Article:
YIN, ZONGYU, REUBEN PARIS, FEDERICO, Stepney, Susan et al. (2023) Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. Machine Learning. pp. 1785-1822. ISSN: 0885-6125

https://doi.org/10.1007/s10994-023-06309-w

Machine Learning (2023) 112:1785–1822
https://doi.org/10.1007/s10994-023-06309-w

Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation

Zongyu Yin[1] · **Federico Reuben**[2] · **Susan Stepney**[1] · **Tom Collins**[2,3]

Received: 25 June 2021 / Revised: 20 October 2022 / Accepted: 27 January 2023 /
Published online: 21 March 2023

Abstract
Deep learning methods are recognised as state-of-the-art for many applications of machine learning. Recently, deep learning methods have emerged as a solution to the task of automatic music generation (AMG) using symbolic tokens in a target style, but their superiority over non-deep learning methods has not been demonstrated. Here, we conduct a listening study to comparatively evaluate several music generation systems along six musical dimensions: stylistic success, aesthetic pleasure, repetition or self-reference, melody, harmony, and rhythm. A range of models, both deep learning algorithms and other methods, are used to generate 30-s excerpts in the style of Classical string quartets and classical piano improvisations. Fifty participants with relatively high musical knowledge rate unlabelled samples of computer-generated and human-composed excerpts for the six musical dimensions. We use non-parametric Bayesian hypothesis testing to interpret the results, allowing the possibility of finding meaningful *non*-differences between systems' performance. We find that the strongest deep learning method, a reimplemented version of Music Transformer, has equivalent performance to a non-deep learning method, MAIA Markov, demonstrating that to date, deep learning does not outperform other methods for AMG. We also find there still remains a significant gap between any algorithmic method and human-composed excerpts.

Keywords Deep learning · Non-parametric Bayesian hypothesis testing · Markov model · Music generation · Comparative evaluation · Listening study

Editor: Tijl De Bie.

✉ Zongyu Yin
zongyu.yin@outlook.com

Extended author information available on the last page of the article

1 Introduction

In the past decade, breakthroughs in artificial intelligence (AI) and deep learning have been established as such through rigorous, comparative evaluations,[1] for example, in computer vision (O'Mahony et al., 2019) and automatic speech recognition (Toshniwal et al., 2018). In the field of automatic music generation (AMG), however, to our knowledge there has been no comparative evaluation to date between deep learning and other methods (Huang et al., 2018; Yang et al., 2017; Dong et al., 2018; Hadjeres et al., 2017; Thickstun et al., 2019; Donahue et al., 2019; Tan and Herremans, 2020).[2] Rather, it appears to have been assumed that deep learning algorithms must have similarly superior performance on AMG. The contribution of this paper concerns the following two fundamental questions:

1. Is deep learning superior to other methods on the task of generating stylistically successful music?[3]
2. Are any computational methods approaching or superior to human abilities on this task?

In recent decades, several methodologies have been applied to tackle music generation tasks, and these methods can be categorised by two musical data representations: raw audio (Mehri et al., 2017; van den Oord et al., 2016) and symbolic tokens (Thickstun et al., 2019; Roberts et al., 2018; Collins et al., 2017; Huang et al., 2018). Here, we focus on symbolic methods for generating polyphonic music.[4] Depending on the underlying generation method, they can be further classified into rule-based approaches (Ebcioğlu, 1990; Bel and Kippen, 1992; Anders and Miranda, 2010; Quick and Hudak, 2013), Markovian sequential models (Cope, 1996; Allan and Williams, 2005; Eigenfeldt and Pasquier, 2010; Collins et al., 2017; Herremans and Chew, 2017), artificial neural networks (Todd, 1989; Mozer, 1994; Hild et al., 1991) and deep learning methods (Oore et al., 2018; Huang et al., 2018; Roberts et al., 2018; Thickstun et al., 2019; Dong et al., 2018). Further details are discussed in Sect. 2.1. Recent deep learning-based systems are claimed, by their authors, display state-of-the-art performance, but this is only in comparison with earlier deep learning-based systems (e.g., Huang et al., 2018; Yang et al., 2017; Dong et al., 2018; Hadjeres et al., 2017; Thickstun et al., 2019; Donahue et al., 2019; Tan and Herremans, 2020).[5] The consequence is an echo chamber, where deep learning for AMG is evaluated in isolation

[1] The term "comparative evaluation" refers to a scenario in which two or more systems are compared with respect to their performance on a well-defined task. The purpose of evaluation is to extend knowledge by improving understanding of how systems or algorithms perform under certain conditions, and whether a method is superior to human abilities on a task and/or compared to other computational approaches. Additional benefits of evaluation include determining the feasibility of methods and providing a basis for characterising and improving upon systems in future work.

[2] Examples of recent evaluations include Yang and Lerch (2020), Sturm et al. (2019) and Janssen et al. (2019) but none constitute direct comparisons between deep learning and other methods.

[3] A "stylistically successful" excerpt can be defined as one that conforms, in a listener's opinion, to the characteristics of some target style.

[4] The term polyphonic refers to music where more than one note may begin or be sustained while others begin or are sustained.

[5] The term state-of-the-art has also been used rather freely in these papers. We argue it should be reserved for describing a system whose performance is statistically significantly better than others, according to the results of a listening study where relevant existing systems encompassing diverse approaches are compared, and the description of the method—especially recruitment and musical background of participants—is detailed enough to enable replication.

from other methods, yet the corresponding papers claim state-of-the-art performance. Here we describe a comparative evaluation across a broader range of music generation algorithms, which enables us to address the question "Are deep learning methods state-of-the-art in the automatic generation of music?"

Evaluation by participants of appropriate expertise,[6] when conducted and analysed in a rigorous manner with respect to research design and statistical methods, has long been considered a strong approach to evaluating generative (music) systems (Ariza, 2009), because it has the potential to reveal the effect of musical characteristics in a system's output on human perception, and it models the way in which student stylistic compositions have been evaluated in academia for centuries (Collins et al., 2016). An alternative to evaluation by listeners is to use metrics such as cross-entropy and predictive accuracy (Huang et al., 2018; Hadjeres and Nielsen, 2020; Johnson, 2017; Thickstun et al., 2019), or distributions of automatically calculated musical features [e.g., pitch class, duration (Yang and Lerch, 2020)], and investigate how such features differ, say, between training data and system output. The automaticity and speed of evaluation by metrics are major advantages, but evaluation by metrics presupposes that the metrics are accurate proxies for the complex construct of music-stylistic success or other musical dimensions. If we knew how to define music-stylistic success as a set of metrics, it would be of great help in solving the challenge of AMG, because the objective function for the system could be obtained and it would be possible to generate music that scored highly according to that definition.

Our review of existing approaches to evaluation finds that the musical dimensions tested in listening studies often vary according to research interests, and so are inconsistent. The performance of deep learning-based systems is often evaluated with loss and accuracy, which do not reflect the stylistic success (or other musical dimensions) of algorithm output. Different evaluations' foci make comparison between models difficult. We argue that although the use of metrics is necessary, it is not sufficient for the evaluation of computer-generated music. Here we address the question "What does the generated music sound like to human listeners of an appropriate level of expertise?" In our listening study (Sect. 5), the performance of four machine learning models is assessed directly by human perception, which is represented by the rating of six musical dimensions. These musical dimensions are derived from previous analyses of classical music (Rosen, 1997): *stylistic success* and *aesthetic pleasure* (Collins et al., 2016, 2017), *repetition*, *melody*, *harmony* and *rhythm* (Hevner, 1936), defined in Sect. 5.2.1.

We apply non-parametric Bayesian hypothesis testing (van Doorn et al., 2020) to the ratings collected from the listening study, to verify hypotheses about differences in performance between systems. The Bayesian hypothesis test is a test between two mutually exclusive outcomes. It allows for the possibility of finding a statistically meaningful *non*-difference in performance between systems; in contrast, the standard frequentist hypothesis testing framework can only fail to reject a null hypothesis of no difference between systems, which is unsatisfactory because this result can also be due to an under-powered test (a more detailed explanation is given in Sect. 2.3). The conclusions that can be drawn from Bayesian hypothesis tests are also complementary and arguably preferable to just describing and displaying statistical features of systems, as provided in Yang and Lerch (2020).

[6] Appropriate expertise can be defined as being familiar with a style of music, but not knowing it so well that one can explicitly recognise excerpts as being from specific pieces.

2 Related work

In this section we review AMG algorithms (see Papadopoulos and Wiggins, 1999; Nierhaus, 2009; Fernández and Vico, 2013 for dedicated surveys). Along with the rapid development of AMG, research on evaluation frameworks has drawn increasing attention (Pearce and Wiggins, 2001, 2007; Agres et al., 2016; McCormack and Lomas, 2020; Yang and Lerch, 2020). There is often a lack of comprehensiveness and standardisation, however, leading to difficulty in comparing between systems. Therefore, we give a review of evaluation frameworks for AMG.

Also, as our work applies non-parametric Bayesian hypothesis testing (van Doorn et al., 2020) to interpret ratings from listening studies, we provide an overview of hypothesis testing in this context.

2.1 Algorithms for automatic music generation

The following review of AMG algorithms is categorised into sequential models, artificial neural networks, and their successor, deep learning approaches. Sequential models, including Markov models, are some of the earliest models, yet are still widely used (Collins et al., 2017; Allan and Williams, 2005). Before this paper, it was not known how these compared in terms of performance to deep learning approaches.

We acknowledge the existence of rule-based approaches (e.g., Hiller Jr and Isaacson, 1957; Xenakis, 1992; Ebcioğlu, 1990; Bel and Kippen, 1992; Steedman, 1984; Aguilera et al., 2010; Navarro et al., 2015), but do not review them here, for the sake of brevity and to focus on machine learning approaches.

2.1.1 Sequential models

Musical dice games (Musikalisches Würfelspiel) of the eighteenth century (Hedges, 1978) are an early example of probabilistic generation applied to Western music. The game begins with a set of prefabricated music components (e.g., notes in bars), from which a "new piece" is formed at random according to the outcome of the dice rolls. This stochastic process can be modeled by Markov models (Norris and Norris, 1998), which were defined a century later. A first-order Markov chain (the simplest type of Markov model) consists of a finite state space, a transition matrix and an initial distribution. For example, one could encode pitch classes into states and assign a transition probability (or derive it empirically from music data) to each pair of states (Collins et al., 2011). The generation process begins with a starting pitch class sampled from the initial distribution, then repeatedly generating transitions between states to obtain a "new" sequence. Ames (1989) and Collins et al. (2011) provide overviews of the application of Markov models to AMG. Conklin and Witten (1995) introduce viewpoints as a means of building a multi-dimensional Markov model, which is then optimised via prediction. Eigenfeldt and Pasquier (2010) propose a real-time system to generate harmonic progressions. This system acts as a composer assistant allowing users' input to influence the continuation selection instead of completely relying on machine selection. Allan and Williams (2005) applies hidden Markov models to chorale harmonisation, where corresponding harmony is inferred with given melody. Cope (1996, 2005) introduces Experiments in Musical Intelligence (EMI), which is a well-known program whose underlying generative mechanism appears to be that of a Markov

model (Cope, 2005, p. 89), and which is said to have generated Bach chorales, Chopin mazurkas, and Mozart operas. The lack of full source code and description of how the model works has attracted criticism and called the EMI project into question (Wiggins, 2008; Collins et al., 2016).

Widmer (2016) states that modelling music with history-based generation approaches, such as Markov models, will always be ineffective because any look-back, attention, or memory capability is inadequate with respect to music's long-term dependencies, which can span minutes and hours.[7] Collins (2011) and Collins et al. (2016, 2017) have made several contributions that comprise nesting a Markov generator in another process that inherits the medium- and long-term repetitive structure from an existing, template piece, such that it is evident—on an abstract level—in the generated output (referred to hereafter as MAIA Markov). MAIA Markov is inspired by EMI, but unlike EMI, the source code has been made available.[8] Its outputs have been the subject of multiple, rigorously conducted listening studies (Collins et al., 2016, 2017), and the starting point for use by artists in the AI Song Contest.[9]

Research by Gjerdingen (1988) on the Classical style suggests excerpts up to 4 bars in length can sound stylistically coherent without structural inheritance. When structural inheritance is required by a MAIA Markov user, it is accomplished by hard-coding a repetitive structure (e.g., reuse of bars 1–4 in bars 5–8) or running a pattern discovery algorithm such as SIARCT (Collins et al., 2013, 2010) to obtain one automatically. In the early version (Collins, 2011; Collins et al., 2016), the algorithm formalises each state as a pair consisting of (1) the beat of the bar on which a note, chord, or rest occurs, and (2) the interval size between MIDI note numbers in that set, referred as a beat-spacing state. Subsequent work (Collins et al., 2017) uses an alternative, beat-relative-MIDI state, due to superior performance: the state instead contains MIDI note numbers relative to an estimated tonal centre.

2.1.2 Artificial neural networks

Here we review methods proposed during what has been referred to as the "AI winter" of the late 1980s and early 1990s. Todd (1989) describes the first application of neural networks to music generation, exploring various symbolic representations of music, and deciding on one-hot vectors for representing musical pitches. Todd demonstrates two different model architectures. The first is based on a feed-forward neural network, taking a fixed time window of melodic information as input to predict the following window. The second comprises a sequential and recursive approach: similar to recurrent neural networks (RNNs) where the output is reentered into the input layer, the model input at each time step consists of a constant melody plan and memory of the melody so far.

The first proper RNN-based music generation model is CONCERT, introduced by Mozer (1994). Inspired by the psychological representation of musical pitch (Krumhansl, 1979), the input representation named PHCCCH is formed by combining pitch information, the chroma circle, and circle of fifths. Hild et al. (1991) introduces HARMONET, a system harmonising in the style of Bach chorales with given melodies. The system is

[7] The deep learning for music literature tends to characterise "long term" as meaning evident temporal dependencies over the course of seconds; not minutes or hours.

[8] https://www.npmjs.com/package/maia-markov.

[9] https://aisongcontest.com.

hybrid, with RNNs generating notes and a second rule-based algorithm checking for "musical rule breaks" relative to the style, such as parallel fifths.

2.1.3 Deep learning

Many deep learning generative models have been proposed recently for symbolic AMG (e.g., Sturm et al., 2015; Yang et al., 2017; Huang et al., 2018; De Boom et al., 2019; Tan and Herremans, 2020). Like many sequential models, some deep learning models represent music as a sequence of tokens, where generation is carried out by repeatedly predicting the next token based on one or more previous tokens. Models based on standard RNNs often cannot effectively learn global musical structure. As an early attempt to address this problem, Eck and Schmidhuber (2002) use long short-term memory networks (LSTMs) to learn melody and chord sequences, and the generated blues music reveals stronger global coherence in timing and structure than using standard RNNs. In addition to different architectures, a point of comparison with approaches from the AI winter is that deep learning researchers for AMG tend to favour minimal (if any) manipulation of the raw symbolic representation. For instance, whereas Mozer (1994) leverages contemporary, empirical research in music cognition that provides evidence for how music is perceived by, shapes, and subsequently interacts with human neural structures, Huang et al. (2018) and others rely on or assume the ability of neural network models to extract non-trivial (or cognitive-like) features without programming for them explicitly.

Oore et al. (2018) serialise polyphonic music and apply RNNs to generate output with expressive timing and dynamics. The serialisation converts notes into four-event sets: note-on, note-off, set-velocity and time-shift. But, as mentioned above, the output of RNN-based AMG models often lacks coherent long-term structure, much like simple Markov models. The multi-head self-attention mechanism of transformer models (Vaswani et al., 2017) shows promise in capturing long-term dependencies, and has been considered as superior to RNNs in many tasks (Devlin et al., 2019). Huang et al. (2018) use the same serialisation method of Oore et al. (2018) to adapt a transformer model to generating music, calling it Music Transformer.[10] Benefiting from the self-attention mechanism, it achieves lower validation loss compared to the RNN of Oore et al. (2018), and also longer-term stylistic consistency than previous RNN-based approaches.

Thickstun et al. (2019) describe an architecture that learns directly from a music data representation called kern (no expressive timing and dynamics). The system combines both convolutional and recurrent layers for modelling horizontal and vertical note relations, and enables multi-instrument generation for music in various classical styles. Mao et al. (2018) explore deep learning models to generate music with diverse styles conditioned by a distributed representation. In addition to sequential generation, Hadjeres et al. (2017) use pseudo-Gibbs sampling to generate music in the style of Bach chorales. Yang et al. (2017) use the piano-roll representation to treat music as images, and train generative adversarial networks (GANs) with convolutional neural networks (CNNs) to generate music. Dong et al. (2018) apply the same overall method, but the generation is performed

[10] Python dependency issues have caused multiple researchers including ourselves to be unable to reuse the published code at https://github.com/magenta/magenta. Using the published code, however, we were able to reimplement the same data preprocessing, the same model architecture with the same hyperparameters, and the same training procedure. Therefore, we continue to refer to this implementation throughout the paper as Music Transformer.

in a hierarchical manner, in order to capture the coherence between tracks and bars. Roberts et al. (2018) use variational autoencoders (VAEs) with LSTMs as both encoder and decoder; the results highlight that it can achieve better performance in capturing long-term dependencies than flat-RNNs baseline models.

2.1.4 Discussion

AMG can be achieved by various approaches, and not all existing work fits neatly into our categories. For example, Herremans and Chew (2017) present MorpheuS, which regards music generation as an optimisation problem and applies a variable neighborhood search algorithm to find solutions with the most appropriate notes.

Common criticisms of deep learning are the lack of interpretability of the learnt weights and the use of various training tricks that increase model performance in practice, and the need for large datasets to train, test, and validate the models, which does not reflect how a human composer can imitate the style of another (but not copy note sequences directly) based on just a few pieces of music. Rule-based systems have been considered expensive in design and not flexible in generating with a wide range of styles; but unlike with neural networks, one can easily trace the behavioural logic (generation decisions) of rule-based and sequential models. In contrast, according to some papers where deep learning has been used, these models can automatically and efficiently learn non-trivial features from complex data (Graves et al., 2013; He et al., 2016; Devlin et al., 2019).

When deep learning AMG models are evaluated, the comparison to other computational systems tends to focus on other deep learning models, overlooking existing *non*-deep learning approaches entirely (Huang et al., 2018; van den Oord et al., 2016; Roberts et al., 2018). We select several systems with different generating strategies to produce music excerpts for conducting the listening study (see Sect. 4.2), to fill the gap in comparative evaluation between deep learning and non-deep learning methods.

2.2 Evaluation of music generation systems

As stated in our introduction, there are multiple benefits to comparative evaluation, yet the amount of effort invested in evaluation methodology is less than that expended on the development of generative systems themselves (Pearce and Wiggins, 2001; Jordanous, 2012). While human-composed stylistic compositions have been evaluated for centuries (e.g., Cambridge, 2010), there is a lack of explicit criteria or standardised methods according to which such evaluations are conducted.

Pearce and Wiggins (2001) provide a generic evaluation framework for AMG systems, which consists of four stages: identifying the goal of a system; defining a critic from examples in the target genre; generating music samples that satisfy the critic; evaluating the generated samples with human judges. Authors' opinions differ on the importance of evaluating the creativity or general details of the generation process itself (putting the outputs to one side): Pearce and Wiggins (2001), Boden (1990), and Cohen (1999) are in favour of such conisderations, while Hofstadter (1995) argues that the internal mechanisms of a generation process can be inferred and appraised via repeated evaluation of its outputs. Pearce et al. (2002) formalise AMG into four areas or activities: algorithmic composition, the design of compositional tools, the computational modelling of musical styles, and the computational modelling of music cognition. In so doing, they attempt to make research aims more specific, and argue for applying evaluation methods appropriate to each area, thus

addressing a prevalent failure in the field of not identifying clear motivations and goals for AMG. Inspired by the work of Pearce and Wiggins (2007), we take a similar approach of combining a listening study and hypothesis testing, to compare the performance of systems according to various musical dimensions. As such, our work falls into their third category: "computational modelling of musical styles".

Regarding a generic evaluation framework for creative systems, Jordanous (2012) proposes SPECS (Standardised Procedure for Evaluating Creative Systems), which consists of a three-stage process: (1) stating what it means for a particular system to be creative, (2) deriving tests based on these statements, and (3) performing the tests. The evaluation is supported by an empirical collection of key components of creativity. A use case of this methodology is presented for a music improvisation system. Although evaluation methods may vary individually, Jordanous (2019) has also surveyed four evaluation methods and summarised five criteria for evaluation methods themselves—in effect evaluating evaluation—which are correctness, usefulness, faithfulness as a model of creativity, usability of the methodology, and generality.

Agres et al. (2016) provides a wide-ranging overview of objective evaluation methodologies for computational creativity, and their application to musical metacreation. These methods are categorised into external evaluation (e.g., Torrance, 1998; Amabile, 1982) and internal evaluation (e.g., Colton et al., 2014; Gardenfors, 2004). An evaluation is considered to be external when the source of judgements or measurements comes from outside of the system itself. Although the review highlights the importance of evaluating creative processes for a certain system, it is still unclear how to establish a standard across systems utilising different generation strategies. Agres et al. (2016) describes the advantage of questionnaire data for identifying which part of a system is not working as well as it might. A potential drawback is listeners' opinions can be altered by the very act of asking them (Schwarz, 1999).

In terms of evaluation by metrics, Yang and Lerch (2020) introduces an evaluation framework for AMG models. Characteristics of a music dataset are determined by extracting musical features (e.g., the number of unique pitch classes within a sample, the inter-onset interval in the symbolic music domain between two consecutive notes), which are then used further to conduct kernel density estimation and obtain the KL divergence between two datasets. In this way, one can investigate whether generated music demonstrates the same properties as the training data. A shortcoming of this approach is the simplicity of the features employed to date. For instance, it is possible to generate material that conforms with respect to basic pitch and rhythmic features, but still falls short of higher-level music-theoretic concepts that have been identified by musicologists (Rosen, 1997; Gjerdingen, 2007), and that may be perceived, implicitly or explicitly, by judges in listening studies.

The features-based evaluation approach (Yang and Lerch, 2020) is relatively recent, whereas most deep learning AMG papers (e.g., Lim et al., 2017; Oore et al., 2018; Huang et al., 2018) evaluate generation performance and compare baselines using training/validation loss. For example, models that repeatedly predict the next tokens (e.g., musical events) are usually based on the concept of a language model in natural language processing. Negative log likelihood/cross entropy loss are used to help models find the statistically correct tokens, and certain advanced sampling processes can be used to adjust the randomness of generated sequences. But for generated music, loss value does not necessarily equate to strong performance along one musical dimension or another, only to the likelihood of being a valid sequence. The criterion used by VAE-based models (Roberts et al., 2018) is a weighted sum of cross entropy and Kullback–Leibler (KL) divergence, where lower KL

divergence weight increases the statistical accuracy of prediction but narrows the diversity of generation. GAN-based models (Yang et al., 2017; Dong et al., 2018) optimise the generator and discriminator via a minimax tradeoff in which the discriminator evaluates generated output according to the probability of it being indistinguishable from training items. With respect to evaluation by human participants, some AMG papers do not include a formal listening study at all (Johnson, 2017; Hadjeres and Nielsen, 2020), while one invites composers to comment on the generated music (Oore et al., 2018). Other AMG papers do include listening studies (Liang, 2006; Lim et al., 2017; Hadjeres et al., 2017; Mao et al., 2018; Roberts et al., 2018; Huang et al., 2018), but there is no standardised approach and sometimes no information is provided as to participant recruitment or musical backgrounds, and the constructs (e.g., "stylistic success") that are operationalised (turned into working definitions and written into questions) also vary widely. For instance, a paper on a harmonisation system may focus solely on whether "correct" chord identities are generated by the system, but omit a question regarding overall stylistic success.

2.3 Bayes factor analyses

A commonly observed characteristic of science, and one that is used to distinguish it from pseudoscience, is that scientific theories consist of falsifiable statements, which—in spite of attempts to falsify them via experiments—appear to stand the test of time (Goodwin and Goodwin, 2016; Popper, 2005, 2014).

Inferential statistics (e.g., *t*-tests, analysis of variance or ANOVA) is the branch of statistics via which conclusions may be inferred from observed data (Aron et al., 2013). Contrast this with descriptive statistics (e.g., mean, variance), which merely enable one to describe datasets. As such, inferential statistics are often used in pursuit of scientific progress, because the inferences made from data observed during experiments can be used to bolster or undermine particular theories.

The two main hypothesis testing frameworks in inferential statistics are called frequentist and Bayesian. A typical frequentist hypothesis test proceeds by stating a null hypothesis H_0 (e.g., no difference in performance between Systems A and B) and an alternative hypothesis H_1 (e.g., a difference in performance between Systems A and B), and determining a cutoff score on some comparison distribution such that the probability of observing this score in an experiment is less than a so-called, arbitrary p-value of .05. When the experiment has been conducted and a sample's score obtained, the experimenter either fails to reject H_0, or finds significant evidence for rejecting H_0 in favour of H_1.

A weakness of this approach is that one can never accept the null hypothesis, as the power of the test (stemming from the experimental design) could be insufficient to find a significant difference that may in truth be present.[11] Failure to reject H_0 is typically associated with a non-result, which is also regarded as a non-publishable result (Aron et al., 2013).

Returning to the example of the performance levels of Systems A and B, in a frequentist approach we can never infer that there is no difference in performance; only fail to reject that there is no difference in performance. Suppose System A was a model for AMG with promising outputs, and System B was original, human-composed music in a target style. Using a frequentist approach, we could never be satisfied (or publish) that System A

[11] Power analyses are used to address this issue, but comprise assumptions that are themselves problematic.

had attained the impressive level of System B, because we cannot infer no difference in performance.

A typical Bayesian hypothesis test also proceeds by stating H_0 and H_1, but, unlike a frequentist approach, *can* provide evidence in support of either null or alternative hypotheses. The fundamental shortcoming of the frequentist approach—of not being able to find in favour of the null hypothesis—is resolved in Bayesian hypothesis testing. Therefore, Bayesian hypothesis testing is better suited for experiments where systems are being compared to one another, and more generally to pursuing a scientific approach where theories consist of falsifiable statements (Dienes, 2014; van Doorn et al., 2020).

The lack of a straightforward likelihood function has made calculation of such tests difficult in the past, but Rouder et al. (2009) began to provide solutions for various common scenarios, at least where assumptions can be made about the normality of the underlying distributions (so-called parametric tests).

The counterpart of a p value in the Bayesian approach is the Bayes factor BF_{10}, which is a likelihood ratio of the marginal likelihood of H_0 and H_1 (Dickey and Lientz, 1970; Wagenmakers et al., 2010), given by the equation

$$BF_{10} = \frac{Pr(\theta_0|H_0)}{Pr(\theta_0|data, H_1)} \tag{1}$$

where θ_0 denotes the parameter of interest. The value of BF_{10} can be interpreted as stated in Table 1 (Lee and Wagenmakers, 2014).

Previous work (including by one of the current authors) tends to treat numeric data such as stylistic success ratings as though they are ratio-scale or at least equal-interval (Pearce and Wiggins, 2007; Collins et al., 2017), but in reality such assumptions are invalid. The data can be assumed only to be ordinal, and therefore a non-parametric Bayes factor test is required. van Doorn et al. (2020) introduce a way to calculate this test by data augmentation with Gibbs sampling, which lets the sample follow a latent normal distribution adapted to the non-parametric scenario. They also provide source code for demonstrating the Bayesian counterparts of three non-parametric frequentist tests: the rank-sum test, the signed rank test, and Spearman's ρ. We use van Doorn et al.'s (2020) method, as it fits our circumstances, with accessible code to conduct the following hypothesis tests.

To our knowledge, this is the first use of such tests in evaluating the performance of machine learning systems. Ideally, the introduction to, and examples of using, the tests that we provide here will lead to wider uptake in the machine learning literature, because the Bayesian approach is preferable to the frequentist one, and van Doorn et al. (2020) offers tests appropriate for use with non-parametric data.

3 Hypotheses

The listening study discussed in Sect. 5.4 consists of two parts differentiated by the target styles of the music therein: Classical string quartet (CSQ) and classical piano improvisation (CPI).[12] Each stimulus is rated according to six musical dimensions, defined in Sect. 5.2.1. Here we provide a list of hypotheses with respect to the performance of various systems for a subset of the styles and musical dimensions. Some of these hypotheses are

[12] The difference between Classical and classical is meaningful and is defined in Sect. 4.1.

theoretically motivated, stemming from our understanding of the music-psychological literature and mechanisms behind the models used to generate stimuli for the listening study (see Sect. 4.2), while others became apparent as we prepared the stimuli for the study.[13]

3.1 System-focused hypotheses

1. *In the CSQ part of study, there will be no significant difference between the stylistic success ratings for MAIA Markov and Music Transformer.*

Both systems learn and generate music with the symbolic representation in a recurrent manner, that is statistically modelling the likelihood of the next state/event by the previous. Based on the generated results, we believe they share a similar strength at generating accurate notes locally. And we assume the local accuracy serves as an important factor of stylistic success when only symbolic representation is considered.

2. *In the CSQ part of the study, MAIA Markov will outperform other models on the repetitive structure rating.*

MAIA Markov forces the generated music to inherit structural patterns. It is supposed to make the repetitive patterns more significant than other systems. Although the multi-headed attention mechanism used in Music Transformer shows efficiency and accuracy in capturing long-term dependencies, we deem the direct indication of repetitive structure to be more manifest than learning and generating a repetitive structure by chance.

3. *In the CSQ part of the study, coupled recurrent model will get lower stylistic success ratings than MAIA Markov and Music Transformer.*

Through our listening experience, the generated excerpts from coupled recurrent model show less variety in pitch range and durations when compared to other models. Thus we doubt it will meet expectations of Classical era.

3.2 Musical dimension-focused hypotheses

4. *Ratings of melodic success will be a driver/predictor of overall aesthetic pleasure.*

Melody, harmony and rhythm are commonly considered as the most fundamental elements when analysing compositions. We aim to find out to what extent those elements are correlated with aesthetic pleasure. Intuitively, we think melody will correlate most strongly, because it is the most memorable and representative element in these styles (Rosen, 1997).

5. *Stylistic success ratings in most cases will be lower than aesthetic pleasure.*

[13] We pre-registered our hypotheses on the Open Science Framework so that we did not fall into the trap of devising them after seeing the listening study results. This trap, called post-hoc hypothesising, runs contrary to a scientific experimental method.

Table 1 Bayes factor interpretation

BF_{10}	Interpreting
> 100	Extreme evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence
1–0.333	Anecdotal evidence for H_0
0.333–0.1	Moderate evidence for H_0
0.1–0.033	Strong evidence for H_0
0.033–0.01	Very strong evidence for H_0
< 0.01	Extreme evidence for H_0

Through listening to the generated excerpts, a short pleasant phrase can be often found locally, but it rarely persists within a coherent stylistic structure. Thus, we aim to verify this phenomenon with participants' ratings.

6. *Music Transformer will outperform other models on melody ratings in both CSQ and CPI parts.*

The weakness of generated music often stems from unexpected (negatively, instead of creatively) notes. We think this phenomenon largely affects melodic aspects. With the prominent performance of multi-headed attention mechanisms, we believe that Music Transformer learns melodies more accurately than other systems.

3.3 Dataset- and participant-focused hypotheses

7. *Among systems taking part in both studies (including the Orig category), the stylistic success ratings received for CSQ will be higher than those for CPI.*

To our best understanding of CSQ and CPI, excerpts of CSQ are less stylistically diverse than those in CPI. Without models being specially adjusted to either style, excerpts generated in CSQ are more likely to be recognised by participants. Thus, they are supposed to receive higher ratings of stylistic success.

8. *Music Transformer will outperform MusicVAE on stylistic success ratings in the CPI part of study but not in the CSQ part.*

As mentioned above, we consider CSQ to be less stylistically diverse, and therefore statistically, the complexity of its dataset is lower than CPI's. With a similar model size and the assumption that Music Transformer is more powerful than MusicVAE, they should have the same performance in the CSQ part of study, but Music Transformer will show better adaptability with the CPI dataset.

9. *Participants with fewer years of musical training tend to give higher ratings.*

Bigand and Poulin-Charronnat (2006) discuss how some music-perceptual abilities result purely from exposure to (experience of) music, whereas others require explicit training. Their study does not include judgement of stylistic success, but we extrapolate to hypothesise that the less experienced our listeners, the less likely they are to notice unidiomatic chords or the absence of appropriate musical schemata (Gjerdingen, 1988, 2007), and so the higher on average their ratings of stylistic success will be compared to more experienced listeners.

4 Systems under test

In this section, we describe two datasets and four AMG systems we (re)implemented to prepare the stimuli or excerpts indicated in Table 2. Supporting materials including stimuli, datasets and code for replicating our outputs can be accessed via https://osf.io/96emr/.

4.1 Datasets

So that our results are not limited to a single musical style, we make use of two separate datasets: Classical string quartets and classical piano improvisation. As previously mentioned, while some AMG research operates in the audio domain, the focus here is on the symbolic domain. The Musical Instrument Digital Interface (MIDI) format is commonly used to describe musical symbolic data, from which features can be extracted and processed, such as attributes of musical notes: ontime, MIDI note number, duration, and velocity,[14] MIDI data may or may not comprise expressive timing and dynamics—it tends to depend on whether or not the data were captured by a human performer playing on a MIDI-enabled instrument. For example, the MAESTRO dataset (Hawthorne et al., 2019) consists of MIDI files collected from virtuosic piano performances, including multiple versions of the same piece with different expressive timing and dynamics. Humdrum (see KernScores[15] for example data) and MusicXML are other widely-used digital musical notations, which encode the original musical scores (sheet music). Timing and dynamic expressivity in these files can be absent entirely or limited to what the composer or editor marked on the score.

4.1.1 Classical string quartet

A Classical string quartet (CSQ) is a piece written for two violins, viola, and cello, following the style of Western art music written in the period 1750–1830. This was a period during which composers Joseph Haydn, Wolfgang Amadeus Mozart, and Ludwig van Beethoven were active, and this music is often regarded as being of an "orderly nature, with qualities of clarity and balance, and emphasising formal beauty rather than emotional

[14] Ontime is the start time of a note, counting in crotchet beats and taking 0 to be bar 1 beat 1 (Collins, 2011, p. 347). MIDI note number is a numeric encoding of pitch, with C4 or "middle C" being MIDI note number 60, C♯4 or D♭4 being 61, etc. Duration is the length of a note, again measured in crotchet beats. Velocity is the term used for the relative loudness of a note, with common ranges being [0, 127] or [0, 1].
[15] http://kern.ccarh.org/.

expression (which is not to say that emotion is lacking)" (Kennedy and Bourne, 2004). Our dataset contains 71 Classical string quartet movements, having a total of 228,021 notes, without expressive timing and dynamics in MIDI format from KernScores. The dataset was formed according to the following filters and constraints:

- string quartet composed by Haydn, Mozart, or Beethoven;
- first movement;
- fast tempo, e.g., Moderato, Allegretto, Allegro, Vivace, or Presto.

4.1.2 Classical piano improvisation

A classical piano improvisation is a piece that usually would be created and played in the moment on the piano by someone who is able to draw on material from "classical" music to varying degrees of abstraction. We use the small "c" for "classical" which subsumes the "Classical" period 1750–1830, and in this study "classical" is taken to mean Western art music, following the conventions of music written in the period 1650–1900, i.e., from the time of J. S. Bach to Brahms. We use the MAESTRO dataset (Hawthorne et al., 2019) for this part of study. It contains 1,276 virtuoso piano performances, with a total of 7.04 million notes, as MIDI data recorded with Yamaha Disklaviers. Unlike the MIDI data collected in KernScores, it contains expressive timing and dynamics.

We define this style based on the contents of the MAESTRO dataset because two of the deep learning models we want to investigate, MusicVAE (standing for variational autoencoders) (Roberts et al., 2018) and Music Transformer (Huang et al., 2018), are introduced in papers that also use this dataset. To give these models the best chance of performing well in the listening study, we include the same dataset as used in the original study (Huang et al., 2018).

4.2 (Re)implementation of models

MAIA Markov (Collins et al., 2017) is the non-deep learning model included in this study. In its current form, this model can only be applied to non-expressive timing data, so it features in the CSQ but not the CPI part of the study. The model structure and state space is described in Sect. 2; we add here that the initial distribution is calculated by extracting estimated phrase beginnings and endings from the input data. Sometimes phrase beginnings and endings are encoded in kern files—transferred from phrase marks in the original score on which the encoding is based, but for the CSQs they are not, so we use rests of longer than one beat to identify phrase beginnings and endings. We use two repetitive structures that are common in Classical music: one where bars 1–4 are repeated in bars 5–8, and a second where additionally bars 1–2 occur again in bars 11–12. Thus, we assume MAIA Markov will perform well on ratings of repetitive structure.

A second model we include in the study, and the first example of a deep learning model, is the coupled recurrent model (Thickstun et al., 2019), which uses an RNN-based hierarchical architecture to achieve polyphonic music generation. It consists of multiple submodels for voices, and generates tokens in steps based on a certain historical window. There is a global state governing all voice generations, so that a generated note in a single voice is not only conditioned by its voice history but also the global state coordinating across voices. As such, the coupled recurrent model should perform well on ratings of

Table 2 Categories of stimuli contained in the CSQ and CPI parts of study

System name (abbreviation)	Classical string quartet (CSQ)	Classical piano improvisation (CPI)
Original (Orig)	25 (4)	25 (4)
Before–After (BeAf)	25 (4)	–
MAIA Markov (MaMa) (Collins et al., 2017, 2016)	25 (4)	–
Coupled Recurrent Model (CoRe) (Thickstun et al., 2019)	25 (4)	–
MusicVAE (MVAE) (Roberts et al., 2018)	25 (4)	25 (4)
Music Transformer (MuTr) (Huang et al., 2018)	25 (4)	25 (4)
Listen to Transformer (LiTr) (Huang et al., 2018)	–	25 (4)
Total	150 (24)	100 (16)

The first value indicates the number of stimuli generated per category; the second value, in parentheses, indicates the number of stimuli selected from the pool per category for each participant

harmony. We use the published scripts (multipart6 as reported as the best version, with the total loss of 12.87) provided with the original paper to generate CSQs. Again, this model is not applicable to expressive data, so it is not included in the CPI part of the study.

The third model we include in the study is MusicVAE (Roberts et al., 2018). The assumption behind VAEs is that each data sample can be represented by a latent code with smaller dimensional size, and all those latent codes lie in a predefined distribution (usually a normal distribution). During the training, the decoder is required to reconstruct the input data by using the latent code sampled from the distribution. On this basis, MusicVAE applies multi-layer bi-directional LSTMs as encoder/decoder to compress/reconstruct sequential musical tokens. In addition, a conductor [as presented by the original paper (Roberts et al., 2018)] is placed between encoder and decoder to prevent posterior collapse, where the decoder breaks free of dependence on the encoder. Regarding the generation, the conductor firstly takes the randomly sampled latent code and produces sub-codes for lower-level slices (e.g., bars/measures). Musical tokens are then recursively generated by decoders with each corresponding sub-code. In this listening study, we reimplemented MusicVAE and generated excerpts for both CSQ and CPI parts of study. The model consists of two encoder/conductor/decoder layers with a hidden size of 1024, it was trained with scheduled sampling (Bengio et al., 2015) and the Adam optimiser (Kingma and Ba, 2014). The loss for training MusicVAE is a weighted sum of cross entropy and KL divergence. The training starts with 20 warm-up epochs, in which the KL divergence loss is multiplied by the weight that increases with the rate of 0.01 for each epoch. The trained model is then obtained by early stopping, with the best validation loss of 2.455 for CPI and 1.42 for CSQ.

Music Transformer (Huang et al., 2018) is said to benefit from the self-attention mechanism, and therefore said to generate reasonably coherent polyphonic material on a short-term basis (several bars of music). The model is configured with six layers, eight heads, a model dimension of 512, and sequence length of 2048. It was also trained with scheduled sampling (Bengio et al., 2015) and the Adam (Kingma and Ba, 2014) optimiser, using cross entropy loss with early stopping. As mentioned in Sect. 2.1.3, we reimplement Music Transformer, achieving the validation loss of 2.2 for CPI and 1.184 for CSQ, and used it to generate excerpts for both the CSQ and CPI versions of the task. The original authors

(Huang et al., 2018) also provide generated results, but these were trained on over 10,000 h of piano recordings from YouTube, transcribed using Onsets and Frames (Hawthorne et al., 2018), which is not the same dataset as used in the original paper. So while a direct comparison between versions is not possible due to the different training sets involved, we randomly select some excerpts from "Listen to Transformer" and include them in the listening study as an additional category.[16]

5 Method

5.1 Participants

For this study, we aim for the majority of participants to have a relatively high level of musical knowledge. While it is possible to achieve ostensibly impressive results with novice users or listeners (Louie et al., 2020), it does not benefit the broader purpose of measuring true progress (or lack thereof) in the field. One of the reasons we recruit relatively expert participants (e.g., music undergraduates) is because they are taught to focus on the relationship between note content and stylistic success, and less on the expressivity or inexpressivity of a particular performance.[17]

A total of 50 participants were recruited, from email lists including music undergraduate students at the University of York, the Magenta Google Group, and the Society for Music Theory. Participants were compensated £10 for an hour of their time. After excluding all data from participants who went through the study so quickly it would have been impossible for them to listen to their assigned excerpts in entirety, 41 participants' submissions are used as the basis for the analysis in Sect. 6.

Participants' age (lower quartile = 20, median = 26, upper quartile = 37) against years of musical training (lower quartile = 9, median = 13, upper quartile = 18) is shown in Fig. 1, and Table 3 shows participants' frequency of playing/singing (median = "Once a day") and attending concerts (median = "Once a month").

5.2 Experimental design

Here we clarify the design decisions made to prevent possible bias in favour of or against particular systems. The selection of stimuli/excerpts in this listening listening is performed based on the principles of balanced incomplete block design (BIBD) (Street and Street, 1986). We aim to have each stimulus from the various systems be rated by sufficiently many participants, so that overall the results are unlikely to be biased towards arbitrary excerpts. Table 2 shows that there are 25 excerpts per system, giving a total of 250 excerpts in total across the CSQ and CPI versions of the listening task. Since it would be unreasonable, and lead to listener fatigue, to ask each participant to rate all of these excerpts, we sample from this excerpt pool to present a subset of excerpts for each participant,

[16] https://magenta.github.io/listen-to-transformer.

[17] For instance, the class descriptor (rubric) for a stylistic composition in the Music Department at University of York mentions "a high degree of immersion in the style...strong musical ideas, set out and developed idiomatically for the performance forces", all of which points to note content rather than how something is performed.

balanced with respect to blocks/categories. A naive sampling process may result in some excerpts selected many more times than others. To mitigate this, we maintain a count of how many times each excerpt is presented as the study runs. Each participant is presented with 40 stimuli (four excerpts from each of the ten categories across both parts of study). We remove an excerpt from the pool when it has been presented to eight participants, and randomly select four from the remaining excerpts per category with each new participant. As reported above, we remove the data of participants who rush through the study, but still the above steps help ensure a strong coverage of ratings for the excerpts from each system.

To prevent order effects, after an introduction page, participants are first redirected at random to either the CSQ or CPI version of the listening task, followed by the complementary version. Within task version, excerpts are shuffled to prevent ordering effects or associating particular groups of stimuli with particular systems.

Table 2 shows the four computational models introduced in Sect. 4.2, as well as "Listen to Transformer". From a design point of view, so that we can compare computational performance with human composers' capabilities, each task version also includes an *Orig* category consisting of excerpts of human-composed music in the respective target style, the contents of which are detailed further in Sect. 5.3. For CSQ, we also make use of some excerpts of string music composed during eras either side of the Classical period, so there is an additional *BeAf* category in this version of the task. This could enable us to shed some light on whether computer-generated excerpts are at least comparable in terms of stylistic success with out-of-period human-composed works (i.e., the computer models lack specificity of period but they do generate generally high-quality music), or if they fall short here also.

5.2.1 Definition of musical dimensions

Each excerpt is rated according to the following six musical dimensions, which are based on music theory (Rosen, 1997) and previous work (Pearce and Wiggins, 2007; Collins et al., 2016):

Stylistic success (Ss): A stylistically successful excerpt can be defined as one that conforms, in a participant's opinion, to the characteristics of the definitions of CSQ and CPI and example excerpts, which were given on the introductory page of the study. (The example excerpts remained accessible throughout the study.)
Aesthetic pleasure (Ap) The extent to which someone finds beauty in something (De Clercq, 2019). The concepts of "stylistic success" and "aesthetic pleasure" are independent in the sense that two people may agree on certain characteristics of a music excerpt that make it stylistically successful, whereas they may disagree on the extent to which they personally like it.
Repetition or self-reference (Re) The reuse, in exact or inexact form, of musical material (e.g., notes, melody, harmony, rhythm) within a piece.
Melody (Me) A succession of notes, varying in pitch, which have an organised and recognisable shape. Melody is horizontal, i.e. the notes are heard consecutively.
Harmony (Ha) The simultaneous sounding of two or more notes; in this sense synonymous with chords. The organisation and arrangement of chords and their relationships to one another, vertically (at the same time) and horizontally (across time) over the course of a piece.

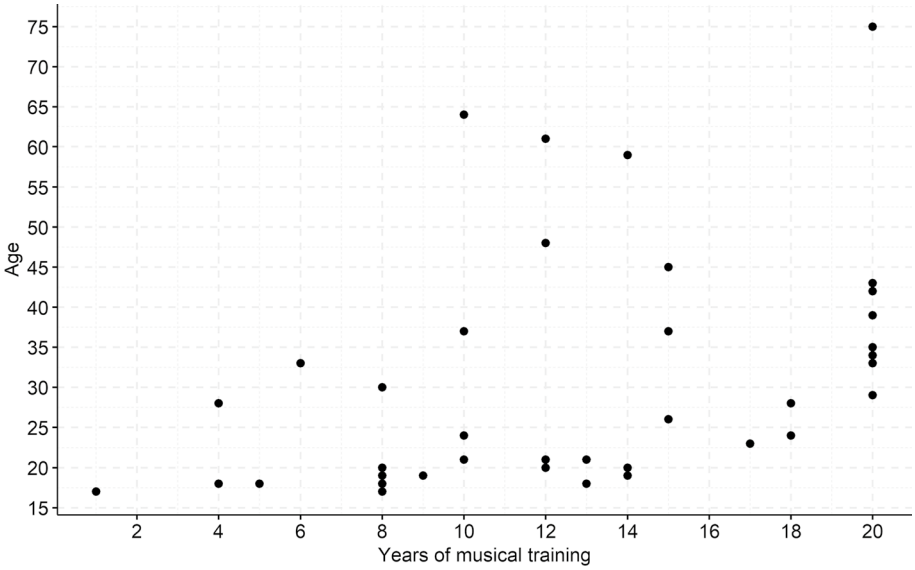


Fig. 1 Participants' age against years of music training

Table 3 Participants' frequency of playing/singing and attending concerts

	Freq. of playing/singing	Freq. of attending concerts
Once a day	25	–
Once a week	8	4
Once every 2 weeks	3	7
Once a month	3	15
Once every 3 months	1	6
Once every 6 months	0	3
Once a year	0	2
Less than once a year	1	3
Never	–	1

Rhythm (Rh) Everything pertaining to the time aspect of music (as distinct from the aspect of pitch), including event or note beginnings and endings, beats, accents, measures, and groupings of various kinds.

5.3 Stimuli

Here we describe the formation of stimuli (music excerpts) for the listening study. As mentioned at the beginning of Sect. 4, all study materials, including the stimuli and interface used, are available via an Open Science Framework repository, which can be consulted for confirming or supplementing the details provided below. We restrict all excerpts to a duration of 20–30 s, aiming to balance the length of the whole procedure while still providing

a good number of excerpts to be tested. For the sake of comparison, participants hear all excerpts played back with high-quality piano sound samples. For CSQ excerpts, which did not have expressive timing or dynamics, the velocity of each note is set to 0.8, in a range [0, 1] that modulates gain on an individual sample; for CPI excerpts, the expressive timing is retained and velocity was normalised using the minimum and maximum values found in the whole piece. The sounds for stimuli are generated in the moment using Tone.js (https://tonejs.github.io/). No effects are added, and no further normalisation is applied to the sampled audio files. The above decisions are in keeping with the playback functionality of desktop music notation software such as MuseScore. This is the key point here: the stimuli are ecologically valid in that they have a sound quality that our participants are familiar with from their use of music software.

There are 150 stimuli in the pool for the CSQ part of study: 25 *Orig* excerpts from seven compositions by Haydn, Mozart, and Beethoven; 25 *BeAf* excerpts from five compositions by Vivaldi (before the Classical era) and Brahms (after the Classical era); 25 generated excerpts for each of the four systems under test; 25 from "Listen to Transformer". There are 100 stimuli in the CPI part of study: 25 original excerpts from five piano improvisations in the classical style[18]; 25 generated excerpts for each of the two systems under test; 25 from "Listen to Transformer".

In terms of human-composed excerpts (the categories *Orig* and *BeAf*), as the duration of a movement (contained in a MIDI file) is in most cases longer than 3 min, we are able to obtain three or four highly contrasting (independent if heard in isolation) excerpts by manually clipping the MIDI files. Thus, the extracted excerpts do not overlap one another; nor do they start or end in the middle of a note. This reduces the size of our training set from 71 to 64 items, but we are also able to use these seven compositions from the CSQ *Orig* category as the validation set during model training. The seven are selected pseudo-randomly to give a balanced representation of the three composers (Haydn, Mozart, and Beethoven) and also of key signatures (mostly major keys).

When running the generation models described in Sect. 4.2, we noticed some systems could generate poor-quality output containing one or more of (a) an incessantly repeated note or notes, (b) unusually long rests, (c) a chunk of noticeable copying from the training set. Thus, we applied some automated filters to address (a), (b), and (c), as otherwise we would have been wasting participants' time with these excerpts. The filters exclude generated excerpts with a large number of repeated notes or long rests, and we then apply the originality report method (Yin et al., 2021) to measure and exclude excerpts that copy too much input in their output. The effect of these filters is non-negligible, with empirical probability of removal in the range 0.05–0.1, depending on filter and model.

As such, no hand- or cherry-picking of computer-generated output occurs when preparing the stimuli for our listening study. We did not generate outputs for *LiTr* (merely reused existing), but the researchers behind this model do explicitly state that the outputs from which we sample are "random, not cherry-picked samples".[19]

[18] Collected from https://www.youtube.com/c/KyleLandryPianoTutorials/channels.

[19] https://magenta.github.io/listen-to-transformer/.

5.4 Procedure

During the listening study, participants go through a web-based questionnaire, listening to the prepared excerpts (see Sect. 5.3) and rating the corresponding musical dimensions using sliders (see Sect. 5.2). Participants are not told which model generated the excerpt, but they do know that some excerpts are model-generated and some are composed by humans. They participate remotely, listening to the excerpts on their own machines with their own headphones or speakers. We asked participants to declare any vision or hearing problems that may impact their responses, but no such issues were reported.

The full procedure consists of an introductory page explaining the task and defining and giving examples of the CSQ and CPI musical styles, followed by (in random order) the CSQ and CPI versions of the tasks themselves, and finishing with a final thank-you page where the payment information is also issued.

For each excerpt, participants are presented with the following instructions on a web form:

1. Rate the following musical dimensions on a scale 1–7 (low–high):
   - stylistic success
   - aesthetic pleasure
   - repetition
   - melody
   - harmony
   - rhythm
2. Indicate time windows or instants, if a very low or high rating has been given to this excerpt and it was due to particular moments rather than an overall impression (optional)
3. A text box for any comments about the excerpt (optional)

6 Results

This section describes analyses of ratings obtained from the listening study and presents and discusses results of the Bayesian hypothesis tests. We use abbreviations for system names and musical dimension rating categories as stated in Table 2 and Sect. 5.2.1.

6.1 Overview

As a consequence of the number of usable datasets provided by participants and the categories and stimulus numbers (Table 2), there are a total of 1,640 sets of ratings, ranging from one to seven. To provide a general overview of the distribution of these ratings, Fig. 2 shows violin plots for all six musical dimensions, for each system and style.

Before addressing the formal hypotheses from Sect. 2.3, we make some informal observations about the results in Fig. 2. Ratings of *Orig* for both CSQ and CPI are higher for all dimensions when compared to generative systems, and *BeAf* has slightly lower ratings overall compared to *Orig*, but still higher than the generative systems. Comparing CSQ and CPI, excerpts generated in CSQ style appear to have higher ratings than those in CPI style.

Particularly for CPI, *LiTr* outperforms *MuTr* in all six musical aspects. Excerpts from both categories were generated by the Transformer Model, so such a large gap in performance raises some concerns as to why, which we will return to in Sect. 7.

6.2 Hypothesis testing

To obtain a formal perspective on ratings and system performance, we conducted the hypothesis tests as introduced in Sect. 2.3, for each hypothesis listed in Sect. 3. Maintaining the same numbering of hypotheses, the following list indicates the outcome of each test, restates the null hypothesis H_0 and alternative hypothesis H_1, and also gives the computed Bayes factor BF_{10} and corresponding interpretation and degree of evidence according to Table 1. It should be noted that computing the Bayes factor with Gibbs sampling can lead to unstable estimates of extreme values, which means the variance of the Bayes factor becomes larger when the degree of evidence is more extreme (Lodewyckx et al., 2011). However, the large variance can be tolerated as it is unlikely to influence the overall interpretation.

6.2.1 System-focused hypotheses

1. *In the CSQ part of study, there will be no significant difference between the stylistic success ratings for MAIA Markov and Music Transformer.*

- Outcome: **There is no difference between *MaMa* and *MuTr* on *Ss* ratings**.
- $\mathbf{H_0}$: In CSQ, there is no difference between *MaMa* and *MuTr* on *Ss* ratings.
- $\mathbf{H_1}$: In CSQ, one of *MaMa* or *MuTr* outperforms the other on *Ss* ratings.
- $\mathbf{BF_{10}}$: 0.18 (moderate evidence for H_0).

2. *In the CSQ part of the study, MAIA Markov will outperform other models on the repetitive structure rating.*

- Outcome: **There is no difference between *MaMa* and *MuTr* on *Re* ratings; *MaMa* outperforms both *MVAE* and *CoRe* on *Re* ratings**.
- $\mathbf{H_0}$: In CSQ, there is no difference between *MaMa* and (a) *MuTr*, (b) *MVAE*, (c) *CoRe* on *Re* ratings.
- $\mathbf{H_1}$: In CSQ, *MaMa* outperforms (a) *MuTr*, (b) *MVAE*, (c) *CoRe* on *Re* ratings.
- $\mathbf{BF_{10}}$: (a) 0.11 (moderate evidence for H_0); (b) 100.58 (very strong evidence for H_1); (c) 134.16 (extreme evidence for H_1).

3. *In the CSQ part of the study, coupled recurrent model will get lower stylistic success ratings than MAIA Markov and Music Transformer.*

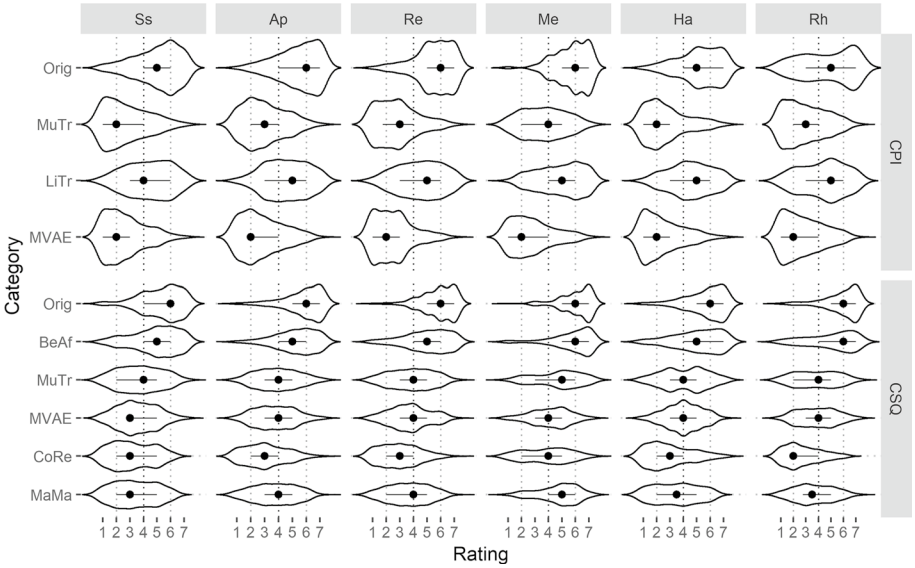


Fig. 2 Rating (1–7) distributions of six musical dimensions: stylistic success (*Ss*), aesthetic pleasure (*Ap*), repetition (*Re*), melody (*Me*), harmony (*Ha*) and rhythm (*Rh*), for two styles Classical string quartets (CPI) and classical piano improvisations (CSQ) across different categories (mostly algorithms—see Table 2 for details). These show violin plots: the envelope shows the distribution of responses; the vertical lines are for reference to the rating scales; the horizontal line goes from lower quartile, through median (dot), to upper quartile. (We do not use mean, variance, etc., as the ratings can only be considered ordinal values)

- Outcome: ***MaMa* and *MuTr* outperform *CoRe* on *Ss* ratings**.
- $\mathbf{H_0}$: In CSQ, *CoRe* shares the same performance with (a) *MuTr* and (b) *MaMa* on *Ss* ratings.
- $\mathbf{H_1}$: In CSQ, (a) *MuTr* and (b) *MaMa* outperforms *CoRe* on *Ss* ratings.
- $\mathbf{BF_{10}}$: (a) 2033.63; (b) 5492.46 (both extreme evidence for H_1).

6.2.2 Musical dimension-focused hypotheses

4. *Ratings of melodic success will be a driver/predictor of overall aesthetic pleasure.*

- Outcome: ***Me* ratings are positively correlated with *Ap* ratings**.
- $\mathbf{H_0}$: *Me* ratings are not positively correlated with *Ap* ratings.
- $\mathbf{H_1}$: *Me* ratings are positively correlated with *Ap* ratings.
- $\mathbf{BF_{10}}$: $1.43e^{284}$ (extreme evidence for H_1).

5. *Stylistic success ratings in most cases will be lower than aesthetic pleasure.*

- Outcome: **Systems perform equally well on *Ss* and *Ap* ratings**. (Ratings were merged across categories for this test.)
- $\mathbf{H_0}$: Systems perform equally well on *Ap* and *Ss* ratings.
- $\mathbf{H_1}$: Systems performs better on *Ap* than *Ss* ratings.
- $\mathbf{BF_{10}}$: 0.016 (very strong evidence for H_0).

6. *Music Transformer will outperform other models on melody ratings in both CSQ and CPI parts.*

- Outcome: ***MuTr* receives similar *Me* ratings as *MVAE* and *MaMa* for both CSQ and CPI parts of the study.**
- $\mathbf{H_0}$: *MuTr* receives similar *Me* ratings as (a) *MVAE* in CSQ, (b) *MaMa* in CSQ and (c) *MVAE* in CPI.
- $\mathbf{H_1}$: On *Me* ratings, *MuTr* outperforms all other systems in both parts of study.
- $\mathbf{BF_{10}}$: (a) 0.213 (moderate evidence for H_0); (b) 0.433 (anecdotal evidence for H_0); (c) 0.554 (anecdotal evidence for H_0).

6.2.3 Dataset- and participant-focused hypotheses

7. *Among systems taking part in both studies (including the Orig category), the stylistic success ratings received for CSQ will be higher than those for CPI.*

- Outcome: **For systems involved in both CSQ and CSI parts of the study, *Ss* ratings are generally higher for CSQ.**
- $\mathbf{H_0}$: (a) *Orig* (b) *MuTr* (c) *MVAE* rate equally well on *Ss* for CSQ and CPI.
- $\mathbf{H_1}$: (a) *Orig* (b) *MuTr* (c) *MVAE* rate higher on *Ss* for CSQ.
- $\mathbf{BF_{10}}$: (a) 254.76 (b) 2247.95 (c) 37036.79 (all extreme evidence for H_1).

8. *Music Transformer will outperform MusicVAE on stylistic success ratings in the CPI part of study but not in the CSQ part.*

- Outcome: ***MuTr* and *MVAE* receive similar *Ss* ratings for both CSQ and CPI parts of the study.**
- $\mathbf{H_0}$: *MuTr* and *MVAE* have similar *Ss* ratings for (a) CPI and (b) CSQ.
- $\mathbf{H_1}$: *MuTr* outperforms *MVAE* on *Ss* ratings for (a) CPI and (b) CSQ.
- $\mathbf{BF_{10}}$: (a) 0.403 (anecdotal evidence for H_0); (b) 0.089 (strong evidence for H_0).

9. *Participants with fewer years of musical training tend to give higher ratings.*

- Outcome: **Years of musical training is not correlated with ratings given.**
- $\mathbf{H_0}$: There is no correlation between years of musical training and ratings.
- $\mathbf{H_1}$: Years of musical training is correlated with ratings.
- $\mathbf{BF_{10}}$: 0.059 (strong evidence for H_0).

Five of the nine outcomes above are as predicted in Sect. 3 (exceptions are #[5, 6, 8, 9]). The outcome of #1 suggests that *MaMa* and *MuTr* are capable of modelling the style of CSQ with the same performance. That is, the strongest-performing non-deep learning method *MaMa* is on a par with the strongest-performing deep learning method *MuTr*. Furthermore, *CoRe* shows inferior performance to *MaMa* according to #3. We can reasonably deduce that *deep learning methods have not surpassed non-deep learning methods in terms of their ability to automatically generate stylistically successful music*.

According to the outcome of #2, *MaMa* and *MuTr* have the same ability to generate repetitive structure. This is interesting because the two methods vary in the mechanism of pattern inheritance. *MaMa* uses a specified repetitive structure or one obtained via a pattern discovery algorithm (Collins, 2011; Collins et al., 2017) to guarantee that generated material inherits certain patterns, whereas *MuTr* achieves similar performance by relying on its representation and network architecture to capture relationships among notes. That is, over 30 s spans of music, what most music analysts would think of as short- or medium-term structure, deep learning methods can achieve the same performance in generating repetitive structure through learning as a non-deep learning approach (Collins, 2011; Collins et al., 2017) achieves through explicit imposition of a structure. Whether such a finding applies at the long-term level, beyond 30 s spans of music, remains to be seen.

For hypothesis #3, *CoRe* was outperformed by both *MuTr* and *MaMa* on *Ss*. For hypothesis #7, systems were overall better at modelling the style of CSQ compared to CPI. For hypothesis #4, we verify that *Me* ratings are positively correlated with *Ap* ratings. Further tests with *Ha* ratings and *Rh* ratings revealed smaller positive correlations.

Hypotheses #[5, 6, 8, 9] did not result as expected. In frequentist statistics, all we would be able to say is that we have failed to refute the null hypothesis. Because we are using a Bayesean approach, we can make stronger claims.

For hypothesis #5 the intention was to investigate whether systems generate generally listenable music but fail to capture certain stylistic aspects. The results indicate that generally this is not the case, and systems have similar performance on *Ss* and *Ap* ratings in both CSQ and CPI.

For hypothesis #8, we compare *MuTr* and *MVAE* on *Ss* ratings separately in CSQ and CPI. We had predicted similar performance in CSQ but superior performance for *MuTr* in CPI, since CPI is a more diverse, complex style to model and *MuTr* has a more powerful architecture than *MVAE*. The results however indicate that the systems have similar performance to each other in both styles. The evidence for similar performance is only anecdotal for CPI, whereas it is strong for CSQ.

The hypothesis and outcome of #6 is somewhat similar to #8, but here specifically for *Me*: we had predicted but do not see in the statistics evidence for *MuTr* having superior performance. In two cases (versus *MaMa* in CSQ, and versus *MVAE* in CPI) the evidence was again only anecdotal. The anecdotal results in #6 and #8 may warrant following up in future work.

The results for hypothesis #9 indicate there is no significant correlation between years of musical training and ratings given. This may be due to range restriction: we recruited only relatively expert listeners, so did not sample from a less expert pool of listeners who are sometimes more generous with stylistic success ratings in such listening studies.

6.3 Musicological analysis

Here we provide and comment on piano-roll (pitch against time) representations of some stimuli that warrant particular attention due to their reception in the listening study.[20] As a complement to the statistical analysis, we analyse some of the open-ended comments received too. Commenting and highlighting sections of stimuli is optional in the listening study; 305 out of a total 1,640 submitted responses have comments. We choose piano roll over staff notation for the following figures as the former is more interpretable for general readers.

Figure 3 shows three excerpts from the CSQ part of the study that received the highest median stylistic success ratings in their respective categories: (a) a sample of *Orig* composed by Haydn[21]; (b) an excerpt generated by *MuTr*; (c) an excerpt generated by *MaMa*. The human-composed excerpt in Fig. 3a has a clear melodic arc indicated by the blue dashed curve: the upper line rises to its highest point in bars 6–8 and falls to a lower register in bars 12–14. A sequential progression can be identified in bars 9–10.[22] The arc and sequence mechanisms are important for creating a sense of narrative or purpose and self-referentiality in music, but it is something rarely demonstrated in the model-generated music. Self-reference or repetitive structure can be recognised in multiple places in Fig. 3a: bars 1 and 2 are very similar to one another; so are bars 12, 13, and 14.

The arc indicated by blue dashed curve in Fig. 3b is inverted compared to that of (a): It begins and ends at relatively high pitches with a dip in the middle, which is less common in music of this period than the other way around, but is potentially preferable to no arc at all. The latter half (bars 8–15) contains ascending, sequential material. While there is no medium-term repetitive structure, bars 1.5–2.5 are similar to bar 5. That said, there is an unusual rhythmic displacement where the long chord in bar 4 ends one quarter of the way through the first beat bar 5. We deduce that the serialisation method used by MuTr makes the generated music vulnerable in aspects of rhythmic coherence. Some participants think (b) starts off well, but the "wonky" harmony means the latter part is perceived as stylistically incoherent.

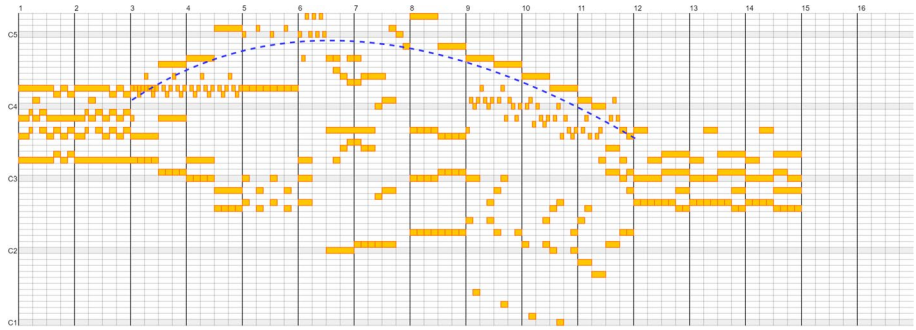
The model-generated excerpt in Fig. 3(c) also demonstrates a repetitive structure, with bars 1–4 repeating in bars 5–8, marked by blue dashed boxes, but it lacks the arc of either (a) or (b).

Figure 4 shows three excerpts from CPI part of the study that received the highest median stylistic success ratings in their respective categories: (a) a sample from *Orig*, (b) an excerpt generated by *MVAE*, and (c) an excerpt generated by *MuTr*. Compared with CSQ, excerpts in CPI contain more short notes resulting in rapid rhythms. Figure 4a contains a motif during the first half of bar 1 that is repeated three times with variations of

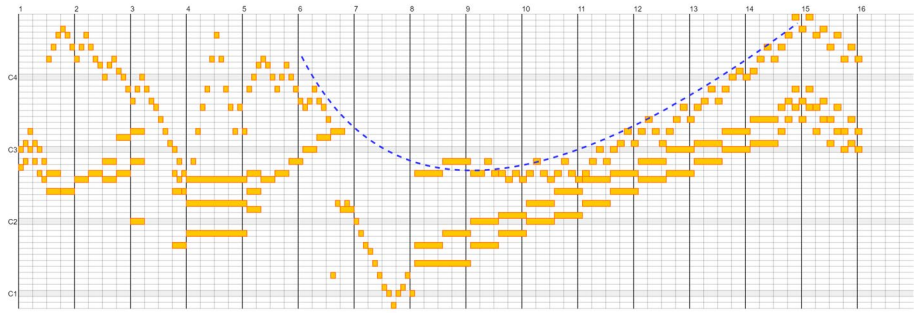
[20] Excerpts can be accessed via https://osf.io/96emr/.

[21] Haydn, Quartet in F Minor, op.20 no.5, first movement, bars 35–48.

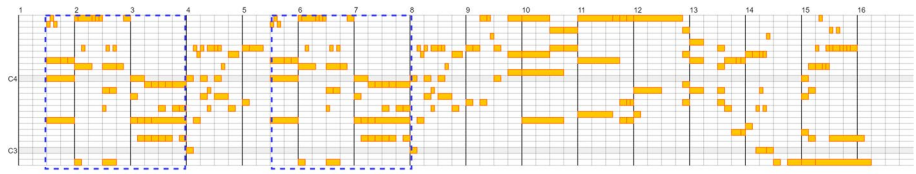
[22] A sequence in music is when a collection of (pitch, rhythm)-pairs is reused by shifting up or down in pitch and later in time by a fixed amount.



(a) *Orig*



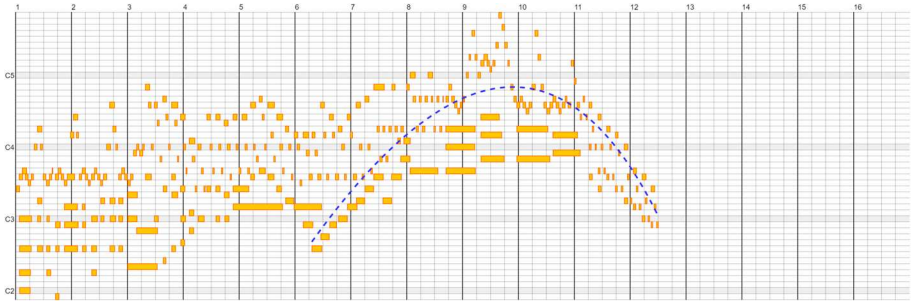
(b) *MuTr*



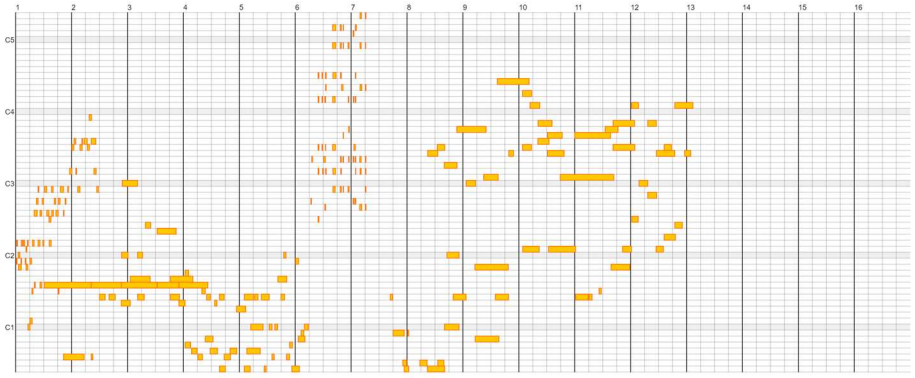
(c) *MaMa*

Fig. 3 Excerpts in CSQ with the highest median *Ss* ratings in their respective categories

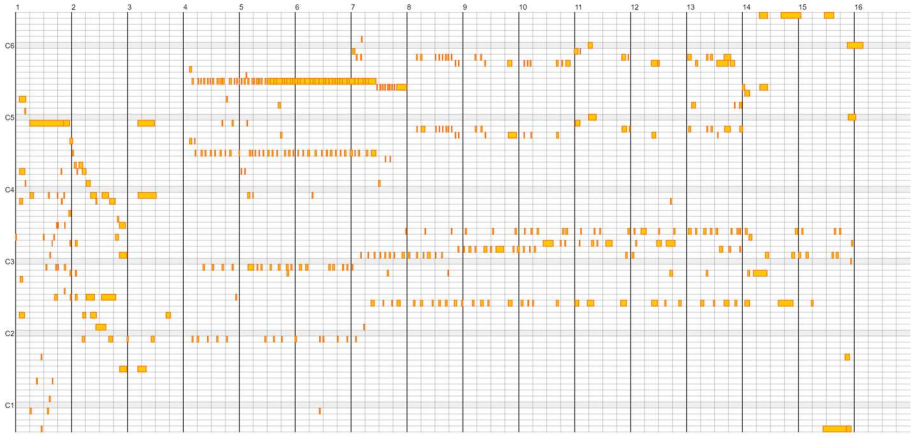
ascending pitches. The same motif is again apparent in bar 10. The excerpt as a whole contains an arc, particularly from bar 6 to the end. Figure 4b shows evidence of some local arcs, but the overall shape is less recognisable. Significant repetitive structures cannot be found, and the sudden rapid chords in bars 6–7 undermine the preceding coherence, although naturally a balance between expected and unexpected musical events needs to be struck. One participant observes that the sixteenth notes in the beginning half "betrayed" the rest of the piece, as they make for an odd pairing. The melody from bar 8 to the end is pleasant yet seems unfinished. Figure 4c demonstrates repetitions in various places like bars 4–6 and 8–10, which focus more on single note repetitions instead of phrases as with other excerpts. It does not have a clear arc either: some participants think it is inconsistent with respect to practices of the classical period but compelling to listen to nonetheless.



(a) *Orig*



(b) *MVAE*



(c) *MuTr*

Fig. 4 Excerpts in CPI with the highest median *Ss* ratings in their respective categories

The excerpt in Fig. 5 attracted some contradictory comments. Some participants indicated this excerpt contains too many repetitions, while others appreciated the crossing of the voices, and the overall sonic effect. Objectively, it highlights the ability of *MuTr* to

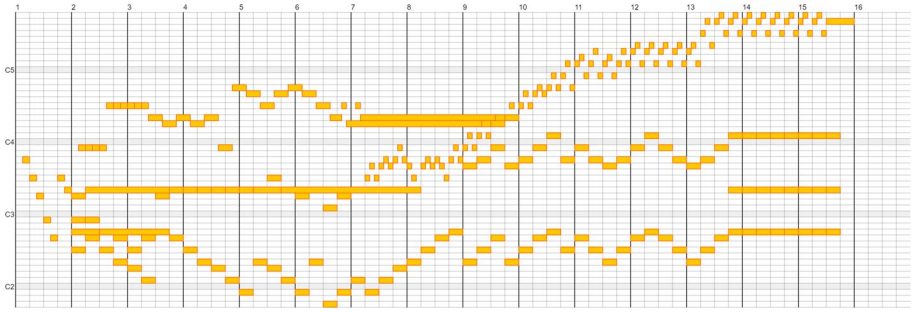


Fig. 5 An example of *MuTr* in CSQ

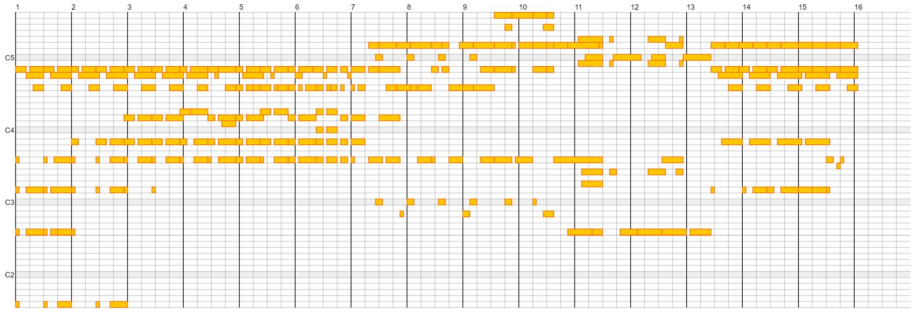


Fig. 6 An excerpt of *LiTr* found mimicking "Carol of the Bells"

generate excerpts with clear, identifiable voices, just as in the target style of Classical string quartets.

Deep learning systems suffer from copying large chunks of the training set in their outputs (Yin et al., 2021). This is the case here. The stimuli of *LiTr* are randomly selected from the web radio station published by the Magenta research group. Some of our participants recognised an excerpt from *LiTr* as shown in Fig. 6 copying the motif of "Carol of the Bells" composed by Mykola Leontovych (1914). This underlines the need for a systematic and widely adopted method to identify copying and potential copyright infringement in the output of AMG systems, for example, the originality report of Yin et al. (2021).

7 Discussion

Deep learning models have come to prominence and have been recognised, due to rigourous comparative evaluation, as state-of-the-art solutions for many tasks in machine learning, particularly in computer vision and natural language processing. Prior to this work, no such comparative evaluation existed for automatic music generation (AMG), where the research effort devoted to evaluation and evaluation methodologies has fallen behind the efforts invested in training and generating with novel network architectures (Pearce and Wiggins, 2001; Jordanous, 2012; Yang and Lerch, 2020).

7.1 Findings

This work reports the results of a comparative evaluation involving appropriately expert judges, for various symbolic AMG systems as well as human-composed material. This comparative evaluation covers four AMG systems: one non-deep learning method (Collins et al., 2017), and three deep learning methods with different generation strategies (Huang et al., 2018; Roberts et al., 2018; Thickstun et al., 2019). We use these systems to learn from datasets, and generate, in two target styles: Classical string quartets (CSQ) and classical piano improvisations (CPI). Combined with human-composed excerpts, we have conducted a listening study that asks participants to rate excerpts according to six musical dimensions on a scale 1–7. We analyse the results in the context of a non-parametric Bayesian hypothesis testing framework.

7.1.1 No advances due to deep learning, and copying problems

Broadly, the listening study results demonstrate that the assumption of superiority of deep learning methods for polyphonic AMG is unwarranted. The non-deep-learning model MAIA Markov (Collins et al., 2017) had the same level of performance as that of the best-performing deep learning models, MusicVAE (Roberts et al., 2018) and Music Transformer (Huang et al., 2018).[23] We also find that a large gap still exists between the stylistic success ratings for the strongest-performing computational models and the human-composed excerpts. Deep learning methods may be improving relative to one another according to metric-only evaluations, but these metrics appear to be of limited use, in that the listening experience of our participants indicates there is no improvement in stylistic success beyond a non-deep-learning method, and there is still an obvious gap in ratings compared to human-composed music.

In recent years, artificial intelligence systems have been increasingly involved in music creation and production, with musicians using them as cues for ideation, harmonisation, synthesis, mixing, and so on. An ethical issue here is that the data-driven property of generative models can lead to those systems copying chunks from original music data, which may result in plagiarism or copyright infringement. According to our previous work (Yin et al., 2021), existing deep learning AMG research neglects to check copying as part of its evaluation process. In that work, we discuss the nature of the language model-based Music Transformer algorithm, and show the extent of replication through various training checkpoints. For the listening study, we apply the same method (Yin et al., 2021) to measure and exclude excerpts that copy too much from the training set. The originality baseline, showing percentage of copying, is constructed with training and validation sets, and any generated excerpt with originality less than the lower 95%-confidence interval of the baseline is removed. In choosing 25 generated excerpts for each category of CSQ, 44% of MusicVAE (Roberts et al., 2018) excerpts and 56% of Music Transformer (Huang et al., 2018) excerpts

[23] Some might say the comparison is unfair because MAIA Markov incorporates repetitive structure explicitly, whereas MusicVAE and Music Transformer do not. But the titles of the MusicVAE and Music Transformer papers are "A hierarchical latent vector model for learning **long-term structure in music**" (Roberts et al., 2018) and "Music transformer: Generating **music with long-term structure**" (Huang et al., 2018) (our emphases), suggesting that the deep neural networks they employ are capable of (a) modelling long-term structure in music, and consequently (b) generating outputs where this long-term structure is evident/perceivable. As such, our comparison and conclusion are fair.

had to be removed due to the copying issue, and new excerpts generated until 25 passed the originality test. On the other hand, no over-copying was identified in excerpts generated by MAIA Markov (Collins et al., 2017) or the coupled recurrent model (Thickstun et al., 2019). With the "Carol of the Bells" example mentioned above in relation to Fig. 6, Music Transformer is again shown to exceed the level of borrowing considered reasonable among human composers. Although quantifying music originality is still a challenge for the development of AI systems and musicology, efforts are being made in the areas of plagiarism detection and benchmarking. For example, Spotify recently applied for a patent for its AI-driven music plagiarism detection method, based on a method for sequence alignment from the 1980 s (Pachet and Roy, 2020; Smith and Waterman, 1981).

Taking these two outcomes together—a non-deep-learning method (MAIA Markov, Collins et al., 2017) performing comparably to the strongest deep learning method (Music Transformer, Huang et al., 2018) in a human listening study where the participants are uninformed of the source of the composition, and the deep learning method raising concerns of ethical violations from direct copying—it seems there is still much for deep learning researchers to consider, improve upon, and evaluate, before a deep learning approach can be considered superior for automatic music generation.

7.1.2 Gulf between 'Listen to Transformer' and 'Music Transformer'

There is a gulf between *LiTr* (Listen to Transformer) and *MuTr* (Music Transformer) across all rating categories. Most likely, this is due to differences in training data: *MuTr* is trained with the MAESTRO dataset (Hawthorne et al., 2019) while *LiTr* is trained with a wider range of piano data: this is evidenced by the generated excerpt that copies "Carol of the Bells" shown in Fig. 6, which is not included in the MAESTRO dataset and is not in the "classical" era. If we had trained our *MuTr* implementation on this wider range of piano data, it may have improved the stylistic success of its outputs, but this wider dataset has not been made publicly available, and it is not clear how using it affects the originality of generated material.

We do not know the stopping criteria for training used for *LiTr*, and differences in these between *MuTr* and *LiTr* could also lead to differences in the results. For *MuTr*, our reimplementation uses the same early stopping technique stated in the original paper. However, our previous work Yin et al. (2021, 2022) demonstrates that systems can start over-fitting (in calculation of originality with respect to the training set) before the epoch with the lowest validation loss. This prompts the question: are the commonly used criteria for early stopping (e.g., validation loss) sufficient for generative models, or do the results need to be further assessed by other criteria (e.g., originality, musical metrics)?

7.2 Limitations

While we are confident in the functional equivalence between our reimplementation of *MuTr* and the original (Huang et al., 2018), we accept it is a slight but unavoidable limitation of this work that it has not been possible for us (or other researchers) to use the original implementation due to dependency issues.

The listening study highlights that the performance of AMG systems can vary according to target style and/or corresponding dataset (outcome of hypothesis #7). To prevent the case where a certain system is optimised for one specific style of dataset, we test with two target styles: Classical string quartets and classical piano improvisations. However, the

coverage of musical styles is still narrow and does not provide a broad basis for claiming our findings apply in general. We observe that the system performance is usually better for CSQ than for CPI. Based on the fixed configuration of model training, we believe that the difference is caused by the stylistic complexity of the compositions in each style: CSQ is narrower in style than CPI, and the music in CSQ shows more regularity that makes composition rules more predictable.

As described in Sect. 5.3, we apply some filters for repeated notes, long rests, and the originality report (Yin et al., 2021) to remove excerpts that copy large chunks of the original training data. Investigating the extent to which the system can be creative, in the sense that it is producing a novel output from its input, is another important aspect of creativity system evaluation (Jordanous, 2019) that is beyond the scope of this study. The 2021 version of the originality report supports only music data with non-expressive timing. Therefore, we applied it only to outputs from the CSQ dataset.

The selected musical dimensions for evaluation are obtained from prominent writings on Western music (Rosen, 1997). Our evaluation framework may not be applicable to other music styles (for example, atonal music or non-Western music), as it is not in their nature to emphasise these same musical dimensions. Even in this context, we could have provided clearer definitions of the terms "melody", "harmony", and "rhythm", and what it means for an excerpt to receive a relatively high rating for "melody", say.

In choosing to include both a model that performs poorly (Coupled Recurrent Model, Thickstun et al., 2019) and human-composed excerpts, we may have made it more difficult for listeners and, consequently, our statistical analyses, to distinguish differences between middling-performance systems. However, it is important to include a number of computational models in the study as well as human-composed music.

7.3 Future work

Feature extraction is a common approach in the field of music information retrieval, where dimensions of music are summarised or quantified (McKay et al., 2018). This could be used to further explore and analyse the relationship between musical components and the various ratings we gathered. The aim of such work would be to identify features that have explanatory power in terms of predicting relatively low or high ratings for generated music, and to help assess the capability of generative systems to model creativity.

The term "style" refers to (at least) two concepts in music: musical style, meaning the period in which a piece of music was written (influencing the choices of which notes to write and how to combine them—crudely, the note data); performance style, meaning the way in which it is performed (characterising the expressive performance parameters—crudely, adjustments in start times and durations of notes, and how strongly/softly they are played). As such, the former is mostly the domain of the composer, while the latter is mostly the domain of the performer. This is an oversimplification, because artists will sometimes take a classical piece, arrange it (change or add notes, instrumentation), and play it in, say, a jazz style. But still, the point remains that musical style and performance style are identifiable, separable concepts. Accordingly, and prior to the arrival of deep learning models for music, researchers tended to investigate these two topics separately. For instance, Conklin and Witten (1995) and Collins et al. (2016) investigate musical style, while Widmer (2002) and Grachten and Widmer (2011) investigate performance style. When Google Magenta started publishing on AMG models (including Roberts et al., 2018; Huang et al., 2018), they took the rather bold step of attempting to model both musical

style and performance style all in one neural network architecture. Given the conclusions drawn from the current study, it might be advisable to separate out these concepts again in future work, at least until the note-level originality issues of deep learning AMG models can be better understood and addressed.

As mentioned before, the field of AMG lacks a comprehensive and standardised evaluation framework, although we claim the current paper goes some way towards filling that gap. In future work, we will use the collected rating data and extracted features to predict ratings for new, previously unheard musical excerpts. As such, the predictive model would mimic the evaluation criteria of our human listeners, have some validity for the target styles of CSQ or CPI, and so remove the need for conducting a new listening study every time a researcher wants to evaluate a new network architecture or minor modifications to an existing model. This predictive approach would contribute to a greater degree of standardisation of evaluation procedures for AMG, as well as make systems more directly comparable. We will also consider integrating the distributional comparison approach of Yang and Lerch (2020) into our evaluation framework.

Considering the amount of data used for both parts of the study, in CSQ we used a relatively small dataset compared to CPI. Although our CSQ dataset represents a substantial amount of Classical music, and certainly enough for a human with general music knowledge to have an idea of the intended style, we would like to follow up on this principle by codifying it, pre-training the model with a large, generic music dataset, which gives the model "general musical knowledge", and then fine-tuning it with a smaller dataset in the target style (Donahue et al., 2019; Conklin and Witten, 1995). This would be a more realistic model of how humans learn to compose, and so we can imagine it may lead to a higher quality of generated outputs from computational models.

In Sect. 6.3, we observe that several generated excerpts do not exhibit a clear arc. While such a shape is not a necessary property for stylistic success, it is at least one that can be perceived from a piano-roll plot, and that could be quantified in future. We also suggest that the slicing method used during data preprocessing could be altered in this regard: slicing a whole piece into sub-sequences of fixed size may result in a large portion of excerpts not having a clear arc, so that the model cannot learn the abstract/high-level pattern because it is not present sufficiently often. In future, we could use certain segmentation methods (Rafael et al., 2009; Rodríguez-López and Volk, 2013) that might cause the training excerpts to have more appropriate beginnings, middles, and ends with respect to learning abstract patterns.

7.4 Conclusion

We present a comparative evaluation of four symbolic AMG systems. To the best of our knowledge, there was no such evaluation comparing deep learning-based and Markov-based approaches prior to this paper. Our evaluation is based on a human listening study and hypothesis testing with a Bayesian method (van Doorn et al., 2020), which contributes to filling a gap in the comparative evaluation of AMG systems, and the wider uptake of Bayesian methods across the evaluation of machine learning systems. The results show that the best deep learning and Markov-based algorithms for automatic music generation perform equally well, and there is still a significant gap to bridge between stylistic success ratings received by the strongest computational models and human-composed music.

This final checklist of suggestions for methodological fortitude is intended to help recover and sustain the accurate measure of progress in the field of AMG, as well as more broadly in the application of machine learning to domain-specific tasks:

1. Conduct a comparative evaluation via a listening (more broadly, participant) study;
2. If your particular technique or approach is *X* (neural networks, say), include at least one system from beyond *X* that can also be used to address the same task and that the literature suggests will be a competitive point of comparison;
3. Having formulated hypotheses about the likely findings of the study, pre-register (state) them via a service such as Open Science Framework;
4. Detail how you recruited and compensated your participants, as well as describing their musical (more generally, domain-specific) backgrounds and level of expertise;
5. Publish the listening study interface and associated stimuli so that other researchers can attempt to replicate your findings;
6. If you collect Likert ratings, use non-parametric statistics to analyse them; use Bayesian rather than frequentist hypothesis testing to evaluate your hypotheses;
7. If you insist on publishing via arXiv or blog (for time-stamping or marketing purposes, respectively), then follow up with submission of the work to a peer-reviewed conference or journal.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by ZY and TC. The first draft of the manuscript was written by ZY, with outer sections being developed by TC. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors did not receive support from any organisation for the submitted work.

Availability of data and materials Data and material used and collected from the study can be accessed via https://osf.io/96emr/.

Code availability Code for running experiments and analyses can be accessed via https://osf.io/96emr/.

Declarations

Conflict of interest The last author is a co-founder of the music collective Music Artificial Intelligence Algorithms, Inc. As such, he has an interest in the strong performance of the MAIA Markov algorithm. To avoid this bias affecting the outcome of the study, we ran the listening study blind and did everything one could reasonably expect to maximise the performance of outputs from other algorithms. Apart from this, the authors have no relevant financial or non-financial interests to disclose.

Ethics approval The listening study has been approved by the Physical Sciences Ethics Committee of the University of York.

Consent to participate Participants involved in the listening study are informed their provided data will be used anonymously for this work.

Consent for publication All authors consent to the publication of this work.

References

Agres, K., Forth, J., & Wiggins, G. A. (2016). Evaluation of musical creativity and musical metacreation systems. *Computers in Entertainment (CIE), 14*(3), 1–33.

Aguilera, G., Galán, J. L., Madrid, R., Martínez, A. M., Padilla, Y., & Rodríguez, P. (2010). Automated generation of contrapuntal musical compositions using probabilistic logic in derive. *Mathematics and Computers in Simulation, 80*(6), 1200–1211.

Allan, M., & Williams, C. K. (2005). Harmonising chorales by probabilistic inference. *Advances in neural information processing systems, 17*, 25–32.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997.

Ames, C. (1989). The Markov process as a compositional model: A survey and tutorial. *Leonardo, 22*(2), 175–187.

Anders, T., & Miranda, E. R. (2010). Constraint application with higher-order programming for modeling music theories. *Computer Music Journal, 34*(2), 25–38.

Ariza, C. (2009). The interrogator as critic: The turing test and the evaluation of generative music systems. *Computer Music Journal, 33*(2), 48–70.

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th ed.). London: Pearson.

Bel, B., & Kippen, J. (1992). *Bol processor grammars* (pp. 366–400). Cambridge, MA: MIT Press.

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings of the 28th international conference on neural information processing systems (Vol. 1, pp. 1171–1179)

Bigand, E., & Poulin-Charronnat, B. (2006). Are we experienced listeners? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100–130.

Boden, M. A. (1990). *The creative mind: Myths and mechanisms*. London: Weidenfield and Nicholson.

Cambridge University Faculty of Music: Music Tripos courses. Cambridge, UK (2010)

Cohen, H. (1999). Colouring without seeing: A problem in machine creativity. *AISB Quarterly, 102*, 26–35.

Collins, T. E. (2011). Improved methods for pattern discovery in music, with applications in automated stylistic composition. Ph.D. thesis, The Open University

Collins, T., & Laney, R. (2017). Computer-generated stylistic compositions with long-term repetitive and phrasal structure. *Journal of Creative Music Systems 1*(2)

Collins, T., Arzt, A., Flossmann, S., & Widmer, G. (2013). SIARCT-CFP: Improving precision and the discovery of inexact musical patterns in point-set representations. In: Proceedings of the international society for music information retrieval conference (ISMIR) (pp. 549–554)

Collins, T., Laney, R., Willis, A., & Garthwaite, P. H. (2011). Chopin, mazurkas and Markov: Making music in style with statistics. *Significance, 8*(4), 154–159.

Collins, T., Laney, R., Willis, A., & Garthwaite, P. H. (2016). Developing and evaluating computational models of musical style. *AI EDAM, 30*(1), 16–43.

Collins, T., Thurlow, J., Laney, R., Willis, A., & Garthwaite, P. (2010). A comparative evaluation of algorithms for discovering translational patterns in baroque keyboard works. In *Proceedings of the international society for music information retrieval conference (ISMIR)* (pp. 3–8)

Colton, S., Pease, A., Corneli, J., Cook, M., & Llano, T. (2014). Assessing progress in building autonomously creative systems. In: ICCC (pp. 137–145). Ljubljana

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research, 24*(1), 51–73.

Cope, D.: Experiments in musical intelligence (Vol. 12). AR editions (1996)

Cope, D. (2005). *Computer models of musical creativity*. Cambridge: MIT Press.

De Boom, C., Laere, S. V., Verbelen, T., & Dhoedt, B. (2019). Rhythm, chord and melody generation for lead sheets using recurrent neural networks. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 454–461). Springer

De Clercq, R. (2019). Aesthetic pleasure explained: De Clercq. *The Journal of Aesthetics and Art Criticism, 77*(2), 121–132. https://doi.org/10.1111/jaac.12636

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1 (Long and Short Papers), pp. 4171–4186)

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics, 41*, 214–226.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., & McAuley, J. (2019). LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Dong, H. W., Hsiao, W. Y., Yang, L. C., & Yang, Y. H. (2018). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the 32nd AAAI conference on artificial intelligence*

Ebcioğlu, K. (1990). An expert system for harmonizing chorales in the style of JS Bach. *The Journal of Logic Programming, 8*(1–2), 145–185.

Eck, D., & Schmidhuber, J. (2002). Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing* (pp. 747–756). IEEE

Eigenfeldt, A., & Pasquier, P. (2010). Realtime generation of harmonic progressions using controlled Markov selection. In *Proceedings of ICCC-X-computational creativity conference* (pp. 16–25)

Fernández, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research, 48*, 513–582.

Gardenfors, P. (2004). *Conceptual spaces: The geometry of thought*. Cambridge: MIT press.

Gjerdingen, R. (1988). *A classic turn of phrase: Music and the psychology of convention*. Philadelphia, PA: University of Pennsylvania Press.

Gjerdingen, R. (2007). *Music in the galant style*. Oxford, UK: Oxford University Press.

Goodwin, C. J., & Goodwin, K. A. (2016). *Research in psychology: Methods and design* (8th ed.). New York, NY: Wiley.

Grachten, M., & Widmer, G. (2011). Explaining musical expression as a mixture of basis functions. In *Proceedings of the 8th sound and music computing conference (SMC 2011)*

Graves, A., Mohamed, A. R., Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649). IEEE

Hadjeres, G., & Nielsen, F. (2020). Anticipation-RNN: Enforcing unary constraints in sequence generation, with application to interactive music generation. *Neural Computing and Applications, 32*(4), 995–1005.

Hadjeres, G., Pachet, F., & Nielsen, F. (2017). Deepbach: A steerable model for Bach chorales generation. In *International conference on machine learning* (pp. 1362–1371). PMLR

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., & Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. Z. A., Dieleman, S., Elsen, E., Engel, J., & Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International conference on learning representations*. https://openreview.net/forum?id=r1lYRjC9F7.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778)

Hedges, S. A. (1978). Dice music in the eighteenth century. *Music & Letters, 59*(2), 180–187.

Herremans, D., & Chew, E. (2017). Morpheus: Generating structured music with constrained patterns and tension. *IEEE Transactions on Affective Computing, 10*(4), 510–523.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology, 48*(2), 246–268.

Hild, H., Feulner, J., & Menzel, W. (1991). Harmonet: A neural net for harmonizing chorales in the style of JS Bach. In *Proceedings of the 4th international conference on neural information processing systems* (pp. 267–274)

Hiller Jr, L. A., & Isaacson, L. M. (1957). Musical composition with a high speed digital computer. In *Audio Engineering society convention 9*. Audio Engineering Society

Hofstadter, D. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books.

Huang, C. Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer: Generating music with long-term structure. In *International conference on learning representations*

Janssen, B., Collins, T., & Ren, I. Y. (2019). Algorithmic ability to predict the musical future: Datasets and evaluation. In *Proceedings of the international society for music information retrieval conference (ISMIR)* (pp. 208–215)

Johnson, D. D. (2017). Generating polyphonic music using tied parallel networks. In *International conference on evolutionary and biologically inspired music and art* (pp. 128–143). Springer

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation, 4*(3), 246–279.

Jordanous, A. (2019). Evaluating evaluation: Assessing progress and practices in computational creativity research. In *Computational creativity* (pp. 211–236). Springer

Kennedy, M., & Bourne, J. (2004). *The concise Oxford dictionary of music*. Oxford: OUP Oxford.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11*(3), 346–374.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Liang, F. (2006). Bachbot: Automatic composition in the style of Bach chorales. Master's thesis, University of Cambridge

Lim, H., Rhyu, S., & Lee, K. (2017). Chord generation from symbolic melody using BLSTM networks. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E. J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology, 55*(5), 331–347.

Louie, R., Coenen, A., Huang, C. Z., Terry, M., & Cai, C. J. (2020). 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* (pp. 1–13)

Mao, H. H., Shin, T., & Cottrell, G. (2018). Deepj: Style-specific music generation. In *2018 IEEE 12th international conference on semantic computing (ICSC)* (pp. 377–382). IEEE

McCormack, J., & Lomas, A. (2020). Understanding aesthetic evaluation using deep learning. In *International conference on computational intelligence in music, sound, art and design (part of EvoStar)* (pp. 118–133). Springer

McKay, C., Cumming, J., & Fujinaga, I. (2018). Jsymbolic 2.2: Extracting features from symbolic music for use in musicological and mir research. In *Proceedings of the international society for music information retrieval conference (ISMIR)* (pp. 348–354)

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., & Bengio, Y. (2017). SampleRNN: An unconditional end-to-end neural audio generation model. In *5th international conference on learning representations (ICLR 2017)*

Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science, 6*(2–3), 247–280.

Navarro, M., Caetano, M., Bernardes, G., Castro, L. N. D., Corchado, J. M. (2015). Automatic generation of chord progressions with an artificial immune system. In *International conference on evolutionary and biologically inspired music and art* (pp. 175–186). Springer

Nierhaus, G. (2009). *Algorithmic composition: Paradigms of automated music generation*. Berlin: Springer.

Norris, J. R., & Norris, J. R. (1998). *Markov chains* (Vol. 2). Cambridge: Cambridge University Press.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., Walsh, J. (2019). Deep learning vs. traditional computer vision. In *Science and information conference* (pp. 128–144). Springer

Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2018). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications, 32*, 1–13.

Pachet, F., & Roy, P. (2020). Plagiarism risk detector and interface. US Patent application number US 2020/0372882 A1

Papadopoulos, G., & Wiggins, G. (1999). AI methods for algorithmic composition: A survey, a critical view and future prospects. In *AISB symposium on musical creativity* (Vol. 124, pp. 110–117). Edinburgh, UK

Pearce, M., Meredith, D., & Wiggins, G. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientiae, 6*(2), 119–147.

Pearce, M., & Wiggins, G. (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 symposium on artificial intelligence and creativity in the arts and sciences* (pp. 22–32). Citeseer

Pearce, M. T., & Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In *Proceedings of the 4th international joint workshop on computational creativity* (pp. 73–80). Goldsmiths, University of London

Popper, K. (2005). *The logic of scientific discovery*. Routledge, London, UK. Original work published 1959

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. Routledge, London, UK. Original work published 1963

Quick, D., & Hudak, P. (2013). Grammar-based automated music composition in Haskell. In *Proceedings of the first ACM SIGPLAN workshop on Functional art, music, modeling and design* (pp. 59–70)

Rafael, B., Oertl, S., Affenzeller, M., & Wagner, S. (2009). Using heuristic optimization for segmentation of symbolic music. In *International conference on computer aided systems theory* (pp. 641–648). Springer

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning* (pp. 4364–4373)

Rodríguez-López, M., & Volk, A. (2013). Symbolic segmentation: A corpus-based analysis of melodic phrases. In *International symposium on computer music multidisciplinary research* (pp. 548–557). Springer

Rosen, C. (1997). *The classical style: Haydn, Mozart, Beethoven*. London: WW Norton & Company.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*(1), 195–197.

Steedman, M. J. (1984). A generative grammar for jazz chord sequences. *Music Perception, 2*(1), 52–77.

Street, A. P., & Street, D. J. (1986). *Combinatorics of experimental design*. Oxford: Oxford University Press, Inc.

Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., & Pachet, F. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research, 48*(1), 36–55.

Sturm, B., Santos, J. F., & Korshunova, I. (2015). Folk music style modelling by recurrent neural networks with long short term memory units. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Tan, H. H., & Herremans, D. (2020). Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Thickstun, J., Harchaoui, Z., Foster, D. P., & Kakade, S. M.: Coupled recurrent models for polyphonic music composition. In *Proceedings of the international society for music information retrieval conference (ISMIR)*

Todd, P. M. (1989). A connectionist approach to algorithmic composition. *Computer Music Journal, 13*(4), 27–43.

Torrance, E. P. (1998). Torrance tests of creative thinking: Norms-technical manual: Figural (streamlined) forms A & B. Scholastic Testing Service.

Toshniwal, S., Kannan, A., Chiu, C. C., Wu, Y., Sainath, T. N., & Livescu, K. (2018). A comparison of techniques for language model integration in encoder–decoder speech recognition. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 369–375). IEEE.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv:1609.03499

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and spearman's ρ. *Journal of Applied Statistics, 47*, 1–23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008)

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*(3), 158–189.

Widmer, G. (2002). Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research, 31*(1), 37–50.

Widmer, G. (2016). Getting closer to the essence of music: The con espressione manifesto. *ACM Transactions on Intelligent Systems and Technology (TIST), 8*(2), 1–13.

Wiggins, G. A. (2008). Computer models of musical creativity: A review of computer models of musical creativity by David Cope. *Literary and Linguistic Computing, 23*(1), 109–116.

Xenakis, I. (1992). *Formalized music: Thought and mathematics in composition* (Vol. 6). New York: Pendragon Press.

Yang, L. C., Chou, S. Y., & Yang, Y. H.: MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. In *Proceedings of the international society for music information retrieval conference (ISMIR)* (pp. 324–331)

Yang, L. C., & Lerch, A. (2020). On the evaluation of generative models in music. *Neural Computing and Applications, 32*(9), 4773–4784.

Yin, Z., Reuben, F., Stepney, S., Collins, T. (2021). A good algorithm does not steal–it imitates: The originality report as a means of measuring when a music generation algorithm copies too much. In *Artificial intelligence in music, sound, art and design: 10th international conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, proceedings 10* (pp. 360–375). Springer International Publishing.

Yin, Z., Reuben, F., Stepney, S., & Collins, T. (2022). Measuring when a music generation algorithm copies too much: The originality report, cardinality score, and symbolic fingerprinting by geometric hashing. *Springer Nature Computer Science, 3*, 340.

Authors and Affiliations

Zongyu Yin[1] · Federico Reuben[2] · Susan Stepney[1] · Tom Collins[2,3]

Federico Reuben
federico.reuben@york.ac.uk
https://www.york.ac.uk/arts-creative-technologies/people/federico-reuben/

Susan Stepney
susan.stepney@york.ac.uk

Tom Collins
tomthecollins@gmail.com
https://tomcollinsresearch.net

[1] Department of Computer Science, University of York, York, UK

[2] School of Arts and Creative Technologies, University of York, York, UK

[3] Music Artificial Intelligence Algorithms, Inc., Davis, California, USA