Evaluating Pretrained Music Embeddings for Cross-Performance Jazz Standard Recognition

Cagri Eser [1]

Abstract

Recognizing jazz standards from audio is a challenging form of tune-level music retrieval: different performances of the same standard may vary in tempo, key, arrangement, instrumentation, improvisational content, and even whether the head melody is present. We study this problem using a curated subset of the Jazz Trio Database designed for cross-performance standard recognition. We compare a from-scratch trained Harmonic CNN baseline against frozen pretrained music representations from recent music understanding foundation models, using both supervised probing and nearest-neighbor retrieval. Our results suggest that from-scratch spectrogram models overfit strongly to training performances, while pretrained embeddings provide better top-k results but are sensitive to performer identity, which can be partially reduced with a lightweight contrastive projection. Our findings motivate jazz standard recognition as a useful stress test for music representation models and as a step toward retrieval-based standard identification. Project page: https://github.com/cagries/tipofmyear.

1. Introduction

Audio recognition systems such as Shazam (Wang, 2003) are highly effective for identifying exact recordings, but recognizing the underlying tune across different performances is a different problem. This distinction is especially important in jazz: a standard such as *Autumn Leaves* can appear in many keys, different tempo, different arrangements, and in many improvisational contexts. In this paper, we study whether modern audio representations support this form of cross-performance tune recognition. The task is difficult for several reasons. First, jazz performances often devote long stretches to improvisation, where cropped local windows may not contain the head melody. Second, different standards can share common harmonic progressions or similar melodic fragments. Third, recordings by the same group can be acoustically and stylistically similar across performances of different standards, which causes problems for nearest-neighbor retrieval methods.

Our work makes three contributions. **First**, we construct a curated filtered benchmark subset from the Jazz Trio Database (Cheston et al., 2024) for standard recognition across performances. **Second**, we compare spectrogram-based training, supervised probing of frozen embeddings, and embedding-based retrieval under the same evaluation protocol. **Third**, we analyze retrieval-based classification and demonstrate that nearest neighbors often retrieve performer identity rather than tune identity, and explore a lightweight supervised contrastive projection to reduce performer-biased retrieval.

2. Related Work

Fingerprinting systems for audio such as Shazam (Wang, 2003) are very effective for exact recording identification from short and noisy excerpts, however they are designed to match a given signal to a recording already present in the database rather than recognizing a new performance of the same underlying composition. With this in mind, jazz standard recognition is therefore closer to cover or version identification (Serrà et al., 2011; Xun et al., 2023), which is concerned with retrieval of different renditions of the same work using features that are invariant to changes in key, tempo, timbre, and arrangement. Recent version-identification systems have increasingly adopted embedding-based retrieval formulations to improve scalability while preserving work-level invariance (Xun et al., 2023).

Recent developments in music foundation models further motivate our approach. Earlier work showed that pretrained audio representations can transfer well across downstream MIR tasks with shallow heads or no fine-tuning, including codified-audio language-model features from Jukebox-

[1]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey. Correspondence to: Cagri Eser <cagri.eser@ceng.metu.edu.tr>.

Accepted to the *Workshop on Machine Learning for Audio* at the 43^{rd} International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.

based MIR (Castellon et al., 2021) and broader comparative studies of supervised and unsupervised music embeddings (McCallum et al., 2022). MERT (Li et al., 2024) scales masked self-supervised pretraining with acoustic and musical teachers and reports strong transfer across various music-understanding tasks, while MuQ uses Mel quantization targets and shows broad downstream gains with improvements with larger-scale pretraining. A closely related reference in approach is Papaioannou et al. (2025), who evaluate multiple audio foundation models through probing and lightweight fine-tuning across culturally diverse tagging datasets. Our setting is complementary because it probes jazz tune identity across performances rather than tag prediction across cross-cultural musical traditions.

This work complements previous approaches in the literature by using jazz standards as a deliberately hard retrieval target where unlike fingerprinting the goal is not exact recording recognition, and unlike generic cover-song benchmarks many query windows may not contain an explicit statement of the melody. Methodologically, it is closest to recent evaluation-first studies of music foundation models that compare probing and lightweight adaptation, however our target is cross-performance jazz standard identification under heavy improvisation.

3. Dataset, Evaluation and Experiment Setup

Dataset construction. Given our focus on jazz performances, we use the Jazz Trio Database (Cheston et al., 2024) and curate a subset of the data intended for standard-level recognition. The JTD dataset has 1294 performances from various standards, however the distribution of the data is long-tailed: more than 85% of standards from the dataset have fewer than 3 performances from distinct bandleader/pianist groups. From this observation, we construct a filtered dataset based on the following approach: we first canonicalize standard titles to merge spelling and naming variants. Afterwards, we collect all standards with **at least** 4 performances from distinct groups. This threshold is chosen to support our leave-one-performance-per-standard evaluation so that in each fold one performance of every standard is held out for testing, while the remaining performances provide training and validation examples. Finally, for each standard, we keep **at most one performance** from any given group, reducing leakage from repeated performances by the same ensemble. The resulting training subset is not perfectly class-balanced, but it greatly reduces the long-tailed structure of the original dataset and ensures multiple performances and a balanced test setup. Figure 1 shows counts of various standards within the subset. With these adjustments, the evaluation dataset contains 16 standards with 79 performances, with more metadata described in Table 1. For each selected recording, we assign a canonical standard label and segment the audio into fixed-length windows. We use the fixed standards list for all reported experiments.

Table 1. Statistics for the curated JTD standard recognition subset.

Statistic	Value
Standards	16
Performances	79
Performances per standard	4.93
Unique Groups	27
Window length / hop	10 seconds / 5 seconds

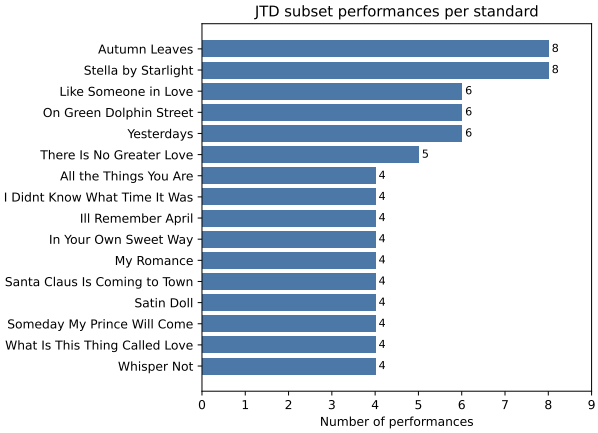

Figure 1. Distribution of performances across the curated JTD subset.

Windowing strategy. Each recording is converted to 24 kHz mono audio and split into 10-second windows with 5-second hop. Each window inherits the standard label of its parent performance. This creates many training examples per performance, but these windows are highly correlated within a recording. Therefore, we report performance-level metrics in addition to window-level metrics.

Splits and evaluation protocol. Given the small number of recordings in the final subset, we use a leave-one-performance-out evaluation protocol. For each evaluation fold, one performance for each standard is held out as the test set, while remaining performances form the training set. Windows from the same held-out performance do not appear in the training set. When training supervised models, one additional random performance from the training side is reserved for validation and hyperparameter tuning. For the final reported fold results, hyperparameters were fixed and models were trained on the full training partition before being evaluated on the held-out performances.

Evaluation Metrics. We report three metrics based on classification and retrieval. Window accuracy measures

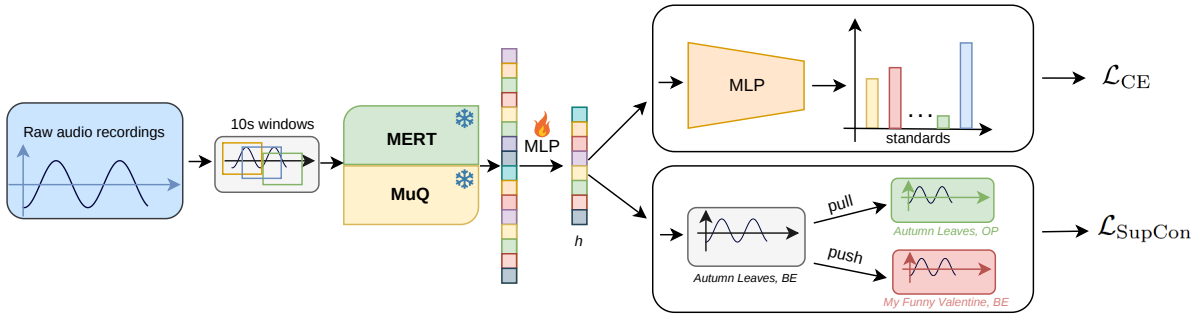

Figure 2. Pipeline for the proposed standard-aware supervised contrastive retrieval approach. Frozen MERT/MuQ embeddings are projected into a retrieval space trained to pull together windows from the same standard across different performances while reducing the same-performer retrieval bias.

Top-1 classification of individual 10-second windows. Performance Top-1 accuracy aggregates predictions over all windows from a held-out performance and evaluates the Top-1 standard prediction, corresponding to the strict standard identification setting. Performance Top-5 reports whether the true standard appears among the five highest-scoring standards after performance-level aggregation, which better reflects a retrieval-based use case for jazz standard recognition. We include window accuracy mainly as a diagnostic metric, since many individual windows may be noisy, or may not contain sufficient melodic or harmonic evidence for a given standard.

4. Methods

From-scratch Harmonic CNN. As a supervised baseline trained directly on the subset, we train a Harmonic CNN-style (Won et al., 2020) model on harmonic time-frequency representations of short audio windows. The model predicts a standard label for each window. At test time, we average window-level class probabilities over each held-out performance and choose the highest-scoring standard.

Frozen pretrained embeddings. For this task, we extract frozen audio embeddings using MERT (Li et al., 2024) and MuQ (Zhu et al., 2025). MERT-v1-95M[1] is a pretrained musical understanding model with 95M parameters, while MuQ[2] is a similar larger model with 300M parameters with an open-weight version trained on the Million Song Dataset (Bertin-Mahieux et al., 2011). For each window, hidden states are mean-pooled over time and concatenated across selected layers. We then train lightweight classifiers on top of these frozen embeddings. Linear probing tests whether standard identity is linearly separable in the representation, while a small MLP classification head is used to test whether a nonlinear classifier improves the performance of the probe.

[1]https://huggingface.co/m-a-p/MERT-v1-95M
[2]https://huggingface.co/OpenMuQ/MuQ-large-msd-iter

kNN-based retrieval. Given that performances from the same standard should follow similar melodies and underlying harmonies in theory, we also evaluate nearest-neighbor retrieval. Here, we build a reference set from embeddings from 10-second windows from the training split with MERT and MuQ encoders and L2-normalize each embedding. For each query window, we retrieve the top-k nearest reference windows and aggregate their standard labels using temperature-scaled weighted voting with cosine similarity. We obtain performance-level scores by averaging window-level scores across all query windows.

5. Results

We compare the results of model training on the filtered dataset in Table 2. In particular, from the results we observe the following interesting properties:

From-scratch training overfits on the curated subset. The Harmonic CNN (Won et al., 2020) baseline trains well but its held-out performance remains close to chance-level under both Top-1 and Top-5 accuracy metrics, with worse-than-random performance on Top-1 accuracy. This suggests that the model learns performance-specific or recording-specific structure rather than a standard-invariant representation. This behavior is expected given the small number of independent performances in the dataset and the relative difficulty of the standard identification problem. The HCNN also shows low window-level accuracy across the test set, which supports these observations.

Pretrained embeddings consistently improve Top-k recognition. Linear and MLP probes on MERT and MuQ embeddings improve over the from-scratch HCNN baseline and over random prediction odds, especially in the performance-level Top-5 accuracy metric. This indicates that audio foundation models indeed learn useful general-purpose representations for improved standard recognition,

Table 2. Standard recognition results on the 16-standard JTD subset with 10-second windows. Random choice yields Top-1 accuracy of $1/16 = 0.0625$ and Top-5 accuracy of $5/16 = 0.3125$. We report mean ± standard deviation across folds.

Method	Representation	Window Acc.	Perf. Top-1	Perf. Top-5
Harmonic CNN (Won et al., 2020)	spectrogram	0.034 ± 0.012	0.031 ± 0.036	0.359 ± 0.079
MERT-v1 (Li et al., 2024)				
w/ linear probe	frozen embedding	0.074 ± 0.056	**0.094** ± 0.081	0.359 ± 0.139
w/ MLP probe	frozen embedding	0.096 ± 0.078	**0.094** ± 0.091	0.422 ± 0.164
w/ kNN retrieval ($k = 5$)	frozen embedding	0.066 ± 0.065	0.063 ± 0.051	0.359 ± 0.180
MuQ (Zhu et al., 2025)				
w/ linear probe	frozen embedding	0.085 ± 0.030	0.078 ± 0.031	**0.469** ± 0.149
w/ MLP probe	frozen embedding	**0.108** ± 0.068	0.078 ± 0.060	0.438 ± 0.102
w/ kNN retrieval ($k = 5$)	frozen embedding	0.060 ± 0.058	0.078 ± 0.079	0.359 ± 0.107

even when Top-1 recognition remains difficult. Although we assumed that the relative high dimensionality of the embeddings may be noisy, in our preliminary experiments dimensionality reduction with PCA before probing did not consistently improve these results.

Retrieval is promising but naive retrieval is limited by performer identity. Nearest-neighbor retrieval eliminates the need for fine-tuning on the dataset toward a Shazam-like system because new standards can be added by inserting embeddings into the reference database. However, retrieval errors show a particular tendency: nearest neighbors are often drawn from the same trio or performer but correspond to different standards. This suggests that the embedding space of pretrained models preserves performer or recording similarity in addition to jazz standard identity. We revisit this idea in Section 6.

6. Ablations

Effect of window length. With the relatively long duration of solos in performances from the dataset and low window accuracies from methods in Table 2, we investigate whether using a longer window length helps overall recognition for pretrained models. Albeit minor and model-specific, results from Table 3 show that increasing the window length most notably increases the Top-1 accuracy of the MuQ + MLP probe approach to a value of 0.125, while being otherwise uninformative for other configurations.

Retrieval failures and improvements. On investigation of the window-based retrieval approach from pretrained embeddings, we notice a particular trend: during retrieval, query windows from a given performance are often matched to windows from the same trio or performer but a different standard. For example, query sections of *Autumn Leaves* from Ahmed Jamal might retrieve embeddings from Ahmed Jamal's *Someday My Prince Will Come*, due to similarity in playing style, volume, or recording composition. Although these properties do represent a degree of similarity, this is not our target standard-focused similarity measure.

As a lightweight adaptation towards this problem, we experiment with the following pipeline. We keep the pretrained audio encoder frozen and train only a lightweight projection model on top of its window-level embeddings. Given a 10-second audio window, we first extract a fixed embedding using MERT or MuQ. This embedding is passed through a two-layer projection MLP to obtain a normalized representation z_i, and an MLP classification head maps z_i to logits over the set of jazz standards. The model is trained with a combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SupCon}}, \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ denotes standard cross-entropy between the MLP head and the target standard label and $\mathcal{L}_{\text{SupCon}}$ represents a supervised contrastive objective (Khosla et al., 2020). For a mini-batch of projected embeddings $\{z_i\}_{i=1}^{N}$, the supervised contrastive loss for anchor i is

$$\mathcal{L}_{\text{SupCon},i} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s_{ip}/\tau)}{\sum_{a \in A(i)} \exp(s_{ia}/\tau)},$$
$$s_{ij} = \text{sim}(\mathbf{z}_i, \mathbf{z}_j), \tag{2}$$

where $P(i)$ is the set of positive examples for anchor i, $A(i) = \{1, \ldots, N\} \setminus i$, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature parameter. In this setup, positives are windows with the same standard label but from a different performance, so that the projection is encouraged to learn cross-performance standard identity rather than recording similarity. After training, we discard the MLP head and use the learned projected embeddings for retrieval. Figure 2 shows the proposed pipeline.

We track the effectiveness of this method through the "Same Group Frequency" feature: while this ratio is high for vanilla

Table 3. Standard recognition results with varying window length. We report mean ± standard deviation across folds.

Method	Window Len.	Window Acc.	Perf. Top-1	Perf. Top-5
MERT-v1 (Li et al., 2024) + linear probe	10s	0.074 ± 0.056	0.094 ± 0.081	0.359 ± 0.139
	20s	0.065 ± 0.054	0.078 ± 0.060	0.375 ± 0.161
MuQ (Zhu et al., 2025) + linear probe	10s	0.085 ± 0.030	0.078 ± 0.031	**0.469** ± 0.149
	20s	0.061 ± 0.015	0.063 ± 0.016	0.453 ± 0.107
MERT-v1 (Li et al., 2024) + MLP probe	10s	0.096 ± 0.078	0.094 ± 0.091	0.422 ± 0.164
	20s	0.088 ± 0.065	0.094 ± 0.081	0.453 ± 0.164
MuQ (Zhu et al., 2025) + MLP probe	10s	0.108 ± 0.068	0.078 ± 0.060	0.438 ± 0.102
	20s	**0.113** ± 0.090	**0.125** ± 0.088	**0.469** ± 0.063

Table 4. Retrieval-based standard recognition results with 10-second windows.

Method	Same Group Freq.	Window Acc.	Perf. Top-1	Perf. Top-5
MERT-v1 (Li et al., 2024)				
w/ kNN probe ($k = 5$)	0.336	0.066 ± 0.065	**0.063** ± 0.051	0.359 ± 0.180
w/ kNN + SupCon ($\lambda = 0.2$)	**0.109**	**0.081** ± 0.078	**0.063** ± 0.051	**0.469** ± 0.120
MuQ (Zhu et al., 2025)				
w/ kNN probe ($k = 5$)	0.328	0.060 ± 0.058	0.078 ± 0.079	0.359 ± 0.107
w/ kNN + SupCon ($\lambda = 0.2$)	**0.156**	**0.095** ± 0.066	**0.109** ± 0.060	**0.438** ± 0.072

kNN-based matching, we observe that this value drops significantly for kNN-based retrieval trained with the auxiliary supervised contrastive loss, and the overall Top-1 and Top-5 accuracy of the models appear to improve as well, as highlighted in Table 4. We set $\lambda = 0.2$ based on preliminary validation experiments and keep it fixed across all folds to avoid any fold-specific tuning.

7. Discussion and Future Work

Our experiments suggest that jazz standard recognition is a challenging stress test for music representations. The difficulty is not merely the number of classes. In our curated dataset, many recordings appear to omit the head or begin in sections where the standard melody is not clearly present, which makes the task much harder than melody recognition. Consequently, a randomly sampled 10-second window may be considered difficult even for a human listener to identify. This makes window-level labels noisy: a solo passage labeled as *Stella by Starlight* may contain little direct evidence of that standard, and we cannot accurately predict what is a "good" target accuracy for a trained model.

Future work. The gap between Top-1 and Top-5 performance suggests that pretrained models often retrieve or classify within a plausible neighborhood but do not reliably rank the correct standard in first place. This motivates two future directions for this area of research: First, parameter efficient fine-tuning of musical foundation models instead of working with frozen embeddings is a very promising candidate for improving recognition. Second, further explicit analysis of the underlying melody and harmony with a given audio sample may further provide useful features towards better generalization of such recognition models.

Limitations. Our methodology and evaluation use local audio windows rather than a full symbolic or harmonic representation of a performance. As a result it only has a limited access to long-range musical context, which is important as many standards share similar local windows but possess a characteristic global flow that is important for identification. Conversely, many standards that share similar progressions (e.g., a 12-bar blues) are identifiable mainly through small but distinctive excerpts which can be rare relative to the duration of an entire performance. In this context, models might benefit from longer-context modeling, melody extraction and harmony identification for better results. Lastly, our curated dataset is intentionally small and controlled, so the results should be interpreted as an exploratory benchmark rather than a definitive ranking of music foundation models.

8. Conclusion

In this study, we present an exploratory study of jazz standard recognition across performances using a curated subset of the Jazz Trio Database. Our results indicate that a from-scratch HCNN network overfits to training performances, while probing based on embeddings from audio encoding models such as MERT and MuQ provide stronger results due

to their pretraining. Our retrieval analysis suggests that the group identity influences embedding-based similarity and retrieval. Ultimately, our results indicate that jazz standard recognition is a realistic but difficult setting for evaluating whether music representations encode tune identity rather than only acoustic or performance similarity.

References

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 591–596, 2011.

Castellon, R., Donahue, C., and Liang, P. Codified audio language modeling learns useful representations for music information retrieval. *arXiv preprint arXiv:2107.05677*, 2021. URL https://arxiv.org/abs/2107.05677.

Cheston, H., Schlichting, J. L., Cross, I., and Harrison, P. M. C. Jazz trio database: Automated annotation of jazz piano trio recordings processed using audio source separation. *Transactions of the International Society for Music Information Retrieval*, 7(1):144–158, 2024. doi: 10.5334/tismir.186. URL https://transactions.ismir.net/articles/10.5334/tismir.186.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R. B., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., and Fu, J. MERT: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107v5*, 2024. URL https://arxiv.org/abs/2306.00107v5.

McCallum, M. C., Korzeniowski, F., Oramas, S., Gouyon, F., and Ehmann, A. F. Supervised and unsupervised learning of audio representations for music understanding. *arXiv preprint arXiv:2210.03799*, 2022. URL https://arxiv.org/abs/2210.03799.

Papaioannou, C., Benetos, E., and Potamianos, A. Universal music representations? evaluating foundation models on world music corpora. *arXiv preprint arXiv:2506.17055*, 2025. URL https://arxiv.org/abs/2506.17055.

Serrà, J., Zanin, M., Herrera, P., and Serra, X. Characterization and exploitation of community structure in cover song networks. *arXiv preprint arXiv:1108.6003*, 2011. URL https://arxiv.org/abs/1108.6003.

Wang, A. L.-C. An industrial-strength audio search algorithm. In *Proceedings of the 4th International Society for Music Information Retrieval Conference*, pp. 7–13, 2003. URL https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf.

Won, M., Chun, S., Nieto, O., and Serra, X. Data-driven harmonic filters for audio representation learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 536–540, 2020.

Xun, J., Zhang, S., Yang, Y., Zhu, J., Deng, L., Zhao, Z., Dong, Z., Li, R., Zhang, L., and Wu, F. DisCover: Disentangled music representation learning for cover song identification. *arXiv preprint arXiv:2307.09775*, 2023. URL https://arxiv.org/abs/2307.09775.

Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., and Chen, X. MuQ: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025. URL https://arxiv.org/abs/2501.01108.