Music-Evoked GEMS-9 Emotions: A Statistical Analysis of the Emotion-to-Music Mapping Atlas (EMMA) Database

Bryan Im

Received August 18, 2025
Accepted September 15, 2025
Electronic access October 15, 2025

This study investigates mean differences among the nine dimensions of the Geneva Emotional Music Scale (GEMS-9) and determines which variables predict those ratings. Using the Emotion-to-Music Mapping Atlas (EMMA) database, an ANOVA confirmed the GEMS-9 model's effectiveness, with post hoc comparisons showing significant mean differences in over half of the emotion rating pairs. Regression analysis revealed a primary finding: liking and familiarity are distinct, non-interchangeable psychological factors. Liking predicts the aesthetic emotions of sublimity, while familiarity predicts the energetic emotions of vitality. This finding provides a more granular understanding of how different factors influence specific emotional dimensions. The findings have significant implications. Theoretically, this research suggests that future emotion models should incorporate liking and familiarity as separate constructs and account for the complex interactions between them. Methodologically, its findings highlight the need for more robust statistical methods and diverse datasets. Practically, the results can refine music information retrieval (MIR) and music emotion recognition (MER) models and guide applications in music therapy toward more personalized and accurate outcomes. The study's focus on classical music is a limitation that may affect the generalizability of the findings.

Keywords: Music, Emotion, Database, Geneva Emotional Music Scale (GEMS), Emotion-to-Music Mapping Atlas (EMMA), ANOVA, Regression, Python

1 Introduction

Music-evoked emotion is a rapidly expanding research area with broad implications for anthropology, psychology, neuroscience, musicology, music therapy, advertising, signal processing, and machine learning. Research suggests that music-evoked emotions involve a complex interaction of psychological, physiological, and neural processes, influenced by both individual and cultural factors. Despite theoretical frameworks [1,2] and significant empirical findings [3–5], disagreement persists regarding the definition, mechanisms, and characterization of music-evoked emotions.

A main challenge in this field is the lack of standardized terminology and methods. As a result, researchers often utilize general emotion models developed for non-musical emotions, such as the circumplex model of affect [6] or basic emotions theory [7]. Because these models are not designed to capture induced emotions, many music-evoked emotion studies resort to measuring perceived emotions. The circumplex model, while simple, lacks the granularity to differentiate nuanced emotional states. Similarly, the utilitarian emotion terms from basic emotions theory are not appropriate for detecting the aesthetic emotions induced by music [8,9]. The Geneva Emotional Music Scale (GEMS) [10] aims to overcome these limitations by providing a domain-specific, evidence-based taxonomy of music-evoked emotions designed to detect induced, rather than perceived, emotions. Derived from a factor analysis, the GEMS contains 45 musically significant emotion terms (GEMS-45) that can be categorized into nine first-order dimensions (GEMS-9), which are further grouped into three second-order factors in a hierarchical structure: (1) sublimity ("wonder," "transcendence," "tenderness," "nostalgia," and "peacefulness"); (2) vitality ("power" and "joyful activation"); and (3) unease ("tension" and "sadness").

While the GEMS model is highly capable of detecting music-evoked emotions, few databases have utilized it in music information retrieval (MIR) and music emotion recognition (MER). Emotify [11], an online game, collected emotion ratings for song excerpts using a modified GEMS framework, but the study's findings are limited by a lack of granularity due to the use of a binary scale. The Emotion-to-Music Mapping Atlas (EMMA) [12], a more recent online database built on the GEMS framework, provides a systematically organized platform for music and emotion data. Despite thorough GEMS-45 annotation, previous regression analyses were limited to higher-order dimensions (sublimity, vitality, unease), failing to reveal subtle differences between constituent GEMS-9 ratings. These limitations of previous GEMS-based studies directly motivated the objectives of this research.

This study aims to address two primary research questions

through hypothesis testing:

1. Do significant differences exist in the mean ratings of the GEMS-9 emotions?

   H_0: The mean ratings for all GEMS-9 emotions are equal.

   Statistical Test: ANOVA.

2. Which variables (familiarity, liking, intraclass correlation coefficient (ICC), music period, and music type) are significant predictors of GEMS-9 emotion ratings?

   H_0: The independent variables do not significantly predict the GEMS-9 emotion ratings.

   Statistical Test: Regression analysis.

To provide a more comprehensive analysis of nuanced GEMS-9 emotion ratings, this study focuses on a single genre: classical music. This genre was chosen because most studies on music and emotion focus on it[13–15], and the music periods required for the regression analyses are clearly defined. Building on the EMMA database's classical music dataset, this study conducted exploratory data analysis on the original continuous variables. Additionally, the influence of newly introduced variables (music period and music type) was examined via confirmatory regression analysis to address the study's second research question.

The findings of this study extend beyond the current limitations of music-evoked emotion research. By clarifying the predictive roles of musical features and individual listener characteristics, this work offers crucial insights for MIR and MER, enabling more accurate and nuanced artificial intelligence (AI) models. Furthermore, the identification of key predictors has direct implications for music therapy, where a deeper understanding of emotional responses to music can inform therapeutic interventions. Ultimately, this research contributes to the interdisciplinary fields of musicology, psychology, and neuroscience by advancing the theoretical understanding of how musical properties and personal factors interact to elicit emotional responses.

2 Methods

2.1 Materials

The data for this study were sourced from the EMMA database. The full database contains GEMS emotion ratings for 817 music excerpts across seven genres, with participants instructed to rate induced (felt) emotions. This study is based on a specific subset: the classical music dataset ($N = 105$), which was part of the initial set of 364 excerpts (classical, hip-hop, and pop) rated by 567 participants using the GEMS-45 emotion items. The data, downloaded as a CSV file, included details on music characteristics (title, composer, excerpt duration, and a link to the YouTube clip), rater characteristics (familiarity, liking, ICC, and number of raters), and emotion ratings (GEMS-9 dimensions and three higher-order dimensions).

The GEMS-9 ratings were calculated as a weighted average of the corresponding GEMS-45 item ratings (scaled from 0 to 100), while the higher-order dimension ratings were a simple average of the corresponding GEMS-9 ratings. Familiarity and liking were measured on a 5-point Likert scale. Familiarity ranged from 1.09 to 4.03 ($M = 1.88$, $SD = 0.64$), while liking had a narrower range from 1.85 to 3.97, with higher average ratings ($M = 3.13$, $SD = 0.40$). The ICC, which assesses inter-rater reliability, was determined using a two-way random effects average measure model[16], as detailed in the original study. The ICC ranged from -0.26 to 0.96 ($M = 0.79$, $SD = 0.19$), with the single negative value being a statistical artifact. The number of raters per music excerpt ranged from 11 to 101 ($M = 28.65$, $SD = 10.02$). Excerpt duration was excluded from the analysis due to a high number of missing values ($n = 36$, 34.3%). The original study reported an average classical music excerpt duration of 47 seconds ($SD = 10$ seconds, $N = 105$).

Initial data from the EMMA database lacked the classical music period information required for this study. The composition year of each excerpt was retrieved from its EMMA webpage profile for period assignment. Each excerpt was then categorized into the five generally accepted classical music periods[17]: Renaissance, Baroque, Classical, Romantic, and Modern. To satisfy the minimum data points per level for ANOVA variance calculation, a single Renaissance excerpt was merged into the closest Baroque period. This resulted in a new independent variable, music period, with four categories: Baroque ($n = 12$, 11.4%), Classical ($n = 15$, 14.3%), Romantic ($n = 54$, 51.4%), and Modern ($n = 24$, 22.9%).

The original data from the EMMA database, filtered by genre and instrumentation, comprised 105 classical music excerpts (87 instrumental and 18 vocal pieces). To validate this data, all 105 excerpts were investigated from song profiles and YouTube clips. The results revealed several errors in categorization, which were corrected to create a new independent variable, music type, with two categorical levels: instrumental ($n = 88$, 83.8%) and vocal ($n = 17$, 16.2%).

2.2 Procedure

The final reorganized dataset identified the variables used in this study. The dependent variables were the GEMS-9 emotion ratings for each of the nine dimensions (wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness; $N = 105$). The independent variables included three continuous variables (familiarity, liking, and ICC), one discrete variable (number of raters), and two categorical variables (music period and music type). The music period variable had four levels (Baroque, Classical, Romantic, and Modern), while the music type variable had two levels (instrumental and

vocal).

This study consisted of two main sections: exploratory data analysis and confirmatory data analysis. The first section examined data characteristics using descriptive statistics, histograms, normality tests, intercorrelations, and contour plots. The second section directly addressed the research questions using ANOVA and regression analyses.

The first confirmatory analysis formally assessed the significance of mean differences among the GEMS-9 emotion ratings using repeated measures ANOVA and post hoc pairwise comparisons. A repeated measures ANOVA was chosen because the GEMS-9 emotion ratings, measured on the same music excerpts, were inherently dependent. After the ANOVA, post hoc pairwise comparisons were performed using a paired-sample t-test with Bonferroni correction to explore specific pairs of emotion ratings with significant mean differences.

The second confirmatory analysis evaluated the statistical significance of all independent variables in predicting the GEMS-9 emotion ratings using multiple linear regression. Dummy coding was used for the four music periods (with Baroque as the reference level) and two music types (with instrumental as the reference level), resulting in eight predictors for each regression model. Prior to creating the models, all nine dependent variables, three continuous independent variables, and one discrete independent variable were standardized, while the four categorical dummy variables remained unstandardized as generally recommended.

The assumptions for repeated measures ANOVA (independence of observations, sphericity, and normality of residuals) and regression (normality, homoscedasticity, no autocorrelation of residuals, and no multicollinearity of independent variables) were tested before each analysis to ensure the validity and reliability of the results. Any violations were handled with appropriate corrective measures, such as the Greenhouse-Geisser correction for sphericity violation.

All statistical analyses and data visualizations were performed using Python (version 3.12.7) with the following libraries: pandas[18] (version 2.2.2), scipy[19] (version 1.13.1), statsmodels[20] (version 0.14.2), pingouin[21] (version 0.5.5), matplotlib[22] (version 3.9.2), and seaborn[23] (version 0.13.2). Specific Python functions are shown in Table 1.

3 Results

3.1 Exploratory Data Analysis

This section summarizes the data characteristics of the nine dependent variables (GEMS-9 emotion ratings) and three continuous independent variables (familiarity, liking, and ICC). Data distributions were examined using descriptive statistics, histograms, and normality tests. The relationships among these continuous variables were then investigated using intercorrelations and contour plots.

3.1.1 Descriptive Statistics

Table 2 presents the descriptive statistics of the GEMS-9 emotion ratings. The mean values of GEMS-9 emotion ratings ranged from 5.23 to 29.04 on a scale from 0 to 100. The GEMS-9 dimension peacefulness exhibited the highest mean ($M = 29.04$, $SD = 18.32$), followed by joyful activation ($M = 25.34$, $SD = 14.29$) and wonder ($M = 25.11$, $SD = 7.01$). The mean of peacefulness was notably higher compared to tenderness ($M = 17.84$, $SD = 12.66$, 95% CI [15.42, 20.26]) and nostalgia ($M = 21.73$, $SD = 13.29$, 95% CI [19.19, 24.27]), as indicated by its non-overlapping 95% confidence interval (CI [25.54, 32.55]). Conversely, sadness showed the lowest mean ($M = 5.23$, $SD = 6.50$), distinctly lower than the other emotion ratings. This observation aligns with the GEMS's acknowledged underrepresentation of negative emotions[10].

Skewness and kurtosis values were examined to assess the asymmetry and peakedness of the data distributions. For skewness, most values were between -1 and 1, indicating near-normal distributions. The primary exceptions were tension (1.33) and sadness (1.96). Notably, all values were positively skewed except for wonder (-0.23), which was the closest to a normal distribution. For kurtosis, all values were within the -1 to 1 range, with the exceptions of tension (1.08) and sadness (4.04). The high kurtosis of sadness suggests its distribution has more extreme outliers, indicating heavier tails and a sharper peak than a normal distribution.

3.1.2 Histograms

Figure 1 shows histograms visualizing data distributions of the GEMS-9 emotion ratings. Curve lines represent smoothed distributions based on the Kernel Density Estimate (KDE)[23]. The data distribution for wonder was near-normal, as its skewness value was closest to 0. Among the other emotions, peacefulness exhibited the widest range (Min = 0.00, Max = 70.75), followed by power (Min = 0.00, Max = 68.45) and nostalgia (Min = 0.00, Max = 60.58). In contrast, sadness displayed the narrowest range (Min = 0.00, Max = 28.61). Notably, its distribution was heavily concentrated at the lower end of the emotion rating scale: over 50% of excerpts ($n = 59$, 56.2%) received sadness ratings below 4, with 26 excerpts (24.8%) rated at 0.

3.1.3 Normality Test

Formal normality tests were conducted to statistically evaluate the distribution of all variables and confirm visual observations. The distributions of wonder (Shapiro-Wilk, $p = .50$; Jarque-Bera, $p = .61$) and transcendence (Shapiro-Wilk, $p = .17$; Jarque-Bera, $p = .42$) were confirmed as normal by both tests.

In contrast, tenderness, power, joyful activation, tension, and sadness were identified as non-normally distributed ($p < .05$) by both tests. For nostalgia and peacefulness, the tests yielded conflicting results: the Shapiro-Wilk test indicated non-normality

Table 1 Statistical Analyses and Corresponding Python Functions

Analysis	Python Function
Data loading	`pandas.read_csv()`
Descriptive statistics	`statsmodels.stats.descriptivestats.describe()`
Correlation	`scipy.stats.spearmanr()`
Regression	`statsmodels.formula.api.ols()`
ANOVA	`pingouin.rm_anova()`
Partial eta-squared (η_p^2)	`pingouin.rm_anova()`
Post hoc test:	
Paired-sample *t*-test	`pingouin.pairwise_tests()`
Normality:	
Shapiro-Wilk test	`scipy.stats.shapiro()`
Jarque-Bera test	`statsmodels.formula.api.ols()`
Omnibus test	`statsmodels.formula.api.ols()`
Homoscedasticity:	
Breusch-Pagan test	`statsmodels.stats.diagnostic.het_breuschpagan()`
White's test	`statsmodels.stats.diagnostic.het_white()`
Autocorrelation:	
Durbin-Watson test	`statsmodels.formula.api.ols()`
Multicollinearity:	
VIF	`statsmodels.stats.outliers_influence.variance_inflation_factor()`
Sphericity:	
Mauchly's test	`pingouin.rm_anova()`
Data visualization:	
Histogram	`seaborn.histplot()`
Interpolation	`scipy.interpolate.Rbf()`
Contour plot	`matplotlib.pyplot.contour()`
Scatter plot	`matplotlib.pyplot.scatter()`

Table 2 Descriptive Statistics of the GEMS-9 Emotion Ratings

GEMS-9 dimension	*M*	*SD*	95% CI	Min	Max	Skewness	Kurtosis
Wonder	25.11	7.01	[23.77, 26.45]	7.18	45.20	−0.23	0.14
Transcendence	23.99	7.48	[22.56, 25.42]	10.64	46.58	0.27	−0.33
Tenderness	17.84	12.66	[15.42, 20.26]	0.00	47.70	0.56	−0.61
Nostalgia	21.73	13.29	[19.19, 24.27]	0.00	60.58	0.45	−0.53
Peacefulness	29.04	18.32	[25.54, 32.55]	0.00	70.75	0.37	−0.90
Power	19.40	15.87	[16.36, 22.43]	0.00	68.45	0.84	0.02
Joyful activation	25.34	14.29	[22.61, 28.08]	2.00	62.36	0.71	−0.16
Tension	14.74	13.05	[12.25, 17.24]	0.00	55.06	1.33	1.08
Sadness	5.23	6.50	[3.99, 6.48]	0.00	28.61	1.96	4.04

Note. $N = 105$. Mean, standard deviation, 95% confidence interval (CI), minimum, maximum, skewness, and kurtosis of the GEMS-9 emotion ratings. The Fisher's kurtosis values are shown (a normal distribution has a kurtosis of 0).

($p = .005$ and $p = .002$, respectively), while the Jarque-Bera test suggested normality ($p = .09$ and $p = .052$, respectively). The Shapiro-Wilk test's higher power to detect non-normality in small to moderate sample sizes likely accounts for this discrepancy[24].

Of the three continuous independent variables, only liking was found to be normally distributed (Shapiro-Wilk, $p = .46$; Jarque-Bera, $p = .39$).

3.1.4 Intercorrelation

Table 3 shows the intercorrelation matrix and reveals numerous significant relationships among the variables. Of the 66 total pairs, 39 (59.1%) had statistically significant correlations.

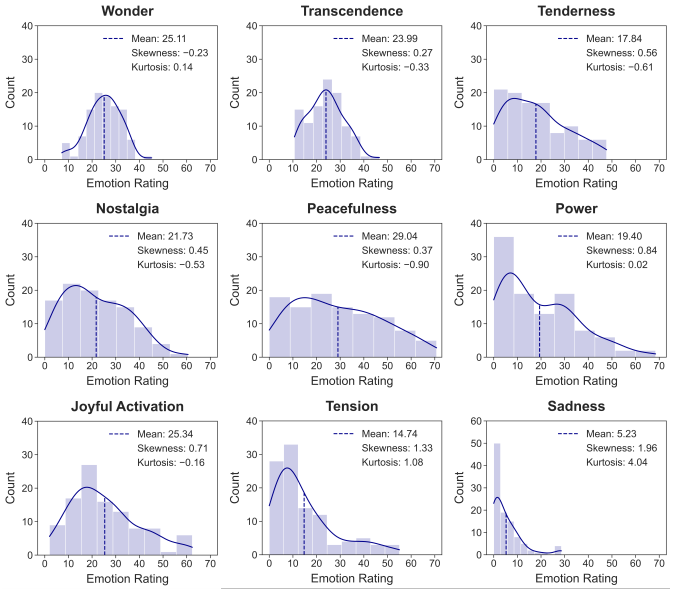

Fig. 1 Distributions of the GEMS-9 Emotion Ratings

A clear pattern of strong relationships emerged among the GEMS-9 emotion ratings. The strongest positive intercorrelations were found between emotions that share similar qualities, such as tenderness and peacefulness ($r_s = .87$, $p < .001$), tenderness and nostalgia ($r_s = .81$, $p < .001$), and nostalgia and peacefulness ($r_s = .80$, $p < .001$). Conversely, strong negative correlations were consistently observed between emotions with opposing qualities, such as tension and peacefulness ($r_s = -.75$, $p < .001$).

The emotion sadness exhibited a distinctive correlation pattern, showing positive relationships with aesthetic emotions such as nostalgia ($r_s = .52$, $p < .001$), while being negatively correlated with energetic emotions such as joyful activation ($r_s = -.62$, $p < .001$).

The independent variables demonstrated notable relationships with the emotion ratings. Liking had strong positive correlations with all five emotions in the higher-order dimension of sublimity (wonder, transcendence, tenderness, nostalgia, and peacefulness), indicating a broad association with these aesthetic emotions. In contrast, familiarity was positively correlated with energetic emotions: power ($r_s = .40$, $p < .001$) and joyful activation ($r_s = .39$, $p < .001$). Liking and familiarity were also positively correlated ($r_s = .40$, $p < .001$).

Finally, the ICC had a positive correlation with liking ($r_s = .32$, $p = .001$) and a negative correlation with familiarity ($r_s = -.25$, $p = .009$). These correlations suggest that inter-rater consensus is more closely aligned with participants' positive aesthetic appraisals of the music than with their prior exposure to it. The ICC had no significant correlations with most of the GEMS-9 emotion ratings, with the exception of a negative correlation with sadness ($r_s = -.26$, $p = .006$).

3.1.5 Contour Plots

Figure 2 displays contour plots of the GEMS-9 emotion ratings as functions of familiarity and liking. Visual inspection of these plots revealed several clear patterns. Both power and joyful activation peaked where both familiarity and liking ratings were high, as indicated by spots in the upper-right corners. This pattern suggests that these emotions are most strongly felt when a piece of music is both familiar and liked.

In contrast, peacefulness showed its highest ratings in the upper-left corner, where familiarity was low but liking was high. This emotion also displayed multiple peaks in the high-liking regions, indicating less dependence on familiarity ratings than on liking. Tension exhibited higher ratings in the lower-left corner (low familiarity, low liking), an observation that aligns with its strong negative correlation with liking ($r_s = -.43$, $p < .001$). Finally, two aesthetic emotions, tenderness and nostalgia, displayed similar profiles, with their highest ratings in the high-liking regions, consistent with their positive correlations with liking ($r_s = .42$, $p < .001$ and $r_s = .49$, $p < .001$, respectively).

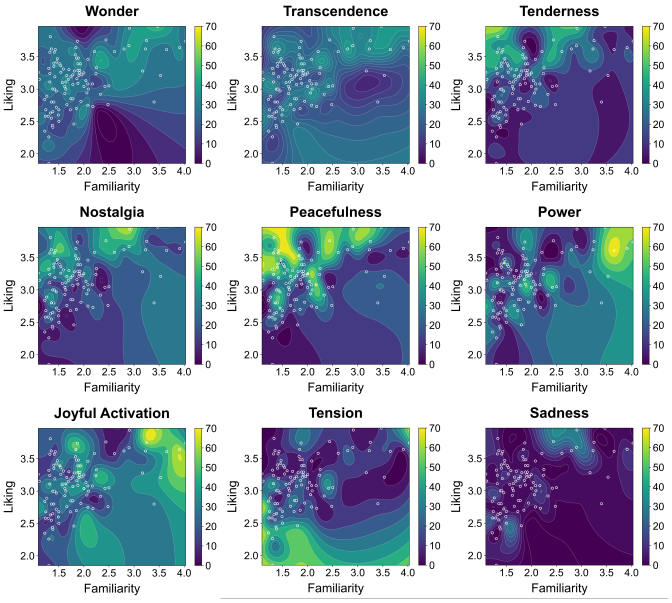

Fig. 2 Contour Plots of the GEMS-9 Emotion Ratings as Functions of Familiarity and Liking

3.2 Confirmatory Data Analysis

This section presents the results of the confirmatory data analysis, which directly addresses the study's research questions. The first part of the analysis examines the differences in mean emotion ratings, and the second part investigates the predictive relationships between the independent variables and the GEMS-9 emotion ratings.

3.2.1 ANOVA and Post Hoc Comparisons

Table 3 Intercorrelation Matrix for GEMS-9 Emotion Ratings, Familiarity, Liking, and ICC

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Wonder	—										
2. Transcendence	.37**	—									
3. Tenderness	.26**	.07	—								
4. Nostalgia	.30**	.09	.81**	—							
5. Peacefulness	.18	−.01	.87**	.80**	—						
6. Power	.05	.14	−.73**	−.67**	−.71**	—					
7. Joyful activation	.27**	.15	−.51**	−.44**	−.44**	.77**	—				
8. Tension	−.18	.06	−.68**	−.62**	−.75**	.45**	.20*	—			
9. Sadness	.05	.14	.50**	.52**	.39**	−.50**	−.62**	−.31**	—		
10. Familiarity	.19	.26**	−.07	−.01	−.06	.40**	.39**	−.14	.01	—	
11. Liking	.48**	.37**	.42**	.49**	.40**	.04	.22*	−.43**	.13	.40**	—
12. ICC	.11	.04	−.02	.06	.09	.04	.10	.06	−.26**	−.25**	.32**

Note. $N = 105$. Table entries are sample Spearman's rank correlation coefficients (r_s). Rows and columns represent the same variables. ICC = intraclass correlation coefficient. $^{*}p < .05$. $^{**}p < .01$.

A repeated measures ANOVA revealed a significant main effect of emotion dimensions on ratings ($F(2.07, 215.74) = 32.33$, $p < .001$, $\eta_p^2 = .24$), indicating at least one significant difference among the means of the nine emotion ratings. Prior to this analysis, the assumption of sphericity was violated ($\chi^2(35) = 736.81$, $p < .001$), which prompted the use of a Greenhouse-Geisser correction ($\varepsilon = .24$) to the degrees of freedom. While the normality of residuals was also violated (Shapiro-Wilk: $W = 0.97$, $p < .001$), the analysis is considered robust to this violation due to the large sample size, as per the Central Limit Theorem (CLT)[25,26].

Post hoc pairwise tests were conducted to further explore these differences. Table 4 shows the post hoc pairwise comparison results, indicating that 20 of 36 total pairs of emotion ratings (55.6%) exhibited significant mean differences. The results confirmed that the mean sadness emotion rating was significantly lower than the other eight emotion ratings. The Hedges' g values for pairs involving sadness ranged from 0.92 to 2.93, indicating substantial mean differences. The largest mean difference occurred between peacefulness and sadness (MD = 23.81, $t(104) = 14.01$, $p < .001$, $g = 1.73$).

To examine the nuanced music-evoked emotions within the higher-order dimension of sublimity, the mean differences among the five aesthetic emotion ratings were investigated. Of the 10 possible pairs, five showed statistically significant mean differences. Tenderness had a significantly lower mean than the other four emotion ratings: wonder (MD = −7.27, $t(104) = -5.70$, $p < .001$, $g = -0.71$), transcendence (MD = −6.15, $t(104) = -4.42$, $p < .001$, $g = -0.59$), nostalgia (MD = −3.89, $t(104) = -5.09$, $p < .001$, $g = -0.30$), and peacefulness (MD = −11.20, $t(104) = -11.45$, $p < .001$, $g = -0.71$). Notably, even similar aesthetic emotion pairs showed statistically significant mean differences. For example, the mean of nostalgia was significantly lower than peacefulness (MD = −7.31, $t(104) = -6.65$, $p < .001$, $g = -0.46$).

Similarly, a statistically significant mean difference was also observed between the energetic emotions within vitality. The mean of joyful activation was significantly higher than power (MD = 5.95, $t(104) = 5.40$, $p < .001$, $g = 0.39$). The results from the ANOVA and post hoc pairwise comparisons demonstrated that the GEMS-9 model effectively detected nuanced and statistically significant differences among emotions induced by classical music.

3.2.2 Regression Analysis

Multiple linear regression was conducted to evaluate the statistical significance of all independent variables in predicting the GEMS-9 emotion ratings. Prior to the analysis, the major assumptions of regression were tested. The normality of residuals was violated for the power, tension, and sadness models (Shapiro-Wilk, Jarque-Bera: $p < .01$), and homoscedasticity was violated for the tenderness model (Breusch-Pagan, White's: $p < .05$). All models met the assumptions of no autocorrelation and no multicollinearity, as indicated by Durbin-Watson statistics ranging from 1.66 to 2.38 and the variation inflation factor (VIF) values ranging from 1.24 to 2.82, respectively.

Table 5 presents the results of nine multiple linear regression analyses. The overall regression models were significant for eight emotion ratings ($F(8, 96) = 2.83$–8.00, $p < .01$), explaining 19.1–40.0% of the variance in the dependent variables ($R^2 = 0.19$–0.40). The model for sadness was not significant ($F(8, 96) = 1.21$, $p = .30$).

The categorical independent variables (music period and type) were not significant predictors in any of the regression models for the GEMS-9 emotion ratings ($p > .05$). A sensitivity analysis was conducted to assess the impact of data preparation on this finding. For this analysis, the single Renaissance excerpt, which had been merged into the Baroque period, was removed, reducing the sample size from $N = 105$ to $N = 104$. The re-estimated models consistently showed that music period and type remained non-significant predictors. The overall findings were robust, with the models for eight emotions remaining significant ($F(8, 95) = 2.64$–8.34, $p < .05$), explaining 18.2–41.3% of the variance ($R^2 = 0.18$–0.41), while the model for sadness remained non-significant ($F(8, 95) = 1.19$, $p = .32$).

The results revealed that liking was the only significant positive predictor for all five aesthetic emotions within sublimity.

Table 4 Post Hoc Pairwise Comparisons of the GEMS-9 Emotion Ratings

A	B	MD	t(104)	p	95% CI	Hedges' g
Wonder	Tenderness	7.27	5.70	$< .001$	[4.74, 9.79]	0.71
Wonder	Power	5.71	3.47	.03	[2.45, 8.97]	0.46
Wonder	Tension	10.36	6.64	$< .001$	[7.27, 13.46]	0.99
Wonder	Sadness	19.88	21.73	$< .001$	[18.06, 21.69]	2.93
Transcendence	Tenderness	6.15	4.42	$< .001$	[3.39, 8.90]	0.59
Transcendence	Tension	9.24	6.49	$< .001$	[6.42, 12.06]	0.87
Transcendence	Sadness	18.76	20.85	$< .001$	[16.97, 20.54]	2.67
Tenderness	Nostalgia	−3.89	−5.09	$< .001$	[−5.40, −2.37]	−0.30
Tenderness	Peacefulness	−11.20	−11.45	$< .001$	[−13.14, −9.26]	−0.71
Tenderness	Joyful Activation	−7.50	−3.33	.04	[−11.97, −3.03]	−0.55
Tenderness	Sadness	12.61	10.93	$< .001$	[10.32, 14.90]	1.25
Nostalgia	Peacefulness	−7.31	−6.65	$< .001$	[−9.49, −5.13]	−0.46
Nostalgia	Sadness	16.50	14.31	$< .001$	[14.21, 18.78]	1.57
Peacefulness	Tension	14.30	5.06	$< .001$	[8.70, 19.90]	0.90
Peacefulness	Sadness	23.81	14.01	$< .001$	[20.44, 27.18]	1.73
Power	Sadness	14.16	7.43	$< .001$	[10.39, 17.94]	1.16
Joyful Activation	Power	5.95	5.40	$< .001$	[3.76, 8.13]	0.39
Joyful Activation	Tension	10.60	5.91	$< .001$	[7.04, 14.16]	0.77
Joyful Activation	Sadness	20.11	11.17	$< .001$	[16.54, 23.68]	1.81
Tension	Sadness	9.51	6.19	$< .001$	[6.47, 12.56]	0.92

Note. Paired-sample t-test results with Bonferroni correction. Only 20 pairs with significant mean differences ($p < .05$) are shown, out of total 36 pairs compared across GEMS-9 emotion ratings. MD represents the mean difference (the mean of A minus the mean of B). The p-values are Bonferroni-corrected for pairwise comparisons.

Conversely, familiarity was the significant positive predictor for energetic emotions (power and joyful activation) within vitality. Additional regression analyses using only familiarity and liking as predictors confirmed that the results for power ($F(2,102) = 13.34$, $p < .001$, $R^2 = 0.21$) and joyful activation ($F(2,102) = 11.10$, $p < .001$, $R^2 = 0.18$) remained unchanged, indicating that familiarity was the only significant positive predictor for these emotions.

4 Discussion

This study successfully demonstrated that the GEMS-9 model is effective in detecting nuanced emotional responses to classical music, as evidenced by the statistically significant mean differences found in more than half of the emotion rating pairs. The analysis of predictive variables revealed distinct roles for liking and familiarity. Liking was a consistent positive predictor for all five aesthetic emotions of sublimity, highlighting the importance of aesthetic appreciation in these responses. In contrast, familiarity was the significant positive predictor for energetic emotions, power and joyful activation, within the vitality dimension, suggesting a strong link between prior exposure and feelings of energy. The categorical variables of music period and type, however, were not found to be significant predictors in any of the regression models. These findings offer crucial insights while also highlighting several limitations that warrant future research.

4.1 Limitations

While this study provides valuable insights, it is important to acknowledge its constraints, which are related to the GEMS model, the EMMA database, and the statistical analysis.

4.1.1 GEMS Model

Despite its utility, the GEMS emotion model has several acknowledged limitations. It has been criticized for an underrepresentation of negative emotions and a potential bias toward Western classical music. Furthermore, the abstract scales of wonder and transcendence have shown inconsistent ratings across studies, suggesting that these concepts may be difficult for listeners to interpret reliably [11,27]. From a practical standpoint, the full 45-item scale can limit its use in large-scale and time-limited studies, such as neuroimaging. Finally, the GEMS is susceptible to potential inaccuracies because its framework relies on participants' subjective self-reports of their induced emotions.

4.1.2 EMMA Database

While the EMMA database represents a significant advancement, it presents key limitations. Its reliance on the GEMS model indicates that it inherits the GEMS's limitations. The database also lacks comprehensive genre representation and, specifically, its classical music subset has structural flaws, in-

Table 5 Multiple Linear Regression Results for Predicting the GEMS-9 Emotion Ratings

	Model								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Model fit									
R^2	0.27	0.19	0.36	0.30	0.38	0.29	0.25	0.40	0.09
Adjusted R^2	0.21	0.12	0.30	0.24	0.33	0.23	0.19	0.35	0.02
$F(8, 96)$	4.51^{**}	2.83^{**}	6.59^{**}	5.20^{**}	7.34^{**}	4.81^{**}	4.01^{**}	8.00^{**}	1.21
Variable									
Familiarity	0.01	0.01	-0.45^{**}	-0.26^{*}	-0.48^{**}	0.60^{**}	0.44^{**}	0.21^{*}	−0.09
Liking	0.50^{**}	0.41^{**}	0.72^{**}	0.63^{**}	0.67^{**}	−0.22	−0.03	-0.67^{**}	0.27^{*}
ICC	0.09	−0.07	−0.07	−0.04	0.04	0.21^{*}	0.23^{*}	0.17	−0.20
Number of raters	−0.06	0.08	−0.14	−0.04	-0.25^{**}	0.10	−0.09	0.24^{**}	0.03
Period: Classical	0.29	0.29	−0.03	0.02	−0.07	−0.25	0.07	0.16	0.32
Period: Romantic	0.01	−0.04	−0.14	−0.05	−0.40	0.23	0.21	0.22	−0.17
Period: Modern	0.29	0.06	0.01	0.11	−0.26	−0.08	−0.13	0.33	0.19
Type: Vocal	0.06	0.48	0.28	−0.09	0.25	0.05	−0.29	−0.36	0.21

Note. $N = 105$. Column headings 1–9 represent the GEMS-9 emotions (1 = Wonder; 2 = Transcendence; 3 = Tenderness; 4 = Nostalgia; 5 = Peacefulness; 6 = Power; 7 = Joyful activation; 8 = Tension; 9 = Sadness). The upper portion of the table shows model fit statistics. The lower portion shows standardized regression coefficients (β) for the continuous and discrete independent variables (familiarity, liking, ICC, and number of raters) and unstandardized regression coefficients (b) for the categorical dummy variables. $^{*}p < .05$. $^{**}p < .01$.

cluding an unbalanced selection of music periods and types that may have weakened statistical findings. Moreover, the downloadable data file lacks crucial information on music period and type, and the categorization rules are unavailable and unvalidated. Finally, the absence of crucial musical features such as tempo, rhythm, and harmony restricts its application in fields such as MIR and MER. These combined factors highlight the need for more diverse and well-structured datasets in future research.

4.1.3 Statistical Analysis

The statistical methodology employed also warrants discussion. The multiple linear regression analysis relied on Ordinary Least Squares (OLS), a method sensitive to multicollinearity. Given the inherent correlation between the key predictors of familiarity and liking, the resulting standardized regression coefficients can be unstable and cannot be reliably used to infer which predictor has a stronger effect. While the regression models successfully accounted for variance in GEMS-9 ratings, it is crucial to recognize that the direct comparison of beta coefficients between these correlated variables is statistically unsound[28]. Therefore, this study focused on evaluating the model's overall predictive power and the significant role of each variable, rather than making causal claims about their relative strengths. The discrepancy in findings between this study and previous research on the predictors for vitality can be explained by differences in regression model composition and multicollinearity. Future research could explore alternative methods to better address this issue and more robustly evaluate the individual contributions of correlated predictors.

4.2 Implications and Future Directions

This study provides several crucial insights with significant theoretical, methodological, and practical implications.

4.2.1 Theoretical Implications

This study's findings demonstrate that liking and familiarity are distinct psychological factors contributing to different emotional responses, a conclusion supported by neuroscientific evidence[3,5,29–31]. Specifically, the strong predictive role of liking for sublimity emotions suggests that aesthetic appreciation and personal preference are paramount for eliciting emotions of beauty and awe. In contrast, the predictive role of familiarity for vitality emotions indicates that prior exposure and memory retrieval are more closely tied to feelings of energy and physiological arousal. This work suggests that to capture a more nuanced understanding of music-evoked emotions, future models should incorporate distinct personal constructs, such as liking and familiarity, and consider their complex interactions, such as the mere exposure effect[32,33], rather than relying on a single personal factor.

4.2.2 Methodological Implications

The results highlight important considerations for future research design. Future studies investigating the roles of highly correlated variables should employ more robust methods, such as relative importance analysis[34,35] or ridge regression[36,37], rather than relying solely on OLS, to provide a more accurate evaluation of each predictor's unique contribution. Additionally, to overcome the limitations of the EMMA database, future research should aim for more diverse and balanced datasets, including a wider range of musical genres and more representative distributions of musical periods and types to reveal potential effects that were not detected in this study. Finally, integrating objective musical features into these analyses is essential for a more comprehensive understanding of how specific acoustic properties interact with listener characteristics to elicit emotional responses.

4.2.3 Practical Implications

These findings have direct applications in music-related fields. In MIR and MER applications, AI models could be refined to incorporate user-specific data on both liking and familiarity,

leading to more personalized and accurate music recommendations. For example, AI models could use a listener's liking scores to generate playlists for reflective or aesthetic experiences, while using familiarity data to create playlists designed to maximize energy and vitality. In music therapy, therapists could use a client's aesthetic preferences to facilitate deeper emotional reflection and well-being, and their pre-existing familiarity to elicit specific physiological responses. Ultimately, this research provides a framework for future studies to move beyond correlational analysis and toward a more predictive understanding of music-evoked emotion, bridging the gap between theoretical models and real-world applications.

5 Conclusion

This study provides strong empirical evidence that the GEMS-9 model is a robust tool for mapping subtle emotional responses to classical music. A key finding lies in distinguishing the distinct roles of liking and familiarity. This work showed that these factors are not interchangeable but instead serve as unique predictors for different emotional dimensions: liking predicts the aesthetic emotions of sublimity, and familiarity predicts the energetic emotions of vitality. This discovery represents a significant advance in the field, challenging previous models that conflated these concepts. The implications are significant both theoretically and practically. Future research should build on this by incorporating these separate constructs into more nuanced emotion models. Real-world applications in music recommendation and therapy can now be refined to create more personalized experiences based on a listener's aesthetic preferences or prior exposure. Ultimately, this work offers a critical step toward a more predictive and nuanced understanding of music-evoked emotion, bridging the gap between theoretical models and practical applications.

References

1 P. Juslin and D. Vstfjll, *Emotional responses to music: The need to consider underlying mechanisms*.

2 P. Juslin, *From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions*.

3 A. Blood and R. Zatorre, *Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion*.

4 S. Koelsch, T. Fritz, D. Cramon, K. Mller and A. Friederici, *Investigating emotion with music: An fMRI study*.

5 V. Salimpoor, M. Benovoy, K. Larcher, A. Dagher and R. Zatorre, *Anatomically distinct dopamine release during anticipation and experience of peak emotion to music*.

6 J. Russell, *A circumplex model of affect*.

7 P. Ekman, *An argument for basic emotions*.

8 K. Scherer and M. Zentner, *Music evoked emotions are differentmore often aesthetic than utilitarian*.

9 P. Juslin, G. Barradas, M. Ovsiannikow, J. Limmo and W. Thompson, *Prevalence of emotions, mechanisms, and motives in music listening: A comparison of individualist and collectivist cultures*.

10 M. Zentner, D. Grandjean and K. Scherer, *Emotions evoked by the sound of music: Characterization, classification, and measurement*.

11 A. Aljanaki, F. Wiering and R. Veltkamp, *Studying emotion induced by music through a crowdsourcing game*.

12 H. Strauss, J. Vigl, P. Jacobsen, M. Bayer, F. Talamini, W. Vigl, E. Zangerle and M. Zentner, *The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts*.

13 D. Vstfjll, *Emotion induction through music: A review of the musical mood induction procedure*.

14 T. Eerola and J. Vuoskoski, *A review of music and emotion studies: Approaches, emotion models, and stimuli*.

15 L. Warrenburg, *Choosing the right tune: A review of music stimuli used in emotion research*.

16 K. McGraw and S. Wong, *Forming inferences about some intraclass correlation coefficients*.

17 J. Burkholder, D. Grout and C. Palisca, *A History of Western Music*.

18 W. McKinney, *Data structures for statistical computing in Python*.

19 P. Virtanen, R. Gommers, T. Oliphant, M. Haberland, T. Reddy, D. Cournapeau and P. Mulbregt, *SciPy 1.0: Fundamental algorithms for scientific computing in Python*.

20 S. Seabold and J. Perktold, *Statsmodels: Econometric and statistical modeling with Python*.

21 R. Vallat, *Pingouin: Statistics in Python*.

22 J. Hunter, *Matplotlib: A 2D graphics environment*.

23 M. Waskom, *Seaborn: Statistical data visualization*.

24 T. Thadewald and H. Bning, *JarqueBera test and its competitors for testing normality: A power comparison*.

25 M. Blanca, R. Alarcn, J. Arnau, R. Bono and R. Bendayan, *Non-normal data: Is ANOVA still a valid option?*

26 U. Knief and W. Forstmeier, *Violating the normality assumption may be the lesser of two evils*.

27 J. Vuoskoski and T. Eerola, *Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences*.

28 W. Greene, *Econometric Analysis*.

29 W. Trost, T. Ethofer, M. Zentner and P. Vuilleumier, *Mapping aesthetic musical emotions in the brain*.

30 C. Pereira, J. Teixeira, P. Figueiredo, J. Xavier, S. Castro and E. Brattico, *Music and emotions in the brain: Familiarity matters*.

31 J. Plailly, B. Tillmann and J. Royet, *The feeling of familiarity of music and odors: The same neural signature?*

32 R. Zajonc, *Attitudinal effects of mere exposure*.

33 R. Zajonc, *Mere exposure: A gateway to the subliminal*.

34 D. Budescu, *Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression*.

35 J. Johnson, *A heuristic method for estimating the relative weight of predictor variables in multiple regression*.

36 A. Hoerl and R. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*.

37 A. Hoerl and R. Kennard, *Ridge regression: Applications to nonorthogonal problems*.