DECAYFM: A PHYSICS-STRUCTURED FOUNDATION MODEL FOR DATA-EFFICIENT RARE PARTICLE DECAY IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying particle decay modes from reconstructed detector measurements is essential to high-energy physics, but building separate classifiers for many channels is costly, particularly when downstream training samples are limited. This setting presents two challenges for transferable representation learning: heterogeneous physical measurements contribute differently across decay modes, and highly imbalanced training frequencies can bias learning toward common channels. We introduce DecayFM, a physics-structured foundation model for data-efficient particle decay identification. DecayFM separately encodes kinematic, energy-deposition, and particle-identification features, integrates them through a particle-set Transformer, and applies paired low-rank adapters whose contributions are weighted by a learned head–tail router. A two-stage training procedure separates shared representation learning from frequency-aware adaptation. We construct the ParticleDecayMode (PDM) dataset series using BESIII simulation, comprising approximately 8.3 million J/ψ decay events for large-scale pretraining and downstream evaluation, including a benchmark of 200 rare decay modes. On rare-mode classification tasks, pretrained DecayFM improves accuracy over the reported from-scratch baseline by up to 11.61 percentage points. In the evaluated few-shot settings, parameter-efficient adaptation with 10-100 training events per class exceeds the accuracy of the corresponding baseline trained with 1,000 events per class. These results support physics-structured pretraining as a promising approach to reducing downstream training-data requirements for rare decay identification.

1 INTRODUCTION

Modern high-energy physics experiments (Benedikt et al., 2025) rely on identifying particle decay modes to reconstruct the underlying physical processes from observed collision events. Given the reconstructed information of final-state particles in an event, particle decay identification aims to determine the decay channel that produced the observed final-state configuration (Sirunyan et al., 2020). This task provides a basis for a broad range of physics analyses, including measurements of decay properties (Yuan, 2021), studies of hadron spectroscopy (Destefanis, 2016), and searches for rare or forbidden processes (Chen et al., 2023).

The Beijing Spectrometer III (BESIII) experiment (Wang, 2006) at the Beijing Electron Positron Collider (BEPCII) provides a particularly rich setting for particle decay identification. BESIII has accumulated approximately 10^{10} J/ψ decay events, where the J/ψ is the ground-state vector charmonium state (Yuan & Olsen, 2019). Its exceptionally large and clean data sample enables studies of hundreds of exclusive decay channels, providing important access to light-hadron spectroscopy and rare or forbidden processes. In particular, rare decays with very small branching fractions yield only limited signal samples, making reliable identification essential for isolating weak signals and improving experimental sensitivity to such processes (Ablikim et al., 2024).

However, the diversity of J/ψ decays also creates a fundamental identification challenge. The occurrence rates of different decay modes can differ substantially, resulting in a highly long-tailed distribution in which a small number of frequent modes dominate the available training samples while

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

many rare modes have only limited supervision (Figure 2). At the same time, different decay modes can rely on different combinations of kinematic, energy-deposition, and particle-identification information extracted from the reconstructed final-state particles. These two forms of heterogeneity, namely statistical imbalance across decay modes and heterogeneous dependence on physical information, make it difficult for a single shared representation to effectively serve both abundant and rare decay modes. This motivates a foundation model paradigm that learns transferable representations across diverse decay modes while adapting efficiently to their distinct data regimes.

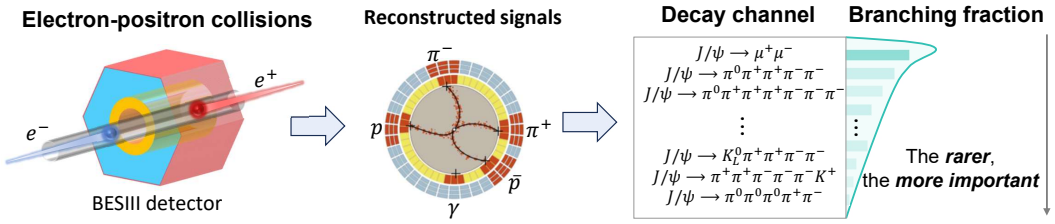


Figure 1: Illustration of the BESIII decay identification scenario. Electron-positron collisions at the J/ψ resonance produce abundant J/ψ events. The J/ψ particles undergo diverse decay processes, and the resulting final-state particles leave detector responses that are reconstructed into different decay channels with highly imbalanced branching fractions. We focus on the identification of rare decay modes.

To address these challenges, we propose a physics-structured foundation model for long-tailed particle decay identification, termed DecayFM, that separates physical representation learning from long-tail adaptation. The key idea is to structure heterogeneous physical information according to its underlying semantics, rather than forcing kinematic, energy, and particle identity information into a single shared representation. These complementary information sources are first encoded in separate representation subspaces and subsequently integrated to capture their interactions, allowing the model to preserve both shared structures across decay modes and information specific to individual physical aspects. Building upon these structured representations, DecayFM introduces dynamic adaptation to distinguish the representation needs of frequent and rare decay modes, using parameter-efficient adapters to allocate specialized transformations while retaining a shared transferable backbone. This design enables a single pre-trained model to jointly recognize diverse decay modes and provides an efficient basis for adapting to rare and previously unseen decay-identification tasks. To support the training and evaluation of DecayFM, we further construct the ParticleDecay-Mode (PDM) dataset series from the BESIII experiment, comprising approximately eight million simulated J/ψ decay events. The dataset provides a unified benchmark spanning diverse decay modes and highly imbalanced occurrence frequencies, supporting both large-scale pre-training and downstream evaluation.

Our contributions are summarized as follows:

- We construct the PDM dataset series from the BESIII experiment, comprising approximately eight million simulated J/ψ decay events. The dataset fills the gap in large-scale, systematic benchmarks for data-driven particle decay identification, enabling unified evaluation of models across diverse and highly long-tailed decay modes.
- We introduce a physics-structured and long-tail-adaptive learning paradigm for particle decay identification, which disentangles kinematic, energy, and particle identity information and dynamically adapts their representations to different data regimes. This enables scalable joint recognition across diverse decay modes while retaining physically interpretable representations.
- We establish DecayFM as a foundation model for particle decay analysis, providing transferable and physics-informed representations that can be efficiently adapted to rare and previously unseen decay-identification tasks through parameter-efficient fine-tuning with limited downstream data.

2 RELATED WORK

2.1 DECAY IDENTIFICATION

Traditional particle decay identification is commonly based on cutoff-based selection or multivariate analysis (MVA). Cutoff-based methods identify candidate events by applying thresholds to selected physical observables, while MVA-based approaches combine physics-motivated variables with classical classifiers such as boosted decision trees (Bhat, 2011). Although effective for individual decay channels, these approaches typically rely on channel-specific feature selection, variable construction, and hyperparameter tuning. As the number of target decay modes increases, such channel-specific designs become increasingly difficult to scale, making it challenging to build a unified model that can jointly distinguish a large number of diverse decay modes, such as hundreds of channels, within a single analysis framework.

2.2 FOUNDATION MODELS FOR PARTICLE PHYSICS

Jet tagging and particle decay identification share a common set-to-label formulation, mapping reconstructed particle-level information to underlying physics categories. This connection motivates us to review foundation-model approaches in jet physics, which can be broadly categorized into the following four paradigms.

Contrastive learning. Contrastive pre-training was an early route to physics-aware jet representations. JetCLR (Dillon et al., 2022a) maps constituents to a symmetry-aware embedding with a transformer encoder, encouraging approximate invariance to rotations, translations, and soft/collinear augmentations. Later work extends contrastive objectives to anomaly detection (Dillon et al., 2022b; 2024) and replaces hand-crafted views with re-simulation (Harris et al., 2025). The resulting embeddings are shared across jets and optimized for tagging or model-agnostic searches, rather than for distinguishing hundreds of exclusive decay modes with highly unequal occurrence rates.

Masked modeling. Masked modeling adapts BERT-style objectives to particle sets. MPM (Golling et al., 2024) masks constituents and recovers VQ-VAE tokens, yielding a backbone that fine-tunes well with limited labels and transfers across classes and datasets. A subsequent study drops tokenization (Leigh et al., 2025), and HEP-JEPA applies a Joint Embedding Predictive Architecture to grouped particle patches (Bardhan et al., 2025). Independently, Kishimoto et al. (Kishimoto et al., 2023) pre-train on unlabeled CMS collision data by masking reconstructed object types for event classification. Masked pre-training can therefore transfer with limited labels on jets or inclusive events, but these backbones remain shared and are not specialized for exclusive, long-tailed decay identification.

Generative pre-training. Generative pre-training is represented by OmniJet-α (Birk et al., 2024), which tokenizes jets with a VQ-VAE, pre-trains an autoregressive transformer for jet generation, and transfers the backbone to jet tagging. It is an early cross-task foundation model for jet physics that unifies generative pre-training with supervised tagging, but its tokens encode only kinematic information and the transferred representation is not adapted to frequent versus rare exclusive decay modes.

Large-scale supervised training. Large-scale supervised training remains the strongest jet baseline. Particle Transformer (Qu et al., 2022) introduces the 100M-jet JetClass corpus and pairwise interaction attention, setting the standard for jet tagging. Subsequent work scales event classification (Ho et al., 2024), scales jet-origin identification with an LLM-style model (Wu et al., 2024), or reuses one supervised backbone across many jet tasks (Mikuni & Nachman, 2025). These models set strong baselines under comparatively balanced category structures and a shared representation, whereas J/ψ decay identification is long-tailed and depends on heterogeneous detector information. This gap motivates a physics-structured foundation model for exclusive decay modes.

3 PRELIMINARIES

We formulate the particle decay identification problem as follows. Let $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ denote a particle decay event, where each reconstructed particle is represented by an M-dimensional feature vector $\mathbf{p}_i \in \mathbb{R}^M$. The decay identification task is formulated as learning a permutation-invariant

mapping:

$$f_\theta : \mathcal{P} \to \mathcal{L}, \tag{1}$$

where $\mathcal{L} = \{1, \ldots, K\}$ denotes the set of decay-mode labels and $K = |\mathcal{L}|$ is the number of target decay modes. The parameter θ specifies the learnable model parameters.

For foundation-model adaptation, we further decompose the model parameters as $\theta = \theta_0 + \Delta\theta$, where θ_0 represents the transferable knowledge acquired during pre-training and $\Delta\theta$ captures task-specific adaptation. Parameter-efficient fine-tuning keeps θ_0 fixed while updating only $\Delta\theta$, thereby preserving the general representations learned across diverse decay modes while adapting the model to the specific characteristics of a downstream decay-identification task.

4 METHODOLOGY

4.1 PHYSICAL MOTIVATION AND STRUCTURING ASSUMPTION

The physical information available for particle decay identification is inherently heterogeneous. The BESIII detector combines four complementary subsystems, namely the drift chamber (DC), time-of-flight system (TOF), electromagnetic calorimeter (EMC), and muon counter (MUC), which provide different measurements of the particles produced in a collision. Their responses are not uniform across decay modes: a given decay mode may be characterized more strongly by a particular detector response, while the same detector subsystem can exhibit different response patterns for different decay modes. Consequently, the reconstructed final-state particle features contain multiple physical factors whose relevance to decay identification varies across modes.

This observation suggests that the heterogeneity in detector information is structured rather than arbitrary. In particular, different aspects of the reconstructed particles, such as kinematics, energy deposition, and particle identity, are supported by distinct detector measurements or their combinations. Encoding all such information into a single undifferentiated representation therefore forces heterogeneous physical factors to compete within the same feature space, potentially making the learned representation overly dependent on dominant patterns in the training distribution. This issue is especially relevant in the long-tailed setting, where frequent decay modes provide substantially more supervision than rare modes.

Motivated by this structure, we make the following structuring assumption: the physical information relevant to decay identification can be organized into complementary subspaces, each associated with a distinct semantic aspect of the reconstructed event, while their interactions remain necessary for determining the final decay mode. Under this assumption, kinematic, energy-related, and particle-identity information should first be encoded separately to preserve their individual physical structures, and only then integrated to model their cross-factor interactions. This provides a principled basis for learning a shared representation without requiring different physical factors to compete for a single latent space, while retaining their complementarity for downstream decay identification.

4.2 DECAYFM ARCHITECTURE

The overall architecture of DecayFM is illustrated in Figure 2. Given a reconstructed decay event $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^N$ defined in Section 3, DecayFM first decomposes each particle feature into three physically distinct subspaces, encodes them independently, and subsequently models their interactions before performing long-tail-aware adaptation. This design follows the physical structuring assumption and separates transferable representation learning from decay-mode-specific adaptation. Accordingly, DecayFM consists of two complementary components: a physics-structured representation module and a long-tail-aware adaptation module.

We first decompose each particle feature into three physically distinct subspaces:

$$\mathbf{p}_i = \left[\mathbf{x}_i^{\mathrm{kin}}; \mathbf{x}_i^{\mathrm{edep}}; \mathbf{x}_i^{\mathrm{pid}}\right], \qquad \mathbf{x}_i^r \in \mathbb{R}^{M_r}, \quad r \in \{\mathrm{kin}, \mathrm{edep}, \mathrm{pid}\}, \tag{2}$$

where

$$M = M_{\mathrm{kin}} + M_{\mathrm{edep}} + M_{\mathrm{pid}}. \tag{3}$$

Accordingly, the event-level inputs are decomposed as:

$$\mathcal{P}^r = \{\mathbf{x}_i^r\}_{i=1}^N, \qquad r \in \{\mathrm{kin}, \mathrm{edep}, \mathrm{pid}\}. \tag{4}$$

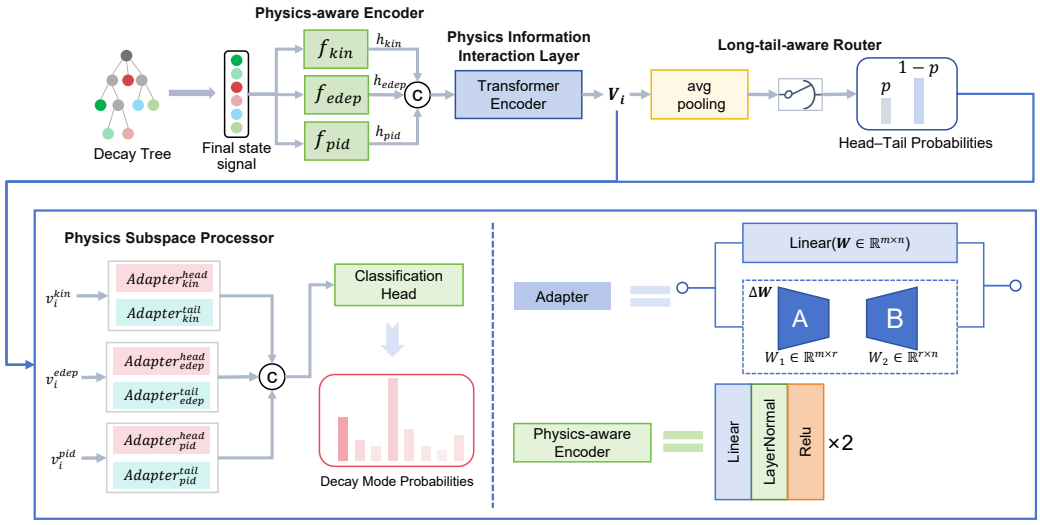


Figure 2: Overview of the DecayFM framework. DecayFM consists of four components: (1) Physics-aware Encoder, independently encodes kinematic, energy-deposition, and particle-identification information; (2) Physics Information Interaction Layer, models interactions among particles and physical subspaces; (3) Long-tail-aware Router, estimates the head-tail regime of the input decay mode; and (4) Subspace Processor, performs regime-specific adaptation with LoRA-based adapters.

Here, the three subspaces represent complementary physical information associated with particle kinematics, detector energy deposition, and particle identification, respectively. This decomposition does not assume independence among the three factors. It provides an explicit structure for preserving their distinct physical semantics before learning their interactions. This is important because different decay modes may exhibit different sensitivities to these physical factors, and forcing heterogeneous information into a single shared projection may obscure such decay-specific responses.

Each physical subspace is processed by an independent Physics-aware Encoder:

$$\mathbf{h}_i^r = f_r(\mathbf{x}_i^r), \qquad f_r : \mathbb{R}^{M_r} \to \mathbb{R}^d, \qquad r \in \{\text{kin}, \text{edep}, \text{pid}\} \tag{5}$$

where d denotes the shared embedding dimension. The independent encoders allow each representation space to learn transformations specialized to its corresponding physical semantics, reducing the need to accommodate heterogeneous information within a single shared projection. We therefore obtain three event-level representations and then jointly contextualized by the Physics Information Interaction Layer (PIIL):

$$\mathbf{H}^r = [\mathbf{h}_1^r, \ldots, \mathbf{h}_N^r]^\top \in \mathbb{R}^{N \times d}, \qquad \mathbf{V} = \text{PIIL}\left([\mathbf{H}^{\text{kin}}; \mathbf{H}^{\text{edep}}; \mathbf{H}^{\text{pid}}]\right). \tag{6}$$

We implement PIIL with a transformer-based interaction module to capture dependencies among particles and across physical information sources. Unlike the independent encoding stage, which preserves the specificity of individual physical factors, PIIL integrates these complementary representations to model their joint configurations. Such cross-factor interaction is essential for decay identification, where the decay mode is often determined not by a single physical observable but by the collective pattern of final-state particles and their heterogeneous detector responses.

The resulting representation $\mathbf{V}$ is retained in the corresponding physical subspaces,

$$\mathbf{V} = \left[\mathbf{V}^{\text{kin}}; \mathbf{V}^{\text{edep}}; \mathbf{V}^{\text{pid}}\right], \tag{7}$$

and is subsequently used for long-tail-aware adaptation. A Long-tail-aware Router first estimates the probability p that the input belongs to the head regime:

$$p = \sigma\left(g_\phi(\text{MeanPool}(\mathbf{V}))\right). \tag{8}$$

Rather than imposing a hard head-tail assignment, the continuous routing probability provides a soft estimate of the data regime, allowing the subsequent representation adaptation to vary smoothly between frequent and rare decay modes.

Each physical subspace is then processed by an independent Physics Subspace Processor. Each processor contains head- and tail-specialized Low-Rank Adaptation (LoRA) (Hu et al., 2021), whose outputs are dynamically combined according to the routing probability:

$$\mathbf{S}^r = pA_r^{\text{head}}(\mathbf{V}^r) + (1-p)A_r^{\text{tail}}(\mathbf{V}^r), \qquad r \in \{\text{kin}, \text{edep}, \text{pid}\}. \tag{9}$$

This design allows the model to specialize each physical representation according to the inferred data regime while preserving the structured organization of the physical information. The use of LoRA further enables this specialization with a small number of trainable parameters, avoiding modification of the shared representation learned during pre-training.

The adapted subspaces are subsequently aggregated for decay-mode classification. Overall, DecayFM combines physics-structured representation learning with long-tail-aware parameter-efficient adaptation, enabling the model to capture decay-specific responses to heterogeneous physical factors while adapting to the distinct data regimes of frequent and rare decay modes. This structured design provides transferable representations across diverse decay modes and improves the characterization of rare modes with limited supervision.

4.3 Training and Adaptation

To decouple representation learning from long-tail adaptation, DecayFM adopts a two-stage optimization strategy. In the first stage, we fix the routing probability to $p = 0.5$ and optimize the Physics-aware Encoders, Physics Information Interaction Layer, and shared parameters of the Physics Subspace Processors using the decay classification objective. This stage learns a transferable representation without long-tail-dependent routing.

In the second stage, we freeze these shared parameters and optimize only the Long-tail-aware Router and LoRA adapters. Let $y_k^{\text{lt}} \in \{0, 1\}$ indicate whether decay mode k belongs to the head or tail regime. For an event with decay label y, the router is supervised by

$$\mathcal{L}_{\text{route}} = -\left[y_y^{\text{lt}} \log p + \left(1 - y_y^{\text{lt}}\right) \log(1-p)\right], \tag{10}$$

where p is the probability of selecting the head branch. The router is optimized jointly with the decay classification loss to select the appropriate adaptation according to the data regime, while the shared representation remains fixed.

5 Dataset

We construct the ParticleDecayMode (PDM) dataset series in the BESIII environment to support large-scale pre-training and systematic evaluation of particle decay identification. The dataset is designed to separate three aspects of the problem: learning transferable representations from diverse decay modes, adapting the pre-trained model to downstream identification, and evaluating recognition under extreme rarity. It consists of three complementary subsets.

PDM-150-Pretrain. PDM-150-Pretrain contains 150 decay modes following a long-tailed distribution, with 5.68 million events in total. We use this subset for foundation-model pre-training, where the large-scale and imbalanced samples expose the model to diverse decay patterns and their naturally varying frequencies. It contains 5.11 million training events and 0.57 million validation events.

PDM-150-Downstream. PDM-150-Downstream contains the same 150 decay modes as PDM-150-Pretrain, with a total of 2.42 million events following the same long-tailed distribution. This subset is used to evaluate downstream adaptation from the pre-trained model under a consistent decay-mode space. It contains 1.94 million training events, 0.24 million validation events, and 0.24 million test events. Keeping the decay-mode space consistent with pre-training allows us to isolate the effectiveness of representation transfer and parameter-efficient adaptation.

PDM-rare-200. PDM-rare-200 is designed to evaluate recognition of rare decay modes under a controlled frequency setting. It contains 200 rare decay modes with approximately uniform numbers

of events per mode, ranging from 1,080 to 1,366, for a total of 245,002 events. The dataset is split into 195,919 training events, 24,486 validation events, and 24,597 test events. By removing the dominant frequency differences between classes, this subset provides a complementary evaluation of the model's ability to recognize rare decay modes rather than relying on their occurrence frequency.

Together, the three subsets provide a unified evaluation protocol for particle decay foundation models: PDM-150-Pretrain supports large-scale representation learning, PDM-150-Downstream evaluates transfer and adaptation, and PDM-rare-200 probes recognition of rare decay modes under a controlled data regime.

6 EXPERIMENTS

We first perform large-scale pre-training on PDM-150-Pretrain to obtain the foundation model. All subsequent experiments involving pre-trained models use this model for initialization and adapt it to downstream tasks with different parameter-update strategies.

To evaluate DecayFM across diverse decay identification tasks, we consider four settings: *in-domain evaluation*, *cross-domain evaluation*, *comparison with other methods*, and *few-shot fine-tuning on rare decay modes*. The first two assess transferability, while the latter two evaluate overall performance and adaptation efficiency on rare modes.

6.1 IN-DOMAIN EVALUATION

We first evaluate the models on PDM-150-Downstream, which contains the same set of decay modes as PDM-150-Pretrain and follows the same long-tailed distribution. We consider classification over the top 10, 20, 50, 100, and 150 decay modes to examine how the benefit of pre-training varies with the number of target classes.

Table 1 shows that pre-training consistently improves in-domain classification, but the gain decreases from 0.0167 for 10 classes to 0.0053 for 150 classes. This limited improvement is expected because DecayFM has already learned representations of these decay modes during pre-training.

The optimal adaptation strategy also varies with the number of target classes. *Decay head+LoRA* achieves the highest accuracy for 10 and 50 classes, while *Only decay head* performs best for 20 classes. For 100 and 150 classes, *Full* fine-tuning achieves the highest accuracy. In contrast, *Decay head+LoRA* shows a noticeable performance drop for 100 and 150 classes, indicating that parameter-efficient adaptation does not provide a consistent advantage for in-domain tasks.

Table 1: Classification accuracy of different downstream adaptation strategies on the in-domain PDM-150-Downstream benchmark. *N-Class* denotes the number of decay modes being classified. *From scratch* denotes training a dedicated model from scratch without pre-training. *Only decay head* freezes the pre-trained backbone and updates only the classification head. *Decay head+LoRA* performs parameter-efficient fine-tuning by updating the classification head and learning weight updates through LoRA. *Full* denotes full fine-tuning of the pre-trained model. Bold indicates the best accuracy in each row, and Δ denotes the accuracy difference between the best-performing strategy and *From scratch*.

N-Class	From scratch	Only decay head	Decay head+LoRA	Full	Δ
10	0.8734	0.8897	**0.8901**	0.8857	0.0167
20	0.7677	**0.7843**	0.7841	0.7798	0.0166
50	0.6714	0.6771	**0.6780**	0.6771	0.0066
100	0.5971	0.6030	0.5669	**0.6046**	0.0075
150	0.5597	0.5640	0.5244	**0.5650**	0.0053

Overall, these results suggest that the in-domain task is already substantially covered by the representations acquired during large-scale pre-training. As the number of target decay modes increases, downstream adaptation yields progressively smaller additional gains. This diminishing improvement is consistent with a saturation trend in the present in-domain setting, suggesting that further

training on the same distribution may provide limited benefits once the foundation model has already acquired substantial task-relevant representations.

6.2 CROSS-DOMAIN EVALUATION

We evaluate cross-domain generalization on PDM-rare-200, which contains 200 rare decay modes with a uniform class distribution and differs substantially from PDM-150-Pretrain in both decay modes and class distribution. As shown in Table 2, all models initialized from the foundation model outperform the dedicated models trained from scratch. The gains are 0.0720, 0.0708, 0.1161, 0.0634, and 0.0407 for 10, 20, 50, 100, and 200 classes, respectively, with the largest improvement reaching 0.1161 for 50-class identification. These results demonstrate the strong cross-domain generalization of DecayFM for rare decay identification.

Table 2: Classification accuracy of different downstream adaptation strategies on the cross-domain PDM-rare-200 benchmark.

N-Class	From scratch	Only decay head	Decay head+LoRA	Full	Δ
10	0.7922	0.8605	**0.8642**	0.8554	0.0720
20	0.6186	0.6721	**0.6894**	0.6865	0.0708
50	0.4325	0.5338	0.5448	**0.5486**	0.1161
100	0.3586	0.4212	**0.4252**	0.4213	0.0666
200	0.3152	0.3443	0.3502	**0.3559**	0.0407

Compared with the in-domain results, where the corresponding gains range from 0.0053 to 0.0167, cross-domain adaptation yields substantially larger improvements. This indicates that the benefit of pre-training becomes more pronounced when the target decay modes differ from those seen during pre-training. The gain is consistent with our physics-structured representation design, which separates heterogeneous physical information into distinct subspaces while jointly modeling their complementary dependencies, enabling the pre-trained representation to retain mode-specific physical information and adapt to unseen decay modes.

Notably, LoRA-based adaptation already achieves competitive performance without updating the pre-trained backbone, reaching 0.8642, 0.6894, 0.5448, 0.4252, and 0.3502 for 10, 20, 50, 100, and 200 classes, respectively. This shows that efficient parameter updates are sufficient to transfer the pre-trained representation to rare decay modes, indicating that our foundation model provides a robust representation space for cross-domain decay identification.

6.3 COMPARISON WITH OTHER METHODS

We compare DecayFM with representative Transformer- and GNN-based methods on both PDM-150-Downstream and PDM-rare-200 tasks. As shown in Table 3, DecayFM consistently achieves superior performance across diverse decay identification settings. These results demonstrate that explicitly structuring heterogeneous physical information into dedicated subspaces and applying specialized processors enables DecayFM to learn more informative representations for different decay modes, particularly under long-tailed and rare-mode scenarios.

Table 3: Comparison with other methods on the PDM-150-Downstream and PDM-rare-200 tasks.

Method	PDM-150-Downstream	PDM-rare-200
MPM (Golling et al., 2024)	0.5411	0.2532
OmniJet-α (Birk et al., 2024)	0.4963	0.2569
JetCLR (Dillon et al., 2022a)	0.4894	0.2653
Vanilla Transformer (Finke et al., 2023)	0.5124	0.2789
HEP-JEPA (Bardhan et al., 2025)	0.5528	0.2971
ParT (Qu et al., 2022)	0.5525	0.3277
DGCNN (Giagu et al., 2022)	0.4712	0.3304
DecayFM (ours)	**0.5650**	**0.3559**

6.4 FEW-SHOT FINE- TUNING ON RARE DECAY MODES

Rare decay modes have limited event yields and labeled data, making dedicated models difficult to train effectively. We therefore evaluate whether DecayFM can transfer its learned representations to rare decay modes under few-shot settings. We consider 10, 20, 50, 100, and 200 rare decay modes, with 1 to 1,000 training samples per class, and compare parameter-efficient fine-tuning (PEFT) of DecayFM with training dedicated models from scratch.

Table 4: Few-shot classification accuracy on rare decay modes with different numbers of training samples per class. *DecayFM-PEFT* denotes parameter-efficient fine-tuning of our foundation model by updating the classification head and LoRA parameters. Bold indicates the best accuracy achieved by the dedicated model and the first *DecayFM-PEFT* result that exceeds this accuracy.

N-Class	Method	Training samples per class							
		1	5	10	50	100	200	500	1000
10	From scratch	0.2368	0.4688	0.5107	0.6554	0.6900	0.7393	0.7541	**0.7722**
	DecayFM-PEFT	0.5378	0.7327	**0.7911**	0.8232	0.8388	0.8446	0.8594	0.8594
20	From scratch	0.1208	0.2844	0.3322	0.5326	0.5635	0.5998	0.6344	**0.6558**
	DecayFM-PEFT	0.3730	0.5771	0.6352	**0.7246**	0.7292	0.7415	0.7448	0.7576
50	From scratch	0.0611	0.1773	0.1946	0.3312	0.3555	0.3836	0.4320	**0.4559**
	DecayFM-PEFT	0.2090	0.3649	0.4301	**0.5222**	0.5396	0.5535	0.5793	0.5824
100	From scratch	0.0450	0.1170	0.1346	0.2292	0.2469	0.2805	0.3213	**0.3614**
	DecayFM-PEFT	0.1336	0.2524	0.3112	**0.3871**	0.4082	0.4184	0.4393	0.4425
200	From scratch	0.0178	0.0677	0.1056	0.1632	0.1887	0.2090	0.2603	**0.2960**
	DecayFM-PEFT	0.0363	0.1732	0.2197	0.2928	**0.3130**	0.3261	0.3420	0.3489

Table 4 shows that *DecayFM-PEFT* consistently outperforms *From scratch* across all evaluated settings, with the largest gains occurring in the low-data regime. More importantly, the *DecayFM-PEFT* reaches the best performance of the dedicated model using only 10, 50, 50, 50, and 100 samples per class for 10, 20, 50, 100, and 200 rare modes, respectively. Compared with the 1,000 samples per class used by the dedicated models to reach their best performance, this reduces the required training data by approximately one to two orders of magnitude.

These results demonstrate that our physics-structured foundation model learns transferable representations across decay modes, enabling efficient adaptation to rare decay modes with limited supervision and making more effective use of rare events. By leveraging representations learned during large-scale pre-training, our model can capture discriminative features from only a small number of rare-decay samples, substantially improving the identification of rare decay modes in high-energy physics experiments.

7 CONCLUSIONS

We introduced DecayFM, a physics-structured foundation model for particle decay identification in high-energy physics. By structuring heterogeneous physical information into semantic-specific representation subspaces and incorporating long-tailed adaptation, DecayFM learns transferable representations from large-scale pre-training and efficiently adapts to diverse decay modes. Experiments across in-domain, cross-domain, and few-shot settings demonstrate improved decay identification and substantially higher sample efficiency for rare decay modes, enabling effective identification with far fewer labeled events than dedicated models trained from scratch. These results provide a foundation-model paradigm for more efficient utilization of rare events in high-energy physics.

Rare decay modes are often constrained by their limited event yields. Building on our foundation model for rare decay identification, future work will explore generative foundation models for particle decays to synthesize rare and previously unseen decay modes, providing new tools for exploring rare physical processes and new physics.

REFERENCES

M. Ablikim, M. N. Achasov, P. Adlarson, O. Afedulidis, X. C. Ai, R. Aliberti, A. Amoroso, Q. An, Y. Bai, O. Bakina, et al. Measurement of the decay $\Xi^0 \to \Lambda\gamma$ with entangled $\Xi^0\bar{\Xi}^0$ pairs. *arXiv preprint arXiv:2408.16654*, 2024.

Jai Bardhan, Radhikesh Agrawal, Abhiram Tilak, Cyrin Neeraj, and Subhadip Mitra. HEP-JEPA: A foundation model for collider physics. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025. Also presented at the ICLR 2025 Workshop on Agentic AI for Scientific Discovery.

Michael Benedikt, Helmut Burkhardt, and M Zykova. Future circular collider feasibility study report. 2025.

Pushpalatha C Bhat. Multivariate analysis methods in particle physics. *Annual Review of Nuclear and Particle Science*, 61(1):281–309, 2011.

Joschka Birk, Anna Hallin, and Gregor Kasieczka. OmniJet-α: the first cross-task foundation model for particle physics. *Machine Learning: Science and Technology*, 5(3):035031, 2024.

Shanzhen Chen, Yiming Li, Wenbin Qian, Zhihong Shen, Yuehong Xie, Zhenwei Yang, Liming Zhang, and Yanxi Zhang. Heavy flavour physics and cp violation at lhcb: a ten-year review. *Frontiers of Physics*, 18(4):44601, 2023.

Marco Destefanis. Hadron spectroscopy at besiii. In *EPJ Web of Conferences*, volume 129, pp. 00017. EDP Sciences, 2016.

Barry M. Dillon, Gregor Kasieczka, Hans Olischläger, Tilman Plehn, Peter Sorrenson, and Lorenz Vogel. Symmetries, safety, and self-supervision. *SciPost Physics*, 12(6):188, 2022a.

Barry M. Dillon, Radha Mastandrea, and Benjamin Nachman. Self-supervised anomaly detection for new physics. *Physical Review D*, 106(5):056005, 2022b.

Barry M. Dillon, Luigi Favaro, Friedrich Feiden, Tanmoy Modak, and Tilman Plehn. Anomalies, representations, and self-supervision. *SciPost Physics Core*, 7(3):056, 2024.

Thorben Finke, Michael Krämer, Alexander Mück, and Jan Tönshoff. Learning the language of qcd jets with transformers. *Journal of High Energy Physics*, 2023(6):184, 2023.

Stefano Giagu, Luca Torresi, and Matteo Di Filippo. Tau lepton identification with graph neural networks at future electron–positron colliders. *Frontiers in Physics*, 10:909205, 2022.

Tobias Golling, Lukas Heinrich, Michael Kagan, Samuel Klein, Matthew Leigh, Margarita Osadchy, and John Andrew Raine. Masked particle modeling on sets: towards self-supervised high energy physics foundation models. *Machine Learning: Science and Technology*, 5(3):035074, 2024.

Philip Harris, Jeffrey Krupa, Michael Kagan, Benedikt Maier, and Nathaniel Woodward. Resimulation-based self-supervised learning for pretraining physics foundation models. *Physical Review D*, 111(3):032010, 2025.

Joshua Ho, Ryan Roberts, Shuo Han, and Haichen Wang. Pretrained event classification model for high energy physics analysis. arXiv:2412.10665, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tomoe Kishimoto, Masahiro Morinaga, Masahiko Saito, and Junichi Tanaka. Pre-training strategy using real particle collision data for event classification in collider physics. In *NeurIPS 2023 Workshop on Machine Learning and the Physical Sciences*, 2023.

Matthew Leigh, Samuel Klein, François Charton, Tobias Golling, Lukas Heinrich, Michael Kagan, Inês Ochoa, and Margarita Osadchy. Is tokenization needed for masked particle modeling? *Machine Learning: Science and Technology*, 6(2):025075, 2025.

Vinicius Mikuni and Benjamin Nachman. OmniLearn: A method to simultaneously facilitate all jet physics tasks. *Physical Review D*, 111(5):054015, 2025.

Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18281–18292. PMLR, 2022.

Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Federico Ambrogi, Thomas Bergauer, Marko Dragicevic, Janos Erö, A Escalante Del Valle, Martin Flechl, Rudolf Fruehwirth, et al. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques. *Journal of Instrumentation*, 15(06):P06005–P06005, 2020.

Yi-Fang Wang. The construction of the besiii experiment. *International Journal of Modern Physics A*, 21(27):5371–5380, 2006.

Hengkui Wu, Panpan Chi, Yongfeng Zhu, Liujiang Liu, Shuyang Hu, Yuexin Wang, Chen Zhou, Qihao Wang, Yingsi Xin, Bruce Liu, Hengyuan Wu, Dahao Liang, Xinglong Jia, and Manqi Ruan. Scaling particle collision data analysis. arXiv:2412.00129, 2024.

Chang-Zheng Yuan. Charmonium and charmoniumlike states at the besiii experiment. *National Science Review*, 8(11):nwab182, 2021.

Chang-Zheng Yuan and Stephen Lars Olsen. The besiii physics programme. *Nature Reviews Physics*, 1(8):480–494, 2019.