

Article

Dimension-Wise Probing of Music Foundation Models for Chinese and Western Music: Timbre, Technique, Key, and Mode

Xueying Bai ¹, Nanmu Hui ², Yanping Li ² and Xiaowei Han ^{3,*}¹ School of Music and Media, Shenyang University, Shenyang 110044, China; xueying@syu.edu.cn² School of Intelligent Science and Information Engineering, Shenyang University, Shenyang 110044, China; huinm@syu.edu.cn (N.H.); 17004@syu.edu.cn (Y.L.)³ School of Mechanical Engineering, Shenyang University, Shenyang 110044, China

* Correspondence: hanxw@syu.edu.cn

Abstract

Music foundation models are increasingly used as feature extractors in music information retrieval; yet, their treatment of Chinese tonal organization has received little controlled evaluation. We probe frozen MERT, MuQ, CultureMERT, and an 88-dimensional librosa baseline on instrument timbre, guzheng playing technique, Western key, and Chinese pentatonic mode. Probes are evaluated with task-specific splits that guard against leakage and with chance-normalized accuracy. Timbre is near ceiling (0.994–0.998), whereas source-aware guzheng-technique accuracy ranges from 0.382 to 0.787. In the unmatched external comparison, GiantSteps key scores 0.441–0.604 and CNPM mode pattern scores 0.294–0.355; this difference is descriptive because the corpora differ in genre, production, provenance, and label distribution. Same-audio CNPM controls identify an important boundary: the TongGong system reaches 0.649–0.734, and absolute tonic reaches 0.433–0.480, whereas their matched five-way mode-pattern labels reach 0.256–0.350 and 0.243–0.366, respectively. Chroma ablations show that pitch-class energy is central to the external key task but does not by itself account for CNPM mode-pattern performance. These results map which musical dimensions are readily decodable from the evaluated frozen representations while avoiding a cultural-capability interpretation of the unmatched cross-corpus comparison.

Keywords: music foundation models; probing; music information retrieval; Chinese music; pentatonic mode; key estimation; representation analysis

1. Introduction

Machine learning now dominates music and audio signal processing, but successful systems still depend on how musical information is represented, how much annotated data is available, and whether learned features add value beyond strong signal-processing baselines [1–3]. In music information retrieval (MIR), self-supervised foundation models have become a default front-end: frozen embeddings from models such as MERT [4] and MuQ [5] rival or surpass task-specific systems on auto-tagging, genre, and instrument recognition. Their strong downstream performance has prompted a complementary line of work that asks what these representations actually encode, using probing classifiers on frozen features. So far, this has been examined almost entirely on Western music, where models have been shown to capture timbre, instrumentation, and a range of music-theoretic concepts. Far less is known about non-Western traditions, whose modal systems and tunings are under-represented in the Western-centric data these models are trained



Academic Editor: George A. Tsihrintzis

Received: 30 June 2026

Revised: 28 July 2026

Accepted: 30 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

on. Recent cross-cultural evaluations have begun to close this gap, but report aggregate tagging metrics that do not isolate which musical dimensions a model fails on or why, and Chinese music has remained largely outside their scope.

This study addresses this gap by probing three music foundation models (MERT, MuQ, and the multiculturally adapted CultureMERT), together with a hand-crafted librosa baseline, keeping every encoder frozen and training only shallow probes on their embeddings. Instead of reducing performance to a single aggregate score, we organize four classification tasks along two musical dimensions: surface attributes (instrument timbre and guzheng playing technique) and pitch-class organization (Western key and Chinese pentatonic mode). GiantSteps key and CNPM mode-pattern results form an external tonal comparison, although the EDM and Chinese-music corpora differ in production conditions, provenance, and label distribution. We complement this comparison with CNPM controls on identical audio subsets and piece-grouped folds, comparing the five-way mode-pattern target with the recorded TongGong system and absolute tonic. The design retains the observed external pattern while separating it from cultural attribution. We report chance-normalized accuracy, leakage-controlled cross-validation, layer-wise profiles, chroma ablations, and individual-model error analyses.

Across the tasks, surface attributes are recovered much more readily than tonal labels. Within CNPM, the TongGong system and absolute tonic are decoded more strongly than the five-way mode-pattern label, so a low pattern score should not be read as the absence of pitch information across all tonal targets. Per-model confusions and chroma ablations provide complementary views of the error structure. These findings persist across the architecture-native hidden states and are not reversed by the non-linear probe. CultureMERT is included as a fixed model comparison; because its continual-pre-training corpus contains Greek, Turkish, and Indian rather than Chinese music, our result does not test whether Chinese-targeted adaptation would help. All code, splits, configurations, and an environment specification will be made publicly available upon publication for reproducibility.

The main contributions of this study are as follows:

- **Cross-corpus dimension-wise probing.** We report Western-key and CNPM mode-pattern results with two surface-attribute tasks. The external comparison is not evidence of a cultural capability difference.
- **Protocol-matched CNPM tonal controls.** On identical CNPM recordings and piece-grouped folds, we compare mode pattern with the TongGong system and absolute tonic to identify what is specific to the five-way target.
- **A leakage trap in MIR evaluation, exposed and controlled.** On guzheng technique, chance-normalized accuracy falls from 0.95 under a random split to 0.59 under a leakage-controlled source-aware split, so part of the apparent performance reflects recording- and synthesis-specific fingerprints rather than the playing technique itself.
- **A cautious error diagnosis.** Individual-model confusions, chroma ablations, and a task-specific CNN contextualize the CNPM pattern result without treating a single confusion pattern as proof of an internal representation mechanism.
- **Data-driven versus hand-crafted features.** Chroma is essential for the external Western-key reference, whereas removing or retaining chroma alone does not account for the CNPM pattern result.
- **A dimension-wise probing framework.** Four tasks across two musical dimensions (surface attributes: instrument timbre, guzheng technique; pitch-class organization: Western key, Chinese pentatonic mode) let representation quality be read per dimension rather than as a single aggregate score.

The remainder of the article is organized as follows. Section 2 reviews probing, cross-cultural evaluation, and tonal analysis; Section 3 details the models, datasets, controls, and statistical protocol; Section 4 reports the results; and Sections 5 and 6 discuss their scope and implications.

2. Related Work

2.1. Probing Music Foundation Models

Self-supervised foundation models, including MERT [4], MusicFM [6], MuQ [5], and the generative Jukebox [7], now provide general-purpose audio representations that, with only a shallow probe, rival task-specific systems across MIR [8,9]. This shift is consistent with the broader MDPI music/audio literature, which frames contemporary MIR as a data-driven field while still emphasizing task-specific representations, annotation scarcity, and the continuing role of hand-crafted audio descriptors [1–3,10]. A parallel line of work asks what these frozen representations encode. SynTheory [11] builds a synthetic, copyright-free benchmark and finds that Jukebox and MusicGen encode music-theoretic concepts such as tempo, intervals, and chords to a degree that varies by layer and model size. Zhou et al. [12] perform a layer-wise analysis of MusicFM and MuQ, reporting that objective acoustic attributes (instrument, pitch) are recovered in lower layers while subjective semantic attributes (genre, emotion) emerge in higher layers. These studies establish probing as the standard lens on music representations, but evaluate almost exclusively on Western material and on isolated music-theoretic concepts rather than on the modal systems of specific non-Western traditions.

2.2. Cross-Cultural Evaluation and Adaptation

A growing body of work probes whether these gains extend beyond Western music. Papaioannou et al. [13] evaluate five foundation models across six corpora (Western, Greek, Turkish, Indian) and report lower aggregate tag-level scores in several non-Western, low-resource settings. This motivates culturally diverse evaluation, but the aggregate scores do not identify a particular musical dimension or isolate corpus effects. To mitigate such gaps, CultureMERT [14] continually pre-trains MERT on a 650-h Greek/Turkish/Indian mix and reports improved non-Western auto-tagging. Both works use aggregate tagging metrics (area under the receiver operating characteristic curve, ROC-AUC; mean average precision, mAP) that pool many tags; neither covers Chinese music, and CultureMERT's adaptation data contains no Chinese tradition. Our study is complementary: it decomposes Chinese-music results by musical dimension and includes CultureMERT as a fixed out-of-domain comparison, not as a test of Chinese-targeted adaptation.

2.3. Tonality, Mode, and Chinese Music Computing

Modeling tonal organization has a long history in MIR, from template-based key finding [15] to convolutional key estimation [16], where genre and tuning shifts are known to degrade accuracy. Adjacent MDPI work on music signal processing also shows that robust musical analysis often rests on carefully designed spectral or time-frequency representations before the downstream model is even chosen [10,17]. The Chinese pentatonic mode is a distinct system: the five modes Gong, Shang, Jue, Zhi, and Yu share a collection of pitches and differ in which degree serves as the tonic [18], so the modal label hinges on identifying the tonal center rather than the set of pitches. Computational work on Chinese music has so far targeted instrument and technique recognition [19] and pentatonic-mode classification with task-specific models [18], much of it built on the recently consolidated CCMusic database [20]. Symbolic-domain evidence echoes a tonal weakness: MoGE [21] reports that MIDIBERT fails to encode emotion–mode correlations and must inject explicit

mode features to recover them. To our knowledge, no prior work probes whether audio foundation models encode Chinese pentatonic mode while separating it from additional tonal labels on the same recordings.

2.4. What Probes Do and Do Not Tell Us

Probing accuracy is a measure of how readily a property can be linearly decoded from frozen features, not a direct readout of what the model knows [22]. Two caveats are central to our interpretation. First, a successful probe shows a property is decodable, but a failing probe is weaker evidence that the property is absent: the information might be present yet non-linearly encoded [23], a concern raised specifically for music, where chord and pitch concepts may not be linearly separable without intervention in the hidden space [24]. Second, probe accuracy is sensitive to probe capacity, class balance, and data splits [22], and high accuracy can arise from memorizing spurious correlations rather than the target property [23]. We address both directly: we pair linear probes with a non-linear (gradient-boosted) probe and find the same low ceiling on tonal tasks; we report chance-normalized accuracy with leakage-controlled splits; and we read the per-class confusion structure rather than accuracy alone, so that our claims rest on how errors are distributed, not on a single number whose level could be a probe artifact.

The closest methodological comparisons address complementary, rather than interchangeable, questions. SynTheory [11] establishes concept- and layer-wise probing with controlled synthetic Western-style inputs, whereas Zhou et al. [12] map layer-dependent acoustic and semantic profiles in MusicFM and MuQ. Neither study tests real Chinese recordings, Chinese pentatonic mode, or more than one tonal label on the same audio. Conversely, Papaioannou et al. [13] provide a broader cross-cultural tag-level evaluation, but their aggregate scores vary jointly with tag vocabulary, corpus, and musical dimension. CultureMERT [14] studies continued pre-training on Greek, Turkish, and Indian music through auto-tagging; because its adaptation and evaluation settings exclude Chinese music, it is an informative out-of-domain model comparison here rather than evidence of Chinese-targeted adaptation.

Among Chinese-music studies, Wang et al. [18] use a task-specific CNN to recognize Chinese pentatonic labels. That work establishes the relevance of supervised mode classification, while the present study asks a different question: what can be decoded from frozen general-purpose representations under five-way mode-pattern labels and piece-grouped splits? Its target construction and split protocol therefore prevent a direct numerical ranking. Taken together, these studies motivate the present design: an external Western-key reference is retained to map dimensions across evaluation settings, while same-audio CNPM controls distinguish mode pattern from other tonal labels without attributing a cross-corpus score difference to culture alone.

3. Materials and Methods

3.1. Overview

We treat each pre-trained model as a frozen feature extractor and train shallow probes on its embeddings, following standard practice in music representation analysis [8,11]. The same pipeline is applied to every (model, task) pair: extract per-layer embeddings, train a probe under a leakage-controlled cross-validation split, and report chance-normalized accuracy with bootstrap confidence intervals. Four tasks are grouped into two musical dimensions, surface attributes (instrument timbre, playing technique) and pitch-class organization (Western key, Chinese pentatonic mode), so that performance can be read per dimension rather than as a single aggregate score (Figure 1).

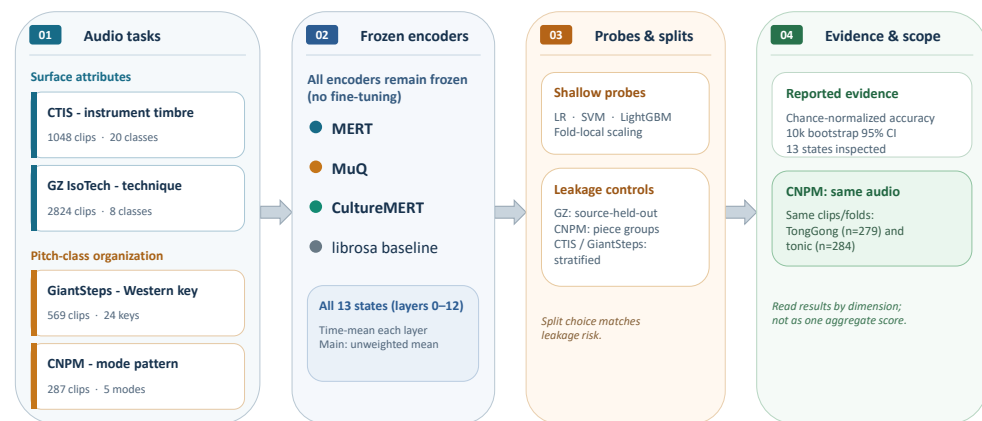


Figure 1. Study overview. Four tasks span surface attributes (instrument timbre and guzheng technique) and pitch-class organization (Western key and Chinese pentatonic mode pattern). Frozen MERT, MuQ, CultureMERT, and librosa features are pooled across all 13 architecture-native states and evaluated with task-specific splits. GiantSteps is an external tonal reference; the CNPM controls compare the mode pattern with the TongGong system and absolute tonic on identical clips and folds.

3.2. Models and Feature Extraction

We probe MERT (m-a-p/MERT-v1-95M, 768 dimensions), MuQ (OpenMuQ/MuQ-large-msd-iter, 1024 dimensions), CultureMERT (ntua-slp/CultureMERT-95M, 768 dimensions), and a hand-crafted librosa baseline. The 13 states are not tuned hyperparameters: all three released encoder implementations expose the embedding output plus 12 transformer-block outputs, and we retain this complete architecture-native set (layers 0–12). We neither add nor select layers. The single-layer analysis in Section 4.7 reports every available state, while the main feature is their unweighted mean. Table 1 specifies the complete downstream classification system; the pretrained encoders remain frozen in every reported experiment. The probe families are logistic regression, a linear support vector machine (SVM), and a Light Gradient Boosting Machine (LightGBM).

All audio is decoded to 24 kHz mono. Clips longer than 30 s are segmented into non-overlapping 30 s windows. To reduce artificial-boundary effects, a 20-ms raised-cosine taper is applied only at internal window edges (not at the original beginning or end); a final segment shorter than 1 s is discarded. Window-level feature vectors are then pooled by their original durations, so a short retained tail cannot receive the weight of a complete window. The same segmentation and duration pooling are applied to learned and librosa features. For each learned clip representation, we time-mean pool each hidden state, yielding 13 vectors $\mathbf{e}^{(0)}, \dots, \mathbf{e}^{(12)}$ of dimension D , and use the across-layer mean $\bar{\mathbf{e}} = (1/13) \sum_l \mathbf{e}^{(l)}$ for the main probe. The librosa baseline is an 88-dimensional vector containing the mean and standard deviation of 20 Mel-frequency cepstral coefficients (MFCCs), a 12-bin short-time Fourier transform (STFT) chromagram, seven spectral-contrast bands, and spectral centroid, bandwidth, roll-off, zero-crossing rate, and root-mean-square (RMS) energy; no temporal deltas are used. It supplies a transparent acoustic reference. Per-layer time pooling and the across-layer mean are

$$\mathbf{e}^{(l)} = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t^{(l)}, \quad l = 0, 1, \dots, 12; \quad \bar{\mathbf{e}} = \frac{1}{13} \sum_{l=0}^{12} \mathbf{e}^{(l)}, \quad (1)$$

where $\mathbf{h}_t^{(l)}$ is the hidden state of layer l at frame t and T is the number of frames. Within each cross-validation fold q , a scaler fitted only on its training set maps $\bar{\mathbf{e}}$ to $\tilde{\mathbf{e}}_q$, and a shallow

classifier g_{ϕ_q} fitted on that same training set produces $\hat{y} = g_{\phi_q}(\tilde{\mathbf{e}}_q)$. Thus no encoder, scaler, classifier, or model-selection quantity receives held-out labels.

Table 1. Frozen representation and classification configuration. All feature scaling and classifier fitting are performed independently inside each training fold.

Component	Configuration
Frozen encoders	MERT and CultureMERT: 13 states $\times$ 768 dimensions; MuQ: 13 states $\times$ 1024 dimensions. Each state is time-mean pooled, then the 13 state vectors are averaged with equal weights. No encoder parameter is updated.
Hand-crafted baseline	One 88-dimensional librosa vector (strict definition in Table 2), extracted per 30-s window and duration pooled exactly as for the learned features.
Fold-local preprocessing	StandardScaler: mean and standard deviation estimated on the training partition only, then applied to its held-out partition.
Logistic regression Linear SVM	Multiclass scikit-learn LogisticRegression(max_iter = 2000). scikit-learn LinearSVC(C = 1.0, max_iter = 10,000, dual = True).
Gradient-boosted trees	LGBMClassifier: multiclass objective; 300 trees; learning rate 0.1; 31 leaves; max depth –1; minimum child samples 20; row and feature subsampling 1.0; L1/L2 penalties 0; random seed 42; all available CPU workers; silent verbosity.
CNN reference (CNPM only)	One deterministic 10-s centre excerpt; 64-bin log-Mel input; convolutional blocks with 24/48/96 channels; global average pooling and a linear classifier; AdamW, 50 fixed epochs, three initialization seeds per held-out fold, and no validation-set model selection.

The 88-dimensional baseline uses librosa 0.10.2 [25]. For every scalar frame sequence $a_d(t)$, we concatenate its temporal mean and population standard deviation, $[\mu_d, \sigma_d]$, across the frames in each 30-s window. The precise descriptor families, dimensions, and parameters are given in Table 2; the released requirements file pins the library version so that these calls are not left to version-dependent defaults.

Table 2. Strict definition of the 88-dimensional librosa baseline at 24 kHz. Unless stated otherwise, spectral transforms use $n_{\text{fft}} = 2048$, a hop length of 512, a Hann window, centered frames, and constant padding. Each listed coefficient or scalar contributes a temporal mean and standard deviation.

Feature family	Dim.	Exact Configuration
MFCC	$20 \times 2 = 40$	20 coefficients from a 128-bin power Mel spectrogram ($f_{\min} = 0$, $f_{\max} = 12$ kHz, Slaney Mel normalization); type-II discrete cosine transform (DCT-II) with orthonormal normalization and lifter 0.
STFT chroma	$12 \times 2 = 24$	12 pitch-class bins; $n_{\text{fft}} = 2048$, hop 512, Hann window, ℓ_{∞} column normalization; tuning is estimated by librosa (tuning = None).
Spectral contrast	$7 \times 2 = 14$	Six octave bands ($f_{\min} = 200$ Hz), 2% quantile, logarithmic (not linear) contrast; the base band plus six bands give seven values per frame.
Spectral scalar descriptors	$5 \times 2 = 10$	Spectral centroid, spectral bandwidth ($p = 2$, normalized), 85% spectral roll-off, zero-crossing rate, and RMS energy. Zero-crossing rate and RMS use frame length 2048 and hop 512; zero-crossing frames use edge padding and RMS uses constant padding, matching librosa 0.10.2.
Total	88	No delta, delta-delta, learned, or label-derived feature is included.

3.3. Tasks and Datasets

The Chinese corpora are drawn from the open CCMusic database [20]. Surface-attribute tasks are (i) the CCMusic instrument-timbre (CTIS) dataset, for which we report the 20 most frequent instrument classes (1048 clips), and (ii) the CCMusic guzheng playing-technique (GZ_IsoTech; hereafter GZ) dataset, with 2824 clips in eight classes from a sampler library (KONTAKT, 1309), a human performer (yrs, 1019), and real recordings (record, 496). The CNPM mode-pattern task has 287 recordings in five labels: Gong (82), Yu (59), Zhi (52), Jue (49), and Shang (45). GiantSteps [26] provides 569 EDM recordings normalized to 24 major/minor keys. It is an external tonal reference, not a matched Chinese–Western control: genre, production, corpus provenance, and label distribution are all different.

CNPM provides additional annotations that permit a matched internal control. We retain TongGong-system classes with at least four clips (nine classes; $n = 279$) and absolute-tonic classes with at least four clips (ten classes; $n = 284$). For each retained subset, we evaluate both the control label and the five-way mode-pattern label on exactly the same clips, with the same four-fold piece-grouped splits. These comparisons test relative decodability of recorded CNPM labels, not whether one musical tradition is intrinsically easier than another.

3.4. Probing and Leakage-Controlled Evaluation

For each main task, we train three probes on features standardized from the training fold only: logistic regression (maximum 2000 iterations), linear SVM ($C = 1.0$, maximum 10,000 iterations), and LightGBM. LightGBM uses the fully specified configuration: multi-class objective, 300 trees, learning rate 0.1, 31 leaves, unlimited depth, minimum child size 20, unit row and feature subsampling, no L1/L2 regularization, seed 42, and all available CPU workers. The same three probes are used for GiantSteps. The main descriptive table reports the largest out-of-fold accuracy among the three and gives every probe in Appendix B; the matched CNPM controls and formal pairwise tests instead use pre-specified logistic regression, avoiding a second best-of-probes selection.

Evaluation uses leakage-controlled cross-validation, with the grouping chosen per dataset to prevent train–test contamination: guzheng technique uses leave-one-source-out (StratifiedGroupKFold grouped by source, three folds), plus a synthetic→real hold-out (train on sampler+performer and test on real recordings); pentatonic mode uses five-fold StratifiedGroupKFold grouped by piece title; GiantSteps uses four-fold stratified cross-validation because the rarest key class has four recordings; and CTIS uses five-fold stratified cross-validation. CTIS has no recording identifier, so its score is an upper bound that may be inflated by within-recording leakage. All splits are stratified by label and generated with seed 42.

3.5. Metrics

Because the four tasks differ in the number of classes (5 to 24), we report chance-normalized accuracy, which rescales accuracy so that chance maps to 0 and perfect accuracy to 1:

$$\text{acc}_{\text{norm}} = \frac{\text{acc} - 1/K}{1 - 1/K}, \quad (2)$$

where K is the number of classes. Accuracy confidence intervals are percentile intervals from 10,000 resamples of the pooled out-of-fold predictions (2.5th/97.5th percentiles); table subscripts are the corresponding bootstrap standard deviations. For the pre-specified same-audio CNPM representation contrasts, we report paired bootstrap intervals for the difference in chance-normalized accuracy and two-sided paired permutation p -values (10,000 sign permutations), with Benjamini–Hochberg correction across the stated family of

contrasts. Appendix A reports the effect differences, confidence intervals, raw p -values, and adjusted p -values. We do not infer significance from confidence-interval overlap. Mode confusions are reported separately for each model; their similarity is described rather than used to pool predictions.

4. Results

4.1. Dimension-Wise Results Across Chinese and Western Evaluation Sets

Table 3 provides the dimension-wise overview, and Figure 2 visualizes the MERT and MuQ subset. Timbre is near the ceiling for all representations, and the technique remains decodable under the source-aware split. The tonal tasks are lower in absolute chance-normalized accuracy. For every representation, GiantSteps Western-key accuracy is higher than CNPM mode-pattern accuracy in this evaluation. Because the corpora are unmatched in genre, production, provenance, and label distribution, we report this only as an external cross-corpus pattern. Table 4 provides the corresponding CNPM controls on identical audio and folds.

Table 3. Chance-normalized probe accuracy across four tasks (best probe; subscripts are bootstrap standard deviations, $B = 10,000$). Sample sizes: CTIS timbre 1048 clips (top-20 instrument classes), GZ technique 2824, GiantSteps key 569, and CNPM mode pattern 287. GZ uses the leakage-controlled source-aware split; CTIS has no recording-level grouping and is an upper bound. GiantSteps and CNPM are unmatched evaluation sets, so their difference is reported as a cross-corpus pattern. Table 4 gives CNPM controls on identical audio and folds; Appendix A reports formal within-label representation contrasts.

Model	CTIS Timbre	GZ Technique	GiantSteps Key (W)	CNPM Mode Pattern
MERT	0.997 _{0.002}	0.584 _{0.010}	0.516 _{0.022}	0.351 _{0.036}
MuQ	0.998 _{0.001}	0.787 _{0.008}	0.604 _{0.021}	0.355 _{0.037}
CultureMERT	0.994 _{0.002}	0.545 _{0.011}	0.441 _{0.022}	0.316 _{0.037}
librosa	0.996 _{0.002}	0.382 _{0.011}	0.461 _{0.022}	0.294 _{0.036}
Chance (normalized)	0.000	0.000	0.000	0.000
Raw chance (1/K)	1/20	1/8	1/24	1/5

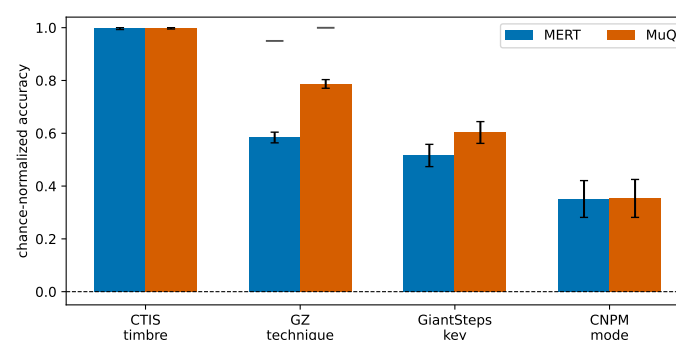


Figure 2. Dimension-wise overview. Chance-normalized accuracy of MERT and MuQ on the four tasks; error bars are bootstrap 95% intervals. For GZ, the bars show the source-aware value and the gray tick marks the random-split upper bound. GiantSteps and CNPM are unmatched evaluation sets, and their score difference is shown as an external comparison.

4.2. Part of the Apparent Performance Reflects Recording Cues

The guzheng-technique results show that high accuracy can partly reflect recording provenance rather than the musical attribute itself (Figure 3). All values in this subsection are chance-normalized. Under a random split, all models classify technique near-perfectly (MERT 0.95, MuQ 1.00, librosa 0.94), but this split lets clips from the same source appear in

both training and test. Under the leave-one-source-out (source-aware) split, accuracy drops sharply: MERT 0.58, MuQ 0.79, CultureMERT 0.55, with librosa falling furthest to 0.38. The official synthetic→real protocol (train on sampler and performer audio, test on real recordings) gives a similar picture (MERT 0.57, MuQ 0.70). The size of the drop, together with the fact that librosa (a pure acoustic feature) falls the most, indicates that part of the apparent technique recognition rides on source- and synthesis-specific acoustic fingerprints. We therefore use the fair source-aware values throughout (Table 3); even under this stricter protocol, the foundation models retain a real advantage over librosa on technique, unlike on the pitch tasks.

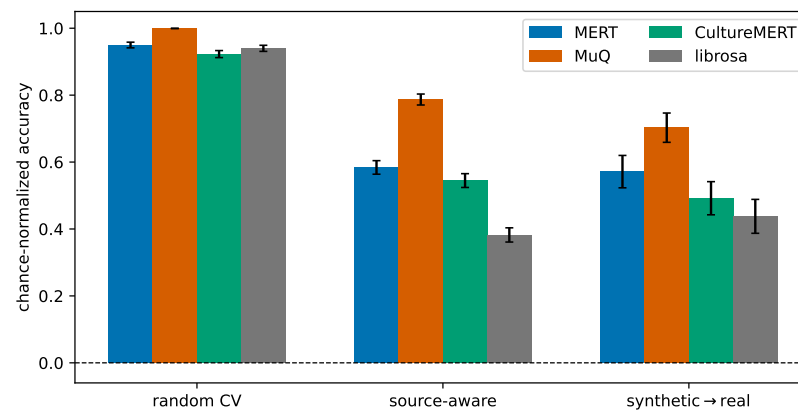


Figure 3. Source and synthesis cues in guzheng technique. Chance-normalized accuracy ($cn = 0$ marks chance, dashed line) is shown under three GZ_IsoTech protocols: random cross-validation (upper bound; same-source clips may share train and test), leave-one-source-out (source-aware), and the official synthetic→real hold-out. The large performance drop, especially for the librosa baseline, indicates that random splits capture source- or synthesis-specific acoustic fingerprints in addition to playing gesture; error bars are bootstrap 95% confidence intervals.

4.3. Same-Audio CNPM Controls Constrain the Tonal Interpretation

Table 4 provides the within-CNPM complement to the external tonal comparison. For each retained CNPM subset, control and pattern are learned from the same audio clips with the same piece-grouped folds. All four representations decode the TongGong system more strongly than the matched mode-pattern label (chance-normalized accuracy 0.65–0.73 versus 0.26–0.35), and the same ordering holds for absolute tonic versus its matched pattern subset (0.43–0.48 versus 0.24–0.37). These contrasts do not rank the musical constructs, because the targets differ in class count and annotation semantics. They do establish that the modest CNPM pattern score is not equivalent to an absence of tonal information in the frozen features and limit how the external score difference can be read.

Table 4. Same-audio CNPM controls using pre-specified logistic regression and four-fold piece-grouped cross-validation. Each pair of columns uses exactly the same clips; values are chance-normalized accuracy. The system has nine retained classes ($n = 279$), tonic ten ($n = 284$), and the pattern has five in both matched subsets. Bootstrap intervals ($B = 10,000$) are shown in Figure 4.

Model	TongGong System	Pattern (Same 279)	Absolute Tonic	Pattern (Same 284)
MERT	0.649	0.350	0.448	0.366
MuQ	0.734	0.350	0.480	0.366
CultureMERT	0.665	0.328	0.433	0.318
librosa	0.710	0.256	0.452	0.243

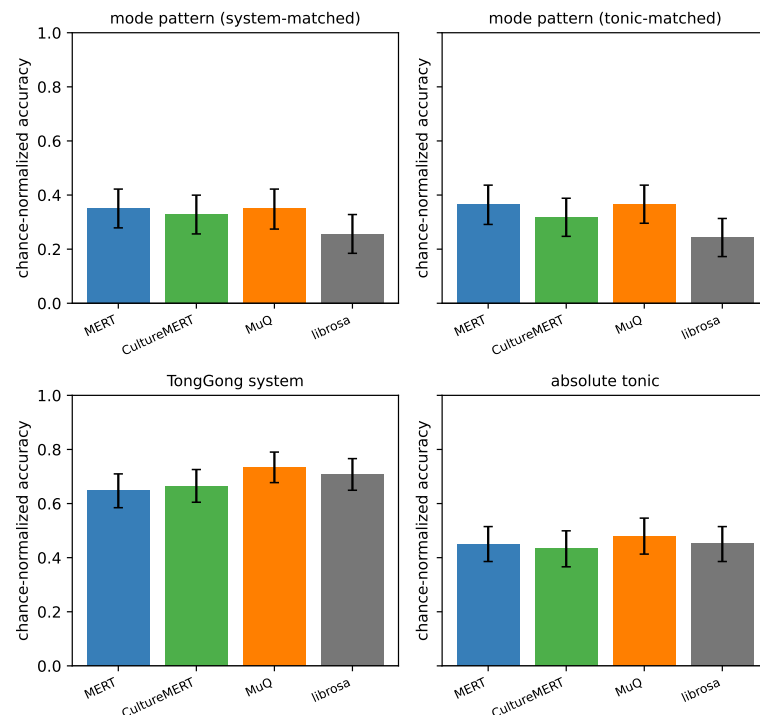


Figure 4. Same-audio CNPM controls. Chance-normalized logistic-regression accuracy with 10,000-resample bootstrap 95% intervals. Within each control/pattern pair, the clips and piece-grouped folds are identical.

4.4. Chroma Is Essential for External Key but Does Not Alone Explain CNPM Pattern

The 88-dimensional librosa baseline was ablated into all features, the 24 chroma summary values alone, and the remaining 64 values. Under the pre-specified logistic-regression protocol, GiantSteps key reaches chance-normalized accuracy 0.461 with all features, 0.582 with chroma alone, and 0.079 without chroma. Thus, the external key reference is largely chroma-driven, as expected. On the CNPM pattern, the corresponding values are 0.229, 0.207, and 0.172. Chroma contributes, but neither retaining only chroma nor removing it reproduces the full baseline. The ablation therefore prevents a simplistic reading of the CNPM result as merely a chroma-presence effect (Figure 5).

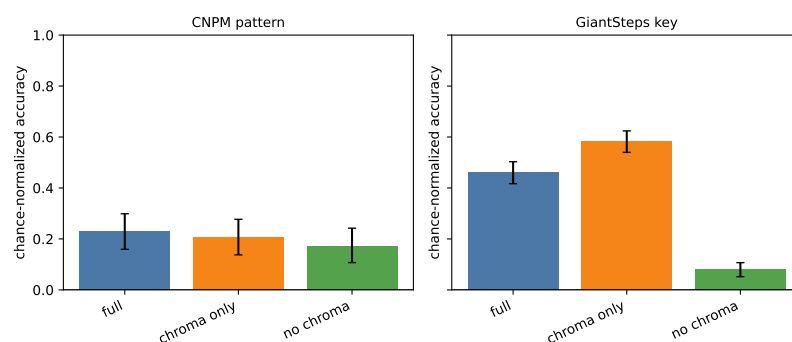


Figure 5. librosa chroma ablation using pre-specified logistic regression. Bars show chance-normalized accuracy with 10,000-resample bootstrap 95% intervals. GiantSteps is an external key reference; CNPM uses the piece-grouped mode protocol.

4.5. A Task-Specific CNN Remains a Reference, Not a Direct Literature Replication

To check whether the frozen-feature result is simply a consequence of using shallow probes, we trained a compact task-specific CNN on the current five-way CNPM target

under the same five-fold piece-grouped protocol. The CNN receives one deterministic 10-s centre excerpt per recording as a 64-bin log-Mel spectrogram; it has three convolutional blocks (24/48/96 channels), global average pooling, and a linear classifier. We train for 50 fixed epochs with AdamW and average three fixed initialization seeds within each held-out fold, without validation-set model selection. Its out-of-fold chance-normalized accuracy is 0.207 (95% bootstrap interval 0.138–0.277; macro-F1 0.340). This small baseline is deliberately not presented as a reproduction of Wang et al. [18]: that work uses a different joint target and data protocol. It shows only that, under the present labels and leakage-controlled grouping, a compact supervised log-Mel CNN does not remove the difficulty of the five-way pattern task.

4.6. Individual-Model Mode Errors Are Descriptive

Figure A1 presents row-normalized CNPM confusions for MERT and MuQ separately. Their broadly similar errors motivate further investigation, but they are not pooled for inference. In particular, the observed confusions are consistent with difficulty in separating some mode labels, yet a confusion pattern alone cannot establish which latent musical variable is or is not encoded. We report them alongside the CNPM controls and ablations without inferring a tonic-specific mechanism.

4.7. The Weakness Holds at Every Layer

The layer-wise profiles serve as an exhaustive robustness check rather than a layer-selection procedure. Figure 6 reports every architecture-native state (layers 0–12) under the same splits as the main tasks. No single layer reverses the overall ordering between the surface tasks and the CNPM mode-pattern task. CultureMERT is compared as the released model adapted on Greek, Turkish, and Indian music. For the Chinese-music tasks examined here, that released model does not yield consistent gains. This limited comparison does not test multicultural adaptation in general or Chinese-targeted adaptation in particular.

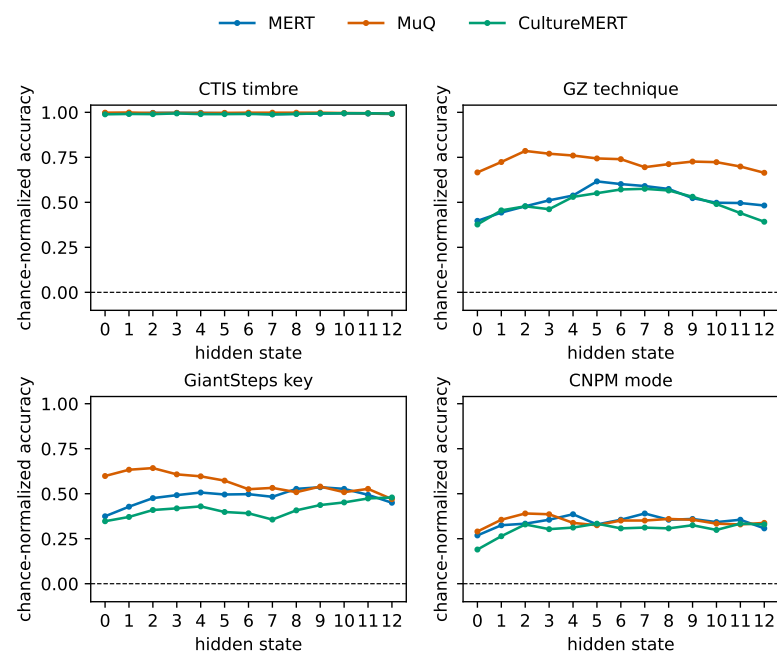


Figure 6. Layer-wise probing. Chance-normalized accuracy of a logistic-regression probe over all 13 architecture-native hidden states (layers 0–12) of each foundation model, one panel per task. Each task uses the same split as the descriptive main table (timbre: top-20 random; technique: source-aware; key: four-fold stratified; mode: piece-grouped). The figure is an exhaustive sensitivity analysis, not a layer-selection step.

5. Discussion

5.1. Main Interpretation

Taken together, the results yield a dimension-wise profile. Across the present Chinese and Western evaluation settings, surface attributes are readily decodable, and Western key receives higher scores than the CNPM mode-pattern task. GiantSteps and CNPM are not matched corpora, so the difference is reported only as an external comparison. Within CNPM, the TongGong system and absolute tonic remain decodable on the same clips and folds; the mode-pattern result should therefore not be characterized as an absence of pitch information. The chroma ablation further shows that the external key reference is predominantly driven by pitch-class energy, whereas the CNPM pattern is not explained by chroma alone. This outcome aligns with evidence that learned and hand-crafted descriptors can be complementary rather than categorically ordered [3,10] and with the need, emphasized by cross-cultural evaluations, to distinguish task-specific evidence from corpus composition [13].

5.2. Evidence Against a Probe Artifact

A low linear-probe score need not prove that information is absent; it could be non-linearly arranged [22,23]. We therefore include a fixed non-linear LightGBM probe, inspect all exposed layers, and report a task-specific CNN baseline on the CNPM labels under the same grouping protocol. These analyses contextualize the frozen-feature result, but none licenses a claim about all possible downstream architectures or fine-tuning procedures. Reporting the best of three probes in the descriptive table also introduces mild selection optimism; the full per-probe table and the fixed-probe controls make this choice visible.

5.3. Limitations

Several limitations bound our claims. GiantSteps is EDM and is not a matched control for CNPM; we therefore use it only as an external tonal reference and do not infer a culture effect from its difference. CNPM is small (287 clips), and its annotations, though useful for same-audio controls, do not constitute a synthetic factorial design: system, tonic, pattern, repertoire, and performance may still covary. CTIS has no recording-level identifier and is an upper bound. GZ source-aware evaluation reveals provenance effects but cannot fully disentangle source from all performance differences. Finally, frozen probing measures what is readily decodable from fixed features, not what a task-specific model or Chinese-targeted fine-tuning could induce.

5.4. Implications and Future Work

The practical implication is conditional. For a CNPM mode-pattern application, frozen embeddings should be benchmarked against transparent tonal descriptors and task-specific baselines rather than assumed to contain the needed distinction. The compact CNN is a protocol-matched reference, not a numerical replication of the task-specific CNN study by Wang et al. [18]; differences in target construction and data split prevent such a ranking. The current CultureMERT comparison cannot answer whether adaptation trained on Chinese music or an objective targeting tonal function would improve the result. A stronger next study would use a deliberately matched or synthetic corpus that independently varies system, tonic, mode pattern, repertoire, instrumentation, and recording condition, then test Chinese-targeted adaptation under that design.

6. Conclusions

We probed three music foundation models and a hand-crafted baseline across surface and tonal tasks. In the present evaluation, Western key receives higher chance-normalized

scores than the CNPM mode-pattern task, but GiantSteps and CNPM are not a matched Chinese–Western control. The CNPM analyses show that mode pattern is less readily decoded than the TongGong system or absolute tonic on the same audio subsets, and that chroma alone does not account for this difference. Leakage-aware GZ evaluation, all-layer inspection, a fully specified feature/classifier pipeline, chroma ablations, and a task-specific CNN reference delimit the result to the present frozen-feature protocol. The study thus provides a reproducible dimension-wise account of both its empirical findings and their scope.

Author Contributions: Conceptualization, X.B. and N.H.; methodology, N.H. and Y.L.; software, Y.L.; validation, N.H., Y.L. and X.H.; formal analysis, N.H. and Y.L.; investigation, X.B., N.H. and X.H.; resources, X.B. and X.H.; data curation, Y.L.; writing—original draft preparation, X.B. and N.H.; writing—review and editing, X.B., N.H., Y.L. and X.H.; visualization, Y.L.; supervision, X.H.; project administration, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available: the CTIS, GZ_IsoTech, and CNPM corpora are distributed through the CCMusic database on Hugging Face (huggingface.co/ccmusic-database), and the GiantSteps key dataset is accessed via the mirdata package. Code, split definitions, configuration files, and the environment specification will be made publicly available upon publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Meaning
MIR	Music information retrieval.
MERT	Acoustic Music Understanding Model with large-scale self-supervised training.
MuQ	Self-supervised music representation learning with Mel residual vector quantization.
CNPM	Chinese pentatonic mode; “mode pattern” denotes the five-way Gong, Shang, Jue, Zhi, Yu target.
CTIS	CCMusic instrument-timbre dataset.
GZ_IsoTech	CCMusic guzheng playing-technique dataset.
EDM	Electronic dance music.
MFCC	Mel-frequency cepstral coefficient.
STFT	Short-time Fourier transform.
CNN	Convolutional neural network.
SVM	Support vector machine.
LightGBM	Light Gradient-Boosting Machine.

Appendix A. Paired Representation Contrasts

Table A1 makes the inferential results for the pre-specified same-audio CNPM representation contrasts explicit. Each difference is the first representation minus the second in chance-normalized accuracy. The correction family contains the 16 listed contrasts; these are within-label representation comparisons, not tests of the unmatched GiantSteps–CNPM cross-corpus difference.

Table A1. Paired representation contrasts for the same-audio CNPM controls. Δ is the difference in chance-normalized accuracy (first representation minus second); intervals are 10,000-resample paired-bootstrap 95% intervals; p values come from two-sided 10,000-sign-permutation tests. p_{BH} is Benjamini–Hochberg adjusted across all 16 listed contrasts.

Matched Label	Contrast	Δ	95% Interval	p	p_{BH}
TongGong system ($n = 279$)	CultureMERT–MERT	0.016	[−0.028, 0.060]	0.582	0.656
TongGong system ($n = 279$)	MERT–librosa	−0.060	[−0.117, −0.004]	0.056	0.137
TongGong system ($n = 279$)	CultureMERT–librosa	−0.044	[−0.101, 0.012]	0.166	0.333
TongGong system ($n = 279$)	MuQ–librosa	0.024	[−0.028, 0.077]	0.461	0.656
Pattern, same system subset ($n = 279$)	CultureMERT–MERT	−0.022	[−0.085, 0.040]	0.582	0.656
Pattern, same system subset ($n = 279$)	MERT–librosa	0.094	[0.018, 0.170]	0.016	0.063
Pattern, same system subset ($n = 279$)	CultureMERT–librosa	0.072	[0.004, 0.139]	0.050	0.137
Pattern, same system subset ($n = 279$)	MuQ–librosa	0.094	[0.022, 0.166]	0.014	0.063
Absolute tonic ($n = 284$)	CultureMERT–MERT	−0.016	[−0.063, 0.031]	0.615	0.656
Absolute tonic ($n = 284$)	MERT–librosa	−0.004	[−0.063, 0.055]	1.000	1.000
Absolute tonic ($n = 284$)	CultureMERT–librosa	−0.020	[−0.082, 0.043]	0.610	0.656
Absolute tonic ($n = 284$)	MuQ–librosa	0.027	[−0.035, 0.090]	0.458	0.656
Pattern, same tonic subset ($n = 284$)	CultureMERT–MERT	−0.048	[−0.114, 0.018]	0.196	0.348
Pattern, same tonic subset ($n = 284$)	MERT–librosa	0.123	[0.044, 0.202]	0.005	0.046
Pattern, same tonic subset ($n = 284$)	CultureMERT–librosa	0.075	[0.000, 0.145]	0.060	0.137
Pattern, same tonic subset ($n = 284$)	MuQ–librosa	0.123	[0.044, 0.202]	0.006	0.046

Appendix B. Per-Probe Results

Table A2 reports the chance-normalized accuracy of all three probes (logistic regression, linear SVM, gradient-boosted trees) for every (model, task) cell, so that the best-probe selection used in Table 3 is fully transparent. The bolded value in each cell is the one reported in Table 3.

Table A2. Per-probe chance-normalized accuracy for every (model, task) cell. Bold marks the best probe per cell, i.e., the value reported in Table 3. All three pre-specified probes are run for the GiantSteps external reference.

Model	Probe	CTIS Timbre	GZ Technique	GiantSteps Key	CNPM Mode Pattern
MERT	LogReg	0.995	0.584	0.516	0.351
	LinearSVC	0.997	0.539	0.420	0.321
	LightGBM	0.978	0.421	0.455	0.325
MuQ	LogReg	0.998	0.686	0.604	0.355
	LinearSVC	0.998	0.787	0.518	0.347
	LightGBM	0.988	0.457	0.529	0.347
CultureMERT	LogReg	0.994	0.545	0.441	0.316
	LinearSVC	0.994	0.494	0.380	0.312
	LightGBM	0.977	0.406	0.387	0.307
librosa	LogReg	0.995	0.354	0.461	0.229
	LinearSVC	0.996	0.382	0.402	0.255
	LightGBM	0.986	0.307	0.435	0.294

Appendix C. Per-Model Confusion Matrices

Figure A1 shows row-normalized confusion matrices for MERT and MuQ separately (piece-grouped split, best-probe out-of-fold predictions). They are shown as descriptive, model-specific error patterns and are not pooled for statistical inference.

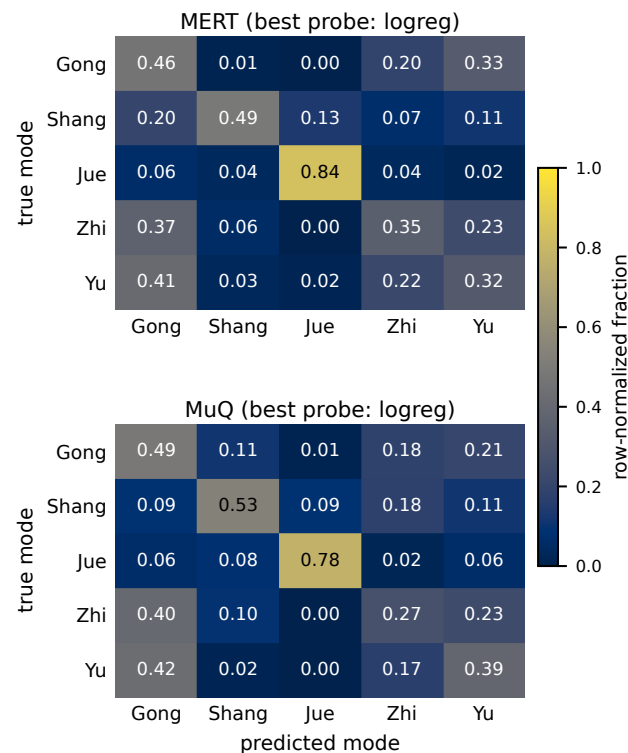


Figure A1. Per-model CNPM confusion. Row-normalized 5×5 confusion matrices for MERT and MuQ separately, on the piece-grouped split with best-probe out-of-fold predictions. The panels are descriptive and are not pooled.

References

1. Lerch, A.; Knees, P. Machine Learning Applied to Music/Audio Signal Processing. *Electronics* **2021**, *10*, 3077. [\[CrossRef\]](#)
2. Ashraf, M.; Abid, F.; Din, I.U.; Rasheed, J.; Yesiltepe, M.; Yeo, S.F.; Ersoy, M.T. A Hybrid CNN and RNN Variant Model for Music Classification. *Appl. Sci.* **2023**, *13*, 1476. [\[CrossRef\]](#)
3. Louro, P.L.; Redinho, H.; Malheiro, R.; Paiva, R.P.; Panda, R. A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition. *Sensors* **2024**, *24*, 2201. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; et al. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
5. Zhu, H.; Zhou, Y.; Chen, H.; Yu, J.; Ma, Z.; Gu, R.; Luo, Y.; Tan, W.; Chen, X. MuQ: Self-Supervised Music Representation Learning With Mel Residual Vector Quantization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2025**, *33*, 3653–3664. [\[CrossRef\]](#)
6. Won, M.; Hung, Y.N.; Le, D. A Foundation Model for Music Informatics. In Proceedings of the ICASSP 2024—IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 1226–1230. [\[CrossRef\]](#)
7. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A Generative Model for Music. *arXiv* **2020**, arXiv:2005.00341.
8. Castellon, R.; Donahue, C.; Liang, P. Codified Audio Language Modeling Learns Useful Representations for Music Information Retrieval. In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR), Online, 7–12 November 2021; pp. 88–96.
9. Yuan, R.; Ma, Y.; Li, Y.; Zhang, G.; Chen, X.; Yin, H.; Zhuo, L.; Liu, Y.; Huang, J.; Tian, Z.; et al. MARBLE: Music Audio Representation Benchmark for Universal Evaluation. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023.
10. Xie, B.; Kim, J.C.; Park, C.H. Musical Emotion Recognition with Spectral Feature Extraction Based on a Sinusoidal Model with Model-Based and Deep-Learning Approaches. *Appl. Sci.* **2020**, *10*, 902. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Wei, M.; Freeman, M.; Donahue, C.; Sun, C. Do Music Generation Models Encode Music Theory? In Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR), San Francisco, CA, USA, 10–14 November 2024; pp. 680–687.
12. Zhou, Y.; Zhu, H.; Chen, H. Layer-wise Investigation of Large-Scale Self-Supervised Music Representation Models. *arXiv* **2025**, arXiv:2505.16306.

13. Papaioannou, C.; Benetos, E.; Potamianos, A. Universal Music Representations? Evaluating Foundation Models on World Music Corpora. In Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR), Daejeon, Republic of Korea, 21–25 September 2025; pp. 303–311.
14. Kanatas, A.N.; Papaioannou, C.; Potamianos, A. CultureMERT: Continual Pre-Training for Cross-Cultural Music Representation Learning. In Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR), Daejeon, Republic of Korea, 21–25 September 2025; pp. 555–564.
15. Izmirli, O. Template Based Key Finding from Audio. In Proceedings of the International Computer Music Conference (ICMC), Barcelona, Spain, 5–9 September 2005; pp. 211–214.
16. Korzeniowski, F.; Widmer, G. End-to-End Musical Key Estimation Using a Convolutional Neural Network. In Proceedings of the 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 966–970. [\[CrossRef\]](#)
17. Driedger, J.; Müller, M. A Review of Time-Scale Modification of Music Signals. *Appl. Sci.* **2016**, *6*, 57. [\[CrossRef\]](#)
18. Wang, Z.; Che, M.; Yang, Y.; Meng, W.; Li, Q.; Xia, F.; Li, W. Automatic Chinese National Pentatonic Modes Recognition Using Convolutional Neural Network. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), Bengaluru, India, 4–8 December 2022; pp. 345–352.
19. Li, D.; Wu, Y.; Li, Q.; Zhao, J.; Yu, Y.; Xia, F.; Li, W. Playing Technique Detection by Fusing Note Onset Information in Guzheng Performance. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), Bengaluru, India, 4–8 December 2022; pp. 314–320.
20. Zhou, M.; Xu, S.; Liu, Z.; Wang, Z.; Yu, F.; Li, W.; Han, B. CCMusic: An Open and Diverse Database for Chinese Music Information Retrieval Research. *Trans. Int. Soc. Music Inf.* **2025**, *8*, 22–38. [\[CrossRef\]](#)
21. Xia, H.; Huang, Z.; Tan, Y.; Song, S. Let the Model Learn to Feel: Mode-Guided Tonality Injection for Symbolic Music Emotion Recognition. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 2182–2190. [\[CrossRef\]](#)
22. Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances. *Comput. Linguist.* **2022**, *48*, 207–219. [\[CrossRef\]](#)
23. Hewitt, J.; Liang, P. Designing and Interpreting Probes with Control Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 2733–2743. [\[CrossRef\]](#)
24. Ma, W.; Li, X.; Xia, G. Do Music LLMs Learn Symbolic Concepts? A Pilot Study Using Probing and Intervention. In Proceedings of the NeurIPS 2024 Workshop on Audio Imagination, Vancouver, BC, Canada, 14 December 2024. Available online: <https://openreview.net/forum?id=uvzw0gS0Nn> (accessed on 29 July 2026).
25. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.W.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25. [\[CrossRef\]](#)
26. Knees, P.; Faraldo, Á.; Herrera, P.; Vogl, R.; Böck, S.; Hörschläger, F.; Le Goff, M. Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections. In Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain, 26–30 October 2015; pp. 364–370. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.