VARIATION TRANSFORMER: NEW DATASETS, MODELS, AND COMPARATIVE EVALUATION FOR SYMBOLIC MUSIC VARIATION GENERATION

Chenyu Gao[1] **Federico Reuben**[1] **Tom Collins**[2,3]
[1] School of Arts and Creative Technologies, University of York, UK
[2] Frost School of Music, University of Miami, FL, USA
[3] MAIA, Inc., Davis, CA, USA
{chenyu.gao,federico.reuben}@york.ac.uk, tomthecollins@gmail.com

ABSTRACT

Variation in music is defined as repetition of a theme, but with various modifications, playing an important role in many musical genres in developing core music ideas into longer passages. Existing research on variation in music is mostly confined to datasets consisting of classical theme-and-variation pieces, and generative models limited to melody-only representations. In this paper, to address the problem of the lack of datasets, we propose an algorithm to extract theme-and-variation pairs automatically, and use it to annotate two datasets called POP909-TVar (2,871 theme-and-variation pairs) and VGMIDI-TVar (7,830 theme-and-variation pairs). We propose both non-deep learning and deep learning based symbolic music variation generation models, and report the results of a listening study and feature-based evaluation for these models. One of our two newly proposed models, called Variation Transformer, outperforms all other models that listeners evaluated for "variation success", including non-deep learning and deep learning based approaches. An implication of this work for the wider field of music making is that we now have a model that can generate material with stronger and perceivably more successful relationships to some given prompt or theme. [1]

1. INTRODUCTION

The term variation refers to "a form founded on repetition, and as such an outgrowth of a fundamental musical and rhetorical principle, in which a discrete theme is repeated several or many times with various modifications" [1]. In western music, variation is a technique in which the theme is repeated but in an alternate form with various modifications in one or more aspects of melody, rhythm, harmony, texture, instrumentation, etc. An example from game music is provided in Figures 1(a) and (b), where the top and bottom staves in (b) are embellished compared to (a), but the melodic and harmonic structure is largely the same.

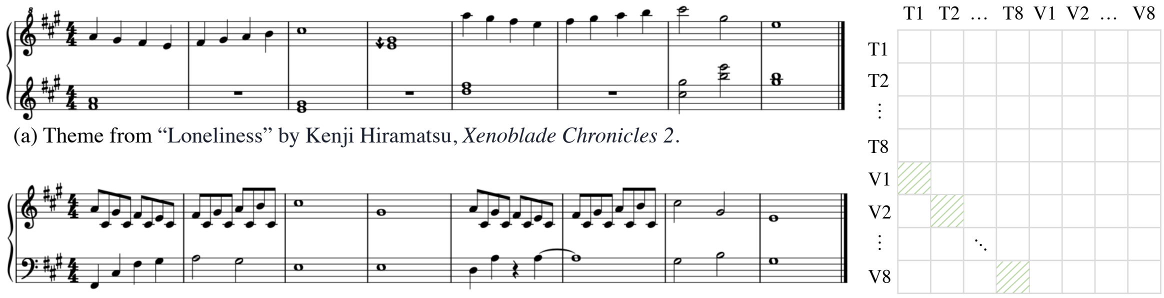
In recent years, a number of music generation algorithms and commercialised artificial intelligence (AI) music generation systems have emerged [2–9], but there are only a few studies focusing on symbolic music variation generation [10–14]. Although some infilling systems claim that they have the potential to generate variations [7, 8, 15, 16], an infilling system may sometimes work to continue writing [17] rather than always varying the theme (i.e., generate content with a strong relationship to the given prompt). Accepting a musical input prompt but destroying the original music idea is a "lack of control" issue, and could frustrate composers [18, 19]. It also leaves the presence and perception of rhetorical or narrative content to serendipity (chance), which goes against the rhetorical principle of the definition of musical variation.

Existing music variation research is mostly confined to datasets consisting of classical theme-and-variation pieces [20] or monophonic folk music [21], and most of existing music variation generation models are also limited to varying melody only [11, 12, 14].

To address the issues above, in this paper we develop both new datasets and models for symbolic polyphonic music variation generation. For data annotation, we develop an algorithm for theme-and-variation extraction, and apply it to annotate two datasets: POP909 [22], and VGMIDI [23]. For model design, we propose both deep and non-deep learning-based models, as another shortcoming of recent research is that evaluations ignore models published prior to c. 2015 – assuming, rather than actually testing whether, deep learning approaches are superior for music generation [13, 14, 24]. Three research questions are addressed: **RQ1:** To what extent can AI models generate successful music variations? **RQ2:** Can deep learning approaches outperform non-deep learning approaches on music variation generation? **RQ3:** Would variation generation tools be useful? We conduct a listening study and feature-based evaluation to address these research questions, and finish by discussing the implications of the study's findings for music generation and the field of MIR.

[1] Demos: https://variation-transformer.glitch.me. Code and datasets: https://github.com/ChenyuGAO-CS/Variation-Transformer.

Attribution: C. Gao, F. Reuben, and T. Collins, "Variation Transformer: New datasets, models, and comparative evaluation for symbolic music variation generation", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.



(a) Theme from “Loneliness” by Kenji Hiramatsu, *Xenoblade Chronicles 2*.

(b) A variation of (a) from “Loneliness ” by Kenji Hiramatsu, *Xenoblade Chronicles 2*.

(c) Measure-level encodings.

Figure 1. (a) Theme from “Loneliness” by Kenji Hiramatsu, *Xenoblade Chronicles 2*; (b) Variation from the same piece; (c) Measure-level encodings based on this theme-variation example, discussed in a subsequent section. T_m denotes the encoding of the mth measure of the theme, while V_m denotes the encoding of the mth measure of the variation. The shaded areas are filled with 1s, while the blank areas are filled with 0s, indicating how we condition the model to attend to specific parts of the theme when generating a variation.

2. RELATED WORK

2.1 Symbolic music generation approaches

Before deep learning models became a popular approach for music generation, many models were based on Markov chains [25–28]. Markov models assume that the current state prediction depends on one (first-order Markov model) or more (second-, third-order models) previous states. They have been used to generate music in many styles, and recent work [29] finds evidence that ratings of the stylistic success of their outputs are on a par with deep learning models such as Music Transformer [30].

Among a large number of deep learning approaches to symbolic music generation [30–40], the most popular architectures are the generative adversarial network (GAN) [41], variational autoencoder (VAE) [31, 42, 43], and transformer [44]. More recently, there are also some attempts to adopt the diffusion model [45, 46] to generate music [39, 40]. But to date, we observed these diffusion-based algorithms suffer from the problem of a lack of structure in long-term music generation.

2.2 Symbolic music variation generation

As a sub-task of symbolic music generation, symbolic music *variation* generation takes a prompt/theme as input, and aims to generate variations where the new material is different to the theme but remains musically relatable. Some variation generation approaches are based on genetic algorithms [10, 11], but drawbacks are that they have only been applied to sequential representations of monophonic music and are reliant on manually designed rules [47].

There are also variation generation methods based on probabilistic methods. For example, an algorithm mentioned in [48] starts with the same first beat as the theme, and subsequent beats are generated by a Markov model. To ensure the generated variation begins and ends somewhere “sensible”, the Markov process can be run forwards and backwards from such start and end points, with a join in the middle of a generated phrase that may break the Markov property [26, 48]. Compared to variation generation, these two methods are more like style-composing. In contrast, an idea in [49] is to decide if each of the states in an existing sequence (such as a theme) should be replaced by another state according to a corresponding probability distribution. However, this approach has only ever been applied to monophonic variation generation.

Compared to non-deep learning approaches, an advantage of deep learning for music generation is that it places less emphasis on the domain knowledge/expertise of the programmer – the trained network weights *should* take on responsibility for generalizing the style/structure of the training data.

There are only a few studies for music variation generation that have adopted deep learning methods. This is because most deep learning approaches are data-driven, and existing theme-and-variation datasets are either relatively small classical music datasets (e.g., the TAVERN dataset [20] with 17 works by Beethoven and 10 by Mozart for a total of 281 variations) or monophonic folk music datasets [21], which restricts the development of deep methods for music variation generation. Although Music Transformer is adopted in [13] for jazz variation generation, the JAZZVAR dataset (502 theme-variation pairs) proposed in this study is still relatively small for deep learning model training. Besides, the lack of listening studies and comparative evaluation makes it difficult to conclude to what extent these models are effective in generating musical variations [13, 14, 24].

3. DATASET

In this section, we introduce our algorithm for automatic extraction of variations from a collection containing annotated themes. We use it to extract theme-and-variations (TVar) pairs from the POP909 [22] and VGMIDI [23] datasets. As a result, two new datasets (POP909-TVar, and VGMIDI-TVar) are constructed.

3.1 Construction of the POP909-TVar dataset

The POP909 dataset [22] contains piano arrangements of 909 Chinese popular songs in MIDI format. We use 809 pieces (∼90%) for training, and 100 (∼10%) for testing.

As repetitive phrase annotations are provided in the POP909 dataset [50], we use these to estimate the lower and upper bounds of the similarity between human-composed themes and variations by utilizing a symbolic fingerprinting-based similarity calculation [51, 52]. The first occurrence of each repetitive pattern is regarded as the theme, and the following occurrences are regarded as variations. For each theme, we record the minimum and maximum similarity scores between it and its variations. The similarity lower bound is the average of the per-theme minimum scores, and the upper bound is defined correspondingly, with values of 53.03 and 70.95, respectively.

Algorithm 1 TVar extraction on the POP909 dataset

```
Input: Repetitive pattern labels ($\mathbf{P}$) and MIDIs ($\mathbf{M}$) of the dataset, similarity upper bound $u$ and lower bound $l$
Output: TVar pairs
 1: for $p \in \mathbf{P}$ do
 2:     Separate the first occurrence of $p$ as the theme $t$, and the subsequent occurrences as an array $\mathbf{V}_{\text{Rep}}$
 3:     $\mathbf{V}_{\text{Match}} \leftarrow$ match_occ $(t, \mathbf{M}, u, l)$
 4:     Push $\mathbf{V}_{\text{Match}}$ into $\mathbf{V}_{\text{Rep}}$
 5:     for $v \in \mathbf{V}_{\text{Rep}}$ do
 6:         if similarity$(t, v) > u$ then
 7:             Filter out $v$
 8:         else
 9:             if Similarity score between $v$ and the previous occurrence $> u$ then
10:                 Filter out $v$
11:             else
12:                 Push $v$ into $\mathbf{V}_{\text{Out}}$
13:     if Occurrence count of $\mathbf{V}_{\text{Out}} \geq 1$ then
14:         return $t, \mathbf{V}_{\text{Out}}$
```

The pseudocode for TVar extraction is given in Algorithm 1. We take the first occurrence of repetitive patterns as themes when applying our algorithm to POP909 (line 2).[2] Variations are extracted from both human-annotated patterns (line 2) and the whole dataset (line 3). When extracting variations from human annotations, we exclude variations whose similarity score is larger than the similarity upper-bound (lines 6-7), since we aim to train models to generate variations where there is new but theme-relatable material. When extracting variations of a theme on the whole dataset, we run a symbolic fingerprinting-based pattern-matching approach [51–53] using the same lower and upper bounds mentioned previously to retain variations (line 3). We also filter out variations that are too similar to existing variations (lines 9-12).

The POP909-TVar dataset is constructed by applying our TVar extraction algorithm to POP909, giving 2,609 TVar pairs in the training set, and 262 TVar pairs in the test set.

3.2 Construction of the VGMIDI-TVar dataset

The VGMIDI dataset [23] contains piano arrangements of game music in MIDI format recorded by human performers.[3] There are three subsets in VGMIDI: the largest has 2,520 MIDI files for music generation model training, the second one has 136 MIDI files with emotion labels, and the third one (272 MIDI files) is for music discriminator training, which involves both human-composed music and fake data. Here, we merged the largest subset and the subset with emotional labels and adhered to the original train-test split, obtaining 2,301 MIDI files for training and 355 for testing.

Compared to popular music, we infer there could be greater scope for new material in variations in game music, so we reduce the similarity lower bound to 30 but keep the similarity upper bound as 70.95. Also, we restrict the extracted variation and the theme to come from the same song. Then, we follow the steps as in Section 3.1 to obtain variations. In contrast to the POP909 dataset, repetitive patterns are not annotated in the VGMIDI dataset, so we run a slice window with size $= 8$ measures and step $= 4$ measures from the beginning to the end of the song to extract theme samples. The similarity between each new theme and previous themes is calculated to filter out theme samples that are too similar (similarity score $>$ upper-bound) to existing themes. The variation extraction function `match_occ()` is applied to each of the theme samples, and then the matched occurrences will be filtered by the same processes as that in Algorithm 1 (lines 5-12). Only theme samples with more than one variation will be retained (lines 13-14 in Algorithm 1).

The VGMIDI-TVar dataset is constructed by applying the above steps to VGMIDI, giving 6,790 TVar pairs in the training set, and 1,040 in the test set.

4. MUSIC VARIATION GENERATION MODELS

In this section, we introduce two new music variation generation models: one is a deep-learning model called Variation Transformer, and the other acts as a non-deep learning baseline called Variation Markov.

4.1 Variation Transformer

Variation Transformer builds on Music Transformer [30], utilizing the REMI representation [54] to encode incoming MIDI files. The design of the relative positional self-attention [55] alleviates the problems of the regular self-attention that attends only locally or at the beginning for a sequence [56] – Music Transformer is used for jazz variation generation in [13]. However, while developing and testing these models, we observed that Music Transformer's ability to understand the measure-wise relationship between theme and variation was not strong enough. For example, when generating a variation of an 8-measure theme, Music Transformer might generate something new in the first 2 measures, but copy large sections of the theme

[2] First occurrences are not always the *archetypal occurrence*, but it is a reasonable assumption [17].

[3] Sources for this dataset are `https://www.vgmusic.com` and `https://www.ninsheetmusic.org`.

in the following measures [52, 57], showing the failure of the Music Transformer model to learn the theme-and-variation relationship. When a human composer creates a variation, commonly each bar of the variation is relatable to the corresponding measure of the theme (recall Figures 1(a) and (b)). Thus, in this study, we propose the measure-level encodings (Figure 1 (c)) and **t**heme-and-**v**ariation **Attention** (tvAttn) to force the transformer architecture to take into account more information about a specific measure of an existing theme when generating the corresponding measure of a new variation, calling our new model the Variation Transformer.

Figure 1(c) shows the measure-level encoding (which we will notate E_{bar}) to capture the relationship between corresponding measures of theme and variation, with a size of $N \times N$, where N is the length of the encoding of theme concatenated with variation. The formula for tvAttn is then

$$\text{tvAttn} = \text{Softmax}\left((1 + \mathbf{w}E_{\text{bar}})\frac{QK^{\top} + S^{\text{rel}}}{\sqrt{D_h}}\right)V, \quad (1)$$

where $\mathbf{w}$ is a learnable parameter, and E_{bar} is the measure-level encodings. Q represents the queries, K is the set of keys, V is the set of values, $1/\sqrt{D_h}$ is a scale factor, and S^{rel} is to encode the relative positional information between each pair of tokens in a sequence.

4.2 Variation Markov

Based on [26, 58] and inspired by [12, 48], we propose a non-deep learning music variation generation strategy based on Markov models. Polyphonic MIDI inputs are represented as states in a state space consisting of beat in the measure and MIDI note numbers relative to estimated tonal center. The transitions between states observed across our training data are stored in a directed graph.

When generating variations, we extract the beginning and end states of each measure of the theme, and run a "scenic pathfinding algorithm" to find replacement states. This algorithm is adapted from Dijkstra's shortest path algorithm [59]. When finding the shortest path between connected vertices u and v in a graph G, Dijkstra's method always updates the distance from the starting vertex u to other vertices with shorter distances. In our scenic version, we insert an extra piece of logic to determine whether the distance from the starting vertex u to another vertex will be updated to a shorter distance with probability $p = .5$. In this way, more varied musical content is generated, because the path connecting u to v that results on each occasion is not necessarily the shortest. We replace each measure from the theme with the scenic path alternative with probability $q = .5$, and if u and v are not connected (due to not being observed in a training data sequence), then we retain the original measure from the theme.

5. EVALUATION

5.1 Experimental design

We conduct a listening study and feature-based evaluation on both POP909-TVar and VGMIDI-TVar datasets. The variation generation ability of three transformer-type models (TTMs) – fast-Transformer (FaTr) [37, 60, 61], Music Transformer (MuTr) [13, 30], and Variation Transformer (VaTr) – and Variation Markov (VaMa) is compared.

For fair comparison, we use the REMI representation [54] to represent MIDI files for all three TTMs, which were trained on A40 GPUs with a batch size of 16 for 10 epochs on each of the two training sets. The learning rate is set as 1×10^{-4} for the first 5 epochs, then decreased to 5×10^{-5} for the last 5 epochs. For model training, we concatenate each theme-and-variation pair, with a `[Separate]` token inserted between the theme and variation. Ten variations were generated by each algorithm with using each theme in the test set as an initial prompt to provide the pool of stimuli for evaluation.

For hypothesis testing, we utilize a Bayes factor analysis (BFA, [62]), where the ratio of the marginal likelihoods of the alternative hypothesis H_1 to the null hypothesis H_0 is calculated, and notated BF_{10}. A large value of BF_{10} suggests there is strong evidence for H_1. Conversely, a small BF_{10} suggests strong evidence for H_0. A table for interpreting BF_{10} values is provided in [63]. BFA is superior to classical (frequentist) hypothesis testing, because of this ability to find evidence in favor of the null, which in (computational) systems testing corresponds to a meaningful non-difference between systems.

5.2 Listening study

Our listening study is approved by the Ethics Committee of the School of Arts and Creative Technologies at the University of York. Our overall design builds on previous listening studies in this domain (e.g., [49]), and our hypotheses are as follows:

1. In terms of variation success, we predict the following ordering of systems: VaTr > MuTr > FaTr > VaMa.
2. TTMs (VaTr, MuTr, and FaTr) achieve better music quality than VaMa.

5.2.1 Participants

We aim to recruit participants with a relatively high level of music knowledge, using student email lists at the University of York, and the C4DM group of the Queen Mary University of London. [4] Participants are compensated £10 Amazon vouchers for the 30 mins it takes to complete the study. After removing responses that were unfinished or submitted too quickly to fully listen to the music, there are 25 responses under analysis.

Participants' mean age is 25 years old, and their mean years of formal musical training is 10 years. Over 90% of participants listen to music daily, and 80% of participants play music/sing at least weekly.

5.2.2 Stimuli

For the listening study, 15 groups of stimuli were picked randomly from POP909-TVar generated outputs, and 15

[4] We follow the consensual assessment technique [49,64] to design our study, which requires participants be experienced in the relevant domain.

from VGMIDI-TVar outputs. In each group, there are one theme and five variations, in which one is composed by a human, and the other four are generated by the models (VaTr, MuTr, FaTr, and VaMa). Each music excerpt is about 30-sec rendered using a piano sound. Each participant listens to 3 groups of music.

5.2.3 Procedure

After informed consent, instructions, and TVar examples, participants listen to a theme and then each variation, rating the musical dimensions of *variation success*, *stylistic consistency*, *similarity*, *creativity*, and *musical quality* on a 1–7 Likert scale, as well as two additional questions – *willingness to use a system that generates this variation* (*willingness*), and *the extent to which this variation sounds like it is composed by a human* (*is human*).[5] An optional free text box for any comments follows the rating scales. After completing the evaluation of all 3 groups of materials, the extent to which the participant finds an algorithmic variation generation tool useful for their creative practice is rated (same scale), and a final optional free text box for any comments is provided. Given the participant and stimulus numbers, each TVar stimulus group was heard by approximately 3 participants, and all presentation orders were randomized to mitigate ordering and fatigue effects.

5.2.4 Results

Participants' ratings for the features mentioned in Section 5.2.2 are shown as violin plots in Figure 2. For the BFA addressing our hypotheses at the top of Section 5.2, results for Hyp. 1 demonstrate that VaTr outperforms all three other algorithms (MuTr, FaTr, and VaMa) on *variation success ratings*, and MuTr outperforms FaTr and VaMa. But there is no difference between FaTr and VaMa. Results for Hyp. 2 indicate that TTMs (VaTr, MuTr, and FaTr) perform better than VaMa on *musical quality* ratings.

In terms of observations of results not tied to particular hypotheses, human-composed variations (Hu) appear to outperform algorithms on all metrics. In addition to *variation success* mentioned above, VaTr achieves higher ratings than other algorithms for *willingness* on both POP909-TVar and VGMIDI-TVar. The TMMs have higher ratings for *stylistic consistency*, *musical quality*, and *is human* than VaMa, but VaMa shows potential for generating creative variations. For POP909-TVar, VaMa and VaTr receive similar *creativity*, which is higher than that of MuTr and FaTr. For VGMIDI-TVar, although VaMa gets lower *creativity* than VaTr, it is still on par with MuTr and FaTr.

Approximately 100 comments are provided explaining the reasons for ratings, from which we find that participants usually consider the success of a variation according to the musical dimensions of pitch, rhythm, structure, dynamics, key signature, and texture, as well as the four more holistic dimensions mentioned in Section 5.2.2 (*stylistic consistency*, *similarity*, *musical quality*, and *is human*). As such, deviations in these musical dimensions (e.g., dissonance, discordant dynamics, confusing structure) during the generation process could lead to unsatisfactory results.

Usually, a lack of stylistic consistency or being too similar/different to the theme will also result in an unsuccessful variation, but sometimes slight alterations (P11) or varying a lot from the theme (P21) can still lead to high ratings.

When considering whether a variation is written by a human composer or generated by AI, participants usually evaluate it in terms of the musical dimensions of rhythmic repetition, and appearance of dissonance, as well as overall musical quality. Lower-quality music seems to be associated with thoughts of being created by machines. But sometimes, even if the variation is recognised as AI-generated, participants are still receptive to it if the creativity and/or quality of the variation is good (P14 and P21).

The distribution of the extent to which participants find an algorithmic variation generation tool useful for their creative practice is: lower quartile = 3, median = 4, upper quartile = 5 on a 1–7 Likert scale. Corresponding comments comprise the following categories: i) benefit of music variation generation AI (MVG-AI) [18, 19, 66], with 8 out of 25 participants mentioning MVG-AI could be beneficial especially for inspiration; ii) concerns about MVG-AI [18, 19, 66], such as the quality and consistency not being sufficient to replace human composers (P7); iii) the clash between "creative ego" and MVG-AI [19], where for example P1 considers composing as creating art that is meaningful to the individual, which should not be done by AI instead. Similarly, P9 and P15 demonstrate wariness of the implications of AI and reluctance to use generative AI [18]; iv) further support/functionality required [18], such as P14 expecting MVG-AI to be able to produce variations that reflect a composers' own style, and P24 thinking composers may have extra requirements for the MVG-AI in terms of emotional or style targets.

5.3 Feature-based evaluation

We use the whole pool of evaluation materials here, in which ten variations were generated by each algorithm for each theme drawn from the test sets. Three musical features are extracted and evaluated at the measure level:

***Similarity score* (*SS*)** [67] gives the similarity between each measure of the generated variation and the corresponding measure of the theme in terms of pitch and rhythm.

***Translational coefficient consistency*(*TC*)** [68] estimates the complexity or music-repetitive structure of an excerpt. A lower TC value means a music excerpt is highly repetitive, and vice versa. Here we calculate the absolute difference between the TC of each measure of the generated variation and the theme.

***Key signature consistency* (*KSC*)** [69] captures the percentage of measures of the generated variation that have the same estimated key as the theme.

The evaluation results are shown in Table 1. We found that VaMa has the highest *SS* and *KSC* for both datasets. Among the TTMs, VaTr has higher values than MuTr and

[5] The *variation success* is mainly to address **RQ1** and **RQ2**. Following existing research [37, 40, 65], we also include other music dimension metrics and *is human*. The *willingness* metric is to address **RQ3**.

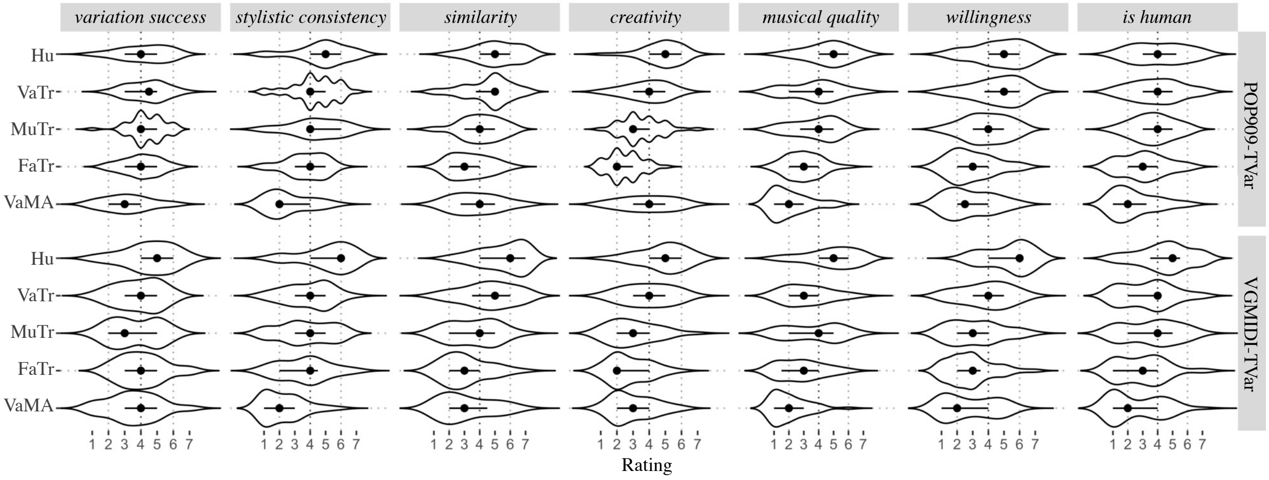


Figure 2. Violin plots for rating (1–7) distributions of seven dimensions on two datasets, where the envelope represents the distribution of responses; the lines indicate rating scales; the horizontal line goes from the lower quartile through the median (point) to the upper quartile.

	POP909-TVar				
Feat.	Hu	VaTr	MuTr	FaTr	VaMa
SS	26.1 (20.6)	19.0 (18.0)	12.0 (15.2)	6.8 (6.6)	15.8 (9.5)
TC	0.10 (0.09)	0.11 (0.10)	0.12 (0.10)	0.12 (0.09)	0.15 (0.11)
KSC	51.7	41.0	29.0	22.4	52.6
	VGMIDI-TVar				
Feat.	Hu	VaTr	MuTr	FaTr	VaMa
SS	25.2 (19.1)	9.7 (16.7)	7.2 (12.9)	4.2 (8.0)	17.4 (14.6)
TC	0.12 (0.13)	0.16 (0.15)	0.17 (0.15)	0.14 (0.13)	0.14 (0.13)
KSC	32.8	22.4	19.7	19.5	52.0

Table 1. The feature evaluation results, with mean and standard deviation (in brackets) for each feature.

FaTr on all three metrics for POP909-TVar, and higher than MuTr and FaTr on *SS* and *KSC* for VGMIDI-TVar. But, VaMa is outperformed by TMMs in most of metrics in the listening study (Section 5.2.4), reflecting that feature-based metrics alone cannot evaluate the performance of models from the human-aesthetic perception of music [70].

6. DISCUSSSION

In this paper, we propose datasets and models for symbolic variation generation. To address our research questions, we run a listening study and feature-based evaluation for both deep and non-deep learning models, as most recent music generation research only compares deep learning approaches. According to our listening study results, human-composed variations outperform algorithms on all metrics, indicating that there is still a gap between human-composed variations and those generated by our proposed algorithms (**RQ1**). One of our proposed models (VaTr) is the strongest for variation generation, which demonstrates the superiority of a deep learning over a non-deep learning approach when the task is as specific as “generate a successful variation of this theme". But our experiment results also show that not all deep learning approaches outperform the non-deep learning approach, especially in creativity (**RQ2**). And so for the less specific task of “generate music in a target style", more research and comparative evaluation is required to establish the superiority of deep learning over alternative music generation methods. We hope that our study encourages researchers to revisit non-deep learning approaches, as well as to test experimentally whether deep learning methods are broadly superior to non-deep learning methods for music-generative tasks. To address **RQ3**, we further explore the extent to which participants in our listening study find MVG-AI useful for their creative practice, with an average rating of 4 on a 1–7 Likert scale, and some of the comments suggest that MVG-AI could lead to powerful tools for inspiration. One of our proposed models VaTr achieves the higher ratings for *willingness* than other models, and a comparable rating for *willingness* as that for human-composed variations on POP909-TVar (Figure 2).

Although the results are promising, there is still plenty of work to do in order to bridge technology and musical creativity. To increase the willingness of users to adopt MVG-AI, it is necessary to improve the quality of music generation and to consider the expectations of users. For example, to mitigate deviations in musical dimensions like dissonance, which lead to unsatisfactory results, adding a post-processing stage could be useful. Some participants mentioned their expectations about personalized AI in our study as well, as in [18, 19]. Using low-rank adaptation techniques [71, 72] to fine-tune a pre-trained model could be a strategy to explore in future. Another topic for future work entails further investigation of the quality of the provided datasets, to validate the reliability of the extracted theme-variation pairs.

Future applications of this work include: being integrated into AI music making systems to enable these systems to generate music with a stronger relationship to the user’s music prompt; being used in video game music domain, either as a tool to provide inspirations for composers, or for in-game generation to reduce listener fatigue [73,74]; and structured music generation [12,24,75].

7. ETHICS STATEMENT

The listening study in this paper is approved by the Ethics Committee of The School of Arts and Creative Technology, University of York. A Participant Consent Form and a Participant Project Information Sheet is included prior to the start of the questionnaire to inform participants of the project and obtain their consent. Participants have the right to withdraw at any time. Each participant's data is protected by anonymization. The data collected involves ratings and comments as described in Section 5.2.2. The demographic information collected only involves participants' age in years, years of formal music training, regularity of playing music or signing, and regularity of listening to music, which are not sufficiently detailed for participants to be identified. No other identifying data are collected. Researchers shuffle the order of their responses, and then record these responses and use anonymized new IDs, which are person 1, person 2, etc. This way, even the researchers will not be able to identify the person after the survey.

Previous work demonstrates that some deep learning approaches that generate music from scratch tend to copy large sections from the training set with a high risk of copyright infringement [52]. In order to mitigate this issue, our models vary the input prompt. Moreover, future work includes further experiments regarding originality of the generation results. Although the training materials come from open-source datasets (POP909 [22], and VGMIDI [23]), it does not mean all the contents are copyright free. There is a possibility of our models to output copyrighted music. Therefore, our models and data are used for academic research only, not for commercially usages.

8. ACKNOWLEDGEMENT

The Viking cluster was used during this project, which is a high performance compute facility provided by the University of York. We are grateful for computational support from the University of York, IT Services and the Research IT team. Chenyu Gao is a PhD student supported jointly by the China Scholarship Council and the University of York.

9. REFERENCES

[1] E. Sisman, "Variations. The New Grove Dictionary of Music and Musicians, edited by Stanley Sadie and John Tyrrell," 2001.

[2] "AIVA," 2016. [Online]. Available: https://www.aiva.ai/

[3] A. Roberts, J. Engel, Y. Mann, J. Gillick, C. Kayacik, S. Nørly, M. Dinculescu, C. Radebaugh, C. Hawthorne, and D. Eck, "Magenta studio: Augmenting creativity with deep learning in ableton live," in *Proceedings of the 7th International Workshop on Musical Metacreation*, 2019.

[4] "Stable Audio," 2023. [Online]. Available: https://www.stableaudio.com/

[5] "Suno," 2023. [Online]. Available: https://www.suno.ai/

[6] "Staccato," 2023. [Online]. Available: https://staccato.ai/

[7] R. B. Tchemeube, J. Ens, C. Plut, P. Pasquier, M. Safi, Y. Grabit, and J.-B. Rolland, "Evaluating human-AI interaction via usability, user experience and acceptance measures for MMM-C: A creative AI system for music composition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 5769–5778.

[8] M. E. Malandro, "Composer's Assistant: Interactive transformers for multi-track MIDI infilling," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023.

[9] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] K. Ishizuka and T. Onisawa, "Generation of variations on theme music based on impressions of story scenes considering human's feeling of music and stories," *International Journal of Computer Games Technology*, vol. 2008, 2008.

[11] V. Arutyunov and A. Averkin, "Genetic algorithms for music variation on genom platform," *Procedia computer science*, vol. 120, pp. 317–324, 2017.

[12] F. Pachet, A. Papadopoulos, and P. Roy, "Sampling variations of sequences for structured music generation." in *Proceedings of 18th International Conference on Music Information Retrieval*, 2017, pp. 167–173.

[13] E. Row, J. Tang, and G. Fazekas, "JAZZVAR: A dataset of variations found within solo piano performances of jazz standards for music overpainting," in *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research*, 2023.

[14] B. Banar, N. Bryan-Kinns, and S. Colton, "A tool for generating controllable variations of musical themes using variational autoencoders with latent space regularisation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, pp. 16 401–16 403.

[15] J. Ens and P. Pasquier, "MMM: Exploring conditional multi-track music generation with the transformer," *arXiv preprint arXiv:2008.06048*, 2020.

[16] R. Guo, I. Simpson, C. Kiefer, T. Magnusson, and D. Herremans, "MusIAC: An extensible generative framework for music infilling applications with multi-level control," in *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, 2022, pp. 341–356.

[17] W. E. Caplin, *Classical form: A theory of formal functions for the instrumental music of Haydn, Mozart, and Beethoven*. Oxford University Press, 1998.

[18] M. Newman, L. Morris, and J. H. Lee, "Human-AI music creation: Understanding the perceptions and experiences of music creators for ethical and productive collaboration," in *Proceedings of 24th International Conference on Music Information Retrieval*, 2023.

[19] K. Worrall and T. Collins, "Considerations and concerns of professional game composers regarding artificially intelligent music technology," *IEEE Transactions on Games*, 2023.

[20] J. Devaney, C. Arthur, N. Condit-Schultz, and K. Nisula, "Theme and variation encodings with roman numerals (TAVERN): A new data set for symbolic music analysis," in *Proceedings of 16th International Conference on Music Information Retrieval*, 2015.

[21] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," in *1st Conference on Computer Simulation of Musical Creativity*, 2016.

[22] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of 21st International Conference on Music Information Retrieval*, 2020.

[23] L. N. Ferreira and J. Whitehead, "Learning to generate music with sentiment," in *Proceedings of 22st International Conference on Music Information Retrieval*, 2021.

[24] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Muller, and Y.-H. Yang, "Theme transformer: Symbolic music generation with theme-conditioned transformer," *IEEE Transactions on Multimedia*, 2022.

[25] T. Collins, R. Laney, A. Willis, and P. H. Garthwaite, "Developing and evaluating computational models of musical style," *AI EDAM*, vol. 30, no. 1, pp. 16–43, 2016.

[26] T. Collins and R. Laney, "Computer-generated stylistic compositions with long-term repetitive and phrasal structure," *Journal of Creative Music Systems*, vol. 1, no. 2, 2017.

[27] A. Eigenfeldt and P. Pasquier, "Realtime generation of harmonic progressions using controlled markov selection," in *Proceedings of ICCC-X-Computational Creativity Conference*, 2010, pp. 16–25.

[28] F. Pachet, P. Roy, and G. Barbieri, "Finite-length markov processes with constraints," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[29] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Deep learning's shallow gains: A comparative evaluation of algorithms for automatic music generation," *Machine Learning*, vol. 112, no. 5, pp. 1785–1822, 2023.

[30] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer: Generating music with long-term structure," in *International Conference on Learning Representations*, 2019.

[31] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *International Conference on Machine Learning*, 2018, pp. 4364–4373.

[32] C. Payne, "Musenet," 2019. [Online]. Available: https://openai.com/research/musenet

[33] N. Zhang, "Learning adversarial transformer for symbolic music generation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[34] J. Jiang, G. G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, "Transformer VAE: A hierarchical model for structure-aware and interpretable music representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 516–520.

[35] H. Zhang, L. Xie, and K. Qi, "Implement music generation with GAN: A systematic review," in *2021 International Conference on Computer Engineering and Application*, 2021, pp. 352–355.

[36] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proceedings of 22nd International Conference on Music Information Retrieval*, 2021.

[37] L. N. Ferreira, L. Mou, J. Whitehead, and L. H. Lelis, "Controlling perceived emotion in symbolic music generation with monte carlo tree search," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 18, no. 1, 2022, pp. 163–170.

[38] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1953–1967, 2023.

[39] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[40] L. Min, J. Jiang, G. Xia, and J. Zhao, "Polyffusion: A diffusion model for polyphonic score generation with internal and external controls," in *Proceedings of 24th International Conference on Music Information Retrieval*, 2023.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[42] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[43] Z. Wang, Y. Zhang, Y. Zhang, J. Jiang, R. Yang, J. Zhao, and G. Xia, "Pianotree VAE: Structured representation learning for polyphonic music," in *Proceedings of 21st International Conference on Music Information Retrieval*, 2020.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[45] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, 2015, pp. 2256–2265.

[46] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[47] P. Doornbusch, "Algorithmic composition: Paradigms of automated music generation," *Computer Music Journal*, vol. 34, no. 3, pp. 70–74, 2010.

[48] D. Cope, *Computer models of musical creativity*. Cambridge: MIT Press Cambridge, US, 2005.

[49] M. T. Pearce and G. A. Wiggins, "Evaluating cognitive models of musical composition," in *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 2007, pp. 73–80.

[50] S. Dai, H. Zhang, and R. B. Dannenberg, "Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music," in *Proceedings of the Joint Conference on AI Music Creativity*, 2020.

[51] A. Arzt, S. Böck, and G. Widmer, "Fast identification of piece and score position via symbolic fingerprinting." in *Proceedings of 13rd International Conference on Music Information Retrieval*, 2012, pp. 433–438.

[52] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Measuring when a music generation algorithm copies too much: The originality report, cardinality score, and symbolic fingerprinting by geometric hashing," *SN Computer Science*, vol. 3, no. 5, p. 340, 2022.

[53] T. Collins, A. Arzt, S. Flossmann, and G. Widmer, "SIARCT-CFP: Improving precision and the discovery of inexact musical patterns in point-set representations." in *Proceedings of 14th International Conference on Music Information Retrieval*, 2013, pp. 549–554.

[54] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1180–1188.

[55] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, 2018, pp. 464–468.

[56] A. Huang, M. Dinculescu, A. Vaswani, and D. Eck, "Visualizing music self-attention," in *Proceedings of NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language*, vol. 1, 2018, p. 4.

[57] R. Batlle-Roca, E. Gómez, W. Liao, X. Serra, and Y. Mitsufuji, "Transparency in music-generative AI: A systematic literature review," 2023.

[58] T. Collins, R. Laney, A. Willis, and P. H. Garthwaite, "Chopin, mazurkas and Markov: Making music in style with statistics," *Significance*, vol. 8, no. 4, pp. 154–159, 2011.

[59] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 50, pp. 269–271, 1959.

[60] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proceedings of the International Conference on Machine Learning*, 2020.

[61] A. Vyas, A. Katharopoulos, and F. Fleuret, "Fast transformers with clustered attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 665–21 674, 2020.

[62] Z. Dienes, "Using Bayes to get the most out of non-significant results," *Frontiers in psychology*, vol. 5, p. 85883, 2014.

[63] M. D. Lee and E.-J. Wagenmakers, *Bayesian cognitive modeling: A practical course*. Cambridge University Press, 2014.

[64] T. M. Amabile, *Creativity in context*. Westview Press, Boulder, Colorado, 1996.

[65] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[66] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-AI music co-creation via AI-steering tools for deep generative models," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[67] Z. Yin, "New evaluation methods for automatic music generation," Ph.D. dissertation, University of York, 2022.

[68] K. Foubert, T. Collins, and J. De Backer, "Impaired maintenance of interpersonal synchronization in musical improvisations of patients with borderline personality disorder," *Frontiers in Psychology*, vol. 8, p. 537, 2017.

[69] C. S. Sapp, "Visual hierarchical key analysis," *Computers in Entertainment*, vol. 3, no. 4, pp. 1–19, 2005.

[70] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[71] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[72] R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, H. Li, P. Gao, and Y. Qiao, "LLaMA-Adapter: Efficient fine-tuning of large language models with zero-initialized attention," in *International Conference on Learning Representations*, 2023.

[73] K. Collins *et al.*, "An introduction to the participatory and non-linear aspects of video games audio," *Essays on sound and vision*, pp. 263–298, 2007.

[74] W. Phillips, *A composer's guide to game music*. MIT Press, 2014.

[75] S. Dai, H. Yu, and R. B. Dannenberg, "What is missing in deep music generation? A study of repetition and structure in popular music," in *Proceedings of 23rd International Conference on Music Information Retrieval*, 2022.