

# Evaluating Low-Dimensional Latent Representations as a Creative Interface for Digital Synthesizers

**Matthew Peachey**

Faculty of Computer Science  
Dalhousie University  
Halifax, Canada  
peacheym@dal.ca

**Sageev Oore**

Faculty of Computer Science  
Dalhousie University  
Halifax, Canada  
sageev@dal.ca

**Joseph Malloch**

Faculty of Computer Science  
Dalhousie University  
Halifax, Canada  
jmalloch@dal.ca

## Abstract

Synthesizers are musical instruments that typically expose a large set of user-adjustable parameters that shape the sound of the instrument. In this paper, we examine how leveraging low-dimensional latent representations of synthesizer patches enables musicians to generate new patches via exploration of a generative model’s latent space. We evaluate two different latent representations, *Latent Coordinates* and *Timbral Representation*, through a mixed-methods user study (n=18). In our study, a mix of novice and experienced musicians engage with both sound matching and sound discovery tasks. Our qualitative results highlight a number of key themes regarding the suitability of our technique with each being supported through related quantitative results. Finally, these results indicate several ways in which this type of support tool may be used in a musician’s existing creative workflow as well as provide a context for discussing the benefits of continuous interactions for generative systems as it relates to creative activities.

## 1 Introduction

Synthesizers are musical instruments that generate audio waveforms using electrical signals and can produce a wide range of different sounds, or timbres, depending on their specific implementation. A synthesizer typically exposes a large number of user adjustable parameters, collectively referred to as a patch, that enables a musician to precisely shape the sound of their instrument. However, the complex and interconnected relationship between parameters and the resulting sound can make it difficult for new users to work with these instruments (Krekovic, 2019). A potential solution to this challenge is leveraging Generative Artificial Intelligence (GenAI), which itself is being increasingly used in the context of many creative activities such as writing (Chakrabarty et al., 2022; Huang and Tan, 2023) as well as visual art such as images and video (Rombach et al., 2022; Polyak et al., 2024), in order to co-create synthesizer patches with generative models.

Though GenAI continues to gain traction in both academia and industry, it is not without drawbacks when considered in the context of human creativity. For instance, text prompts are currently the de facto standard for interacting with contemporary generative models. However, in the context of creative work, literature highlights several limitations of this approach, particularly due to its reliance on discrete interactions (where a user synchronously exchanges information with a system). The limitations include a lack of artist control over the system (McCormack et al., 2023), a dependence on a trial and error for creativity (Dang et al., 2022), a reliance on natural language as an input modality (Akverdi and Baykal, 2024; McCormack et al., 2023) and putting a hindrance on creative ideation (Rajcic et al., 2024). Furthermore, Davis et al. (2024) highlights that it is also unfair to ask users to describe in language an output which they are not aware the system is capable of producing. Therefore, we believe that it is crucial to provide an alternative method for steering the output of a generative system that is based on continuous interactions (where uninterrupted user input is responded to by a system over a period of time (Doherty and Massink, 1999)), such that a human artist’s creative process and workflows are well supported when generating new patches with GenAI.

This paper presents a new method for exploring the sonic possibilities of a synthesizer that leverages continuous interaction and exploration of the latent space of a Variational Autoencoder trained on a corpus of synthesizer patches. This is achieved by integrating our previously introduced *Low-dimensional Latent Representations of Synthesizer Patches* (Peachey et al., 2023) into a lightweight, real-time, user interface that aims to fit within existing musical workflows. We evaluate our approach with a mixed-methods user-study and highlight the quantitative and qualitative results of this study. The results of this study provide a baseline context for discussing how the use of continuous interactions supports the creative process of musicians working with GenAI tool as well as provides a standalone support tool for exploring timbral capabilities of a musical synthesizer.

## 2 Related Work

### 2.1 Synthesizers and Timbre

Synthesizers available today use many different synthesis paradigms (e.g., subtractive, additive, FM, etc.) and form-factors (e.g., software, modular, semi-modular, standalone, etc.). However, one thing that most synthesizers have in common is the notion of a *patch*, which is the set of user-adjustable *parameters* used by musicians to precisely shape the sound of their instrument. The type of parameters available to a user, and subsequently the range of timbral options, are typically dependant on a synthesizer’s specific implementation details. The parameters exposed will typically range from simple control over pitch and amplitude of a waveform (as seen in very early or relatively simple synths) to hundreds of interconnected parameters in increasingly complex modern synthesizers.

Timbre is a term used in the fields of music & audio to describe a sound in terms of everything except its pitch and loudness (Wessel, 1979). Timbre is a concept that can be difficult to understand and describe, as there are many instances of different language or terminology used to describe the same timbral quality. For instance, Pearce et al. (2017) conducted a thorough review of timbre related literature, identifying over 1,000 different timbral descriptors. However, they were able to group these terms into eight core descriptors, noting that the most commonly identified timbral attributes are hardness, depth, and brightness (Pearce et al., 2017). These types of timbral terms are useful for musicians and sound-designers to describe what qualities they wish for a sound to possess, with phrases such as “*make that synth sound brighter*” or “*those drums need more depth*” being common place for those types of creative workflows. Software tools such as AudioCommons’ Timbral Models<sup>1</sup> exist for analysing audio clips in terms of these core timbral descriptors.

### 2.2 Generative Models

Generative Machine Learning models, much like unsupervised learning architectures, aim to learn an underlying structure of a dataset. However, generative models, as the name suggests, are uniquely capable of generating realistic data (such that new data could’ve been a plausible member of the original dataset) based on that learned structure (Foster, 2022). One such generative model is the Variational Auto Encoder (VAE), a model with the goal of learning a lower-dimensional representation, called a latent space, and then generating new realistic data by sampling from that latent space (Kingma and Welling, 2013).

VAEs feature an Encoder-Decoder architecture, where high-dimensional data is encoded to a latent space and then latent vectors (i.e. a point in the latent space) are decoded back into high-dimensional space. As shown in Figure 1, once a VAE is properly trained, a latent vector can be sampled and fed directly into the decoder network, generating a new instance of high-dimensional data of the expected form. Furthermore, a well formed latent space has several characteristics that when leveraged properly can help users navigate and explore possible generated outputs. For instance, data points that are similar to one another in the high-dimensional representation should be placed near each other in the latent space whereas those that are dissimilar should be positioned further apart. Furthermore, smoothly interpolating between two points in the latent space should also result in a smooth interpolation between high-dimensional data, enabling precise continuous control over a model’s outputs.

---

<sup>1</sup>[https://github.com/AudioCommons/timbral\\_models](https://github.com/AudioCommons/timbral_models)

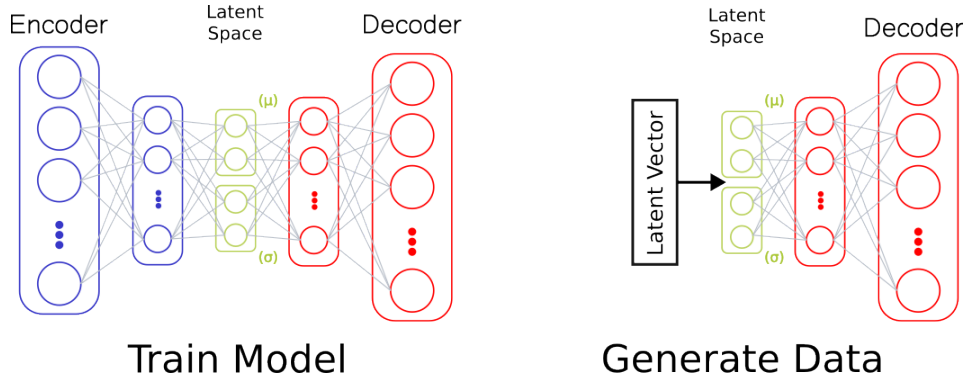


Figure 1: In a typical Variational Autoencoder architecture, an encoder neural network attempts to learn a low-dimensional representation of the statistical distribution (both mean and standard deviation) of a dataset. Conversely, a decoder neural network aims to reconstruct data in the same format based on the sampling of a latent-vector from the latent space. Once the model is trained by showing it many example data points, a new latent vector can be sampled from the latent space and passed into the decoder network, resulting in the generation of a realistic data point.

### 2.2.1 Generative Models for Musical Creativity

Generative models such as VAEs and Generative Adversarial Networks (GANs) have previously been applied to many aspects of research on musical creativity, including audio synthesis and music generation (Engel et al., 2019; Dhariwal et al., 2020; Bińkowski et al., 2019; Roberts et al., 2018a; Weber et al., 2019), drum sequence generation (Warren and Çamcı, 2022; Evans et al., 2024), learning musically meaningful sequences of MIDI notes (Roberts et al., 2018b), and in-painting musical pieces (Pati et al., 2019). Furthermore, Vaillant et al. (2021) has previously used VAEs with large latent spaces to assist with synthesizer patch programming and further extended their work by focusing on regularizing their VAE’s latent space based on timbre (Le Vaillant and Dutoit, 2024). Roche et al. (2021) also used timbre-based regularization of a VAE to implement perceptually relevant control over audio synthesis. While data availability is often a challenge faced by researchers, Vigliensoni et al. (2020) demonstrated how VAEs can be used to generate musical rhythms despite training their model on a dataset of limited size.

Leveraging direct manipulation of a latent space to support creative activities has also been previously explored. Murray-Browne and Tigas (2021) demonstrated this with *Latent Mappings* where they trained a VAE on motion-capture data and then mapped values of the latent space to parameters of a synthesizer such that when a dancer’s movement was fed back through their VAE’s encoder, that new latent vector would update their synthesis parameters (Murray-Browne and Tigas, 2021). Furthermore, other researchers have relied on interactive ML techniques as a mapping strategy between human performance and latent space exploration (Vigliensoni et al., 2023; Zheng et al., 2024). It has also been previously argued that leveraging explainable AI techniques to explain how latent spaces of generative models work under the hood is key to helping users of varying levels of musical expertise engage with generative models for musical creativity (Bryan-Kinns et al., 2022). Therefore, it is clear that engaging with generative models through exploration of a latent space (either directly or via a accompanying mapping strategy) is an important area of research at the intersection of GenAI and musical creativity.

Finally, we reflect on a number of previously run studies in the space of synthesizer patch creation as we designed our own evaluation plan. Yee-King and Roth (2008) evaluated the use of genetic algorithms for matching parameters to synthesizer patches using a sound matching evaluation (n=10) and Le Vaillant et al. (2020) explored graphical techniques for preset searching using interpolation (n=28) again using a set of sound matching activities. Finally, (Macret and Pasquier, 2014) also relied on a sound matching procedure for evaluating their PresetGen software (n=14). From these previous studies, we see that research is highly focused on evaluating techniques for helping musicians find a specific target sound, yet do not necessarily place an emphasis on exploration and discovery using their respective tools.

### 2.3 Latent Representations of Synthesizer Patches

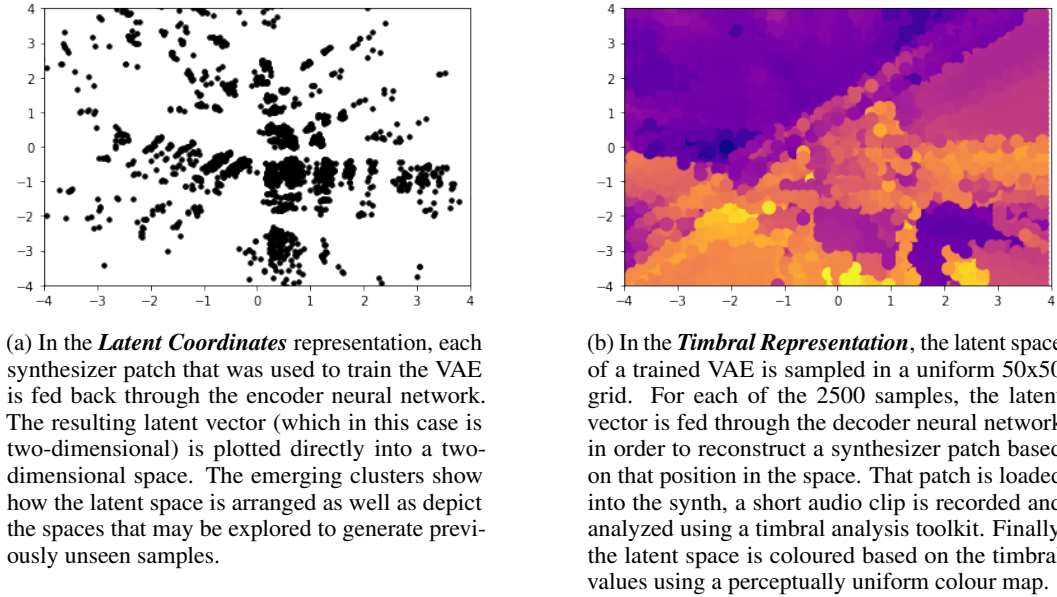


Figure 2: Examples of the two latent representations of synthesizer patches used in this study. In practice, the exact visualizations may change depending on the data used to train a VAE, but the procedure for deriving each representations remains consistent.

The focus of this paper is an evaluation of leveraging a VAE with a very low-dimensional (2D) latent space that has been trained on a collection of synthesizer patches as a support tool for generating synthesizer patches. We previously introduced two latent representations of synthesizer patches derived by training a VAE on patches for an open-source software subtractive synthesizer, AmSynth Dowell (2022), examples of which are shown in Figure 2 (Peachey et al., 2023). The first latent representation introduced is called *Latent Coordinates*, where the training set is projected back onto a 2D plane by encoding and plotting each data point as shown in Figure 2a. The second latent representation is called *Timbral Representation*, where the 2D latent space is uniformly sampled, with each sample being decoded into a synthesizer patch. Each of those decoded patches are used to record an audio clip whose timbral value is used to colour the latent space as shown in Figure 2b.

## 3 Methodology

We designed a within-subjects user study in order to evaluate the effectiveness of integrating two previously proposed latent representations of synthesizer patches into a real-time user-interface. Figure 3 depicts the system architecture of our GUI which was implemented as a web-application for this study. Our web-app sends real-time mouse event coordinates (formatted as a two-dimensional latent vector) to the VAE using WebSockets, at which time the VAE generates parameter values based on that input vector and automatically updates the synthesizer using the libmapper (Malloch et al., 2013) library integrated with a forked version of amSynth. Mouse click events are also sent via websockets and trigger a ‘note-on’ MIDI message which are also consumed by the synthesizer.

The objective of this study is to gather both quantitative and qualitative data in order to derive a full range of insights about using low-dimensional latent representations as part of a creative workflow. While we do collect and report on quantitative metrics as part of this study, the intention of our methodology is to ensure, as much as possible, that all users, regardless of previous experience with synthesizers, are able to engage with our system in the most similar way. This decision was made such that each of our participants could provide the most meaningful qualitative feedback regarding our approach, regardless of their previous experience, in order to better understand the usability of our system as well as improve future iterations of this work.

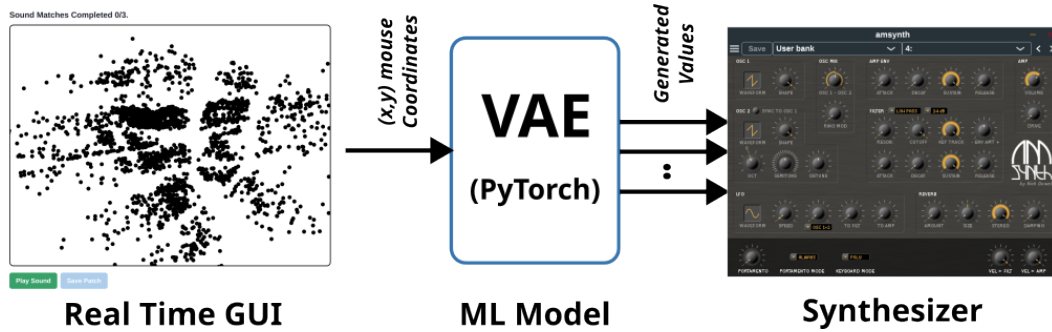


Figure 3: Low-Dimensional Latent Representations are integrated into a real time GUI and evaluated with a user study. A user’s mouse coordinates are captured and passed to the VAE via a web-socket connection which generates synthesizer parameter values based on that latent vector. Parameter values on the synthesizer are updated in real time via the libmapper library based on the output of the VAE. Finally, mouse click events are also communicated via web-sockets and trigger a ‘note-on’ MIDI message such that an external keyboard is not required.

### 3.1 Participants

We recruited eighteen participants for our user study through a combination of social media posts, university mailing lists, and direct recruitment. Because we were interested in understanding how musicians with varying levels of experience felt about our system, we did not impose an explicit inclusion requirement of any level of musical expertise. Rather, we began by recruiting from a general pool of potential participants and asked about their experience working with synthesizers and timbre as part of a “musical demographics” questionnaire at the end of our study. We chose this method of recruitment in order to run our study with a mix of novices and experienced users such that our results would be as generalizable to members of the research community focused on the intersection of music and HCI as possible.

In the context of this study, we consider a novice to be any participant that does not have any previous experience using or performing with a synthesizer nor experience in speaking or thinking about sound in terms of timbre. The final breakdown of our participant population includes 11 participants who self-identified as novices as well as 7 participants who identified having prior experience with either using synthesizers or interpreting timbral qualities of sound.

### 3.2 Study Tasks

Each participant was asked to complete a number of tasks during the session. The first task was an introductory training task in which participants were instructed to explore the interfaces for generating synthesizer patches, including both TIMBRAL REPRESENTATION and LATENT COORDINATES, as well as working directly with the synthesizer itself. During this self-guided training time, participants were encouraged to ask any questions about the generative system, but were told that questions about how the synthesizer’s inner working would not be answered. This training time helped users of all previous experience levels get comfortable with the different types of sounds that could be heard by generating new patches using the latent representations, as well as how to interact directly with the synthesizer UI.

The first study task was focused on sound matching, where a participant would listen to a target sound and their goal was to recreate that sound using each of the interfaces. The target sounds used for this task were comprised of both sounds taken directly from the latent space (i.e. able to be matched exactly using the latent representations) and from the synth UI (i.e. there was not necessarily an exact match to find). Furthermore, each sound was recorded using a single C4 note to avoid any pitch differences being perceived by participants, with C4 also being used for sound-matching via each interface. Participants were instructed to match three target sounds per interface, the order of which was counterbalanced according to a Latin-square to mitigate against the learning effect. This task was included to help us better understand the effectiveness of our system, in terms of both time to complete the task as well as users’ accuracy when matching sounds.

The second study task was focused on sound discovery, where there was no target sound and participants were instead encouraged to find sounds that were interesting or unique to them. In this task, participants were again asked to generate three sounds per each of the three interfaces, the order of which was again counterbalanced to mitigate against the learning effect. This task was included to allow users to experience being inspired by potentially surprising outcomes of the system, similar to how musicians are often inspired by ideation and exploration when working with their instruments.

### 3.3 Questionnaires and Interview

Once each participant had completed both the sound matching and sound discovery tasks, we moved on to a post session questionnaire and interview phase. We first asked the participants to fill out a brief musical demographics form which asked about their experience working with synthesizers or prior knowledge about timbre.

Next, we administered a custom questionnaire that was focused on gathering participant’s opinions about our system as well as their opinion about their performance working with that system. These questions are made available in Appendix A, are measured using a five point likert scale (numerical values 1-5), and are used to support the qualitative findings that we discuss in a later section. We also administer a modified Creativity Support Index (CSI) survey (Cherry and Latulipe, 2014) which, similar to both Wastnidge and Erdem (2024) and Tchemeube et al. (2023), we removed the two questions associated with collaboration, as well as the paired-factors components of the traditional CSI. This was done to prevent our users being skewed by seemingly unrelated questions as our tool does not rely on any collaboration between users. We administered two copies of the modified CSI, one for each latent representation, such that we could report on each representation individually.

Finally, we administered a semi-structured interview consisting of six questions which are listed in Appendix C. These questions were intentionally open-ended in an attempt to capture as many qualitative thoughts about the systems for our participants as possible. The interview lasted approximately 10 minutes for each participant, but final times varied depending on how the conversation with a participant naturally flowed. Audio recordings and transcriptions were taken for these interviews, with key points being extracted as codes for thematic analysis which is reported on in the following section.

## 4 Results

As previously stated, the primary objective of this study is to derive qualitative insights about the use of low-dimensional representations for generating high-dimensional synthesizer patches. Therefore, we present the following results of a thematic analysis of our qualitative data, with each key theme being supported by the associated quantitative results (Appendix B) we gather from participants.

### 4.1 Quality of Generated Patches

The first main theme to result from our qualitative analysis concerned the quality of the patches generated by the VAE. A large number of responses addressed the quality of the generated patches, revealing mixed feelings among the group of participants. Firstly, many participants expressed that that the quality of the patches were good, e.g.:

*“The quality was good, I think. Not quite to the level of other professional-grade audio plugins, but still good.” (P8)*

*“Compared to the patches I have experience with, they were of the same quality.” (P17)*

*“The quality from the VAE was similar, if not the same, as the quality from the synthesizer.” (P1)*

Participants also observed the variety of sounds that could be generated while leveraging the low-dimensional representations as part of a creative workflow, e.g.:

*“The quality was very good, I tried many different types of sounds.” (P5)*

*“I thought there was good representation of [available] sounds.” (P16)*

*“There was a mix of sounds.” (P3)*

However, participants also commented on the fact that the low-dimensional latent representations limited the total range of possible sounds. Furthermore, some participants commented that the quality of the sounds were not satisfactory, e.g.:

*“There were only so many sounds, most were fairly basic. It was a compressed representation.” (P7)*

*“Some sounds were not so usable, it varied across the space.” (P10)*

*“Some of them were quite chaotic.” (P18)*

This result may be explained due to the compressed nature of the low-dimensional representations. However, we see that even when sounds are of lower quality, participants comment on them being better than the noise which could be generated by simply randomizing the parameters of a synthesizer (which would essentially be the effect of sampling from a latent space of an untrained VAE).

The qualitative results regarding the sound quality of generated patches are supported by quantitative results from the questionnaire. We observe that most participants found there to be a wide range of timbres available via the latent space ( $\mu = 4.5$ ,  $\sigma = 0.62$ ). Furthermore, of this wide range of timbres, participants stated that they found them to be at least somewhat musically pleasing ( $\mu = 3.44$ ,  $\sigma = 0.98$ ), with only a single participant strongly disagreeing. Therefore, we conclude that a low-dimensional representation of synthesizer patches is sufficient to generate musically interesting patches. However, we also acknowledge the opportunity to improve on this model, either through the use of a latent space with more dimensions or through more diverse training data in order to generate a wider range of timbral outputs for users to explore.

## 4.2 Challenges working with Synthesizers

The next main theme that was identified speaks to the challenges that users faced when working with synthesizers and how the low-dimensional representations worked towards mitigating these challenges. Firstly, participants commented on how synthesizers can be confusing to work with due to the complexity of their interfaces, e.g.:

*“I usually find the standard synths to be very user unfriendly and often counter intuitive.” (P17)*

*“The labels on the synth UI did not communicate what they did.” (P15)*

*“I knew I wasn’t going to be able to use the synthesizer, it was too complicated when I first saw it.” (P4)*

Secondly, participants commented on the lengthy time requirement that is needed to become proficient with these instruments, e.g.:

*“Knowing the Synth takes some time. Using it to match sounds takes a lot of effort.” (P5)*

*“I’d need like hours at this desk to figure out how to work the synthesizer.” (P7)*

Finally, participants also noted that gaining a level of proficiency with the instrument would naturally make for a better creative experience, but that the ability to quickly find sounds that they were interested in was a major benefit provided by our system, e.g.:

*“I think if someone is a professional they’re prefer this one [standard synth], but I think the [latent representation] will be easier for me and so I’d use that one.” (P15)*



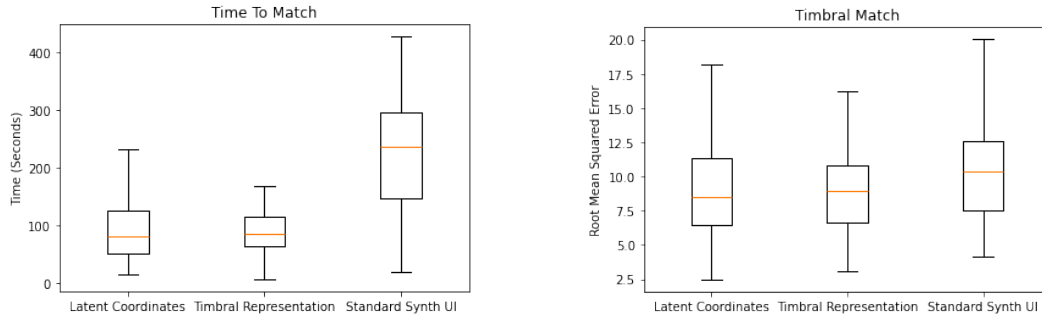
“Someone who is used to creating voices may rather use [the synth UI], but for me because I didn’t know how [to use the synth UI], the latent representation was better.” (P4)

“It will help people who have no experience with a synthesizer because you can quickly create different sounds.” (P14)

“I was able to explore more sounds than with the synthesizer.” (P12)

We see from these participant responses that it can be challenging to begin working with synthesizers due to the complexity of the relationship between parameters and the time commitment required to become proficient with them. These findings are supported by the questionnaire results: when asked if they felt that they did a good job finding and matching target sounds using the synthesizer UI directly, most tended to disagree ( $\mu = 2.5$ ,  $\sigma = 0.99$ ). In contrast, participants tended to agree that they did a good job of the same task using the latent representations ( $\mu = 3.33$ ,  $\sigma = 0.91$ ).

Interestingly, results differ between the sound-discovery task and the sound-matching task. When asked again if they felt they had done a good job discovering new sounds, participants tended to agree for both of the latent representations ( $\mu = 4.05$ ,  $\sigma = 0.80$ ), as well as the standard AmSynth UI ( $\mu = 3.61$ ,  $\sigma = 0.92$ ). This result may be due to the fact that participants were likely to be more satisfied with the sounds they were able to produce, regardless of representation, when not comparing it to a target sound but rather focusing on what was interesting to them.



(a) *Time-to-match* for each representation as measured by the web-based study environment. While the two latent representations take approximately the same amount of time, both are faster than when users interacted with the standard synthesizer interface.

(b) *Timbral-match* for each representation as measured by the AudioCommons’ Timbral Models and averaged across each timbral characteristic. The latent representations on average perform better than the standard synthesizer UI, but by a much smaller margin than the *time-to-match* metric.

Figure 4: Box-plots showing the results for *time-to-match* and *timbral-match* for both latent representations as well as the standard synthesizer interface.

We measured the completion time for each task (*time-to-match*) as well as the accuracy of the matched patch in terms of timbral qualities (*timbral-match*) using the AudioCommons’ Timbral Models using Root Mean Squared Error (RMSE) averaged across each timbral characteristic. We used a Repeated Measures Analysis of Variance (ANOVA), which according to Norman (2010) is valid even for small sample sizes, to test for statistical significance when comparing *time-to-match* and *timbral-match* across our latent representations and AmSynth’s standard interface. Results show that there is a significant difference when considering time-to-match ( $F(2, 30) = 28.65$ ,  $p < .001$ ,  $\eta^2 = .46$ ), with post-hoc analysis using Bonferroni correction showing a significant difference between LATENT COORDINATES and synth interface ( $p < .001$ ) as well as TIMBRAL REPRESENTATION and synth interface ( $p < .001$ ) but with no statistical difference between latent representations. However, for the timbral match metric we do not see a statistical difference between latent representations and the synth interface. Furthermore, we use a Multivariate Analysis of Variance (MANOVA) test to compare the combined effect on *time-to-match* and *timbral-match*. That result showed a significant multivariate effect (Wilks’  $\Lambda = .255$ ,  $F(2, 140) = 204.49$ ,  $p < .001$ ) with post-hoc comparison via uni-variate ANOVA confirming that *time-to-match* differed significantly while *timbral-match* did not.



### 4.3 Leveraging the Structure of the Latent Space

Another important theme we identified was how participants observed clusters in the latent space. As we’ve previously discussed, it was an explicit decision to use low-dimensional latent spaces for this research, both to understand the limits of the VAE architecture as well as to promote efficient visualization and direct navigation throughout the latent space. Participants were not explicitly informed about clusters emerging in the space, yet is an area of user reflection that is important to discuss, e.g.:

*“When matching sounds, I was able to navigate through regions to find a sound that was close to the target sound” (P3)*

When discussing the LATENT COORDINATES representation. Other participants also observed the benefits of sounds in the latent space being clustered, e.g.:

*“dots that were closer to one another produced kind of similar sounds.” (P3)*

*“I was picking up on these different little clouds where you can expect types of sounds.” (P10)*

*“I could find a cluster of sounds that I liked in general, but struggled to find a precise sound within that cluster.” (P16)*

The structure of the latent space presented the opportunity to integrate continuous interactions into our real-time GUI. Users quickly picked up on the fact that clusters in the space were often associated with patches/timbres that were similar to one another and that by making small movements in the space they could explore the small differences in patches as encoded by the VAE. One participant explicitly talked about how changes to the sound, as triggered by exploration of the latent space, served as a primary inspiration to them:

*“I found myself performing based on live changes to the sound rather than any discrete point in the LS” (P13)*

Users reported that they found visual feedback for each representation to be helpful when using the tool. Participants responded positively for both LATENT COORDINATES ( $\mu = 3.94, \sigma = 1.06$ ) and TIMBRAL REPRESENTATION ( $\mu = 3.94, \sigma = 0.87$ ), indicating that the visual feed back about the structure of the space was helpful. This result is encouraging with respect to future design iterations that continue to leverage a continuous interaction first approach to working with generative models in creative contexts.

### 4.4 User Preference of Synthesizer Patch Representation

Another key theme reflects on users’ preferences with respect to using each Latent Representation when creating synthesizer patches. Responses varied among participants, but analysis suggests that each representation was well suited for participants with certain ways of thinking about the system. For instance, with regards to the Timbral Representation, participants spoke to how the colouring of the latent space guided their attention to certain regions, including areas of the latent space that may have been interpreted as ‘blank’ in the LATENT COORDINATES representation, e.g.:

*“In Timbral Representation you could know the difference between sounds because of the colours even though I didn’t know the exact mapping. I knew bright colour meant high something and dark colour meant low something. (P6)”*

*“I preferred the Timbral Representation because there is another dimension that helps me see a pattern. Colours of the same hue were producing similar sounds.” (P18)*

In similar fashion, other participants offered insights into the LATENT COORDINATES representation and their perceived benefits of being able to view and explore throughout the clusters that emerged in the latent space, e.g.:

*“I liked the latent coordinates, I found the clusters were more representative of the sounds that were produced.” (P16)*

*“I preferred the Latent Coordinates because the clustering was hugely helpful.” (P8)*

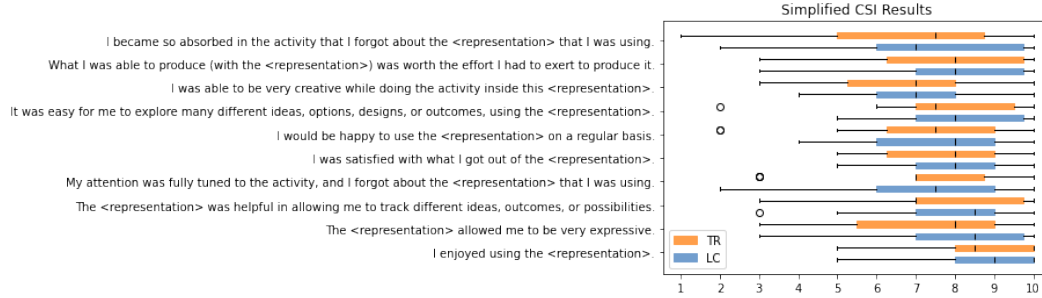


Figure 5: Results of the modified CSI for both of the timbral representations. Orange boxes represent the average responses per question for the TIMBRAL REPRESENTATION; blue boxes represent the same for LATENT COORDINATES. We calculated that on average the Latent Coordinates representation received a better score from participants when summed across all 10 modified CSI questions. In total, the Latent Coordinates scored a 77.56 whereas the Timbral Representation scored 74.61, suggesting that participants had a slight preference for the former in terms of creativity support.

These participant insights with respect to the benefits of each representation are supported by the results of the modified CSI that was administered to each participant. As shown in Figure 5, while both representations performed well overall, a slight edge was given to the LATENT COORDINATES representation (achieving a score of 77.56) over the TIMBRAL REPRESENTATIONS (achieving score of 74.61) with respect to creativity support. These results suggest that a user’s preference may be based on previous experience or expectations, rather than any intrinsic quality of either latent representation.

#### 4.5 Using Latent Representations as a Support Tool

The last major theme that was identified concerns participants’ feelings about incorporating this kind of tool into their own creative workflows. Firstly, participants reflected on how the latent representation based interface enabled them to find sounds that they found interesting, e.g.:

*“I would use this as a look similar to a colour-palette selector when working on web-development. I would use it like I was finding a general colour i was interested in, and then slowly tweaking the sliders to find the exact shade I wanted.” (P1)*

*“I would use this to explore sounds, find a sound, and then use that sound with the whole keyboard.” (P16)*

Secondly, participants discussed how this kind of interface worked well as a method for being inspired by new sounds or variants of sounds they had already heard, e.g.:

*“I would use a tool like this as an inspiration piece, for playing a neat role of breaking out of the box.” (P10)*

*“I’d use this tool as a better way to adventure off towards variants of a given sound.” (P8)*

Overall, participants determined, for a variety of individual reasons, that they would use a tool like this when engaging with synthesizers. This result is supported by the questionnaire response where we observe that participants tend to agree that this type of tool has a place as part of their own musical workflow ( $\mu = 3.94, \sigma = 0.54$ ). Therefore, we expect users with varying creative workflows to be able to find a place for a tool such as this when making music with synthesizers.

## 5 Conclusion & Future Work

In this paper, we present the a mixed-methods user study that aims to evaluate the use of low-dimensional latent representations of synthesizer patches when integrated into a real-time creativity support tool. In our study, 18 participants engaged with both sound matching and sound discovery tasks before providing quantitative and qualitative feedback, the output of which resulted in a number of key themes and supporting metrics that speak to the usability of our system.

We have identified a number of limitations for this research. First, we intentionally chose to use a VAE architecture that relies on a very low-dimensional (2D) latent space, rather than using dimensionality reduction techniques to facilitate exploration of a larger dimensional latent space as has been previously explored in other work. Secondly, we acknowledge the challenge in presenting quantitative results from a population that includes both novices and experienced users, as less experience users would be expected to take longer to fully understand the sonic options of a synthesizer than users who have worked with that instrument before.

There are several important aspects of future work that will continue the development of latent representations as a creative support tool for synthesizer patch generation. Specifically, we will explore additional VAE architectures that leverage a larger dimensional (yet still relatively small compared to the parameter space of a chosen synthesizer) latent space with the expectation that this model will generate higher quality and more diverse patches. It will therefore also be important to develop visualizations that effectively represent the state of the instrument, including the region of the latent space that a user is currently exploring, such that users of all levels of expertise can engage with future iterations of latent representations in a way that compliments their existing creative workflow. Furthermore, integrating our approach into DAW based tools will allow us to continue exploring real-life creative workflows of practicing musicians.

Finally, we will continue to explore more broadly how the concept of continuous interactions can be applied to the rapidly expanding field of GenAI, especially in the context of creativity. As creativity support tools that leverage generative systems continue to become more prevalent, we believe it is important to design interaction methods that not only enable human performance, but celebrates and encourages the differences in ideas, styles and techniques that are fundamental to the human creative experience.

## 6 Ethics Statement

This research has been completed with approval from Dalhousie University’s Research Ethics Board (REB File #: 2023-6769). Furthermore, the latent representations of synthesizers that are used in this research are derived from the latent space of a VAE trained on a corpus of synthesizer patches that were included with the open-source synthesizer AmSynth. We give credit to the contributors of the AmSynth project and creators of the original patches that were used to train that VAE. Finally, this research has been funded by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-03845) and we thank them for their support.

## References

- Akverdi, C. and Baykal, G. E. (2024). Generative ai tools in design fields: Opportunities and challenges in the ideation process. In *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction*, NordiCHI ’24 Adjunct, New York, NY, USA. Association for Computing Machinery.
- Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K. (2019). High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*.
- Bryan-Kinns, N., Banar, B., Ford, C., Reed, C., Zhang, Y., Colton, S., Armitage, J., et al. (2022). Exploring xai for the arts: Explaining latent space in generative music.
- Chakrabarty, T., Padmakumar, V., and He, H. (2022). Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*.

- Cherry, E. and Latulipe, C. (2014). Quantifying the creativity support of digital tools through the creativity support index. *ACM Trans. Comput.-Hum. Interact.*, 21(4).
- Dang, H., Mecke, L., Lehmann, F., Goller, S., and Buschek, D. (2022). How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.
- Davis, R. L., Wambsganss, T., Jiang, W., Kim, K. G., Käser, T., and Dillenbourg, P. (2024). Fashioning creative expertise with generative ai: Graphical interfaces for design space exploration better support ideation than text prompts. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.
- Doherty, G. and Massink, M. (1999). Continuous interaction and human control. In *Proceedings of the XVIII European annual conference on human decision making and manual control*, pages 80–96.
- Dowell, N. (2022). amsynth. <https://github.com/amsynth/amsynth>.
- Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.
- Evans, N., Haki, B., and Jordà Puig, S. (2024). Groovetransformer: a generative drum sequencer eurorack module.
- Foster, D. (2022). *Generative deep learning*. "O'Reilly Media, Inc."
- Huang, J. and Tan, M. (2023). The role of chatgpt in scientific communication: writing better scientific review articles. *American journal of cancer research*, 13(4):1148.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Krekovic, G. (2019). Insights in habits and attitudes regard-ing programming sound synthesizers: A quantitative study. In *Proceedings of the 16th Sound and Music Computing Conference*.
- Le Vaillant, G. and Dutoit, T. (2024). Latent space interpolation of synthesizer parameters using timbre-regularized auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Le Vaillant, G., Dutoit, T., and Giot, R. (2020). Analytic vs. holistic approaches for the live search of sound presets using graphical interpolation. In *NIME*, pages 227–232.
- Macret, M. and Pasquier, P. (2014). Automatic design of sound synthesizers as pure data patches using coevolutionary mixed-typed cartesian genetic programming. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pages 309–316.
- Malloch, J., Sinclair, S., and Wanderley, M. M. (2013). Libmapper: (a library for connecting things). In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 3087–3090.
- McCormack, J., Cruz Gambardella, C., Rajcic, N., Krol, S. J., Llano, M. T., and Yang, M. (2023). Is writing prompts really making art? In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pages 196–211. Springer.
- Murray-Browne, T. and Tigas, P. (2021). Latent mappings: Generating open-ended expressive mappings using variational autoencoders. In *NIME 2021*. PubPub.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, 15:625–632.
- Pati, A., Lerch, A., and Hadjeres, G. (2019). Learning to Traverse Latent Spaces for Musical Score Inpainting. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 343–351, Delft, The Netherlands. ISMIR.

- Peachey, M., Oore, S., and Malloch, J. (2023). Creating latent representations of synthesizer patches using variational autoencoders. In *2023 4th International Symposium on the Internet of Sounds*, pages 1–7. IEEE.
- Pearce, A., Brookes, T., and Mason, R. (2017). Timbral attributes for sound effect library searching. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society.
- Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.-Y., Chuang, C.-Y., Yan, D., Choudhary, D., Wang, D., Sethi, G., Pang, G., Ma, H., Misra, I., Hou, J., Wang, J., Jagadeesh, K., Li, K., Zhang, L., Singh, M., Williamson, M., Le, M., Yu, M., Singh, M. K., Zhang, P., Vajda, P., Duval, Q., Girdhar, R., Sumbaly, R., Rambhatla, S. S., Tsai, S., Azadi, S., Datta, S., Chen, S., Bell, S., Ramaswamy, S., Sheynin, S., Bhattacharya, S., Motwani, S., Xu, T., Li, T., Hou, T., Hsu, W.-N., Yin, X., Dai, X., Taigman, Y., Luo, Y., Liu, Y.-C., Wu, Y.-C., Zhao, Y., Kirstain, Y., He, Z., He, Z., Pumarola, A., Thabet, A., Sanakoyeu, A., Mallya, A., Guo, B., Araya, B., Kerr, B., Wood, C., Liu, C., Peng, C., Vengertsev, D., Schonfeld, E., Blanchard, E., Juefei-Xu, F., Nord, F., Liang, J., Hoffman, J., Kohler, J., Fire, K., Sivakumar, K., Chen, L., Yu, L., Gao, L., Georgopoulos, M., Moritz, R., Sampson, S. K., Li, S., Parmeggiani, S., Fine, S., Fowler, T., Petrovic, V., and Du, Y. (2024). Movie gen: A cast of media foundation models.
- Rajcic, N., Llano Rodriguez, M. T., and McCormack, J. (2024). Towards a diffractive analysis of prompt-based generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. (2018a). A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, pages 4364–4373. PMLR.
- Roberts, A., Engel, J., Raffel, C., Simon, I., and Hawthorne, C. (2018b). Musicvae: Creating a palette for musical scores with machine learning.
- Roche, F., Hueber, T., Garnier, M., Limier, S., and Girin, L. (2021). Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 4:52–66.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Tchemeube, R. B., Ens, J., Plut, C., Pasquier, P., Safi, M., Grabit, Y., and Rolland, J.-B. (2023). Evaluating human-ai interaction via usability, user experience and acceptance measures for mmm-c: a creative ai system for music composition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*.
- Vaillant, G. L., Dutoit, T., and Dekeyser, S. (2021). Improving synthesizer programming from variational autoencoders latent space. In *2021 24th International Conference on Digital Audio Effects (DAFx)*, pages 276–283.
- Viglienconi, G., Fiebrink, R., et al. (2023). Steering latent audio models through interactive machine learning.
- Viglienconi, G., McCallum, L., and Fiebrink, R. (2020). Creating latent spaces for modern music genre rhythms using minimal training data. *International Conference on Computational Creativity (ICCC'20)*.
- Warren, N. and Çamcı, A. (2022). Latent drummer: A new abstraction for modular sequencers.
- Wastnidge, A. and Erdem, C. (2024). Deep Steps: A Generative AI Step Sequencer. *AIMC 2024 (09/09 - 11/09)*. <https://aimc2024.pubpub.org/pub/odrhfyym>.

- Weber, A., Alegre, L. N., Tørresen, J., and Castro da Silva, B. (2019). Parameterized melody generation with autoencoders and temporally-consistent noise. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 174–179. Universidade Federal do Rio Grande do Sul.
- Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer music journal*, pages 45–52.
- Yee-King, M. and Roth, M. (2008). Synthbot: An unsupervised software synthesizer programmer. In *ICMC*.
- Zheng, S., Del Sette, B. M., Saitis, C., Xambo Sedo, A., Bryan-Kinns, N., et al. (2024). Building sketch-to-sound mapping with unsupervised feature extraction and interactive machine learning.

## Appendix A | Custom Questionnaire

Question
1. I found the patches I generated to be musically pleasing?
2. I did a good job finding & matching target sounds using the latent space explorer?
3. I did a good job finding sounds new sounds using the latent space explorer?
4. I found there was a wide range of timbres to explore throughout the latent space?
5. I found the visual feedback of Latent Coordinates to be helpful when searching for sounds?
6. I found the visual feedback of Timbral Representation to be helpful when searching for sounds?
7. I did a good job finding & matching target sounds using the standard synthesizer UI?
8. I did a good job finding new sounds using the standard synthesizer UI?
9. I found the timbres discovered throughout the latent space to be similar to one another?
10. I would use this type of tool as part of my own synthesizer workflow.

## Appendix B | Questionnaire Results

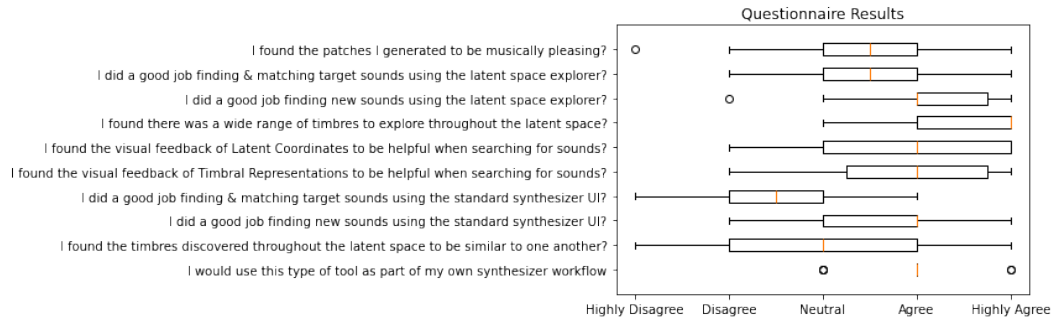


Figure 6: Results for our custom questionnaire computed for all eighteen participants. This questionnaire was measured via a five point likert scale with numeric values ranging from 1-5. We observe that participants tended to respond that they found there to be a wide range of timbres to explore throughout the latent space, and of that many were musically pleasing. Most importantly, we see that participants typically agreed that they would use this type of tool as part of their own musical workflow.

## Appendix C | Semi-structured Interview Questions

Question
1. What do you think about the quality of synth patches you generated using the VAE-based system?
2. Did your initial expectations about the synth and/or latent space match the real outcomes?
3. Did you prefer searching the latent space for matching sounds or for finding new sounds?
4. Do you have a preference of the visualizations used in the latent space explorer? If so, what did you prefer?
5. Can you see room for a tool like this in your musical workflow?
6. Is there anything else about the session you would like to comment on?