Foundation Model for Physics Analysis at BESIII

CLEAR (Contrastive Learning for Event Attention-based Representation)

Jingde Chen, Zijie Shang, Tong Liu, Beijiang Liu, Ke Li

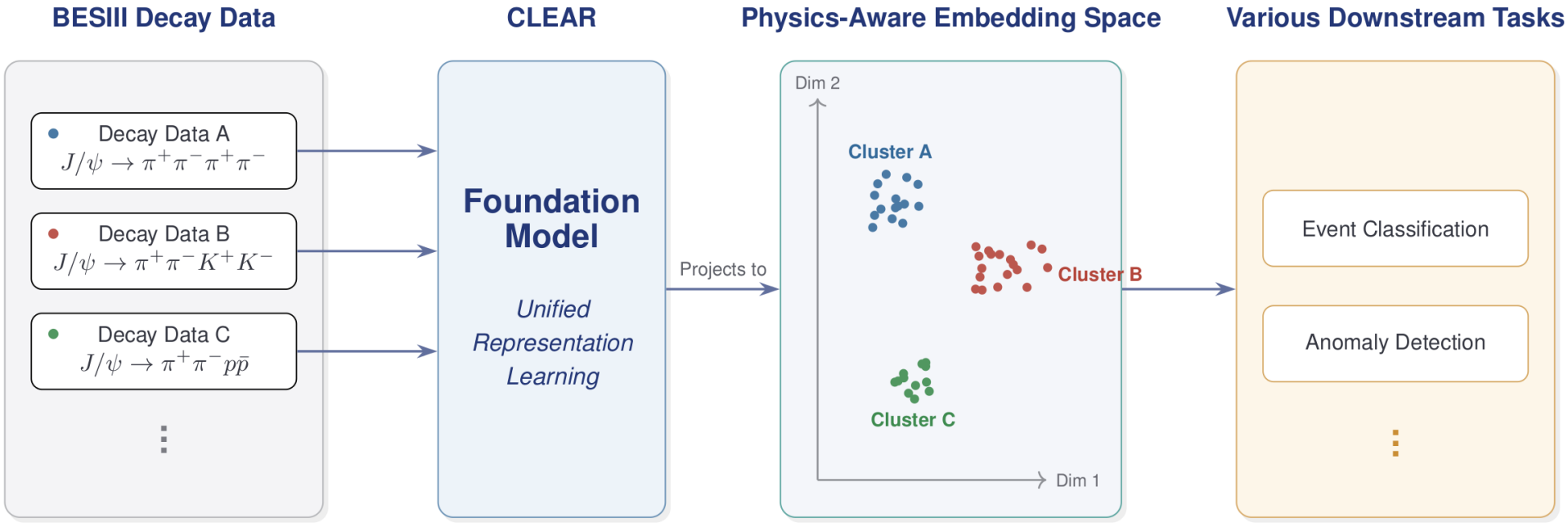
Institute of High Energy Physics, Chinese Academy of Sciences

Quantum Computing and Machine Learning Workshop
Shenzhen, 19 Aug 2026

Introduction

Foundation Model in BESIII:

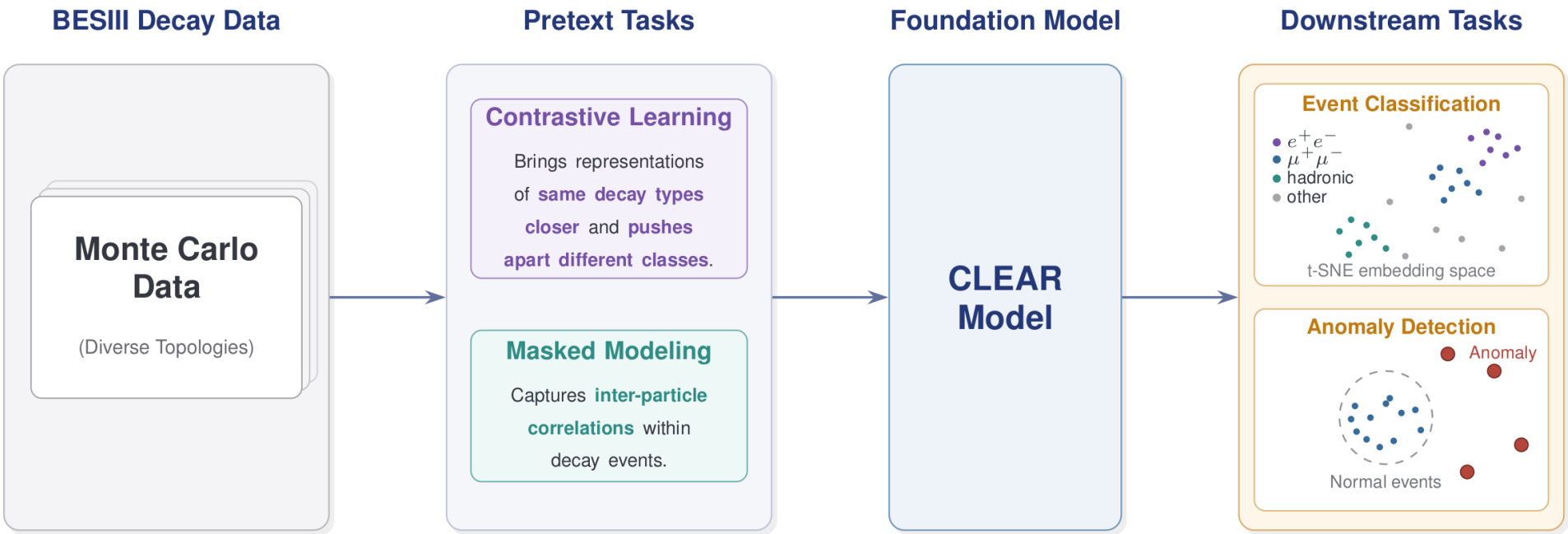
★ **Research Goal**: Develop the **CLEAR** model at BESIII to learn a **unified representation** that **naturally clusters all decay topologies** in BESIII, enabling the model to organize diverse physics processes into a physics-aware embedding space to adapt various downstream tasks.



Introduction

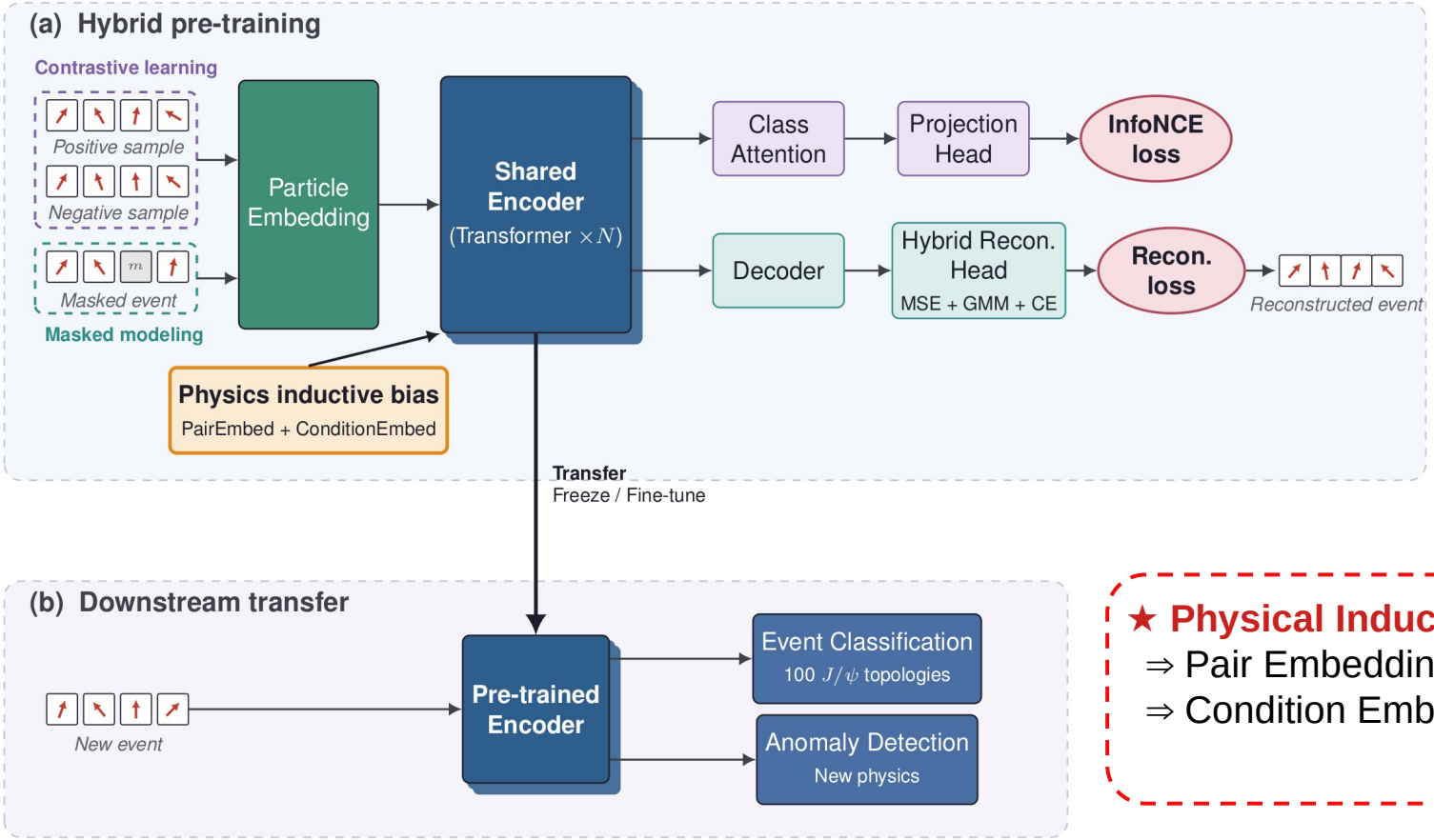
Pretext Tasks and Downstream Tasks:

- Pretext tasks: **Contrastive Learning** + **Masked Modeling**
- Downstream tasks: Decay event classification + Anomaly Detection



Research Method

CLEAR Structure:



★ **Physical Inductive Bias**

⇒ Pair Embedding (Pair-level)

⇒ Condition Embedding (Event-level)

Research Method

Loss Function:

- **Hybrid pre-training objective**

$$\mathcal{L}_{\text{pretrain}} = w_1 \mathcal{L}_{\text{MAE}} + w_2 \mathcal{L}_{\text{Con}}$$

Masked modeling

$$\mathcal{L}_{\text{MAE}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{GMM-NLL}} + \mathcal{L}_{\text{CE}}$$

- **MSE** — ordinary continuous variables
- **GMM-NLL** — multimodal continuous variables
- **Cross-entropy** — categorical variables

Masked one valid particle only.

Contrastive learning

$$\mathcal{L}_{\text{Con}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\boldsymbol{z}_i^a \cdot \boldsymbol{z}_i^p / \tau\right)}{\sum_{j=1}^{B} \exp\left(\boldsymbol{z}_i^a \cdot \boldsymbol{z}_j^p / \tau\right)}$$

- $\tau = 0.1$
- **Positive** — another event from answer pool
- **Negatives** — the other $B - 1$ events in the batch

Supervised contrastive learning.

- **Curriculum weighting**

Epochs 1–20	→	Epochs 21–60	→	Epochs 61–100
$1.0\,\mathcal{L}_{\text{MAE}} + 0.3\,\mathcal{L}_{\text{Con}}$		$1.0\,\mathcal{L}_{\text{MAE}} + 0.7\,\mathcal{L}_{\text{Con}}$		$0.8\,\mathcal{L}_{\text{MAE}} + 1.0\,\mathcal{L}_{\text{Con}}$
Reconstruction first		Joint learning		Representation refinement

Research Method

From InfoNCE to Representation Geometry:

$$\mathcal{L}_{\text{InfoNCE}} \xrightarrow[\text{Asymptotically [1]}]{} \underbrace{\ell_{\text{align}}}_{\text{positive attraction}} + \underbrace{\ell_{\text{unif}}}_{\text{negative repulsion}}$$

Positive attraction

$$\ell_{\text{align}} = \mathbb{E}_{(x,y)\sim p_{\text{pos}}} \left\| f(x) - f(y) \right\|_2^2$$

- Positive pairs: independent events from the same class.
- Minimizing ℓ_{align} forms compact local neighborhoods.

Pull positive event representations together.

Negative repulsion

$$\ell_{\text{unif}} = \log \mathbb{E}_{x,y\sim p_{\text{data}}} \exp\left[-2\left\| f(x) - f(y) \right\|_2^2\right]$$

- All normalized events define the global distribution.
- Minimizing ℓ_{unif} spreads features and prevents collapse.

Repel representations at the global distribution level.

- The decomposition becomes exact only in the asymptotic limit of an infinitely large batch, i.e., infinitely many negative samples.
- Hypothesis: classification should favor compact clusters (smaller λ), anomaly detection broader coverage (larger λ) — the two tasks pull in opposite directions.

[1]. T. Wang and P. Isola, Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, ICML 2020

Research Method

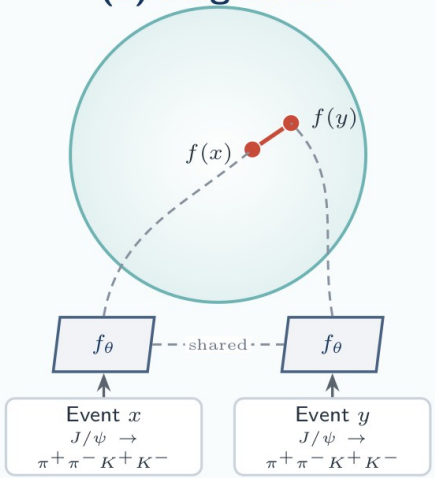
Alignment and Uniformity:

◎ Alignment: Closeness of features from positive pairs.

◎ Uniformity: Normalized representations over the hypersphere, preventing global collapse.

◎ A Good Representation: **Linearly separable** and **clusterable**. (kNN)

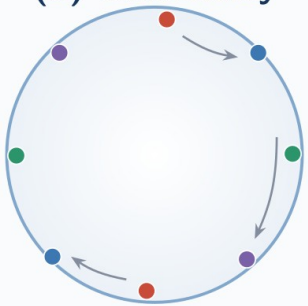
(a) Alignment



Two independent events are sampled from the same decay class.

$$\ell_{\text{align}} = \mathbb{E}_{(x,y)\sim p_{\text{pos}}} \left\| f(x) - f(y) \right\|_2^2$$

(b) Uniformity

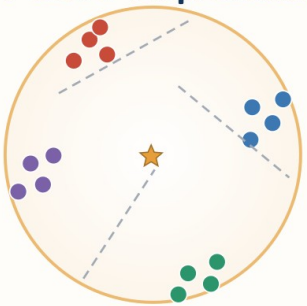


- $J/\psi \to \pi^+\pi^-K^+K^-$
- $J/\psi \to \pi^+\pi^-\pi^+\pi^-$
- $J/\psi \to \pi^+\pi^-p\bar{p}$
- $J/\psi \to \pi^+\pi^-n\bar{n}$

100 decay modes used in pre-training.

$$\ell_{\text{unif}} = \log \mathbb{E}_{x,y\sim p_{\text{data}}} \exp\left[-2\left\| f(x) - f(y) \right\|_2^2\right]$$

(c) A Good Representation



- Each decay mode forms a tight spherical cap ⇒ linearly separable (linear probe).
- Uniform coverage leaves empty regions ⇒ anomalous events become detectable.

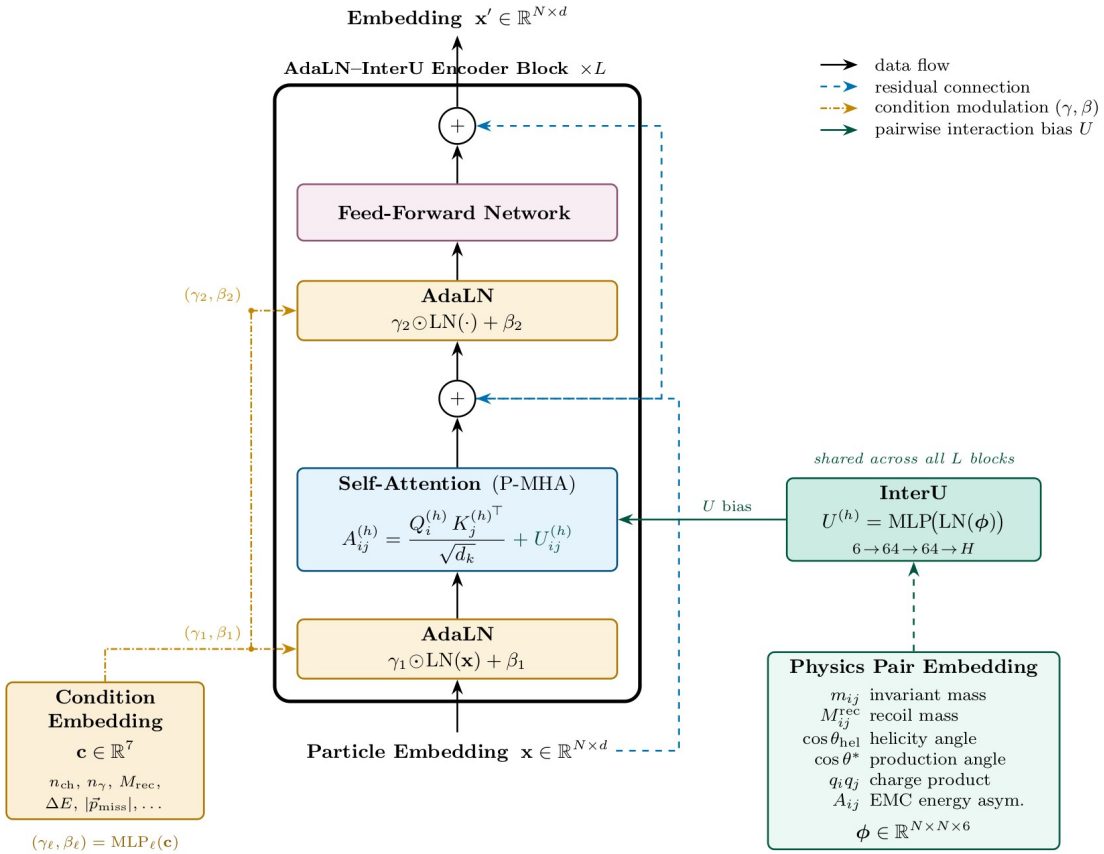
$$\mathcal{L}_{\text{con}}(\lambda) = \frac{\ell_{\text{align}} + \lambda\, \ell_{\text{unif}}}{1+\lambda}$$

Research Method

Physical Inductive Bias for Encoder:

◉ **Event-level feature**:
- ⇒ Number of the total charged tracks
- ⇒ Number of the total neutral tracks
- ⇒ Charged w/o EMC match
- ⇒ Neutral w/o EMC match
- ⇒ M_recoil (Recoil invariant mass)
- ⇒ E_miss (Missing energy)
- ⇒ |p_miss| (Missing momentum)

◉ **Pair-level feature**:
- ⇒ m_ij (Two-body invariant mass)
- ⇒ M_rec (Recoil invariant mass)
- ⇒ cosθ_hel (Helicity angle)
- ⇒ cosθ* (Production polar angle)
- ⇒ q_i·q_j (Charge product)
- ⇒ A_ij (EMC energy asymmetry)



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 100** most frequent classes are used for model pretraining.

◎ Training dataset: 100 classes of J/ψ decay events (**2M; 5M; 10M samples**)

◎ 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_dof']

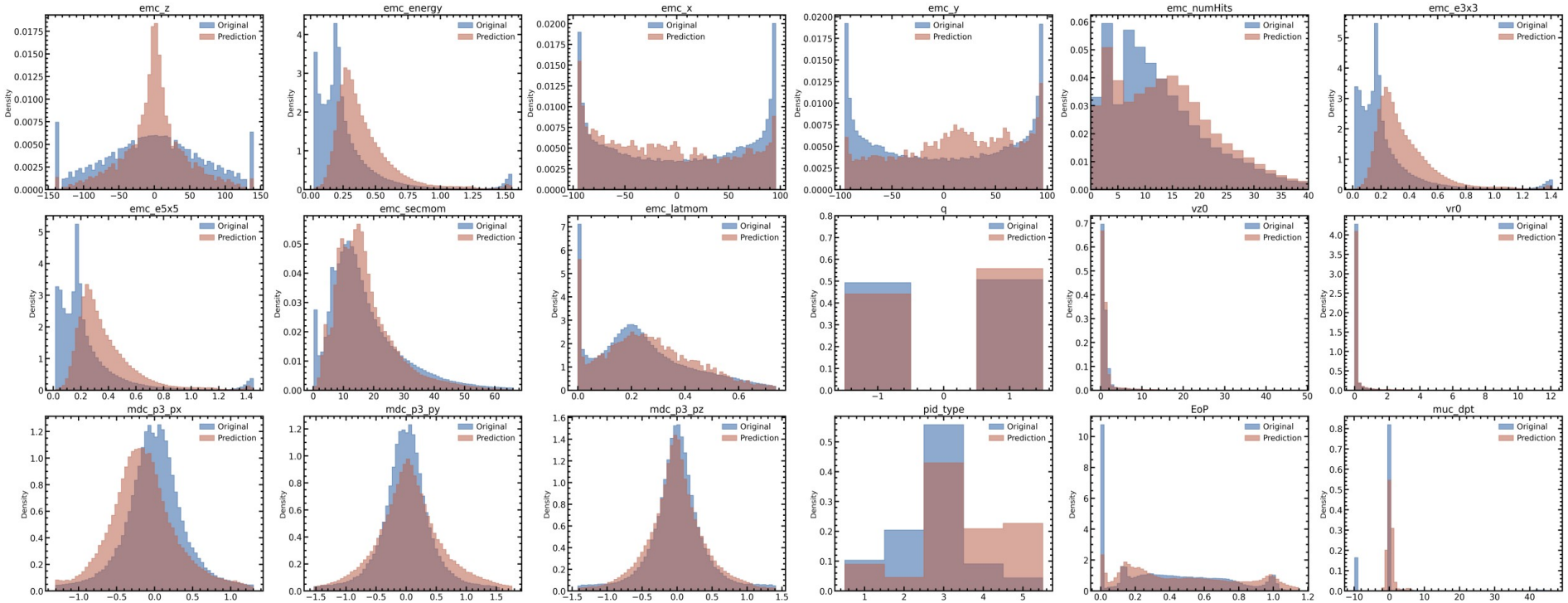
Downstream Tasks:

◉ Task 1 – **100 classes classification task**.

◉ Task 2 – **Anomaly Detection**.

Research Results

Prediction Results:

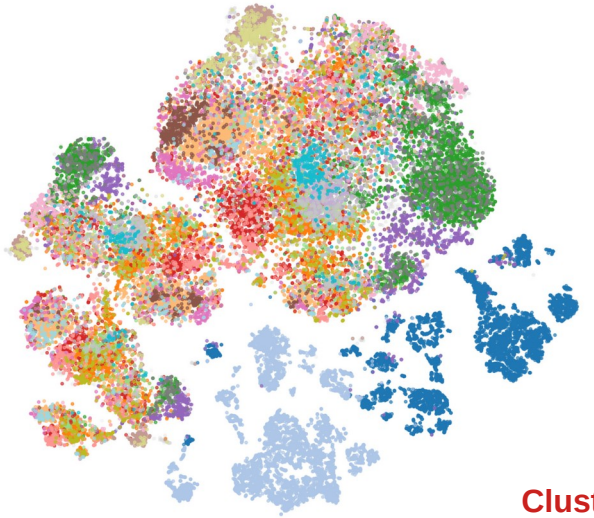


- The model reasonably reproduces the overall distributions of most continuous variables.
- Further improvement is needed for discrete variables and features with complex non-Gaussian distributions.

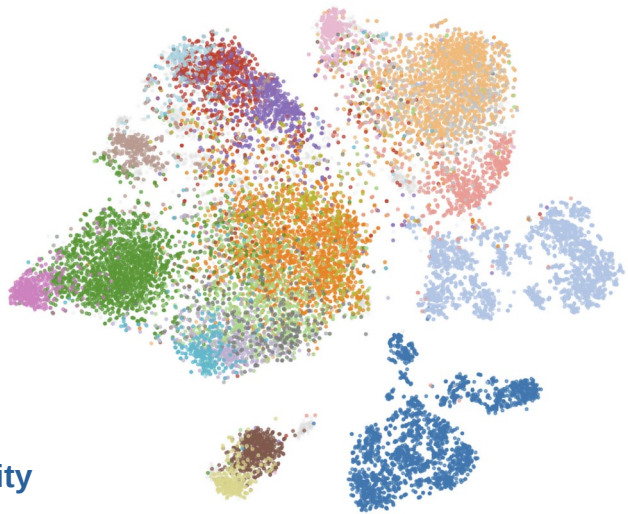
Research Results

Ablation Study:

◎ Without Contrastive



◎ With Contrastive



t-SNE

Clusterability and **Linear separability**

	k-NN	Linear
w/o Contrastive	47.39%	58.77%
w/ Contrastive	**54.33%**	**59.79%**

■ Based on the t-SNE visualizations, incorporating contrastive learning yields measurably **improved representation quality**.

Research Results

λ-sweep test for classification:

	Decomposed $\mathcal{L}_{\text{Con}}(\lambda) = (\ell_{\text{align}} + \lambda\,\ell_{\text{unif}})/(1+\lambda)$				
Probe	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 1.0$	$\lambda = 10.0$
k-NN	—	58.99	61.01	**63.26**	62.06
Linear	—	64.60	65.91	**66.85**	65.35
MLP-2L	—	66.03	66.92	**67.56**	66.12
MLP-3L	—	65.90	66.84	**67.48**	66.07

Top-1 accuracy (%), 100-class task, frozen encoder, epoch-100 checkpoints.
$\lambda = 0$: collapsed at epoch 11.

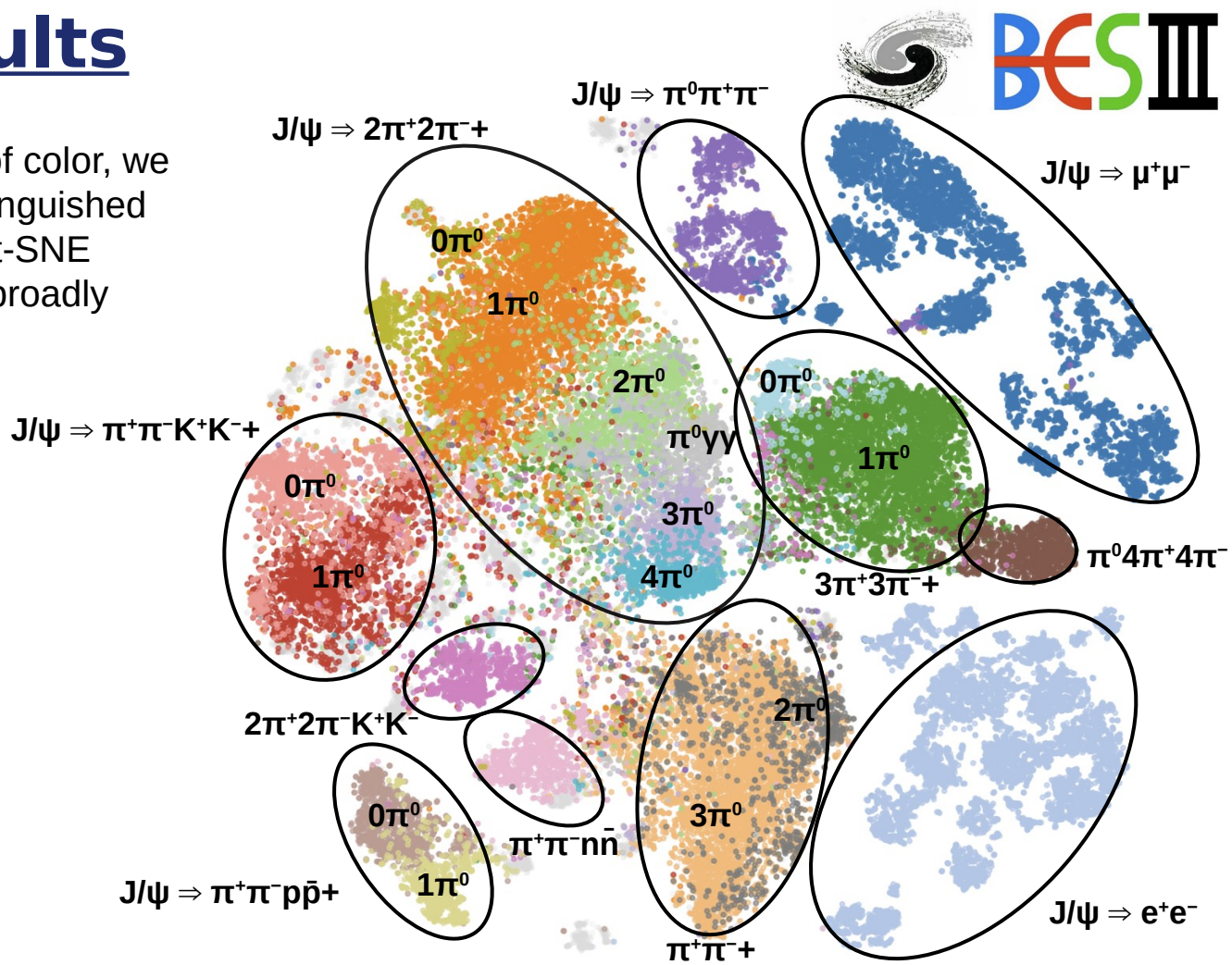
- All four probes achieve their **highest accuracy at λ = 1**, suggesting an interior optimum with a balanced alignment–uniformity trade-off.
- The **alignment-only λ = 0 collapses** at epoch 11, demonstrating that uniformity is essential for stable representation learning.
- Less expressive probes—particularly **k-NN**—are more **sensitive to λ**.

  ⇒ λ directly shapes the local neighborhood geometry of the representation space.

Research Results

Clustering Results:

■ Due to the limited availability of color, we present the **top 20 classes** distinguished by their final-state decay. In the t-SNE figure, these categories can be broadly grouped into **11 clusters**.



Research Results

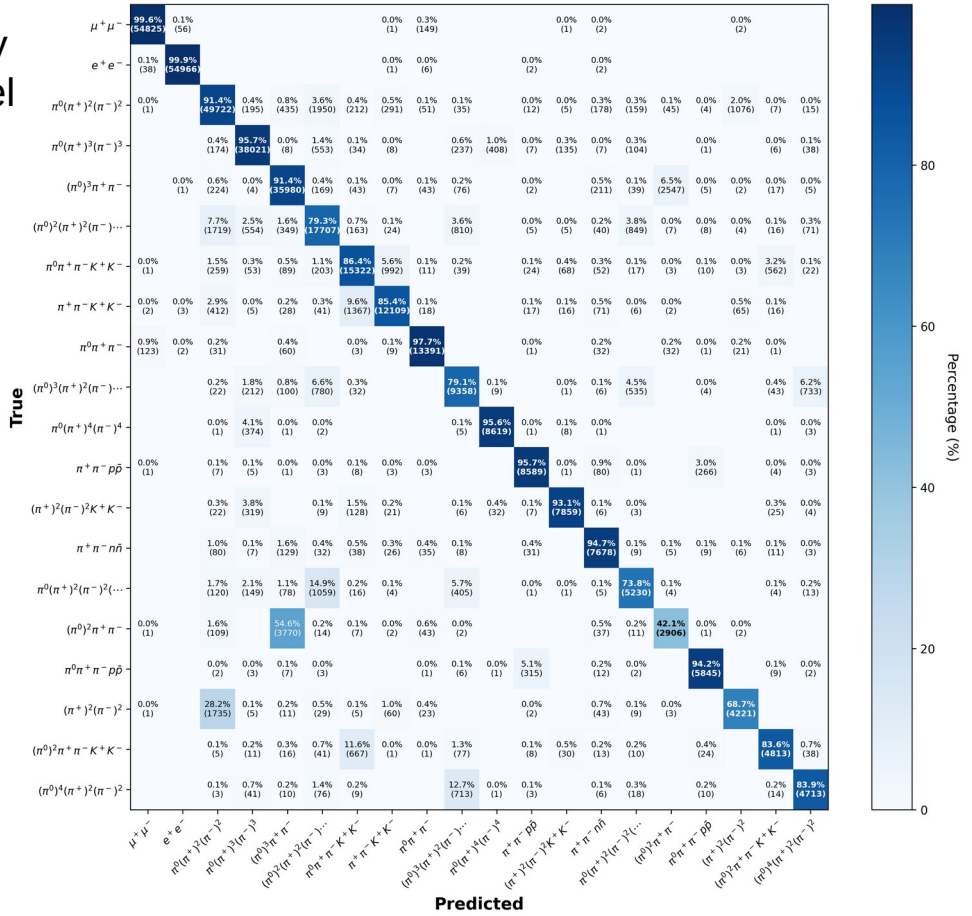
Classification Results:

■ Due to the limited size of the confusion matrix, only the **top 20 classes** are presented. Overall, the model achieves satisfactory classification performance for most categories.

■ However, several cases still exhibit relatively high misclassification rates (exceeding 20%), including:

◎ J/ψ ⇒ **2π⁰**(π⁺π⁻) misclassified as **3π⁰**(π⁺π⁻)

◎ J/ψ ⇒ (2π⁺2π⁻) misclassified as **π⁰**(2π⁺2π⁻)



Research Results

Fine-tuning strategies:

- Baseline: **Frozen encoder**
  ⇒ k-NN / Linear / 2L MLP / 3L MLP

- Strategy A: **CA-Only Fine-Tuning**
  ⇒ Freeze SA encoder; **unfreeze only Class Attention**.

- Strategy B: **End-to-End Fine-Tuning** with differential learning rate (LR).

- Strategy C: **Multitask Fine-Tuning**
  ⇒ Joint training with reconstruction loss.

- Strategy D: **Two-Stage Fine-Tuning**
  ⇒ Stage 1: train CA + Head only until convergence.
  ⇒ Stage 2: unfreeze full model with very low LR.

$$\mathrm{Acc} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathbf{1}[y_{\mathrm{pred},i} = y_{\mathrm{true},i}]$$

Strategy	Acc (100-class)
k-NN	63.26%
Linear	66.85%
MLP-2L	67.56%
MLP-3L	67.48%
[A] CA-Only	68.71%
[B] E2E	70.24%
[C] Multi-Task	**70.26%**
[D] Two-Stage	69.85%

CLEAR pretrained with $\lambda = 1$, epoch-100.

Research Results

OOD and Scorer:

■ **Near-OOD**: seen final state, unseen resonance path; **Far-OOD**: final state never seen in pre-training.

■ Five scorers, frozen embedding, 50k-event reference set, per-class AUC

★ Mahalanobis: **best on 6/8 classes**, **never worst**, and stable in the reference-set size — adopted as the primary read-out.

	Near-OOD				Far-OOD			
Scorer	13	356	192	906	229	1447	214	240
Mahalanobis	**0.77**	**0.77**	**0.69**	**0.73**	0.89	0.78	**0.87**	**0.87**
kNN-distance	0.69	0.69	0.62	0.63	0.83	**0.84**	0.78	0.78
Ensemble	0.73	0.70	0.58	0.68	**0.95**	0.73	0.82	0.81
IsolationForest	0.60	0.57	0.37	0.49	**0.95**	0.49	0.65	0.62
Recon-MSE	0.47	0.41	0.42	0.51	0.54	0.49	0.42	0.42

AUC at $\lambda = 1$. Values are rounded to two decimal places. **Bold** indicates the best among the five scorers shown, based on the unrounded AUC.

13: $J/\psi \to \pi^0\pi^+\pi^- p\bar{p}$ $(\pi^+\Delta^+\bar{\Delta}^{--})$

356: $J/\psi \to 4\pi^0 2\pi^+ 2\pi^-$ $(\pi^0\eta\, b_1^0)$

192: $J/\psi \to 3\pi^0 2\pi^+ 2\pi^-$ $(\pi^0\eta'\, b_1^0)$

906: $J/\psi \to \pi^+\pi^- p\bar{p}$ $(\Delta^{++}\bar{\Delta}^{--})$

229: $J/\psi \to \pi^0 p\bar{p}$

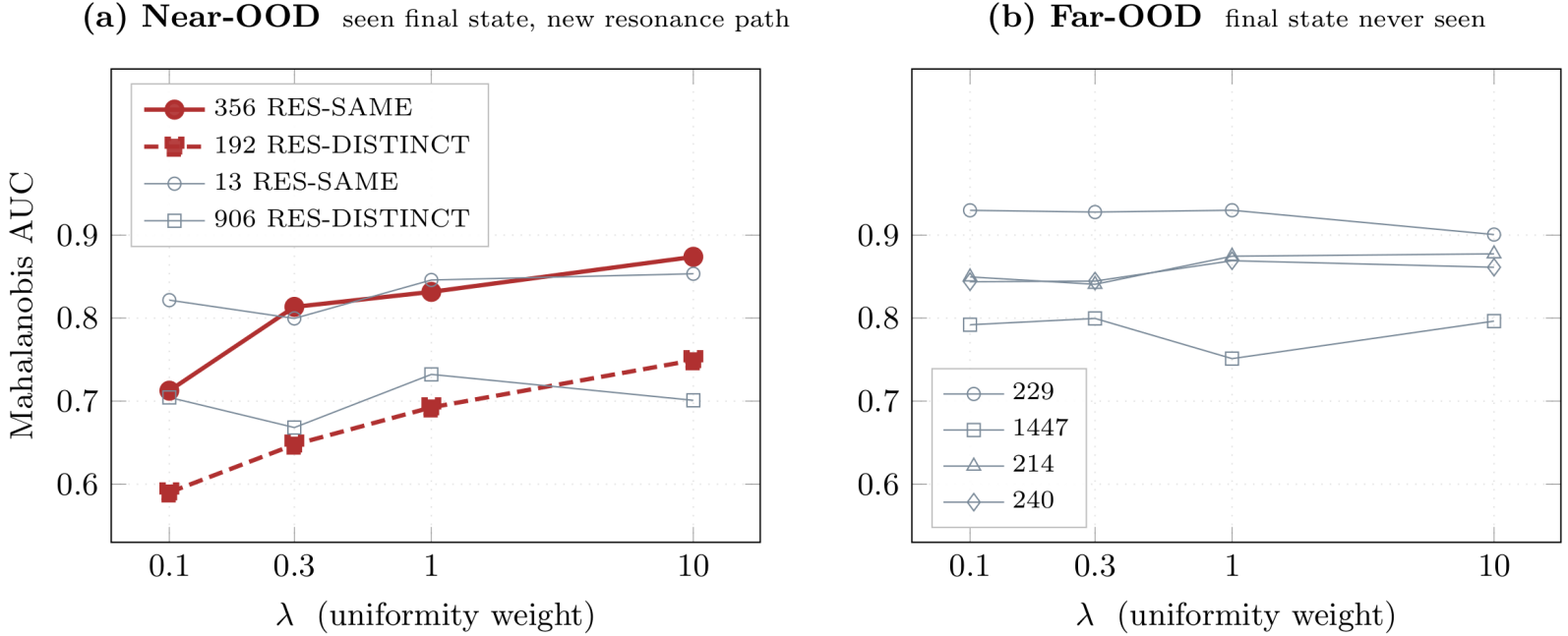
1447: $J/\psi \to K_L^0 K_L^0 2\pi^+ 2\pi^-$

214: $J/\psi \to \pi^0 4\pi^+ 3\pi^- K^-$

240: $J/\psi \to \pi^0 3\pi^+ 4\pi^- K^+$

Research Results

λ-sweep for Anomaly Detection:

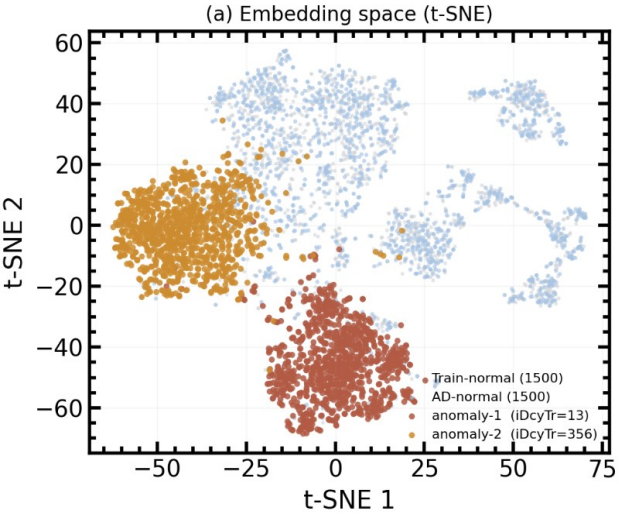


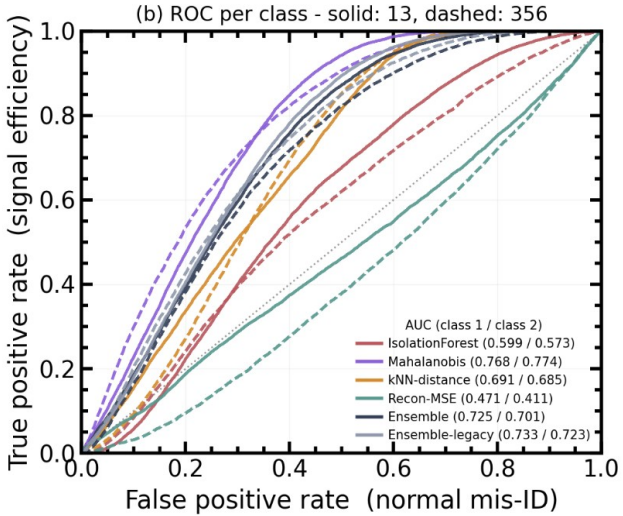
■ For the near-OOD classes, the **Mahalanobis** AUC rises **almost monotonically with λ**: a broader, more uniform embedding helps expose unseen resonance paths.

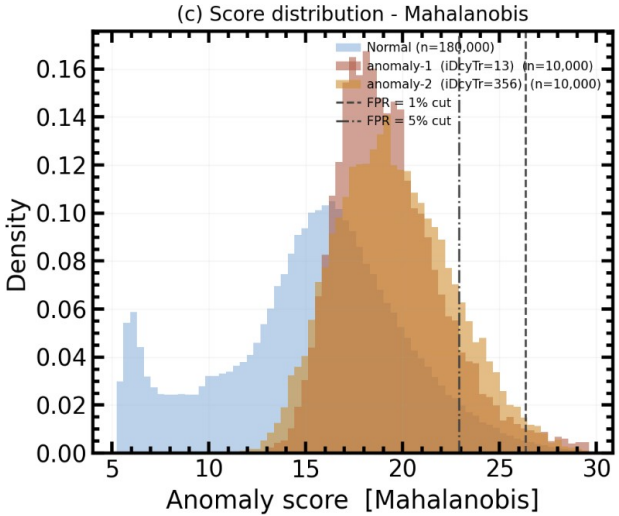
■ Far-OOD performance shows no consistent monotonic dependence on λ, suggesting that uniformity is not the controlling factor in this regime.

Research Results

Near-OOD Anomaly Detection Results:





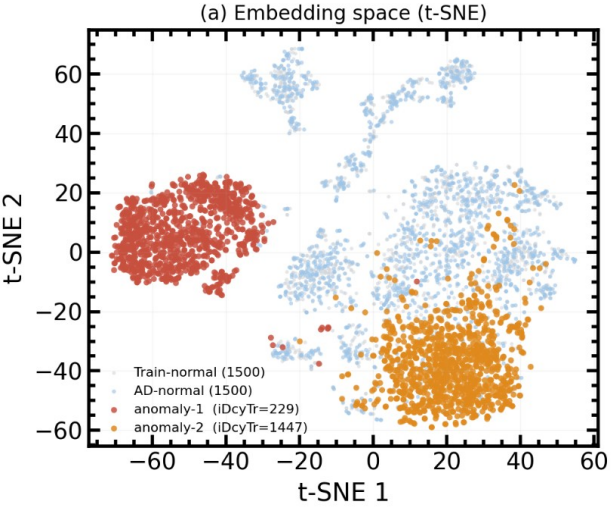


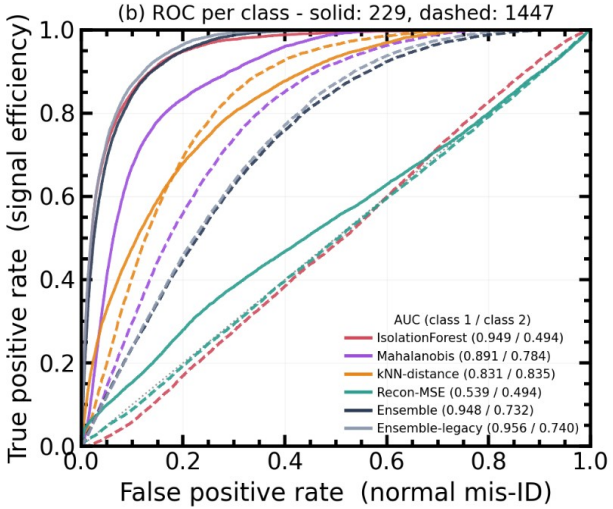
■ Near-OOD: Anomaly-1 = J/ψ ⇒ $\pi^0\pi^+\pi^-p\bar{p}$ ($\pi^+\Delta^+\bar{\Delta}^{--}$) (iDcyTr 13) and Anomaly-2 = J/ψ ⇒ $4\pi^0 2\pi^+ 2\pi^-$ ($\pi^0\eta\, b_1^0$) (iDcyTr 356)

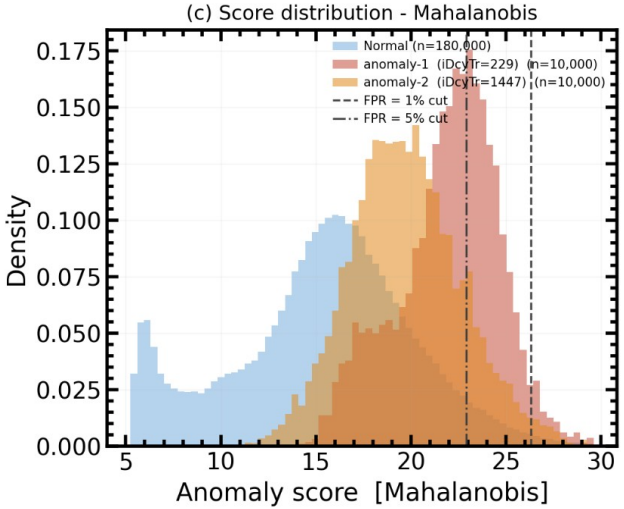
■ Both anomaly classes form compact clusters distinct from normal events with the same final states, indicating sensitivity to intermediate resonance structures beyond final-state composition. **Mahalanobis** distance **performs best and consistently across both classes**.

Research Results

Far-OOD Anomaly Detection Results:







■ Far-OOD: Anomaly-1 = J/ψ ⇒ $\pi^0 p\bar{p}$ (iDcyTr 229) and Anomaly-2 = J/ψ ⇒ $K^0_L K^0_L 2\pi^+ 2\pi^-$ (iDcyTr 1447).

■ The two anomaly classes occupy distinct embedding regions but exhibit different geometric characteristics. IsolationForest and ensemble methods perform well on class 229, whereas **kNN distance is the only method achieving an AUC above 0.83 for both**.

Research Results

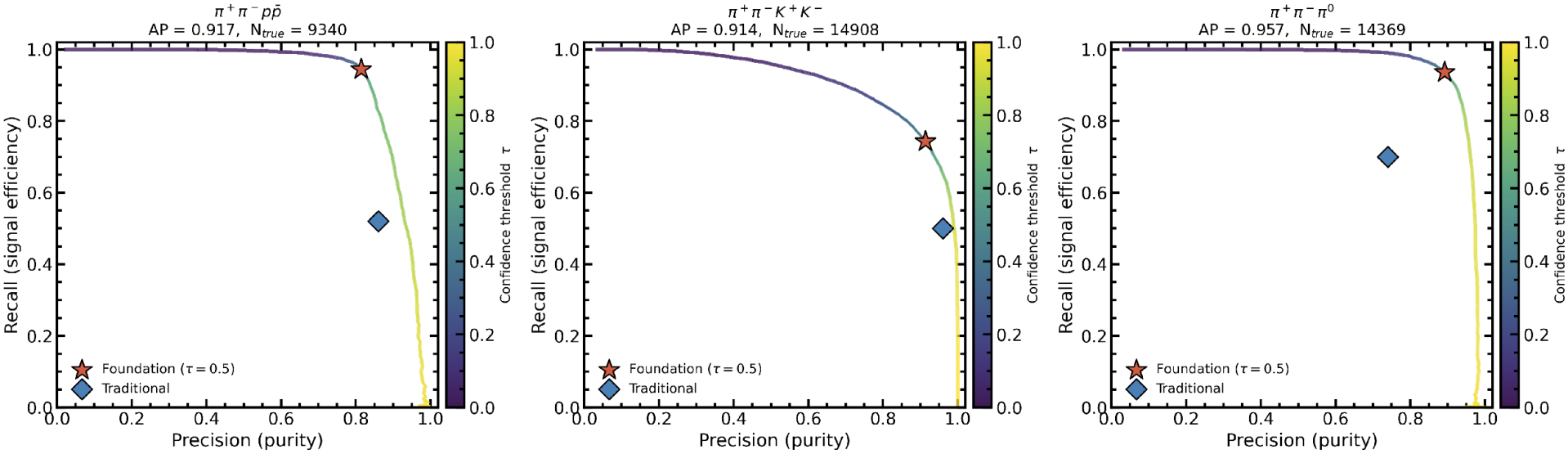
Data-Scaling Study:

	100 Class Classification Accuracy (%)					
Pre-training Data	**MLP-2L**	**MLP-3L**	**CA-Only**	**E2E**	**Multi-Task**	**Two-Stage**
2M	67.56	67.48	68.71	70.24	70.26	69.85
10M	68.91	68.53	72.43	74.41	74.42	74.82
Improvement (Δ)	+1.35	+1.05	+3.72	+4.17	+4.16	+4.97

◎ With model capacity fixed, scaling the pre-training data from 2M to 10M events consistently improves performance across all downstream strategies, with gains ranging from 1.05% to 4.97% and a best accuracy of 74.82%.

Research Results

Comparison with Standard Approaches:



◎ Across the three tested decay channels, **CLEAR achieves higher recall** than traditional physics-based selections at comparable precision.

Conclusions

1. The CLEAR model is **a foundation model for BESIII** event analysis: a single pre-trained encoder (masked modelling + supervised contrastive learning) learns a physics-aware representation that transfers to both event classification and anomaly detection.

2. Across the three tested decay channels, **CLEAR achieves higher recall** than traditional physics-based selections at comparable precision.

3. Beyond classification, the pretrained representation transfers to anomaly detection: far-OOD final states **form their own distinct clusters** and are more readily separated, while near-OOD events that share a seen final state remain substantially harder.

Future Work:

1. Incorporate physics constraints such as energy-momentum conservation directly into the model.

2. Scale pretraining from 10M toward 100M events across more topologies.

Thank you for listening!