A Hierarchical Framework for Automatic Human-Centric Musical Affective Design

Frederico G. Pedrosa[1]

[1]Federal University of Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brazil

Abstract

Traditional computational models for music affective recognition primarily rely on bidimensional frameworks, assuming orthogonal independence between Valence and Arousal. However, these linear architectures consistently fail to differentiate affective states in low-energy states. This study investigates the validity of the Hierarchical Model of Affect (HMA) by contrasting algorithmic acoustic decoding (MERT and CLAP architectures) with human perceptual construction across three distinct databases (n = 200, n = 1,802, n = 216). Methodological triangulation across four mathematical paradigms of variance decomposition, network analysis, reflective modeling, and formative modeling identified a substantive and unipolar General Factor (GF) that precedes qualitative differentiation. Results demonstrate that while state-of-the-art Audio Transformers operate as linear magnitude detectors (r = 0.99 between signal and construction), human perception exhibited a functional dissociation within this framework in Serene states, where listeners actively suppress physical magnitude to construct relaxation. The proposed HMA framework significantly outperformed the traditional 2D model in predictive accuracy (ΔAIC=86), yielding a 17 percent gain in valence prediction within high-density datasets. Furthermore, Generalized Linear Mixed Models revealed that the perception of the intensity scaffold is significantly modulated positively by the General Factor of Personality and negative by musical expertise ($p < 0.001$), suggesting these traits as relevant parameters for user-centered affective design. We conclude that human-centric computational design requires a transition from signal-mirroring architectures to hierarchical systems that independently manage global informational magnitude and qualitative residuals. These findings provide a calibrated blueprint for the development of generative and evaluative creative tools that align algorithmic output with the idiosyncratic sensitivity of the receiver.

Keywords

Computational Design, Affective Computing, Hierarchical Model of Affect, Music Emotion Recognition, Music Information Retrieval

1. Introduction

The study of affective structure has been primarily dominated by the bidimensional circumplex model, which organizes emotional states into a space defined by the orthogonal dimensions of Valence and Arousal (2D) [1, 2]. While this paradigm has facilitated significant advancements in Music Information Retrieval (MIR) and Creative Design [3], it presents persistent limitations in the discrimination of low-arousal states. Specifically, computational systems often struggle to distinguish between quadrants of positive and negative valence when activation levels are low, as both Serenity and Sadness [1] are frequently subsumed into a singular low energy category [5, 6].

This limitation is largely rooted in the linear operationalization of arousal within MIR design, where acoustic magnitude, measured through proxies such as Loudness and Spectral Centroid, is treated as a direct correlate of perceived activation [7, 8]. However, recent meta-analyses of Music Emotion Recognition (MER) reveal a persistent "accuracy ceiling" in this paradigm, where Arousal is predicted with significantly higher precision than Valence [9]. This performance gap suggests a fundamental misalignment between current linear computational architectures and the non-linear, hierarchical structure of human affect [10, 3].

Computational Design and Computer-aided Creativity (CDCC) @ ICCC'26, June 29–July 03, 2026, Coimbra, Portugal
fredericopedrosa@ufmg.br (F. G. Pedrosa)
https://github.com/FredPedrosa/Intensity_DT (F. G. Pedrosa)
0000-0002-0682-0734 (F. G. Pedrosa)

[1]Within the scope of this work, the Happiness quadrant corresponds to high valence and high arousal; Fear/Anger to low valence and high arousal; Sadness to low valence and low arousal; and Serenity to high valence and low arousal [4].

Addressing these challenges suggests the benefit of a shift from raw signal processing toward a psychologically grounded computational design. We propose the Hierarchical Musical Affective (HMA) framework, which posits that affective intensity is a foundational organizational structure [11] that acts as a gatekeeper for qualitative nuances. This approach aligns with modern constructionist views, where emotions are emergent meaning-making acts constructed from a continuous neurophysiological Core Affect [12, 13].

Through a three-stage methodological triangulation, this research investigates whether a hierarchical architecture can resolve the systematic scarcity of affective representations and the resulting misclassification of states in the low-arousal, low-valence quadrant — where nuanced emotions often collapse into a generic category of low energy — termed the Melancholy Gap [14]. By contrasting human perceptual data and large-scale international benchmarks with state-of-the-art AI architectures (MERT and CLAP), while evaluating individual differences within the Brazilian musical landscape, we aim to provide a more parsimonious design for human-centric musical affective systems.

2. Related Work

2.1. Limitations of the Circumplex Model in Computational Design

The 2D model imposes a significant structural constraint on affective modeling [5]. Although seminal psychometric studies identified a dominant unipolar factor in affective data, where all items exhibit high positive loadings, this component was traditionally relegated to a secondary status. It was either treated as a measurement artifact to be statistically suppressed through procedures such as ipsatization [1, 15] or defined as a purely mathematical derivative of Euclidean distance [11]. To resolve this, we draw upon the Theory of Constructed Emotion, which proposes that affective states are emergent meaning-making acts constructed from Core Affect: a foundational, continuous neurophysiological state [12, 13].

Within this constructionist framework, the General Factor (GF)[2] of affect is re-evaluated as the primary organizational scaffold for qualitative distinctions. Recent meta-analyses in Music Emotion Recognition (MER) corroborate this hierarchical necessity, revealing a persistent "accuracy ceiling" where Arousal is predicted with significantly higher precision ($r = 0.81$) than Valence ($r = 0.67$) [9]. This performance gap suggests that current models may behave as "horses" in music content analysis—referencing the Clever Hans effect [19]—where systems appear to solve complex affective tasks while actually relying on simpler, confounding low-level cues like acoustic magnitude to track energy.

Neuroscientific evidence further supports this functional dissociation, demonstrating that the brain represents affective intensity and valence in spatially nonoverlapping subpopulations of voxels connected to distinct large-scale networks [20]. Specifically, intensity is anchored in the ventral attention (salience) network, while valence is processed via limbic and default mode networks. These findings validate intensity as a fundamental and independent dimension of core affect, providing a biological basis for the HMA framework's claim that global saliency must be modeled independently from qualitative residuals.

2.2. Measurement Paradigms

Following Russell's (1980) original triangulation, we employed Principal Component Analysis (PCA) to investigate the latent structure of affective ratings. While Russell utilized Multidimensional Scaling (MDS) for semantic similarity tasks, he explicitly relied on PCA to identify the primary dimensions of self-reported affect, demonstrating that both techniques converge on a nearly identical two-dimensional

[2]The concept of a General Factor originates from Spearman's landmark work on intelligence [16], identifying a single underlying dimension for positive correlations between diverse tasks. This hierarchical architecture has been extended to domains such as the General Factor of Personality (GFP) [17] and the *p*-factor in psychopathology [18]. In this study, we identify a similar unipolar factor of saliency anchoring the valence-arousal manifold.

circumplex ($r > .94$). In our study, PCA was prioritized to isolate the General Factor (GF) of variance (1980).

However, traditional affective modeling has predominantly relied on **reflective measurement models**, which assume that a latent emotional state is the primary cause of observed indicators [21]. Under this reflective tradition, Confirmatory Factor Analysis (CFA) remains the standard statistical practice for estimating latent traits by modeling the common variance shared among manifest responses. Yet, when these paradigms encounter a dominant GF, it is historically treated as a descriptive artifact or methodological noise — often isolated via PCA and statistically suppressed to maintain the theoretical independence of Valence and Arousal [1]. This approach forces the common factor model as a gold standard, which may be restrictive when dealing with complex psychological constructions where causality is not necessarily unidirectional [22].

In contrast, recent advancements in **Network Psychometrics** propose that affective states emerge from the mutual interactions and reinforcing dynamics between observable indicators, rather than latent causes [23]. Tools such as Exploratory Graph Analysis (EGA) and its hierarchical version (hEGA) enable the identification of modular clusters, revealing how affective qualities organize into stable communities [24]. Complementing this, **Partial Least Squares Structural Equation Modeling (PLS-SEM)** represents a shift toward a causal-predictive paradigm based on composite models [22]. While reflective models assume that the construct exists independently of its measures, the formative-composite perspective offered by PLS-SEM aligns with the Theory of Constructed Emotion [12]. Within this framework, affect is not the underlying cause of symptoms, but the emergent result of integrating neurophysiological signals and contextual markers. Consequently, the GF of affect is re-evaluated as a primary metric of saliency, an intensity that dictates the construction of the emotional experience.

Mathematically, PLS-SEM is uniquely suited for this investigation as it maximizes the explained variance (R^2) of dependent constructs, outperforming covariance-based methods in identifying relationships within complex hierarchical models where the goal is theory development and predictive accuracy rather than simple covariance fitting [22]. By integrating the four paradigms for variance isolation (PCA), systemic organization (EGA), reflective modeling (CFA), and formative construction (PLS-SEM), this research provides a methodological triangulation. This architecture allows us to investigate whether affective structure remains invariant across fundamentally different mathematical frameworks, moving computational design from passive signal mirroring to an active modeling of human perceptual construction.

2.3. Generative Psychometrics and Latent Representations

The emergence of Large Language Models (LLMs) has birthed Generative Psychometrics, where high-dimensional embeddings can act as synthetic participants to interrogate internalized cultural knowledge [25, 26]. While Audio Transformers (e.g., MERT, CLAP) are adept at identifying dense acoustic patterns [27, 28], they often function as mechanical sensors that map signal magnitude to arousal in a strictly linear fashion. Empirical evidence shows that even multimodal pipelines demonstrate a 70% misclassification rate in specific affective categories [3].

2.4. The Challenge of Linear Bias in Audio Transformers

State-of-the-art models like MERT (Music undERstanding model with large-scale self-supervised Training; [28]) and CLAP (Contrastive Language-Audio Pretraining; [27]) have significantly improved MIR tasks through self-supervised learning. However, these models are primarily optimized for signal reconstruction, which reinforces a linear dependency on acoustic magnitude. In human listening, the relationship between physical magnitude and subjective impact is often dissociated, particularly in low-arousal states, a phenomenon linear models are ill-equipped to represent. A meta-analysis of MER studies (2014–2024) confirms a persistent accuracy ceiling in this paradigm, where arousal is predicted with significantly higher precision (r = 0.81) than valence (r = 0.67) [9]. This performance gap, and the fact that simpler linear models often outperform deep neural networks in regression [10, 29], points to a

fundamental misalignment between current computational architectures and the qualitative structure of human affect. Moving beyond raw signal processing towards a psychologically grounded computational design necessitates accounting for latent organizational layers that may govern how qualitative nuances are anchored in the listener's perception.

3. Methodology

This research was conducted through a three-stage methodological framework designed to investigate the structural organization of musical affect. The investigation transitioned from a controlled experimental setting to large-scale computational generalization and, finally, to a methodological triangulation focused on individual differences.

3.1. Study 1: Controlled Experimental Discovery

The initial stage utilized the human perceptual data from the Musical Emotion Evaluation Test (MEET; [30]). The sample consisted of 200 participants, aged 18 to 43 years ($Mean = 27.27$; $SD = 5.90$), strategically balanced between professional musicians ($n = 100$) and non-musicians ($n = 100$). Following the consensus in music psychology [31], musicians were defined as individuals with at least six years of formal music training, recruited from higher education music institutions. The non-musician group consisted of individuals with no history of formal music education; those with less than two years of informal practice were retained in this category to ensure a clear distinction between professional/advanced proficiency and general listeners. Participants evaluated musical stimuli through a forced-choice emotional identification task and rated the dimensions of Valence and Arousal using the Self-Assessment Manikin (SAM) pictorial scale on a 1-to-9 range. From an original pool of 116 stimuli designed to represent the affective quadrants, 38 excerpts validated through Item Response Theory (IRT) were selected for this investigation [4]. This selection ensured the inclusion of clear prototypes for Happiness, Fear/Anger, Sadness, and Serenity.

The analytical procedure followed a multi-paradigm approach to identify the primary structure of these ratings: a) An unrotated PCA was applied to the unified matrix of 76 indicators (38 stimuli across 2D) to evaluate the presence of a dominant, unipolar organizational component; b) We tested the structural fit (CFA) of standard reflective models to determine if qualitative affective states could be identified as independent latent causes of the responses; c) To investigate the modular organization of perception, EGA was used to identify how affective qualities cluster into stable, non-linear communities; d) Finally, we employed a two-stage hierarchical formative model with PLS-SEM. For that, in the first stage, the EGA-identified modules were modeled as first-order formative constructs; in the second stage, these modules were integrated to form a second-order GF.

To enable a direct structural comparison between human perception and algorithmic decoding, the 38 musical stimuli were also processed through the MERT deep learning model to extract 768-dimensional acoustic embeddings. These representations were subjected to the same analytical pipeline employed for the human ratings (CFA, PCA, EGA, and PLS-SEM) to identify potential structural similarities. This parallel design was intended to reveal whether the machine's native latent architecture follows the same hierarchical organization as human listening or if it relies on a different logic of signal magnitude.

This combination of methods allowed for a rigorous comparison between reflective and formative architectures, as well as AI and human affective representation, testing whether if they are better represented as independent coordinates or as a mediated hierarchy organized by a primary magnitude detector.

To establish nomological validity within the PLS-SEM framework – a requirement for formative measurement – the hierarchical General Factor (GF) was modeled to predict specific criterion variables. For the human dataset, the GF was used to predict Emotional Identification Accuracy (the participants' ability to correctly identify the target affective quadrant). In the parallel AI architecture, the GF was tested for its ability to decode and categorize musical stimuli from the MEET database. This predictive

design ensures that the emergent hierarchical structure is a functional representation of affective decoding.

3.2. Study 2: Large-Scale Computational Generalization

The second stage scaled the previous findings to a high-dimensional computational context using the Database for Emotional Analysis in Music (DEAM; [6]). This dataset comprises 1,802 musical excerpts from diverse Western popular genres (e.g., rock, pop, electronic, country, and jazz), providing a benchmark for dynamic Music Emotion Recognition. The ground-truth annotations were crowdsourced through Amazon Mechanical Turk (MTurk), with each excerpt rated by a minimum of 5 to 10 unique workers who passed qualification tests to ensure their understanding of the valence-arousal space. In total, the dataset consolidates thousands of subjective reports, providing a statistically stable reference for human affective perception across a wide demographic of online listeners. To determine whether the hierarchical organization offers a more efficient architecture for automatic design than the 2D model, we performed a predictive duel across two state-of-the-art AI architectures: MERT for acoustic features and CLAP for multimodal (semantic-acoustic) representation.

This stage employed comparative mixed-effects modeling to establish which framework provides superior informational parsimony and predictive accuracy across a broad range of musical genres. By contrasting the HMA with traditional linear assumptions in a large-scale dataset, we aimed to verify the robustness of the GF when applied to automated Music Emotion Recognition (MER) tasks, specifically testing its ability to resolve classification challenges in high-density acoustic manifolds.

3.3. Study 3: Methodological Invariance and Individual Calibration

The third stage involved a new database of 216 participants to evaluate individual calibration and methodological stability. The research protocol was conducted in accordance with the Declaration of Helsinki and observed the Brazilian National Health Council resolutions CNS/MS 510/16 and 466/2012. Institutional ethical approval was obtained, and all participants provided informed consent. The sample was characterized by a diverse demographic profile regarding gender (116 female, 82 male, 2 non-binary, and 1 trans participant), age of ($M = 39.56$ years, $SD = 14.66$), education (30 participants held Doctorates, 25 held Master's degrees, 113 had Undergraduate or Specialized degrees, and 48 had completed High School) and musical background (80 participants with over five years of formal musical study and 91 with no prior music education). To ensure ecological relevance, musical excerpts were selected to represent each of the five dimensions of musical preferences according to the RITMO [3].

Individual differences in the receivers were mapped through standardized psychometric scales to account for variability in affective perception. Personality traits were assessed using the Big Five Inventory-2 (BFI-2; [32]), utilizing the Brazilian Portuguese adaptation [33], which allowed for the extraction of the general factor of Personality (GFP). Mental health indicators of distress were measured using the Depression, Anxiety, and Stress Scale (DASS-21; [34, 35]) to derive a general factor of psychological distress (p-factor). To capture granular musical affective responses, we developed a specific instrument consisting of 10 items mapped onto the dimensions of the circumplex model. Participants rated their perception of Arousal (Active, Energized, Attentive, Relaxed, and Sleepy) and Valence (Pleasurable, Happy, Good Mood, Unpleasant, and Sad) on a 5-point Likert-type scale for each musical excerpt.

This stage aimed to stress the stability of the HMA framework through a methodological triangulation involving four distinct analytical paradigms to extract the organizational structure of affect: variance decomposition (PCA), network analysis (hEGA), reflective modeling (CFA), and formative modeling (PLS-SEM). To ensure that the GF was not an artifact of a specific statistical approach, we evaluated the convergent validity between scores derived from these frameworks. This multi-method approach provided the basis for a comparative performance analysis using mixed-effects models, aiming to identify the predictive stability of human traits and expertise over the latent affective architecture.

3.4. Sample independence

The samples for Study 1 and Study 3 were entirely independent, with no overlapping participants. While the human perceptual data, in Studies 1 and 3, was collected within the Brazilian cultural context, the framework's robustness was cross-evaluated in Study 2 using the international DEAM dataset [6]. This inclusion supports the generalizability of the HMA framework, suggesting that its predictive gains represent broader properties of human-AI affective alignment rather than idiosyncratic cultural artifacts.

3.5. Item Calibration and Refinement

To ensure the structural integrity of the affective architecture, we performed a preliminary diagnostic stage to evaluate the contribution of each indicator to the GF. Based on the psychometric performance and directional alignment, the measurement models were refined. For example, in PLS-SEM the item Sleepy was identified as a non-significant indicator of the GF and was excluded from the final structural models. Additionally, the item Relaxed was inverted to maintain unipolar consistency with the magnitude of the signal. This recursive refinement process ensured that the methodological triangulation was performed on a statistically stable set of indicators, reflecting the core dimensions of activation and pleasure without the interference of redundant or misaligned variables.

3.6. Analysis Strategy and Model Selection

The analytical pipeline was distributed between Python for acoustic signal processing and R (v.4.5.0, [36]) for structural modeling. The latent organization of affect was estimated through four distinct mathematical paradigms to ensure methodological invariance: (1) Frequentist (PCA for variance decomposition); (2) Network (hEGA); (3) Reflective (CFA); and (4) Formative (PLS-SEM). The convergent validity of these representations was verified through correlation analysis of the resulting scores across 2,160 observations (216 participants evaluating 10 stimuli each).

The quality of the models was assessed against established psychometric thresholds. For reflective models, we targeted a Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) above 0.90, and a Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) below 0.08—metrics that indicate the degree of similarity between the observed and predicted covariance matrices. For the formative-hierarchical models, reliability and validity were established through Cronbach's alpha (α) and composite reliability (ρ_C) greater 0.70, alongside Average Variance Extracted (AVE) above 0.50 to ensure the convergent validity of affective modules. For network-based models, structural stability was verified through bootstrap procedures with a threshold of 0.75, ensuring the identified modular clusters are not artifacts of sampling error.

Comparative model performance was evaluated using Generalized Linear Mixed Models (GLMM), incorporating random intercepts for participants and stimuli to account for the nested and idiosyncratic nature of the data. To identify the most parsimonious architecture, we utilized the Akaike Information Criterion (AIC) — which balances model fit against complexity — and Akaike weights, that quantify the probability of a given model be the best representation among the candidates. This selection process was further supported by the variance explained by fixed factors (Marginal R^2), the variance explained by the full model, including individual differences (Conditional R^2), and the Root Mean Square Error (RMSE) to assess predictive accuracy. This multi-metric approach ensures that the chosen framework best bridges the gap between physical signal magnitude and human perception.

4. Results

4.1. Study 1: Structural Identification and the Collapse of Reflective Modeling

The investigation of human ratings in the MEET database provided the first empirical evidence of the inadequacy of traditional linear-reflective designs for musical affect. Initial attempts to validate the

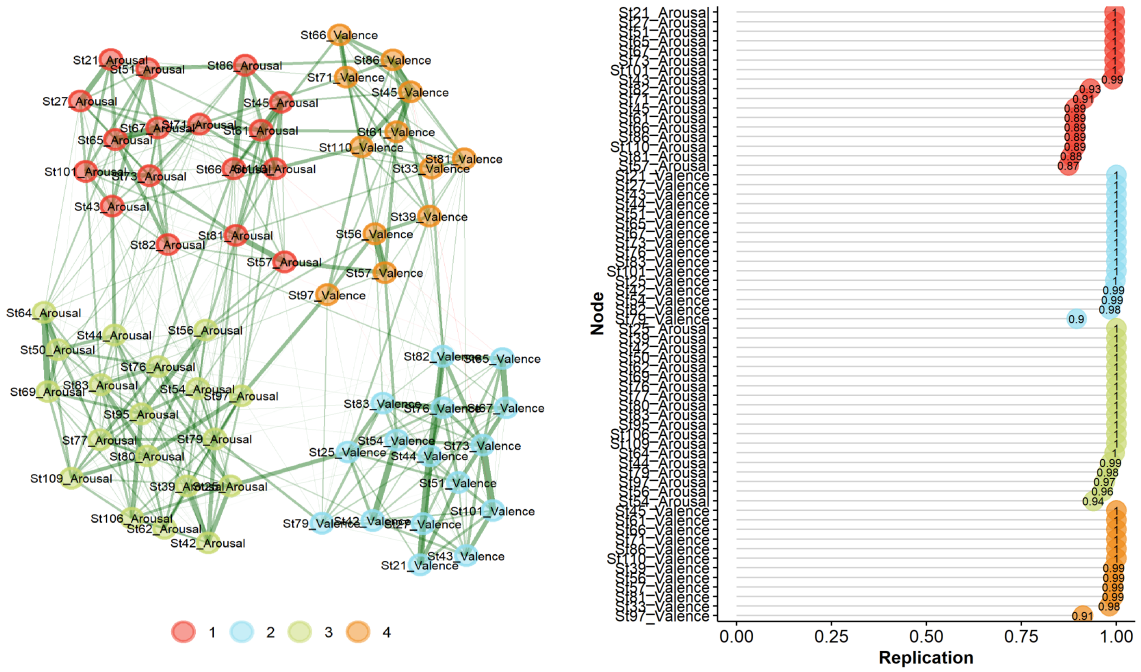

Figure 1: Network Topology and Stability of Human Affective Modules.

Note: (Left) Exploratory Graph Analysis (EGA) identifying four stable modular communities in human perception: (1) High Energy, (2) Aversion, (3) Low Energy, and (4) Pleasure. (Right) Stability results from bootEGA (500 iterations) demonstrating high structural replicability (stability ≈ 1.00) for the identified clusters.

bidimensional structure through CFA revealed systemic inconsistencies (*CFI* and $TLI < 0.90$; *RMSEA* and $SRMR > 0.10$). Furthermore, models specifying Valence and Arousal as independent latent causes frequently failed to achieve mathematical convergence or produced unstable parameters (e.g., negative variances), indicating that the reflective assumption via reflexive modeling is statistically ill-suited for this data.

In contrast, the unrotated PCA identified a dominant organizational structure. The first principal component (PC1) emerged as a unipolar dimension, explaining 24.5% of the total variance, with all 76 indicators presenting positive loadings ($M = 0.469$, $SD = 0.161$). This finding demonstrates that human perception initially integration the signal's magnitude into a primary metric. As shown in Table 1, while PC1 is strictly unipolar, PC2 and PC3 exhibit bipolar distributions, representing the qualitative nuances that emerge only after the isolation of the GF.

Table 1
Descriptive Statistics of Factor Loadings (PCA) for the First Three Components.

Component	Mean (M)	SD	Min.	Max.	Range	Prop. Var.
PC1	0.469	0.161	0.124	0.718	0.594	24.5%
PC2	0.102	0.362	-0.403	0.649	1.052	10.7%
PC3	-0.003	0.220	-0.424	0.377	0.801	6.6%

Note: PC1 reflects the General Factor (GF) of Intensity, characterized by strictly positive loadings across all 76 indicators. PC2 and PC3 represent the qualitative nuances that emerge after intensity isolation. The full loading matrix for all indicators will be provided in a supplementary repository.

EGA results indicated that human processing is subdivided into four functionally distinct modular communities: (1) High Energy, (2) Aversion/Negativity, (3) Low Energy, and (4) Pleasure/Positivity. The stability of these modules demonstrated absolute structural replicability (1.00 for all nodes). This suggests that while there is a unified product (PC1), the brain segments the information into stable

categorical niches to manage Valence and Arousal independently.

The PLS-SEM demonstrated substantive nomological validity, with the GF serving as a significant predictor of Emotional Identification Accuracy ($\beta = 0.534$, $p < 0.001$, $R^2 = 0.285$; Figure 2). Bootstrap analysis confirmed the stability of the formative weights.

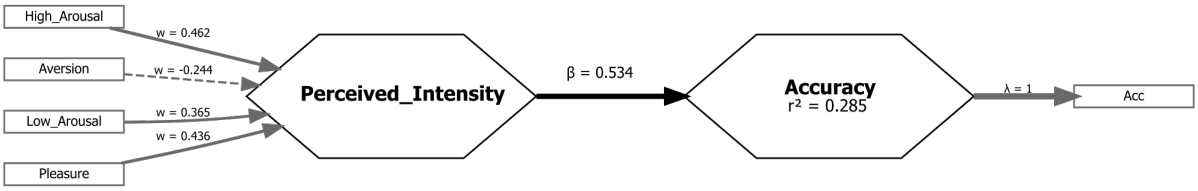

Figure 2: Hierarchical Formative Model of Perceived Intensity.

Note: Perceived Intensity (General Factor of Affect in Music) is formed by the four modular communities and acts as a significant predictor of Emotional Identification Accuracy. Dashed lines indicate the non-significant contribution of the Aversion module to the primary intensity scaffold.

The correlation analysis between signal components and formative modules revealed a highly specialized structural mapping, rather than a generalized redundancy. While the first principal component ($PC1_{human}$) functioned as a significant energy gatekeeper for most dimensions ($p < 0.001$), it showed no significant relationship with the Aversion module ($p = 0.129$). This suggests that negative affective saliency is independent of the primary acoustic magnitude.

Furthermore, the qualitative specialization of the modules was confirmed by the selective significance of the residual components. The Pleasure module demonstrated a specific alignment with $PC3_{human}$ ($p = 0.001$), while showing no dependency on lower-order variance such as $PC5_{human}$ ($p = 0.683$). In contrast, the High Arousal module was the only construct sensitive to the full range of signal variance, including $PC5_{human}$ ($p < 0.001$). This pattern of selective significance supports the hypothesis that the HMA framework successfully identifies how different "acoustic sensors" in the human mind are tuned to specific dimensions of the musical signal, allowing for a more nuanced and accurate computational design than traditional linear models (Table 2).

Table 2
Correlation Matrix between HMA Formative Modules and Human Principal Components.

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$
General Factor	0.663*	-0.422*	0.529*	-0.166*	-0.070
High Arousal	0.449*	-0.233*	0.158*	0.060	0.362*
Aversion	-0.130	0.192*	-0.113	0.222*	0.119
Low Arousal	0.390*	-0.240*	0.001	-0.157	0.062
Pleasure	0.349*	0.249*	0.219*	-0.243*	-0.026

Note: PC = Principal Component. *Significant at $p < 0.05$.

4.1.1. Algorithmic Acoustic Architecture (MERT)

Unlike the human results, PCA performed on the AI embeddings showed that the $PC1_{AI}$ explains 87.69% of the total acoustic variance. This component is strictly unipolar, with all 38 musical stimuli presenting positive loadings ranging from 0.14 to 0.17. These results identify the AI's primary processing layer as a dimension of a GF, which captures the total sonic energy and informational density of the signal (Table 3).

The lack of modular granularity in the AI was further confirmed by Exploratory Graph Analysis (EGA), which identified a single global community with 100 percent stability across bootstrap samples.

The algorithmic construction of affect demonstrated a high structural compression compared to human perception. While the human system operates through four distinct and balanced functional

Table 3
Descriptive Statistics of Factor Loadings (PCA) for the MERT Algorithmic Architecture.

Component	Mean	*SD*	Min.	Max.	Range	Prop. Var.
PC1	0.935	0.039	0.801	0.971	0.170	87.7%
PC2	0.004	0.153	-0.209	0.386	0.595	2.3%
PC3	0.001	0.118	-0.249	0.468	0.718	1.4%

Note: SD = Standard Deviation; Min. = minimum; Max. = maximum; Prop. Var. = proportional variation.

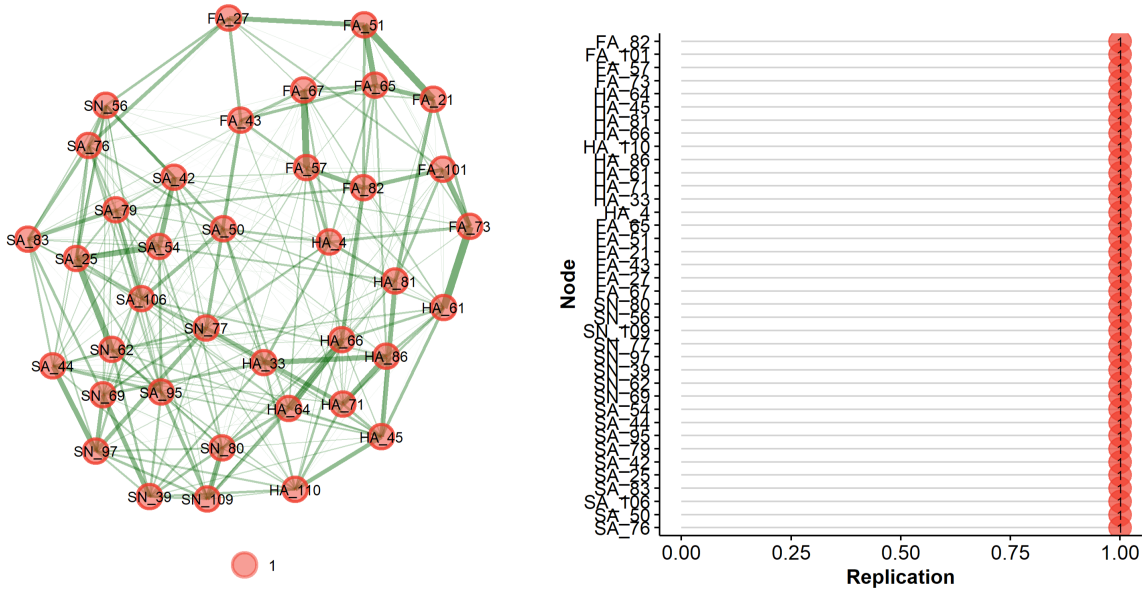

Figure 3: Network Topology and Stability of MERT Algorithmic Representations.

Note: (Left) Exploratory Graph Analysis (EGA) identifying a single, monolithic community in the AI latent space. (Right) Stability results from bootstrap samples demonstrating the lack of modular granularity. FA = Fear/Anger; HA = Happy; SN = Serenity; SA = Sadness.

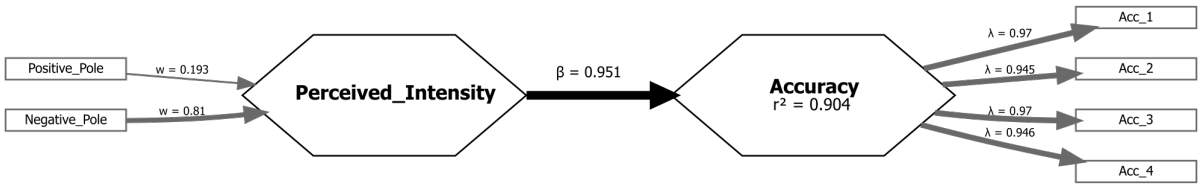

Figure 4: Hierarchical Formative Model of Algorithmic Intensity (MERT).

Note: PLS-SEM validation for the AI architecture. The model exhibits extreme redundancy ($R^2 = 0.904$), functioning as a near-perfect mirror of acoustic magnitude. Note the weights primarily driven by the Negative Pole ($w = 0.81$) compared to the non-significant contribution of the Positive Pole, suggesting the AI functions primarily as a tension detector.

modules, the PLS-SEM analysis of the MERT model revealed a monolithic architecture. Even when forced into a two-pole structure (Positive vs. Negative), the AI demonstrated high correlation between poles ($r = 0.997$, $p < 0.001$), treating all signals as a single vector of magnitude.

Crucially, the formative weights showed a bias towards GF's construction in the AI was driven, exclusively by the Negative Pole ($weight = 0.88$, $p < 0.05$), while the Positive Pole contributed no significant weight ($weight = 0.12$, $p = 0.68$). This suggests that state-of-the-art Audio Transformers function primarily as tension detectors, being unable to integrate aesthetic reward into their saliency mapping.

The triangulation of the MERT model revealed an architecture of absolute redundancy, where mathematical signal decomposition and affective construction are virtually identical. The correlation between the first principal component ($PC1_{AI}$) and the hierarchical formative GF reached a robust closeness ($r = 0.996$, $p < 0.001$). This high correlation indicates that, unlike humans, who share only 43.5% of variance between signal and perception, the AI architecture closely reflects the acoustic signal's magnitude, leaving zero room for subjective or mediated construction.

Furthermore, the redundancy between the positive and negative poles ($r = 0.997$) indicates a total collapse of affective categories into a single vector of force. While human modules are specialized and stable, the AI's architecture is unable to distinguish between pleasure-driven or energy-driven magnitude, treating variations as raw informational density. Finally, the near-zero correlation between the AI's formative constructs and the qualitative residuals ($PC2_{AI}$ and $PC3_{AI}$) suggests that the machine shows limited sensitivity to qualitative nuances that humans tend to decouple from the GF.

4.1.2. Cognitive Dissociation

The results of GLMM testing the interaction between AI GF and Intended Emotion confirmed that AI Intensity is a significant general predictor of the human response ($b = 0.11$, $t = 5.55$, $p < 0.001$). However, a critical functional dissociation was identified in low-arousal states. The interaction analysis revealed that the Serenity category significantly moderates this relationship, presenting a negative coefficient ($b = -0.36$, $t = -10.99$, $p < 0.001$). As shown in the interaction plot (Figure 5), while Fear/Anger, Happiness, and Sadness exhibit positive slopes, in which increased acoustic magnitude leads to higher perceived intensity, the slope for Serenity is sharply inverted.

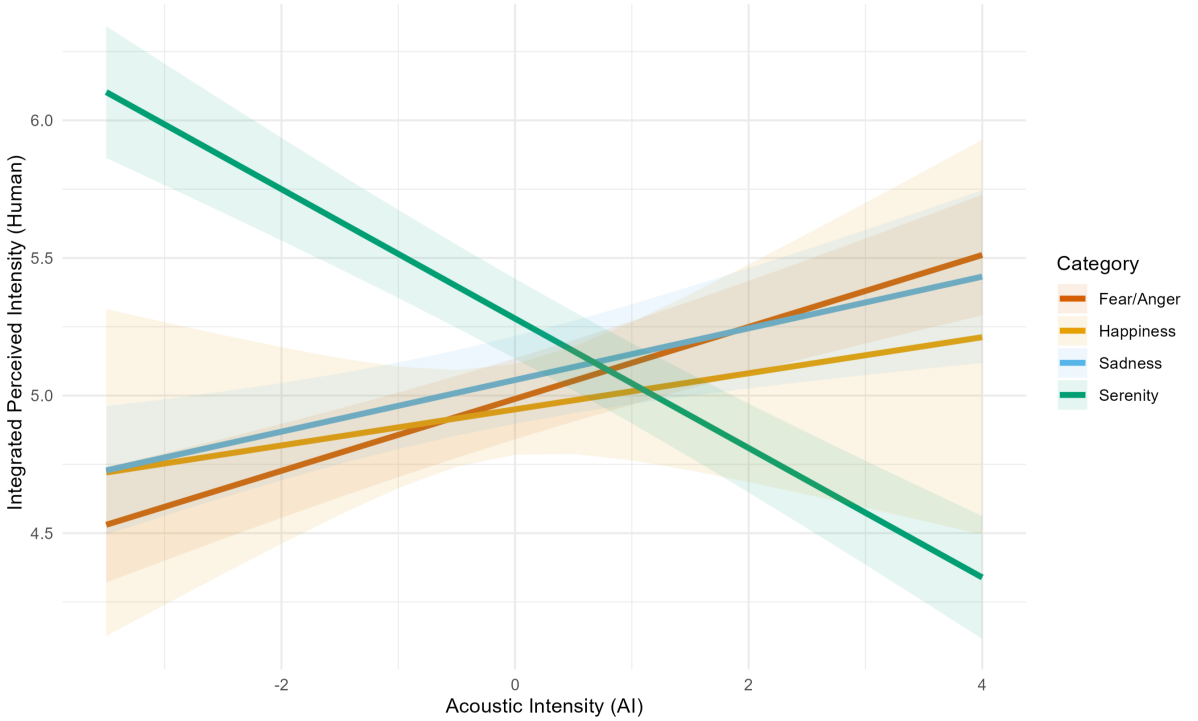

Figure 5: Functional Dissociation and the Serenity Filter.

Note: Shaded areas represent 95% Confidence Intervals (CI).

While fixed effects of signal and category explained only a marginal portion of the variance ($R^2_{marginal} = 0.014$), the inclusion of individual participant variability increased the model's explanatory capacity to 37.1% ($R^2_{Conditional}$). This disparity suggests that, although the hierarchical organization of intensity is a stable structural feature, the final magnitude of the experience is governed by idiosyncratic filters. These results demonstrate that current linear AI models suffer from an difficulty in disentangle low-energy

states, failing to account for the categorical suppression mechanisms that humans employ to distinguish peace from melancholy.

The test of intensity as a Euclidean distance from the neutral center showed that, although the relationship is statistically significant due to the large sample size, the correlation was weak ($r = 0.20$, $p < 0.001$) and the geometric model explained only 3.6% of the perceived intensity variance ($R^2_{marginal} = 0.036$).

4.2. Study 2: Large-Scale Generalization and Predictive Improvement

The comparative performance between the traditional 2D model and the HMA framework on the DEAM dataset ($n = 1,802$) is presented in Table 4. The HMA framework outperformed the Russell 2D model across both AI architectures, with an Akaike weight > 0.999, indicating a high probability of being a superior representation of the data.

For the acoustic-only architecture (MERT), the HMA provided a 17% relative gain in valence prediction accuracy (R^2 increase from 0.219 to 0.256; $\Delta AIC = 86$). In the multimodal architecture (CLAP), which incorporates semantic features, the HMA also yielded a significant improvement ($R^2 = 0.241$; $\Delta AIC = 32.1$). While the inclusion of semantic-acoustic embeddings in CLAP already provided a higher baseline for valence compared to MERT, consistent with recent surveys on the importance of lyrics and high-level descriptors for valence prediction [37], the hierarchical decoupling of intensity remained a critical factor for model optimization.

Table 4
Comparison of Model Performance Indices on the DEAM Dataset.

Architecture	Model	AIC	Akaike Weights	R^2	RMSE
MERT	Russell (2D)	5254.3	< 0.001	0.219	1.037
	HMA (Hierarchical)	5168.3	> 0.999	0.256	1.012
CLAP	Russell (2D)	5237.5	< 0.001	0.226	1.033
	HMA (Hierarchical)	5205.4	> 0.999	0.241	1.023

Note: AIC = Akaike Information Criterion; R^2 = Coefficient of Determination; RMSE = Root Mean Square Error.

4.3. Study 3: Methodological Invariance and Human Calibration

The correlation between scores derived from PCA, hEGA, CFA, and PLS-SEM revealed a stable latent architecture, with all coefficients remaining high (between $r = 0.94$ and 0.98, $p < 0.001$). This methodological invariance provides relevant evidence that this GF is a substantive structural property of musical affect.

The predictive superiority of the HMA was further tested through a comparison of performance indices using GLMM. As shown in Table 4, the formative-hierarchical approach (PLS-SEM) provided the best fit for the data. The statistical evidence in favor of the HMA framework is noteworthy. The PLS-SEM model reached an Akaike weight that suggests it is highly probable to be the most accurate representation of the data compared to reflective, network, or frequentist alternatives.

As shown in Table 5, the results of predictive stability of human traits over the extracted intensity scores reveal a convergence: the predictors of affective intensity remain consistent regardless of the extraction method employed.

The stability analysis identifies the GFP as the primary modulator of the affective experience. Across all extraction methods, GFP emerged as a significant and positive predictor ($p \leq 0.001$), suggesting that individuals with higher personality organization perceive the Intensity GF with greater sensitivity. Furthermore, a consistent linear effect of education was identified ($p < 0.02$), suggesting that academic expertise facilitates a more calibrated anchoring of the intensity signal.

Another stable finding was the negative impact of age on intensity perception, which remained significant in most models. This suggests a progressive sensory or cognitive attenuation of the affective

Table 5
Comparative Performance of Intensity Extraction Methods in GLMM.

Model	AIC	Akaike Weights	$R^2_{conditional}$	$R^2_{marginal}$	RMSE
PLS-SEM (Formative)	5133.8	> 0.999	0.510	0.056	0.678
CFA (Reflective)	5156.1	< 0.001	0.501	0.047	0.685
PCA (Variance)	5187.0	< 0.001	0.492	0.061	0.691
hEGA (Network)	5267.4	< 0.001	0.465	0.057	0.710

Note: AIC = Akaike Information Criterion; $R^2_{conditional}$ = Variance explained by both fixed and random effects; $R^2_{marginal}$ = Variance explained by fixed effects only; RMSE = Root Mean Square Error.

signal over time. Crucially, the Intraclass Correlation Coefficients (ICC) ranged from 0.44 to 0.48, indicating that nearly half of the variance in intensity perception is tied to the individual listener rather than the acoustic stimulus. Beyond that, the statistical support for the PLS-SEM model suggests that a human-centric computational design must be stratified.

Table 6
Stability of Predictors across Different Extraction Methods (GLMM Results).

	PCA			hEGA			CFA			PLS-SEM		
Predictors	Est	SE	p	Est	SE	p	Est	SE	p	Est	SE	p
(Intercept)	0.09	0.19	0.635	0.11	0.19	0.564	0.14	0.19	0.481	0.06	0.20	0.772
GFP	0.17	0.05	<.001	0.15	0.04	0.001	0.16	0.05	<.001	0.17	0.05	<.001
Age	-0.01	0.00	0.064	-0.01	0.00	0.054	-0.01	0.00	0.023	-0.00	0.00	0.121
Education [L]	-0.30	0.11	0.005	-0.26	0.10	0.011	-0.30	0.11	0.005	-0.31	0.11	0.004
Sex [male]	0.18	0.08	0.025	0.15	0.08	0.045	0.18	0.08	0.020	0.19	0.08	0.023
Random Effects												
σ^2		0.52			0.55			0.52			0.51	
ICC		0.46			0.43			0.46			0.47	
$R^2_{Marginal}$		0.068			0.056			0.074			0.072	
$R^2_{Conditional}$		0.492			0.466			0.501			0.511	

Note: Est = Parameter Estimate (Beta); SE = Standard Error; p = p-value; σ^2 = Residual variance; ICC = Intraclass Correlation Coefficient; $R^2_{marginal}$ = Variance explained by fixed effects; $R^2_{Conditional}$ = Total variance explained by the model.

4.3.1. Formative Affective Model

The initial measurement model, incorporating the full set of five items for both Arousal (Active, Energized, Attentive, Relaxed, and Sleepy) and Valence (Pleasurable, Happy, Good Mood, Unpleasant, and Sad), showed significant internal inconsistencies. Specifically, the Arousal construct failed to meet established reliability thresholds ($\alpha = 0.48$, $\rho_C = 0.61$), suggesting that a raw lexical mapping is insufficient for capturing the latent structure of musical intensity.

Diagnostic analysis identified "Sleepy" as a non-significant indicator and "Relaxed" as having a misaligned directional effect. To improve the model's calibration, "Sleepy" was removed, and "Relaxed" was inverted to align with the unipolarity of the Intensity. Similarly, in the Valence construct, the items "Unpleasant" and "Sad" exhibited near-zero weights, suggesting a hedonic asymmetry in how participants construct global saliency from these stimuli.

The optimized formative model resulted in a substantial increase in reliability and explanatory power. The final Arousal construct (Active, Energized, Attentive, and Inverted Relaxed) achieved good internal consistency ($\alpha = 0.86$, $\rho_C = 0.90$, $AVE = 0.70$), as well as the refined Valence construct (Pleasurable, Happy, Good Mood, and Unpleasant) ($\alpha = 0.87$, $\rho_C = 0.90$, $AVE = 0.72$). The structural relationship between these refined modules provided an R^2 of 0.726, indicating that 72.6% of the variance

in qualitative construction is anchored in the energetic-hedonic interplay. These refinements suggest that human-centric affective design requires a calibrated formative architecture that moves beyond simple dictionary-based categorizations to capture the emergent nature of musical impact.

Face validity of the proposed hierarchical structure, assessed through a correlation analysis between the raw additive sums of the affective items and the latent scores generated by the formative PLS-SEM model, revealed a robust association. The correlation for the Arousal construct reached $r = 0.985$ ($p < 0.001$), while Valence showed $r = 0.967$ ($p < 0.001$). Furthermore, the general Intensity Factor showed a high convergence with the total sum of all indicators ($r = 0.982$, $p < 0.001$).

5. Discussion

The results of this investigation suggest that current computational models of musical affect suffer from a structural misalignment with human perception. By contrasting the monolithic architecture of Artificial Intelligence with the modular and hierarchical construction of human listeners, we identify a blueprint for the next generation of human-centric affective design.

5.1. The Ontological Nature of the General Factor

The central find of this investigation is the identification of a dominant GF that precedes the qualitative distinctions of the circumplex model. Our triangulation across four mathematical paradigms (PCA, hEGA, CFA, and PLS-SEM) provides robust evidence for its substantive nature. We argue that this GF represents Affective Intensity (or Saliency) based on three convergent evidence: (1) its strictly unipolar nature across all studies; (2) its role as an energetic gatekeeper, sharing 44% of its variance with the signal's first principal component ($r = 0.66$); and (3) its near-perfect alignment with the acoustic component of AI models ($r = 0.99$). Crucially, this biological dissociation is supported by neuroimaging evidence showing that the brain represents intensity and valence in spatially nonoverlapping subpopulations of voxels [20]. This identifies intensity not as a geometric byproduct of valence and arousal, but as a foundational layer anchored in the ventral attention (Salience) network, preceding qualitative construction.

5.2. Intensity as a Foundation, Not a Coordinate

A major question was whether this GF of Intensity could be reduced to a geometric byproduct of the 2D plane (Euclidean distance). Our results provide a negative answer. The weak correlation between the GF and the Vector Theory of Intensity ($r = 0.20$, explaining only 3.6% of variance) suggests that Intensity is not merely an emergent property of valence and arousal. Instead, our data supports a hierarchical architecture where Intensity is proposed as the principal structure of the Hierarchical Affective Model. This may explain why the reflective CFA model failed to converge in Study 1: qualitative dimensions tend to collapse when the foundational Intensity layer is not properly accounted for.

5.3. The Serenity Paradox

The most critical evidence for the HMA framework lies in the identified functional dissociation within low-arousal states. General Linear Mixed Models revealed that while acoustic magnitude (PC1) is a positive predictor of perceived arousal for Sadness ($b = 2.44$, $p < .001$), this relationship is significantly moderated and inverted for Serenity (interaction $b = -2.72$, $p = .038$).

As illustrated in the interaction analysis (Figure 5), human listeners employ an active a cognitive mechanism that suppresses physical magnitude to construct a state of relaxation—the "Serenity Filter". In contrast, current AI models like MERT function as strictly linear magnitude detectors, failing to distinguish between the stillness of peace and the heaviness of melancholy. This finding provides a mathematical explanation for the Melancholy Gap: traditional models collapse these categories because they cannot account for the non-linear suppression humans apply to the intensity scaffold.

Consequently, effectively designing human-centric affective systems requires architectures capable of managing global informational magnitude and qualitative residuals as independent, non-linear parameters.

5.4. Design Implications for Human-Centric AI

The gain in valence prediction achieved by the HMA framework in Study 2 suggests that the accuracy "ceiling" in MER [9] can be addressed through architectural shifts. For computational design, this necessitates:

1. **Hierarchical Decoupling:** Generative systems should treat Intensity (GF) and qualitative nuances (Residual 2D) as independent control parameters.
2. **Listener Calibration:** Since the perception of the GF is modulated by personality (GFP) and expertise, creative tools should be "receiver-aware." Professional tools should prioritize control over qualitative residuals, whereas consumer-level systems should focus on the stability of the hierarchical structure.

By transitioning from signal mirroring to affective construction, we provide a framework that aligns computational magnitude with the subjectivity of the human listener, aiming to ensure that the machine's creative intent matches the felt meaning of the experience.

5.5. Establishing a Reliable Metric for Hierarchical Affect

A critical requirement for computational design is the identification of a reliable and reproducible metric. Our results indicate that the GF of Intensity in music is a substantive dimension that can be consistently extracted through multiple paradigms. The methodological triangulation showed that scores derived from frequentist, network, and formative models share over 90 percent of their variance ($r > 0.94$).

For the development of creative tools, we identify PLS-SEM in its formative specification (Mode B) as a plausible architecture, given its informational parsimony. However, the high correlation ($r > 0.96$) between the formative scores and simple additive sums suggests that the HMA framework is also highly intuitive. This implies that designers can implement high-resolution back-end models using formative weighting while maintaining streamlined user interfaces based on additive inputs, without sacrificing the structural integrity of the affective measurement.

5.6. MERT as a Manifestation of the Acoustic Latent Space

The behavior of the MERT model, explaining 87.69% of the total variance in a single unipolar component, suggests that its primary latent space functions as a high-fidelity manifestation of informational magnitude. This high degree of variance compression, coupled with the near-perfect correlation between signal and construction ($r = 0.99$), aligns with the Clever Hans effect described by Sturm [19]. It appears that the AI achieves accuracy in arousal primarily by acting as a mechanical sensor of sound density. This mapping may explain the persistent **accuracy ceiling** in MER meta-analyses, where arousal is predicted with significantly higher precision than valence [9]. Current architectures appear to reach a performance plateau because they mirror signal force rather than modeling the non-linear suppression mechanisms that humans employ to bridge the Melancholy Gap.

5.7. Multimodal Integration and the Valence Challenge

The results from the CLAP architecture in Study 2 indicate that multimodal representations offer a slight but significant gain in valence prediction accuracy (6.6%) compared to purely acoustic models. This improvement is consistent with recent surveys suggesting that the inclusion of semantic information is critical for addressing the valence "accuracy ceiling" [37]. While the DEAM dataset contains a substantial proportion of vocal tracks, the CLAP model's advantage stems from its ability to leverage semantic-acoustic embeddings that capture high-level associations beyond raw signal. However, even

with these multimodal inputs, the 2D model remains prone to categorical collapse in low-arousal states. We conclude that while semantic cues can refine valence, a hierarchical decoupling of intensity is still required to ensure that generative tools can distinguish the quietude of peace from the stillness of melancholy.

6. Conclusion

By identifying the General Factor (GF) of intensity as a foundational organizational structure, we provide a scalable blueprint for receiver-aware computational design. Our findings suggest that overcoming the Melancholy Gap in MER requires a transition from signal-mirroring architectures to hierarchical systems capable of replicating human non-linear suppression mechanisms, particularly in serene states.

Ultimately, these results indicate that effective affective design must move beyond flat representations to independently manage global informational magnitude and qualitative residuals. Future research should prioritize the integration of HMA-based architectures into generative pipelines and explore the stability of these hierarchical layers across diverse cultural contexts and emerging neural architectures. By aligning algorithmic output with the idiosyncratic sensitivity of the listener, we ensure that the next generation of creative tools can finally match the felt meaning of the human experience.

Acknowledgments

This work was supported by the National Council for Scientific and Technological Development (CNPq) through the PIBITI/CNPq program and the internal grant EDITAL PRPq–06/2025. The author would like to thank the undergraduate researchers Yasmin Rezende and Suellen Martins for their valuable contribution to data collection. Special thanks to Professors Flávio Figueiredo and Lucas Nascimento from the Music Oriented Systems and AI for Creativity (MOSAIC) group, Professor Marília Nunes-Silva and Professor Carmen Flores-Mendoza from the Laboratory of Individual Differences Assessment (LADI) for their critical feedback and institutional support.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.5 in order to: perform sentence polishing and rephrasing to improve clarity and academic style, and assist with LaTeX and BibTeX formatting. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. A. Russell, A circumplex model of affect, Journal of Personality and Social Psychology 39 (1980) 1161–1178.

[2] J. Posner, J. A. Russell, B. S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Development and Psychopathology 17 (2005) 715–734.

[3] F. G. Pedrosa, song_sent_scores: Computational design for charting dynamic emotion in songs with a multimodal circumplex framework, in: Proceedings of the Hybrid Workshop on Computational Design and Computer-aided Creativity (CDCC @ ICCC'25), 2025.

[4] M. Nunes-Silva, A. R. Da Luz, C. G. Schlottfeldt, P. S. R. Martins, F. G. Pedrosa, Validation of musical stimuli for emotion evaluation: The musical emotion evaluation test, Journal of New Music Research (2025). doi:10.1080/09298215.2025.2607364.

[5] T. Eerola, J. K. Vuoskoski, A comparison of the discrete and dimensional models of emotion in music, Psychology of Music 39 (2011) 18–49.

[6] A. Aljanaki, Y.-H. Yang, M. Soleymani, Developing a benchmark for emotional analysis of music, PLoS ONE 12 (2017) e0173392. doi:10.1371/journal.pone.0173392.
[7] Y.-H. Yang, H. H. Chen, Music Emotion Recognition, CRC Press, 2012.
[8] A. Dash, K. Agres, Ai-based affective music generation systems: A review of methods and challenges, ACM Computing Surveys 56 (2024).
[9] T. Eerola, C. J. Anderson, A meta-analysis of music emotion recognition studies, ACM Computing Surveys 58 (2026) Article 248. doi:10.1145/3796518.
[10] S. Beveridge, D. Knox, Popular music and the role of vocal melody in perceived emotion, Psychology of Music 46 (2018) 411–423. doi:10.1177/0305735617713834.
[11] R. Reisenzein, Pleasure-arousal theory and the intensity of emotions, Journal of Personality and Social Psychology 67 (1994) 525–539.
[12] L. F. Barrett, How emotions are made: The secret life of the brain, Houghton Mifflin Harcourt, 2017.
[13] J. A. Russell, L. F. Barrett, Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant, Journal of Personality and Social Psychology 76 (1999) 805–819.
[14] Y. V. R. Rezende, S. V. Silva, R. H. A. Coelho, F. G. Pedrosa, Validade ecológica do modelo circumplexo de afeto: o campo semântico da música brasileira no youtube, in: Anais do 11º Nas Nuvens... Congresso de Música, 2025.
[15] A. Tellegen, D. Watson, L. A. Clark, On the dimensional and hierarchical structure of affect, Psychological Science 10 (1999) 297–303.
[16] C. Spearman, "general intelligence," objectively determined and measured, The American Journal of Psychology 15 (1904) 201–292. doi:10.2307/1412107.
[17] J. Musek, A general factor of personality, Journal of Research in Personality 41 (2007) 1213–1233. doi:10.1016/j.jrp.2007.02.003.
[18] A. Caspi, R. M. Houts, J. Belsky, S. J. Goldman-Mellor, H. Harrington, H. Hogan, P. Ramrakha, R. Poulton, T. E. Moffitt, The p factor: One general psychopathology factor in the structure of psychiatric disorders?, Clinical Psychological Science 2 (2014) 119–137. doi:10.1177/2167702613497473.
[19] B. L. Sturm, The "horse" inside: Seeking causes behind the behaviors of music content analysis systems, Computers in Entertainment (CIE) 14 (2017) 1–32. doi:10.1145/2967507.
[20] S. A. Lee, J.-J. Lee, J. Han, M. Choi, T. D. Wager, C.-W. Woo, Brain representations of affective valence and intensity in sustained pleasure and pain, Proceedings of the National Academy of Sciences (PNAS) 121 (2024) e2310433121. doi:10.1073/pnas.2310433121.
[21] J. F. Hair, G. T. M. Hult, C. M. Ringle, M. Sarstedt, A primer on partial least squares structural equation modeling (PLS-SEM), 3rd ed., Sage, 2022.
[22] M. Sarstedt, J. F. Hair Jr, C. M. Ringle, "pls-sem: indeed a silver bullet" – retrospective observations and recent advances, Journal of Marketing Theory and Practice 30 (2022) 261–272. doi:10.1080/10696679.2022.2056488.
[23] D. Borsboom, A. O. J. Cramer, Network analysis: An integrative approach to the structure of psychopathology, Annual Review of Clinical Psychology 9 (2013) 91–121.
[24] H. F. Golino, A. P. Christensen, EGAnet: Exploratory Graph Analysis, 2025. URL: https://doi.org/10.32614/CRAN.package.EGAnet, cRAN – R Project.
[25] L. L. Russell-Lassalandra, A. P. Christensen, H. Golino, Generative psychometrics via ai-genie: Automatic item generation and validation via network-integrated evaluation, PsyArXiv (2025). doi:10.31234/osf.io/fgbj4_v2.
[26] J. Suárez-Álvarez, Q. He, N. Guenole, D. D'Urso, Using artificial intelligence in test construction: A practical guide, Psicothema 38 (2026) 1–12. doi:10.70478/psicothema.2026.38.01.
[27] B. Elizalde, S. Deshmukh, M. A. Ismail, H. Wang, Clap: Learning audio concepts from natural language supervision, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.
[28] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, J. Fu, Mert: Acoustic music understanding model with large-scale self-supervised training, in: International Conference on Learning Representations (ICLR), 2024. URL: https://doi.org/10.48550/arXiv.2306.00107.

[29] L. Xu, Z. Sun, X. Wen, Z. Huang, C. Chao, L. Xu, Using machine learning analysis to interpret the relationship between music emotion and lyric features, PeerJ Computer Science 7 (2021) e785. doi:10.7717/peerj-cs.785.
[30] M. Nunes-Silva, Musical emotion evaluation test (meet): Secondary database, 2023.
[31] J. Zhang, M. Susino, G. E. McPherson, E. Schubert, The definition of a musician in music psychology: A literature review and the six-year rule, Psychology of Music 48 (2020) 389–409. doi:10.1177/0305735618804038.
[32] C. J. Soto, O. P. John, The next big five inventory (bfi-2): Developing tidy, hierarchical, and robust questionnaires to measure personality domains and facets, Journal of Personality and Social Psychology 113 (2017) 117–134.
[33] A. J. Passos, O. P. John, C. J. Soto, Adaptação brasileira do big five inventory-2 (bfi-2), Avaliação Psicológica 20 (2021) 50–61. doi:10.15689/ap.2021.2001.18956.05.
[34] S. H. Lovibond, P. F. Lovibond, Manual for the Depression Anxiety Stress Scales, 1995.
[35] R. C. B. Vignola, A. M. Tucci, Adaptation and validation of the depression, anxiety and stress scale (dass-21) among brazilian adults, Journal of Affective Disorders 155 (2014) 104–109. doi:10.1016/j.jad.2013.10.031.
[36] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2025. URL: https://www.R-project.org/.
[37] J. Kang, D. Herremans, Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges, IEEE Transactions on Affective Computing 16 (2025) 2545–2559. doi:10.1109/TAFFC.2025.3583505.