View Reviews

Paper ID
40

Paper Title
Multitrack Music Transformer: Learning Long-Term Dependencies in Music with Diverse Instruments

Track Name
Papers

Reviewer #1

Questions

2. I am an expert on the topic of the paper.
Agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly agree

5. The paper discusses, cites and compares with all relevant related work
Agree

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Agree

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Agree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the

scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Agree

16. Please explain your assessment of reusable insights in the paper.
The details are clearly written in the paper.
According to the authors, the source code will be publicly available.

17. Write ONE line (in your own words) with the main take-home message from the paper.
The authors made attempts to generate multi-track MIDI-based musical pieces using Transformer.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Disagree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Strong accept

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
I think that it is well done work. It should be accepted in ISMIR 2022.
But, it was difficult to interpret comparisons between your models and
baseline models (MMM, REMI+), because the data representation etc.
adopted in MMM and REMI+ were not described in details.

Strength:
- The details of data representation are clearly written
- The generated pieces (available at the web page) are of high quality.
Weakness:
- Details of MMM and REMI+ were not described.
The authors clearly describe what are the difference between your models
and these baseline models and discuss how these differences caused
the differences of the experimental results.
- It should be to add an objective evaluation from musical point of view,
e.g., the counts of dissonant chords.

Reviewer #2

Questions

2. I am an expert on the topic of the paper.
Strongly agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly agree

5. The paper discusses, cites and compares with all relevant related work
Disagree

6. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen)
All relevant related work is cited, but some connections to it are missed (see main review), in particular to the closely related representation used by [50], and to the representations used by [8,9], which compress sequences without any independence assumptions. These works are mentioned only briefly and are not compared to.

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Strongly agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Agree

11. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen)
I have doubts about the correctness of some of the conclusions, as detailed in my main review:
- The values in Fig. 3b are not necessarily comparable across fields, and as a result, the conclusions in Section 4.4 may be incorrect.
- The analysis done in section 5 has issues that also lead to possibly incorrect conclusions.
Other than that, the content seems rather scientifically correct.

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Disagree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that

are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Strongly Disagree (Well-explored topic, task, or application)

15. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

16. Please explain your assessment of reusable insights in the paper.
The design choices are mostly well-motivated. However, as for the experiments, I do not really find them to provide strong reusable insights:
- The validation loss curves in Fig. 3a are quite expected.
- The values in Fig. 3b are not necessarily comparable across fields, and as a result, the corresponding conclusions in Section 4.4 may be incorrect (this is detailed in my other comments in the main review).
- The observations in Section 4.7 (from Fig. 5) are somewhat interesting, but also quite expected.
- My overall takeaway from both the objective and subjective evaluation is that the strong independence assumption introduced in this paper is harmful to generation quality (which is a reusable insight, but not too unexpected). The results unfortunately do not allow to see whether a more compact representation such as the one proposed here leads to better generation quality (e.g. through better long-range dependency modelling).
- The analysis done in Section 5 does offer some interesting insights and in my opinion has the potential to be used more widely. However, it also has some issues (some of them again leading to possibly incorrect conclusions) which are detailed in my main review.

17. Write ONE line (in your own words) with the main take-home message from the paper.
A new compact representation for multi-track generation is proposed, along with a new musically motivated technique for analyzing Transformer attention.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Strong reject

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
The paper proposes Multitrack Music Transformer (MTMT), a multi-instrument music generation model that uses a compact representation. The model is evaluated using objective metrics and a listening test and offers "competitive quality" compared to two baseline models while benefiting from reduced inference time. The authors also propose a new technique for analyzing Transformer self-attention in a musically meaningful way.

The paper reads well, describes with clarity almost all the details needed to understand the methods, and discusses relevant literature. However, I have doubts about the novelty of the proposed method, I do not find the presented evaluation results particularly convincing in terms of the proclaimed benefits of the method, and I also question the correctness of some of the conclusions.

The main novelty of MTMT according to the authors appears to be the proposed representation, where each position in the sequence corresponds to a note, as opposed to some other representations that include explicit timing events. This leads to a shortening of the sequences, leading to improved generation speed and supposedly better modelling of long-term dependencies. However, the novelty of this idea is limited in my opinion, given that:
- It seems like an incremental improvement on top of the compound word representation [6], which is already a compressed representation. The only major difference apparently is that timing and note information is combined and treated as one unit. (Unfortunately, it is hard to see what exactly this difference is just from reading this paper, since the compound word representation is never described, even though it is stated that the proposed representation is based on it.)
- A very similar representation, called OctupleMIDI, is already used by MusicBERT [50]. This paper could then be viewed simply as using a simplified version of OctupleMIDI in an autoregressive generation scenario, as opposed to masked language modelling as in MusicBERT. However, this connection is not discussed in the paper.

The main proclaimed advantage of this method is that it allows generating longer sequences, which should help in modelling long-term dependencies. However, the method unfortunately did not outperform the baselines in the objective and subjective evaluation results, and no other experiment is provided to demonstrate better long-term dependency modelling. As for the ability to simply generate long sequences and/or improved generation speed, this is already possible using models such as Transformer-XL or linear-complexity transformers, which have been applied to music previously ([9] and Liutkus et al., 2021).

Moreover, a major drawback of the proposed method is that it "cannot model the interdependencies between the six output heads as it predicts each field independently". This is an issue shared by the compound word Transformer, and only exacerbated here due to also predicting the timing independently of the other attributes. While this allows for modelling longer sequences, it probably comes at a cost of reduced output quality, as noted by the authors and evidenced by both the objective and subjective evaluation results. This is unfortunate, as it prevents us from seeing whether the availability of longer context improves the long-term dependency modelling.

Given that this is yet another paper in a series proposing a new representation for music generation, I think this is a missed opportunity to try to at least partially remedy the mentioned independence assumption. For example, some fields can be predicted jointly as in [8,9], where each token in the vocabulary is a combination of different variables (e.g. instrument, velocity, pitch). I believe testing a scheme like this at least for comparison could greatly improve the paper.

The other main contribution presented in the paper, and in my opinion maybe the most interesting part of the paper, is the self-attention map analysis method proposed in section 5. The idea proposed here is to visualize attention maps not by absolute indices, but by musically meaningful relations, e.g. beat difference, pitch difference. I think such methods have a potential to be used on any data where there are explicit relations between different elements (e.g. graphs), or where position (time) is not linear. However, in this paper, the idea is not too developed and has the following issues:
- Mean relative attention is defined as the total attention between pairs of elements with a "musical offset" of k, normalized over all possible k. This means that it will be biased towards values of k that occur frequently in the dataset. Thus, the conclusion on L355 that the model "pays less attention to pitches that form a dissonant interval with the current note" is not valid, as the pattern seen in Fig. 6 (c1) and (c2) could be simply reflecting the fact that dissonant intervals occur less frequently in the test examples.
- The authors do not state which of the 6 layers of the model is used for the visualization. Attention maps can be very different for different layers.
- There are methods that integrate information from different layers (Abnar and Zuidema, 2020), which would have been a better fit. Attention maps of a single layer can be hard to interpret or even misleading.

- This method could have been used to explore the hypothesis that the proposed representation would improve long-term dependency modelling. However, the authors unfortunately chose to only apply it to the main proposed MTMT-APE model and not to the baselines.
- Plots (b1), (b2), (c1), (c2) are dominated by position/pitch difference 0. This makes me wonder if the authors are really excluding the current position when computing mean relative attention (i.e. considering s>t in eq. (1) and not s>=t). Otherwise, this would probably be simply reflecting the fact that Transformers tend to attend most strongly to the current position.

Other remarks that might help improve the paper:
- L87: "of a certain difference from the query" – unclear expression
- L249: I think the more likely reason is simply the different number of possible values for each field. To get comparable loss values, consider normalizing them either by the log of the number of possible values of each field, or by the overall entropy of the field.
- Figure 4 is too small when printed. Instead, consider adding piano rolls to the website alongside the audio examples.
- L317: What is the music education level of the participants?
- L332: Which statistical test was used and with what inputs exactly?
- The conditional independence assumption, resulting in the fact that the model "cannot model the interdependencies between the six output heads" (different features of a note) as stated in Section 4.8, is an important and strong assumption and in my opinion should be mentioned much earlier in the paper.
- Eq. (1): It should be stated what (d) stands for.

References:
Abnar and Zuidema, 2021: https://arxiv.org/abs/2005.00928
Liutkus et al., 2021: http://proceedings.mlr.press/v139/liutkus21a.html

Reviewer #3

Questions

2. I am an expert on the topic of the paper.
Disagree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the “Special Call” column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly agree

5. The paper discusses, cites and compares with all relevant related work
Strongly agree

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 template, has at most 6 pages of technical content followed by “n” pages of references, references are well formatted). If you

selected “No”, please explain the issue in your comments.
Yes

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Strongly agree

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Disagree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Strongly agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated “Strongly Agree” and “Agree” can be highlighted, but please do not penalize papers rated “Disagree” or “Strongly Disagree”. Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Strongly Disagree (Well-explored topic, task, or application)

15. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

16. Please explain your assessment of reusable insights in the paper.
Using a more compact representation for symbolic music generation.

17. Write ONE line (in your own words) with the main take-home message from the paper.
The paper presents a slightly more compact word representation than the Compound Word Transformer, and demonstrates that the transformer network attends more to the notes that are 4N beats away.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Disagree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Strong reject

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
I believe that the paper presents little novelty on top of the Compound Word Transformer (CWT). Also, CWT isn't included in the baseline methods to be compared with.

The results in Table 1 are unclear. What is "Size (M)"? The text says that the baseline models need multiple passes to create a single note. However, this shouldn't be related to the results in Table 1 because the samples in Table 1 are created autoregressively. Hence, unless a maximum number of tokens is specified, the length of a sample depends on when an "end-of-sequence" token is generated. Hence, it is not clear to me what causes the different values in this table.

Looking at the results on Table 2 and 3, the proposed method underperforms compared to the baseline methods.

View Meta-Reviews

Paper ID
40

Paper Title
Multitrack Music Transformer: Learning Long-Term Dependencies in Music with Diverse Instruments

Track Name
Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. I am an expert on the topic of the paper.
Strongly agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Agree

5. The paper discusses, cites and compares with all relevant related work.
Disagree

6. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen)
There could have been more consideration given to previous work conducting listening studies in the evaluation of computer-generated music.

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Disagree

11. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen)
See my comments relating to statistical significance and hypothesis testing.

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of

existing methods from other research fields are considered novel at ISMIR conferences.
Disagree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Disagree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Agree

16. Please explain your assessment of reusable insights in the paper.
The main reusable insight is that if a more compact representation is used, a transformer model can generate content containing more long-term dependencies.

17. Write ONE line (in your own words) with the main take-home message from the paper.
If a more compact representation is used, a transformer model can generate content containing more long-term dependencies.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Disagree

21. Overall evaluation (to be completed before the discussion phase): Please first evaluate before the discussion phase. Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines.
Strong reject

22. Main review and comments for the authors (to be completed before the discussion phase). Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
A lot of work has gone into this paper.

But I find there are fundamental flaws in terms of what has been evaluated and how it has been evaluated that make it difficult for me to accept this paper for ISMIR.

The evaluation does not select appropriate comparisons:
1. How do your generated results compare to real, human-composed excerpts in terms of stylistic success ratings in a listening study?
2. How do your generated results compare to well-known models in this field, such as Music Transformer, and Markov models predating the introduction of deep learning, in terms of stylistic success ratings in a listening study?

It is in the paper's favor that a listening test was conducted, but it was not clear why terms such as coherence and richness were used. This should have been justified in relation to listening study methodologies and terms established in previous research.

Pearce, M., Wiggins, G.: 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 (2001)

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

The paper does not include an evaluation of the originality of generated passages: have you checked whether generated passages contain chunks of training data?

I have listened carefully to many of the examples provided at
https://dezimynona.github.io/mtmt/
I think the outputs of this model would be evaluated below earlier models such as Music Transformer or a Markov model in a listening study evaluation.

The results of the listening test are inconclusive: REMI+ is not statistically significantly higher in coherence than MTMT as the confidence intervals overlap. Therefore, to say that "the REMI+ model outperforms the MTMT-APE model" is incorrect. You acknowledge as much later in the same paragraph. But neither can these results be taken to mean that "our proposed model provide[s] -a- competitive performance against the baseline models". In a frequentist hypothesis testing framework, failure to reject the null hypothesis of "no difference between systems" is just that; you cannot infer a meaningful non-difference.

While it is probably advisable to use piano-roll notation in an ISMIR paper, because not all readers will cope with staff notation, the scale used for Figures 2 and 4 is too small to be of use.

---

23. Final recommendation (to be completed after the discussion phase) Please give a final recommendation after the discussion phase. In the final recommendation, please do not simply average the scores of the reviewers. Note that the number of recommendation options for reviewers is different from the number of options here. We encourage you to take a stand, and preferably avoid "weak accepts" or "weak rejects" if possible.
Reject

---

24. Meta-review and final comments for authors (to be completed after the discussion phase)
This paper has been reviewed by four reviewers including myself. My original review should be included somewhere among this feedback package.

We find that a lot of work has gone into this submission. However, the reviews and discussion phase confirmed major shortcomings in the work that meant we did not recommend it for inclusion in ISMIR 2022. If the issues and improvements highlighted below could be addressed, an updated version of this paper would be more likely to succeed at a future ISMIR.

Lacking appropriate points of comparison (i.e., missing or ignoring algorithms from other researchers)

Several reviewers found that the selection of algorithms for a comparative evaluation was quite insular, missing or ignoring other relevant work. Reviewer 2 has written in most detail about this, which hopefully will be useful in designing an improved comparative evaluation in future.

I would also note it is typical to include human-composed excerpts in an evaluation tool, to measure/show the gap between what can be achieved with machine learning compared to expert human composers. Or if you decide not to do this, justify this decision explicitly, backed up by references as to why.

Lacking appropriate terms and statistical analysis of ratings

It was not clear why listeners rated outputs according to terms such as coherence and richness. This should have been justified in relation to listening study methodologies and terms established in previous research. I have included a few references in my own review, but a Google Scholar cited-by search will reveal more recent examples too.

With respect to the statistical analysis of ratings, the lowest/highest ratings possible were not stated (a 1-5 or 1-7 Likert scale is typical). It is also useful to know about the listeners' musical backgrounds and expertise. While confidence intervals are provided, the authors make some mistakes with interpreting these results (my own review contains more details).

Long-term dependencies were addressed before deep learning was applied to music generation

It is worth noting that before and alongside deep learning models emerging as a method for generating music, other authors have addressed the issue that computational music generation methods often lack the ability to achieve long-term dependencies in their outputs.

Cope says "In music, what happens in measure 5 may directly influence what happens in measure 55, without necessarily affecting any of the intervening material" (2005, p. 98)

Collins et al. (2017), following in Cope's footsteps, provide an algorithm with code for embedding a local generative process in a global structure, thus ensuring output contains long-term dependencies.

Overall, we think it is important to clearly define what you mean musically by "long-term dependencies" (how long is "long"?), and to look beyond deep learning papers to other music generation papers when reviewing previous attempts to address this important issue.

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

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