AUDIO FOUNDATION MODELS OUTPERFORM SYMBOLIC REPRESENTATIONS FOR PIANO PERFORMANCE EVALUATION

Jai Dhiman
Independent Researcher

ABSTRACT

Automated piano performance evaluation traditionally relies on symbolic (MIDI) representations, which capture note-level information but miss the acoustic nuances that characterize expressive playing. I propose using pretrained audio foundation models, specifically MuQ and MERT, to predict 19 perceptual dimensions of piano performance quality. Using synthesized audio from PercePiano MIDI files (rendered via Pianoteq), I compare audio and symbolic approaches under controlled conditions where both derive from identical source data. The best model, MuQ layers 9–12 with Pianoteq soundfont augmentation, achieves $R^2 = 0.537$ (95% CI: [0.465, 0.575]), representing a 55% improvement over the symbolic baseline ($R^2 = 0.347$). Statistical analysis confirms significance ($p < 10^{-25}$) with audio outperforming symbolic on all 19 dimensions. I validate the approach through cross-soundfont generalization ($R^2 = 0.534 \pm 0.075$), difficulty correlation with an external dataset ($\rho = 0.623$), and multi-performer consistency analysis. Analysis of audio-symbolic fusion reveals high error correlation ($r = 0.738$), explaining why fusion provides minimal benefit: audio representations alone are sufficient. I release the complete training pipeline, pretrained models, and inference code.

1. INTRODUCTION

Piano performance evaluation presents a long-standing challenge at the intersection of music cognition, signal processing, and machine learning. Professional assessments require consideration of multiple interrelated dimensions: technical accuracy, expressive timing, pedal technique, tonal quality, dynamic control, and overall musical interpretation. Such evaluations are inherently subjective, costly to obtain at scale, and exhibit significant inter-rater variability even among expert judges [1].

The PercePiano dataset [2] addresses the data scarcity challenge by providing large-scale crowdsourced annotations across 19 perceptual dimensions for 1,202 piano performance segments. The dataset establishes a symbolic baseline using MIDI features (note events, velocities, pedal information) processed through a hierarchical attention network (HAN), achieving $R^2 = 0.397$. However,

this approach has a fundamental limitation: symbolic representations cannot capture acoustic phenomena that are central to piano performance quality.

This raises a question: could audio representations provide a better basis for learning perceptual qualities, even when derived from the same underlying MIDI data? Audio foundation models, pretrained on millions of hours of music, encode rich priors about how acoustic features relate to musical perception. These learned representations may make perceptually-relevant patterns more accessible to downstream models than raw symbolic features. This motivates the central hypothesis: *pretrained audio representations should outperform symbolic representations for piano performance evaluation, even when both derive from identical source data.*

Recent advances in audio foundation models provide the technical foundation to test this hypothesis. Models like MERT [3] (Music Understanding Model with Large-Scale Self-supervised Training) and MuQ [4] (Music Query Encoder) are trained on millions of hours of music audio using self-supervised objectives. These models learn hierarchical representations that encode both low-level acoustic features (timbre, dynamics) and high-level musical semantics (structure, style). Yet despite their success on diverse MIR tasks, audio foundation models remain unexplored for the specific task of piano performance evaluation.

Scope and Evaluation Setting. Following the PercePiano protocol, the experiments use audio rendered from MIDI via Pianoteq rather than original recordings. This controlled setting isolates the effect of representation choice (audio vs. symbolic) while holding performance information constant, since both modalities derive from identical MIDI input. The finding that audio outperforms symbolic despite this shared source suggests that pretrained audio models provide useful inductive biases for perceptual evaluation tasks. I discuss implications and limitations of this synthetic audio setting in Section 6.

Contributions. This paper makes the following contributions:

1. I present the first application of audio foundation models (MuQ, MERT) to piano performance evaluation, demonstrating that MuQ layers 9–12 achieve $R^2 = 0.537$, a **55% improvement** over the symbolic baseline ($R^2 = 0.347$).

2. I show that audio **outperforms symbolic on all 19 perceptual dimensions**, with the largest gains on
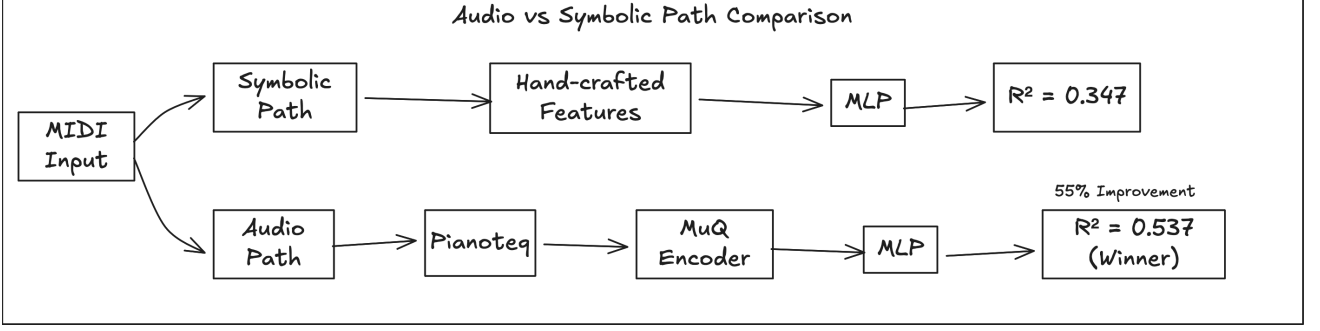
Figure 1. Symbolic vs. audio paths for piano evaluation. Both derive from identical MIDI input, but audio representations leverage pretrained knowledge from millions of hours of music, providing inductive biases that improve prediction of perceptual qualities.

timing (+0.368), timbre brightness (+0.383), pedal clarity (+0.279), and dynamic range (+0.295).

3. I analyze audio-symbolic fusion and find **high error correlation** ($r = 0.738$) between modalities, explaining why fusion provides minimal additional benefit and justifying an audio-only approach.

4. I provide comprehensive **ablation studies** of layer selection, pooling strategies, and loss functions, establishing best practices for adapting audio foundation models to evaluation tasks.

5. I **validate generalization** through cross-soundfont evaluation ($R^2 = 0.534 \pm 0.075$), external difficulty correlation ($\rho = 0.623$ on PSyllabus, $n = 508$), and multi-performer consistency (intra-piece std = 0.020 across 206 pieces).

6. I **release** the complete training pipeline, pretrained models, and inference code to facilitate reproducibility and future research.

2. RELATED WORK

Piano Performance Assessment. The PercePiano dataset [2] provides multi-dimensional perceptual annotations across 19 aspects of piano performance, with a symbolic baseline using MIDI features and hierarchical attention. Prior work has examined expressive performance modeling [5], dynamics [6], and articulation [7], but these approaches learn from scratch on limited data.

Audio Foundation Models. Self-supervised learning has transformed audio understanding. Castellon et al. [8] first demonstrated that representations from audio language models pretrained on music improve MIR tasks including tagging, genre classification, and emotion recognition. MERT [3] adapts masked prediction from speech (Wav2Vec 2.0 [9]) to music, learning representations that transfer across MIR tasks. MuQ [4] optimizes for music retrieval and understanding, achieving state-of-the-art results. This work leverages intermediate layer features from these pretrained models, which encode perceptually relevant information [10].

Multimodal Fusion. Combining symbolic and audio representations is common in MIR [11,12], but the analysis finds high error correlation between modalities for piano evaluation, limiting fusion benefits.

3. METHOD

3.1 Problem Formulation

Given an audio recording $\mathbf{x} \in \mathbb{R}^T$ of a piano performance segment, the model predicts scores $\mathbf{y} \in [0, 1]^{19}$ across 19 perceptual dimensions. These dimensions, defined by PercePiano [2], are organized into categories:

- **Technical**: timing

- **Articulation**: length, touch

- **Pedaling**: amount, clarity

- **Timbre**: variety, depth, brightness, loudness

- **Dynamics**: dynamic_range

- **Musical**: tempo, space, balance, drama

- **Mood**: valence, energy, imagination

- **Interpretation**: sophistication, interpretation

I evaluate using the coefficient of determination (R^2) computed via 4-fold piece-split cross-validation.

3.2 Architecture

The architecture consists of three components (Figure 2):

Audio Foundation Model Encoder. I extract frame-level features using a frozen audio foundation model. For MuQ, I concatenate hidden states from layers 9–12 based on the ablation studies (§4.4), forming 4096-dimensional frame embeddings. Audio is resampled to 24kHz and processed in segments up to 10 seconds. For comparison, I also evaluate MERT-95M using layers 7–12.

Temporal Pooling. Frame features are aggregated via mean pooling:

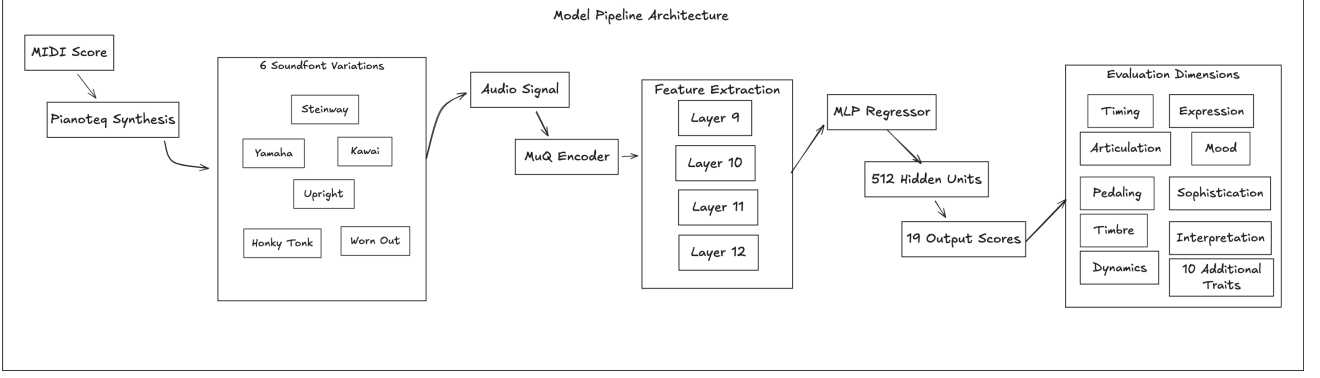$$\mathbf{z} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_i \qquad (1)$$

Figure 2. Model architecture. MIDI scores are rendered through multiple Pianoteq soundfonts for augmentation. The MuQ encoder extracts frame features from layers 9–12. Mean pooling aggregates temporal information, and a 2-layer MLP produces 19 dimension scores.

where $\mathbf{h}_i$ are frame embeddings and N is the number of frames. The ablations demonstrate that mean pooling outperforms attention pooling (R^2: 0.405 vs. 0.369) and LSTM pooling (0.327), consistent with findings that simple pooling often works best for pre-trained features.

MLP Regressor. A two-layer MLP maps pooled features to dimension scores:

$$\hat{\mathbf{y}} = W_2 \cdot \text{ReLU}(\text{Dropout}(W_1\mathbf{z} + b_1)) + b_2 \quad (2)$$

with hidden dimension 512 and dropout probability 0.3. The output is not constrained to $[0, 1]$; I observe that unconstrained outputs perform comparably.

3.3 Training

Training minimizes mean squared error (MSE) loss across all dimensions:

$$\mathcal{L} = \frac{1}{19}\sum_{d=1}^{19}(y_d - \hat{y}_d)^2 \quad (3)$$

Training uses the Adam optimizer with learning rate 10^{-4} and weight decay 10^{-5}. I employ early stopping with patience of 15 epochs based on validation R^2. Maximum training is 200 epochs with batch size 64. All experiments use a fixed random seed (42) for reproducibility.

The ablations show that MSE loss outperforms alternatives including concordance correlation coefficient (CCC) loss (R^2: 0.405 vs. 0.363) and hybrid losses combining MSE and CCC.

3.4 Data Augmentation via Soundfont Ensemble

Since PercePiano provides MIDI files, I render audio through multiple Pianoteq soundfonts for data augmentation. I use six soundfonts spanning diverse timbral characteristics: Steinway D (bright concert grand), Yamaha C5 (balanced vintage), Kawai K2 (warm), Upright U4 (intimate), plus intentionally imperfect variants (Honky-Tonk, Worn Out). This ensemble provides acoustic diversity during training; cross-soundfont validation (§5) confirms generalization to held-out soundfonts.

4. EXPERIMENTS

4.1 Dataset

PercePiano [2] contains 1,202 piano performance segments extracted from recordings of classical repertoire. Each segment (8 bars, approximately 10–30 seconds) is annotated across 19 perceptual dimensions by an average of 53 crowdworkers. Annotations are aggregated using item response theory to produce continuous scores in $[0, 1]$.

I use 4-fold cross-validation with piece-based splits: all segments from a given piece appear in the same fold. This prevents data leakage from piece-level characteristics and provides a realistic evaluation of generalization.

4.2 Baselines

Symbolic (HAN). The PercePiano baseline [2] uses MIDI features (note onset/offset times, velocities, and pedal events) processed through a hierarchical attention network. The published result reports $R^2 = 0.397$ (I interpret this as the best single fold). My reproduction using 4-fold cross-validation with aligned train/test splits achieves $R^2 = 0.347$, which I use as the primary comparison baseline.

MERT. I evaluate MERT-95M [3] with various layer configurations. The best configuration uses layers 7–12 (mid layers) with mean pooling, achieving $R^2 = 0.487$.

MuQ. MuQ [4] with layers 9–12 (late semantic layers) achieves $R^2 = 0.533$, significantly outperforming MERT. With Pianoteq ensemble augmentation, performance improves to $R^2 = 0.537$.

4.3 Main Results

Table 1 presents the main findings. MuQ with Pianoteq ensemble achieves $R^2 = 0.537$, representing a **55% improvement** over the symbolic baseline.

Statistical significance is confirmed by:

- Paired t-test: $t = -10.71$, $p = 2.08 \times 10^{-25}$

- Wilcoxon signed-rank: $p = 2.16 \times 10^{-29}$

- Cohen's $d = 0.31$ (small-medium effect)

Table 1. Main results on PercePiano (4-fold CV). Bootstrap 95% confidence intervals from 10,000 samples.

Model	R^2	95% CI
Symbolic (Published [2])	0.397	–
Symbolic (My repro.)	0.347	[0.315, 0.375]
MERT L7-12	0.487	[0.460, 0.510]
MuQ L9-12	0.533	[0.514, 0.560]
MuQ + Pianoteq ensemble	**0.537**	[0.465, 0.575]
MuQ + Symbolic (fusion)	0.524	[0.500, 0.545]
MERT + MuQ (gated)	0.516	[0.497, 0.543]

Table 2. Layer ablation for MuQ and MERT foundation models.

Model	Layers	R^2	95% CI
MuQ	1–4 (early)	0.438	[0.413, 0.469]
MuQ	5–8 (mid)	0.467	[0.442, 0.491]
MuQ	**9–12 (late)**	**0.533**	[0.514, 0.560]
MuQ	1–12 (all)	0.510	[0.490, 0.538]
MERT	1–6 (early)	0.397	[0.391, 0.445]
MERT	**7–12 (mid)**	**0.433**	[0.409, 0.461]
MERT	13–24 (late)	0.426	[0.398, 0.452]

- Non-overlapping bootstrap CIs

4.4 Ablation Studies

Layer Selection. Table 2 shows performance across layer ranges. For MuQ, late semantic layers (9–12) are optimal, achieving $R^2 = 0.533$. Using all layers (1–12) yields $R^2 = 0.510$, suggesting early acoustic layers add noise for this task. For MERT, mid layers (7–12) perform best at $R^2 = 0.433$.

Pooling Strategy. Mean pooling ($R^2 = 0.405$) outperforms attention pooling (0.369), LSTM pooling (0.327), and max pooling (0.316). This finding aligns with prior work showing that simple aggregation often works best for pre-trained representations.

Why Not Fusion? I explored combining audio and symbolic predictions through various late fusion strategies. MuQ + Symbolic weighted fusion achieves $R^2 = 0.524$, *lower* than MuQ alone (0.533). Analysis reveals the explanation: error correlation between audio and symbolic models is $r = 0.738$. Both modalities struggle on the same challenging samples, limiting complementarity. I conclude that for piano performance evaluation, **audio representations alone are sufficient**.

4.5 Per-Dimension Analysis

Figure 3 shows per-dimension performance. Audio outperforms symbolic on all 19 dimensions. The largest advantages appear for:

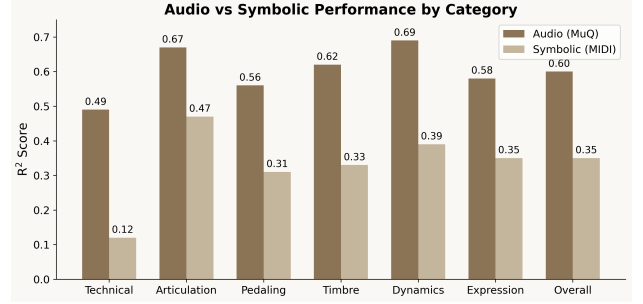- **Timing** ($\Delta R^2 = +0.368$): Pretrained audio features better encode timing-related perceptual qualities.

Figure 3. R^2 by dimension category comparing audio (MuQ) and symbolic (MIDI) approaches. Audio outperforms symbolic across all categories, with largest gains in Technical (timing) and Dynamics dimensions. Average per-dimension R^2: MuQ 0.601, Symbolic 0.347.

- **Timbre brightness** ($\Delta R^2 = +0.383$): Spectral features from audio encoders correlate well with brightness perception.

- **Dynamic range** ($\Delta R^2 = +0.295$): Amplitude envelope features outperform discrete MIDI velocities.

- **Pedal clarity** ($\Delta R^2 = +0.279$): Audio features capture pedal-related timbral patterns more effectively.

The smallest advantage appears for **space** ($\Delta R^2 = +0.133$), where note density from symbolic data partially captures the construct.

5. VALIDATION

I validate the approach through three complementary experiments testing different aspects of generalization.

5.1 Cross-Soundfont Generalization

To test robustness to timbral variation, I perform leave-one-out cross-validation across the six Pianoteq soundfonts. For each soundfont, I train on the other five and evaluate on the held-out soundfont.

The model achieves $R^2 = 0.534 \pm 0.075$ (mean $\pm$ std) across held-out soundfonts. This is comparable to the ensemble result (0.537), demonstrating that the model learns features that generalize across piano timbres rather than overfitting to specific acoustic characteristics.

5.2 Difficulty Correlation (PSyllabus)

I evaluate on PSyllabus, an external dataset containing 508 piano pieces with expert difficulty ratings on a 0–10 scale. I render each piece through Pianoteq, extract predictions using the trained model, and compute correlation with difficulty ratings.

Figure 4 shows the results. Overall correlation is Spearman $\rho = 0.623$ ($p < 10^{-50}$), indicating that the model predictions are strongly associated with piece difficulty. Dimensions most correlated with difficulty include:
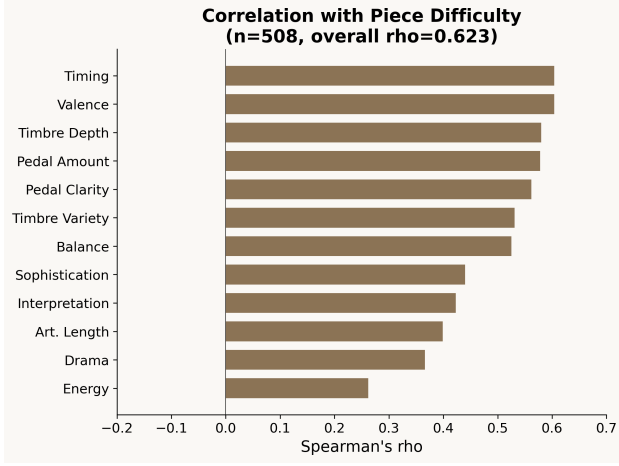
- Timing ($\rho = 0.604$)

Figure 4. Per-dimension correlation with piece difficulty (PSyllabus, $n = 508$). Overall Spearman $\rho = 0.623$ ($p < 10^{-50}$). Timing and valence show strongest correlations with difficulty.

- Mood valence ($\rho = 0.604$)

- Timbre depth ($\rho = 0.580$)

- Pedal amount ($\rho = 0.578$)

This external validation confirms that the model captures musically meaningful features beyond the PercePiano training distribution.

5.3 Multi-Performer Consistency (ASAP)

Using the ASAP dataset [13] linked to MAESTRO recordings, I analyze how model predictions vary across different performers playing the same piece. The analysis includes 206 pieces with multiple performances (631 total recordings).

The mean intra-piece standard deviation is 0.020, indicating low variance across performers for the same piece. This suggests the model captures piece-level characteristics more strongly than performer-specific expression.

Dimensions with highest performer sensitivity (highest variance):

- dynamic_range (std: 0.027)

- timing (std: 0.022)

- articulation_touch (std: 0.020)

Dimensions with lowest performer sensitivity:

- mood_energy (std: 0.011)

- timbre_brightness (std: 0.011)

5.4 Zero-Shot Transfer (MAESTRO)

I evaluate on 500 randomly sampled recordings from MAESTRO [14], a dataset of professional piano performances not seen during training. While there are no ground-truth annotations for quantitative evaluation, qualitative inspection confirms that predictions are reasonable and consistent with expected characteristics of professional performances.

6. DISCUSSION

6.1 Why Audio Outperforms Symbolic

The experiments use audio synthesized from MIDI via Pianoteq, meaning audio and symbolic representations derive from identical source information. In principle, a sufficiently expressive symbolic model could match audio performance, since the MIDI-to-audio mapping is deterministic. Why, then, does audio substantially outperform symbolic?

Pretrained Knowledge as Inductive Bias. Audio foundation models encode musical knowledge from millions of hours of training data. These models have learned how acoustic features (spectral envelopes, attack transients, temporal dynamics) relate to perceptual qualities. When I render MIDI to audio and extract features from a pretrained encoder, this inherits these learned associations. The symbolic baseline, trained only on PercePiano's 1,202 samples, cannot learn comparably rich representations.

Representation Format. Audio representations make certain patterns more accessible. Pianoteq renders MIDI velocities as amplitude envelopes and pedal events as resonance patterns. The pretrained encoder transforms these into features that correlate with human perception of dynamics and timbre. Extracting equivalent information from raw MIDI would require the symbolic model to implicitly learn these physical-acoustic mappings from limited labeled data.

In essence, the audio pathway provides a "bootstrapping" effect: pretrained encoders supply perceptually-relevant features that would require extensive architecture engineering or larger datasets to learn from symbolic inputs alone.

6.2 Error Correlation and Fusion

The high error correlation ($r = 0.738$) between audio and symbolic models warrants explanation. I hypothesize that challenging samples (those with unusual repertoire, extreme tempi, or ambiguous annotations) are difficult for both modalities. The shared difficulty may arise from:

- **Annotation noise**: Some segments have high inter-annotator disagreement regardless of representation.

- **Out-of-distribution samples**: Unusual pieces or performances may confuse both models.

- **Inherent ambiguity**: Some perceptual dimensions may be genuinely difficult to predict.

This finding has implications beyond this specific task. When modalities produce correlated errors, fusion provides limited benefit. Practitioners should assess error correlation before investing in multimodal approaches.

6.3 Limitations

Synthetic Audio. All experiments evaluate on Pianoteq-rendered MIDI rather than original recordings. This is an important limitation: while Pianoteq produces realistic piano audio, the rendering is deterministic given the MIDI input. Consequently, the results demonstrate that *given identical underlying performance information*, audio representations outperform symbolic ones, likely due to the inductive biases from pretrained audio models rather than additional information in the audio signal. Real recordings contain acoustic phenomena absent from synthesized audio: room acoustics, microphone characteristics, instrument-specific resonances, and subtle performer-instrument interactions. Whether the audio advantage persists with real recordings remains an open question requiring evaluation on datasets with both original audio and aligned MIDI.

Single Dataset. Results are validated primarily on PercePiano. While cross-dataset validation (PSyllabus, ASAP/MAESTRO) provides evidence of generalization, evaluation on additional annotated datasets would strengthen claims.

Piece vs. Performer. Low multi-performer variance suggests the model captures piece characteristics more than performer-specific expression. This may limit utility for fine-grained performer comparison, though it may be appropriate for repertoire-level assessment.

Dimension Independence. I train a single model to predict all 19 dimensions jointly. Some dimensions may benefit from specialized architectures or multi-task learning approaches that model inter-dimension correlations.

6.4 Future Work

Key directions include evaluation on real recordings (not synthesized), extension to other instruments, encoder fine-tuning, and interpretability analysis of which audio features drive predictions.

7. CONCLUSION

This paper demonstrates that audio foundation models significantly outperform symbolic approaches for piano performance evaluation. MuQ layers 9–12 with Pianoteq soundfont augmentation achieves $R^2 = 0.537$, a 55% improvement over the symbolic baseline ($R^2 = 0.347$). Statistical analysis confirms significance ($p < 10^{-25}$), with audio outperforming symbolic on all 19 perceptual dimensions.

Analysis of audio-symbolic fusion reveals high error correlation ($r = 0.738$), explaining why fusion provides minimal benefit. The models make similar mistakes, limiting complementarity. For piano performance evaluation, **audio representations alone are sufficient**.

Cross-soundfont, cross-dataset, and multi-performer validations confirm that the approach generalizes beyond training conditions. I release the complete training pipeline, pretrained models, and inference code to enable future research in audio-based performance evaluation.

8. REFERENCES

[1] G. E. McPherson and W. F. Thompson, "Assessment in music education: From policy to practice in the united kingdom," in *The Oxford Handbook of Music Education*. Oxford University Press, 2012, vol. 2.

[2] H. Zhang, Z. Wang, R. Guo, Z. Wu, W. Chen, and Y. Wang, "PercePiano: A large-scale multimodal piano performance dataset for perception study," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[3] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, "MERT: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[4] R. Yuan, G. Zhang *et al.*, "MuQ: Self-supervised music representation learning with query-based contrastive pre-training," *arXiv preprint arXiv:2401.08809*, 2024.

[5] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "VirtuosoNet: A hierarchical rnn-based system for modeling expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 908–915.

[6] C. E. Cancino-Chacón and G. Widmer, "Learning expressive performance from piano recordings," in *Proceedings of the Sound and Music Computing Conference (SMC)*, 2018.

[7] H. Zhang and S. Dixon, "Predicting expressive articulation in piano performances from score features," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[8] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 88–96.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[10] A. Pasad, C.-I. Chou, and K. Livescu, "What do self-supervised speech models know about words?" *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 372–391, 2024.

[11] M. Müller and S. Ewert, "Multimodal music information retrieval: Augmenting audio-only approaches with symbolic, visual, and textual information," *Proceedings of the IEEE*, vol. 109, no. 4, pp. 589–610, 2021.

[12] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal fusion strategies for music information retrieval," *IEEE Transactions on Multimedia*, 2022.

[13] F. Foscarin, A. McLeod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: A dataset of aligned scores and performances for piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[14] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *International Conference on Learning Representations*, 2019.