



<http://www.diva-portal.org>

This is the published version of a paper presented at *AI Music Creativity*.

Citation for the original published paper:

**Sturm, B. (2022)**

**The Ai music generation challenge 2021: Summary and results**

**In:**

**N.B. When citing this work, cite the original published paper.**

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-326350>

# The Ai Music Generation Challenge 2021: Summary and Results

Bob L. T. Sturm\*

Tal, Musik och Hörsel, School of Electrical Engineering and Computer Science  
KTH Royal Institute of Technology, Lindstedtsvägen 24, Stockholm, Sweden SE-100 44  
bobs@kth.se

## Abstract

We discuss the design and results of *The Ai Music Generation Challenge 2021* and compare it to the challenge of the previous year. While the 2020 challenge was focused on the Irish double jig, the 2021 challenge was focused on a particular kind of Swedish traditional dance music, called *slängpolska*. Six systems participated in the 2021 challenge, each generating a number of tunes evaluated by five judges, all professional musicians and experts in the music style. In the first phase, the judges reject all tunes that are plagiarised, or that have incorrect meter or rhythm. In the second phase, they score the remaining tunes along four qualities: dancability, structure coherence, formal coherence, and playability. The judges know all the tunes are computer generated, but do not know what tunes come from what systems, or what kinds of machine learning and data are involved. In the third stage, the judges award prizes to the top tunes. This resulted in five tunes garnering first and second prizes, four of which come from one particular system. We perform a statistical analysis of the scores from all judges, which allows a quantitative comparison of all factors in the challenge. Finally, we look to the 2022 challenge.

## 1 Introduction

*The Ai Music Generation Challenge* is a public machine learning challenge having three aims: 1) to promote meaningful approaches to evaluating music artificial intelligence (Ai); 2) to see how music Ai research can benefit from traditional music, and how traditional music can benefit from music Ai research; 3) to facilitate discussions about the ethics of music Ai research applied to traditional music practices. Each challenge focuses on a specific domain of traditional music practice, selected based on the availability of data, as well as human experts for evaluating generated material.

The first occurrence of the challenge in 2020 (Sturm and Maruri-Aguilar, 2021) focused on a form of Irish traditional dance music (double jig), and involved seven systems including a benchmark – *folk-rnn* (Sturm et al., 2016). The reference collection consists of 365 double jigs from *The Dance Music of Ireland: O’Neill’s 1001* (1907). In four stages of evaluation, four human judges (all experts in Irish traditional music) evaluated tunes selected randomly from up to 10,000 generated by each system. Five tunes were selected from each system, giving a total of 35 tunes. In the first stage, the judges individually rejected a tune if: it is plagiarised, or it does not exemplify the double jig rhythm, or its pitch range is not characteristic, or its mode and use of accidentals are not characteristic. For the remaining tunes, the second stage involved the judges individually rating each in five qualities: melody, structure, playability, memorableness, and interestingness. The third stage involved the judges meeting to discuss their observations, and to lobby for favorite tunes. They awarded prizes to two tunes, one generated by the benchmark. The fourth stage involved the judges considering the consistency of a system in generating tunes of high quality. This did not result in any award.<sup>2</sup>

---

\*<https://www.kth.se/profile/bobs>

<sup>2</sup>A film about the 2020 challenge featuring two judges playing the award-winning tunes can be seen here: <https://youtu.be/KSoSyoEx6hc>



Figure 1: “Flickornas Michaelidans”, a slängpolska from Småland, notated for fiddle.

Participants of *The Ai Music Generation Challenge 2021*<sup>3</sup> were to build an artificial system that generates the most plausible *slängpolska* – a form of Swedish traditional dance music. As in the 2020 challenge, submissions to the 2021 challenge were evaluated by a panel of human judges, all professional musicians and experts in traditional Swedish music.<sup>4</sup> We selected this style from expert elicitation with one of the judges. Slängpolska is a common kind of polska heard in Sweden, and there exists many digitized examples in the online Scandinavian folk music resource <http://folkwiki.se>. This style has a 3/4 meter with an even pulse in performance. Unlike a traditional Irish double jig, there is no set metric structure for slängpolska. In the example shown in Fig. 1, the first part has four bars and the second has six; but other slängpolska can have two or more parts consisting of 14 bars or more. In performance, a slängpolska is repeated as many times as desired, sometimes with improvised second voices.

In the following, we review the design and outcomes of the 2021 challenge. The winning tunes are notated with reflections of the judges. We statistically analyze the judging scores to compare the factors of the challenge. We discuss several points of comparison with the 2020 challenge, and look forward to the challenge of the coming year.

## 2 The Ai Music Generation Challenge 2021

### 2.1 Design

The design of the 2021 challenge closely follows that of the 2020 challenge, but with a few differences motivated by the change of music style, and observations from the previous challenge (Sturm and Maruri-Aguilar, 2021). First, each participant is required to submit a collection of 1,000 tunes generated by their system, rather than 10,000. The larger number was motivated in 2020 by a desire to make human curation infeasible; but we felt an order of magnitude reduction could still achieve this while reducing the total computational effort. Second, the tunes in a submission must be rendered as MIDI and notation (such as ABC, musicXML, or staff). Judges of the 2020 challenge evaluated only staff notation, which they found to be burdensome. For the 2021 challenge, the judges evaluated each tune based on notation, midi, and an mp3 soundfile synthesized with the same appropriate tempo for the style, and using the same piano soundfont. Each participant is still required to submit a brief technical document describing their system, and linking to code and models for reproducibility.

The five judges in 2021 evaluated a collection tunes selected at random from the submissions. Originally, each submission was to contribute five tunes; but we increased this to ten because the number of participants was only six. We selected nine tunes at random from each submission. Unlike in the previous challenge, we let the 2021 participants elect one tune from their collection for evaluation. We hoped that this would motivate each participant to critically engage with their generated tunes in reference to the style, which could positively influence the engineering of their systems. We also made public a brief introductory video about slängpolska,<sup>5</sup> with two judges discussing what they will be looking for when evaluating tunes. In total, each judge evaluated 60 tunes generated by five submissions and a benchmark system.

The 2021 challenge consisted of four stages. In the first stage, each judge individually reviews each tune and rejects it if: 1) they detect plagiarism; or 2) the meter is not characteristic of a slängpolska; or 3) the rhythm is not characteristic of a slängpolska. We decided these criteria by expert elicitation with one of the judges. In the second stage, for all tunes not rejected, each judge individually grades tunes (A = excellent, B, C, D, F = failure) in four qualities: Danceability, Stylistic coherence, Formal

<sup>3</sup><https://github.com/boblsturm/aimusicgenerationchallenge2021>

<sup>4</sup>The judges were: Sven Ahlbäck, Eva Blomquist-Bjämborg, Lena Jonsson, Anders Löfberg, and Olof Misgeld.

<sup>5</sup>[https://play.kth.se/media/AIMGC2021/0\\_kgu3qwog](https://play.kth.se/media/AIMGC2021/0_kgu3qwog)

System ID	Approach
Benchmark (B)	<i>folk-rnn</i> fine-tuned on Swedish traditional music (Hallström et al., 2019) seeded with the start token and 3/4 meter token
Jönköping (J)	Markov chain then genetic algorithm with fitness function based on music structure
Kalmar (K)	Transformer architecture with templates derived from existing Swedish tunes and rejection sampling
Oskarshamn (O)	Iterative elaboration of a template guided by principles of music theory
Småland (S)	as Benchmark, but with beam search and curation by “artificial critic”
Växjö (V)	Transformer architecture trained with Irish data, fine-tuned with Swedish data, and using rejection sampling (Casini and Sturm, 2022)

Table 1: Summary of systems participating in the 2022 challenge.

coherence, Playability. On the grading sheet provided to the judges, the stylistic and formal coherence are defined as follows, coming from expert elicitation with one of the judges:

- *Stylistic coherence*: “Does the rhythmic and melodic structure of the melody evolve in a way which makes up a coherent structure? Do the rhythms fit into the same metrical general structure typical of slängpolska? Do the melodic motifs and phrases express stability, consistency and contrast with respect to melodic motion in a consistent way similar to slängpolska?”
- *Formal coherence*: “Do the sections and parts of the melody make up a coherent structure, where the sections are combined and are balanced in a manner which is similar to traditional slängpolska? Are the themes related and contrast in a way which creates a whole and a coherent story sequence of segments? Is there a structural coherence in the sense that motifs and sub-segments are consistent with higher level segments, like phrases, sections and repeats, in a way which is consistent with slängpolska?”

We specifically tell the judges to consider the following in their evaluation of the tunes:

- “Some tunes may be notated with slight errors, e.g., a measure missing a quaver, or a missing repeat sign. Please do not consider those as marks against the tune.”
- “Some tunes may be in keys that are not characteristic. Please do not consider those as marks against the tune. If it can be transposed to a more characteristic key then that’s good.”
- “Please try to complete your assessment of each tune in 5 minutes on average. Some may be easy, but others may take some time.”

The third stage of the challenge involves the judges meeting in person to elect their favorite tunes in the collection and discuss their observations. Together they decide on which tunes to award prizes (or none at all). Finally, the fourth stage involves the judges playing select tunes for dancers, who then elect their favorites (or none at all).

## 2.2 Results

The 2021 challenge attracted 12 participants, but only six submitted in the end including the benchmark. Table 1 summarises the six systems of the challenge. Two systems (B and S) use long short-term memory networks (LSTM). Two systems use transformer architectures (K and V). The remaining two systems use principles of music theory to guide the generation of tunes, either through a genetic algorithm (J) or a tree-based rule structure (O).

Table 2 shows the results from the judges for all tunes, with grades coded as following: A is 5, B is 4, C is 3, D is 2 and F is 1. (One judge used grades E and F, reflecting the grading scale used in Sweden. Herein, we consider grades E and F equivalent.) No tunes were rejected due to plagiarism, but several were rejected based on meter and rhythm. This was especially problematic for systems J and K. Several tunes from these systems had 2/4 and 4/4 meters. All judges are unanimous when rejecting tunes by meter, but for only one tune do all judges agree that the rhythm is unacceptable: 339 by J. This appears to be due to a prevalence of dotted quaver-semiquaver pairs, which some judges find uncharacteristic of the style, while others allow it.

Judges A and E were unable to meet in person with the others in stage 3, but specified to us in private their top three tunes beforehand, with judge A also mentioning a fourth (117 by V) that they felt

	Tune no.	Judge A				Judge B				Judge C				Judge D				Judge E			
		Stage 1	Dancability	S-coherence	F-coherence	Playability	Stage 1	Dancability	S-coherence	F-coherence	Playability	Stage 1	Dancability	S-coherence	F-coherence	Playability	Stage 1	Dancability	S-coherence	F-coherence	Playability
<i>Benchmark (B)</i>	12		4	5	4	5		3	3	4	4		4	5	4	3		5	5	5	5
	14		3	2	1	3		1	1	1	1		2	2	2	3		4	2	4	4
	45		3	2	3	3	R						2	2	2	3	R		4	1	3
	326		5	5	5	5		3	3	4	3		4	4	4	4		5	5	4	5
	433		4	3	4	4		2	1	1	2		3	3	2	3		5	3	4	5
	573		3	3	4	4	R						2	2	4	3		2	2	3	4
	633		5	5	5	5		2	2	2	2		4	3	3	3		5	5	5	5
	672		5	4	4	4		5	4	5	4		3	3	3	3		5	3	5	5
	736		3	3	3	4		2	1	2	2		4	3	3	4		2	2	4	4
	946		4	3	3	4		2	1	1	2		3	3	4	4		2	2	4	3
<i>Jönköping (J)</i>	116	R					R					R					R				
	291		1	1	1	3							1	1	1	1					
	339	R					R					R					R				
	387	M					R					M					R				
	493	M					M					M					M				
	592		1	1	1	1							1	1	1	2					
	596		3	3	3	2							3	3	3	3					
	676	M					R					R					M				
862	R					R						R					R				
894		1	1	1	1							1	1	1	2						
<i>Kalmar (K)</i>	8		3	1	1	3		1	1	1	2		1	1	1	3					
	111	M					M					M					R				
	161	R					R						2	4	4	4					
	441	M					M					M									
	464	M					M					M									
	506		3	1	3	3							2	1	1	2					
	556	M					R					M									
	804	M					M					M									
	932		4	3	3	4		2	1	1	2		2	2	2	3					
	955		3	3	4	4	R						3	4	4	4		R			
<i>Oskarshamn (O)</i>	101		3	3	3	3		4	3	3	3		3	3	1	2					
	247		3	2	2	3		3	3	4	3		2	2	2	3					
	283		2	2	3	3		2	2	3	3		2	2	3	3					
	368		3	2	3	4		2	2	2	1		2	2	3	3					
	403		3	2	2	4		3	2	3	3		2	3	3	2					
	411		2	2	2	3		2	2	3	2		3	2	3	2					
	473		2	2	2	3		3	3	4	4		2	2	2	3					
	602		3	3	4	4	R						2	2	2	3					
	636		4	3	3	4	R						2	3	2	2					
	639		4	3	3	4		2	2	3	2		2	2	2	2					
	23		5	4	5	5		5	5	5	5		3	4	5	4					
	28		4	3	4	3		4	3	3	3		3	3	3	3					
	43		4	4	4	4		4	4	4	3		4	4	3	4					
199		4	4	5	5		5	5	5	5		4	3	2	3						
328		3	3	4	5		4	4	4	3		4	3	3	4						
413		5	5	5	5		3	3	3	3		4	5	4	4						
505		4	4	3	4		4	4	4	3		3	4	3	3						
716		4	4	4	4		4	3	3	3		3	3	3	3						
979		4	4	4	5		1	1	1	1		4	4	5	4						
980		5	4	5	5		3	4	3	3		4	4	3	4						
<i>Växjö (V)</i>	108		4	3	3	4		3	3	3	3		3	3	2	3					
	117		5	5	5	5		4	4	4	4		3	4	5	3					
	263		5	4	5	5		4	4	3	4		3	4	5	3					
	267		5	5	5	5		5	5	5	5		4	5	5	4					
	463		4	4	3	4		2	2	2	2		3	3	2	3					
	553		4	4	4	4		5	4	5	5		3	5	5	4					
	576		5	4	5	5		5	4	5	5		4	4	5	4					
	738		5	5	5	5		5	5	5	5		4	5	5	4					
	751		5	4	4	4		5	5	5	5		4	5	4	4					
	900		4	4	4	5		4	4	4	4		3	4	2	3					
	Mean		3.65	3.19	3.46	3.92		3.24	2.97	3.24	3.13		2.83	3.10	2.96	3.10		2.78	2.88	3.63	3.76
Median		4	3	4	4		3	3	3	3		3	3	3	3		2	3	4	4	
Std. Dev.		1.12	1.20	1.24	1.01		1.28	1.33	1.34	1.21		0.95	1.19	1.27	0.81		1.35	1.00	1.39	1.04	

Table 2: Judge ratings of tunes generated by submitted systems summarised in Tab. 1. In Stage 1, tunes marked “M” means reject due to uncharacteristic meter; and “R” means reject due to uncharacteristic rhythm. Highlighted tune numbers are those elected by participants for evaluation. Highlighted ratings of a tune denote it was singled out by the judge as a favorite.

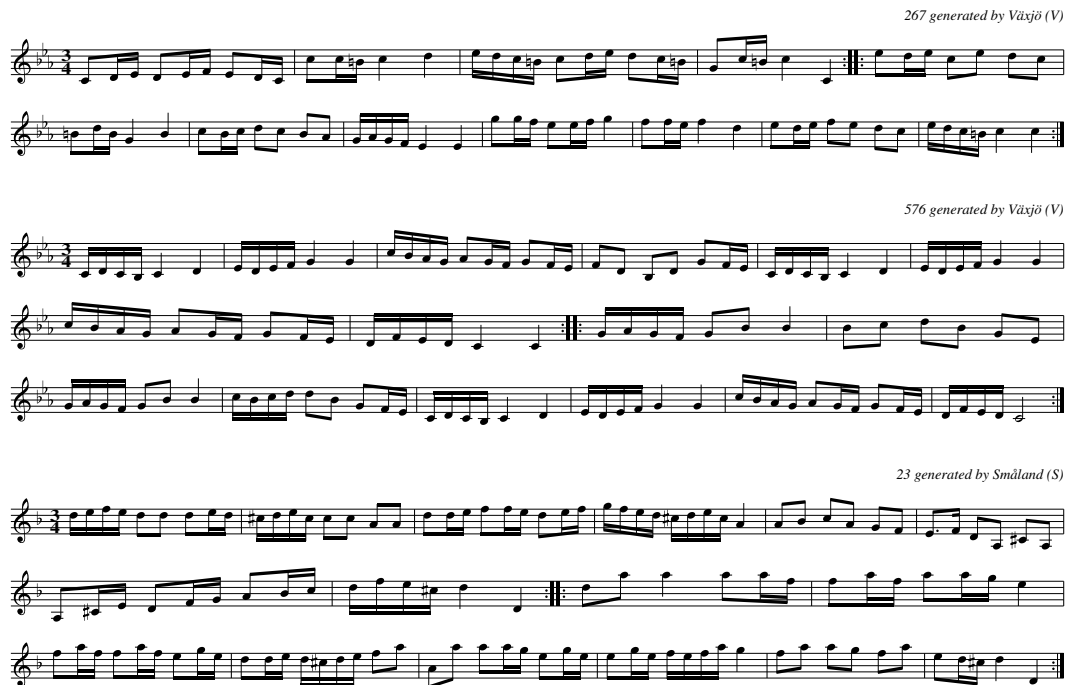


Figure 2: These three tunes were awarded first prize.

was close to the top position. During the in-person meeting of judges B, C and D, judge B elected four tunes, judge C elected six tunes, and judge D elected five tunes. The judges in general found the quality to be so high that identifying the best three tunes was very difficult. Tunes generated by system B garnered only two mentions, and those by S received five mentions; but those from V received fifteen. The judges found it very difficult to identify the single best tune, so through discussion and performance, the judges decided to award first prize to three tunes, and second prize to two tunes. We award honorable mentions to seven tunes elected by only one judge.<sup>6</sup>

Of the three tunes awarded first prize, two are generated by system V (267 and 576), and one is generated by S (23), shown in Fig. 2. For tune 267 by system V, judge A remarks, “This tune is longer than a lot of the tunes. Not so much repetition and the melody takes interesting detours. Going back to the Bb in measure 7 seems a bit sudden, but it still works in the context I think.” Judge B remarks, “‘Typical’ tonality. You get the feeling that you can find this tune in a *spelmansbok*. Easy to learn and remember.”<sup>7</sup> For tune 576 by system V, judge A remarks, “Not a typical slängpolska melodically but a good melody and rhythm.” Judge B remarks, “Gives nice energy for the dance. Nice tune.” Judge C remarks, “nice simple melody, very convincing, bra enkel låt.”<sup>8</sup> For tune 23 by system S, judge A remarks, “Interesting melody. A nice surprise with the C# in the end of the A-part.” Judge B remarks, “This tune feels authentic. However the last two notes in bar 7 could be a little problem for me.” Judge D remarks, “The Bb in the A-part measure 7 is unusual and not in style.” Judge E says, “second part does not match first”.

The judges awarded second prize to two tunes, both generated by system V (553 and 751), shown in Fig. 3. About 553 judge A remarks, “It’s a bit confusing in the beginning because you feel it as 4/4 but then it ‘resolves’ to 3/4 again. It makes it interesting when playing with the meter.” Judge C remarks, “simple but effective, coherent, enkel!” Judge D remarks, “Sounds like a Danish tune or a menuette. Cool.” Judge E remarks, “could be traditional!” About 751, judge A remarks, “Not a very common key. A part ends surprisingly on G.” Judge B remarks, “The repeat of the motif in bar 6 and 7 in the second part is nice. Nice polska. The tune feels authentic.” Judge C remarks, “Simple but coherent, Very repetitive but in a nice coherent way”. Judge D remarks, “More 1-3 feel in the rhythm

<sup>6</sup> All tunes awarded a prize can be found at the challenge website in footnote 3.

<sup>7</sup> A *spelmansbok* is a book of tunes for a folk music player.

<sup>8</sup> *bra enkel låt* means “good simple tune”.



Figure 3: These two tunes were awarded second prize.

but still even. Feels like a tune, maybe not the most brilliant one but still. Better played in D minor.”<sup>9</sup> Judge E remarks, “good! within the box, makes tonal sense – even a progression in the second part”.

We award “Honorable Mentions” to seven tunes selected as a favorite by only one judge. Three are generated by system V (117, 263, 738), two are generated by S (413, 980), and two others by B (12, 326). Judge D elected 12: “Nice tune. Sounds like a tune from Skåne. Or like a Quadrille but in 3/4”.<sup>10</sup> Judge E remarks that this tune is “close to plagiarism”, but does not specify the real tune that may be copied. Judge A elected 738, saying: “It’s nice that it’s in minor. I can hear this being played for dancing at a festival by a group of fiddlers. I like when the motif moves in the beginning of the B-part.” Judge A elected 326, saying: “Happy, cute and simple tune. I can hear a *spelmanslag* play this one. Nice scale motions and phrase endings.”<sup>11</sup>

The fourth and final stage of the 2021 challenge remains to be completed. This will involve a live performance of the top five generated tunes for dancers, who will then vote for their favorite tunes among them. This is scheduled to happen in November 2022 in Stockholm.

### 2.3 Statistical Analysis

Figure 4 shows the marginal distributions of ratings for all tunes, judges, systems, and qualities. The distributions for tunes clearly show large differences in performance between the systems, with B, S, and V generating tunes scoring highest. Judge E rated tunes mostly at either extreme of the scale (1 and 5) whereas judge C stayed mostly in the middle (2–4). It appears judge E usually rated structural coherence at the extremes, and gave the most top marks in playability. Judge A appears to be the most forgiving over all qualities, and judge C the least. Tunes generated by S and V show the highest frequency of top ratings across qualities, with those generated by B showing ratings more spread across all ratings. Finally, judges A and E clearly prefer tunes generated by systems B, S and V.

Figure 5 shows the parameter estimates for fixed effects models of quality ratings,  $y_{jqt} = \mu_q + \beta_{tq} + \beta_{jq} + \epsilon_{jqt}$ , where  $y_{jqt}$  is the rating by judge  $j$  in quality  $q$  for tune  $t$ ,  $\beta_{tq}$  is the effect of tune in quality  $q$ ,  $\beta_{jq}$  is the effect of judge in quality  $q$ ,  $\epsilon_{jqt}$  is the residual, and  $\mu_q$  is a reference value. In this case, we take judge A rating tune 14 by system B in the given quality as the reference. Tune 14 was rather poorly rated by all judges, and so it is no surprise that most fixed effects in each quality are positive. We also see that most fixed effects of judges in each quality are negative, also showing the forgiving grading of judge A. By and large, the fixed effects of the tunes awarded first (23, 267, 576), and second place (553, 751) are among the highest in all qualities. In all four qualities, we find significant differences between at least two judges ( $p < 0.017$ ). This is different from the 2020 challenge (Sturm and Maruri-Aguilar, 2021) in which we detect significant differences between judges in all qualities except for melody.

## 3 Discussion

System V (Casini and Sturm, 2022), using a transformer architecture, is clearly superior to the long short-term memory network used in both the benchmark (B) (Hallström et al., 2019), and the modified

<sup>9</sup>“1-3 feel” refers to a kind of polska that is missing the second beat.

<sup>10</sup>Skåne is a region in southern Sweden.

<sup>11</sup>A *spelmanslag* is an ensemble of traditional musicians.

benchmark using beam search and rejection sampling (S). The engineering of system V involved several months of tuning and analysis, with intermediate stages of qualitative evaluation considering the qualities of slängpolska. This appears to have paid off in terms of better quality output with fewer parameters than B and S. It is unfortunate that most tunes generated by J and K were rejected due to meter and rhythm criterion. Indeed, the elected tunes from these systems are not in the correct meter. A simple check for a 3/4 time signature would have addressed this problem.

During stage 3, judges B, C and D all remarked on general qualities of the tunes. Judge B remarked that they felt they could identify which tunes came from the same system, because they had a certain “personality.” Judges B and C mentioned a “British feeling” with some of the tunes. This coincides with the fact that training material for systems B, S and V includes Irish traditional music, with fine-tuning on Swedish slängpolskor. Judge B also remarked that the tunes they evaluated in this challenge are considerably more impressive than the AI-generated music they experienced in the 1980s in a research project at Uppsala University.

The use of notation, MIDI and sound files in the 2021 challenge resulted in a significant reduction of time spent judging. The judges approximated their total time spent in the first two stages as: 6 hours (B), 8 (C), 10 (A, D), and 12 (E). The judges in the 2020 challenge used about the same amount of time to evaluate 35 tunes (25 notated and 10 audio-rendered MIDI) (Sturm and Maruri-Aguilar, 2021). Further time could have been saved in 2021 if meter rejection was done before sending tunes to the judges (which would only involve looking at the time signature). Judges also mentioned that “playability” as a category was hard to evaluate, since it is related both to the tune and the ability of a performer. The suggestion is to omit this, similar to our finding from the 2020 challenge.

The focus on Irish and Swedish traditional music for the first two challenges comes from the overarching project of which they are part.<sup>12</sup> This project seeks to analyze the impacts of artificial intelligence technology on music practices, and vice versa, with specific case studies in Ireland and Sweden. When it comes to the chosen style for the 2021 challenge, this was made through consultation with one of the judges. The slängpolska is a common form of Swedish traditional dance music, which brings with it the availability of data and experts. As organizers, we did not provide any specific dataset for training, but did point to the folkwiki website, as well as the dataset used in Hallström et al. (2019). We wanted participants of this challenge to consider the entire machine learning pipeline for music style emulation, rather than just engineering a system for a given dataset that is already cleaned and formatted. This can make the challenge more time-consuming for a participant, and, together with the rather niche practice of traditional music, perhaps reduces the attraction of participating – unlike the more popular annual *AI Song Contest*.<sup>13</sup>

There are of course numerous ways to meaningfully evaluate music generation systems (Sturm and Ben-Tal, 2017; Ens and Pasquier, 2018; Sturm et al., 2018; Yin et al., 2021). *The Ai Music Generation Challenge* seeks to engage music practitioners of the specific tradition being computationally modeled, and to motivate participants to make use of this expert knowledge. The judges are concerned only with how well a tune put in front of them fits within their practice, regardless of what kinds of algorithms and data are used in its generation. The major contribution of the challenges is not the generated tunes themselves, or even the resulting systems, but the process of applying machine learning to living music traditions. Each challenge is meant to provide an organized and highly focused application area for participants, showcasing the use of experts in evaluating music generation systems.

*The Ai Music Generation Challenge 2022* returns to Irish traditional dance music, this time focused on the reel: a very popular form, commonly notated in 2/2 or 4/4, and often performed with a strong backbeat. *The Dance Music of Ireland: O’Neill’s 1001* (1907) contains 350 reels, and will serve as a collection for comparison. Unlike the previous challenges, there are three subchallenges for 2022: 1) build a system that generates the most plausible reels; 2) build an artificial judge that predicts the responses of the human judges; and 3) build a system that generates titles for given tunes. The first 2022 subchallenge will have three stages, evaluating ten tunes selected from each submission – one elected by each participant and nine sampled at random. The first stage will be rejection by plagiarism, rhythm, and mode/accidentals. The second stage will be evaluation along two qualities: structure and melody. The third stage will involve the election of top tunes by the judges. The third subchallenge will be evaluated by the judges as well, for the titling of award-winning tunes. The

---

<sup>12</sup>MUSAiC <https://musaiclab.wordpress.com>

<sup>13</sup><https://www.aisongcontest.com/>

second subchallenge will be evaluated by a yet-to-be-determined distance measure between the scores of the judge and the artificial judges averaged over all evaluated tunes.

## Acknowledgments and Disclosure of Funding

This challenge is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism).

## References

- Casini, L. and Sturm, B. L. T. (2022). Tradformer: A transformer model of traditional music transcriptions. In *Proc. Int. Joint Conf. Artificial Intell.*
- Ens, J. and Pasquier, P. (2018). A cross-domain analytic evaluation methodology for style imitation. In *Proc. Int. Conf. Computational Creativity*, Salamanca, Spain.
- Hallström, E., Mossmyr, S., Sturm, B. L., Vegeborn, V. H., and Wedin, J. (2019). From jigs and reels to schottisar och polskor: Generating Scandinavian-like folk music with deep recurrent networks. In *Proc. Sound and Music Computing Conf.*
- O’Neill, F. (1907). *The Dance Music of Ireland: O’Neill’s 1001*. Chicago.
- Sturm, B. L. and Ben-Tal, O. (2017). Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *J. Creative Music Systems*, 2(1).
- Sturm, B. L., Ben-Tal, O., Monaghan, U., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., and Pachet, F. (2018). Machine learning research that matters for music creation: A case study. *J. New Music Research*, 48(1):36–55.
- Sturm, B. L., Santos, J. F., Ben-Tal, O., and Korshunova, I. (2016). Music transcription modelling and composition using deep learning. In *Proc. Conf. Computer Simulation of Musical Creativity*, Huddersfield, UK.
- Sturm, B. L. T. and Maruri-Aguilar, H. (2021). The Ai Music Generation Challenge 2020: Double jigs in the style of O’Neill’s “1001”. *Journal of Creative Music Systems*.
- Yin, Z., Reuben, F., Stepney, S., and 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 *Proc. EvoMUSART*.

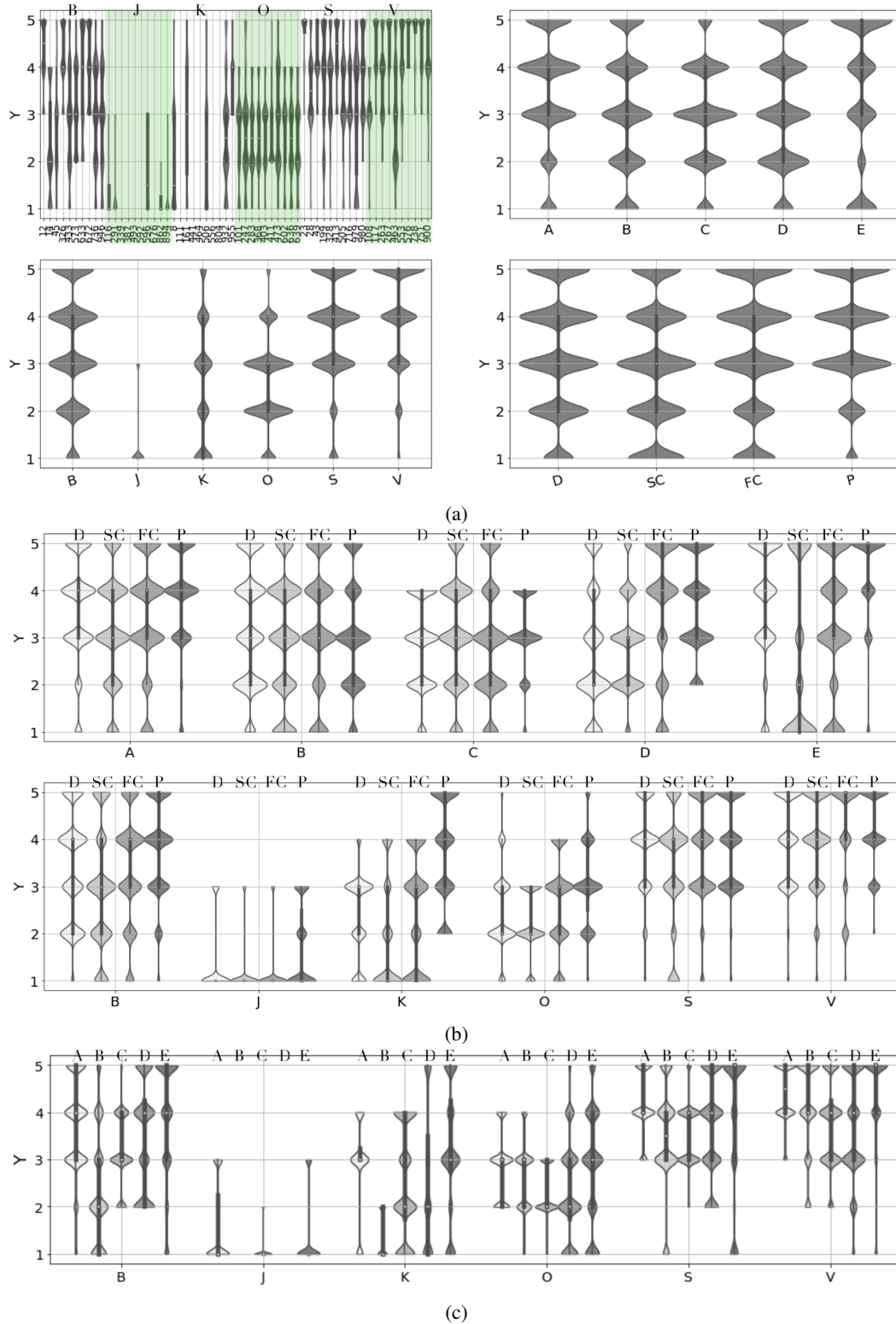


Figure 4: Violin plots of the marginalised ratings ( $Y$ ): (a) for tune (highlighting demarcates systems, labeled), judge, quality, and system (clockwise from top-left); (b) for joint ratings of quality (danceability (D), structural coherence (SC), formal coherence (FC), and playability (P)) and judge or system (top to bottom); (c) for joint rating of judge (A–E) and system. The width of each violin is scaled by the count in that bin.

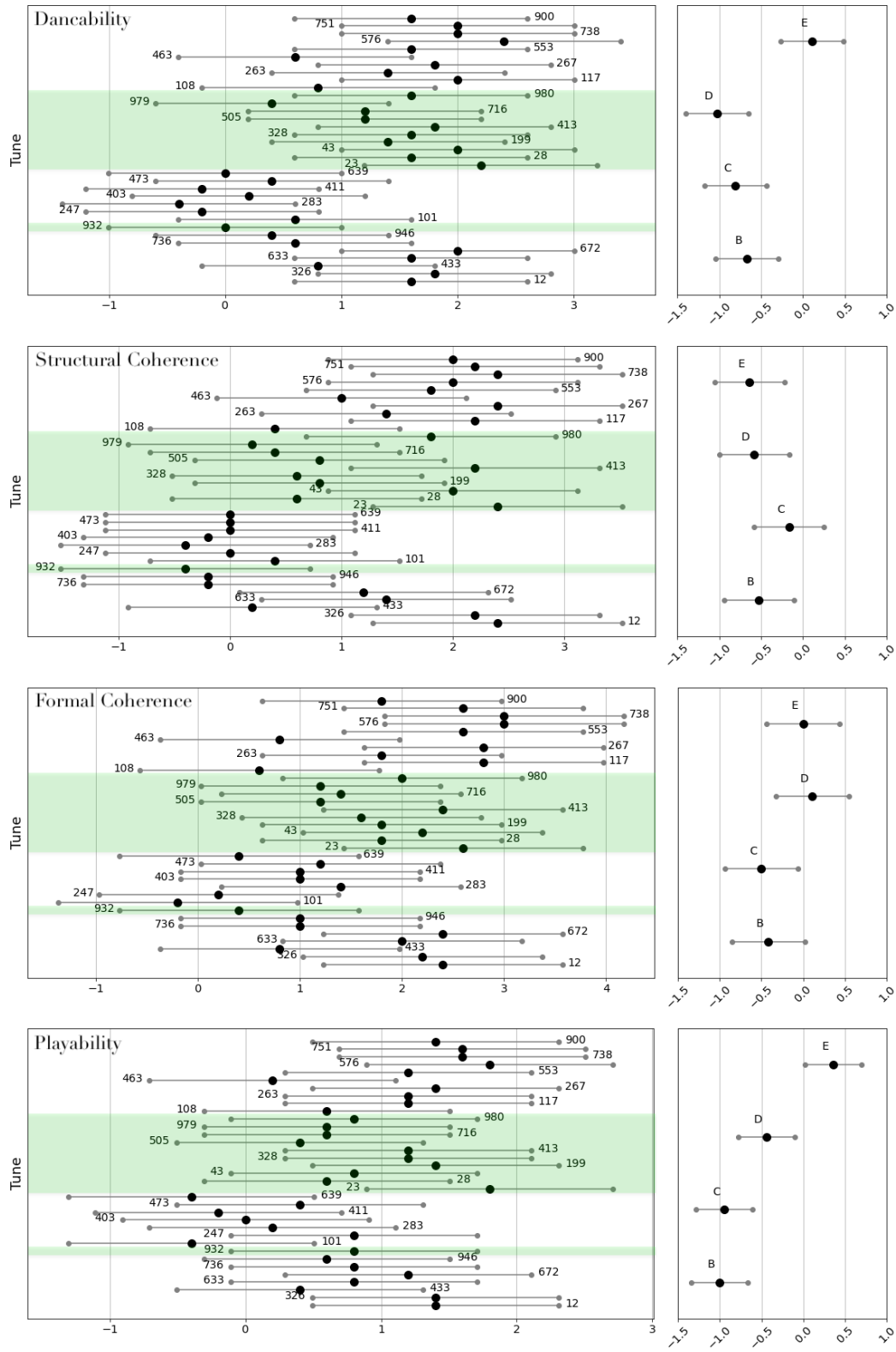


Figure 5: For all tunes passing the rejection criteria in Stage 1, estimates of the parameters of fixed effects models,  $y_{jqt} = \mu_q + \beta_{tq} + \beta_{jq} + \epsilon_{jqt}$ , and their 95% confidence intervals for tunes ( $\beta_{tq}$ , left) and judges ( $\beta_{jq}$ , right) with respect to tune 14 (B) and judge A in the each quality (labeled). Highlighted blocks in tune plots at left demarcate those generated by the same system, from bottom to top: B, K, O, S, and V.