Evaluating CLAP and MERT for Fine-Grained Cymbal Classification: A Multi-Stage Representation Analysis

Michael Starakis *, Maximos Kaliakatsos-Papakostas, Chrisoula Alexandraki *

Posted Date: 24 March 2026

doi: 10.20944/preprints202603.1837.v1

Keywords: audio foundation models; CLAP; MERT; cymbal classification; audio embeddings; dimensionality reduction; kNN probing; clustering alignment; confound analysis; leakage-safe evaluation

 doi:10.20944/preprints202603.1837.v1

Article

Evaluating CLAP and MERT for Fine-Grained Cymbal Classification: A Multi-Stage Representation Analysis

Michael Starakis *, Maximos Kaliakatsos-Papakostas and Chrisoula Alexandraki *

Department of Music Technology and Acoustics, Hellenic Mediterranean University, 74133 Rethymno, Greece
* Correspondence: ddk216@edu.hmu.gr (M.S.), chrisoula@hmu.gr (C.A.)

Abstract

This study presents a representation-centric evaluation of audio foundation models for fine-grained musical instrument analysis, focusing on cymbal classification. A confound-aware comparison of CLAP and MERT embeddings is conducted to examine how each latent space supports recoverability of acoustically and semantically relevant information. To support this analysis, a five-stage evaluation framework is introduced, comprising geometric inspection, leakage-safe probing, and unsupervised clustering, complemented by confound diagnostics formulated as a horizontal task across all stages. The methodology is applied to a challenging cymbal dataset characterized by hierarchical labels, class imbalance, and subtle acoustic variation. Results reveal a target-dependent profile of representational strengths rather than a single overall winner. CLAP exhibits stronger variance concentration and more label-consistent local neighborhood organization, and it clearly outperforms MERT on fine-grained, strike-related targets. MERT, however, retains a small but consistent advantage on higher-level cymbal-type classification. Unsupervised analyses show that these advantages reflect local neighborhood structure rather than strong global cluster formation, and confound diagnostics indicate that size-related information remains largely type-mediated. Overall, the findings underscore the importance of structured, multi-stage evaluation for disentangling embedding geometry, recoverability and confound effects, while demonstrating the complementary strengths of AFMs in complex audio classification settings.

Keywords: audio foundation models; CLAP; MERT; cymbal classification; audio embeddings; dimensionality reduction; kNN probing; clustering alignment; confound analysis; leakage-safe evaluation

1. Introduction

Audio Foundation Models (AFMs) have shifted research emphasis from training narrowly specialized networks toward evaluating pre-trained representations under frozen conditions [1–3]. In this setting, embedding quality cannot be reduced to downstream classifier performance alone, but is instead tied to latent space geometry, stability of local and global structure, recoverability of labels under group-aware generalization, and sensitivity to confounding factors. Consequently, contemporary evaluation of AFMs requires a representation-centric perspective rather than an accuracy-only focus [1–4].

Despite growing literature on pre-trained audio and music models, many comparative studies focus on broad domains such as speech [2], music information retrieval and generative music tasks [3], or musical instrument classification [5,6]. Fine-grained domains concerned with subtle timbral variation remain largely unexplored, even though they provide particularly demanding testbeds for different pretraining paradigms [3]. Cymbal sounds constitute a suitable edge-case domain, combining strong attacks, metallic texture, high inharmonicity, and prolonged decay [7–9]. Encoding such signals requires preserving both micro-acoustic cues (e.g., transient onsets, high-frequency

 doi:10.20944/preprints202603.1837.v1

spectral texture) and macro-acoustic characteristics (e.g., timbral identity, decay structure). Perceptual attributes such as brightness, wash, dryness, and attack guide musicians' choices during performance and instrument selection, making the cymbal domain a bridge between perceptual experience, instrument properties, and the measurable behavior of AFMs [3,6].

This study investigates whether embeddings produced by CLAP and MERT under frozen conditions organize cymbal-related acoustic information in a useful, generalizable, and interpretable manner. Beyond downstream performance, the focus is on whether the latent spaces exhibit coherent structure and support reliable encoding of physical and perceptual sound attributes [3,10,11]. A representation-centric, multi-stage evaluation framework is employed, combining geometric visualization, leakage-safe kNN probing, and unsupervised clustering. These complementary analyses assess exploratory geometry, predictive label recoverability under group-aware generalization, intrinsic categorical structure, and robustness to leakage as well as shortcut learning [4,5].

The evaluation uses a sound dataset comprising 2,800 cymbal hits from 101 instruments. It is highly imbalanced, group-structured and partially confounded; evaluation is therefore group-aware and target-hierarchical. Cymbal family type (i.e. crash, splash, etc.) serves as the primary target, while striking related attributes are used as fine-grained secondary descriptors. Specifically, striking zone and stick material are treated as main secondary targets, whereas strike type and precise striking point on cymbal surface are considered auxiliary descriptors. The diameter of the cymbal is treated as a diagnostic, confound-sensitive target because of its structural dependency with the family type. The objective is to compare CLAP and MERT as frozen embedding extractors using converging evidence on latent space geometry, separability, intrinsic organization, and confound robustness in this fine-grained percussive domain.

The study makes three main contributions. First, it presents a fine-grained, confound-aware comparison of CLAP and MERT embeddings, examining how each latent space encodes acoustic and semantic structure. Second, it introduces a structured five-stage evaluation framework, within which geometric inspection, leakage-safe probing, and unsupervised clustering form the core analytical stages, while confound diagnostics are formulated as a horizontal task spanning the full pipeline, thereby enabling converging evidence beyond single-metric evaluation. Third, it demonstrates the applicability of this framework to a challenging musical instrument dataset, providing insights into embedding behavior under conditions of hierarchical targets, label imbalance, and subtle acoustic variation. Collectively, these contributions support both the applied task of cymbal classification and the broader methodological study of AFM evaluation.

The remainder of this article is structured as follows. Section 2 reviews related work, outlining the rationale for comparing self-supervised and contrastive approaches, specifically CLAP and MERT, for cymbal classification, and discussing representative efforts in benchmarking audio foundation models (AFMs) on common downstream music tasks. Section 3 introduces the cymbal dataset used to evaluate the frozen embeddings and defines leakage-safe split design and a target variable hierarchy to account for confounding factors. Section 4 describes the five-stage methodology, designed to assess geometric structure, quantitative recoverability on unseen cymbals, and the extent to which the embedding spaces exhibit intrinsic category-aligned organization versus more diffuse, manifold-like structure. As the first two stages are preparatory, Section 5 presents the results of the final three stages. Finally, Section 6 summarizes the main findings and their implications for tracing perceptual sound qualities, addresses constraints and critical considerations, and outlines directions for future work.

2. Related Work

Recent work on Audio Foundation Models (AFMs) has been shaped primarily by two pretraining paradigms: Self-Supervised Learning (SSL) and contrastive audio–text alignment [3,10,11]. SSL models such as MERT learn representations by predicting masked or teacher-derived acoustic targets, often combining multi-task setups with inductive biases related to pitch and

 doi:10.20944/preprints202603.1837.v1

harmonic structure [11–13]. As a result, SSL embeddings often tend to preserve music-acoustic regularities and continuous variations suitable for fine-grained distinctions. On the other hand, contrastive models such as CLAP align audio and text embeddings via InfoNCE objectives, thereby tending to favor semantically oriented macro-level distinctions that support zero-shot classification and broad categorization, while in some cases potentially reducing sensitivity to subtle acoustic nuances [10,14,15].

These differences suggest that SSL and contrastive embeddings may organize latent spaces differently, with the balance between semantic alignment and fine-grained acoustic variation depending on the pretraining objective, architecture, and evaluation setting. This makes comparative evaluation in fine-grained, transient-sensitive domains, such as cymbal audio, particularly informative. Previous studies frame this as a trade-off between general acoustic fidelity and semantic alignment rather than as a winner-takes-all comparison [3,10,11].

Alongside model development, standardized frameworks for evaluating frozen embeddings have become influential. HEAR evaluates frozen encoders using lightweight downstream predictors across diverse tasks, establishing reproducible protocols for cross-model comparison [4]. MARBLE adapts this to Music Information Retrieval (MIR), with 18 tasks and 12 datasets, constrained frozen-backbone evaluation, and restricted hyperparameter search to reduce variance [5]. ARCH provides a unified protocol across speech, music, and acoustic events, reinforcing the value of cross-domain evaluation [16]. Dimensionality reduction and embedding-space visualization are widely used as exploratory tools rather than as primary evidence [4,5,17].

Furthermore, recent work on large audio-language models, such as MMAU, emphasizes reasoning and multimodal understanding [17]. While relevant for advanced audio reasoning, these approaches are less directly applicable here, as the focus is on embedding quality and structure rather than multimodal reasoning.

Taken together, prior work highlights two key points: first, SSL and contrastive AFMs differ in the type of information preserved and latent-space organization [10,11,14,15]; second, benchmark-inspired frozen evaluation frameworks provide reproducible, task-sensitive comparison logic [4,5,17]. Existing benchmarks, however, remain broad and coarse-grained, lacking a hierarchically structured framework combining geometric inspection, leakage-safe probing, unsupervised clustering alignment, and confound diagnostics. The present work addresses this gap by adopting a benchmark-inspired frozen-evaluation philosophy specialized into a fine-grained, cymbal-centered, multi-stage diagnostic framework designed to assess performance, latent-space structure, generalization, and interpretive validity.

3. Dataset Design and Experimental Considerations

The present study is based on a curated dataset of cymbal audio recordings, which is used as a representation-level test bed for the comparative evaluation of CLAP and MERT. The sounds have been derived from the commercial Studio Cymbals sound pack by Audio Animals Ltd. (https://www.audioanimals.co.uk/shop/sample-shop/bundles/studio-cymbals - 2025 version) and were organized into a structured metadata format, so that each recording is linked to physical and performance-related properties of the instrument.

The methodological value of the dataset lies not only in its size, but primarily in the combination of four characteristics: rich tagging, multiple hits per physical instrument, visible class imbalance, and the presence of potentially confounded label relationships. This internal structure makes the dataset particularly suitable not merely for straightforward classification, but for a more controlled assessment of what kinds of acoustic information may be reflected in the embeddings, how robustly they generalize to unseen instruments, and to what extent the observed results remain valid under leakage-aware and shortcut-aware evaluation. In this sense, the dataset serves as a useful benchmark-inspired environment for the comparative study of CLAP and MERT.

3.1. Dataset Description: Labels and Statistics

 doi:10.20944/preprints202603.1837.v1

The dataset comprises 2,800 unique audio recordings, each representing an individual cymbal strike. Each recording is accompanied by metadata describing the physical identity of the instrument, including *manufacturer*, *model_name*, *cymbal_type*, *diameter*, *cymbal_id*, as well as performance-related attributes, namely *hit_point*, *hit_zone*, *stick_type*, and *stick_material*. In addition, technical metadata, including *filesize* and *duration*, are provided for each audio sample.

In terms of diversity, the dataset covers 101 unique physical cymbals (*cymbal_id*), 18 manufacturers, 48 model names, 5 selected cymbal types (crash, splash, hi-hat, ride, china), 13 diameter values ranging from 6″ to 21″, 7 hit points on 3 hit zones (bow, bell, edge), 2 striker types (mallet, stick), and 3 striker materials (wood, nylon, plastic). Figure 1 illustrates hit zones and hit points on the left as well as the four strikers corresponding to two types and three materials on the right. Pictures have been derived from the Audio Animals Ltd. page describing the original studio cymbals sound pack. This composition combines categorical diversity, physical variation, and performance-related variability, thereby allowing embedding evaluation across multiple potentially relevant levels of acoustic information.

Figure 1. The different hit points and strikers for the cymbal sounds used in the dataset. (**a**) Seven hit points distributed on three hit zones (bow, bell, edge). (**b**) The four strikers used for cymbal excitation in the dataset. From left to right: nylon mallet, wood mallet, plastic stick, and wood stick.

The main labels considered in the present study are *cymbal_type*, *hit_zone*, *stick_material*, and *diameter*. The label *cymbal_type* serves as the primary target, reflecting macro-acoustic categorization, whereas *hit_zone* and *stick_material* are treated as finer-grained secondary targets related to strike location and excitation characteristics. The *diameter* label is included as a diagnostic target linked to a physical property of the instrument and is interpreted more cautiously in later analysis because of its structural relation to *cymbal_type*.

The composition of the dataset is summarized in Table 1. Class distributions are imbalanced. For the *cymbal_type*, the audio samples are distributed as: 1176 crash, 588 splash, 504 hi-hat, 420 ride, and 112 china. For *hit_zone*, the bow category dominates with 2000 samples, whereas bell and edge contain 400 samples each. *Stick_type* is perfectly balanced (mallet: 1400, stick: 1400), whereas *stick_material* shows moderate imbalance (wood: 1400, nylon: 700, plastic: 700). This distribution justifies the use of macro-oriented metrics, such as Macro-F1, and, secondarily, Balanced Accuracy, so that evaluation is not disproportionately driven by majority classes.

Table 1. Compact summary of the composition of the dataset used for evaluation.

Descriptor	Value
Source	Studio Cymbals (Audio Animals)
Total recordings	2800
Unique Physical Cymbals (cymbal_id)	101
Manufacturers / Models	18 / 48

 doi:10.20944/preprints202603.1837.v1

Cymbal Types	5 (crash, splash, hi-hat, ride, china)
Cymbal Diameter	13 distinct values (6″ – 21″)
Hit-point / Hit-zone labels	7 (5 bow area, 1 bell, 1 edge) / 3 (bow, bell, edge)
Stick-type / Stick-material labels	2 (mallet, stick) / 3 (wood, nylon, plastic)
Duration range	1.06 s – 28.58 s (mean: 7.17 s; median: 6.51 s)
Evaluation Implication	Group-aware splitting by *cymbal_id*

Another important property is the organization of the dataset around physical instrument identities. For the 101 *cymbal_id* values, multiple hits are available (from 7 to 28 per cymbal), which means that the dataset is not only label-structured but also group-structured. This makes it suitable for evaluating generalization to unseen physical cymbals under a group-aware protocol.

Finally, recording duration shows substantial variability, ranging from 1.06 s to 28.58 s (mean: 7.17 s; median: 6.51 s). This variability is particularly relevant in the cymbal domain, where informative content is not confined to the initial attack but extends into the decay tail. For this reason, the temporal heterogeneity of the raw recordings makes duration standardization during preprocessing necessary (refer to section 4.1), to ensure a fair comparison between the two model families.

3.2. Leakage Safe Grouping and Acoustic Confounds

A central methodological challenge of the present dataset is the risk of group leakage. Because each physical cymbal (*cymbal_id*) is represented by multiple hits recorded under different striking conditions, a conventional sample-level split could place recordings from the same instrument in both the training and the test sets. In such a case, performance estimates could be artificially inflated, as the model might exploit instrument-specific signatures rather than genuinely generalizable acoustic properties of the target variables. For this reason, the study adopts grouping by *cymbal_id*, so that evaluation reflects generalization to unseen physical instruments rather than recognition of repeated sources.

Beyond leakage, the dataset also contains label dependencies that may induce shortcut-like interpretations if all targets are treated as fully independent. We therefore performed a dataset-level dependency screening using bias-corrected Cramér's V, Normalized Mutual Information (NMI), and False Discovery Rate (FDR)-controlled significance testing. Cramér's V quantifies the strength of association between categorical variables, NMI measures the amount of shared information between label distributions, and FDR control adjusts p-values to limit false positives arising from multiple comparisons.

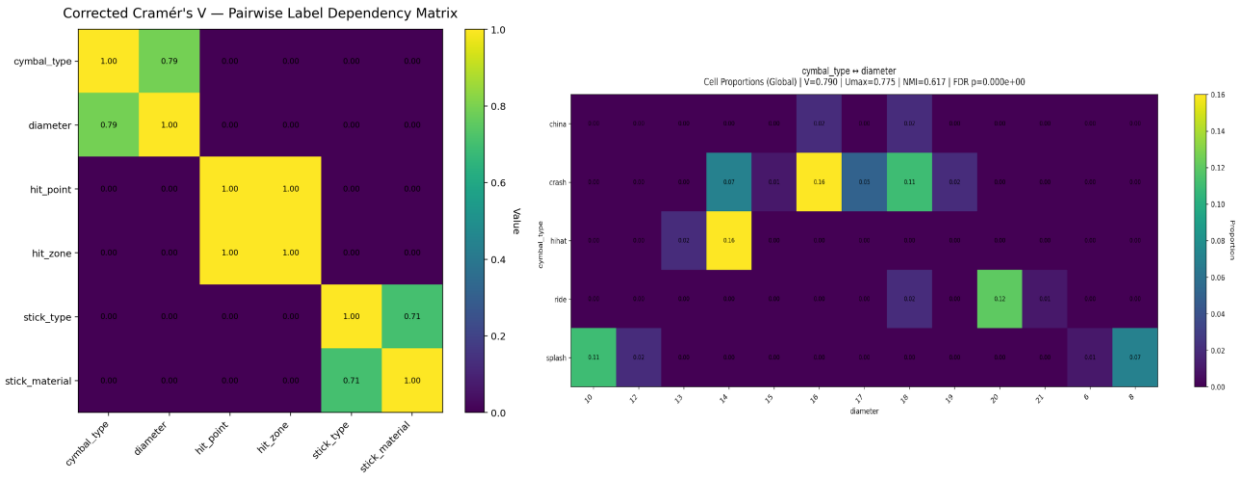
Table 2 shows the results of this screening for labels involved in potentially confounding relationships. Of the fifteen (15) tested label pairs, only three (3) remained significant after correction, whereas the remaining twelve (12) showed no substantial evidence of dependency. The strongest dependencies were concentrated in three clearly interpretable domains: strike-location hierarchy (*hit_point* vs. *hit_zone*), instrument structure (*cymbal_type* vs. *diameter*), and excitation-related metadata coupling (*stick_type* vs. *stick_material*). By contrast, representative pairs such as *hit_point* vs. *stick_type* and *diameter* vs. *stick_material* showed no evidence of dependency.

 doi:10.20944/preprints202603.1837.v1

Table 2. Pairwise label dependencies identified through dataset-level confound screening.

Rank	Label Pair	Corrected Cramér's V	NMI	FDR adjusted p	Relationship Type	Interpretive Role
1	*hit_point* / *hit_zone*	0.999	0.581	0.00e+00	hierarchical dependency coarse-vs-fine label hierarchy	coarse-vs-fine label hierarchy
2	*cymbal_type* / *diameter*	0.790	0.617	0.00e+00	hierarchical dependency main confound-sensitive case	main confound-sensitive case
3	*stick_type* / *stick_material*	0.707	0.400	4.93e-304	confound_like secondary metadata dependency	secondary metadata dependency
4	*hit_point* / *stick_type*	0.000	0.000	1.00e+00	no_evidence	no substantial dependency
5	*Diameter* / *stick_material*	0.000	0.000	1.00e+00	no_evidence	no substantial dependency

These dependencies are schematically illustrated in Figure 2. Figure 2a provides a global overview of the dependency structure, showing that strong label dependence is concentrated in only three pairs, while Figure 2b complements this overview by illustrating the *cymbal_type* vs. *diameter* relation, which constitutes the main confound-sensitive case in the present study. The plot shows that diameter is not freely distributed across cymbal categories. It instead follows a non-random and structured organological pattern.



(**a**) (**b**)

Figure 2. Pairwise dependency matrices for the main categorical labels. (**a**) Bias-corrected Cramér's V matrix showing strong dependencies for *hit_point* vs. *hit_zone*, *cymbal_type* vs. *diameter*, and *stick_type* vs. *stick_material*. (**b**) Global contingency plot for *cymbal_type* vs. *diameter*.

Accordingly, the *cymbal_type* vs. *diameter* relation is treated as the main target-relevant confound-sensitive case in subsequent evaluation stages. Although it was heuristically typed as a

 doi:10.20944/preprints202603.1837.v1

hierarchical dependency, it is methodologically the most consequential anti-shortcut example, because strong recoverability of diameter may partly reflect its structural dependence on type rather than an entirely independent encoding of size. Accordingly, *diameter* is treated throughout this study as a diagnostic and confound-sensitive target rather than as a primary endpoint equivalent to the main classification labels.

The other two significant relations play a different interpretive role. The *hit_point* vs. *hit_zone* pair is best interpreted as a coarse-to-fine label hierarchy rather than as a representative confound case, since *hit_point* encodes finer strike-location detail whereas *hit_zone* captures a broader regional abstraction. In contrast, *stick_type* vs. *stick_material* is treated as a secondary metadata dependency, indicating that apparent sensitivity to *stick_material* may partly reflect structurally coupled excitation-related metadata.

Accordingly, Stage 3 centers confound-aware visual control on the diameter-within-type setting as the most representative anti-shortcut analysis.

Overall, controlling both leakage and confounding is essential for an interpretable CLAP–MERT comparison. The leakage-safe split design ensures evaluation on unseen physical cymbals, whereas the dependency screening clarifies which targets can be interpreted more directly and which require hierarchy-aware or confound-sensitive claims. In this framework, the main methodological implication concerns diameter, while *hit_zone* and *stick_material* are interpreted more cautiously because of annotation hierarchy and secondary metadata coupling, respectively.

4. Methodological Framework

Building on the dataset design and the considerations discussed in the previous section, the present study adopts a controlled, multi-stage methodology for the comparative evaluation of CLAP and MERT. This approach is motivated by the nature of the data corpus: it is richly tagged, group-structured and imbalanced, making a single metric or a single type of evidence insufficient for robust evaluation. Furthermore, the cymbal domain, with its strong transients, metallic texture, and complex attack–decay relationships, provides a suitable case study for assessing AFMs in capturing fine-grained timbral characteristics.

The pipeline proceeds from the preparatory stages of audio standardization (Stage 1) and model selection (Stage 2) to three parallel complementary branches providing exploratory geometric inspection of the embedding space (Stage 3), main quantitative evaluation via leakage-safe kNN probing (Stage 4) and supporting unsupervised clustering alignment (Stage 5). This design separates visual indications from quantitatively supported conclusions, reducing the risk of over-interpretation and ensuring that the final claims are grounded in converging evidence across multiple stages. As each stage employs different workflows, design elements, and evaluation metrics, Annex A presents a schematic overview of the five-stage comparative evaluation framework, assisting the reader in navigating its multiple analytical perspectives.

4.1. Stage 1: Audio Pre-Processing

Each foundation model has different requirements: CLAP uses 10 second windows at rate of 48 kHz [10], while MERT uses 30 seconds windows at a sampling rate of 24 kHz [11]. To ensure a fair evaluation and prevent inconsistencies or artifacts arising from on-the-fly processing, all dataset samples were processed through a five-step pipeline:

1. Downmixing of stereo to mono signals;
2. Dual-branch resampling (48 kHz for CLAP, 24 kHz for MERT);
3. Conservative silence trimming;
4. Fixing duration at 10 s;
5. Peak normalization at −1 dBFS.

 doi:10.20944/preprints202603.1837.v1

The conversion of stereo signals to mono was performed by sample averaging. Resampling was performed separately for each model using the high-quality method for band-limited sinc interpolation used by the librosa python package [19].

Following resampling, conservative silence trimming was performed to expand detected non-silent intervals with asymmetric safety margins (50 ms pre-onset, 2.0 s post-offset). This choice was motivated by the fact that cymbal classification depends not only on attack information, but also on low-level tail energy, shimmer, and decay behaviour, all of which may remain acoustically informative even near the noise floor. To suppress boundary artifacts introduced by trimming and padding, a short 2 ms fade-in and a 10 ms fade-out were applied, allowing to preserve the sharpness of the initial transient while safely attenuating any possible boundary clicks.

All files were standardized to a fixed duration of 10 s. Shorter samples were zero-padded at the end, while longer samples were truncated with a 10 ms fade-out at the new endpoint. This fixed-window policy ensures that all inputs span the same temporal extent, reducing duration as a potential confound in representation quality. The adopted 10 s duration therefore represents a controlled compromise for comparability across models. The use of zero-padding, rather than repeat-padding, was again intentional, to avoid introducing artificial temporal structure that might distort the acoustic envelope of cymbal sounds.

After standardizing the duration, peak normalization (−1 dBFS) was chosen over loudness normalization, because cymbal recordings are strongly driven by transients, and LUFS-style normalization could distort the relationship between onset energy and low-level decay. Finally, the preprocessed samples were stored as 24-bit PCM WAV files.

Overall, Stage 1 established a controlled, branch-specific yet evaluation-consistent audio set, ensuring that subsequent analyses of embedding geometry, probing performance, clustering behavior, and confound structure would rely on inputs that are standardized, traceable, and explicitly designed for a fairness-aware comparison between CLAP and MERT.

4.2. Stage 2: Model Selection and Extraction of Embeddings

Before the main CLAP–MERT comparison, a systematic model-selection stage was conducted to ensure evaluation of the most suitable configurations. Multiple CLAP checkpoints were tested, while MERT used a fixed backbone with alternative embedding readout schemes. This procedure ensured that the final comparison reflects best-performing configurations rather than incidental implementation choices.

Thus, this stage aims to carry out internal model selection within each model family, identifying the best-versus-best configuration and yielding two locked embedding spaces—one for CLAP and one for MERT—for comparison in the subsequent stages. Here, "locked" denotes that, after model selection, the embedding configuration for each model family is fixed and reused unchanged throughout all downstream analyses. The workflow of Stage 2 comprises the following steps:

1. Load Stage 1 branch-specific audio dataset.
2. Construct leakage-safe grouped folds (SGFK / grouping by *cymbal_id*).
3. Sweep CLAP checkpoints and MERT readout variants.
   - CLAP: 4 checkpoints
   - MERT: 1 backbone × 5 layer/readout sources × 3 pooling strategies
4. Extract frozen candidate embeddings.
5. Evaluate with cosine kNN probing (k = 5 / StandardScaler fit on TRAIN only).
6. Aggregate fold-wise metrics and rank candidates
   - Primary: Macro-F1
   - Tiebreakers: Balanced Accuracy, Accuracy
7. Lock final CLAP and MERT winners.

 doi:10.20944/preprints202603.1837.v1

To choose the best model configuration, the curated dataset of Stage 1 was split into training/test sets via *StratifiedGroupKFold*, using *cymbal_id* as the grouping variable, so that all hits originating from the same physical cymbal remained exclusively in either the training or the test set. At the same time, stratification aimed to preserve, as far as possible, the label distribution of the remaining labels: *cymbal_type*, *diameter*, *hit_zone*, *hit_point*, *stick_material*, and *stick_type* across folds. Test samples were evaluated using a k-nearest neighbors (kNN) probe with k = 5 and cosine distance applied to embeddings extracted from the training set. The *StandardScaler* was fitted only on the training fold and applied to the corresponding test fold prior to distance calculation, thereby ensuring feature-space normalization without data leakage.

For both models, the primary evaluation metrics were **Accuracy**, **Balanced Accuracy**, and **Macro-F1** per target variable. Candidate configurations were ranked using an aggregate criterion defined as the unweighted mean of Macro-F1 across targets. Secondary tiebreakers were the mean Balanced Accuracy and the mean Accuracy, ensuring stable selection when configurations exhibited similar primary performance.

For CLAP, a checkpoint sweep was performed over four candidate variants: *laion/clap-htsat-unfused*, *laion/clap-htsat-fused*, *laion/larger_clap_general*, and *laion/larger_clap_music*. Model checkpoints were obtained from the LAION CLAP repository (https://huggingface.co/laion). For each checkpoint, one 512-dimensional embedding was extracted per sample in frozen inference mode, using common extraction settings across all targets, and the leakage-safe probing protocol was subsequently applied. Table 2 reports the performance of the candidate checkpoints, in the form of per-label values as fold-averaged scores under *5-fold StratifiedGroupKFold* with grouping by *cymbal_id*. Candidate ranking follows aggregate mean Macro-F1 as the primary criterion, with mean Balanced Accuracy and mean Accuracy used as secondary tiebreakers. Under this criterion, laion/clap-htsat-fused was selected as the final CLAP configuration.

The evaluation of the CLAP checkpoints is reported in Table 3. Within the CLAP candidates, *laion/clap-htsat-fused* achieved the strongest aggregate profile and was selected as the final CLAP configuration. It was followed by *laion/larger_clap_general* and *laion/clap-htsat-unfused*, whereas *laion/larger_clap_music* performed at a clearly lower level across target variables. This pattern suggests that, for isolated cymbal signals, more general-purpose or sound-event-oriented CLAP checkpoints are better suited than a more music-specialized alternative. At the same time, the margin among the top three CLAP candidates was not large, indicating that the CLAP family was overall strong on the present task, with *clap-htsat-fused* offering the most balanced aggregate profile under the predefined ranking criterion. The substantially lower performance of *larger_clap_music* further suggests that isolated cymbal hits are less well aligned with embeddings designed to support broader audio-event and audio-language distinctions more effectively [10,18].

Table 3. The performance of CLAP candidate checkpoints across target variables.

CLAP Model laion/	*cymbal_type* Macro-F1	*diameter* Macro-F1	*hit_zone* Macro-F1	*hit_point* Macro-F1	*stick_material* Macro-F1	*stick_type* Macro-F1	Mean Macro-F1	Mean BalAcc	Mean Acc
clap-htsat-fused	**0.5480 ± 0.034**	0.2051 ± 0.026	**0.7368 ± 0.024**	**0.4080 ± 0.031**	0.7015 ± 0.024	**0.8251 ± 0.010**	**0.5707**	**0.5896**	**0.6250**
larger_clap_general	0.5295 ± 0.024	**0.2212 ± 0.018**	0.7098 ± 0.046	0.3777 ± 0.022	**0.7098 ± 0.028**	0.8205 ± 0.018	0.5614	0.5823	0.6242
clap-htsat-unfused	0.5433 ± 0.034	0.2135 ± 0.018	0.7184 ± 0.032	0.3794 ± 0.027	0.6912 ± 0.027	0.8096 ± 0.022	0.5592	0.5807	0.6188

 doi:10.20944/preprints202603.1837.v1

larger_clap_music	0.4275 ± 0.075	0.1733 ± 0.038	0.4606 ± 0.035	0.2058 ± 0.008	0.4949 ± 0.019	0.6471 ± 0.013	0.4015	0.4284	0.480 5

For MERT, the backbone was fixed to *m-a-p/MERT-v1-330M* (https://huggingface.co/m-a-p/MERT-v1-330M), while the sweep was defined over a unified set of candidate readout configurations. Specifically, frame-level representations were extracted from multiple layer sources—*last, layer_23, mean_last4* and *layer_22*—and evaluated under three temporal aggregation (pooling) strategies: *mean, max, and mean_std*. The *m-a-p/MERT-v1-330M* backbone comprises 24 layers and 1024 feature dimensions. Therefore, *last* refers to the 24th layer, *layer_23* to the 23rd layer, *mean_last4* refers to a layer produces by the last four layers (i.e. 21st – 24th) and *layer_22* refers to the 22nd layer. These four redouts combined with the three pooling strategies resulted in twelve configurations.

Similarly to CLAP, each configuration represented every audio sample as a 1024-dimensional embedding obtained in frozen inference mode. Again, performance was evaluated using 5-fold *StratifiedGroupKFold* cross-validation, with grouping by *cymbal_id* to prevent leakage across cymbals and to obtain label-wise metrics. This design enabled a controlled comparison of temporal pooling strategies, representational depth, and layer aggregation on the separability of frozen MERT embeddings, while keeping the pretrained backbone fixed.

Table 4 reports on the evaluation of the four best performing MERT configurations. Within the MERT family, the best-performing configuration was not the baseline last-layer readout, but the layer-aware configuration *mean_last4+max*. The second-ranked configuration, *layer22+max*, was very close, followed by *layer23+max* and *last+max*. This result is methodologically important because it shows that representational depth and layer aggregation materially affect the quality of frozen MERT embeddings for cymbal-related tasks. Furthermore, the fact that all the best performing MERT configurations relied on max pooling may indicate that max-based temporal aggregation is advantageous for transient-heavy cymbal signals, plausibly because it preserves strong local activations associated with attack-related information more effectively than averaging-based schemes. At the same time, the small margin between *mean_last4+max* and *layer22+max* indicates that the winning MERT configuration should be interpreted as the best-performing option under the predefined criterion, rather than as a uniquely dominant solution [11].

Table 4. The performance of the four best performing configurations of MERT across target variables.

MERT Model	*cymbal_type* Macro-F1	*diameter* Macro-F1	*hit_zone* Macro-F1	*hit_point* Macro-F1	*stick_material* Macro-F1	*stick_type* Macro-F1	Mean Macro-F1	Mean BalAcc	Mean Acc
mean_last4+max	**0.5668 ± 0.087**	**0.2179 ± 0.035**	0.6227 ± 0.040	0.3018 ± 0.025	**0.5509 ± 0.015**	0.6965 ± 0.014	**0.4927**	**0.5163**	0.552 2
layer22+max	0.5602 ± 0.081	0.2162 ± 0.036	**0.6239 ± 0.052**	**0.3053 ± 0.025**	0.5348 ± 0.013	**0.7009 ± 0.013**	0.4902	0.5150	**0.552 3**
layer23+max	0.5185 ± 0.078	0.2079 ± 0.019	0.6056 ± 0.043	0.3024 ± 0.022	0.5166 ± 0.023	0.6716 ± 0.022	0.4704	0.4969	0.539 0
last+max	0.4908 ± 0.082	0.2067 ± 0.012	0.6017 ± 0.059	0.3013 ± 0.026	0.5249 ± 0.024	0.6513 ± 0.015	0.4627	0.4881	0.537 6

Table 5 reports on the target-wise comparison of the best performing candidates of both CLAP and MERT. When comparing between the two final winners, laion/clap-htsat-fused and

 doi:10.20944/preprints202603.1837.v1

mean_last4+max, CLAP remains the overall winner according to the primary ranking criterion, with higher aggregate mean Macro-F1 and mean Balanced Accuracy. At the label level, CLAP outperformed MERT on *hit_zone*, *hit_point*, *stick_material*, and *stick_type*, whereas MERT achieved slightly higher mean performance on *cymbal_type* and *diameter*. This pattern is methodologically and substantively informative, suggesting that the two models differ not only in overall separability, but also in the kind of information their latent spaces make more readily recoverable [3].

CLAP's superior performance on *hit_zone*, *hit_point*, *stick_material*, and *stick_type* suggests that its contrastive latent space is particularly effective for strike-dependent and timbre-sensitive discrepancies, i.e. where and how the cymbal is struck, and with which implement. MERT's slight advantage on cymbal type and diameter may indicate better preservation of broader spectral or category-level information in the selected readout. However, these differences should be interpreted cautiously, given the small effect size of diameter and the confound-sensitive structural relationships between diameter, family type, and other dimensions. Taken together, Tables 3–5 show that model selection materially shaped the final CLAP–MERT comparison: *laion/clap-htsat-fused* emerged as the strongest CLAP configuration, whereas the best MERT solution was the layer-aware *mean_last4+max* rather than the default last-layer baseline. At the same time, Stage 2 should be interpreted as an internal model-selection step rather than as an independent confirmatory layer.

Table 5. Label-wise comparison of the best performing configurations of CLAP and MERT.

Target	laion/clap-htsat-fused		mean_last4+max		Best Model
	Macro-F1	Balanced Accuracy	Macro-F1	Balanced Accuracy	
cymbal_type	0.5480 ± 0.034	0.5663 ± 0.045	**0.5668 ± 0.087**	**0.5660 ± 0.072**	MERT
diameter	0.2051 ± 0.026	0.2976 ± 0.025	**0.2179 ± 0.035**	**0.3136 ± 0.036**	MERT
hit_zone	**0.7368 ± 0.024**	**0.7228 ± 0.019**	0.6227 ± 0.040	0.6380 ± 0.045	CLAP
hit_point	**0.4080 ± 0.031**	**0.4188 ± 0.027**	0.3018 ± 0.025	0.3260 ± 0.024	CLAP
stick_material	**0.7015 ± 0.024**	**0.7070 ± 0.024**	0.5509 ± 0.015	0.5575 ± 0.018	CLAP
stick_type	**0.8251 ± 0.010**	**0.8253 ± 0.010**	0.6965 ± 0.014	0.6967 ± 0.013	CLAP
Mean	**0.57074**	**0.58964**	**0.49275**	**0.51631**	**CLAP**

For the selected configurations the locked embedding spaces (512-dimentions for CLAP and 1024-dimensions for MERT) were exported in both raw and L2-normalized form for downstream geometric and statistical analyses. Overall, Stage 2 ensured that the CLAP–MERT comparison was grounded in systematically selected, audit-ready configurations rather than arbitrary model variants, thereby establishing a controlled and reproducible basis for the subsequent stages of the present methodology.

4.3. Stage 3: Dimensionality Reduction and Geometric Inspection

The aim of this stage may be articulated as geometric inspection of the locked embedding spaces, and its output as a few visualization plots combined with quantitative metrics that can serve for hypothesis generation to be tested in Stage 4. Concisely, the workflow comprises the following steps:

1. Load Stage 2 winners.
2. Lock embeddings and metadata.
3. Choose preprocessing mode
   - P0 raw / P1 L2 / P2 StandardScaler / P3 StandardScaler → L2
4. Define target and evidential role

 doi:10.20944/preprints202603.1837.v1

- core / supporting / exploratory / diagnostic

5. Create leakage-aware group splits
   - GDSS (canonical visuals by cymbal_id) / SGKF (diagnostic supervised views)
6. Fit reducers on TRAIN.
   - e.g., PCA, UMAP, MDS, PaCMAP, t-SNE, LDA, supervised UMAP, as appropriate to the evidential role
7. Project TEST in shared space.
8. Produce evidence-ranked visuals.
   - core: PCA-2D, UMAP-2D (unsupervised)
   - supporting: MDS, PaCMAP, UMAP sensitivity panels
   - exploratory: t-SNE, full-set descriptive DR
   - diagnostic: LDA, supervised UMAP
9. Compute projection QC metrics.
   - explained_variance_2d for PCA, pca90_components, pca90_retained_variance, train_trustworthiness, test_trustworthiness, raw_stress, Stress-1 for MDS, trustworthiness for other reducers Compute original-space probes.
   - P@K, R@K, mAP@K; group-aware and same-cymbal-excluded evaluation
10. Run diagnostic confound control
    - Diameter-within-Type, UMAP small multiples, confound-aware local-structure visualization
11. Export figures, audits, and registries.

Specifically, a methodological protocol was established to govern both the derivation as well as the interpretation of DR projections. Visualizations are organized into four rankings depending on the type of evidence they provide: core (2), supporting (4), exploratory (2), and diagnostic (2) evidence. Thus, ten (10) visualizations were produced for each of the six (6) target variables. Prior to deriving visualizations, preprocessing and leakage-aware split was performed. Table 6 provides a comprehensive summary of the protocol used per evidence-ranking. It details the DR algorithms employed, the corresponding preprocessing modes, the data splitting and transformation strategies used to prevent label leakage, and the metrics used to assess the geometric integrity of each projection.

Table 6. A summary of the protocols applied per evidence-ranking.

Evidence Ranking	DR / Visualization Technique	Preprocessing	Split	Eval. Metrics	Selection Rationale
Core	PCA-2D (unsupervised)	P3	GSS (TR-fit / TE-transf.)	explained_variance_2d, pca90_components, pca90_retained_variance, trustworthiness	Linear baseline for global structure
	UMAP-2D (unsupervised)			train_trustworthiness, test_trustworthiness	Main non-linear view of local structure
Supporting	MDS (unsupervised)	P3 +PCA-50;	GSS (TR-fit / TE-OOS)	raw stress, normalized Stress-1 trustworthiness	Robustness checks across alternative reducers

 doi:10.20944/preprints202603.1837.v1

	PaCMAP (unsupervised)			trustworthiness	
	UMAP sensitivity panel (unsupervised)	P3	GSS (TR-fit / TE-transf.)	train_trustworthiness, test_trustworthiness	Parameter stability check
	Diameter-within-type UMAP small multiples (unsupervised)			trustworthiness; supportive reading with within-type retrieval	Confound-aware control for diameter
Exploratory	t-SNE sweeps (unsupervised)	P3 + PCA-50;	Full set (no split)	trustworthiness	Exploratory inspection of manifold structure
	UMAP/PCA (unsupervised)	P3		descriptive reading	
Diagnostic	LDA (supervised)	P2	SGKF (TR-fit / TE-transf.)	test_trustworthiness	Diagnostics reveal label-related structure
	UMAP (supervised)	P2+P3			

Core-evidence visualizations carry the main geometric narrative of the analysis. For this evidence, PCA-2D was chosen to examine whether the representation contains a strong linear, or more broadly global signal, while UMAP-2D was used as a complementary visualization aiming at reading local neighborhood structure, and examining whether the embedding space organizes local neighborhoods with semantic or fine-grained acoustic coherence.

S**upporting visualizations** serve as robustness checks. As such, Muti-Dimensional Scaling (MDS) is used as a supporting projection for examining the global geometry of the embedding space, while PaCMAP is used as a reducer-robustness view, allowing to test whether a pattern observed in UMAP (core evidence) also persists under a different non-linear reducer. Agreement between UMAP and PaCMAP strengthens confidence in the geometric reading, while disagreement points to possible reducer sensitivity. Also, given the structural dependence between *cymbal_type* and *diameter*, supporting visualizations also include a diameter-within-type visual control as its main confound-aware inspection, allowing size-related structure to be examined under an anti-shortcut constraint.

Exploratory visualizations are intended for hypothesis-generating. They are based on t-SNE sweeps to explore local micro-structures. In parallel, descriptive UMAP and PCA projections on the full dataset are produced to provide the big-picture manifold inspection, visual reading of density structure, and identification of possible bridges, overlaps, or smaller-scale structures that may not be immediately visible in the train/test projections.

Finally, **diagnostic** evidence serves the purpose of examining whether a linear supervised mapping can reveal separation that is not visible in the unsupervised geometry. Diagnostic evidence facilitates Linear Discriminant Analysis (LDA) as a linear supervised overview of the embedding space. A weak unsupervised structure combined with a more coherent LDA organization indicates that the relevant information exists in the embedding space but is not naturally aligned to emerge without supervision. In parallel, supervised UMAP functions as a nonlinear diagnostic method while adhering to the same fit-on-train / transform-on-test protocol.

Preprocessing is standardized across four modes: P0 (raw embeddings), P1 (L2-normalized embeddings), P2 (*StandardScaler*), and P3 (*StandardScaler* + *L2-normalized* embeddings).

 doi:10.20944/preprints202603.1837.v1

Standardization reduces the dominance of features with larger scale, whereas L2 normalization aligns the geometry of the space with cosine-based neighborhood logic. This ensures that observed differences reflect embedding quality rather that preprocessing inconsistencies. Core, supporting, and exploratory visuals use P3, while diagnostic visuals use P2. However, in some cases, a prior reduction of the embedding space to 50 principal components was required before the main visualization step and subsequently to preprocessing. This is denoted as PCA-50 on Table 6, and was deemed necessary for stabilizing and denoising pre-processing step for selected reducers.

Leakage control is ensured through group-aware splits by *cymbal_id*. Core and supporting projections use a canonical *GroupShuffleSplit* (GSS), while supervised diagnostics rely on *StratifiedGroupKFold* (SGKF). TR-fit / TE-transf. denotes projection of TEST samples through the fitted projection model itself. TR-fit / TE-OOS indicates out-of-sample placement of TEST samples into the TRAIN-derived low-dimensional space, typically via interpolation rather than a native transform function. Full set indicates descriptive projections produced on the complete dataset without a train/test split.

Beyond visual inspection, Stage 3 also supports by projection-quality metrics and retrieval-style geometry probes in the original embedding space, including Precision@K, Recall@K, and mAP@K under leakage-aware query–gallery separation. These probes provide quantitative grounding for local neighborhood patterns suggested by the
DR visualizations and serve as a methodological bridge to the leakage-safe probing
of Stage 4.

4.4. Stage 4: Leakage-Safe kNN Probing

This stage aims to encode predictive recoverability under group-aware generalization to unseen physical cymbals, providing a formal assessment of whether the geometric structure identified in Stage 3 translates into stable recoverability in the original high-dimensional space. The workflow of this stage can be summarized as:

1. Load Stage 2 winners.
2. Select one embedding view per run.
3. Define target variables.
4. Apply group-aware split by *cymbal_id* (SGKF / GKF fallback).
5. Check fold admissibility.
6. Fit scaler on TRAIN only.
7. Run cosine kNN for K-sweep (K = {1, 3, 5, 11, 21}).
8. Compute fold-wise metrics
   - Primary: Macro-F1, Balanced Accuracy
   - Secondary: MCC, Accuracy
9. Aggregate summary statistics (mean, SD, and CI95) and diagnostic outputs.
10. Select Best-K (Macro-F1 → Balanced Accuracy → smaller K).
11. Run diameter-within-type control (subgroup grouped probing; anti-shortcut check).
12. Export predictive separability results.

Stage 4 evaluates local label consistency using a **cosine k-nearest neighbors** (kNN) probe. Cosine similarity emphasizes angular proximity, providing a robust neighborhood definition in high-dimensional spaces. As a minimally parametric method, kNN measures recoverability directly from neighborhood structure rather than a learned decision boundary, preserving the native geometry of the frozen embeddings and reducing the risk that a stronger downstream classifier masks representational weaknesses. Implementation uses cosine distance, uniform weighting, and brute-force search to maintain direct dependence on local structure.

Because the dataset contains multiple strikes from the same physical cymbal, a sample-level split would cause instrument-identity leakage. To prevent this, folds are constructed by grouping on

 doi:10.20944/preprints202603.1837.v1

cymbal_id, ensuring that no instrument appears in both TRAIN and TEST and thus targeting generalization to unseen cymbals. When label coverage permits, splits use *StratifiedGroupKFold* to maintain class balance while preserving group isolation; otherwise, a controlled fallback to *GroupKFold* is applied and explicitly logged. Label encoding is performed within each fold using TRAIN labels only, and TEST labels absent in the corresponding TRAIN set are excluded from supervised evaluation, as recoverability cannot be meaningfully assessed for unseen classes.

Preprocessing is strictly fold-contained: when used, *StandardScaler* is fitted on the TRAIN fold and applied to TEST, preventing information leakage. Evaluation is performed under a fixed embedding view for the final Stage 2 winners, ensuring identical representational conditions across model families. Although both raw and L2-normalized exports are available for traceability, each Stage 4 run uses one explicitly selected view. Probing performance is assessed via a K-sweep ($K \in \{1, 3, 5, 11, 21\}$) to examine recoverability across local scales. Best-K is selected separately for each model–target pair, using **mean Macro-F1** as the primary criterion, **Balanced Accuracy** as the secondary criterion, and the smallest K as a deterministic tiebreaker. This choice is motivated by the fact that the dataset is class-imbalanced and group-structured. Primary metrics are Macro-F1, which evaluates class-wise recoverability under imbalance, and Balanced Accuracy, which reflects recall consistency across classes. Secondary metrics include **Matthews Correlation Coefficient (MCC)** and overall **Accuracy**, reported for completeness but not used for ranking. Fold-wise results are summarized with mean, standard deviation, and 95% confidence intervals as descriptive indicators of central tendency and stability. For transparency, the pipeline exports fold records, class summaries, confusion matrices, and paired CLAP–MERT fold-wise deltas as additional diagnostics.

The analysis prioritizes *cymbal_type* as the primary target and strike-related variables as secondary targets, with *diameter* retained only as a diagnostic control. This is necessary because physical size is structurally linked to cymbal family and may be recovered indirectly via type-mediated shortcuts. Stage 4 therefore applies **within-type probing** as a confound-aware test, assessing whether *diameter* remains informative once the dominant macro-level confound is constrained. The procedure is explicitly **anti-shortcut**, distinguishing genuine diameter-related structure from family-driven dependency. Implementation uses subgroup-specific grouped evaluation, probing each within-type subset separately while enforcing minimum subgroup size, sufficient label coverage, and retention of only admissible folds

4.5. Stage 5: Unsupervised Clustering Alignment

Stage 5 was designed as supporting unsupervised inspection evidence, intended to examine whether the frozen embedding spaces exhibit intrinsic, label-blind organization that aligns post-hoc with known cymbal categories. Therefore, this stage allows distinguishing between true category-aligned cluster structures and more diffuse, manifold-like neighborhoods. The workflow is delineated in the following list.

1. Load Stage 2 winners.
2. Lock embeddings and metadata.
3. Define representation level and target.
4. Create sample-level and cymbal-level views.
5. Apply geometry-aware preprocessing
- P1 L2 normalization / PCA-50 stabilization / optional post-PCA L2
6. Run label-blind clustering
- Spherical k-means as the main angular partition
- probe with oracle K
- HDBSCAN as density-/manifold-aware
- Agglomerative clustering as cosine-compatible hierarchical clustering).
7. Compute external alignment metrics

 doi:10.20944/preprints202603.1837.v1

- AMI as primary
- ARI, V-measure
- Homogeneity
- Completeness

8. Compute internal diagnostics

- Silhouette Score,
- Davies–Bouldin
- HDBSCAN noise ratio

9. Inspect contingency and overlay outputs
10. Supporting intrinsic-structure evidence rather than direct unsupervised "classification" benchmarking
11. Run diagnostic confound control (Diameter-within-Type / within-type clustering and alignment checks).
12. Export summaries and figures.

To preserve comparability across the study, Stage 5 operates exclusively on the final embedding representations of the models selected in Stage 2. The analysis remains aligned with the target hierarchy established in the preceding stages. Clustering was examined first at the native sample level and then at the cymbal level through aggregation by *cymbal_id*, to test whether the observed structure persists after reducing the influence of repeated strikes from the same physical instrument. This robustness step was methodologically important, as clustering at the individual-hit level may partly reflect shared instrument identity rather than more generalizable organization of the embedding space. Labels intrinsically tied to the strike level were evaluated only where their interpretation remained valid after aggregation. Accordingly, Stage 5 was treated as a stage-locked structural audit of the same representational space previously examined geometrically in Stage 3 and quantitatively in Stage 4.

Before clustering, the embeddings were passed through a geometry-aware preprocessing pipeline designed to preserve cosine-consistent structure while improving numerical stability in high-dimensional space. The default sequence consisted of L2 normalization, PCA reduction to a maximum of 50 components, and, where required, post-PCA L2 normalization. Initial L2 normalization ensured that similarity was driven primarily by vector direction rather than magnitude, consistent with cosine-based embedding analysis. PCA was used as a stabilization step to reduce high-dimensional noise, improve neighborhood reliability, and make clustering less sensitive to sparsity effects. Retaining up to 50 principal components represented a practical compromise between compression and structural preservation, while explained variance was recorded for transparency. A second L2 normalization after PCA was retained to preserve angular consistency, particularly for clustering methods that rely on direction-based similarity. This protocol was applied symmetrically to both models as a common analytical discipline rather than a model-specific optimization.

Three complementary clustering methods were applied in a strictly label-blind manner. **Spherical k-means**, a centroid-based variant for normalized vectors, served as the primary angular clustering probe, as it is well suited to embedding spaces in which cosine-like geometry is more informative than raw Euclidean distance. The number of clusters was fixed to the known number of categories for each target (Oracle K), allowing Spherical k-means to function as a controlled test of whether the embedding space naturally forms compact, directionally coherent, and category-aligned groups when cluster cardinality matches the true class count. **HDBSCAN** was used as the primary density-based method because it identifies clusters only where sufficiently dense and coherent regions exist, without forcing all samples into an assignment. As a result, ambiguous or boundary points can remain unassigned as noise, providing additional information on whether the embedding space is organized into well-defined category islands or a more diffuse, manifold-like structure with

 doi:10.20944/preprints202603.1837.v1

transitional regions. **Agglomerative** clustering with cosine-compatible settings was included as a supporting hierarchical method to examine whether the structure of the embedding space is expressed more naturally through nested partitions than through a single flat decomposition. In all cases, labels were introduced only after clustering for alignment analysis and were never used to form the clusters, ensuring that Stage 5 remained a genuine test of intrinsic organization rather than a disguised form of supervised evaluation.

Cluster–label correspondence was quantified primarily with **Adjusted Mutual Information (AMI),** which was used as the main external alignment metric because it is chance-corrected and comparatively robust to class imbalance and differing cluster cardinalities. **Adjusted Rand Index (ARI)** was reported as a secondary reference metric, providing a complementary view of partition agreement based on pairwise sample assignments. **V-measure**, together with homogeneity and completeness, was used to diagnose fragmentation and merging effects, indicating whether cluster solutions tended to over-segment individual classes or merge multiple classes into shared partitions. Internal metrics, including **silhouette score** and **Davies–Bouldin index**, were treated only as geometric diagnostics of compactness and separation, indicating how tightly grouped and well-separated the clusters were in the representation space, rather than as evidence of semantic validity. For HDBSCAN, the noise ratio was retained as an additional diagnostic, since high noise levels may indicate diffuse or manifold-like organization rather than well-formed compact clusters. Visual outputs were used strictly as interpretive support rather than standalone evidence, with contingency heatmaps inspecting cluster purity and mixing, cluster-colored low-dimensional overlays relating cluster assignments to the Stage 3 projections, and hierarchical plots serving as supporting diagnostics for the overall structure of the embedding space.

Finally, diameter was treated, as in the preceding stages, as a confound-sensitive diagnostic target and was re-examined within fixed cymbal_type subsets. This within-type analysis tested whether any apparent size-related alignment persisted after constraining the dominant effect of cymbal family, thereby serving as a complementary anti-shortcut control.

5. Results

This section presents the results of the CLAP - MERT comparison across the final stages (3-5) of the evaluation framework. Stages 1 and 2 are preparatory; therefore, their outcomes are reported within the corresponding subsection of the methodology (4.1 and 4.2 respectively).

5.1. Stage 3: Geometry and Visualization Findings

The presentation of the evaluation results follows the target variable hierarchy, with *cymbal_type* as the primary target, *hit_zone* and *stick_material* as the main fine-grained targets, *stick_type* and *hit_point* as secondary descriptors, and *diameter* as a confound-aware diagnostic target. For each label, all ten DR visualizations of Table 6 were generated. The following subsections focus on the most informative ones in terms of interpretability, used for the subsequent confirmatory quantitative analyses of Stage 4.

5.1.1. Macro-Level Taxonomy: Geometric Structure by Cymbal Type

For the primary target *cymbal_type*, the core geometric evidence is anchored to two projections: PCA-2D and unsupervised UMAP-2D under the canonical train–test visual split and shared P3 preprocessing. PCA provides the deterministic linear baseline for assessing variance concentration, whereas unsupervised UMAP serves as the main non-linear view of local neighborhood organization. Together, these views define the principal macro-level comparison between CLAP and MERT in Stage 3.

The evaluation of these two geometries is shown on Table 7. The metrics on the columns, from left to right are: PCA EV-2D is the explained variance of the 2D PCA projection, PCA Dims (90%) is the number of principal components required to explain 90% of the total variance, PCA EV-90 reports

 doi:10.20944/preprints202603.1837.v1

on the total variance captured by PCA Dims (90%), UMAP TR-Trust and UMAP TE-Trust is the trustworthiness of the training points and the test points on the UMAP projection respectively, while P@5 is the precision at 5 nearest neighbors, mAP@10 the mean average precision at 10 nearest neighbors, P@20 the precision at 20 nearest neighbors and mAP@20 the mean average precision at 20 nearest neighbors.

Table 7. Core Stage 3 geometry summary for the primary target *cymbal_type*.

Model	PCA EV-2D	PCA Dims (90%)	PCA EV-90	UMAP TR-Trust	UMAP TE-Trust	P@5	mAP@10	P@20	mAP@20
CLAP	**0.3045**	**27**	0.9036	0.9771	**0.9164**	**0.5439**	**0.6336**	**0.5102**	**0.5952**
MERT	0.1769	547	0.9001	**0.9890**	0.8763	0.5034	0.5737	0.4850	0.5504

The clearest macro-structural contrast emerges in the PCA baseline. CLAP reaches approximately 90% retained variance with only 27 principal components, whereas MERT requires 547 components to reach a comparable threshold. Additionally, the 2D PCA projection retains substantially more explained variance for CLAP than for MERT (0.3045 vs. 0.1769). This indicates that the dominant directions of variation are much more concentrated in CLAP, supporting a more compact macro-geometric organization.

The two PCM-2D projections are depicted on Figure 3. As the linear baseline of Stage 3, this view summarizes the dominant variance structure of the embedding spaces. Visually, CLAP forms a broader and more evenly distributed configuration across the 2D plane, whereas MERT appears more compressed, with stronger central class mixing. This pattern is consistent with the stronger PCA compactness of CLAP, as reflected in its higher PCA EV-2D and markedly lower PCA Dims (90%). This pattern should be interpreted as evidence of stronger variance concentration rather than as direct proof of superior class separability.

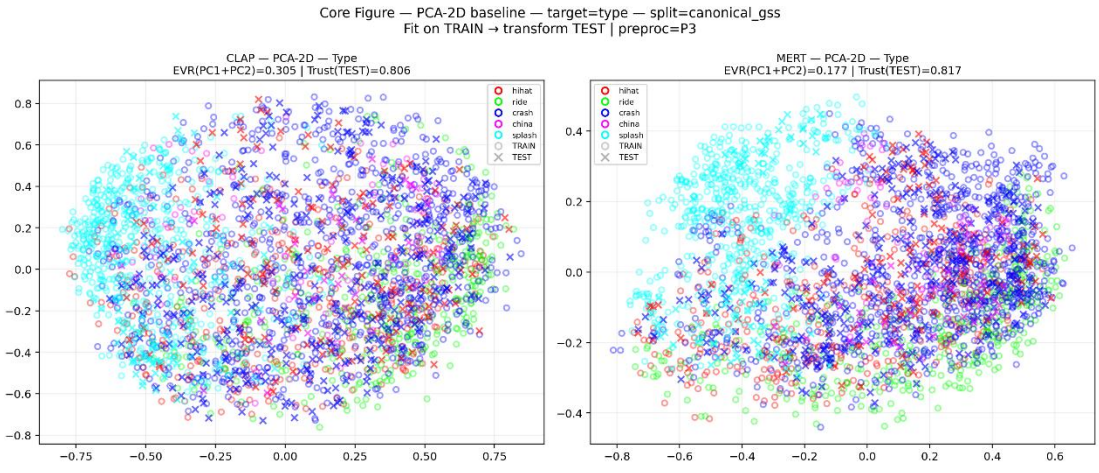


Figure 3. PCA-2D projection for *cymbal_type* under the canonical train-test visual split.

The unsupervised UMAP view reinforces this interpretation at the level of local geometry. Under the canonical train-fit / test-transform protocol, CLAP achieves higher TEST trustworthiness than MERT (0.9164 vs. 0.8763), indicating that its low-dimensional neighborhood structure more faithfully reflects the original embedding space. This pattern is further supported by the original-space retrieval probes for *cymbal_type*, where CLAP also outperforms MERT under group-aware exclusion, according to the measures of P@5 , mAP@10, P@20 and mAP@20.

The UMAP 2D projection is depicted on Figure 4. As the main non-linear view of Stage 3, this projection summarizes local neighborhood organization in the embedding spaces. Visually, CLAP

 doi:10.20944/preprints202603.1837.v1

shows a more spatially extended and more locally coherent arrangement, whereas MERT appears more compact and more mixed in the central regions. This contrast is consistent with the higher UMAP test trustworthiness of CLAP and its stronger retrieval-based neighborhood consistency in the original embedding space.

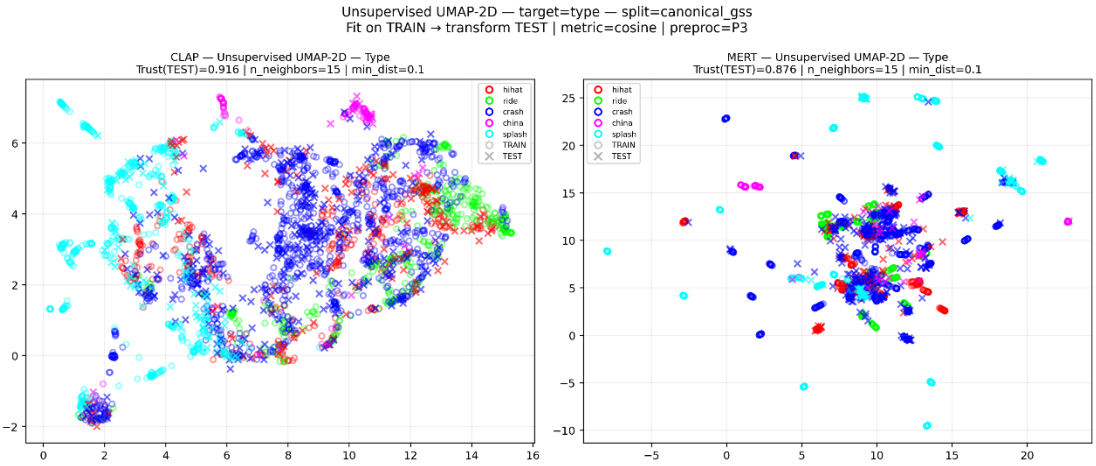


Figure 4. Unsupervised UMAP-2D projection for *cymbal_type* under the canonical train-test visual split.

Taken together, the PCA compactness metrics, the UMAP trustworthiness results, and the original-space retrieval probes converge on the same macro-level conclusion. For cymbal_type, CLAP shows stronger variance concentration, higher projection trustworthiness under the selected UMAP protocol, and more label-consistent local neighborhoods than MERT.

5.1.2. Micro-Level Acoustic Traits: Hit Zone and Articulation Cues

Concerning fine-grained acoustic traits, the results focus on strike-related articulation and stick-mediated timbral variation, emphasizing on local neighborhood organization. Here, the interpretation is driven primarily by the unsupervised UMAP views and the original-space retrieval probes, which are more informative for fine-grained neighborhood structure. All UMAP views in concerning this level of information, use the same unsupervised projection and differ only in their target-specific label overlays.

The clearest micro-level pattern in Stage 3 is that the CLAP advantage becomes more pronounced for strike-related targets than for the primary macro-level target, type, as depicted on Table 8. CLAP yields higher retrieval scores than MERT across all reported micro-level targets, with the strongest margins observed for *hit_zone*, *stick_type*, and *stick_material*, indicating more label-consistent local neighborhood organization for strike-related and stick-related acoustic traits.

Table 8. Original-space retrieval summary for the fine-grained secondary targets in Stage 3 under group-aware exclusion.

Target	CLAP P@5	MERT P@5	CLAP mAP@10	MERT mAP@10
hit_zone	0.7746	0.6727	0.8235	0.7448
stick_material	0.6554	0.5037	0.7284	0.6188
stick_type	0.7919	0.6145	0.8360	0.7049
hit_point	0.3968	0.3245	0.5231	0.4629

CLAP superiority is most evident for *hit_zone*, which emerges as one of the strongest fine-grained findings in the geometric analysis. Under group-aware exclusion, CLAP achieves substantially higher retrieval performance than MERT, while the corresponding core unsupervised

 doi:10.20944/preprints202603.1837.v1

UMAP view is consistent with a more coherent and more clearly separated local arrangement of the three strike-zone classes. Collectively, these results suggest that CLAP organizes hit-zone-related local neighborhoods more consistently than MERT in both the projected space and the original embedding space.

Figure 5 shows the colored overlay for label *hit_zone* of the UMAP visualization of Figure 4 for this target. Compared to MERT (on the right), CLAP (on the left) forms more coherent and more clearly separated local neighborhoods for the three strike-zone classes, supporting stronger local organization of hit-zone-related information in the embedding space.

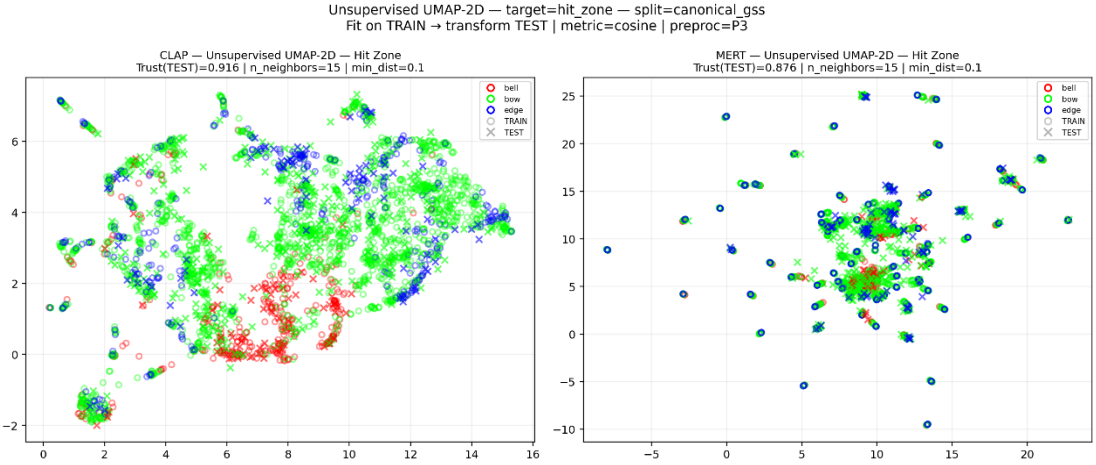


Figure 5. Unsupervised UMAP-2D projection under the canonical train-test visual split, shown with *hit_zone* label overlay.

A similar pattern extends to *stick_material*, which serves here as the second main fine-grained target. Although the margin here is lower than for *hit_zone*, the retrieval evidence remains clearly in favor of CLAP. This replication across more than one strike-related variable is important, because it suggests that the CLAP advantage is not confined to a single fine-grained label but extends across multiple local descriptors linked to playing conditions and timbral variation.

Figure 6 shows the overlay of the label *stick_material.* Also here, CLAP shows more structured and more locally coherent neighborhoods than MERT across the three material categories, consistent with stronger fine-grained organization of material-related acoustic cues.

 doi:10.20944/preprints202603.1837.v1

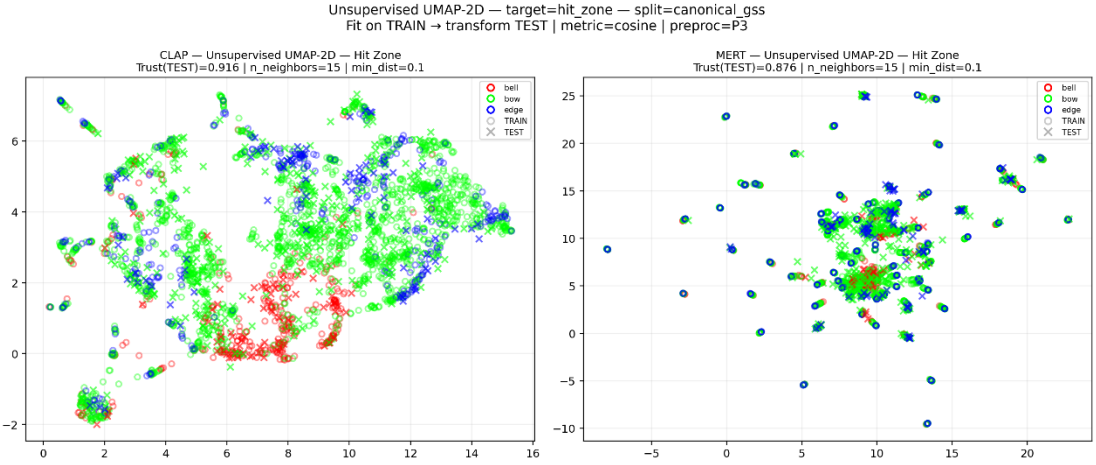


Figure 6. Unsupervised UMAP-2D projection under the canonical train-test visual split, shown with *stick_material* label overlay.

The auxiliary secondary descriptors further refine this micro-level picture. For *stick_type* CLAP shows one of the strongest retrieval margins observed in Stage 3 (P@5 = 0.7919 vs. 0.6145; mAP@10 = 0.8360 vs. 0.7049), consistent with especially strong organization of stick-related local neighborhoods. By contrast, *hit_point* also favors CLAP, but with a more moderate gap (P@5 = 0.3968 vs. 0.3245; mAP@10 = 0.5231 vs. 0.4629), suggesting that this descriptor remains more difficult for both models even though the local geometric advantage still points in the same direction. Accordingly, these two targets are best interpreted as supporting micro-level descriptors rather than as co-equal anchors of the subsection.

This is evident from the label overlay of Figure 7. In this diagram, CLAP exhibits clearer local grouping of the two stick-type categories than MERT, consistent with stronger neighborhood-level organization of stick-related signal in the embedding space.

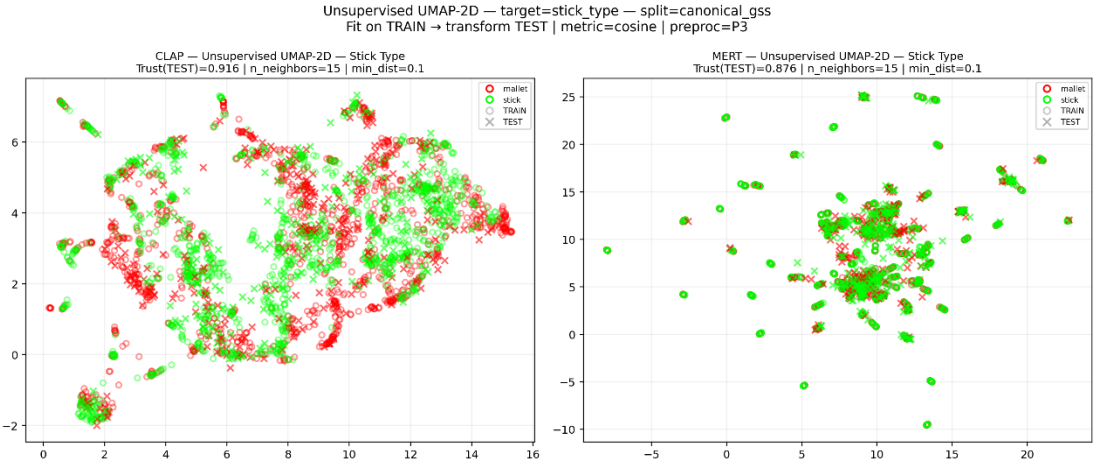


Figure 7. Unsupervised UMAP-2D projection under the canonical train-test visual split, shown with *stick_type* label overlay.

Taken together, the fine-grained results converge on a clear geometric interpretation: the strongest Stage 3 advantage of CLAP does not appear only at the broad category level but becomes even more visible in local neighborhoods related to strike articulation and stick-mediated acoustic variation. This makes the micro-level branch of Stage 3 particularly important for the overall comparative narrative of the study, because it suggests that CLAP is especially effective at organizing the subtle local structure needed to separate fine-grained cymbal-playing attributes. As with the preceding macro-level findings for cymbal type, these results remain geometric and will be evaluated more strictly in Stages 4 and 5.

 doi:10.20944/preprints202603.1837.v1

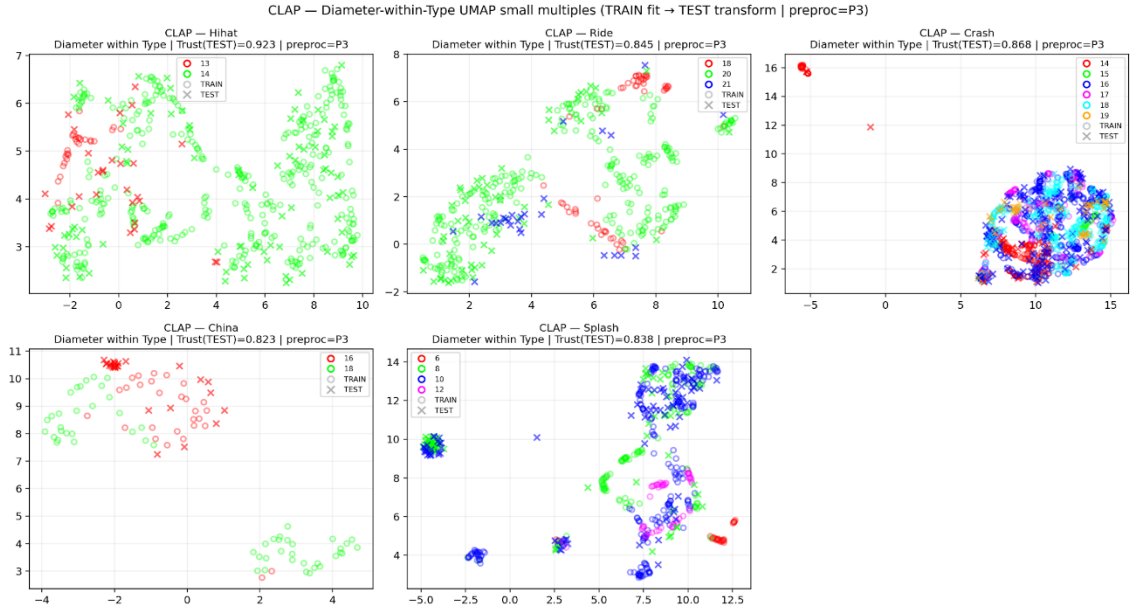
5.1.3. Confound-Aware Control: Diameter as a Type-Mediated Target

Given the confound-sensitive relation between *diameter* and *cymbal_type* (section 3.2), the diameter is treated as a diagnostic target rather than as an ordinary success case. The global retrieval results, shown on Table 9, already motivate this caution. Under group-aware exclusion, diameter remains comparatively weak for both models, with CLAP reaching P@5 = 0.2787 and mAP@10 = 0.3865, and MERT reaching P@5 = 0.3019 and mAP@10 = 0.3762. Relative to the stronger retrieval patterns observed for *cymbal_type*, *hit_zone*, *stick_type*, and *stick_material*, these values indicate that size-related local structure is substantially less stable and less clearly organized in the original embedding space. Thus, the global diameter signal should not be interpreted as direct evidence of robust independent size representation.

Table 9. Original-space retrieval summary for the diagnostic target variable *diameter*.

Model	P@5	mPA@10	P@20	mAP@20
CLAP	0.2787	0.3865	0.2612	0.3588
MERT	0.3019	0.3762	0.2810	0.3541

Accordingly, the main evidence in this subsection comes from the diameter-within-type views rather than from the unrestricted global projection. Figure 8 depicts the UMAP visualizations for diameter-within type using CLAP (top) and MERT (bottom). This confound-aware control tests whether size-related organization remains visible once *cymbal_type* is held fixed, thereby functioning as an anti-shortcut safeguard for the interpretation of diameter-related geometry. Within fixed cymbal families, the color-coded diameter labels remain only partially distinguishable, and the strength of this organization varies across panels. Compared with MERT, CLAP generally shows more continuous and more locally coherent within-type diameter patterns, although the overall structure remains weaker and more heterogeneous than that observed for the primary and strike-related targets. At the same time, this advantage remains conditional and should not be over-interpreted, because the family-specific patterns are neither uniform nor strong enough to support a simple success narrative for diameter.



 doi:10.20944/preprints202603.1837.v1

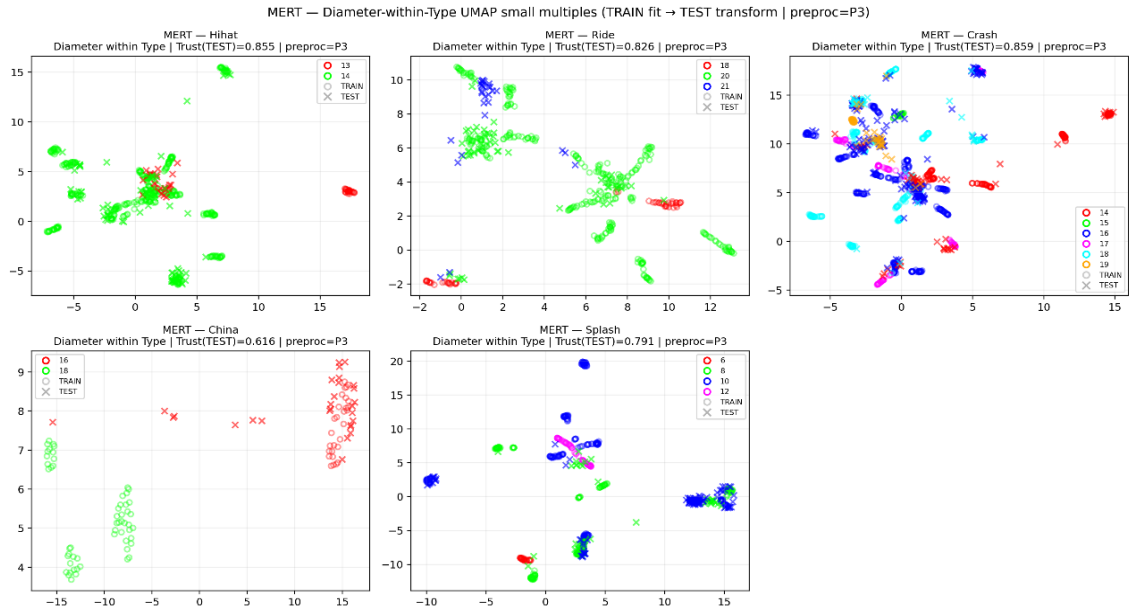


Figure 8. Diameter-within-type UMAP small multiples for the diagnostic target diameter, with CLAP shown on top and MERT on bottom.

Taken together, these results position *diameter* as a diagnostic control rather than as a central positive finding of Stage 3. Its role is to constrain interpretation, test for shortcut-sensitive structure, and clarify whether any apparent size-related organization survives once type is controlled. Under this confound-aware reading, the most defensible conclusion is that diameter-related structure is limited, heterogeneous, and partly type-mediated in both models, although CLAP may retain somewhat more coherent residual organization in selected within-type settings. As with the preceding findings, these results remain geometric and will be evaluated more strictly in the downstream stages.

5.2. Stage 4: Leakage-Safe Probing Results

Overall, Stage 4 reveals a clear and differentiated comparative pattern. CLAP outperforms MERT on five of the six evaluated targets, i.e. *hit_zone*, *stick_material*, *stick_type*, *hit_point*, and diagnostically on *diameter*, while MERT retains a modest advantage only on the primary macro-family target type. Thus, the results do not support a single overall winner, but rather a target-dependent split of representational strengths between the two embedding spaces.

5.2.1. Predictive Separability of the Macro-Target

The primary target of the study is cymbal_type, as it reflects the macro-acoustic family structure and carries the greatest interpretive weight. Under the leakage-safe grouped probing protocol, MERT shows a modest but consistent advantage over CLAP. As shown in Table 10, at their optimal K values, MERT peaks at K = 11, while CLAP peaks at K = 5. Although moderate, this difference is consistent across all primary metrics, indicating an advantage for MERT on the macro-family target.

Table 10. Leakage-safe group-aware probing performance for cymbal_type in the original embedding space.

Model	Best-K	Macro-F1	Balanced Accuracy	MCC	Accuracy
CLAP	5	0.548 ± 0.033	0.566 ± 0.044	0.465 ± 0.074	0.612 ± 0.058
MERT	11	0.577 ± 0.082	0.578 ± 0.067	0.490 ± 0.081	0.625 ± 0.056

 doi:10.20944/preprints202603.1837.v1

As shown in Figure 9 the K-sweep analysis, further shows that CLAP reaches optimal performance within a narrower neighborhood scale (K=5), whereas MERT peaks at a broader one (K=11). This suggests that CLAP preserves stronger local consistency, while MERT benefits from more stable broader family-level organization in recognizing cymbal types.

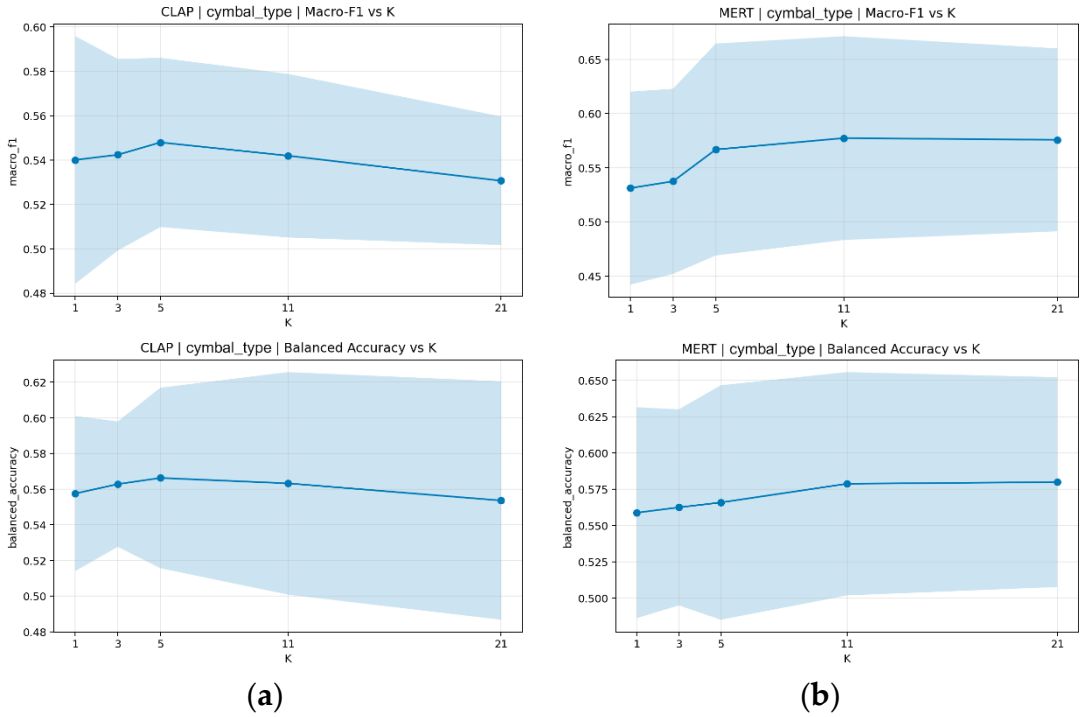


Figure 9. Macro-F1 (top) and Balanced Accuracy scores across neighborhood sizes (K = 1, 3, 5, 11, 21) for CLAP and MERT on the primary target *cymbal_type*. The shaded error bands indicate ±1 standard deviation (SD) across the five folds, reflecting model stability across test sets. **(a)** CLAP reaches its best results and K=5. **(b)** MERT peaks at K=11.

This MERT advantage on the macro-target becomes especially informative when contrasted with the Stage 3 findings. During geometric inspection, CLAP appeared visually more compact and more readable for *cymbal_type*, showing stronger local-neighborhood evidence in the original-space retrieval layer. Yet Stage 4 demonstrates that this geometric clarity does not automatically translate into superior predictive recoverability on unseen data. The fact that geometric readability and leakage-safe separability are related distinct properties of an embedding space strongly justifies the need for the present multi-stage evaluation framework. Overall, the results for *cymbal_type* provide the main qualification against an overly simplified overall ranking of the two models: CLAP, as shown in the following sections, is clearly stronger for fine-grained strike-related information, whereas MERT remains slightly stronger for family-level cymbal categorization under leakage-safe probing.

5.2.2. Fine-Grained Recoverability of Strike-Related Targets

CLAP's strongest advantage in Stage 4 appears on the fine-grained strike-related targets hit_zone and stick_material, the study's main secondary targets. As shown in Table 11, it outperforms MERT across all primary metrics, indicating more recoverable local structure for labels tied to transient cues, attack characteristics, and fine timbral variation. This suggests that the embedding space encodes not only the vibrating object but also its excitation pattern, with CLAP capturing strike-dependent information more effectively within local neighborhoods.

The same tendency is demonstrated by the supporting secondary targets. Although *stick_type* and *hit_point* are not treated as core fine-grained claims of equal weight, they provide useful auxiliary confirmation of the CLAP advantage on strike-related information. In Stage 4, CLAP outperforms

 doi:10.20944/preprints202603.1837.v1

MERT for both labels, *stick_type* and *hit_point*. The label *stick_type* appears to be more easily recoverable, while *hit_point* remains more demanding, although it still favors CLAP.

These findings are also closely aligned with those of Stage 3. As shown on Figure 5 (*hit_zone*), Figure 6 (*stick_material*) and Figure 7 (*stick_type*), CLAP demonstrated stronger neighborhood consistency that MERT for the same targets. Thus, the stronger local organization observed in Stage 3 translates into stronger supervised recoverability in the original embedding space.

Table 11. Leakage-safe group-aware probing performance for the fine-grained strike-related targets in the original embedding space.

Target	Role	CLAP Macro-F1	MERT Macro-F1
hit_zone	Main secondary	0.7368	0.6227
stick_material	Main secondary	0.7015	0.5557
stick_type	Supporting secondary	0.8329	0.7166
hit_point	Supporting secondary	0.4307	0.3182

These observations are further confirmed by the figures that follow. Specifically, Figure 10 shows the K-Sweep analysis for the secondary targets within the two models. In both cases, CLAP remains stronger than MERT across the K-sweep and remains stable across a broad range of neighborhood sizes. Furthermore, Figure 11 depicts the 5-fold-averaged values of Macro-F1 for *hit_zone* when evaluated within each cymbal family type. CLAP remains stronger for the crash, hi-hat, and ride cymbal families, while becoming approximately equivalent to MERT only in splash. Equivalently, Figure 12 depicts the same measures for *stick_material*. In this case across all available families, CLAP remains consistently stronger than MERT across all family types. Please note that the china cymbal family is not included in the diagram due to its reduced representation in the dataset (i.e. 112 samples of four distinct physical cymbals) resulting in insufficient group folds.

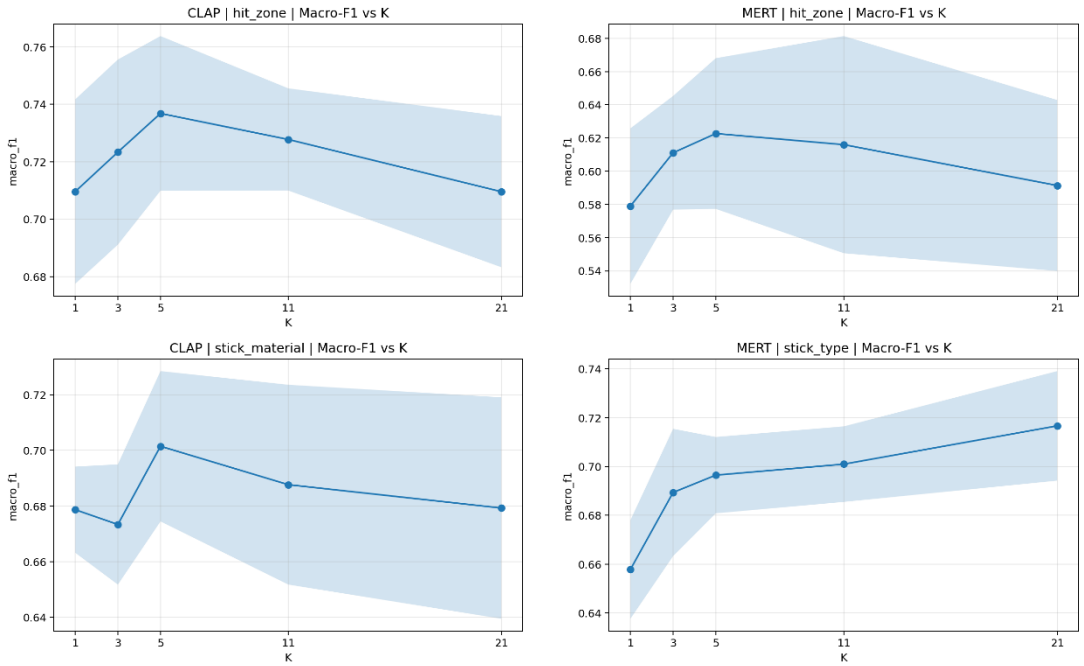


Figure 10. Macro-F1 (top) and Balanced Accuracy scores across neighborhood sizes (K = 1, 3, 5, 11, 21) for **(a)** CLAP and **(b)** MERT on the main secondary targets *hit_zone* (top) *and stick_material* (bottom). The shaded error bands indicate ±1 standard deviation (SD) across the five folds, reflecting model stability across test sets.

 doi:10.20944/preprints202603.1837.v1

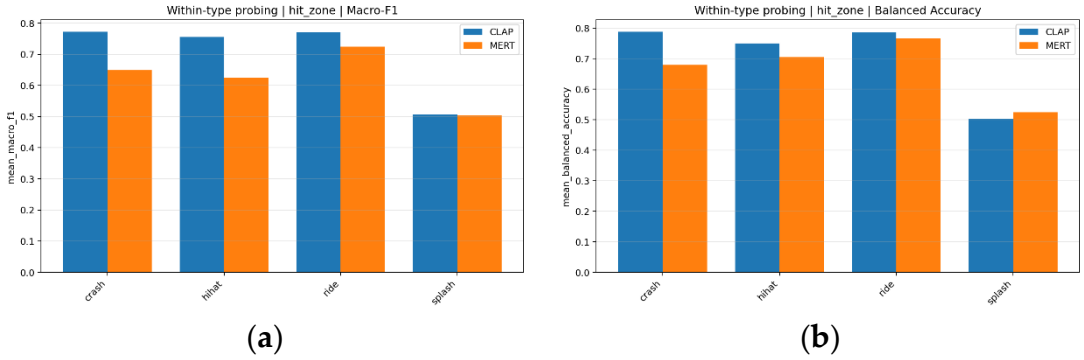


Figure 11. Fold-averaged **(a)** Macro-F1 and **(b)** Balanced Accuracy for the *hit_zone* label under grouped evaluation within each fixed cymbal family.

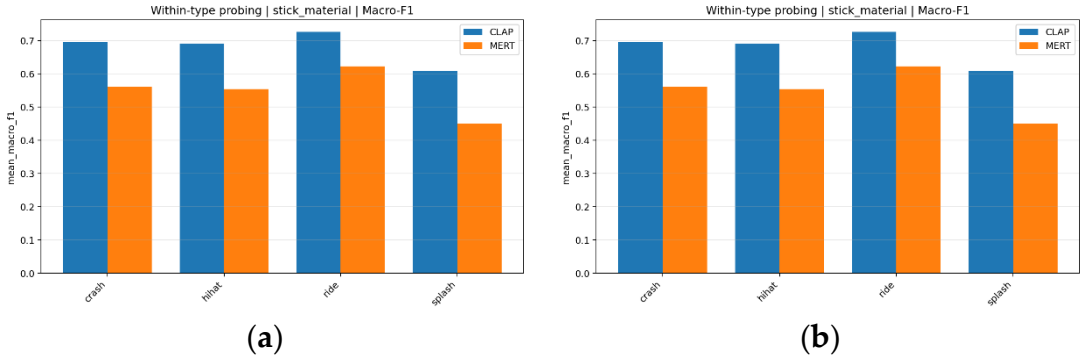


Figure 12. Fold-averaged **(a)** Macro-F1 and **(b)** Balanced Accuracy for the *stick_material* label under grouped evaluation within each fixed cymbal family.

5.2.3. Confound-Aware Probing: Within-Type Diameter Control

Unlike the other labels, diameter is not treated as a standard prediction target but rather as a diagnostic, confound-sensitive variable. This is because physical size is closely tied to the cymbal family and may therefore be inferred indirectly through type-related shortcuts rather than through genuinely size-specific acoustic features. For this reason, the results for diameter are reported separately and should be interpreted with caution.

This caution is further supported by the grouped probing results. In the unrestricted evaluation setting, only 2 out of the 5 folds were valid for both models. The remaining folds were excluded because the TEST partitions contained diameter labels that were not present in the corresponding TRAIN folds (specifically 15, 6, and 21 inches). While these exclusions serve as necessary methodological safeguards, they also highlight that diameter does not provide the same stable or reliable basis for comparison as the other targets.

As shown in Table 12, the confound-aware value of this target becomes clear in the within-type analysis. When probing is restricted within fixed cymbal families, the pattern becomes mixed: CLAP performs better in ride and splash, whereas MERT is slightly better within crash and hi-hat. No metrics are provided for china cymbals, as the relatively small number of samples leads to insufficient group representation, which in turn produces inadmissible folds.

Unlike *hit_zone* or *stick_material*, which retained a relatively coherent CLAP advantage under within-type control (Figure 11 and Figure 12 respectively), *diameter* does not show a stable cross-family pattern. Instead, it behaves as a boundary-condition label, indicating that any apparent size-related recoverability is contingent, family-dependent, and vulnerable to confounding by macro-category structure.

 doi:10.20944/preprints202603.1837.v1

Table 12. Within-type probing results for *diameter* across cymbal families under grouped leakage-safe evaluation.

Cymbal Family	CLAP Macro-F1	MERT Macro-F1	CLAP Balanced Accuracy	MERT Balanced Accuracy
crash	0.2567	0.2877	0.3249	0.3629
hi-hat	0.4786	0.4948	0.7232	0.7470
ride	0.4454	0.3610	0.7321	0.6845
splash	0.3898	0.3050	0.5365	0.4375
china	—	—	—	—

This result is consistent with Stage 3, which had already suggested a mixed and confound-sensitive picture for diameter. Stage 4 makes that caution stricter by showing that, once leakage-safe grouped probing and within-type control are imposed, the diameter pattern becomes clearly less stable than the corresponding patterns for the strike-related targets. In this sense, the analysis of diameter shows that the pipeline clearly distinguishes robust comparative findings from targets that should remain diagnostic only.

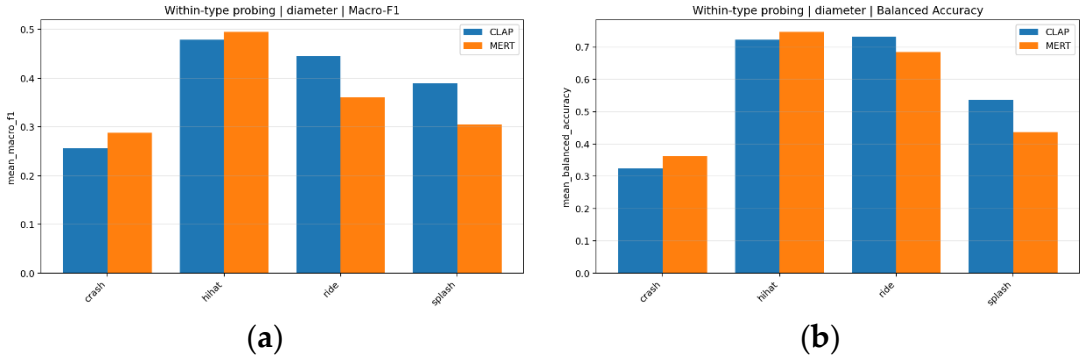


Figure 13. Within-type probing performance for *diameter* across cymbal families. **(a)** fold-averaged Macro-F1 and **(b)** fold-averaged Balanced Accuracy.

5.3. Clustering Alignment Results

This section presents the Stage 5 clustering alignment results, beginning with *cymbal_type* (subsection 5.3.1) and then turning to the finer-grained strike-related targets (subsection 5.3.2). Next, in subsection 5.3.3, it interprets the findings from an algorithmic perspective, contrasting relatively clean partition-based solutions with more diffuse manifold-like organization. Subsection 5.3.4 examines robustness across representation levels by comparing native sample-level clustering with cymbal-level aggregation by *cymbal_id*. Finally, subsection 5.3.5, considers diameter as a confound-sensitive diagnostic case, asking whether any apparent size-related alignment remains interpretable once the dominant effect of cymbal family is controlled.

5.3.1. Macro-Structural Emergence: Unsupervised Alignment of Cymbal Type

The central interpretive question of Stage 5 concerns *cymbal_type*, as this is the most important target in the study and the main point of divergence between the preceding stages. Stage 3 favored CLAP in terms of macro-geometric compactness and projection-level readability (refer to section 5.1.1), whereas Stage 4 showed a modest but consistent advantage for MERT under grouped probing (section 5.2.1). Stage 5 therefore examines whether this MERT advantage reflects stronger unsupervised family-level clustering or whether it remains mainly a probing effect.

 doi:10.20944/preprints202603.1837.v1

Table 13 summarizes the representative Stage 5 clustering solutions used to interpret the macro-structural behavior of *cymbal_type* in CLAP and MERT. It contrasts sample- and cymbal-level results and distinguishes relatively clean flat partitions from high-noise or strongly fragmented solutions, to assess whether the modest Stage 4 MERT advantage reflects stronger intrinsic family-level clustering or mainly more favorable probe-level recoverability. At the sample level, neither model forms fully discrete family clusters, although both show non-trivial alignment under partition-based clustering. For CLAP, the strongest visible flat solutions for *cymbal_type* remain modest, with agglomerative cosine slightly exceeding spherical k-means, while HDBSCAN yields weaker and less stable solutions. For MERT, one HDBSCAN configuration reaches a higher AMI, but only with a very high noise ratio and strong fragmentation (59 clusters). This suggests that the apparent MERT advantage does not simply correspond to cleaner natural partitioning, but partly to a more diffused and fragmented structure.

Table 13. Representative Stage 5 clustering results for *cymbal_type* for sample- and cymbal-level analysis.

Model	Level	Algorithm	AMI	ARI	V-measure	Noise ratio	Number of clusters	Interpretive role
CLAP	sample	Spherical k-means	0.1637	0.1077	0.1653	0.0000	5	Best visible sample-level spherical solution
CLAP	sample	Agglomerative cosine	0.1751	0.1464	0.1768	0.0000	5	Best flat partition at sample level
MERT	sample	HDBSCAN	0.2606	0.0735	0.2800	0.6080	59	Highest sample-level AMI, but strongly fragmented
MERT	sample	Agglomerative cosine	0.1704	0.1244	0.1722	0.0000	5	Best flat partition at sample level
CLAP	cymbal median	Agglomerative cosine	0.3362	0.2182	0.3783	0.0000	5	Clearest cymbal-level family partition
MERT	cymbal median	Agglomerative cosine	0.2492	0.1487	0.2952	0.0000	5	Best visible MERT cymbal-level flat solution

Accordingly, Stage 5 does not support a picture of sharply separated *cymbal_type* islands for either model at the sample level. Instead, the cymbal family type structure appears only moderately cluster-aligned and remains partly algorithm-dependent, especially under density-based solutions. The most defensible reading is that *cymbal_type* is represented in both embedding spaces as a moderate macro-structural tendency, more consistent with partially organized family regions than with clean, fully separated categories.

However, clustering separability becomes clearer at the cymbal level as shown on Figure 14. The *cymbal_type* target is organized more naturally there than at the raw sample level, suggesting that repeated-hit variability partly obscures family structure when individual strikes are clustered

 doi:10.20944/preprints202603.1837.v1

directly. Figure 14 presents the clusters formed on the cymbal-median, which denotes an aggregated instrument-level representation in which all strike-level embeddings belonging to the same physical cymbal (*cymbal_id*) are combined by taking the median value in each embedding dimension, yielding a single robust embedding per cymbal. This indicates that macro-family information is more stable as an instrument-level property than as a strike-level partition.

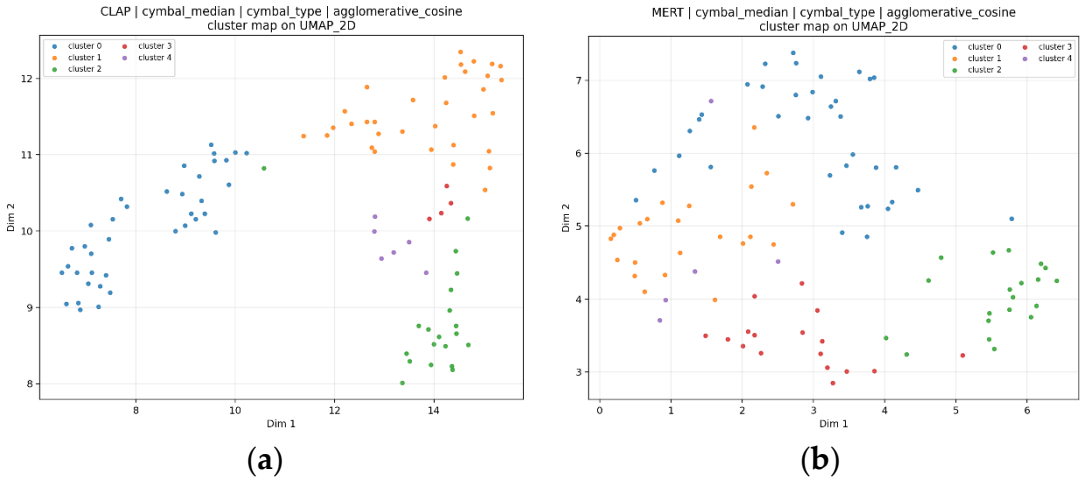


Figure 14. Cluster-colored low-dimensional overlays for the selected cymbal_median *cymbal_type* solutions of **(a)** CLAP and **(b)** MERT.

Overall, Stage 5 shows that family-level information is indeed present in the unsupervised geometry, but not in the form of uniformly clean, low-noise clustering. The safest conclusion is that *cymbal_type* is encoded in both embedding spaces as a moderate macro-structural signal rather than as a set of clean, fully separated clusters. At the sample level, MERT shows stronger alignment only in selected high-noise HDBSCAN settings, whereas at the cymbal level CLAP yields the clearest flat family partition under agglomerative cosine.

5.3.2. Fine-Grained Clustering Alignment for Strike-Related Targets

Stage 5 showed that the strike-related targets (*hit_zone*, *stick_material*, *stick_type*, and *hit_point*) do not emerge as strong, globally separated unsupervised clusters, even though Stages 3 (section 5.1.2) and 4 (sections 5.2.2) had already shown that these targets were more recoverable—especially in CLAP—under retrieval- and probing-based evaluation. Table 14 shows the evaluation metrics of these clusters. For each target, it lists the best-performing algorithm and the corresponding external alignment metrics, together with the noise ratio where applicable. At the sample level, the strongest CLAP solutions remained low in absolute terms. Thus, although alignment is not zero, none of these targets forms clean, trait-specific unsupervised clusters in the embedding space.

Table 14. Best Stage 5 clustering alignment results for strike-related targets across CLAP and MERT.

Target	CLAP best algorithm	CLAP AMI	CLAP ARI	CLAP V-measure	CLAP Noise ratio	MERT best algorithm	MERT AMI	MERT ARI	MERT V-measure	MERT Noise ratio
hit_zone	Agglomerative cosine	0.0668	0.0704	0.0676	0.0000	Agglomerative cosine	0.0667	0.0163	0.0674	0.0000

 doi:10.20944/preprints202603.1837.v1

stick_material	HDBSCAN (mcs = 15)	0.0482	0.0271	0.0508	0.8754	Spherical k-means	0.0946	0.0711	0.0952	0.0000
stick_type	HDBSCAN (mcs = 20)	0.0259	0.0028	0.0274	0.8896	Agglomerative cosine	0.0177	0.0228	0.0179	0.0000
hit_point	Spherical k-means	0.0735	0.0464	0.0766	0.0000	Spherical k-means	0.0847	0.0593	0.0878	0.0000

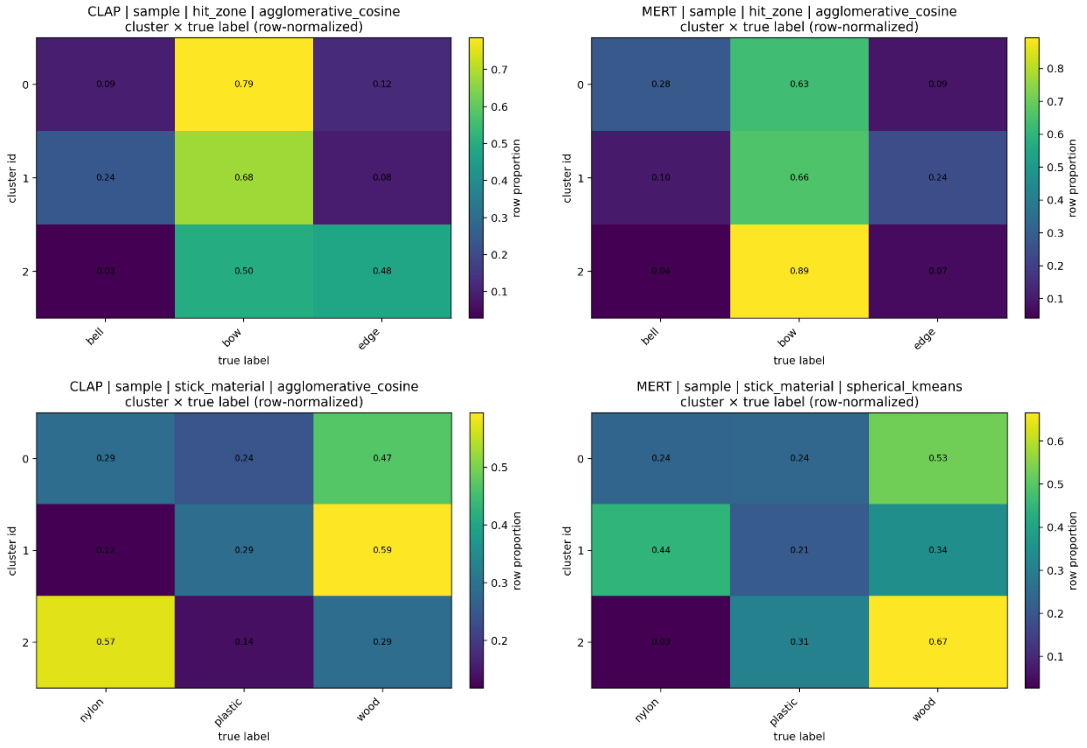


Figure 15. Normalized contingency heatmaps for the selected *hit_zone* (top) and *stick_material* solutions (bottom) for **(a)** CLAP and **(b)** MERT clusters.

This finding is important because it qualifies, rather than overturns, the earlier picture in favor of CLAP. In Stage 3, CLAP showed stronger neighborhood organization for the strike-related labels, and Stage 4 confirmed stronger grouped leakage-safe recoverability for the same targets. Stage 5 shows that this advantage does not translate into equally strong cluster emergence. Instead, the fine-grained signal appears to be expressed mainly as local or partially label-aligned structure rather than as sharply separated global partitions. In other words, the CLAP advantage remains real, but its structural form is closer to neighborhood consistency than to clean unsupervised category formation.

 doi:10.20944/preprints202603.1837.v1

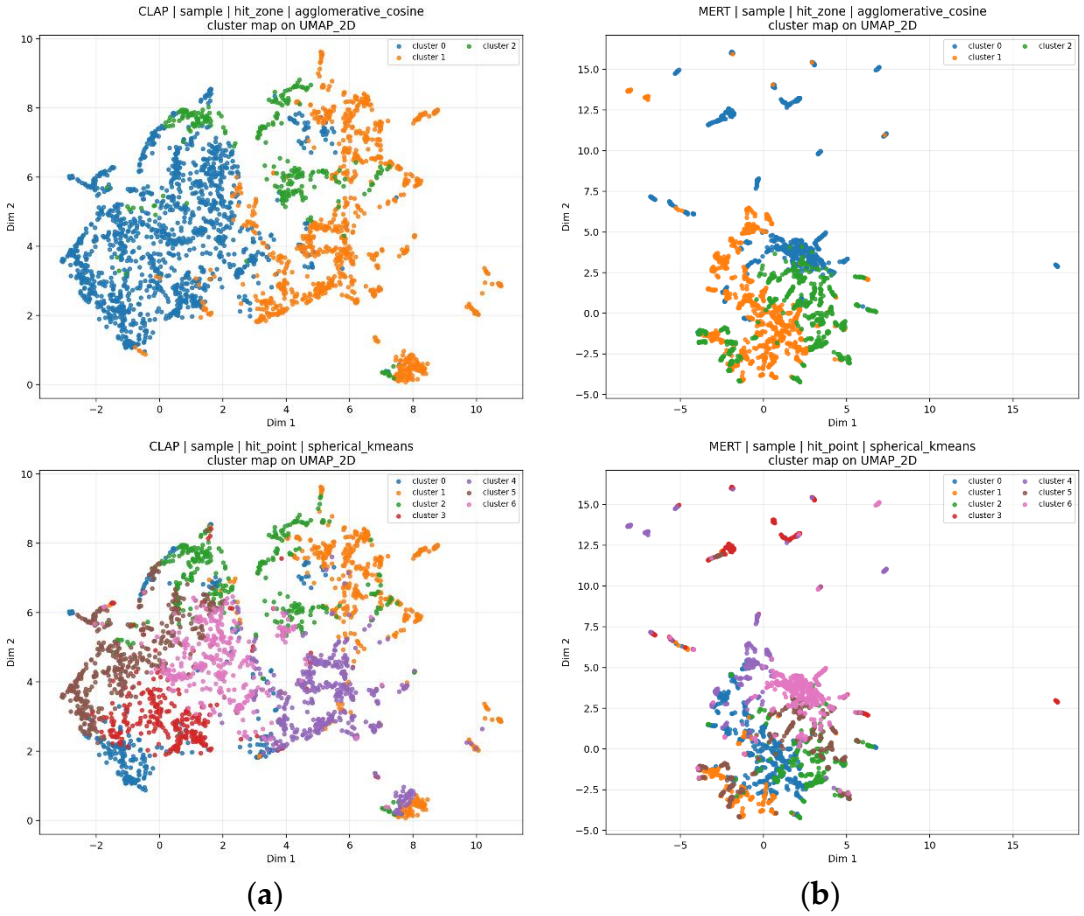


Figure 16. Cluster-colored low-dimensional overlays for *hit_zone* using the agglomerative clustering method (top) and the *hit_point* using spherical k-means clustering for **(a)** CLAP and **(b)** MERT.

The supporting targets, therefore, define the limits of this pattern even more clearly. For *stick_type*, Stage 4 had shown one of the strongest CLAP margins, yet Stage 5 did not reveal correspondingly strong unsupervised clustering, making it one of the clearest cases where high recoverability does not imply high cluster emergence. For *hit_point*, the
Stage 5 findings also remain weak, consistent with its role as the most difficult fine-grained target in Stages 3 and 4. In both cases, the safest interpretation is that the information is present and usable—especially in CLAP—but too diffuse to form stable unsupervised partitions.

5.3.3. Algorithmic Reading: Compact Partitions vs Manifold-Like Structure

Across the representative Stage 5 cases, the strongest and most interpretable alignments appear more consistently under partition-based methods (spherical k-means and agglomerative clustering) than under density-based clustering (HDBSCAN). More specifically, as shown in Table 15, HDBSCAN often yields lower alignment, elevated noise, or fragmented solutions, suggesting that the embedding spaces do not generally resolve into a small number of dense, well-isolated category islands. This pattern is already visible in the *cymbal_type* results, where HDBSCAN can produce non-trivial AMI in isolated cases, but only together with very high noise or strong fragmentation. Such cases should not be read as evidence of clean natural clustering. Rather, they indicate that some label alignment can arise within diffuse, unevenly populated, or weakly consolidated regions of the space.

 doi:10.20944/preprints202603.1837.v1

Table 15. Representative algorithm-level patterns in Stage 5 clustering behavior.

Algorithm	Representative case	Representative result	Structural role in Stage 5	Noise / fragmentation	Main structural reading
Spherical k-means	CLAP; sample, *cymbal_type*	AMI = 0.1637 ARI = 0.1077 V = 0.1653	Probes flat partitioning under fixed cluster cardinality and typically yields weak-to-moderate but interpretable label alignment	No noise by design	Weak-to-moderate but interpretable flat partition
Agglomerative cosine	CLAP; cymbal-median, *cymbal_type*	AMI = 0.3362 ARI = 0.2182 V = 0.3783	Probes cosine-compatible hierarchical partitioning and often gives the clearest cymbal-level family partitions	No noise by design	Clearest cymbal-level family partition
HDBSCAN	MERT; sample, *cymbal_type*	AMI = 0.2606 ARI = 0.0735 V = 0.2800	Probes density-based structure and often yields lower alignment or only apparently favorable solutions under high noise and fragmentation	Noise = 0.6082; 59 clusters	Diffuse, high-noise, fragmented alignment

This algorithmic contrast helps explain why Stage 5 is often more conservative than Stage 4. A target may be strongly recoverable under leakage-safe kNN probing and still fail to form equally strong unsupervised clusters if its information is encoded primarily through locally coherent neighborhoods rather than sharply separated global partitions. Stage 5 therefore clarifies that recoverability does not necessarily imply strong intrinsic cluster emergence. This is especially relevant for the strike-related targets, whose cross-stage pattern increasingly supports the presence of partial substructure embedded within broader acoustic manifolds rather than fully discrete trait-specific islands.

Overall, from the perspective of the clustering algorithms, the Stage 5 results show that neither CLAP nor MERT organizes the targets as uniformly compact, low-noise unsupervised clusters. Instead, the algorithmic comparison points to a mixture of moderate partitionability and manifold-like organization, whose relative balance varies by target and representation level. This contrast is illustrated in Figure 17, where a relatively coherent CLAP cymbal-median agglomerative partition is set against a diffuse, high-noise MERT sample-level HDBSCAN configuration. Under this reading, partition-based methods are more informative for identifying weak-to-moderate but interpretable label alignment, whereas HDBSCAN is especially useful as a diagnostic of where the structure remains diffuse, fragmented, or insufficiently density-supported to form clean unsupervised clusters.

 doi:10.20944/preprints202603.1837.v1

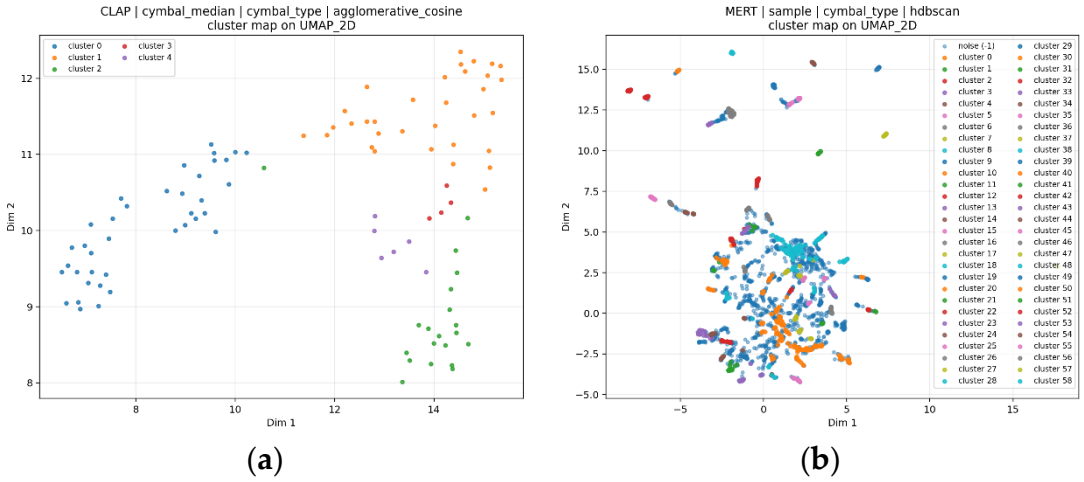


Figure 17. Comparative examples of partition-like and manifold-like organization in Stage 5 *cymbal_type* for **(a)** CLAP and **(b)** MERT.

5.3.4. Robustness Across Representation Levels: Sample vs Cymbal-Level Analyses

The comparison across representation levels serves as a meaningful robustness filter for distinguishing stable instrument-level organization from sample-local effects tied to repeated strikes. Under this view, *cymbal_type* became more coherent at the cymbal level than at the raw sample level, indicating that macro-family structure is more stable as a property of the instrument itself than as a partition at the level of individual strikes.

As shown in Table 16, at the aggregated cymbal level, examined through both cymbal-mean and cymbal-median representations, *cymbal_type* became more coherent than at the raw sample level. This pattern was especially clear in CLAP, where agglomerative cosine rose from AMI = 0.1751 at sample level to 0.3362 at both aggregated levels. In MERT, the aggregated agglomerative solutions were not uniformly higher than the best sample-level HDBSCAN result (AMI = 0.2606), but they were structurally cleaner and more interpretable because they avoided the extreme noise and fragmentation of the sample-level density-based solution. This supports the view that part of the apparent instability at sample level is introduced by strike-level variability rather than by the absence of family-level structure itself.

Table 16. Cross-level interpretability and stability of Stage 5 clustering alignment across sample- and cymbal-level representations.

Target	CLAP strongest algorithm	MERT strongest algorithm	Cross-level status	Interpretive note
cymbal_type	Agglom. cosine (Sample): AMI = 0.1751 Agglom. cosine (Cymbal-mean): AMI = 0.3362 Agglom. cosine (Cymbal-median): AMI = 0.3362	HDBSCAN (Sample): AMI = 0.2606 Agglom. cosine (Cymbal-mean): AMI = 0.2624 Agglom. cosine (Cymbal-median): AMI = 0.2492	Stable; clearer at cymbal level	Family structure persists across levels and becomes more coherent after aggregation, especially in CLAP.
hit_zone	Agglom. cosine (Sample): AMI = 0.0668	Agglom. cosine (Sample): AMI = 0.0667	Not interpretable after aggregation	Sample-native target; omitted at cymbal level because it becomes non-unique.

 doi:10.20944/preprints202603.1837.v1

stick_material	HDBSCAN (mcs = 15) (Sample): AMI = 0.0482	Spherical k-means (Sample): AMI = 0.0946	Not interpretable after aggregation	Valid at strike level only; cymbal-level omission is methodological.
stick_type	HDBSCAN (mcs = 20) (Sample): AMI = 0.0259	Agglom. cosine (Sample): AMI = 0.0177	Not interpretable after aggregation	Fine-grained strike label; not meaningful after grouping by cymbal_id.
hit_point	Spherical k-means (Sample): AMI = 0.0735	Spherical k-means (Sample): AMI = 0.0847	Not interpretable after aggregation	Sample-local structure only; not a valid cymbal-level target.

By contrast, *hit_zone*, *stick_material*, *stick_type*, and *hit_point* were omitted at the cymbal-mean and cymbal-median levels because, after aggregation across multiple strikes of the same instrument, they no longer correspond to a single unique and semantically valid value per cymbal. Their absence at cymbal level should therefore not be interpreted as an ordinary null result or simple information loss, but as a methodological consequence of target invalidity after aggregation. In this sense, Stage 5 distinguishes sample-local organization from more stable cymbal-level structure, rather than treating all targets as equally valid across representation levels.

Taken together, *cymbal_type* benefits from stabilization at the cymbal level, reinforcing its status as a macro-structural and instrument-level property, whereas the strike-related labels are better understood as sample-native structure that is not expected to survive unchanged after grouping by *cymbal_id*. Under this reading, the contrast between sample and cymbal level serves not only as a robustness check, but also to define the proper interpretive scope of the results by distinguishing stable instrument-level clustering from structure that remains meaningful primarily at the level of individual strikes.

5.3.5. Diameter as a Confound-Sensitive Diagnostic Case

The relevant Stage 5 question for diameter, as a confound-sensitive diagnostic target, was not whether some global clustering alignment could be observed, but whether such a signal remained interpretable once the dominant effect of cymbal family was constrained.

At the global cymbal-level view, diameter still yields non-trivial clustering alignment in both models. In the cymbal-median results, CLAP reaches AMI = 0.2517 with spherical k-means and 0.2285 with agglomerative cosine, whereas MERT reaches 0.2232 and 0.2526, respectively, under the same methods. These values indicate that some size-related structure is present at the unrestricted global level. However, they must be interpreted cautiously, because the diagnostic question is not whether global diameter alignment exists at all, but whether it remains interpretable once the stronger family scaffold is held fixed. This contrast between global and within-type behavior is summarized in Table 17.

Table 17. Cross Global versus within-type clustering alignment for diameter.

Model	Global diameter alignment	Within-type behavior	Within-type coverage	Robust across families?	Final interpretation
CLAP	Spherical k-means (cymbal_median): AMI = 0.2517	Limited and inconsistent	Sparse and family-limited	No	Global size-related alignment is present, but

 doi:10.20944/preprints202603.1837.v1

MERT	Agglomerative cosine (cymbal_median): AMI = 0.2285 Spherical k-means (cymbal_median): AMI = 0.2232 Agglomerative cosine (cymbal_median): AMI = 0.2526	Mixed and uneven	Sparse, mixed, and family-dependent	No	not robust after family control. Global diameter structure is not supported as a stable family-independent dimension.

The within-type branch is therefore the more decisive diagnostic component. If diameter were represented more independently, some stable alignment would be expected to persist within fixed cymbal families. Instead, the observed pattern is non-robust, with some families omitted in aggregated settings and no clear evidence of a consistent cross-family rule. This does not mean that size-related information is entirely absent from the embeddings; rather, it suggests that such information is not sufficiently organized to support reliable unsupervised clustering once the broader timbral-family scaffold is removed. A representative within-type example (Figure 18) is consistent with this reading, showing that residual size-related alignment, when present, remains partial and family-specific rather than clean or generalizable across families.

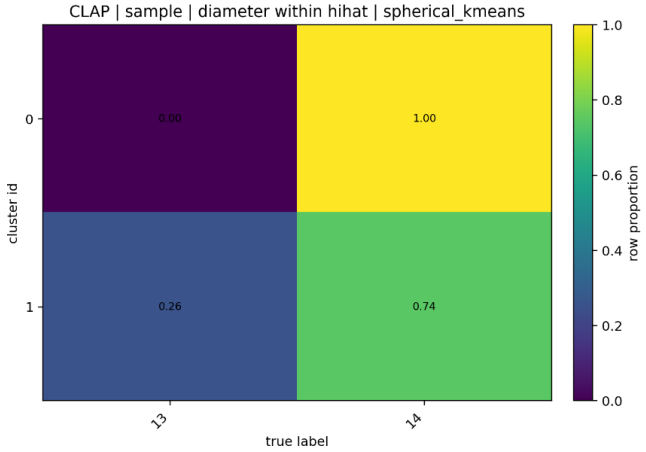


Figure 18. Within-type clustering example for diameter under family-restricted evaluation.

Taken together, Stage 5 provides only limited support for the claim that either model encodes physical size as a stable and independently clusterable dimension. The most defensible conclusion is that some apparent global diameter signal exists, but much of it remains plausibly mediated by *cymbal_type*, non-robust, or family-dependent

6. Discussion

The discussion section is organized into three subsections. The first, Cross-Stage Evidence Synthesis, presents the consolidated findings of the evaluation framework. The second explicitly addresses the constraints and limitations of the study, and the final subsection discusses the implications for the target domain (musical acoustics, with a focus on cymbal sounds) and outlines directions for future research.

6.1. Summary: Cross-Stage Evidence Synthesis

The combined reading of Stages 3–5 shows that the two embedding families organize cymbal-related information differently, and that the CLAP–MERT comparison cannot be reduced to a single global winner. Instead, a clearly target-dependent representational profile emerges: MERT retains a small but consistent advantage on the primary macro-family target (*cymbal_type*), whereas CLAP is

 doi:10.20944/preprints202603.1837.v1

more strongly supported on the strike-related targets (*hit_zone*, *stick_material*, *stick_type*, and *hit_point*). By contrast, *diameter* remains a diagnostic and confound-sensitive case rather than a stable comparative endpoint.

Table 18 summarizes the integrated interpretation of the main targets across the three representational analysis layers of the study. For each target, it reports the Stage 4 comparative winner, the Stage 4 evidence pattern, within-type persistence where methodologically applicable, the degree of agreement with Stage 3, the main clarification sought from Stage 5, the principal Stage 5 finding, and the final integrated interpretation after combining geometric inspection, grouped leakage-safe recoverability, and unsupervised clustering alignment.

The clearest integrated pattern concerns the strike-related targets. Here, the findings converge in favor of CLAP, while also constraining how this advantage should be interpreted: it does not reflect the clean unsupervised emergence of discrete categories, but rather stronger local organization, that is, a more coherent and partially label-aligned local substructure. In other words, the advantage of CLAP lies more in the local representational coherence of strike-related attributes than in the formation of clean global clusters.

The most limited and ambiguous picture concerns *diameter*. Across all stages, the relevant signal appears weaker, more heterogeneous, and more vulnerable to confounding factors than either *cymbal_type* or the strike-related labels. The safest interpretation is that some size-related information may be present, but remains largely type-mediated, non-robust, or family-dependent, rather than constituting a clean and independently organized representational dimension.

The case of *cymbal_type* remains more complex and represents the main interpretive tension of the framework. Family-level information is present in both models, but not in the form of uniformly clean, low-noise clustering. The combined reading suggests that MERT retains a narrow advantage in grouped recoverability for *cymbal_type*, whereas CLAP shows cleaner unsupervised family partitioning under selected conditions, particularly at the cymbal level. The main conclusion, therefore, is not which model "wins" overall on *cymbal_type*, but rather that the two models preserve different forms of macro-family structure.

Table 18. Cross-stage synthesis of target-wise findings across all stages.

Target	Stage 4 winner	Stage 4 evidence	Within-type persistence	Stage 3 agreement	Stage 5 question	Stage 5 finding	Integrated interpretation
cymbal_type	**MERT**	Modest but consistent grouped advantage	Not a Stage 4 subgroup control; clearer at cymbal level in Stage 5	**Partial / mixed**	Probe-level edge vs. intrinsic family clustering	Both models carry type information; CLAP cleaner at cymbal level, MERT stronger in selected probing settings	**Split result**: MERT for grouped recoverability, CLAP for cleaner unsupervised family partitioning
hit_zone	**CLAP**	Clear best-K	**Yes**; stronger	**Strong**	Local neighborhoo	Weak global clustering;	**Robust CLAP**

 doi:10.20944/preprints202603.1837.v1

		advantage; positive paired deltas	in crash, hi-hat, ride; near-equal in splash		ds vs. unsupervised subclusters	stronger local structure	**advantage**, mainly local rather than cluster-based
stick_material	**CLAP**	Clear best-K advantage; positive paired deltas	**Yes**; stronger across available families	**Strong**	Material-related clustering within family	Limited unsupervised alignment	**Strong CLAP recoverability**, but weak intrinsic clustering
stick_type	**CLAP**	Strong best-K advantage; positive paired deltas	**Yes**, where evaluable	**Strong**	Recoverability vs. unsupervised implement clustering	Limited cluster emergence	**Consistent CLAP advantage**, mostly in recoverability
hit_point	**CLAP**	Clear relative advantage; lower absolute scores	**Yes**, but weaker overall	**Strong**, for a harder target	Whether weak recoverability reflects diffuse structure	Weak and diffuse signal	**Supporting CLAP result**, but low intrinsic clustering strength
diameter	**No stable winner**	Small signal; partial fold validity; mixed pattern	**No stable cross-family pattern**	**Mixed / confound-sensitive**	Residual within-type size structure vs. type-mediated signal	Some global alignment; weak and mixed within-type structure	**Diagnostic only**; not a stable comparative endpoint

At the methodological level, the main contribution of the framework lies in its ability to distinguish systematically among geometric readability, supervised recoverability, and intrinsic clustering structure as related but distinct representational properties. In this sense, the proposed multi-stage scheme does not merely summarize individual findings but also functions as a safeguard against over-interpretation by enabling a more precise and controlled synthesis of what is confirmed, what is qualified, and what remains unresolved.

6.2. Constraints and Critical Reflections

Despite the methodological care taken in the present framework, several limitations should be acknowledged explicitly. First, the dataset originates from a single curated commercial source. This strengthens the internal comparability of the CLAP–MERT comparison but limits the external

 doi:10.20944/preprints202603.1837.v1

validity of the findings to alternative cymbal corpora, potentially assembled through different recording conditions, or less controlled acquisition environments. In this sense, the present framework should be understood primarily as a controlled comparative setting rather than as a universal benchmark of generalization.

Second, the standardization of all samples to a fixed 10 s window should be regarded as a controlled compromise adopted for fairness-aware comparison, rather than as a universally neutral choice for all model families. This shared constraint improves comparability, but it is applied to models with different native temporal assumptions, as CLAP operates natively on 10 s windows at 48 kHz, whereas MERT was designed for longer temporal context at 24 kHz. The present evaluation therefore reflects a controlled fairness design rather than necessarily the optimal operational regime of each model.

Third, Stage 2 should be interpreted as an internal model-selection step rather than as an independent confirmatory layer. The sweep procedure was methodologically necessary to compare the two embedding families under best-versus-best configurations. At the same time, the final configurations were determined within the same general evaluation regime rather than through a fully nested procedure. Accordingly, Stage 2 strengthens the internal validity of the comparison, but does not function as a fully independent evidential layer in the same way as Stages 3–5.

Taken together, these limitations do not undermine the value of the framework, but they do define the scope of its claims more precisely. The study proposes a methodologically transparent scheme for the representation audit of Audio Foundation Models (AFMs) in a controlled setting, without claiming to constitute a final benchmark for universal AFM ranking.

6.3. Perceptual Implications and Future Work

Beyond the comparative findings themselves, the present study also carries implications of how AFM embeddings may be used to trace perceptually meaningful properties of cymbal sound. The results suggest that frozen embeddings can preserve information related to perceptually meaningful sound qualities, particularly when these are associated with performance technique and subtle variations in timbre. In the cymbal domain, where the physical identity of the instrument, the manner of performance, and musical perception are closely intertwined, the systematic differentiation observed for variables related to playing and striking behavior indicates that the representations capture not only broad categories, but also aspects of sound linked to both its production and its perception.

At the same time, the study provides a clearer indication of how this research line may be extended. Frozen embeddings constitute a strong starting point for representation audit and initial mapping of the latent space, but not necessarily the final level of modeling for perceptual sound qualities. The present findings suggest that some signals are captured mainly as local structure rather than as fully emergent categories, while others remain more sensitive to confounds. This suggests that a next step could examine forms of task-aware adaptation, that is, targeted adjustment of embeddings to the requirements of specific perceptual or acoustical objectives, through light fine-tuning or Parameter-Efficient Fine-Tuning (PEFT) approaches, such as LoRA-based adaptation, when a more direct link between the representations and perceptual descriptors is required.

Within the same perspective, further progress is also likely to require richer target variables. The variables used here were appropriate for describing macro-family identity, strike-related properties, and a controlled diagnostic target; however, the next step could also turn toward more direct perceptually grounded descriptors and perceptual annotations, such as brightness, dryness, wash, attack sharpness, or stick definition. Such a shift would bring the evaluation of embeddings closer to the way musicians hear, describe, and compare cymbals. Without such human-centered labels, the analysis remains closer to physical metadata and acoustically motivated proxy targets. By contrast, the inclusion of musically and perceptually grounded annotations would make it possible to examine more directly whether AFM embeddings reflect not only what a cymbal is or how it was struck, but also how it sounds and how it is described by musicians.

 doi:10.20944/preprints202603.1837.v1

Finally, a natural direction for future work is the expansion of the experimental framework itself: more datasets, different recording chains, broader coverage of brands and performance contexts, and cross-dataset validation. At the same time, it would be valuable to incorporate additional AFMs, suitable baseline systems, and a more formal inferential layer, including paired significance testing, effect sizes, and multiple-comparison correction, so that the statistical assessment of model differences can become more rigorous and systematic. In this sense, the present study does not close the question of how AFM embeddings capture the acoustic and perceptual properties of cymbals but rather helps define the next step more clearly: a transition from frozen diagnostic mapping toward more perceptually grounded, annotation-rich, and methodologically expanded evaluation frameworks.

7. Conclusions

This study presented a multi-stage comparison of CLAP and MERT for cymbal classification, combining geometric inspection, leakage-safe grouped probing, and unsupervised clustering analysis. The results showed that the comparison cannot be reduced to a single overall winner but instead reveals a clear target-dependent representational profile: MERT retained a small but consistent advantage on cymbal family type, whereas CLAP was more strongly supported on the strike-related targets.

More broadly, the study showed that geometric readability, supervised recoverability, and intrinsic clustering structure should not be treated as equivalent properties. Under this reading, the advantage of CLAP is best understood mainly as stronger local representational organization for strike-related information, whereas the advantage of MERT on *cymbal_type* remains primarily a grouped probing effect rather than evidence of uniformly cleaner family-level clustering. Diameter, in turn, proved to be a confound-sensitive diagnostic case rather than a stable comparative endpoint.

Overall, the proposed framework offers a methodologically transparent way to evaluate how Audio Foundation Model embeddings organize musically meaningful information in a controlled setting. Rather than claiming universal AFM superiority, the present work provides a more differentiated account of what each embedding family captures and how that information is structurally organized.

Author Contributions: Conceptualization, Michael Starakis and Chrisoula Alexandraki; methodology, Michael Starakis and Maximos Kaliakatso-Papakostas; validation, Michael Starakis, Chrisoula Alexandraki and Maximos Kaliakatso-Papakostas; formal analysis, Michael Starakis; investigation, Michael Starakis; data curation, Michael Starakis; writing—review and editing Michael Starakis and Chrisoula Alexandraki; visualization, Michael Starakis; supervision, Chrisoula Alexandraki and Maximos Kaliakatso-Papakostas; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The cymbal sound dataset used in this study contains proprietary audio recordings that cannot be publicly shared. However, the associated metadata, including file identifiers, labels, physical instrument characteristics, and annotation information, are available upon reasonable request from the corresponding author. Researchers seeking to reproduce or extend the analyses reported in this manuscript may access the metadata under the same conditions.

Acknowledgments: The authors acknowledge Audio Animals Ltd for providing the cymbal sound dataset at a reduced price for research purposes. The company had no role in the study design, data analysis, or interpretation of results.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202603.1837.v1

Abbreviations

The following abbreviations are used in this manuscript:

AFMs	Audio Foundation Models
ARCH	Audio Representation Benchmark
CLAP	Contrastive Language-Audio Pretraining
DR	Dimensionality Reduction
HEAR	Holistic Evaluation of Audio Representations
LDA	Linear Discriminant Analysis
MARBLE	Music Audio Representation Benchmark for Universal Evaluation
MCC	Matthews Correlation Coefficient
MERT	Acoustic Music Understanding Model With Large-Scale Self-Supervised Training
MMAU	Massive Multi-Task Audio Understanding
PCA	Principal Components Analysis
UMAP	Uniform Manifold Approximation and Projection

Appendix A

Figure A1 provides a schematic overview of the five-stage comparative evaluation framework, conceived to enable a valid comparison of CLAP and MERT. The diagram is intended to assist the reader in navigating the methodological stages described in Section 4 and the corresponding results in Section 5. Due to its complexity and length, the figure could not be included in the main text.

 doi:10.20944/preprints202603.1837.v1

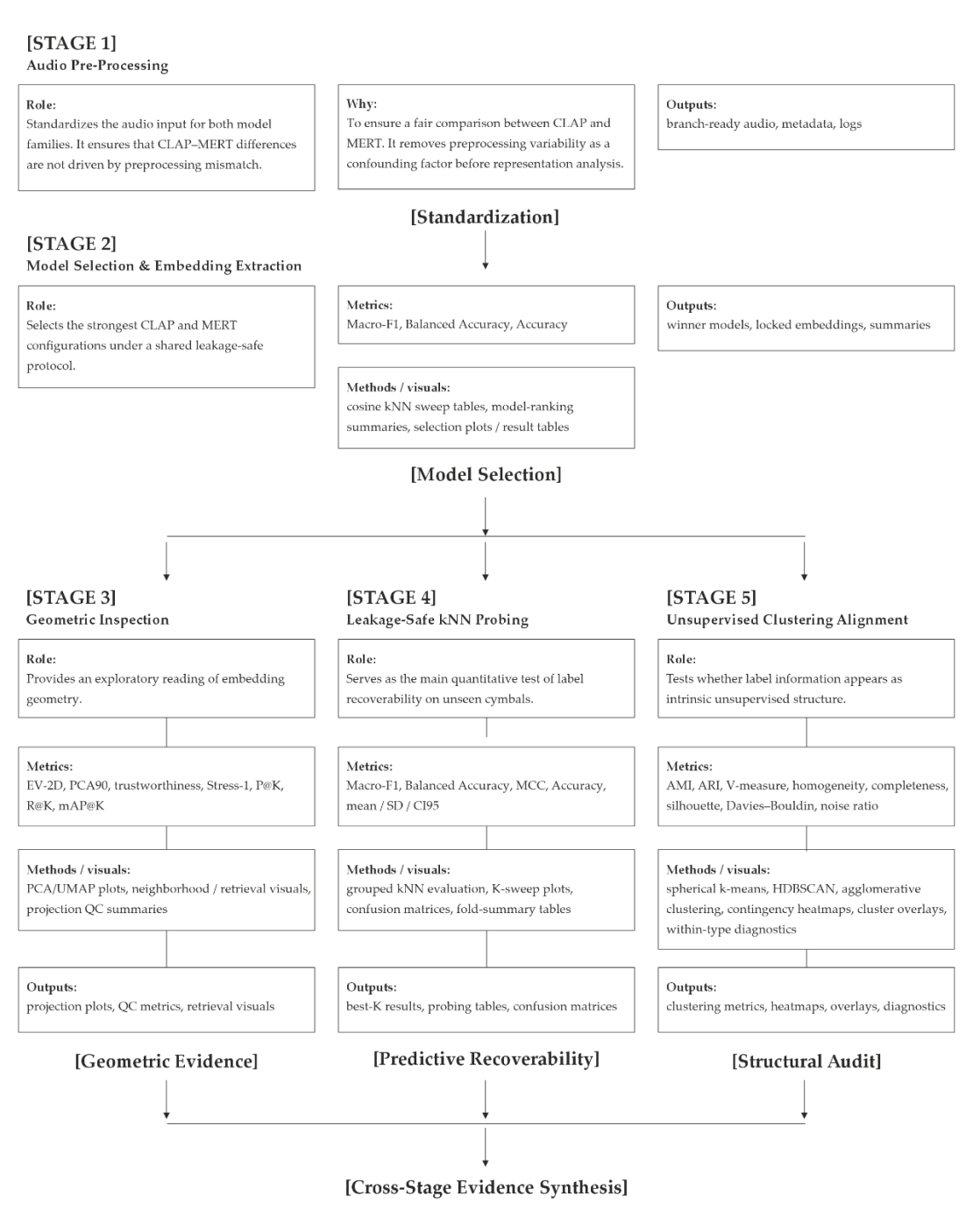


Figure A1. The five-stage methodological framework followed for the comparison of CLAP and MERT. Each stage is combined with information concerning its role within the framework, the evaluation metrics the visualization tools, its output as well as its purpose, namely Standardization, Model Selection, Geometric Evidence, Predictive recoverability and Structural Audit. Collectively the framework proceeds towards Cross-Stage Evidence Synthesis, representing the consolidated result of the comparison.

 doi:10.20944/preprints202603.1837.v1

References

1. Shi, Q.; Zhou, J.; Lin, B.; Cui, J.; Zeng, G.; Zhou, Y.; Wang, Z.; Liu, X.; Luo, Z.; Wang, Y.; Liu, Z. UltraEval-Audio: A Unified Framework for Comprehensive Evaluation of Audio Foundation Models. arXiv 2026. https://doi.org/10.48550/ARXIV.2601.01373.
2. Yang, S.; Chang, H.-J.; Huang, Z.; Liu, A. T.; Lai, C.-I.; Wu, H.; Shi, J.; Chang, X.; Tsai, H.-S.; Huang, W.-C.; Feng, T.; Chi, P.-H.; Lin, Y. Y.; Chuang, Y.-S.; Huang, T.-H.; Tseng, W.-C.; Lakhotia, K.; Li, S.-W.; Mohamed, A.; Watanabe, S.; Lee, H. A Large-Scale Evaluation of Speech Foundation Models. IEEE/ACM Trans. Audio Speech Lang. Process. 2024, 32, 2884–2899. https://doi.org/10.1109/TASLP.2024.3389631.
3. Ma, Y.; Øland, A.; Ragni, A.; Del Sette, B. M.; Saitis, C.; Donahue, C.; Lin, C.; Plachouras, C.; Benetos, E.; Shatri, E.; Morreale, F.; Zhang, G.; Fazekas, G.; Xia, G.; Zhang, H.; Manco, I.; Huang, J.; Guinot, J.; Lin, L.; Marinelli, L.; Lam, M. W. Y.; Sharma, M.; Kong, Q.; Dannenberg, R. B.; Yuan, R.; Wu, S.; Wu, S.-L.; Dai, S.; Lei, S.; Kang, S.; Dixon, S.; Chen, W.; Huang, W.; Du, X.; Qu, X.; Tan, X.; Li, Y.; Tian, Z.; Wu, Z.; Wu, Z.; Ma, Z.; Wang, Z. Foundation Models for Music: A Survey. arXiv 2024. https://doi.org/10.48550/ARXIV.2408.14340.
4. Turian, J.; Shier, J.; Khan, H. R.; Raj, B.; Schuller, B. W.; Steinmetz, C. J.; Malloy, C.; Tzanetakis, G.; Velarde, G.; McNally, K.; Henry, M.; Pinto, N.; Noufi, C.; Clough, C.; Herremans, D.; Fonseca, E.; Engel, J.; Salamon, J.; Esling, P.; Manocha, P.; Watanabe, S.; Jin, Z.; Bisk, Y. HEAR: Holistic Evaluation of Audio Representations. arXiv May 29, 2022. https://doi.org/10.48550/arXiv.2203.03022.
5. Yuan, R.; Ma, Y.; Li, Y.; Zhang, G.; Chen, X.; Yin, H.; Zhuo, L.; Liu, Y.; Huang, J.; Tian, Z.; Deng, B.; Wang, N.; Lin, C.; Benetos, E.; Ragni, A.; Gyenge, N.; Dannenberg, R.; Chen, W.; Xia, G.; Xue, W.; Liu, S.; Wang, S.; Liu, R.; Guo, Y.; Fu, J. MARBLE: Music Audio Representation Benchmark for Universal Evaluation. arXiv 2023. https://doi.org/10.48550/ARXIV.2306.10548.
6. Prabavathy, S. Classification of Musical Instruments Sound Using Pre-Trained Model with Machine Learning Techniques. AJES 2020, 9 (1), 45–48. https://doi.org/10.51983/ajes-2020.9.1.2369.
7. Werner, K.; Vergara, E. F.; Paul, S.; Cordioli, J. A. Timbre Aspects of Ride Cymbals: Sound Coloration Analysis Using Psychoacoustic Models and Subjective Evaluation; Salt Lake City, Utah, 2015; p 035004. https://doi.org/10.1121/2.0000224.
8. Kaselouris, E.; Paschalidou, S.; Alexandraki, C.; Dimitriou, V. FEM-BEM Vibroacoustic Simulations of Motion Driven Cymbal-Drumstick Interactions. Acoustics 2023, 5 (1), 165–176. https://doi.org/10.3390/acoustics5010010.
9. Brezas, S.; Kaselouris, E.; Orphanos, Y.; Bakarezos, M.; Papadogiannis, N. A.; Dimitriou, V. On the Correlation of Cymbals' Vibrational Behavior and Manufacturing Processes. Applied Sciences 2025, 15 (3), 1425. https://doi.org/10.3390/app15031425.
10. Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; Dubnov, S. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Rhodes Island, Greece, 2023; pp 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095969.
11. Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; Gyenge, N.; Dannenberg, R.; Liu, R.; Chen, W.; Xia, G.; Shi, Y.; Huang, W.; Wang, Z.; Guo, Y.; Fu, J. MERT: Acoustic Music Understanding Model with Large-Scale Self-Supervised Training. In The Twelfth International Conference on Learning Representations; 2024.
12. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv October 22, 2020. https://doi.org/10.48550/arXiv.2006.11477.
13. Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language. arXiv October 25, 2022. https://doi.org/10.48550/arXiv.2202.03555.
14. Guzhov, A.; Raue, F.; Hees, J.; Dengel, A. AudioCLIP: Extending CLIP to Image, Text and Audio. arXiv June 24, 2021. https://doi.org/10.48550/arXiv.2106.13043.
15. Parulekar, A.; Collins, L.; Shanmugam, K.; Mokhtari, A.; Shakkottai, S. InfoNCE Loss Provably Learns Cluster-Preserving Representations. arXiv February 15, 2023. https://doi.org/10.48550/arXiv.2302.07920.

 doi:10.20944/preprints202603.1837.v1

16. Quatra, M. L.; Koudounas, A.; Vaiani, L.; Baralis, E.; Cagliero, L.; Garza, P.; Siniscalchi, S. M. Benchmarking Representations for Speech, Music, and Acoustic Events. In 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW); 2024; pp 505–509. https://doi.org/10.1109/ICASSPW62465.2024.10625960.
17. Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; Manocha, D. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. arXiv October 24, 2024. https://doi.org/10.48550/arXiv.2410.19168.
18. Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; Glass, J. Listen, Think, and Understand. arXiv 2023. https://doi.org/10.48550/ARXIV.2305.10790.
19. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and Music Signal Analysis in Python; Austin, Texas, 2015; pp 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003.
20. Papaioannou, C.; Benetos, E.; Potamianos, A. Universal Music Representations? Evaluating Foundation Models on World Music Corpora. arXiv June 20, 2025. https://doi.org/10.48550/arXiv.2506.17055.