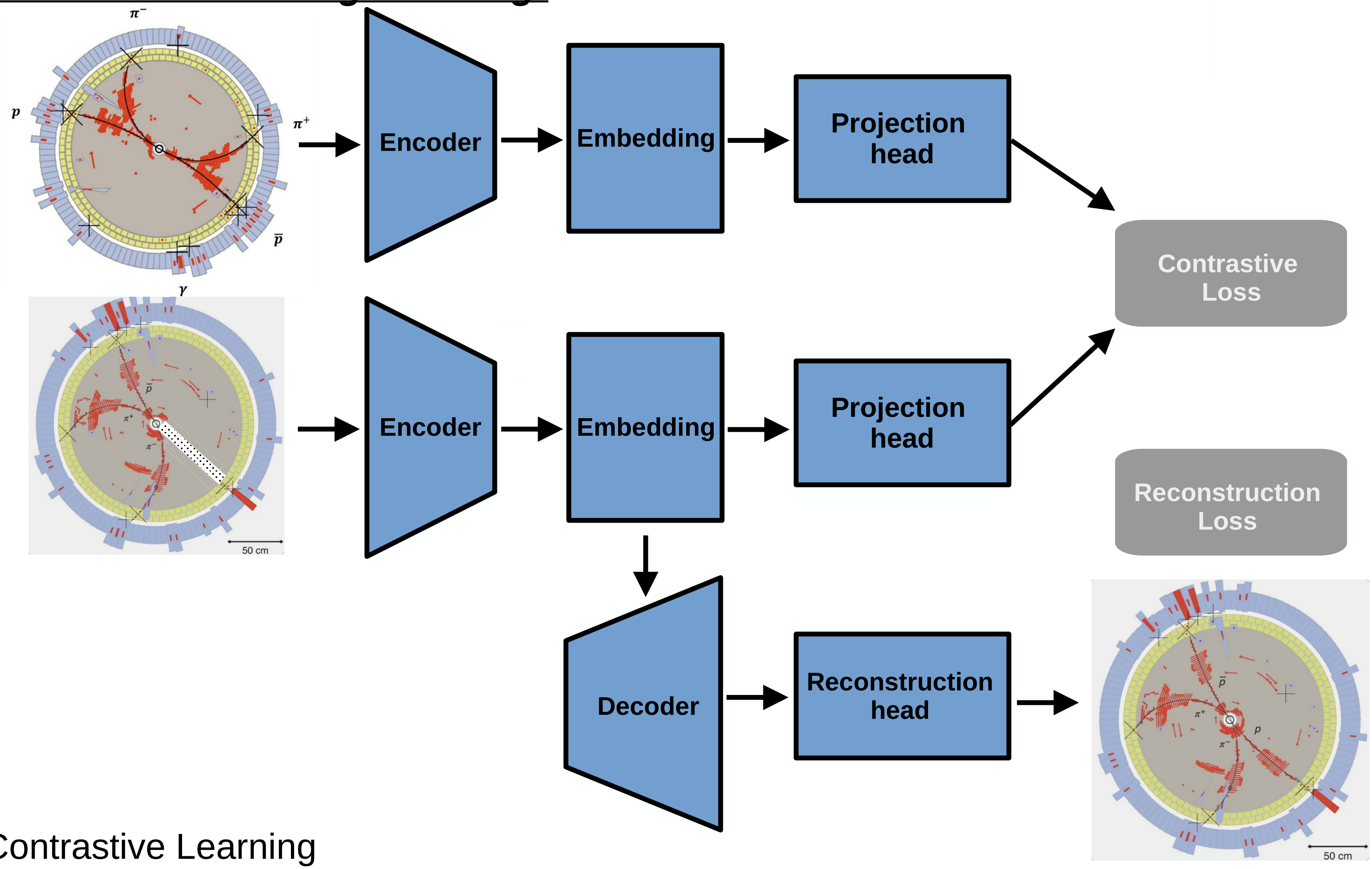
⇒ 300K samples from the top 99 classes of J/ψ decay classes.

◉ Task 2 – Classification (**99 classes classification task**)

⇒ 300K samples from the top 99 classes of J/ψ decay classes. (50% labeled, 50% unlabeled)

Research Method

Contrastive + Masking Learning:



◉ Contrastive Learning

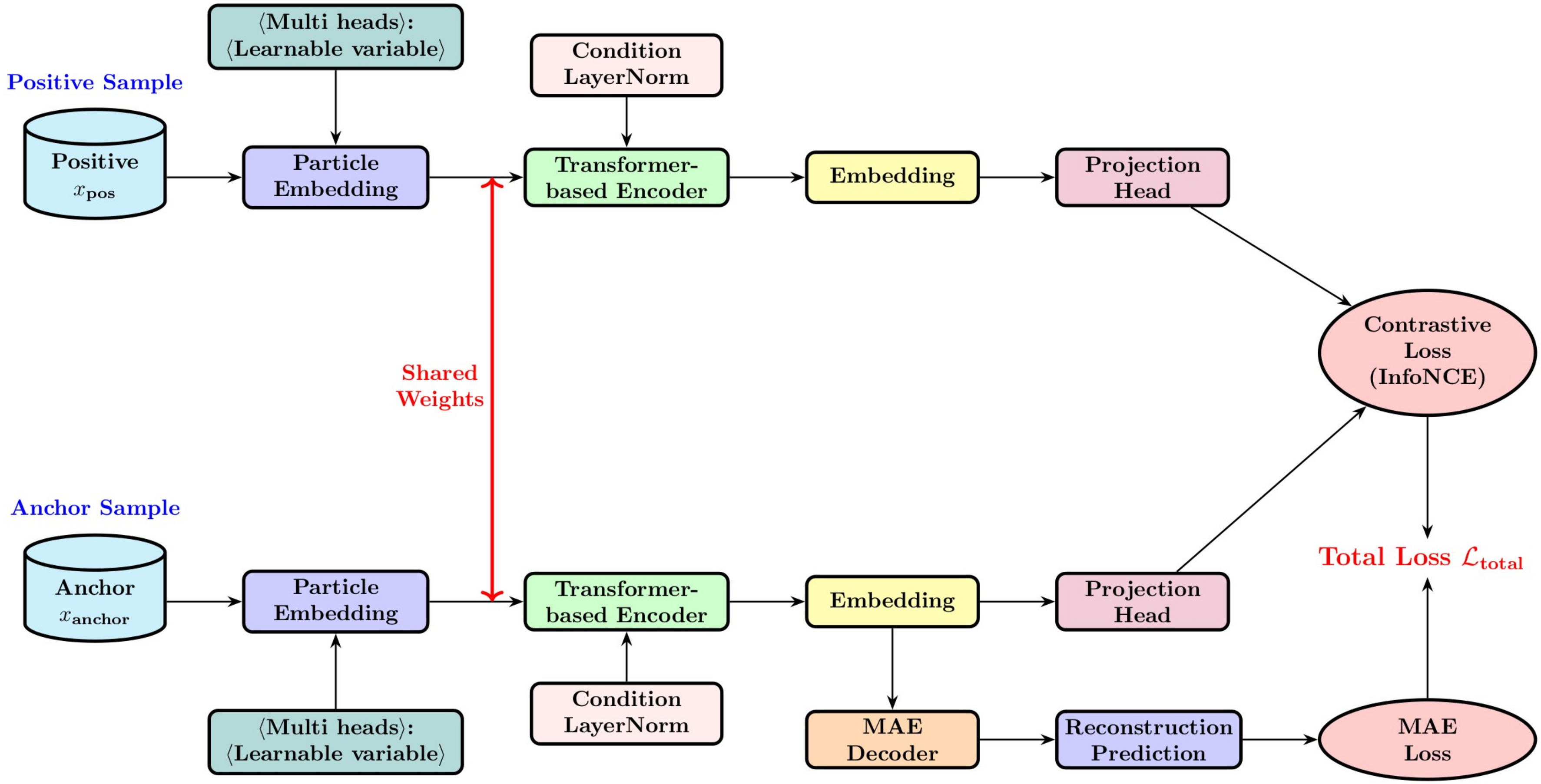
⇒ Brings representations of **same decay types closer**, and **push apart of different classes**.

◉ Masking Learning

⇒ Captures **inter-particle correlations** within decay events.

Research Method

Particle-MSE Structure:



Condition LayerNorm:

- Condition 1: Number of the total charged tracks.
- Condition 2: Number of the total neutral tracks.
- Condition 3: Number of the charged tracks with further selection.
- Condition 4: Number of the neutral tracks with further selection.

Part 3: Research Results

Research Results

Training Results:

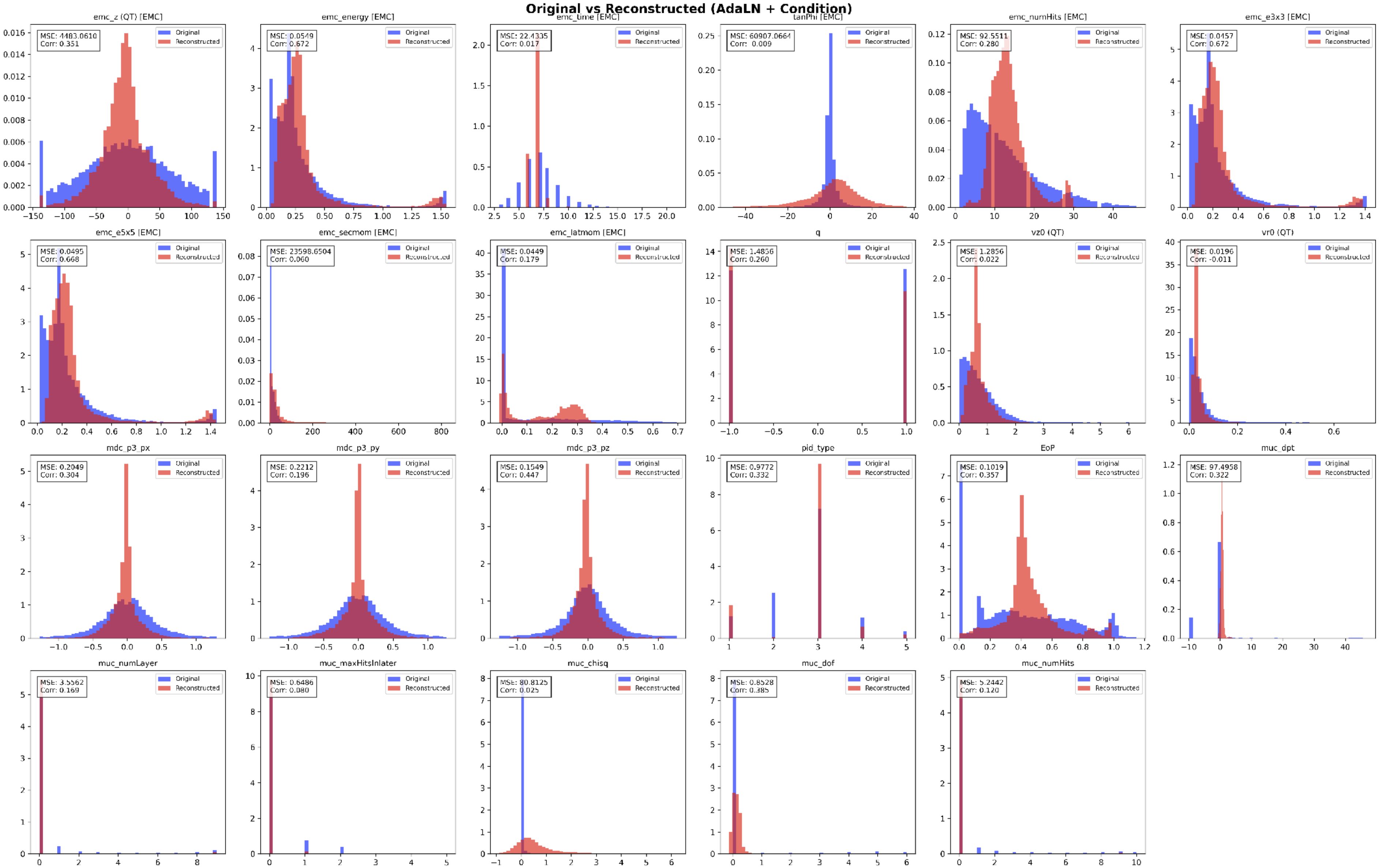
■ Total Loss: Total loss shows a steady downward trend over the first 50 epochs. however, it **does not yet appear to have converged**. Training will be extended to 100 epochs for further optimization.

■ MAE Loss (Reconstruction Loss): Reconstruction loss decreases overall but exhibits **substantial instability** during training.

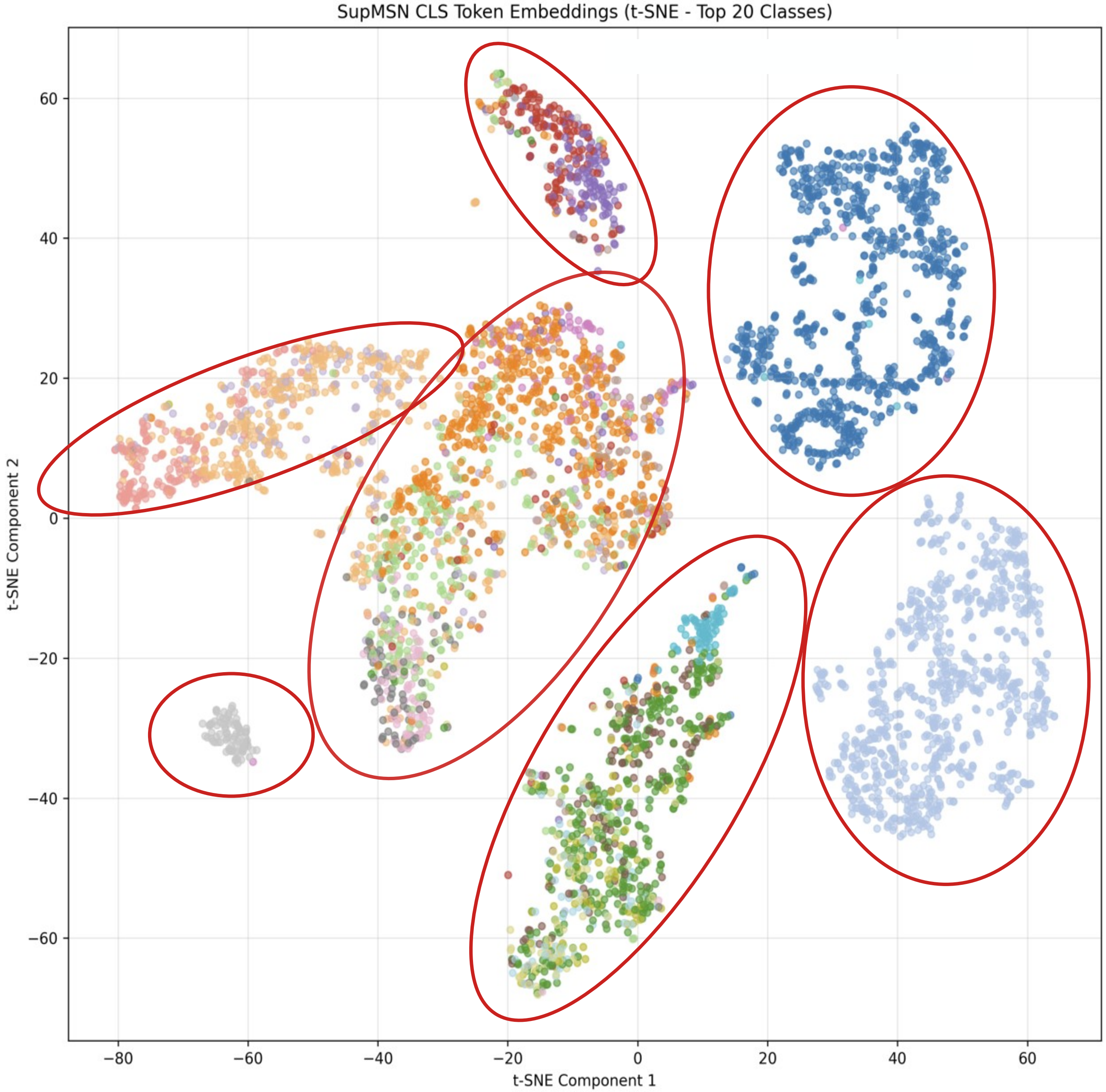
■ Contrastive Loss: Contrastive loss performs well, **benefiting from the supervised contrastive learning**. Future plans include exploring self-supervised contrastive way.

Research Results

Reconstruction Results:



Research Results



Classification results:

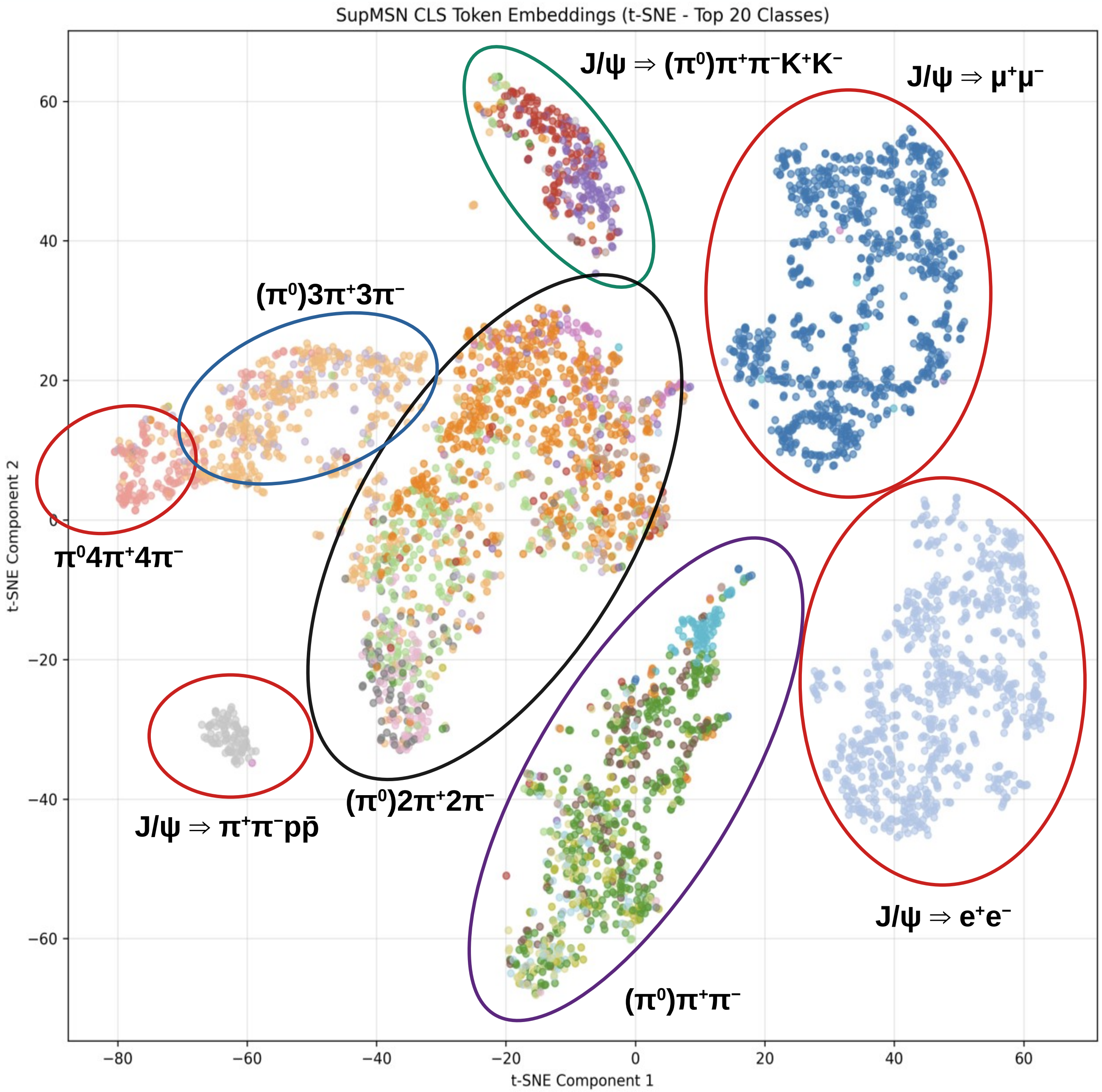
■ For the classification task, the embedding space was derived from representations of events drawn from the top 20 decay classes.

■ In the t-SNE visualization, **7 clusters emerge**.

■ Among these, **3 clusters** are composed almost of **a single type of decay event**.

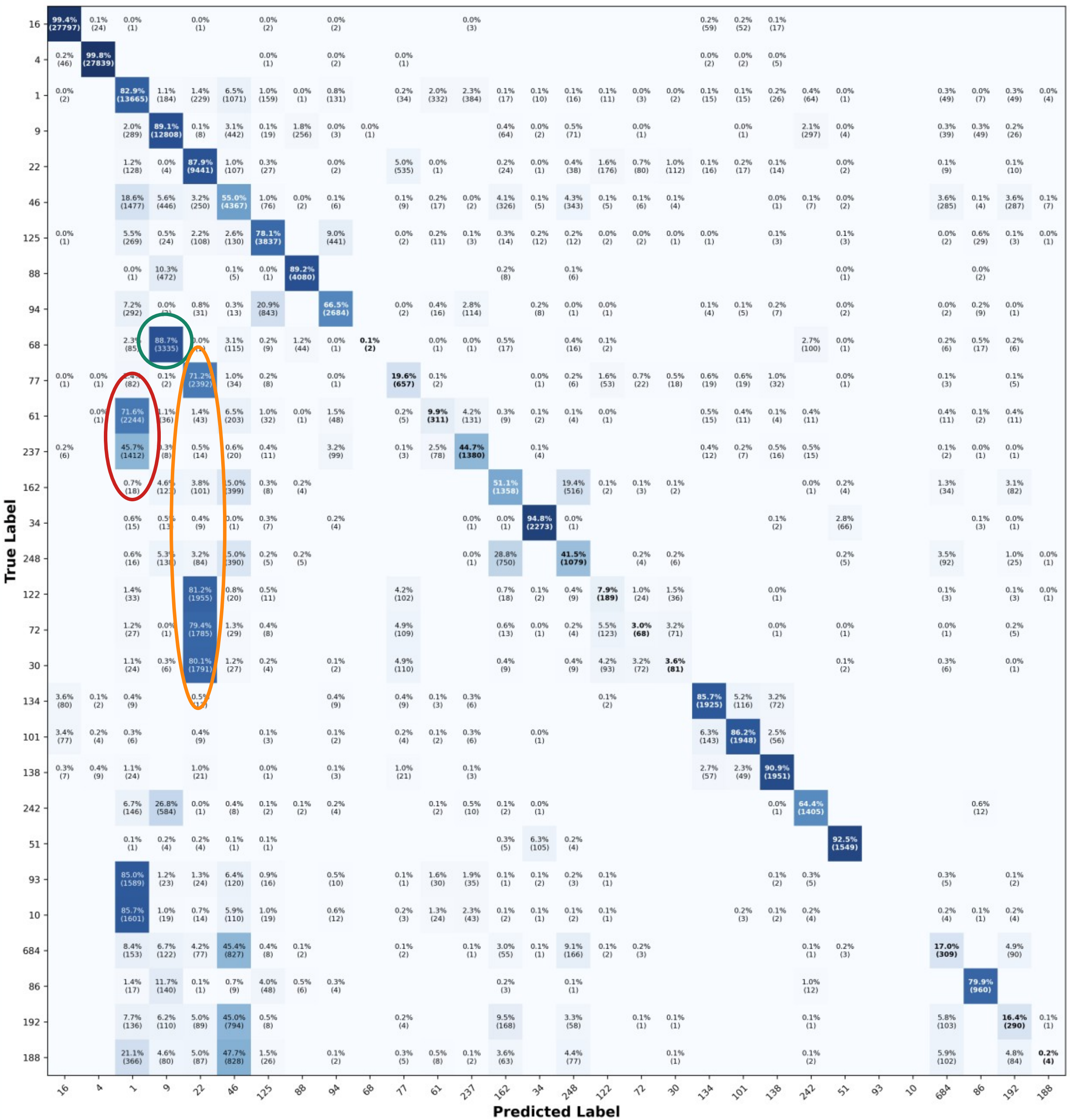
■ The remaining **4 clusters** contain events from **more than three different decay modes**.

Research Results



- Cluster 1:
  - ◎ $J/\psi \Rightarrow \pi^0\pi^+\pi^-K^+K^-$
  - ◎ $J/\psi \Rightarrow \pi^+\pi^-K^+K^-$
- Cluster 2:
  - ◎ $J/\psi \Rightarrow \pi^0\pi^+\pi^+\pi^+\pi^-\pi^-\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^+\pi^+\pi^-\pi^-\omega$, $\omega \Rightarrow \pi^0\pi^+\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^+\pi^+\pi^+\pi^+\pi^-\pi^-\pi^-\pi^-$
- Cluster 3:
  - ◎ $J/\psi \Rightarrow \pi^0\pi^+\pi^+\pi^-\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^+\pi^+\pi^-\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0\pi^+\pi^+\pi^-\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0\pi^-b^+_1$, $b^+_1 \Rightarrow \pi^+\omega$, $\omega \Rightarrow \pi^0\pi^+\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0\pi^-b^-_1$, $b^-_1 \Rightarrow \pi^+\omega$, $\omega \Rightarrow \pi^0\pi^+\pi^-$
- Cluster 4:
  - ◎ $J/\psi \Rightarrow \pi^+\rho^-$, $\rho^- \Rightarrow \pi^0\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0\pi^0\pi^+\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0\pi^+\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0h_1(1170)$, $h_1(1170) \Rightarrow \pi^0\rho^0$, $\rho^0 \Rightarrow \pi^+\pi^-$
  - ◎ $J/\psi \Rightarrow \pi^0\pi^0h_1(1170)$, $h_1(1170) \Rightarrow \pi^-\rho^+$, $\rho^+ \Rightarrow \pi^0\pi^+$

Research Results



Classification results:

■ C68 ⇒ C9: (88.7%)

C9 = $\pi^0\pi^+\pi^+\pi^+\pi^-\pi^-\pi^-$

C68 = $\pi^+\pi^+\pi^-\pi^-$**ω**, **ω** ⇒ $\pi^0\pi^+\pi^-$

■ C61 ⇒ C1: (71.6%)

■ C237 ⇒ C1: (45.7%)

C1 = $\boldsymbol{\pi^0}\pi^+\pi^+\pi^-\pi^-$

C61 = $\pi^+\pi^-$**ω**, **ω** ⇒ $\pi^0\pi^+\pi^-$

C237 = $\pi^+\pi^+\pi^-\pi^-$

■ C77 ⇒ C22: (71.2%)

■ C122 ⇒ C22: (81.2%)

■ C72 ⇒ C22: (79.4%)

■ C30 ⇒ C22: (80.1%)

C22 = $\boldsymbol{\pi^0}\pi^0\pi^0\pi^+\pi^-$

C77 = $\pi^0\pi^0\pi^+\pi^-$

C122 = $\pi^0\pi^0h_1(1170)$, **h_1 (1170)** ⇒ $\pi^0\rho^0$, $\rho^0 \Rightarrow \pi^+\pi^-$

C72 = $\pi^0\pi^0h_1(1170)$, **h_1 (1170)** ⇒ $\pi^-\rho^+$, $\rho^+ \Rightarrow \pi^0\pi^+$

C30 = $\pi^0\pi^0h_1(1170)$, **h_1 (1170)** ⇒ $\pi^+\rho^-$, $\rho^- \Rightarrow \pi^0\pi^-$

★ not sensitive to **neutral track** and **resonance state**.

Research Results

Comparison with Standard Approaches:

Table 1: Performance Comparison on Different Decay Channels

	Foundation Model			Traditional Method		
Decay Channel	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
$J/\psi \to \pi^+\rho^-, \rho^- \to \pi^0\pi^-$	0.72	**0.79**	0.75	0.72	0.74	0.73
$J/\psi \to \pi^-\rho^+, \rho^+ \to \pi^0\pi^+$	0.70	**0.79**	0.74	0.71	0.74	0.72
$J/\psi \to \pi^0\rho^0, \rho^0 \to \pi^+\pi^-$	0.82	**0.87**	0.85	0.82	0.82	0.82
$J/\psi \to \pi^+\pi^-\pi^0$	0.74	**0.76**	0.74	0.74	0.70	0.71

◎ For all events that are well separated by traditional methods, our foundation model achieves comparably high classification accuracy.

◎ Compared with traditional physics-based methods, **foundation models achieve higher recall at the same precision level**.

◎ Traditional physics-based methods can **further increase precision above 99% by sacrificing recall**, which **foundation models currently cannot achieve**.

Part 4: Conclusions

Conclusions

1. We successfully developed a **foundation model** for high-energy physics collider data — the **Particle Masked Siamese Encoder (Particle MSE)** — built upon a transformer-based architecture. The model achieves promising performance across various downstream tasks, including **clustering, event-level classification and anomaly detection**.

2. There remains room for improvement for reconstruction in pretext task, particularly in **reproducing fine structures**.

3. The Particle MSE–based model shows **limited sensitivity to neural track** and **resonance states**, for event-level classification.

Future Work:

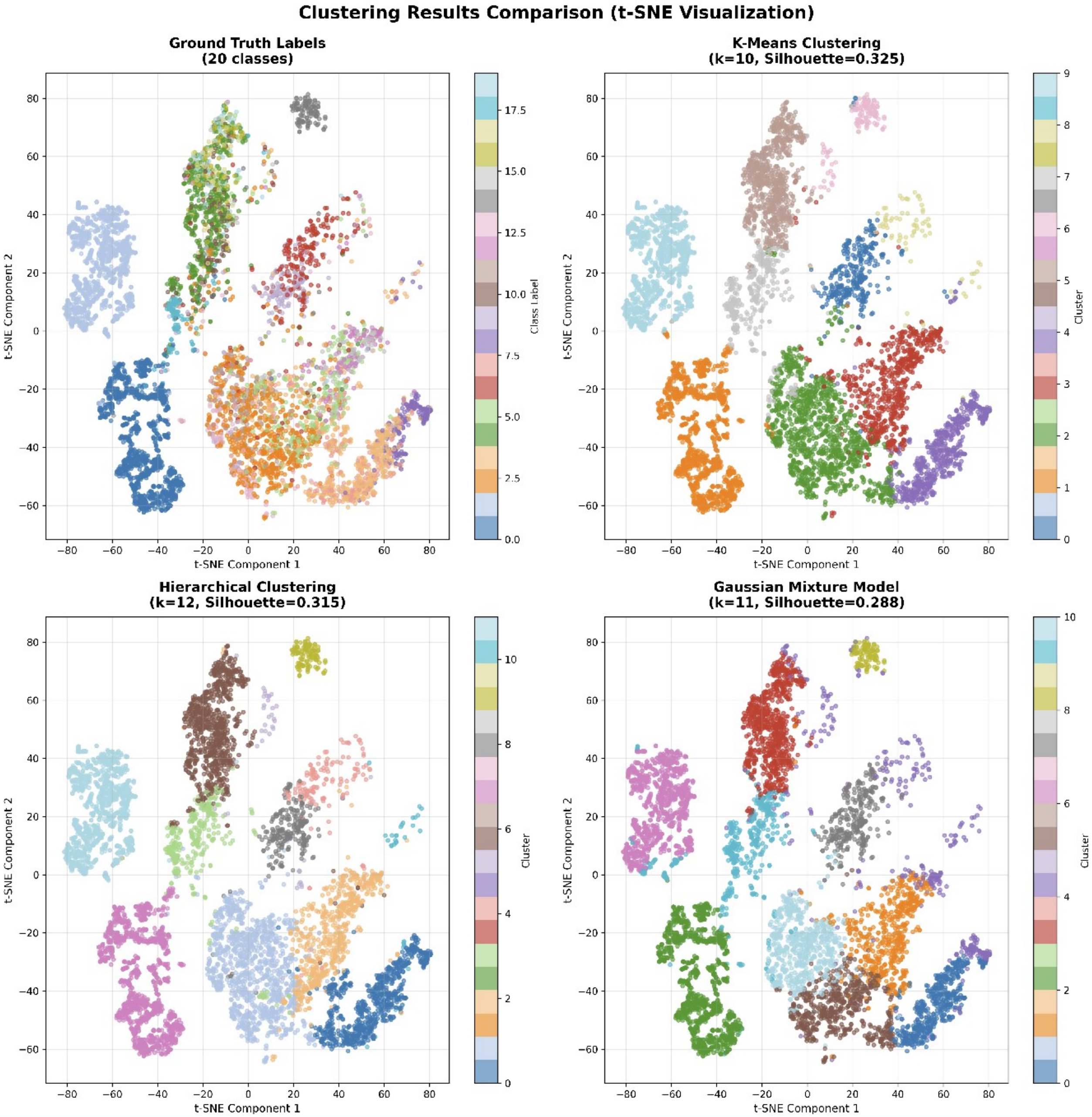
1. Address the model's insensitivity to neutral tracks and resonance states such as dynamic contrastive learning and hard mining.

2. Explore self-supervised contrastive learning approaches.

3. Explore more advance masking approaches.

Thank you for listening!
Any question?

Appendix



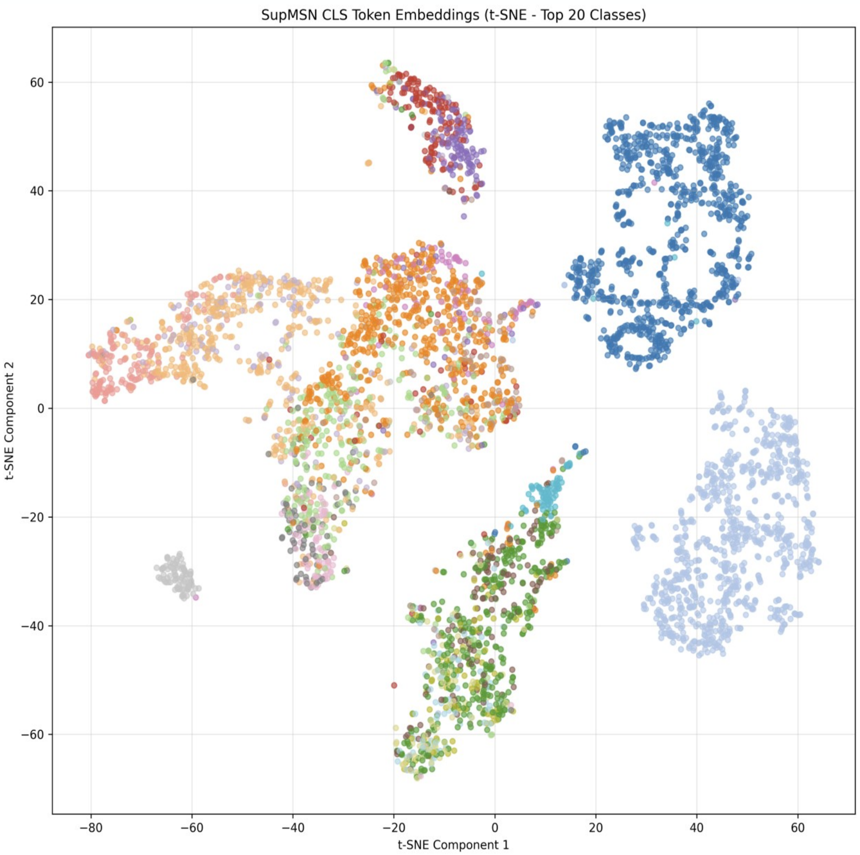
__Clustering Results:__

◉ The model can generally distinguish the decay-event timings **for the three clusters.**

✗However, **for the central four clusters**, the patterns, which have multiple decay events **remain difficult to separate**.

✗In future work, **Dynamic contrastive learning** will be incorporated to enhance clustering performance and improve class-level separability.

Appendix



Inter-U