DO MUSIC GENERATIVE MODELS UNDERSTAND MUSICALITY? AUTOMATIC MUSIC EVALUATION WITH MODEL-INTRINSIC SIGNALS

Xiaosha Li
Georgia Institute of Technology
xiaosha@gatech.edu

Chun Liu
ByteDance Inc.
chun.liu@bytedance.com

Ziyu Wang
New York University / MBZUAI
ziyu.wang@nyu.edu

ABSTRACT

Current music generative models can produce high-quality music, but does this ability imply that they "understand" how good the music is, and is that understanding aligned with human evaluation? Previous attempts to use the likelihood of a generative model to evaluate music, an approach commonly used in text, have proven unsuccessful, leading researchers to rely on standalone supervised music evaluation models. In this paper, we answer this question affirmatively: we show that a models intrinsic signals—derived from its hidden representations and predictions— are strongly correlated with human ratings. In particular, we study MusicGen and consider three types of features: (1) prediction loss, (2) prediction entropy, and (3) concepts extracted from the model using sparse autoencoders (SAE). Using these features, we train a lightweight prediction model to estimate subjective ratings. We evaluate these features both individually and in combination. We hypothesize that these signals parallel the listening process: the temporal and frequency-domain structure of loss and entropy reflect listeners' expectation and surprise, while gradient directions in SAE latent space predict perceived quality. Experiments on five human-evaluation benchmarks spanning continuous ratings and pairwise preferences confirm this hypothesis, with SAE latents carrying most of the predictive signal. [1]

1. INTRODUCTION

Generative music models have advanced rapidly, yet judging the musicality of their outputs without human listeners remains an open problem. Recent years have seen a growing body of human-evaluation datasets for music generation—MusicEval [1], SongEval [2], MusicPref [3], AIME [4], and Music Arena [5]—and of audio-based verifiers that attempt to approximate these ratings, most notably Meta Audiobox Aesthetics [6] and the music-instruction benchmarks CMI-Bench [7] and its reward-model variant [8]. These approaches share a common assumption: musicality is inferred solely from the generated audio, treating the music model as a black box. Such systems automate the rating step but leave the deeper question untouched: *why* does a human listener rate one clip as good and another as bad? Aesthetics research suggests no general answer exists; what can be measured is rating behavior in controlled settings.

Music cognition offers a more principled vantage point. Human appreciation has long been modeled as two exposure-driven processes: repeated exposure shapes preference toward the familiar (the Mere Exposure Effect [9, 10]), and real-time listening continually forms and re-evaluates predictions, with surprises such as the deceptive cadence carrying affective charge [11, 12]. Under this model of cognition—one among several—judging music reads as an alignment between the listener's exposure-shaped representations and expectations and the audio they encounter.

An autoregressive generative music model parallels parts of this process. It is trained on large amounts of music data, and its next-token logits encode an expectation over what the music should do next. This leads us to the central question of this paper: *can a generative music model judge how good its own output sounds to a human, using only its own intrinsic signals—without ever leaving its internal state to consult an external audio verifier?* The broader version is just as natural: a generation model that can *generate* highly-rated music—can it also *evaluate* it in its own traces?

We answer this question empirically. We extract three intrinsic signals from a single pretrained autoregressive music generator—per-token loss, next-token predictive entropy, and sparse-autoencoder (SAE) latents of the hidden states—and align them to human ratings from five diverse benchmarks (MusicEval, SongEval, MusicPref, AIME, Music Arena) under one unified protocol. Empirically, loss and entropy together isolate an *overconfident-wrong* scenario—narrow bet, high surprise—whose share is the dominant local correlate of low ratings; SAE latents independently organise clips along a good–bad axis without human labels. Gemini-2.5-Flash audio captioning corroborates both findings, with a professional music producer agreeing on the latent concepts. Concretely, our contributions are:

[1] Code: https://github.com/YoEv/MEva; demo page with audio examples: https://yoev.github.io/MEva; evaluator checkpoints: https://huggingface.co/Dev4IC/MEva-checkpoints.

- **Intrinsic-signal alignment framework.** Model-internal signals from a single pretrained music generator align with human ratings across five benchmarks under one protocol.

- **Loss–entropy mismatch as a low-rating signature.** The mismatch isolates an *overconfident-wrong* scenario that is the dominant local correlate of low ratings, corroborated by audio captioning.

- **Representational evidence from SAE.** SAE latents, trained without ratings, separate highly- from poorly-rated clips, corroborated by audio captioning and a professional producer.

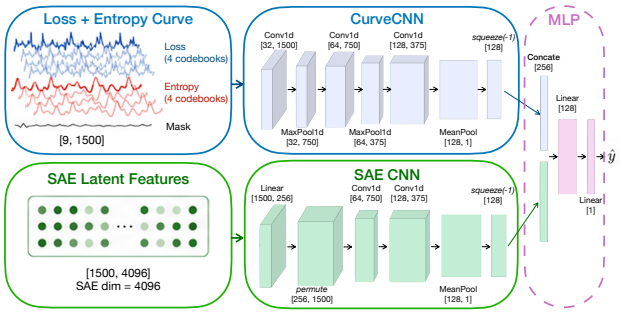

Figure 1. Hybrid prediction model: one 1D-CNN encoder per intrinsic signal, late concatenation (Eq. 3), and a shared MLP head producing $\hat{y} \in [1, 5]$.

2. RELATED WORK

We connect three strands of music evaluation—audio-based verifiers, human-rated benchmarks, and the cognitive science of preference—to a fourth largely missing for music: model-internal signals (loss, entropy, sparse representations) standard in language modeling and interpretability.

Several recent benchmarks score generated music directly from the output waveform. Audiobox Aesthetics [6] evaluates generated music along two complementary perspectives: production-side quality (Production Complexity, Production Quality) and content-side perception (Content Enjoyment, Content Usefulness). CMI-Bench [7] reframes traditional MIR annotations as music instruction-following tasks for audio-text LLMs, while CMI-RewardBench [8] introduces a reward-modeling benchmark for generated music across musicality, text–music alignment, and compositional multimodal instruction alignment.

Recent benchmarks span expert continuous ratings on text-to-music clips (MusicEval [1]) and full songs (SongEval [2]); pairwise head-to-head battles (MusicPref [3], Music Arena [5], the latter analogous to Chatbot Arena [13]); and the survey-style AIME [4], in which each clip appears in twelve pairwise comparisons per rating axis. Their varied supervision formats (MOS, multi-axis ratings, pairwise outcomes) are unified in Section 4. Both families judge from the waveform alone, leaving the *why* of preference unaddressed. In the survey of Lerch et al. [14], our method falls under objective, reference-free quality prediction trained on human ratings—differing from existing predictors in taking as input the generator's internal signals rather than audio embeddings.

Of the many mechanisms music-cognition research proposes, two map naturally onto an autoregressive model: repeated exposure builds preference-shaping representations of timbre, tonality, and structure (the Mere Exposure Effect [9, 10]), and real-time listening generates expectations whose violations carry affective weight [11, 12]. A pretrained autoregressive music model instantiates both: its hidden states encode representations shaped by training exposure [15, 16], and its next-token logits are a learned expectation [17].

If musicality rests on prediction and familiarity, the natural place to read it off is the generator's own internal signals. Li et al. [18] show that, in music LLMs, mean loss does not align with human-rated musicality and local minima do not align with preferred regions; what does carry signal is the shape of the loss curve over time. Parallel work on language models shows that perplexity (and therefore negative log-likelihood) cannot by itself distinguish confidently-correct from confidently-incorrect predictions [19]. This aligns with a long-standing observation in the calibration literature: modern neural networks tend to be overconfident [20], and penalizing overconfident output distributions acts as an implicit regularizer [21].

Entropy thus enters naturally as a complementary signal. It has been used as a reliability indicator at decoding time [22]. At the token level, per-token loss outperforms entropy as an error indicator [23]; at the sequence level, the average predictive entropy of the softmax outperforms mean loss as a quality indicator [24]. The two are therefore complementary at different temporal scales, motivating an encoder that consumes both curves jointly.

Sparse autoencoders (SAEs) [25–27] map dense hidden activations to a sparser latent space, decomposing entangled representations into more localized, interpretable units. Singh et al. [15] apply this to a generative music model and recover sparse latents whose activations correspond to specific, musically meaningful concepts; using them as inference-time steering vectors yields controllable changes in generated audio. We reuse the same latent readout but repurpose it for evaluation, asking whether the latents that Singh et al. interpret concept-by-concept also organize clips along a good-versus-bad axis aligned with human ratings. Unlike loss and entropy—summaries of the output distribution—SAE features probe the model's internal representation, explained in Section 3.1.

3. METHODS

From a pretrained autoregressive music model we extract three intrinsic signals—token-level loss ℓ_t, predictive entropy H_t, and sparse-autoencoder (SAE) latents z_t—each chosen to mirror, at the level of the model, one of the listener-side processes invoked in Section 1, and train a

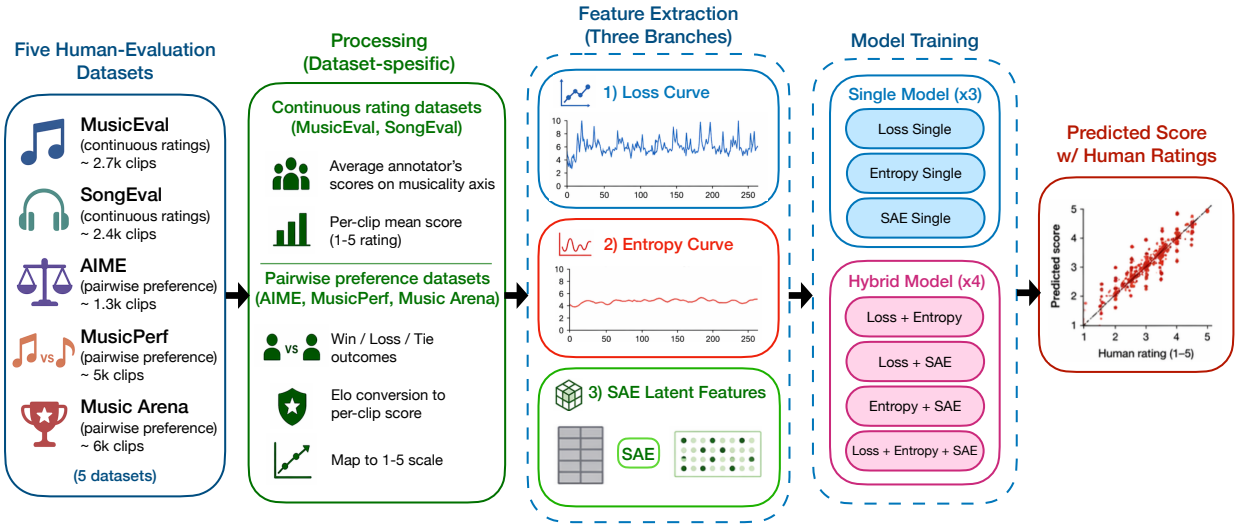

Figure 2. Experimental flowchart: human-evaluation data and preprocessing, per-clip loss / entropy / SAE time–series features, the hybrid neural network and its single- and two-signal ablations, and comparison of predicted quality against human-musicality labels.

single neural network that maps them to a human musicality rating.

3.1 Model-Intrinsic Signals

Given a generator p_θ and a token sequence $\{x_t\}_{t=1}^{T}$, we extract three complementary per-step signals, all computed under teacher-forcing on held-out audio.

Loss curve. Similar to a listener's *surprise*, the loss ℓ_t varies as a deviation from what the model expects, given as the negative log-likelihood of the ground-truth token under its next-token distribution:

$$\begin{aligned} \ell_t &= -\log p_\theta(x_t \mid x_{<t}), \\ \mathbf{l} &= [\ell_1, \ldots, \ell_T]. \end{aligned} \tag{1}$$

$\mathbf{l}$ is the time-resolved loss curve.

Predictive entropy curve. Like how uncertain a listener is about what comes next, H_t is the Shannon entropy of $p_\theta(\cdot \mid x_{<t})$—low when the model concentrates on a few continuations, high when it spreads across many:

$$\begin{aligned} H_t &= -\sum_{k=1}^{K} p_\theta(k \mid x_{<t}) \log p_\theta(k \mid x_{<t}), \\ \mathbf{h} &= [H_1, \ldots, H_T]. \end{aligned} \tag{2}$$

SAE latents. Where ℓ_t and H_t summarise only the output distribution, z_t probes the model's internal state and—mirroring the disentangled concepts (timbre, register, harmony) a listener acquires from exposure—re-expresses each hidden activation as a sparse combination of latent features that each tend to align with a single interpretable musical concept. For a hidden activation $u_t \in \mathbb{R}^{d_{\text{in}}}$, the SAE produces sparse codes $z_t = f_{\text{enc}}(u_t) \in \mathbb{R}^{d_{\text{sae}}}$ ($d_{\text{sae}} > d_{\text{in}}$) and reconstructions $\hat{u}_t = f_{\text{dec}}(z_t)$, trained by Singh et al. with the standard reconstruction-plus-ℓ_1 objective. Because only a few entries of z_t are active per step, the SAE acts as an overcomplete sparse dictionary that turns dense polysemantic activations into a sparse combination of more disentangled, concept-aligned features [15]. We reuse the music-SAE of Singh et al. [15] without retraining.

3.2 Prediction Model

Our main neural network is a three-branch *hybrid* architecture that consumes all three intrinsic signals together (Figure 1). Each signal $s \in \{\ell, H, z\}$ is fed to a dedicated 1D convolutional encoder f_s that produces a per-branch representation $r_s \in \mathbb{R}^{d_r}$. The three branch representations are combined by late concatenation,

$$r = [r_\ell;\, r_H;\, r_z], \tag{3}$$

and a shared Multi-Layer Perceptron (MLP) head maps r to the predicted clip-level rating $\hat{y} \in [1, 5]$. Single- and two-signal ablations drop branches accordingly; encoder sizes and training are in Section 4.3.

3.3 Training Objective Across Datasets

We train all variants of the neural network against a single per-clip target $y_i \in [1, 5]$ with one mean-squared-error objective,

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{4}$$

regardless of supervision format. MusicEval and SongEval supply y_i directly as the per-clip average expert rating. MusicPref, AIME and Music Arena release only pairwise win/loss/tie outcomes, which we convert to per-clip continuous ratings with the Elo scheme [29]: each competitor's scalar rating is updated after every match by an amount proportional to the gap between observed and rating-implied outcomes, and iterating across all matches yields a system-level continuous ranking analogous to chess. This is the same kind of system-level ranking AIME [4] and the

Group	Model	MusicEval		SongEval		AIME		MusicPref		Music Arena		All benchmarks	
		r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Baseline	Aesthetics-CE	0.62	0.64	0.57	0.57	0.41	0.38	0.04	0.03	0.24	0.26	0.20	0.21
	Aesthetics-CU	0.62	0.64	0.63	0.70	0.48	0.45	0.04	0.04	0.22	0.24	0.18	0.19
	Aesthetics-PC	0.08	0.04	0.32	0.27	0.01	0.01	0.01	0.00	0.15	0.23	0.07	0.09
	Aesthetics-PQ	0.56	0.59	0.61	0.65	0.42	0.40	0.03	0.03	0.33	0.40	0.20	0.23
	Mean Loss	-0.02	-0.07	0.24	0.11	-0.10	-0.09	-0.30	-0.41	-0.08	-0.07	-0.08	-0.13
Single	SINGLE-LOSS	0.55/0.55	0.56/0.55	0.60/0.58	0.61/0.61	0.58/0.62	0.56/0.60	**0.95**/0.95	**0.90**/0.89	0.66/0.63	0.71/0.71	0.64/0.67	0.62/0.65
	SINGLE-ENTROPY	0.66/0.63	0.68/0.65	0.66/0.72	0.66/0.73	0.69/0.66	0.68/0.62	0.94/0.95	0.89/0.90	0.66/0.63	0.74/0.70	0.70/0.69	0.70/0.67
	SINGLE-SAE	0.79/0.77	0.80/0.79	0.85/0.85	0.85/0.84	0.82/0.85	0.82/0.85	0.78/0.98	0.71/0.93	0.78/0.79	0.84/0.84	0.73/0.81	0.73/0.80
Hybrid	LOSS+ENTROPY	0.59/0.57	0.61/0.59	0.69/0.65	0.70/0.67	0.67/0.67	0.63/0.64	0.94/0.96	0.90/0.91	0.67/0.66	0.74/0.73	0.63/0.69	0.62/0.67
	LOSS+SAE	0.78/**0.79**	0.80/**0.80**	0.85/**0.86**	0.83/0.85	0.84/**0.86**	0.84/**0.87**	0.94/0.99	0.86/0.94	0.78/**0.80**	0.83/**0.86**	0.79/0.82	0.78/0.81
	ENTROPY+SAE	0.78/0.78	0.79/0.79	**0.87**/0.86	**0.87**/**0.85**	**0.85**/0.85	**0.85**/0.85	0.89/0.99	0.85/0.94	0.77/0.79	0.84/0.85	0.78/**0.83**	0.79/**0.81**
	LOSS+ENTROPY+SAE	**0.79**/0.77	**0.80**/0.79	0.84/0.84	0.83/0.83	0.82/0.85	0.81/0.84	0.94/**0.99**	0.89/**0.94**	**0.79**/0.79	**0.84**/0.85	**0.81**/0.82	**0.80**/0.80

Table 1. Pearson r and Spearman ρ vs. human evaluation; trained cells are MusicGen-small / large. Trained rows: held-out test split per benchmark; Audiobox rows: zero-shot on the full set; *All benchmarks*: union of the five held-out splits. Per size, **bold** is the column best and underline those tied with it (bootstrap, $p \geq 0.05$ [28]).

Chatbot-Arena-style Music Arena [5] also report across music generation systems—a property the standard alternatives do not share: pairwise margin (hinge/ReLU) [30] and RankNet [31] losses are training objectives rather than data conversions, and Bradley–Terry [32] fits the same competitor-strength likelihood as Elo but only as a batch MLE, with no per-match increment of the kind we use to assign each clip its own rating from a single battle. Per-dataset details are in Section 4.2.

4. EXPERIMENTS

We apply the methods of Section 3 on five human-evaluation datasets, comparing against audio-only and model-intrinsic baselines on held-out test splits along the pipeline shown in Figure 2.

4.1 Datasets

We use five public human-evaluation benchmarks. MusicEval [1] and SongEval [2] release per-clip 1–5 expert ratings; MusicPref [3], AIME [4], and Music Arena [5] release pairwise win/loss/tie outcomes (Music Arena adds a both-bad option). We keep only the musicality dimension—Musical Impression, Overall Musicality, Music Quality, Musicality, and Music Arena's overall-preference vote, respectively—and drop orthogonal axes (text–audio alignment, fidelity, vocal naturalness, song-structure clarity). Splits are train/val/test with fixed seeds, stratified by generator system. Together the five benchmarks span $\approx$241 h of rater-aligned audio: MusicEval (2,748 clips, $\approx$17 h), SongEval (2,399 clips, $\approx$140 h), AIME (1,300 clips, $\approx$3.6 h), MusicPref (5,040 clips, $\approx$20 h), and Music Arena (6,080 clips, $\approx$61 h).

4.2 Data Processing

Per-clip targets. MusicEval and SongEval provide continuous ratings directly: we average the musicality axis across annotators ($\sim$5 experts for MusicEval, 4 for SongEval) to obtain $y_i \in [1, 5]$. For the three pairwise benchmarks we apply the Elo scheme of Section 3.3 [29]. AIME exposes 12 comparisons per clip, enough to fit per-clip Elo directly. MusicPref and Music Arena expose only one battle per clip, so we first fit Elo at the system level (each generator plays many matches) and assign each clip its system rating plus a single-match adjustment $K\,(S - E)$, avoiding collapse into three win/tie/loss bins. Here S is the observed score (win 1, tie 0.5, loss 0; Music Arena's both-bad enters as -0.5 for both clips, ranking it below a loss), E the rating-implied expectation, and K the update step (24 at the system level, 16 per clip). All Elo ratings are robust-affinely mapped to the shared 1–5 scale. This rescoring preserves ordinal content: per-clip/system Elo tracks empirical win rate at $r \geq 0.86$ across all three benchmarks.

Clip windowing. Feature extraction must operate on exactly the audio each rater heard. MusicEval ($\sim$30 s) and MusicPref ($\sim$10 s) use the native released clip. AIME uses the rater-aligned 10-s window. SongEval rates full songs (95–598 s, median 205 s): we use the full released audio under a 360-s safety cap, which trims 31 of the 2,399 clips (0.28% of the audio). For Music Arena we keep audio up to the shortest of reported listening time, released duration, and 180 s, dropping raters under 2 s and systems with fewer than 60 clips (6,080 remain). Both corpora are then split into non-overlapping 30-s chunks; we concatenate the per-chunk loss / entropy / SAE feature streams and uniformly mean-pool along time to T=1500 frames. The 30-s chunk matches the 30-s / 1500-token context of MusicGen-Small [17]; T=1500 is the fixed CNN input length (Section 4.3). Loss and entropy curves use four parallel codebook channels plus a binary validity mask $m_t \in \{0, 1\}$, giving $(5, T)$ per curve and $(9, T)$ when both are stacked.

4.3 Model Training

Architecture. Loss and entropy are from MusicGen-Small (teacher-forced); SAE codes are layer-12 activations via Singh et al. [15] (d_{in}=1024, d_{sae}=4096). A 1D CNN maps the stacked $(9, 1500)$ loss/entropy curves (four codebook channels each plus mask) to $\mathbb{R}^{128}$; a parallel path compresses SAE codes 4096$\to$256, applies two stride-2 Conv1d blocks, and pools to $\mathbb{R}^{128}$. A two-layer MLP on the concatenation (Eq. (3)) yields $\hat{y} \in [1, 5]$. The full LOSS+ENTROPY+SAE hybrid has $\approx$ 1.24 M parameters ($\sim$84% in the SAE compress); single- and two-signal variants reuse the same encoders.

Configurations. For every benchmark we train the seven feature combinations in Table 1—the three single-

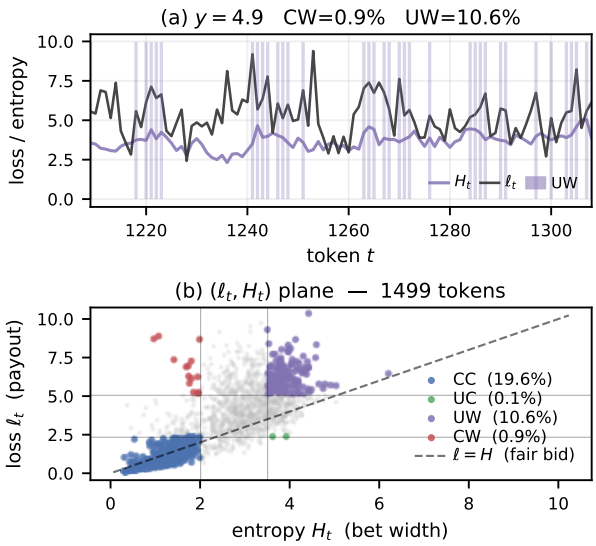

Figure 3. Example clip (y=4.9). **(a)** 100-token window with the densest co-occurrence of CW and UW tokens. **(b)** Full-length (ℓ_t, H_t) wager plane for the same clip.

signal variants, the three pairwise hybrids, and the full LOSS+ENTROPY+SAE hybrid—under identical splits, preprocessing and seeds, so cross-row comparisons isolate the effect of the input signals. Training minimises $\mathcal{L}_{\text{reg}}$ (Eq. (4)) with Adam (lr 10^{-4}, batch 8–64) and early-stops on the best-validation checkpoint.

4.4 Baselines

We compare the hybrid neural network of Section 3.2 and its single- and two-signal ablations against two reference groups. The first is *audio-only aesthetics*: every test clip is scored by Meta Audiobox Aesthetics on its four axes—Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), Production Quality (PQ)—with a per-axis affine map $\hat{z}_i = \text{clip}(az_i + b, 1, 5)$ fit on validation only. The second is *mean loss*: a minimal model-intrinsic baseline that regresses the human rating on the clip's scalar mean token loss.

4.5 Results

We report Pearson r (linear agreement on the 1–5 scale) and Spearman ρ (rank agreement) between predicted and human scores. Trained models use each benchmark's held-out test split; Audiobox is zero-shot on the full annotated set; and the *All benchmarks* column merges the five held-out splits into one evaluation set. Within each column and model size, we **bold** the best score and underline those statistically tied with it (paired clip-level bootstrap, $p \geq 0.05$ [28]).

Every intrinsic model beats the Audiobox aesthetics axes and the mean-loss baseline on the same clips, with SINGLE-SAE the strongest single source except on MusicPref, where loss leads. Loss and entropy are individually predictive but add little beyond SAE: the curves' temporal structure, not a scalar summary, carries the signal—without ever processing the audio.

per-clip descriptor	r	ρ
$\tilde{\pi}_{\text{UC}}$ (uncertain-correct)	$-0.12^{\dagger}$	$-0.12^{\dagger}$
$\tilde{\pi}_{\text{CW}}$ (overconfident-wrong)	$-0.18^{\ddagger}$	$-0.19^{\ddagger}$
$\tilde{\pi}_{\text{UW}}$ (uncertain-wrong)	$+0.29^{✠}$	$+0.32^{✠}$
$\tilde{\pi}_{\text{UW}} - \tilde{\pi}_{\text{CW}}$	$+0.25^{✠}$	$+0.27^{✠}$
$\overline{\ell - H}$ (excess surprise)	$+0.30^{✠}$	$+0.35^{✠}$

Table 2. Loss–entropy wager (significance of r and ρ). † $p<0.05$; ‡ $p<0.01$; $^{✠}$ $p<0.001$.

Combining everything is not always best. Grouping models by significance in each column (paired clip-level bootstrap [28]; bold/underline in Table 1), every benchmark has a top group set significantly apart from the rest, yet that group usually holds several tied models: its single best beats the runner-up in only 4 of 24 columns—three on the small model, one on the large. The group's makeup shifts by benchmark—SAE-based models lead on most, while on MusicPref loss and entropy lead instead—and SINGLE-SAE reaches the top group on most benchmarks, more consistently at large scale. SAE carries most of the signal, and no combination dominates.

5. ANALYSIS

Beyond the aggregate predictive performance reported in Section 4.5, we ask *what* each intrinsic signal reads about the generator. We run three short analyses, each on the MusicEval clean test split (n=270 clips), where loss, entropy and SAE-latent curves are aligned on the same token timeline.

5.1 Loss–Entropy Calibration: Prediction as a Wager

Following Huron's probabilistic account of musical anticipation [12], we treat each generation step as a wager whose *bet width* is summarised by the predictive entropy H_t, and whose *payout* is the token loss ℓ_t; for a categorical next-token distribution these are the usual entropy / cross-entropy (negative log-likelihood) pair [33]. Splitting each clip at the 25th and 75th percentiles (Q_{25}/Q_{75}) of its own ℓ_t and H_t distributions labels each token with one of four *wager scenarios*—confident-correct (CC: low ℓ, low H), uncertain-correct (UC: low ℓ, high H), overconfident-wrong (CW: high ℓ, low H), and uncertain-wrong (UW: high ℓ, high H); tokens inside the central interquartile band are unlabelled. Each clip's *scenario share* $\tilde{\pi}_q$ is the fraction of its labelled tokens in scenario q. Figure 3 illustrates the four scenarios on a clean-good example.

Across all MusicEval clean test clips (n=270), three findings emerge from Pearson r / Spearman ρ between $\tilde{\pi}_q$ and human rating (Table 2). **Overconfident-Wrong** ($\tilde{\pi}_{\text{CW}}$) depresses rating (r=−0.18, p<0.01): the model bet narrow on one token yet was surprised—a local signature of audible distortion. **Uncertain-Wrong** ($\tilde{\pi}_{\text{UW}}$) lifts rating (r=+0.29, p<0.001): the model had already spread its bet across several candidates yet was *still* surprised—wide bet *and* high payout together mark moments the

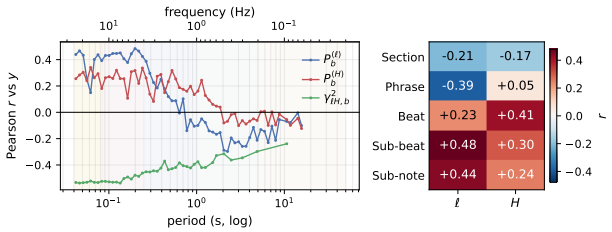

Figure 4. Spectral correlates of rating: Pearson r between per-band power of ℓ_t, H_t and y across log-spaced rate bands. **Left:** 64 fine bands. **Right:** five-band roll-up.

model could not foresee yet a listener accepts. **Uncertain-Correct** ($\tilde{\pi}_{\mathrm{UC}}$) reverses sign ($r{=}{-}0.12$, $p{<}0.05$): wide bet but low payout—uncertainty without surprise does not signal quality. A blind Gemini-2.5-Flash audit corroborates: $6/10$ CW windows are audibly disruptive vs. $3/10$ UW. The contrast $\tilde{\pi}_{\mathrm{UW}}{-}\tilde{\pi}_{\mathrm{CW}}$ collapses both effects into a single scalar ($r{=}{+}0.25$, $p{<}0.001$), and the simplest clip-level summary, $\overline{\ell{-}H}^{(i)} = \frac{1}{T_i}\sum_t(\ell_t - H_t)$ [33], reaches $r{=}{+}0.30$ ($p{<}0.001$): higher-rated clips are systematically more surprising than the model's own uncertainty predicts.

5.2 Temporal–Spectral Study

Section 5.1 summarised each clip by the distribution of (ℓ_t, H_t) values; here we analyse the curves' temporal dynamics. We apply an FFT to each curve $s_t \in \{\ell_t, H_t\}$, writing $F_s(k)$ for the Fourier magnitude at frequency bin k, and define the relative band power $P_b^{(s)} = \sum_{k\in b}|F_s(k)|^2 / \sum_k |F_s(k)|^2$ and $\log(1{+}P_b^{(s)})$ is the per-clip regressor against rating y. Bins are log-spaced between 0.05 and 25 Hz, rolled up into five named bands (*Section* ∼5–50 s, *Phrase* ∼1–5 s, *Beat* ∼0.25–1 s, *Sub-beat* ∼80–250 ms, *Sub-note* ∼40–80 ms). In addition we estimate the magnitude-squared coherence $\gamma^2_{\ell H}(f) \in [0,1]$ between the two curves using Welch's averaged-periodogram method [34]—a frequency-resolved measure of how linearly ℓ_t predicts H_t at rate f. Fig. 4 reports the per-band Pearson r against y: (left) 64 fine bands for $P_b^{(\ell)}$, $P_b^{(H)}$, and $\gamma^2_{\ell H,b}$ and (right) the same P_b rolled up to the five named bands.

Three musical patterns emerge. (i) **Rapid phrase-rate change is bad** ($r{\approx}{-}0.39$ in the *Phrase* band): a surprise track that keeps re-starting every few seconds reads as formally unstable. (ii) **Beat-rate variation is good** ($r{\approx}{+}0.41$ for H at the beat, $r{\approx}{+}0.48$ for ℓ at the sub-beat): a clear rhythmic pulse aligns with perceived groove. (iii) **Sub-note ℓ–H coupling is bad** ($r{\approx}{-}0.54$ across 40–190 ms): surprise and uncertainty oscillating together at the millisecond grain is the signature of glitchy audio.

5.3 SAE Latents: Good vs. Bad vs. Neutral Concepts

To characterise *what* these latents encode—and to verify that the SAE branch's predictions rest on audible musical traits rather than dataset artifacts—we compute for each selective SAE latent k a per-clip gradient attribution

rank	ρ_k	VLM caption	producer concept
good-aligned ($\rho_k > 0$)			
g-1	+0.54	classical, string	string classical
g-2	+0.49	folk, upbeat	folk guitar texture
g-5	+0.45	folk, calm	choir + plucked folk
g-7	+0.44	folk, melancholic	flute texture
g-9	+0.43	ambient, calm	ambient pads + vocal tex.
g-10	+0.43	pop, upbeat	clear pop with vocal+beat
near-null ($\lvert\rho_k\rvert < 0.01$)			
m$_+$3	+0.005	rock, energetic	loud rock, single long note
m$_+$1	+0.001	pop, energetic	clear melody+accomp.
m$_-$1	−0.001	classical, melanch.	harsh sound, bad groove
m$_-$3	−0.002	hip-hop, energetic	bad groove, noise
bad-aligned ($\rho_k < 0$)			
b-9	−0.43	classical, playful	dissonant wide-range
b-8	−0.45	— (silence)	EOS / delay tokens
b-7	−0.46	classical, melanch.	dissonant violin texture
b-5	−0.47	hip-hop, tense	random perc., harsh hi-freq
b-4	−0.48	pop, upbeat	bad vocal+noisy perc.
b-3	−0.48	— (silence)	EOS / delay tokens
b-2	−0.60	rock, energetic	big noise, random trumpet
b-1	−0.61	classical, calm	random hi-freq noise

Table 3. Representative SAE latents ranked by rating-aligned attribution ρ_k, each annotated by Gemini-2.5-Flash and by a professional music producer. Thin red/blue bars in the ρ_k column visualise $|\rho_k|$.

$A_k^{(i)} = \sum_t(\partial\hat{y}_i/\partial z_t^{(i,k)})\, z_t^{(i,k)}$ from the trained Single-SAE model and rank latents by $\rho_k = \mathrm{Pearson}(A_k^{(\cdot)}, y)$ ($365/4096$ latents are active enough to admit a finite ρ_k). The top-activating 2-s windows of three groups—good-aligned (10 largest $+\rho_k$), bad-aligned (10 largest $-\rho_k$), and near-null ($|\rho_k| < 0.01$, 5 per sign)—are captioned by Gemini-2.5-Flash [35] and, without seeing these captions or ρ_k, by a professional music producer[2] (Table 3). On the good side producer and VLM agree on conventional musical content; on the bad side the VLM still issues clean genre labels while the producer reaches for distortion terms (*noise*, *harsh hi-freq*, *EOS/delay tokens*). The producer's musical-vs.-broken judgement matches the ρ_k ranking, so gradient attribution recovers the same quality axis a trained listener would.

6. CONCLUSION

A music generative model's own intrinsic signals already track which of its outputs listeners rate highly. Loss, entropy, and SAE latents from a single frozen MusicGen align with human ratings on five public benchmarks and outperform audio-only aesthetics baselines: loss–entropy mismatch flags the overconfident-wrong moments that disrupt perceived quality, and SAE latents organise clips along a good–bad axis without ever seeing a rating. Intrinsic signals are thus a viable—and previously underused—axis for automatic music evaluation.

7. ACKNOWLEDGEMENTS

We thank Professor Alexander Lerch for his guidance and his review of this paper.

[2] Six years of professional experience in electronic, pop, and orchestral production.

8. REFERENCES

[1] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[2] J. Yao, G. Ma, H. Xue, H. Chen, C. Hao, Y. Jiang, H. Liu, R. Yuan, J. Xu, W. Xue, H. Liu, and L. Xie, "SongEval: A benchmark dataset for song aesthetics evaluation," *arXiv preprint arXiv:2505.10793*, 2025.

[3] Y. Huang, Z. Novack, K. Saito, J. Shi, S. Watanabe, Y. Mitsufuji, J. Thickstun, and C. Donahue, "Aligning text-to-music evaluation with human preferences," in *Proc. of the 26th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2025.

[4] F. Grötschla, A. Solak, L. A. Lanzendörfer, and R. Wattenhofer, "Benchmarking music generation models and metrics via human preference studies," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[5] Y. Kim, W. Chi, A. N. Angelopoulos, W.-L. Chiang, K. Saito, S. Watanabe, Y. Mitsufuji, and C. Donahue, "Music arena: Live evaluation for text-to-music," in *Advances in Neural Information Processing Systems 38 (NeurIPS), Creative AI Track*, 2025.

[6] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025.

[7] Y. Ma, S. Li, J. Yu, E. Benetos, and A. Maezawa, "CMI-Bench: A comprehensive benchmark for evaluating music instruction following," in *Proc. of the 26th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2025, pp. 416–425.

[8] Y. Ma, H. Xia, H. Gao, W. Chen, Y. Ye, Y. Yang, S. Chang, M. Ding, Y. Li, R. Yuan, S. Dixon, and E. Benetos, "CMI-RewardBench: Evaluating music reward models with compositional multimodal instruction," *arXiv preprint arXiv:2603.00610*, 2026.

[9] R. B. Zajonc, "Attitudinal effects of mere exposure," *Journal of Personality and Social Psychology*, vol. 9, no. 2, Pt.2, pp. 1–27, 1968.

[10] E. G. Schellenberg, I. Peretz, and S. Vieillard, "Liking for happy- and sad-sounding music: Effects of exposure," *Cognition and Emotion*, vol. 22, no. 2, pp. 218–237, 2008.

[11] L. B. Meyer, *Emotion and Meaning in Music*. Chicago, IL: University of Chicago Press, 1956.

[12] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[13] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot arena: An open platform for evaluating LLMs by human preference," in *Proc. of the 41st Int. Conf. on Machine Learning (ICML)*, 2024, pp. 8359–8388.

[14] A. Lerch, C. Arthur, N. Bryan-Kinns, C. Ford, Q. Sun, and A. Vinay, "Survey on the evaluation of generative models in music," *ACM Computing Surveys*, vol. 58, no. 4, 2025.

[15] N. Singh, M. Cherep, and P. Maes, "Discovering and steering interpretable concepts in large generative music models," in *Int. Conf. on Learning Representations (ICLR)*, 2026.

[16] M. Müller-Eberstein, R. van der Goot, B. Plank, and I. Titov, "Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[17] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

[18] X. Li, C. Liu, and Z. Wang, "When noise lowers the loss: Rethinking likelihood-based evaluation in music large language models," *arXiv preprint arXiv:2602.02738*, 2026.

[19] P. Veličković, F. Barbero, C. Perivolaropoulos, S. Osindero, and R. Pascanu, "Perplexity cannot always tell right from wrong," *arXiv preprint arXiv:2601.22950*, 2026.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. of the 34th Int. Conf. on Machine Learning (ICML)*, 2017, pp. 1321–1330.

[21] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. E. Hinton, "Regularizing neural networks by penalizing confident output distributions," in *Int. Conf. on Learning Representations (ICLR), Workshop Track*, 2017.

[22] Z. Qiu, Z. Ou, B. Wu, J. Li, A. Liu, and I. King, "Entropy-based decoding for retrieval-augmented large language models," *arXiv preprint arXiv:2406.17519*, 2024.

[23] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," in *Int. Conf. on Learning Representations (ICLR)*, 2021.

[24] M. Fomicheva, S. Sun, L. Yankovskaya, F. Blain, A. Tamchyna, M. Fishel, N. Popović, M. López, V. Chaudhary, A. Fan, Y. Graham, B. Haddow, M. Huck, A. N. Valentino, M. Turchi, and L. Specia, "Unsupervised quality estimation for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 639–657, 2020.

[25] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," in *Int. Conf. on Learning Representations (ICLR)*, 2024.

[26] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah, "Towards monosemanticity: Decomposing language models with dictionary learning," Transformer Circuits Thread, 2023. [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features

[27] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet," Transformer Circuits Thread, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity

[28] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[29] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing, 1978.

[30] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN)*, 1999, pp. 97–102.

[31] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005, pp. 89–96.

[32] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[34] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[35] Gemini Team, Google, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," arXiv preprint arXiv:2507.06261, 2025.