

UNIVERSAL MUSIC REPRESENTATIONS? EVALUATING FOUNDATION MODELS ON WORLD MUSIC CORPORA

Charilaos Papaioannou^{1,2,3}

Emmanouil Benetos²

Alexandros Potamianos^{1,3}

¹ School of ECE, National Technical University of Athens, Greece

² Centre for Digital Music, Queen Mary University of London, UK

³ Archimedes, Athena Research Center, Greece

cpapaioan@mail.ntua.gr

ABSTRACT

Foundation models have revolutionized music information retrieval, but questions remain about their ability to generalize across diverse musical traditions. This paper presents a comprehensive evaluation of five state-of-the-art audio foundation models across six musical corpora spanning Western popular, Greek, Turkish, and Indian classical traditions. We employ three complementary methodologies to investigate these models' cross-cultural capabilities: probing to assess inherent representations, targeted supervised fine-tuning of 1-2 layers, and multi-label few-shot learning for low-resource scenarios. Our analysis shows varying cross-cultural generalization, with larger models typically outperforming on non-Western music, though results decline for culturally distant traditions. Notably, our approaches achieve state-of-the-art performance on five out of six evaluated datasets, demonstrating the effectiveness of foundation models for world music understanding. We also find that our targeted fine-tuning approach does not consistently outperform probing across all settings, suggesting foundation models already encode substantial musical knowledge. Our evaluation framework and benchmarking results contribute to understanding how far current models are from achieving universal music representations while establishing metrics for future progress.

1. INTRODUCTION

The notion of music as a “universal language” remains contested among scholars [1, 2]. While some musical elements transcend cultural boundaries, traditions have evolved with distinct characteristics and semantic content [3,4]. This tension between universality and cultural specificity presents a complex challenge that modern artificial intelligence approaches offer a novel lens to investigate.

Foundation models have emerged as a transformative paradigm across artificial intelligence (AI) domains [5],

including music and audio [6–8]. In music information retrieval (MIR), these multipurpose models perform diverse tasks from beat tracking to automatic tagging [9, 10]. Though implicitly claiming a form of universality, they largely neglect cultural dimensions while training predominantly on Western-centric data [10]. This raises a critical question: to what extent do foundation models actually provide universal music representations that generalize across diverse musical traditions?

In this work, we evaluate five state-of-the-art audio models across six corpora spanning Western popular, Greek, Turkish, and Indian classical traditions, to quantitatively assess their cross-cultural capabilities and contribute to discussions about the universality of musical representations. We focus on automatic music tagging as our evaluation task and employ three complementary methodologies: (i) probing, which uses the models as frozen feature extractors with a trainable classifier, (ii) targeted supervised fine-tuning to assess adaptation potential, and (iii) multi-label few-shot learning to evaluate performance in low-resource scenarios common with world music collections.

Our evaluation reveals both promising cross-cultural transfer capabilities as well as remaining gaps in universal music understanding, due to the decrease in performance for culturally distant domains and especially in low-resource scenarios. The contributions of this work can be summarized as follows:

- This is the first comprehensive evaluation, to the best of our knowledge, of foundation models across culturally diverse music corpora.
- We propose a methodological evaluation framework that integrates few-shot learning with traditional approaches, enabling systematic assessment of model representations under different training setups.
- State-of-the-art results have been achieved by our approaches in five out of six datasets.
- We have optimized multi-label few-shot learning, significantly reducing inference time and making it practical for large numbers of classes.
- Our code is being made available¹ for reproducibility and to promote research on world music.



© C. Papaioannou, E. Benetos, and A. Potamianos. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** C. Papaioannou, E. Benetos, and A. Potamianos, “Universal Music Representations? Evaluating Foundation Models on World Music Corpora”, in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.*, Daejeon, South Korea, 2025.

¹ <https://github.com/pxaris/FM-music-tagging>

2. RELATED WORK

Foundation models. Foundation models for music have emerged by leveraging large-scale self-supervised or contrastive learning on extensive audio datasets, enabling them to capture rich musical features applicable across diverse tasks. Representative works include JukeMIR [11], which explored representations from the Jukebox generative model [12], MULE [13], a self-supervised model pre-trained on MusicNet dataset, and Music2Vec [14], which utilized masked prediction strategies with student-teacher approaches. Subsequent advancements like MusicFM [9] have scaled up both model size and training data, demonstrating effectiveness across multiple benchmark tasks.

The landscape of current foundation models encompasses several architectural approaches: masked acoustic modeling, MERT [6], contrastive audio-text learning such as LAION-CLAP [7], and unified audio understanding with models like Qwen-Audio [8]. Despite their impressive performance on standard benchmarks, their cross-cultural generalization capabilities remain largely unexplored, particularly regarding their effectiveness across diverse musical traditions beyond Western contexts.

Automatic world music tagging. Automatic music tagging - predicting metadata such as genre, mood, and instrumentation from audio signals - is typically referred to as music auto-tagging [15–18] and constitutes a multi-label classification problem. Architectures addressing this task have evolved from convolutional models like VGGish [19] and Musicnn [20] to transformer-based approaches like AST [21] and more recent foundation models [9].

Research on world music computational analysis has grown in recent years [22], with studies focused on specific traditions including Turkish makam recognition [23, 24], Indian classical music classification [25], and analysis of Iranian and Korean traditional music [26, 27]. While a recent study applied auto-tagging across diverse musical datasets [28], this is the first time to the best of our knowledge where a comprehensive evaluation of foundation models on world music corpora is being conducted.

To address the challenges of imbalanced tags and limited data inherent in world music research, we employ Label-Combination Prototypical Networks (LC-Protonets) [29] for few-shot learning. This approach extends Prototypical Networks [30] by creating prototypes for each label combination, rather than generating one prototype per label. While established benchmarks for evaluating representations on downstream tasks typically employ probing and fine-tuning methodologies [31–33], our work incorporates few-shot learning as a complementary evaluation approach, assessing foundation models’ capabilities in low-resource scenarios.

3. METHODOLOGICAL FRAMEWORK

Our methodological framework systematically evaluates whether foundation models can effectively represent musical characteristics across diverse cultural traditions. As shown in Figure 1, we employ three complementary

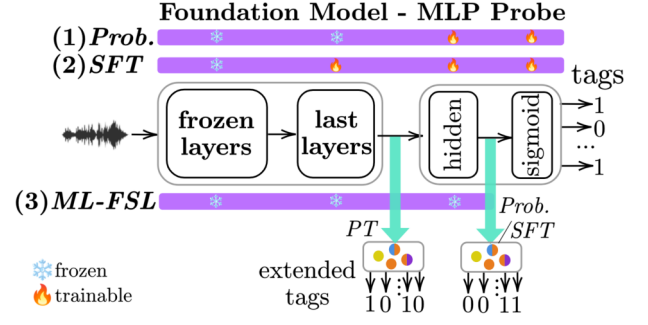


Figure 1. Architectural overview of our evaluation framework showcasing three methodologies: (1) Probing (*Prob.*), (2) Supervised Fine-Tuning (*SFT*), and (3) Multi-Label Few-Shot Learning (*ML-FSL*). The diagram indicates feature extraction points used by *ML-FSL* from either Pre-Trained (*PT*), trained *Prob.* or *SFT* models.

methodologies: probing (*Prob.*), supervised fine-tuning (*SFT*), and multi-label few-shot learning (*ML-FSL*). Probing trains only an MLP classifier on frozen model representations, while *SFT* makes the model’s last layers trainable alongside the MLP. *ML-FSL* extracts representations from three contexts, i.e., pretrained model (*PT*), trained probing model (*Prob.*) and fine-tuned model (*SFT*) to evaluate performance on extended tag sets under data scarcity conditions.

3.1 Models

For our evaluation, we selected five state-of-the-art audio models spanning different architectures, pre-training approaches, and parameter scales:

MERT. We evaluate two variants of MERT [6]: MERT-95M² and MERT-330M³ with 95M and 330M parameters respectively. These transformer-based models employ masked acoustic modeling, using an acoustic and a musical teacher, during pre-training. MERT-95M consists of 12 layers, while MERT-330M has 24 layers.

LAION-CLAP. We include two variants: CLAP-Music⁴ (CLAP-M), trained exclusively on music data, and CLAP-Music&Speech⁵ (CLAP-M&S), which incorporates additional speech data [7]. Both utilize HTS-AT [34] for audio encoding, a transformer-based model with 4 groups of swin-transformer blocks [35], with 68M audio-specific parameters within a larger 194M parameter model.

Qwen2-Audio. The largest model in our evaluation framework, Qwen2-Audio⁶ [36], contains 637M audio-specific parameters within an 8.4B parameter architecture and features 32 transformer layers [37] in its audio tower.

VGGish. As a baseline comparison, we include VGGish [17, 38], a 3.6M parameter end-to-end model trained via supervised learning on mel-spectrograms to predict tags. For VGGish, we report results from the literature for the

² <https://huggingface.co/m-a-p/MERT-v1-95M>

³ <https://huggingface.co/m-a-p/MERT-v1-330M>

⁴ https://huggingface.co/laion/larger_clap_music

⁵ https://huggingface.co/laion/larger_clap_music_and_speech

⁶ <https://huggingface.co/Qwen/Qwen2-Audio-7B>

same experimental setup used in our work [28, 29] rather than running new experiments.

3.2 Datasets

Our evaluation spans diverse traditions from six music datasets. For Western music, we utilize MagnaTagATune [39] (25,863 clips) and FMA-medium [40] (25,000 tracks). For world music traditions, we incorporate the Lyra dataset [41] with 1,570 recordings of Greek folk music, and three collections from the CompMusic project [42]: the Turkish-makam corpus [43, 44] (5,297 recordings) as well as Hindustani [45] (1,204 recordings) and Carnatic [45] (2,612 recordings) of Indian classical music.

Following [28], we set maximum audio durations to achieve similar sizes between datasets and prepare their metadata for the auto-tagging task. For Probing and Supervised Fine-Tuning, we use the standard tag sets, i.e., 50 tags for MagnaTagATune, 30 for Lyra and Turkish-makam, and 20 for the rest of the datasets. Our ML-FSL experiments use extended tag sets that include previously unseen classes, summing up to: 80 tags for MagnaTagATune, 60 for Lyra and Turkish-makam, 40 for FMA-medium and Carnatic, and 35 for Hindustani, consistent with [29].

3.3 Evaluation methodologies

Probing. Our first methodology (*Prob.*) evaluates how well foundation models inherently represent musical characteristics across cultures. We employ probing, where the model remains frozen while only training a classifier on top of the extracted representations. Specifically, we implement a shallow Multi-layer Perceptron (MLP) with a single hidden layer of 512 units followed by a sigmoid classification layer, optimized with binary cross-entropy loss.

Supervised Fine-Tuning. To evaluate adaptation potential, we implement targeted supervised fine-tuning (*SFT*) by unfreezing a subset of model parameters. For MERT-95M, we unfreeze the last two transformer layers, while for MERT-330M only the last layer. For both CLAP models, we unfreeze the last group of swin-transformer blocks of the audio encoder along with the normalization and two projection layers. In Qwen2-Audio, we fine-tune the last layer of the audio tower along with the normalization layer before multi-modal projection. These choices were constrained by RAM limitations affecting both trainable parameters and hyperparameter tuning. We use the same trainable MLP Probe architecture as in the Probing experiments, initializing it with the weights learned during that phase. This weight initialization strategy helps maintain previously learned knowledge while adapting to new domains, mitigating potential catastrophic forgetting issues [46]. We also employ learning rate warmup and cosine scheduling to ensure stable adaptation [47].

Multi-Label Few-Shot Learning. Our third methodology (*ML-FSL*) evaluates performance in low-resource scenarios by employing an optimized version of LC-Protonets [29] that is detailed in subsection 3.4. We extract representations from three different contexts: directly from the pre-trained model (*PT*), from the hidden layer of the trained

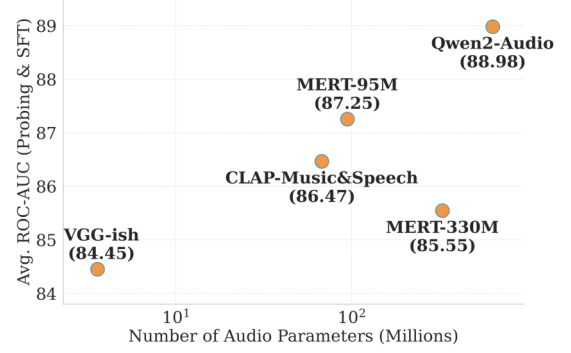


Figure 2. Relationship between model size and performance, averaged over Probing and Supervised Fine-Tuning (SFT) tasks. The x-axis represents the number of audio-specific parameters on a logarithmic scale, while the y-axis reports the mean ROC-AUC (%) across all datasets.

MLP Probe (*Prob.*), and from the fine-tuned model (*SFT*). Notably, this methodology involves no additional training during few-shot evaluation; the model acts as a frozen feature extractor that maps both the few examples and the unknown items to an embedding space where classification occurs utilizing the LC-Protonets approach.

3.4 Multi-label few-shot learning optimization

While the LC-Protonets method [29] offers significant performance advantages for multi-label few-shot learning, its computational complexity increases substantially with the number of labels due to the exponential growth of label combinations. In this work, we introduce an optimization that significantly improves inference efficiency while maintaining identical classification results.

The original approach creates an LC-Prototype (LCP) for each label combination (LC-class) derived from the power sets of the few available examples’ labels. Each available example is called a *support item* and it is defined by $(\mathbf{x}_i, \mathbf{y}_i)$, with \mathbf{x}_i being its input feature vector and \mathbf{y}_i the set of its labels. For the set of support items S , the set of all LC-classes L is computed as $L = \bigcup_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \mathcal{P}(\mathbf{y}_i)$, where $\mathcal{P}(\mathbf{y}_i)$ is the power set of the labels of the i -th support item, excluding the empty set. For each LC-class L_j , with $j = 1, 2, \dots, |L|$, the LCP representation \mathbf{p}_j is computed by averaging the embeddings of all support items that include L_j in their power sets:

$$\mathbf{p}_j = \frac{1}{|S_j|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S_j} f_\theta(\mathbf{x}_i), \quad (1)$$

where $S_j = \{(\mathbf{x}_i, \mathbf{y}_i) \in S \mid L_j \in \mathcal{P}(\mathbf{y}_i)\}$, and f_θ the embedding mapping model.

Our key insight is that multiple LC-classes often share identical LCP representations despite representing different label combinations. This occurs because the same set of support items contributes to multiple label combinations derived from their power sets. For example, if a support item with labels $\{A, B, C\}$ is the only item contributing

to both $\{A, B\}$ and $\{B, C\}$ LCPs, these LCPs will have identical representations.

We exploit this redundancy by maintaining a dictionary structure that maps unique LCP representations to their corresponding sets of LC-classes:

$$\text{UniqueLCPs} = \{\mathbf{p}_m \mapsto \{L_j \mid \mathbf{p}_j = \mathbf{p}_m\}\}, \quad (2)$$

where $j = 1, 2, \dots, |L|$ and $m = 1, 2, \dots, M$ with M being the number of unique LCPs and $M \ll |L|$. During inference, instead of computing distances between a *query item*, unseen during training, and all possible $|L|$ LCPs, we only compute distances to the M unique LCP representations. For the nearest unique LCP, we then select the label combination with the maximum cardinality, consistent with the original LC-Protonets method.

Our experiments show that this approach yields speed improvements of 10× for datasets with 20 labels, scaling to more than 100× for datasets with 60 labels, while producing identical classification results to the original method. We apply this optimization to the LC-Protonets repository⁷, making it practical for large label sets.

4. EXPERIMENTAL SETUP

Experiments and resources. We conducted 5 runs with different random seeds for both Probing and ML-FSL tasks, but a single run for SFT due to computational constraints. SFT trainable parameters varied: 14M for MERT-95M, 13M for MERT-330M, 25M for CLAP models, and 56M for Qwen2-Audio. All experiments ran on an NVIDIA RTX A5000 GPU, and we used Qwen2-Audio in half-precision (FP16) in all our methodologies to fit in this card. Most SFT training completed within 24 hours, with only 3 out of 30 experiments extending to about 36 hours.

Dataset processing. We standardized Turkish-makam, Hindustani, and Carnatic datasets to approximately 200 hours each, matching MagnaTagATune and FMA-medium durations [28], while Lyra remained at its original 80 hours. We followed the training, validation, and test splits from [17, 28]. For ML-FSL, evaluation items came exclusively from test sets [29] to prevent data leakage.

Model-specific configurations. Each foundation model required specific preprocessing: MERT models use 30-second windows at 24kHz, CLAP models 10-second windows at 48kHz, and Qwen2-Audio 30-second windows at 16kHz. All audio was converted to mono and resampled to the model’s required rate.

Representation extraction strategies. For MERT models, we extract representations by summing the average, across time, hidden states of the last four layers of the models. For CLAP models, we extract them from the audio projection layer which takes as input the average pooled layer representation of the last hidden state. For Qwen2-Audio, we use the last hidden state embeddings averaged across all layers of the whole model, when passing a simple text prompt that includes nothing but the respective tags for audio processing, i.e., `<|audio_bos|><|AUDIO|><|audio_eos|>`.

⁷<https://github.com/pxaris/LC-Protonets>

Model	Params Audio/Total	ROC-AUC (%)		mAP (%)	
		<i>Prob.</i>	<i>SFT</i>	<i>Prob.</i>	<i>SFT</i>
VGG-ish [28]	3.6M/3.6M	84.45		50.56	
MERT-95M	95M/95M	87.25 _{0.32}	87.26	52.25 _{0.42}	52.68
MERT-330M	330M/330M	85.40 _{0.68}	85.69	49.62 _{0.83}	50.47
CLAP-M	68M/194M	71.52 _{1.14}	78.96	29.98 _{1.07}	40.41
CLAP-M&S	68M/194M	86.78 _{0.31}	86.15	53.12 _{0.87}	51.99
Qwen2-Audio	637M/8.40B	88.59 _{0.47}	89.37	56.48 _{0.63}	58.73

Table 1. Model performance comparison averaged across all datasets for Probing and SFT tasks. Values are averaged over multiple runs with subscripted standard deviations. Bold values indicate best performance per column.

These representation extraction strategies, number of fine-tuned layers, and other design choices of our method were optimized through preliminary experiments.

Hyperparameters. For Probing, we used Adam optimizer [48] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with learning rate 10^{-3} , batch size 16, early stopping patience 10, and maximum 200 epochs. For SFT, we used AdamW [49] with identical β parameters but learning rate 10^{-4} , model-specific batch sizes (to fit maximum available resources) with gradient accumulation to simulate batch size 16 across all setups, patience 5, and maximum 30 epochs. We applied learning rate warmup and cosine scheduling for the first 5% of SFT epochs. ML-FSL evaluations used cosine distance with an N -way K -shot setup, with N being the number of extended tags per dataset and K equal to 3 examples per label in all experiments. We also attempted Low-Rank Adaptation [50] initially but abandoned it due to extensive hyperparameter tuning requirements across our 5×6 experimental matrix.

Evaluation metrics. For the Probing and SFT methodologies, we report area under the receiver operating characteristic curve (ROC-AUC) and mean average precision (mAP). These metrics are particularly well-suited for multi-label classification tasks [51] and are consistent with prior work in music tagging [17, 28]. For ML-FSL evaluation, we report macro-F1 (M-F1) and micro-F1 (m-F1) scores, which align with the LC-Protonets evaluation framework [29]. F1 score is the harmonic mean of the precision and recall scores. Macro-F1 gives equal weight to all classes, while micro-F1 accounts for class imbalance by calculating metrics globally across all instances.

5. RESULTS

5.1 Probing and Supervised Fine-Tuning

Table 1 presents the performance of the evaluated foundation models averaged across all datasets for both Probing and SFT tasks. Overall, Qwen2-Audio achieves the highest performance with 88.59% ROC-AUC and 56.48% mAP in Probing, further improving to 89.37% ROC-AUC and 58.73% mAP after fine-tuning. This is followed by MERT-95M and CLAP-Music&Speech with comparable performance, while CLAP-Music shows significantly lower performance without speech data in its training corpus.

Model	MagnaTagATune		FMA-medium		Lyra		Turkish-makam		Hindustani		Carnatic	
	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP
VGG-ish [28]	91.23	45.82	88.89	49.49	80.97	48.06	86.96	56.39	84.77	60.82	73.92	42.78
<i>Probing (Prob.)</i>												
MERT-95M	90.46 _{0.10}	44.16 _{0.21}	91.68 _{0.08}	51.43 _{0.43}	85.61 _{0.66}	53.34 _{0.61}	88.22 _{0.23}	57.89 _{0.34}	86.59 _{0.52}	60.26 _{0.56}	80.96 _{0.35}	46.41 _{0.35}
MERT-330M	89.66 _{0.16}	41.73 _{0.59}	90.78 _{0.11}	48.85 _{0.32}	84.65 _{0.78}	51.81 _{0.59}	85.37 _{0.64}	52.45 _{1.12}	84.23 _{1.36}	58.78 _{2.08}	77.73 _{1.03}	44.07 _{0.31}
CLAP-M	80.07 _{0.21}	25.82 _{0.13}	77.42 _{0.15}	22.89 _{0.38}	64.18 _{1.29}	31.16 _{0.43}	77.31 _{0.51}	38.77 _{1.00}	68.69 _{4.05}	33.43 _{4.21}	61.47 _{0.60}	27.83 _{0.30}
CLAP-M&S	92.41 _{0.05}	48.54 _{0.16}	94.05 _{0.08}	59.13 _{0.54}	87.25 _{0.18}	56.94 _{0.51}	86.49 _{0.27}	54.69 _{0.36}	82.61 _{1.14}	55.70 _{3.29}	77.85 _{0.13}	43.73 _{0.35}
Qwen2-Audio	91.17 _{0.13}	45.58 _{0.21}	96.60 _{0.07}	73.38 _{0.28}	86.44 _{0.81}	53.50 _{0.65}	86.64 _{0.42}	53.38 _{0.79}	88.45 _{0.83}	62.42 _{0.99}	82.22 _{0.56}	50.59 _{0.88}
<i>Supervised Fine-Tuning (SFT)</i>												
MERT-95M	90.62	44.52	91.70	51.74	84.89	53.62	87.50	57.91	88.20	61.47	80.64	46.83
MERT-330M	89.55	41.93	91.12	49.56	84.74	52.54	86.17	53.80	85.49	61.33	77.05	43.66
CLAP-M	88.54	39.26	88.37	42.04	71.97	38.14	79.82	42.49	75.65	45.01	69.39	35.51
CLAP-M&S	91.77	47.54	92.86	57.11	85.35	52.86	86.69	54.93	83.73	56.91	76.51	42.58
Qwen2-Audio	92.03	48.27	97.02	75.94	87.57	57.04	87.95	56.10	88.32	64.35	83.35	50.66
(Previous) SOTA	92.7	46.54	92.4	53.7	85.4	54.3	87.7	57.7	86.5	63.1	77.0	43.9

Table 2. Model performance on individual datasets for Probing and SFT tasks. For Probing, values are averaged over multiple runs with subscripted standard deviations, while SFT results are from single runs. Bold values indicate best performance per metric and dataset. SOTA values are from [52] for MagnaTagATune and [28] for the rest of the datasets.

Figure 2 illustrates the relationship between model size (audio-specific parameters) and ROC-AUC performance, averaged across datasets and both Probing and SFT tasks. A generally positive correlation is revealed, with similar trends observed in both methodologies. Qwen2-Audio (637M parameters) consistently outperforms smaller models, achieving 88.98% average ROC-AUC score. Surprisingly, MERT-95M (87.25%) outperforms the much larger MERT-330M (85.55%). This is worth noting as [33] reported that both models performed on par for auto-tagging tasks, suggesting that our common representation extraction strategy for both MERT models may not optimally leverage the larger model’s capacity. Another potential explanation is that MERT-95M has been trained on open data whereas MERT-330M has been trained with additional proprietary data with a strong Western bias [6].

When examining Probing (*Prob.*) performance across individual datasets, in Table 2, we observe a consistent pattern of decreasing performance for music traditions that are culturally distant from the data used to pre-train the respective foundation models. Western music datasets (MagnaTagATune and FMA-medium) consistently achieve the highest performance across all models, with ROC-AUC values reaching 96.60% for Qwen2-Audio on FMA-medium. Greek (Lyra) and Turkish (makam) music datasets show moderate performance, while Indian classical music (Hindustani and Carnatic) datasets typically exhibit the lowest performance. This cultural performance gap is especially pronounced for CLAP-Music, where the ROC-AUC drops from 80.07% for MagnaTagATune to 61.47% for Carnatic.

Applying Supervised Fine-Tuning (*SFT*) generally improves performance across all models and datasets, with an average gain of 1-2% in ROC-AUC for most models. Notably, CLAP-Music shows the largest improvement with *SFT*, indicating greater adaptation potential despite lower absolute performance. For other models, the modest gains suggest that they require broader fine-tuning to further shift their pre-trained representations towards different cultures.

Importantly, our approaches achieve state-of-the-art performance in five out of six datasets, with MagnaTa-

Model	M-F1			m-F1		
	30.18			55.09		
VGG-ish [29]	<i>PT</i>	<i>Prob.</i>	<i>SFT</i>	<i>PT</i>	<i>Prob.</i>	<i>SFT</i>
MERT-95M	23.90 _{1.52}	28.05 _{1.74}	28.28 _{1.80}	46.59 _{1.57}	52.16 _{1.43}	52.56 _{1.63}
MERT-330M	23.03 _{1.12}	28.48 _{1.40}	28.51 _{1.28}	45.11 _{1.29}	51.78 _{1.51}	51.80 _{1.46}
CLAP-M	17.71 _{1.20}	18.43 _{1.40}	21.58 _{1.13}	38.80 _{1.37}	39.97 _{1.20}	46.57 _{1.20}
CLAP-M&S	28.23 _{1.36}	29.22 _{1.09}	30.27 _{1.90}	51.59 _{1.54}	53.32 _{1.31}	54.43 _{1.27}
Qwen2-Audio	25.98 _{1.36}	30.96 _{1.26}	32.00 _{1.41}	49.97 _{1.41}	55.66 _{0.82}	56.85 _{1.23}

Table 3. ML-FSL performance averaged across datasets on extended tag sets. Results show macro-F1 (M-F1) and micro-F1 (m-F1) across contexts (*PT*, *Prob.*, *SFT*). Values are means with subscripted standard deviations. Bold indicates best performance per column.

gATune being the only exception. However, their consistent performance decrease towards diverse cultures, suggests that their representations are still biased toward Western musical traditions.

5.2 Multi-label few-shot learning

Table 3 presents the ML-FSL evaluation results averaged across all datasets using extended tag sets. The results show consistent performance improvements moving from pre-trained models (*PT*) to trained probing models (*Prob.*) and then to supervised fine-tuned models (*SFT*) across all foundation models. The substantial gap between macro-F1 and micro-F1 metrics indicates considerable class imbalance in the extended tag sets, while the increased standard deviation stems from the support set sampling which can significantly impact the classification performance.

Qwen2-Audio demonstrates the best overall performance in the ML-FSL task with 32.00% macro-F1 and 56.85% micro-F1 after fine-tuning, followed closely by CLAP-Music&Speech with 30.27% macro-F1 and 54.43% micro-F1. Notably, even the best foundation model’s performance (Qwen2-Audio) is comparable to a VGG-ish feature extractor trained via supervised learning on standard tags for each dataset. This stands in contrast to the Probing and SFT settings (Table 1), where foundation models clearly outperform VGG-ish, showing that ML-FSL tasks

Model	MagnaTagATune		FMA-medium		Lyra		Turkish-makam		Hindustani		Carnatic	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish [29]	26.40	37.31	29.12	45.37	46.05	69.03	30.07	56.22	31.33	58.38	18.13	64.25
<i>Pre-Trained models (PT)</i>												
MERT-95M	18.76 _{1.04}	28.37 _{1.38}	16.24 _{0.64}	35.37 _{0.94}	46.87 _{2.59}	66.07 _{2.25}	20.69 _{1.77}	40.95 _{1.80}	25.87 _{2.45}	51.50 _{1.92}	14.97 _{0.64}	57.26 _{1.10}
MERT-330M	18.17 _{0.78}	26.99 _{1.36}	16.24 _{0.69}	31.15 _{1.51}	44.22 _{1.45}	65.48 _{1.57}	20.14 _{2.01}	39.71 _{1.95}	25.08 _{1.40}	50.14 _{1.08}	14.32 _{0.41}	57.21 _{0.28}
CLAP-M	13.10 _{0.84}	20.00 _{1.15}	9.65 _{0.29}	19.77 _{1.31}	33.56 _{2.88}	57.14 _{1.49}	14.33 _{1.10}	32.12 _{1.37}	21.06 _{1.63}	47.38 _{1.60}	14.55 _{0.43}	56.42 _{1.30}
CLAP-M&S	25.90 _{0.55}	36.55 _{0.61}	28.78 _{1.66}	42.95 _{2.02}	48.03 _{2.02}	69.04 _{1.54}	24.19 _{1.73}	47.13 _{2.22}	26.29 _{1.20}	54.50 _{1.57}	16.19 _{1.02}	59.38 _{1.30}
Qwen2-Audio	21.29 _{0.51}	32.09 _{0.26}	29.76 _{2.23}	47.50 _{1.86}	39.99 _{1.05}	64.24 _{1.07}	19.89 _{1.71}	42.27 _{1.88}	28.42 _{1.96}	55.92 _{1.70}	16.55 _{0.69}	57.82 _{1.67}
<i>Trained Probing models (Prob.)</i>												
MERT-95M	23.77 _{0.85}	34.71 _{1.03}	24.62 _{1.19}	42.96 _{1.30}	45.80 _{2.76}	68.16 _{1.81}	26.14 _{1.73}	50.00 _{0.70}	30.75 _{2.95}	56.41 _{2.18}	17.25 _{0.98}	60.70 _{1.55}
MERT-330M	24.48 _{0.59}	34.78 _{1.45}	25.21 _{0.76}	40.65 _{1.76}	47.92 _{3.26}	70.15 _{2.18}	26.97 _{1.61}	50.47 _{1.13}	29.25 _{1.55}	53.77 _{1.82}	17.06 _{0.61}	60.85 _{0.69}
CLAP-M	14.84 _{0.49}	22.67 _{1.00}	11.55 _{0.50}	22.72 _{1.48}	34.85 _{4.03}	57.73 _{1.37}	16.68 _{0.81}	36.00 _{1.22}	18.77 _{1.42}	44.96 _{0.95}	13.87 _{1.16}	55.74 _{1.15}
CLAP-M&S	26.90 _{0.47}	37.62 _{0.93}	31.14 _{1.28}	46.53 _{1.59}	47.10 _{0.89}	69.77 _{0.53}	25.58 _{1.59}	49.70 _{1.39}	28.11 _{1.38}	56.43 _{2.19}	16.46 _{0.92}	59.88 _{1.25}
Qwen2-Audio	26.79 _{0.40}	37.65 _{0.21}	39.49 _{1.02}	56.30 _{0.82}	42.52 _{1.81}	67.10 _{1.13}	26.09 _{1.65}	51.59 _{1.20}	31.62 _{1.26}	60.08 _{0.40}	19.25 _{1.40}	61.24 _{1.14}
<i>Supervised Fine-Tuned models (SFT)</i>												
MERT-95M	24.46 _{0.79}	35.28 _{0.90}	24.94 _{1.18}	42.78 _{1.44}	45.51 _{3.74}	67.93 _{2.72}	26.16 _{1.87}	49.76 _{1.54}	30.40 _{2.15}	56.39 _{1.68}	18.18 _{1.08}	63.19 _{1.48}
MERT-330M	23.78 _{0.65}	33.67 _{0.91}	24.94 _{1.21}	39.95 _{1.77}	48.50 _{2.75}	70.06 _{2.23}	26.84 _{1.51}	50.29 _{1.25}	30.56 _{1.31}	55.25 _{1.58}	16.43 _{0.27}	61.57 _{1.04}
CLAP-M	22.15 _{0.51}	32.67 _{1.22}	19.61 _{0.79}	34.81 _{0.99}	30.46 _{2.04}	55.86 _{2.02}	20.66 _{1.69}	45.80 _{1.13}	21.95 _{1.31}	50.74 _{1.14}	14.63 _{0.45}	59.53 _{0.67}
CLAP-M&S	26.28 _{0.50}	37.23 _{1.09}	30.27 _{1.56}	46.57 _{1.61}	48.09 _{4.74}	69.93 _{2.28}	28.91 _{1.75}	53.87 _{1.56}	31.27 _{2.47}	57.41 _{0.74}	16.82 _{0.37}	61.55 _{0.34}
Qwen2-Audio	27.67 _{0.25}	38.57 _{0.18}	40.10 _{1.29}	57.17 _{0.95}	44.13 _{2.45}	68.34 _{2.38}	27.61 _{2.37}	53.98 _{1.55}	32.52 _{1.23}	60.26 _{0.89}	19.97 _{0.87}	62.76 _{1.43}

Table 4. ML-FSL performance on extended tag sets per dataset. Results show macro-F1 (M-F1) and micro-F1 (m-F1) across three contexts. Values are means with subscripted standard deviations. Bold indicates best performance per column.

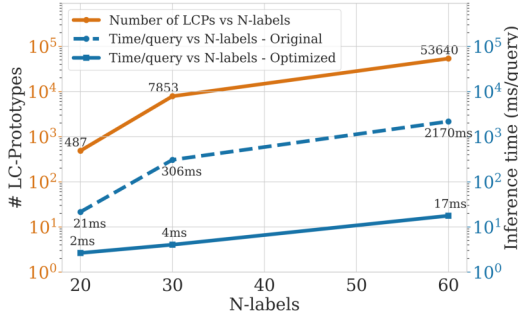


Figure 3. Scalability metrics of the LC-Protonets method, averaged across all datasets. The x -axis represents the number of labels, the left y -axis shows the number of LCPs, and the right y -axis indicates the inference time per item with both y -axes using the same logarithmic scale.

remain challenging for them despite their extensive pre-training. Supervised learning of a VGG-ish model on extended tag sets has not been conducted in the literature, likely due to the scarcity of examples for infrequent tags.

When examining the ML-FSL results per dataset in Table 4, we observe that only on Western datasets (MagnaTagATune and FMA-medium) does the best foundation model (Qwen2-Audio) achieve significantly better performance than the VGG-ish baseline. For Turkish-makam, VGG-ish representations actually outperform foundation models, while for Lyra, Hindustani, and Carnatic, the results are comparable. This pattern provides additional clear evidence of the implicit Western-centric bias integrated into models due to their pre-training data.

LC-Protonets optimization. Figure 3 illustrates the scalability metrics of our optimized LC-Protonets approach compared to the original method, averaged across datasets. As the number of labels increases from 20 to 60, the number of LC-Prototypes grows exponentially, from approximately 500 to over 50,000. This growth leads the origi-

nal method to a corresponding increase from 21ms to over 2,000ms inference time per query item (dashed blue line). However, our optimization (solid blue line), leveraging the unique prototypes, mitigates the computational complexity issues, requiring only 2ms in the 20 labels cases and rising to no more than 20ms for 60 labels, a 100 \times improvement.

6. CONCLUSIONS

In this paper, we examined the universality of music representations in foundation models through a comprehensive methodological framework evaluating five state-of-the-art audio models across six world music corpora. Although these models achieved better performance than previous models for diverse music traditions, we found clear indicators of Western-centric bias.

Our incorporation of ML-FSL tasks particularly revealed this limitation. When faced with these challenging scenarios, foundation models performed on par with significantly smaller and simpler models, with performance notably degrading further on non-Western datasets.

To further enable ML-FSL evaluation, we substantially optimized the computational complexity of the utilized method, by forming unique prototypes representing multiple label combinations. We demonstrated that this change makes it practical for large sets of labels, a typical condition when studying world music datasets.

Future work could extend our methodological framework by incorporating Low-Rank Adaptation (LoRA) and implement broader supervised fine-tuning to investigate further cultural adaptation. More tasks can also be included such as mode estimation, exploring the analogies between key on Western cultures and makam or raga recognition in other cultures.

We hope this work brings attention to the cultural dimensions of foundation models while providing a framework for quantitatively assessing progress toward truly universal musical representations.

7. ACKNOWLEDGMENTS

We would like to thank the reviewers for their valuable and constructive feedback, which helped us improve our study. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

8. REFERENCES

- [1] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, R. M. Howard *et al.*, “Universality and diversity in human song,” *Science*, vol. 366, 2019.
- [2] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie, “Statistical universals reveal the structures and functions of human music,” *Proceedings of the National Academy of Sciences*, vol. 112, pp. 8987 – 8992, 2015.
- [3] S. E. Trehub, J. Becker, and I. Morley, “Cross-cultural perspectives on music and musicality,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, 2015.
- [4] E. H. Margulis, P. C. M. Wong, C. Turnbull, B. M. Kubit, and J. D. McAuley, “Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, 2022.
- [5] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel *et al.*, “On the opportunities and risks of foundation models,” *CoRR*, vol. abs/2108.07258, 2021.
- [6] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, “MERT: acoustic music understanding model with large-scale self-supervised training,” in *ICLR*. OpenReview.net, 2024.
- [7] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP*. IEEE, 2023, pp. 1–5.
- [8] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *CoRR*, vol. abs/2311.07919, 2023.
- [9] M. Won, Y. Hung, and D. Le, “A foundation model for music informatics,” in *ICASSP*. IEEE, 2024, pp. 1226–1230.
- [10] Y. Ma, A. Øland, A. Ragni, B. M. Del Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Quinton *et al.*, “Foundation models for music: A survey,” *CoRR*, vol. abs/2408.14340, 2024.
- [11] R. Castellon, C. Donahue, and P. Liang, “Codified audio language modeling learns useful representations for music information retrieval,” in *ISMIR*, 2021, pp. 88–96.
- [12] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” *CoRR*, vol. abs/2005.00341, 2020.
- [13] M. C. McCallum, F. Korzeniewski, S. Oramas, F. Gouyon, and A. F. Ehmann, “Supervised and unsupervised learning of audio representations for music understanding,” in *ISMIR*, 2022, pp. 256–263.
- [14] Y. Li, R. Yuan, G. Zhang, Y. Ma, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, E. Benetos, N. Gyenge, R. Liu, and J. Fu, “Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning,” *CoRR*, vol. abs/2212.02508, 2022.
- [15] K. Choi, “Deep neural networks for music tagging,” Ph.D. dissertation, Queen Mary University of London, UK, 2018.
- [16] T. Kim, J. Lee, and J. Nam, “Sample-level CNN architectures for music auto-tagging using raw waveforms,” in *ICASSP*. IEEE, 2018, pp. 366–370.
- [17] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, “Evaluation of cnn-based automatic music tagging models,” *CoRR*, vol. abs/2006.00751, 2020.
- [18] J. Lee, J. Park, K. L. Kim, and J. Nam, “Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms,” *CoRR*, vol. abs/1703.01789, 2017.
- [19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, “CNN architectures for large-scale audio classification,” in *ICASSP*. IEEE, 2017, pp. 131–135.
- [20] J. Pons and X. Serra, “musicnn: Pre-trained convolutional neural networks for music audio tagging,” *CoRR*, vol. abs/1909.06654, 2019.
- [21] Y. Gong, Y. Chung, and J. R. Glass, “AST: audio spectrogram transformer,” in *Interspeech*. ISCA, 2021, pp. 571–575.
- [22] M. Panteli, “Computational analysis of world music corpora,” Ph.D. dissertation, Queen Mary University of London, UK, 2018.
- [23] E. Demirel, B. Bozkurt, and X. Serra, “Automatic makam recognition using chroma features,” in *8th International Workshop on Folk Music Analysis*, 2018, pp. 19–24.

- [24] K. K. Ganguli, S. Sentürk, and C. Guedes, “Critiquing task- versus goal-oriented approaches: A case for makam recognition,” in *ISMIR*, 2022, pp. 369–376.
- [25] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya *et al.*, “Classification of indian classical music with time-series matching deep learning approach,” *IEEE Access*, vol. 9, pp. 102 041–102 052, 2021.
- [26] B. Nikzat and R. C. Repetto, “KDC: an open corpus for computational research of dastgāhi music,” in *ISMIR*, 2022, pp. 321–328.
- [27] D. Han, R. C. Repetto, and D. Jeong, “Finding tori: Self-supervised learning for analyzing korean folk song,” in *ISMIR*, 2023, pp. 440–447.
- [28] C. Papaioannou, E. Benetos, and A. Potamianos, “From west to east: Who can understand the music of the others better?” in *ISMIR*, 2023, pp. 311–318.
- [29] —, “LC-Protonets: Multi-label few-shot learning for world music audio tagging,” *IEEE Open Journal of Signal Processing*, vol. 6, pp. 138–146, 2025.
- [30] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” in *NIPS*, 2017, pp. 4077–4087.
- [31] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, M. Henry, N. Pinto, C. Noufi *et al.*, “HEAR: holistic evaluation of audio representations,” in *NeurIPS (Competition and Demos)*, ser. Proceedings of Machine Learning Research, vol. 176. PMLR, 2021, pp. 125–145.
- [32] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee *et al.*, “SUPERB: speech processing universal performance benchmark,” in *Interspeech*. ISCA, 2021, pp. 1194–1198.
- [33] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, L. Zhuo, Y. Liu, J. Huang, Z. Tian, B. Deng, N. Wang, C. Lin, E. Benetos, A. Ragni *et al.*, “MARBLE: music audio representation benchmark for universal evaluation,” in *NeurIPS*, 2023.
- [34] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, “HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection,” in *ICASSP*. IEEE, 2022, pp. 646–650.
- [35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*. IEEE, 2021, pp. 9992–10 002.
- [36] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, “Qwen2-audio technical report,” *CoRR*, vol. abs/2407.10759, 2024.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, 2017.
- [38] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.
- [39] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging,” in *ISMIR*, 2009, pp. 387–392.
- [40] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *ISMIR*, 2017, pp. 316–323.
- [41] C. Papaioannou, I. Valiantzas, T. Giannakopoulos, M. A. Kaliakatsos-Papakostas, and A. Potamianos, “A dataset for greek traditional and folk music: Lyra,” in *ISMIR*, 2022, pp. 377–383.
- [42] X. Serra, “Creating research corpora for the computational study of music: the case of the compmusic project,” in *Semantic Audio*. Audio Engineering Society, 2014.
- [43] B. Uyar, H. S. Atli, S. Sentürk, B. Bozkurt, and X. Serra, “A corpus for computational research of turkish makam music,” in *DLfM@JCDL*. ACM, 2014, pp. 1–7.
- [44] S. Sentürk, “Computational analysis of audio recordings and music scores for the description and discovery of ottoman-turkish makam music,” Ph.D. dissertation, Pompeu Fabra University, Spain, 2017.
- [45] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, “Corpora for music information research in indian art music,” in *ICMC*. Michigan Publishing, 2014.
- [46] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, “Overcoming catastrophic forgetting in neural networks,” *CoRR*, vol. abs/1612.00796, 2016.
- [47] K. Gupta, B. Thérien, A. Ibrahim, M. L. Richter, Q. Anthony, E. Belilovsky, I. Rish, and T. Lesort, “Continual pre-training of large language models: How to (re)warm your model?” *CoRR*, vol. abs/2308.04014, 2023.
- [48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

- [49] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *CoRR*, vol. abs/1711.05101, 2019.
- [50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *ICLR*. Open-Review.net, 2022.
- [51] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
- [52] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, “Mulan: A joint embedding of music audio and natural language,” in *ISMIR*, 2022, pp. 559–566.