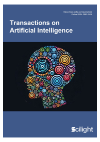
Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Article

EMelodyGen: Emotion-Conditioned Melody Generation in ABC Notation with Musical Feature Templates

Monan Zhou [1], Xiaobing Li [1], Feng Yu [1] and Wei Li [2,3,*]

[1] Department of Music AI and Information Technology, Central Conservatory of Music, Beijing 100031, China
[2] School of Computer Science and Technology, Fudan University, Shanghai 200433, China
[3] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China
* Correspondence: weili-fudan@fudan.edu.cn

How To Cite: Zhou, M.; Li, X.; Yu, F.; et al. EMelodyGen: Emotion-Conditioned Melody Generation in ABC Notation with Musical Feature Templates. *Transactions on Artificial Intelligence* **2025**, *1*(1), 199–211. https://doi.org/10.53941/tai.2025.100013

Received: 11 August 2025
Revised: 11 September 2025
Accepted: 11 Setptmber 2025
Published: 16 September 2025

Abstract: The EMelodyGen system mainly focuses on melody generation in ABC notation controlled by emotional conditions. To overcome the scarcity of emotional labeled sheet music, we utilize statistical correlations derived from small-scale symbolic music datasets with emotion labels and music psychology conclusions to guide subsequent feature extraction, emotional control and automatic annotation. We then automatically annotate a large, well-structured sheet music collection with rough emotional labels, convert the annotated dataset into ABC notation format, and apply data augmentation to address label imbalance, resulting in the creation of a dataset named Rough4Q. We demonstrate that our system backbone pre-trained on Rough4Q can achieve up to 99% music21 parsing rate. Our emotional control parameters, categorized into directly modifiable, embedding, dual-stage, and guidance features, can be selected and assembled to design customized emotional control templates that can lead to a 91% alignment in emotional expression in blind listening tests. Ablation studies further validate the impact of these control conditions on emotional accuracy.

Keywords: music generation; melody; ABC notation; LLM

1. Introduction

Automatic music generation involves three hierarchical levels: score generation, performance generation, and audio generation [1]. Works in symbolic music generation with emotional condition such as [2–5] focus on the first two levels. To validate the effectiveness of the current level, variables from subsequent levels are typically controlled. For instance, piano score generation often employs default performance expressions for audio rendering verification. Many recent emotion-conditioned music generation approaches such as [6–11] use MIDI data for symbolic music generation, while our work attempts to generate melodies in sheet music controlled by emotional conditions. For the selection of sheet format, although XML and ABC notation are equivalent and can be converted to each other without loss, we still choose to use ABC notation for its higher musical information density compared to XML. This is because ABC notation contains exclusively musical information throughout, whereas XML is essentially a configuration file that includes numerous structural symbols. In the field of ABC notation music generation, Tomasz Michal Oliwa previously explored genetic algorithms for rock music composition [12]. More recently, approaches such as Tunesformer [13], abcMLM [14], and MelodyT5 [15] have utilized Transformer-based language models to generate music in ABC notation. Among these, MelodyT5 is a pre-trained model suitable for both generation and understanding tasks, abcMLM is also based on a Transformer encoder-decoder architecture. So we selected Tunesformer, which is a Transformer decoder-only model, as the backbone of our system due to its focusing on generation and relative lightweight nature. The effectiveness of this model in generating sheet music relies heavily on the quality of the training data. Models fine-tuned on well-structured sheet music data are more likely to produce error-free sheet music. In contrast, models trained by disorganized sheet music may generate erroneous sheet music, which cannot be properly rendered into audio.

However, well-structured sheet music data with emotional labels is scarce. Notable datasets such as EMOPIA [16] and VGMIDI [17] consist of MIDI data with emotional annotations rather than XML/ABC sheet music. Although these MIDI files can be converted to XML by tools such as music21 [18] or MuseScore, the resulting sheet music is often disorganized, which impacts the quality of the generated music. Our subsequent experiments in Section 4.1 found that models fine-tuned on these datasets had music21 parsing rates of only 28% and 75%, respectively, indicating inadequate quality for further experiments. To address this, we utilized prior knowledge from music psychology literature [19–22] and statistical correlations between significant music features and emotional labels to guide feature extraction and emotional control. We automatically annotated large-scale and well-structured sheet music collection, in XML/ABC format with rough emotional labels by programming for training and further research. Prior to processing, we conducted data sampling and reviewed these datasets to ensure that their sheet music is well-structured. Additionally, after converting the data into uniform ABC notation slices, we tested the model fine-tuned with this data and achieved an music21 parsing rate of up to 99% for the generated sheet music.

Although above operation ensures a certain degree of quality, the automatically and roughly annotated emotional labels still contain noise. Therefore, we treat these rough emotional labels merely as markers for recording specific musical feature embeddings, and utilize important musical features summarized from statistical prior knowledge as control conditions for emotion. The qualitative conclusions from music psychology [19–22] and the quantitative distributions derived from statistical tests serve as references for selecting these conditions. We developed a set of musical feature control parameters based on these insights, which serve as control inputs for emotion-conditioned generation. These features can be categorized as follows:

- Directly modifiable features: features that can be adjusted directly at the model output stage, such as octave, volume, etc.;
- Embedding features: features that are captured as embeddings and require deep learning for the model to interpret, such as pitch range, average pitch (avg pitch), pitch standard deviation (pitch SD), melodic ascending or descending (direction), etc.;
- Dual-stage features: features that are both embedded and require oversight at the output stage, such as key, mode, tempo, etc.;
- Guidance features: features that are only used for guiding the control of highly related or alternative features, for instance, the root mean square of rendered audio from output sheet music (RMS) guides the volume control.

Typically, the second category arises because these features are inherent in the ABC notation header information and are controllable at the output stage. The third category often involves features that are difficult to manipulate directly or whose deep meanings are embedded in melody or texture, necessitating reliance on deep learning. For the last category, features not only within the current category can guide the control of other features, but some features from other categories that are not currently utilized can also serve as guidance for controlling other features. We selected five representative features from the first three categories to construct our control parameter template, the rest are used for reference or guidance. Using this template for condition control, the emotional alignment of the generated music with human expectations reached 91% in blind listening tests. Additionally, ablation experiments were conducted to assess the contribution of these control conditions to overall emotional expression.

2. Methodology

In our research workflow, we first merge and process the emotion-labeled datasets EMOPIA and VGMIDI, which are not suitable for sheet music training but statistical correlation analysis. Based on these statistical results and relevant literature in music psychology [19–22], we derive preliminary conclusions that guide the feature extraction, feature selection, rough automatic annotation, and emotional control for the subsequent training of large-scale sheet music data. Finally, we fine-tune the backbone network using the large-scale augmented sheet music data with emotion-related feature embeddings. This process involves various methods, such as standardizing the emotion label systems of EMOPIA and VGMIDI, which entails the conversion between the valence/arousal (V/A) values [23] and Russell 4Q [24] emotion space. Additionally, statistical tools for correlation analysis and the design of the backbone structure with the loss function are also involved.

2.1. V/A Values to Russell 4Q

The V/A values is a theoretical model used to describe and analyze emotional states. This emotional model is based on two primary dimensions: valence and arousal. Valence reflects the degree of positivity or negativity of an emotion, while arousal reflects the level of calmness or intensity of the emotion.

Russell 4Q emotional model divides the two-dimensional space defined by V/A values into four quadrants [25] shown as Figure 1. Given that emotion labels in datasets are indexed starting from 0, the formula for converting V/A values to Russell 4Q labels is given by Equation (1).

$$\mathcal{Q}(v,a) = I_{v<0}I_{a\geq 0} + 2I_{v<0}I_{a<0} + 3I_{v\geq 0}I_{a<0} \tag{1}$$

where I is the indicator function, and v and a represent the valence and arousal values, respectively. In our analysis, the EMOPIA dataset uses the Russell 4Q label system, while VGMIDI employs V/A values. We applied Equation (1) to convert the VGMIDI label system to the Russell 4Q space, after which we merged it with EMOPIA. The purpose of this merging process was to enrich the dataset used for analysis, thereby making the analytical conclusions more generalizable.

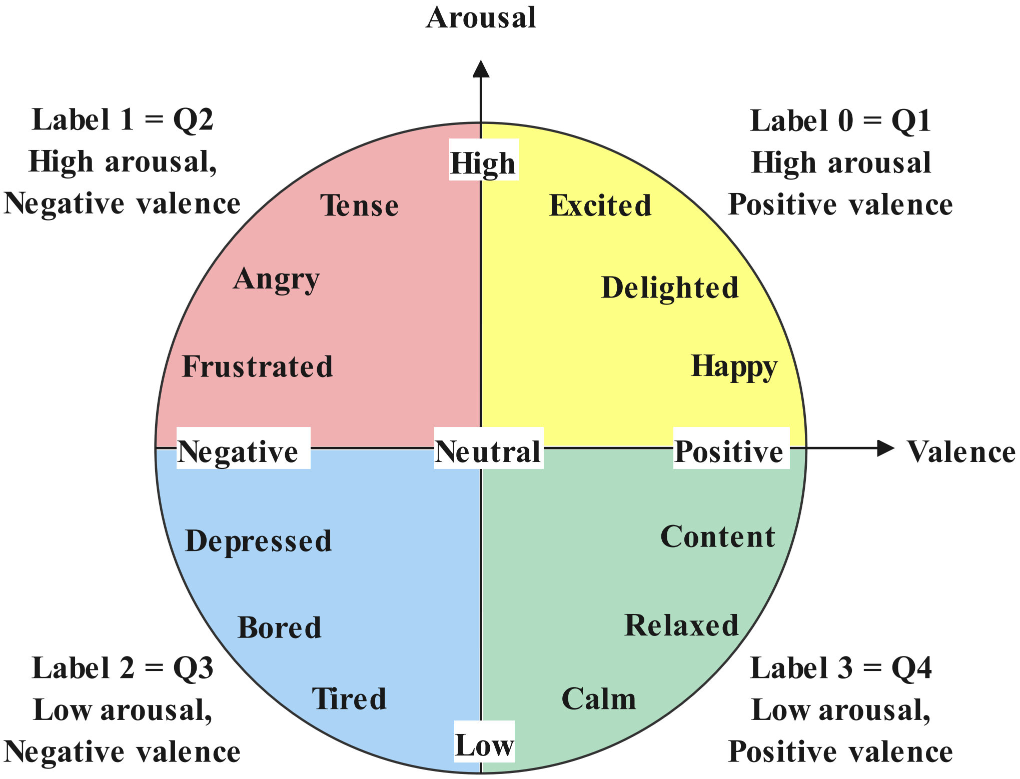


Figure 1. The projection of the emotional space based on V/A values onto Russell 4Q model with typical emotion adjectives.

2.2. Pearson Correlation Coefficient

The Pearson correlation coefficient [26] is a statistical measure used to assess the linear relationship between two variables, it quantifies the strength and direction of the linear relationship between two variables, with values ranging from -1 to +1. To compute the correlation coefficient r between two variables $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, the formula is given by Equation (2).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where x_i and y_i are the observed values of variables X and Y, and $\bar{x}$ and $\bar{y}$ are the means of X and Y, respectively. The correlation coefficient r indicates the direction of the relationship between the variables: a positive r signifies a positive correlation, while a negative r signifies a negative correlation. The absolute value $|r|$ represents the strength of the correlation, with values closer to 1 indicating a stronger correlation and values closer to 0 indicating a weaker correlation. A common threshold is 0.3, where $|r| > 0.3$ suggests a meaningful correlation, whereas $|r| \leq 0.3$ indicates a weak or negligible correlation. The correlation coefficient r is typically accompanied by a p-value, which is a statistical measure used in hypothesis testing to determine the significance of the observed result. The formula for calculating the p-value is given by Equation (3).

$$p = 2 \cdot \left(1 - \mathcal{T}\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, n-2\right)\right) \tag{3}$$

where $\mathcal{T}$ represents the t-distribution. A smaller p-value indicates a higher statistical significance of the observed correlation. Typically, a threshold of 0.05 is used: $p < 0.05$ suggests that the linear relationship between the variables is likely not due to chance, whereas $p \geq 0.05$ suggests that the observed correlation may be the result of random variation.

We used the aforementioned tools to perform statistical analysis on the merged data from EMOPIA and VGMIDI, calculating the correlation between emotional dimensions and features of sheet music. For the emotional dimensions, we chose valence and arousal. Due to the merging process, the Russell 4Q label system has lost the specific V/A values and only retains their positive or negative signs. Therefore, we categorized both valence and arousal into two levels: low and high, with values of 0 and 1, respectively. Specifically, data in the Q1 and Q4 quadrants were assigned a valence value of 0, while rests were assigned a value of 1. For arousal, data in the Q1 and Q2 quadrants were assigned a value of 0, while rests were assigned a value of 1.

Regarding the selection of features, we extracted eight features: key, mode, tempo, direction, avg pitch, pitch range, pitch SD, and RMS from the sheet music or its rendered audio. The specific meanings of these features are detailed in Section 1. However, the calculation of avg pitch, pitch SD and direction requires further explanation. Assuming a melody $M = \{(p_1, d_1), (p_2, d_2), \ldots, (p_n, d_n)\}$ consists of n notes, where p_i and d_i represent the pitch and duration of the i-th note respectively, the avg pitch (denoted as $\bar{p}$) is the weighted average of pitch by duration [27]. The formula for calculating the avg pitch is given by Equation (4).

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i d_i}{\sum_{j=1}^{n} d_j} \tag{4}$$

Based on $\bar{p}$, we further calculated the pitch SD using the formula given in Equation (5), which represents the weighted standard deviation of pitch by duration.

$$pitchSD = \sqrt{\frac{\sum_{i=1}^{n} (p_i - \bar{p})^2 d_i}{\sum_{j=1}^{n} d_j}} \tag{5}$$

For the feature direction, since it describes a musical characteristic at the level of phrases rather than entire pieces, we approach it by analyzing the duration of ascending and descending segments statistically. By comparing these durations, we determine the overall tonal direction of the piece. Specifically, if the total duration of ascending segments is greater, the piece is labeled as having an ascending tonal direction. Otherwise, the piece is labeled as having a descending tonal direction.

We calculated the Pearson correlation coefficients and p-values between V/A and the eight features. The results are presented in Table 1. These findings will guide the subsequent data processing and experimental design for emotional control. To provide a more intuitive representation of the distribution of emotions across the features, we also plotted distribution charts for the 4Q emotion categories across the eight features. For the features mode and direction, which both have only two options, we used bar charts. For the remaining six features, which are either continuous variables or discrete values with multiple options, we employed Gaussian kernel density estimation (KDE) plots [28], as illustrated in Figure 2.

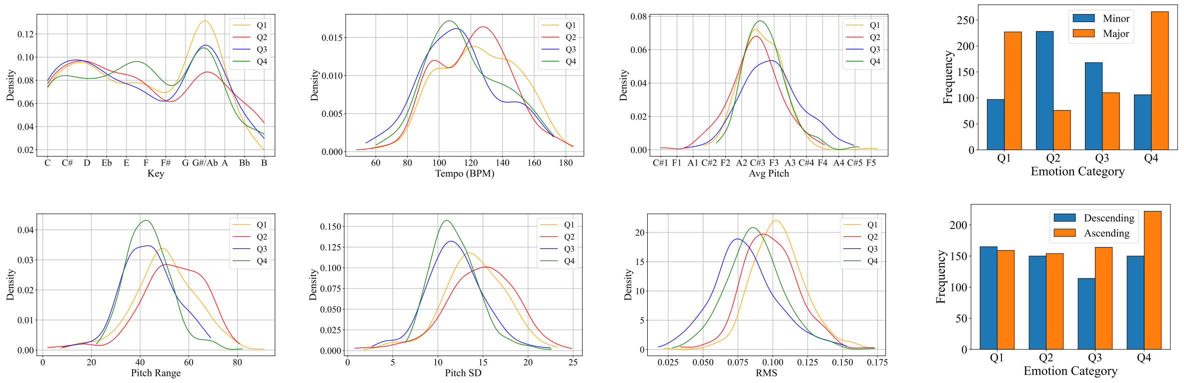


Figure 2. Gaussian KDE charts for Russell 4Q emotions over the six music-related features: key, tempo, average pitch, pitch range, pitchSD, and RMS, respectively (the six subplots on the left side); bar charts of Russell 4Q emotion frequency over modes and directions (the two subplots on the right side).

Table 1. Pearson correlation statistics between emotions and features for the merged data from EMOPIA and VGMIDI.

Emotion Dimension	Music-Related Feature	Correlation Coefficient	Relevance	*p*-Value	Confidence Level
valence	key	+0.0123	weak positive	6.594×10^{-1}	$p \geq 0.05$ insignificant
	mode	+0.3850	positive	2.018×10^{-46}	$p < 0.05$ significant
	tempo	+0.0621	weak positive	2.645×10^{-2}	$p < 0.05$ significant
	direction	+0.0010	weak positive	9.709×10^{-1}	$p \geq 0.05$ insignificant
	avg pitch	+0.0102	weak positive	7.161×10^{-1}	$p \geq 0.05$ insignificant
	pitch range	−0.0771	weak negative	5.794×10^{-3}	$p < 0.05$ significant
	pitch SD	−0.0676	weak negative	1.568×10^{-2}	$p < 0.05$ significant
	RMS	+0.1174	weak positive	2.597×10^{-5}	$p < 0.05$ significant
arousal	key	−0.0007	weak negative	9.809×10^{-1}	$p \geq 0.05$ insignificant
	mode	−0.0962	weak negative	5.748×10^{-4}	$p < 0.05$ significant
	tempo	+0.1579	weak positive	1.382×10^{-8}	$p < 0.05$ significant
	direction	−0.0958	weak negative	6.013×10^{-4}	$p < 0.05$ significant
	avg pitch	−0.1818	weak negative	5.919×10^{-11}	$p < 0.05$ significant
	pitch range	+0.3276	positive	2.324×10^{-33}	$p < 0.05$ significant
	pitch SD	+0.3523	positive	1.179×10^{-38}	$p < 0.05$ significant
	RMS	+0.3800	positive	3.558×10^{-45}	$p < 0.05$ significant

2.3. Backbone Network

We selected pre-trained Tunesformer as the backbone network, which is designed specifically for generating ABC notation music. The input data format is divided into two parts: control code and ABC chars. The latter represents the conventional ABC notation music, with its data structure detailed in the document ABC Music Notation. The former consists of three markers: S (number of sections), B (number of bars), and E (edit distance similarity), with specific meanings provided in [13].

The backbone network primarily consists of two sets of Transformer decoder structures, with a notable feature being the incorporation of music bar-level patch structures. This design significantly enhances the overall data throughput of the network. Consequently, the data input to the backbone is first converted into patch-level data through a patchilizer, then processed through nine layers of patch-level decoders. At the beginning of each patch-level decoder, the data undergoes a linear projection layer followed by the addition of positional embeddings. Finally, the data is processed through three layers of char-level decoders before reaching the output.

For the pre-training process, we consider a dataset D consisting of pairs (X, Y), where X is the input musical score and Y is the target musical score. Each score is represented as a sequence of bar patches $\{B_1, B_2, \ldots, B_m\}$, with each bar patch B_i further decomposed into a sequence of characters $\{c_1, c_2, \ldots, c_n\}$. The backbone is trained to predict each character token of the target score based on the input score and the previously generated tokens in an autoregressive manner. Formally, the pre-training objective is to minimize the cross-entropy loss across all tokens in the target sequence, whose loss function is shown in Equation (6).

$$\mathcal{L}_{CE}(\theta) = - \sum_{(X,Y) \in D} \sum_{i=1}^{m} \sum_{j=1}^{n} log P_\theta(c_j^i | X, B_{<i}, c_{<j}^i) \tag{6}$$

where c_j^i denotes the j-th character in the i-th bar patch of score Y, $B_{<i}$ includes all bar patches before the i-th bar patch, $c_{<j}^i$ refers to characters before the j-th character in the current patch, and P_θ represents the probability distribution function of the backbone, parameterized by θ, of predicting the correct character.

Building upon this backbone, we introduced emotion-related features embedding at the input stage for fine-tuning and added emotion-related features control at the output stage to achieve emotion-conditioned generation. The loss function used during fine-tuning is the same as that used during pre-training; however, the meanings of X and Y are specific to this stage. During fine-tuning, X represents a prompt derived from the fusion of emotion-related features embedding and control code, while Y corresponds to the subsequent ABC notation music following this prompt. The overall system architecture, integrating the backbone structure with the emotion control module, is illustrated in Figure 3, where the reservations are a collective term for the placeholders allocated for potential future additions of all categories of features. The specific sound library of the renderer in our system is the default piano soundfont of MuseScore 4.2.1 (Acoustic Grand Piano) with a 44.1 kHz sample rate.

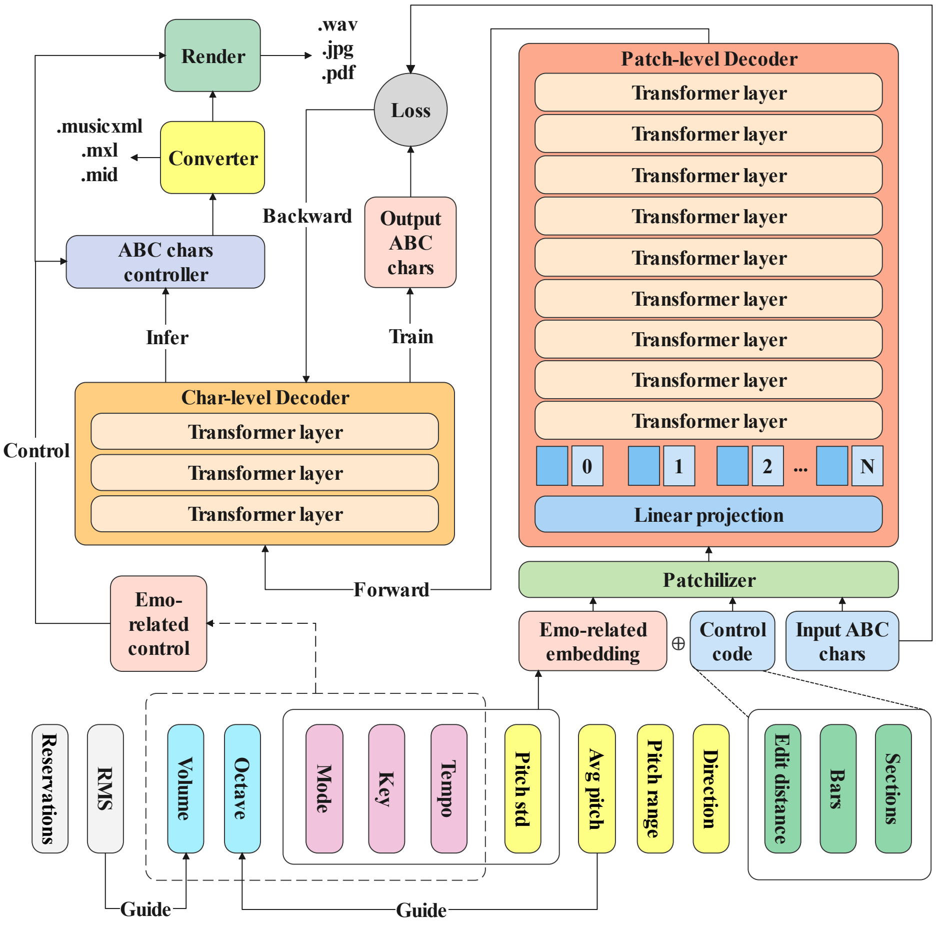


Figure 3. The overall system architecture with training and inference branches of the backbone, whose below part outlines the features currently in use (inside bounding boxes) and those planned for future implementation (outside bounding boxes).

3. Datasets

Based on the previous sections, we understand that our research process involves performing statistical analysis first and then fine-tuning the backbone network. However, the datasets and their processing methods for these two stages are distinct. For the statistical analysis, we merged the emotion-labeled datasets EMOPIA and VGMIDI to perform feature extraction and determine the correlation between emotions and features. This resulting dataset is referred to as the analysis dataset.

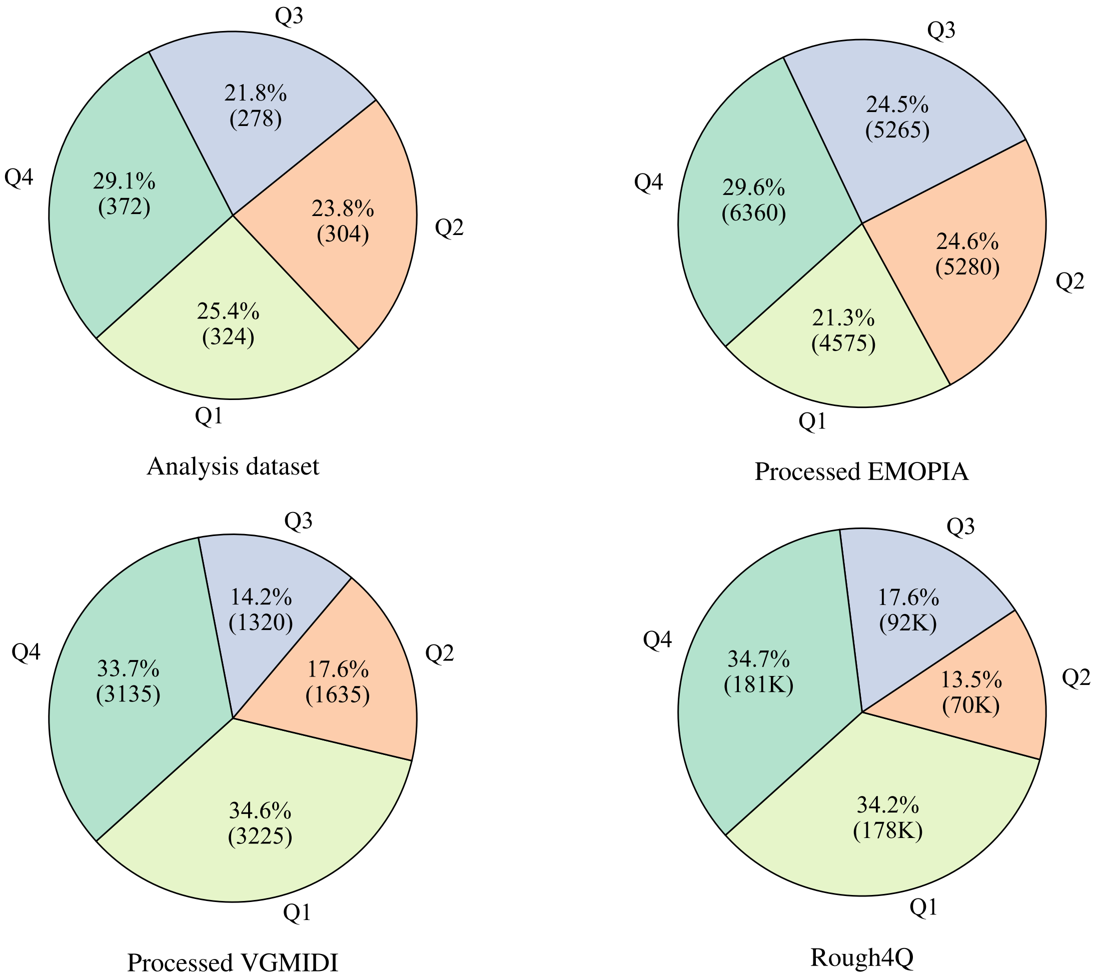
For the fine-tuning phase, we initially converted EMOPIA and VGMIDI to ABC notation format to create two datasets that are consistent with the data format used during the backbone pre-training stage. These datasets are used to explore the music21 parsing rate of the generated musical scores after fine-tuning the backbone model with them. Subsequently, we combined several well-structured music score datasets, and after feature extraction and the addition of rough emotion labels, we created a dataset called Rough4Q. These datasets and their corresponding data processing scripts have all been made available on HuggingFace. This section will provide a detailed description of these datasets.

3.1. Analysis Dataset

The analysis dataset is derived from merging EMOPIA and VGMIDI, with VGMIDI transformed into the Russell 4Q label system as detailed in Section 2.1. This dataset consists of 11 columns: the first three columns are emotion label columns, specifically label (Russell 4Q emotions), valence (low = 0 or high = 1), and arousal (low = 0 or high = 1); the remaining eight columns represent features, which are key (one of 12 keys: *C, C#, D, Eb, E, F, F#, G, G#/Ab, A, Bb, B*), mode (minor = 0 or major = 1), direction (descending = 0 or ascending = 1), avg pitch, pitch range, pitch SD, tempo, and RMS.

For feature extraction, the key, mode, direction, avg pitch, pitch range, and pitch SD features were directly computed using the music21 toolkit at the symbolic level. However, for tempo, since it often defaults to 120 BPM when extracted from MIDI files, which does not reflect the actual tempo, we used MuseScore 4.2.1 to render these MIDI files into WAV format with a 44.1 kHz sample rate by its default piano soundfont (Acoustic Grand Piano). Subsequently, we used the librosa [29] library to estimate the tempo and obtain more accurate and distinguishable data. The librosa library also calculated the RMS of the rendered audios.

It is noted that the extraction of the latter two features (tempo and RMS) from rendered audio is less efficient compared to the first six features. However, since the combined dataset comprises only 1278 pieces of music, the rendering time for these features is acceptable for the analysis phase. Consequently, we have constructed the analysis dataset, with its distribution according to the Russell 4Q classification shown in the first pie chart of Figure 4. For the statistical correlation analysis, we computed the correlations between the second and third columns (valence and arousal) and the remaining eight columns (features).



Analysis dataset

Processed EMOPIA

Processed VGMIDI

Rough4Q

Figure 4. Pie charts of proportions of different emotions in processed datasets.

3.2. Processed EMOPIA and VGMIDI

The processed EMOPIA and processed VGMIDI datasets will be used to evaluate the music21 parsing rate of music scores generated by fine-tuning the backbone with existing emotion-labeled datasets. Therefore, it is essential to ensure that the processed data is compatible with the input format required by the pre-trained backbone.

We found that the average number of measures in the dataset used for pre-training backbone is approximately 20, and the maximum number of measures supported by the pre-trained backbone input is 32. Consequently, we converted the original EMOPIA and VGMIDI data into XML scores filtering out erroneous items and segmented them into chunks of 20 measures each. Each chunk was appended with an ending marker to prevent the model from generating endlessly in cases of repetitive melodies without seeing a terminating mark. For the ending segments of the scores, if a segment exceeded 10 measures, it was further divided; otherwise, it was combined with the previous segment. This approach ensures that the resulting score slices do not exceed 30 measures, thereby guaranteeing that all slices are within the maximum measure limit supported by backbone, with an average of approximately 20 measures.

It is noted that when converting MIDI to XML using current tools, repeat sections cannot be folded back. In fact, after converting the dataset used for pre-training backbone into MIDI and expanding all repeat sections, the average number of measures was approximately 35. However, due to the maximum measure limit supported during pre-training, repeat markers were not expanded at that stage, and since repeat markers themselves occupy only two characters, we could not use 35 measures as the slicing unit even for MIDI data.

Subsequently, we converted the segmented XML slices into ABC notation format, performed data augmentation by transposing to 15 keys, and extracted the melodic lines and control codes to produce the final processed EMOPIA and processed VGMIDI datasets. Both datasets have a consistent structure comprising three columns: the first column is the control code, the second column is ABC chars, and the third column contains the 4Q emotion labels inherited from the original dataset. The total number of samples is 21,480 for processed EMOPIA and 9315 for processed VGMIDI, which were split into training and test sets at a 10:1 ratio. The pie charts showing the label distribution are presented in Figure 4 under the second to third subplots. It is noted that, as indicated by the statistical conclusions in Table 1, there is almost no correlation between emotion and key. Therefore, the data augmentation by transposing to 15 keys is unlikely to significantly impact the label distribution.

3.3. Rough4Q Dataset

The Rough4Q dataset is a large-scale dataset created by automatically annotating a substantial amount of well-structured sheet music based on conclusions from correlation statistics provided in Table 1. The data sources for this dataset, detailed in Table 2, include both scores in XML series (XML / MXL / MusicXML) and ABC notation format scores. It is noted that not all datasets within the data source include chord markings. Since this paper focuses solely on melody generation, the absence of chord information is not a significant concern for the current study. After filtering out erroneous or duplicated scores and consolidating these into a unified XML format, we utilized music21 to rapidly extract features. Due to the high volume of data, we chose a few representative and computationally manageable features for approximate emotional annotation.

Table 2. Comparison of source datasets for Rough4Q by size in ascending order.

Dataset	Size	Average Bars	Main Genre	With Chord Mark	Original Format	Published Year
Midi-Wav Bi-directional Pop Music [30]	111	40	Pop	no	MusicXML	2021
JSBach Chorales [31]	366	17	Classic	yes	MXL	2010
Nottingham [32]	1,015	21	Folk	yes	ABC notation & MIDI	2011
Wikifonia [33]	6,394	41	Mixed	yes	MXL	2018
Essen Folk Song [34]	10,369	11	Folk	no	ABC notation	2013
IrishMAN [13]	216,281	20	Folk	partial	ABC notation & XML	2023

According to the correlation statistics in Table 1, valence is significantly positively correlated only with mode. Therefore, mode was selected as the feature for determining the valence dimension, with minor mode classified as low valence and major mode as high valence. For arousal, it is significantly positively correlated with pitch range, pitch SD, and RMS. Given that RMS calculation requires audio rendering, which is impractical for large-scale automatic annotation, it was excluded. Among the features pitch range and pitch SD, the correlation between arousal and pitch SD is stronger. Moreover, pitch SD not only partially reflects pitch range but also indicates the intensity of musical variation, providing a richer set of information. Therefore, we tentatively select pitch SD as the benchmark for determining the arousal dimension, classifying scores below the median as low arousal and those above the median as high arousal. This approach yields a rough Russell 4Q label based on the V/A quadrant.

This rough labeling with noise primarily serves to record the state of mode and pitch SD as emotion-related embeddings, ensuring consistency with the format of the two datasets in Section 3.2. Following this, we applied the same data processing methods as those described for the two datasets, preserving labels while segmenting the scores. Notably, the IrishMAN in Table 2 was also the dataset used for backbone pre-training. But it discards scores longer than 32 measures, leading to a significant loss of data. In contrast, our segmentation approach preserves these longer scores.

We discovered that the data were highly imbalanced after processing, with the quantities of Q3 and Q4 labels differing by an order of magnitude from the other categories. To address this imbalance, we performed data augmentation by transposing Q3 and Q4 categories across 15 different keys only. As a result of these processes, we ultimately obtained the Rough4Q dataset, which now comprises approximately 521K samples in total and is split into training and test sets at a 10:1 ratio. The distribution of different categories within the dataset is illustrated in the last pie chart of Figure 4.

4. Experiments

Strictly speaking, the statistical correlation tests presented in Section 2.2 should be included in the current section. However, since their conclusions impact the design of the subsequent sections up to the current one, they have been retained in their original location. Instead, the details of the comparison experiments on calculating music21 parsing rates across different datasets for backbone fine-tuning, as well as the ablation experiments controlling for various emotion-related features, will be presented in this section.

4.1. Music21 Parsing Rate

music21 parsing rates refer to the proportion of musical scores that can be successfully parsed by music21 without errors. This metric helps identify and filter out erroneous scores, which may cause rendering failures and compromise the generation quality. We fine-tuned the backbone using the processed EMOPIA, processed VGMIDI, and Rough4Q datasets introduced in Section 3. Since these datasets share the same structure, comprising three columns: control code, ABC chars, and 4Q label, we parsed the 4Q label into string forms and merged them with the control code column. Given that the attention mechanism of the Transformer decoder is rightward [35], we embedded the 4Q label to the left of the control code. To facilitate the model's understanding of the additional emotion labels, we used the same format for embedding the emotion labels as that in the ABC notation header, where the flag "A:" was previously unused in the current dataset and originally denoted *area* but is now repurposed to signify *affection*. For instance, embedding the Q1 label to the left of the control code results in "A:Q1\n".

We fine-tuned the backbone on these three datasets using a single H800 GPU in a Linux environment, with a batch size of 1 and an early stopping training strategy. Training was halted once the evaluation loss during fine-tuning dropped below the minimum evaluation loss observed for the pre-trained model, and the model weights demonstrating the best performance were saved. This approach helps mitigate overfitting and prevents the model from generating melodies that are too similar to those in the training set. All trained weights are publicly available on HuggingFace. We used the three fine-tuned models obtained as described above for inference, generating 100 pieces of ABC notation from each model. The music21 parsing rates of these scores were then calculated, whose statistical results are presented in Table 3.

Table 3. The comparison of sample sizes for the experiments and music21 parsing rates among outputs from backbones fine-tuned by processed EMOPIA, processed VGMIDI, and Rough4Q.

Dataset		Processed EMOPIA	Processed VGMIDI	Rough4Q
$\frac{sample\ size}{total\ size} = sampling\ rate\ (\%)$	Q1	28.852	40.930	0.742
	Q2	25.000	80.734	1.886
	Q3	25.071	100.000	1.435
	Q4	20.725	42.105	0.729
	Total	24.581	56.683	1.014
music21 parsing rate (%)		28	75	99

The quality of a piece of music is a relatively subjective metric that requires extensive subjective testing to minimize bias and obtain reliable results. While it cannot be fully defined by error-free state of a score, having an error-free score is only one of the necessary conditions for high quality. Although it does not completely represent quality, it is a more objective measure and a prerequisite for high quality. Therefore, this experiment reflects the quality of the generated results to a certain extent.

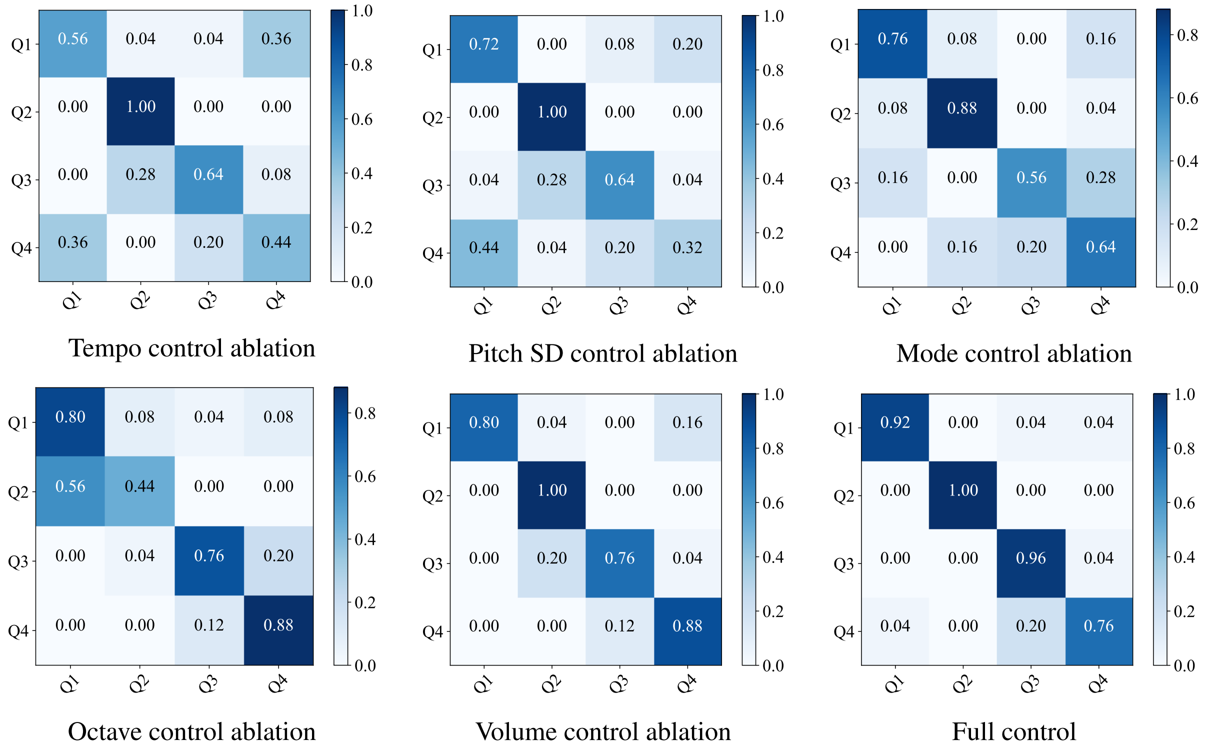
4.2. Ablation Study

Based on the previous experimental results, models fine-tuned by processed EMOPIA and processed VGMIDI were found to have unsatisfactory music21 parsing rates for emotion-conditioned melody generation. Therefore, in subsequent experiments, we used a backbone fine-tuned with Rough4Q for further research. Based on the statistical analysis presented in Table 1, we selected the following five emotion control features to design the emotion-conditioned generation template: mode, tempo, pitch SD, volume, and octave. Here, volume is guided by the RMS statistic, and octave is guided by the avg pitch. The specific control template is as follows: mode, tempo, and pitch SD are managed through embedding, with additional corrections applied at the output stage to enforce adherence to mode and tempo. If the generated result does not meet the expected mode and tempo, its mode and tempo values will be forcibly replaced by the desired parameters. Based on the statistical analysis in Table 1, valence is positively correlated with mode. Furthermore, major is predominantly distributed in the high valence quadrant, while minor is predominantly distributed in the low valence quadrant. Therefore, for generating high valence music, the mode is set to major, and for low valence music, it is set to minor. Arousal is weakly positively correlated with tempo, so for high arousal music, a tempo value above the median is selected; additionally, according to Figure 2, the tempo distribution peak for Q2 music is higher than for Q1, so a higher tempo is set for Q2 compared to Q1. Arousal is positively correlated with pitch SD, thus for high arousal music, a higher pitch SD is selected. Although avg pitch has not been incorporated into embedding, it guides the octave control feature, which is replaced by octave transposition operations (e.g., to create a low avg pitch effect, an octave drop is used). Finally, RMS, which guides volume, is positively correlated with arousal; therefore, for high arousal music, the volume is increased to enhance RMS.

Based on the above analysis and guidance from music psychology conclusions [19–22] on feature control template design, we created an emotion control template as follows:

- Q1: major mode, high pitch SD, tempo between 160-184 BPM (corresponding to *Allegro – Vivace*), no change in octave, volume increased by 5 dB;
- Q2: minor mode, high pitch SD, tempo between 184-228 BPM (corresponding to *Presto – Prestissimo*), octave lowered by 2 octaves, volume increased by 10 dB;
- Q3: minor mode, low pitch SD, tempo between 40-69 BPM (corresponding to *Largo – Adagio*), octave lowered by 1 octave, volume unchanged;
- Q4: major mode, low pitch SD, tempo between 40-69 BPM, no change in octave and volume.

Using this template, we generated 25 pieces of music for each emotion category, totaling 100 pieces. Under blind conditions, we arranged for three groups of listeners to listen to the pieces and label them according to their perceived 4Q emotion in four-alternative forced-choice (4AFC) tasks, each group consisted of 10 music enthusiasts with a limited background in music and had to select one of the four quadrants. To minimize subjective bias [36], we employed a best-of-three format: if at least two of the three groups identified a piece as a specific emotion, that emotion was considered the true emotion. If all three groups provided three different responses, we randomized and replaced members to retake the test until all discrepancies were resolved, following above steps, we finally used up 40 listeners. After obtaining the listeners' labels, we compared them with the emotion condition in prompts to test the effectiveness of the emotion control template. The emotion generation accuracy of the current template was found to be 91%. Additionally, we conducted ablation experiments on the five control conditions within the template in the same way, with all results in Table 4 and confusion matrices in Figure 5.



Tempo control ablation

	Q1	Q2	Q3	Q4
Q1	0.56	0.04	0.04	0.36
Q2	0.00	1.00	0.00	0.00
Q3	0.00	0.28	0.64	0.08
Q4	0.36	0.00	0.20	0.44

Pitch SD control ablation

	Q1	Q2	Q3	Q4
Q1	0.72	0.00	0.08	0.20
Q2	0.00	1.00	0.00	0.00
Q3	0.04	0.28	0.64	0.04
Q4	0.44	0.04	0.20	0.32

Mode control ablation

	Q1	Q2	Q3	Q4
Q1	0.76	0.08	0.00	0.16
Q2	0.08	0.88	0.00	0.04
Q3	0.16	0.00	0.56	0.28
Q4	0.00	0.16	0.20	0.64

Octave control ablation

	Q1	Q2	Q3	Q4
Q1	0.80	0.08	0.04	0.08
Q2	0.56	0.44	0.00	0.00
Q3	0.00	0.04	0.76	0.20
Q4	0.00	0.00	0.12	0.88

Volume control ablation

	Q1	Q2	Q3	Q4
Q1	0.80	0.04	0.00	0.16
Q2	0.00	1.00	0.00	0.00
Q3	0.00	0.20	0.76	0.04
Q4	0.00	0.00	0.12	0.88

Full control

	Q1	Q2	Q3	Q4
Q1	0.92	0.00	0.04	0.04
Q2	0.00	1.00	0.00	0.00
Q3	0.00	0.00	0.96	0.04
Q4	0.04	0.00	0.20	0.76

Figure 5. Confusion matrices between human blind listening emotions of generated music and emotion prompts under full control and ablation options, in which vertical axes represent the emotion prompts, while the horizontal axes represent the emotions labeled by the participants.

Table 4. Performances of the generation model by comparing human blind listening emotions of generated music and emotion prompts with ablation comparsion in accuracy ascending order.

Ablation	Accuracy	F1-Score	Precision	Recall
Tempo	0.660	0.649	0.648	0.66
Pitch SD	0.670	0.648	0.656	0.670
Mode	0.710	0.708	0.713	0.710
Octave	0.720	0.712	0.740	0.720
Volume	0.860	0.859	0.871	0.860
-	0.910	0.909	0.916	0.910

5. Conclusions

From all the previous experimental results, it is evident that although the EMOPIA and VGMIDI datasets contain emotion labels, converting them to ABC notation format for fine-tuning our backbone does not guarantee error-free melody generation. Therefore, a more reliable approach is to first ensure that the generated scores are error-free before applying emotion-conditioned control. In this approach, the statistical correlations between emotion and features obtained from EMOPIA and VGMIDI datasets provide useful guidance for emotion-conditioned control. For instance, the statistical conclusions for features such as key, mode, tempo, pitch range, pitch SD, and RMS are generally consistent with music psychology findings [19–22]. However, for features like direction and avg pitch, the statistical conclusions based on the current data range do not fully align with music psychology. For instance, angry music is characterized by high pitch and ascending pitch, which should correspond to the Q2 quadrant with higher arousal, but the statistical results show a weak negative correlation between arousal and direction or avg pitch. This discrepancy might be due to the complexity of emotional judgments, as direction and avg pitch do not form simple, contiguous regions in the emotional space. For instance, tense music in the Q2 quadrant can also feature low avg pitch [21]. Consequently, direction and avg pitch were not included in the embedding for this study; instead, octave, which is guided by avg pitch and easily controlled at the output stage, was used as a proxy for avg pitch. The remaining four control features, which are tempo, pitch SD, mode, and volume, were selected based on their alignment with music psychology conclusions derived from the previous six features. Since key was validated to have no relation to emotion, RMS was substituted with volume control, and pitch range was not used in the final selection after considering it against pitch SD.

Based on the five controlled features including tempo, pitch SD, mode, octave, and volume, a set of emotion-conditioned control templates was designed. The effectiveness of this template was validated through a blind listening test, demonstrating that an end-to-end approach with emotion embedding is not quite necessary for generating emotion-conditioned music. Instead, using feature engineering combined with statistical correlations and music psychology conclusions can achieve an emotion generation accuracy of 91%. This study explores the feasibility of using the ABC notation system for emotion-conditioned melody generation decoupling the backbone and features from labels. Furthermore, ablation experiments indicate that the five selected control features all contribute to the accuracy of emotion-conditioned control and are crucial for the process.

6. Limitations

However, our current work still faces several limitations. For instance, conclusions derived from statistical correlations only provide a rough guide for designing emotional templates and do not fully reflect the true distribution of features within the emotional space. Additionally, due to the relatively small amount of data and the concentration on pop and game music styles, our analysis results are susceptible to Simpson's Paradox [37], this heavily weighted bias towards folka music in Rough4Q would also affect the model's ability to generate melodies of other styles. Furthermore, melody generation based on emotional control templates often results in music that is concentrated on a few specified emotions, rather than representing the complete emotional quadrant. For instance, when aiming to generate music for the Q2 quadrant, specifying templates may lead to a concentration on tense music, whereas anger and some other emotions also fall within Q2. Although this approach allows for high precision in 4Q representation, it may lead to a lack of emotional diversity in the generated music.

To address these issues, we have released an application demonstration on HuggingFace based on the inference code of our generation system. This demonstration enables users to design and specify emotional templates, utilizing large-scale data to progressively refine feature distributions for greater accuracy. Additionally, the music21 parsing rate is merely a necessary condition for quality but does not fully reflect the true quality of the generated melodies. Future work could incorporate reinforcement learning feedback in the demonstration to adjust the system's generation quality based on user-generated evaluations. Furthermore, while this study focuses on melody, chords are a crucial factor influencing musical emotion. Therefore, our demonstration also includes an option to add chords, and their impact will be considered in future research.

Author Contributions

M.Z.: contributed to conceptualization, methodology, software development, data curation, and original draft preparation, writing, reviewing, editing and validation. X.L.: conceptualization, reviewing and supervision. F.Y.: conceptualization, reviewing and supervision. W.L.: conceptualization, reviewing, validation, editing and supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Social Science Fund of China (21ZD19), and the Special Program of the National Natural Science Foundation of China (T2341003).

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

The code of this work is available at https://github.com/monetjoe/EMelodyGen.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Ji, S.; Yang, X.; Luo, J. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *Acm Comput. Surv.* **2023**, *56*, 1–39.
2. Xin, Y. MusicEmo: transformer-based intelligent approach towards music emotion generation and recognition. *J. Ambient. Intell. Humaniz. Comput.* **2024**, *15*, 3107–3117.
3. Bao, C.; Sun, Q. Generating Music With Emotions. *IEEE Trans. Multimed.* **2023**, *25*, 3602–3614.
4. Ji, S.; Yang, X. Emomusictv: Emotion-conditioned symbolic music generation with hierarchical transformer vae. *IEEE Trans. Multimed.* **2023**, *26*, 1076–1088.
5. Zhu, H.; Wang, S.; Wang, Z. Emotional music generation using interactive genetic algorithm. In Proceedings of the 2008 International Conference on Computer Science and Software Engineering, Wuhan, China, 12–14 December 2008.
6. Zheng, L.; Li, C. Real-Time Emotion-Based Piano Music Generation using Generative Adversarial Network (GAN). *IEEE Access* **2024**, *12*, 87489–87500.
7. Huang, J.; Chen, K.; Yang, Y.H. Emotion-driven Piano Music Generation via Two-stage Disentanglement and Functional Representation. In Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR, San Francisco, CA, USA, 10–14 November 2024.
8. Zhang, J.; Fazekas, G.; Saitis, C. Fast diffusion gan model for symbolic music generation controlled by emotions. *arXiv* **2023**, arXiv:2310.14040.
9. Sulun, S.; Davies, M.E.; Viana, P. Symbolic music generation conditioned on continuous-valued emotions. *IEEE Access* **2022**, *10*, 44617–44626.
10. Ferreira, L.N.; Mou, L.; Whitehead, J.; et al. Controlling Perceived Emotion in Symbolic Music Generation with Monte Carlo Tree Search. In Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE, Pomona, CA, USA, 24–28 October 2022.
11. Grekow, J.; Dimitrova-Grekow, T. Monophonic music generation with a given emotion using conditional variational autoencoder. *IEEE Access* **2021**, *9*, 129088–129101.
12. Oliwa, T.M. Genetic algorithms and the abc music notation language for rock music composition. In Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation; Association for Computing Machinery: New York, NY, USA, 12–16 July 2008.
13. Wu, S.; Li, X.; Yu, F.; et al. TunesFormer: Forming Irish Tunes with Control Codes by Bar Patching. In Proceedings of the 2nd Workshop on Human-Centric Music Information Retrieval 2023 Co-Located with the 24th International Society for Music Information Retrieval Conference (ISMIR), Milan, Italy, 10 November 2023.
14. Casini, L.; Jonason, N.; Sturm, B.L.T. Investigating the Viability of Masked Language Modeling for Symbolic Music Generation in abc-notation. In *Artificial Intelligence in Music, Sound, Art and Design*; Johnson, C., Rebelo, S.M., Santos, I., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 84–96.

15. Wu, S.; Wang, Y.; Li, X.; et al. MelodyT5: A Unified Score-to-Score Transformer for Symbolic Music Processing. In Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR, San Francisco, CA, USA, 10–14 November 2024.
16. Hung, H.T.; Ching, J.; Doh, S.; et al. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR, Online, 7–12 November 2021.
17. Ferreira, L.; Whitehead, J. Learning to Generate Music With Sentiment. In Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, Delft, The Netherlands, 4–8 November 2019.
18. Cuthbert, M.S.; Ariza, C. Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands, 9–13 August 2010.
19. Scherer, K.R.; Coutinho, E. How music creates emotion: A multifactorial process approach. In *The Emotional Power of Music, Multidisciplinary Perspectives on Musical Arousal, Expression, and Social Control*; Oxford University Press: Oxford, UK, 2013; pp. 121–145.
20. Thompson, W.F.; Quinto, L. Music and emotion: Psychological considerations. *Aesthetic Mind Philos. Psychol.* **2011**, 357–375.
21. Granot, R.Y.; Eitan, Z. Musical tension and the interaction of dynamic auditory parameters. *Music. Percept.* **2011**, *28*, 219–246.
22. Juslin, P.N.; Sloboda, J.A. Music and Emotion. In *The Psychology of Music*; Academic Press: Cambridge, MA, USA, 2001.
23. Russell, J.A. Measures of Affect: A Review. In *Emotion: Theory, Research, and Experience*; Plutchik, R., Kellerman, H., Eds.; Academic Press: Cambridge, MA, USA, 1989; Volume 4, pp. 83–111.
24. Russell, J. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161.
25. De, S.; Dutta, P. *Computational Intelligence for Human Action Recognition*; CRC Press: Boca Raton, FL, USA, 2020.
26. Pearson, K. II. Mathematical contributions to the theory of evolution. II. Skew variation in homogeneous material. *Proc. R. Soc. Lond.* **1985**, *57*, 257–260.
27. Müller, J.M. Computational Approaches to Symbolic Music Analysis: Weighting by Note Duration. *J. Comput. Musicol.* **2021**, *29*, 45–60.
28. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
29. McFee, B.; Raffel, C.; Liang, D.; et al. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conference 2015 (SciPy 2015), Austin, TX, USA, 6–12 July 2015.
30. Liu, Z.; Li, Z. *Music Data Sharing Platform for Computational Musicology Research (CCMUSIC DATASET)*; Zenodo: Beijing, China, 2021. https://doi.org/10.5281/zenodo.5676893,
31. Wu, S.; Li, X.; Sun, M. Chord-conditioned melody harmonization with controllable harmonicity. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
32. The Nottingham Music Database. 2011. Available online: https://ifdo.ca/ seymour/nottingham/nottingham.html (accessed on 16 September 2025).
33. Simonetta, F. *Enhanced Wikifonia Leadsheet Dataset*; Zenodo: Beijing, China, 2018. https://doi.org/10.5281/zenodo.1476555.
34. Essen Folk Song Database. 2013. Available online: https://ifdo.ca/ seymour/runabc/esac/esacdatabase.html (accessed on 16 September 2025).
35. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.
36. Kang, C.; Lu, P.; Yu, B.; et al. EmoGen: Eliminating Subjective Bias in Emotional Music Generation. *arXiv* **2023**, arXiv:2307.01229.
37. Simpson, E.H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser.* **1951**, *13*, 238–241.