Latent Terrain: Adapting Neural Audio Autoencoders as Design Materials in NIME

Shuoyang Jasper Zheng
Centre for Digital Music, Queen Mary University of London
London, UK
shuoyang.zheng@qmul.ac.uk

Keigo Yoshida
Graduate School of Media and Governance, Keio University
Kanagawa, Japan
Keigoyoshida@keio.jp

Nico García-Peguinho
Centre for Digital Music, Queen Mary University of London
London, UK
n.a.garcia-peguinho@qmul.ac.uk

Jiatong Liu
Creative Computing Institute, University of the Arts London
London, UK
jiatong.liu37@outlook.com

Dan Hearn
Creative Computing Institute, University of the Arts London
London, UK
d.hearn@arts.ac.uk

Anna Xambó Sedó
Centre for Digital Music, Queen Mary University of London
London, UK
a.xambosedo@qmul.ac.uk

Nick Bryan-Kinns
Creative Computing Institute, University of the Arts London
London, UK
n.bryankinns@arts.ac.uk

Abstract

Neural audio autoencoders, a deep learning method for sound synthesis, are increasingly popular in AI-enhanced NIMEs. It is timely to explore as a community how this technological opportunity has opened a domain of design in NIME. This paper focuses on one compelling technique of using autoencoders for sound synthesis: navigating their latent space as a generative sound space. We introduce *Latent Terrain*, a Max/MSP tool package designed to tailor latent spaces into corpus-based sound spaces for NIMEs. We describe the rationale and development process of Latent Terrain to offer insights into the use of autoencoders in a material-oriented crafting space of musical interface design. We present an annotated portfolio resulting from a collaborative artistic exploration of Latent Terrain with four NIME makers, to showcase the design possibilities opened by autoencoders. We reflect on our practice-based account to discuss the challenges and opportunities of enabling neural audio autoencoders as design materials for AI-enhanced NIMEs.

Keywords

Autoencoder, Neural Audio Codec, AI as Material, Digital Musical Instrument, Machine Learning, Practice Research

1 Introduction

Neural audio autoencoders (or neural audio codecs), a deep learning method for sound synthesis [16, 22], have become increasingly popular in developing AI-enhanced sound and music applications [18, 23]. Research in Music Information Retrieval yields models that can effectively learn compact latent representations from audio for sound analysis and synthesis [51]. In the field of NIME, such a unique affordance is enabling new ways of approaching musical interaction design [79, 95]. Neural audio autoencoders are increasingly engaged as interactive, exploratory, and expressive computational materials [36, 47, 54, 55, 65, 76, 87], expanding the ways of using and appropriating AI in musical interfaces. Given the multifaceted challenges and opportunities of neural audio autoencoders, it is timely to explore as a community how they have opened a domain of design [26] in NIME.

One approach to unpack this domain of design is through the lens of *material-oriented musical interaction* [44] – that is, acknowledging technologies' role in shaping creative actions, and treating them as the site for exploring musical activities [44, 97]. In particular, it draws attention to how makers craft and experiment with technologies to explore their use "at play" in interaction design [17, 64]. NIME research has employed this material-oriented view to observe the emerging affordances of tools and technologies in musical activities [59, 96].

However, we argue that artists and designers' hands-on engagement with neural audio autoencoders remains constrained for various reasons. First, technical AI research is typically positioned outside audio production and creative coding environments, offering limited means for NIME crafting [18, 64]. Second, limited resources such as technical support and documentation raise the barrier to learning and using the underlying models and data [15, 19]. Last, the opaque and high-dimensional nature of model parameters [13] makes their sonic behaviours harder to understand than conventional NIME materials. In other words, the need for tools and resources that enable artist-led and practice-led engagement with autoencoders has emerged.

This paper focuses on one compelling technique of using autoencoders, the *decoder-only* technique that involves overriding latent representations with user inputs to use the decoder as a sound synthesis model [87]. As a platform to explore nuanced control over AI generative models [79, 95], the decoder-only technique allows musicians to "navigate" or "walk" in the latent space as a sound space for musical expressions [47, 65]. We contribute *Latent Terrain*,[1] a Max/MSP package with a repository of resources for sound synthesis with neural audio autoencoders.

NIME '26, June 23–26, 2026, London, UK

[1] Project homepage: https://jasper-zheng.github.io/nn_terrain/

Latent Terrain allows makers to tailor autoencoders into customised sound spaces that yield spectrally complex and varied musical phrases,[2] aiming to bridge autoencoders with broader NIME materials. We use this package to explore our research question: *What are the challenges and opportunities of using neural audio autoencoders as design materials for NIME makers?*

We offer an account of how we transform our technical experimentation of neural audio autoencoders into the Max/MSP package. Inspired by practice-led works that embrace artistic contextualisation of the artefact during its iteration [5, 27, 70], we engage in a collaborative exploration with four artists and musicians to illustrate a design space of Latent Terrain in NIME. This leads to the creation of an annotated portfolio [26] that offers insight into the design thinking and experience behind four artistic pieces. Together, these technical practices and artistic practices allow us to reflect on the opportunities in NIME that have opened by neural audio autoencoders, and the challenges of adapting them as materials in a crafting space.

2 Background

Here we introduce the technical background of autoencoders, the decoder-only technique, and the corpus-based method to build musical interfaces. We also introduce our perspective, the material-oriented exploration of neural audio autoencoders in Digital Musical Instrument (DMI) design.

2.1 Audio Latent Space and Latent Trajectories

There have been comprehensive reviews of autoencoders and latent spaces [79, 95]. Here we summarise their essential background. An audio latent space is a multi-dimensional vector space learned from a corpus of audio data [28, p. 501], typically learned by an autoencoder. The components of an autoencoder include (i) an encoder that compresses audio waveforms into a sequence of vectors in the latent space, referred to as a **latent trajectory**, and (ii) a decoder that takes the latent trajectory to reconstruct audio waveforms, which aims to resemble the original waveforms.

An emerging way of using autoencoders for sound synthesis in NIME design is a **decoder-only** approach that involves overriding the latent vectors with continuous user input to create synthetic latent trajectories [41, 55, 87, 89]. This way of treating the decoder as a generative model has been referred to as *latent space navigation* or *latent space walk* by practitioners in neighbouring fields [2], and widely adopted in NIME design [65, 76]. However, as "black box" AI models, latent spaces do not encode musical controls in easy-to-understand ways [13, 46]. Consequently, existing practices typically rely on the musician to "blindly" explore in the space and discover ways of using it [47]. In addition, general-purpose autoencoders can have a latent dimensionality that exceeds the upper limit of usable dimensions in an instrument [40]. These pose the challenge of designing interfaces with low-dimensional controls to access latent spaces, which we aim to address in this paper.

2.2 Corpus-Based Sound Spaces

Research in eXplainable AI (XAI) has explored ways of projecting the high-dimensional latent space into a low-dimensional control space [13] for accountable and steerable generation. A common approach is using *Dimensionality Reduction* (DR) to extract low-dimensional representations of latent spaces. The use of DR can be seen in a range of works on corpus-based sound spaces [62, 63], aiming to use sound descriptors to build and access a sound corpus and preserve essential characteristics of it [34]. Recent works have explored the creation and interaction of sound spaces for a range of musical applications [1, 25, 62, 83]. Other sound space mapping methods such as musical parameter mapping [85] and latent mapping [29, 45] are less relevant to our work because they do not emphasise the use of audio corpora.

The goal of our proposed sound space construction method, Latent Terrain, is related to but distinct from DR. It also aims to build low-dimensional control interfaces for high-dimensional latent spaces. However, rather than clustering latent codes in the interface as typical DR approaches do, we aim to construct "maps" for latent spaces that resemble the *high-dynamic* nature of latent trajectories, which in effect offer higher spectral complexity for richer and more varied musical phrases.

2.3 Material-Oriented DMI Design

The applications of neural audio autoencoders in DMI design have focused on supporting functional tasks such as timbre transfer [18], conditional synthesis [21, 46], and text-to-audio [23]. These contrast our goal of exploring neural audio autoencoders as design materials, which we introduce below.

To position neural audio autoencoders as design materials, we take on Mudd's view of material-oriented musical interaction [44], which treats materials' resistance and constraints to be revealed and worked with to explore them as creative mediations [27] for DMI design. This reflects Barad's notion of intra-action [9, p. 189], which locates agencies in the entangled coupling between the designer and the technologies. Works that foreground this material-oriented view can embrace practice-related approaches that involve iterative crafting, prototyping, and experimenting with the material [70, 96]. For instance, early contextualisation of artistic practice during the technical development process [5, 70], or considering technical research as practice to be documented by first-person reflexive accounts [53].

Following this material-oriented view, we define our goal as treating autoencoders themselves as the site for exploring musical activities to understand their materiality in NIME.

2.4 Material-Oriented Exploration of Neural Audio Autoencoders

Various NIME examples illustrate how DMI designers unpack the materiality of neural audio autoencoders to explore their creative possibilities. For instance, the embodied and somatosensory aspects of latent representations have inspired works on movement-sound interactions [42, 47] and studies that look into performing techniques in latent spaces [55, 99], and the autonomous feature of latent spaces has been explored in live improvisations [7, 78], in which their uncertain and generative characteristics [10] are used as a source for music creativity. Moreover, various new interfaces for music composition [82], notation [8, 39, 99], and sound design [36] have been explored. Our approach focuses on constructing sound spaces for latent spaces and their use in musical interface design.

3 Algorithm Design and Experiments

Our exploration started by experimenting with the sound space construction method. This section focuses on technical challenges of the method, such as the mapping accuracy, underfitting and overfitting, and low-latency. We outline our design principles, algorithmic method, and data-driven experiments.

[2] See Section 3.1.2 for details.

3.1 Design Principles of Latent Terrain

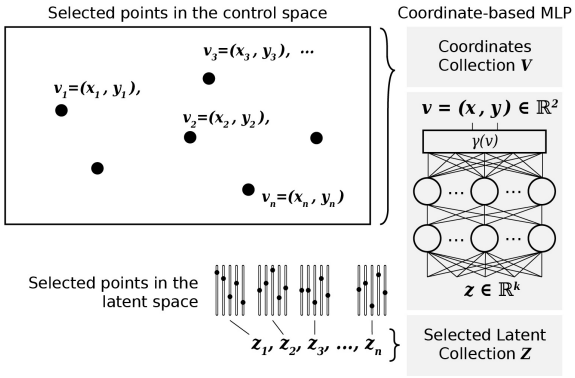
In Section 2.2 we reviewed notable approaches for constructing latent sound space, below we follow up with two interaction design principles that are unique in our work.

3.1.1 Constructing sound spaces with latent trajectories. To offer intuitive controls over sound synthesis with latent spaces, we aim to create a coordinate-to-latent mapping that maps a control space to the latent space. Our initial design was inspired by the work of Vigliensoni and Fiebrink [87] (illustrated in Figure 1a). They used a Multi-Layer Perceptron (MLP) as the mapping model, which takes control space coordinates as inputs and is trained to output latent vectors. The core idea is that the MLP is trained with interactive machine learning [24]. In practice, the users explore and find interesting points in the latent space as targets, and select points in the control space that should map to these targets to create MLP training data [87]. At each timestep during inference, the trained MLP maps a control space coordinate to a latent vector, which is then reconstructed to audio waveforms by the decoder.

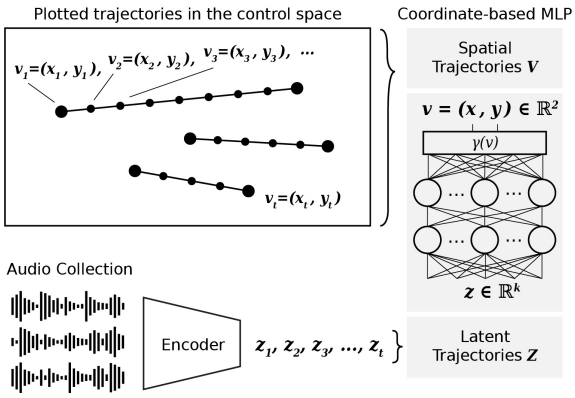
We extend the work of Vigliensoni and Fiebrink [87] from modelling latent vectors to modelling latent trajectories (see Section 2.1 for *latent trajectory*). This enables the sound space to be constructed by long audio samples rather than single audio frames, therefore allowing longer musical phrases to be captured by the sound space. As shown in Figure 1b, the training data of our MLP are (i) sequences of control space coordinates (**spatial trajectories**) plotted by the user, and (ii) sequences of latent vectors (**latent trajectories**) encoded from an audio collection, together forming the coordinate and latent pairs.

To state this problem precisely, Figure 1b uses a 2-dimensional control space as an example. Given a neural audio autoencoder with k-dimensional latent space, a c-channel audio segment $x \in \mathbb{R}^{c \times T}$ can be encoded to a discrete latent sequence $Z = E(x) = \{z_1, z_2, ..., z_t\} \in \mathbb{R}^{k \times t}$, where T and t denote time at audio sampling rate and time at latent frame rate, respectively. The ratio between T and t is defined by the autoencoder's compression ratio. A discrete coordinate sequence $V = \{v_1, v_2, ..., v_t\} \in \mathbb{R}^{2 \times t}$ in the control space is arranged by the user, where $v_t = (x_t, y_t) \in \mathbb{R}^2$. Our goal is to train a MLP model $f_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}^k$ with paired V and Z. Trained f_θ parameterises a latent trajectory as a function of continuous coordinates v. Models tackling similar tasks have been referred to as coordinate-based MLPs [80] or Compositional Pattern Producing Networks (CPPNs) [72, 81].

3.1.2 The high-frequency nature of latent trajectories. However, the practical challenge of training $f_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}^k$ lies in the high-frequency nature of how audio samples traverse the latent space. The latent trajectory shown in Figure 2 (top, dots) demonstrates that a sequence of vectors in a latent dimension can contain high-frequency components, in which latent values change rapidly. These components are crucial in synthesising phrases that have dynamically changing structure or timbre [21]. However, when mapping these components to coordinates in dense Euclidean spaces, such as 1D linear timelines or 2D planes, standard MLPs can fail to model frequent changes in the sequence [80], illustrated in Figure 2 (top, the orange solid curve). This difficulty of approximating high-frequency components has been referred to as the phenomenon of **spectral bias** in machine learning literature [57]. As a result, modelling latent trajectories in this way typically results in sound spaces that are overly smooth and lack musical dynamics. Although this way of smoothly interpolating between sounds has indeed fostered remarkable NIME designs [76, 86], we hope to explore an alternative approach that can model high-dynamic musical phrases, to offer it as a variation to NIME designers.



(a) Vigliensoni and Fiebrink's method [87].



(b) Our method supports the previous work [87], and extends to modelling spatial trajectories paired with latent trajectories encoded from samples.

Figure 1: Comparing the coordinate-to-latent mapping method by Vigliensoni and Fiebrink [87] with our method.

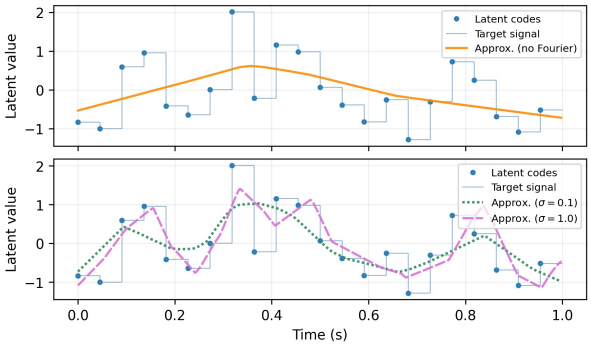


Figure 2: Latent codes of a 1-sec audio encoded by the autoencoder in Stable Audio Open [23], showing the first latent dimension. Codes are converted to a latent signal by a zero-order hold. MLPs are trained to parameterised the signal given time t. A standard MLP (top: solid curve) have difficulty approximating the target signal. In contrast, MLPs with Fourier features (bottom: dashed and dotted curves) yield results that overcome "spectral bias".

Table 1: Our selection of autoencoders/codecs.

Autoencoder	Year	Latent dim.	Latent type[5]	Regularisation[6]	Comp.[7]	In Max[8]	Used in NIME design
RAVE[1]	2021	Various[9]	Continuous	Various[9]	2048	Yes	[8, 18, 30, 36, 41–43, 47, 54, 55, 65, 66, 68, 69, 73, 76, 86, 88, 93, 98]
Music2Latent[2]	2024	64	Continuous	Variational	4096	Yes	[48, 82]
Stable Audio Open[3]	2024	64	Continuous	Variational	2048	Yes	[7, 20, 75]
FlowDec[4]	2025	128	Discrete	RVQ [38]	1764	No	None

[1] Realtime Audio Variational autoEncoder (RAVE) [16], Max/MSP version available at https://github.com/acids-ircam/RAVE
[2] Music2Latent (M2L) [51] is a consistency autoencoder. Our Max/MSP implementation: https://github.com/jasper-zheng/music2latent-scripted
[3] The autoencoder in Stable Audio Open (SAO) 1.0 [23]. Our Max/MSP implementation: https://github.com/jasper-zheng/streamable-stable-audio-open
[4] FlowDec [90] is a neural audio codec with discrete latent codes, with underlying codecs from [38]. We only used the NDAC-25 configuration due to its low latent dimensionality and low latent frame rate (25 Hz).
[5] Latent type: Specifies whether the latent representations are in a continuous space or quantised into discrete codes.
[6] Regularisation: Defines how latent representations are learned during training, affects how information can be distributed in a latent space.
[7] Here we define Compression Ratio as the proportion by which the audio sampling rate is reduced to the latent frame rate, different from [52].
[8] Whether there is a version of that autoencoder implemented for realtime continuous inference in Max/MSP, by the time of writing.
[9] We used two RAVEs with Wasserstein regularisation and 16 latent dimensions, and one with variational regularisation and 4 latent dimensions.

3.2 Modelling High-Frequency Latent Trajectories with Fourier Features

Here we address the challenge formulated in Section 3.1.

3.2.1 Fourier feature mapping. Fourier feature mapping is a technique that allows neural networks to learn high-frequency functions in dense Euclidean coordinates [80], proposed by Rahimi and Recht [58], primarily applied it in computer vision for image and geometry synthesis [81]. Here we apply it to our coordinate-based MLP. Precisely, in a d-dimensional control space with coordinates $v \in \mathbb{R}^d$, we sample a matrix $B \in \mathbb{R}^{m \times d}$ from $\mathcal{N}(0, \sigma^2)$ with mapping size m and Gaussian scale σ. Before passing v to the MLP, we apply the Fourier feature mapping $\gamma(v)$:

$$\gamma(v) = \left[\sin(2\pi Bv),\ \cos(2\pi Bv) \right] \tag{1}$$

Equivalently, writing $b_k \in \mathbb{R}^d$ for the k-th row of B shows that inputs in the control space is transformed by a series of sinusoidal functions with varying frequencies:

$$\begin{aligned} \gamma(v) = \big[&\sin(2\pi b_1^\intercal v),\ \ldots,\ \sin(2\pi b_m^\intercal v), \\ &\cos(2\pi b_1^\intercal v),\ \ldots,\ \cos(2\pi b_m^\intercal v) \big] \end{aligned} \tag{2}$$

In the following sections, we show our experiments on harnessing it for latent trajectory regression.

3.2.2 Experiments and results. We selected a range of autoencoders to form a plurality of materials, presented in Table 1, aimed at covering those that (i) have open-source pre-trained models (ii) have been used in NIME for artistic practices, (iii) vary in configurations including dimensionality, regularisation, and compression ratio. We use the three autoencoders that have a Max/MSP version later in our NIME making.

To stay focused on the practical use of our mapping approach, we put the majority of our technical evaluation in the Appendix. In summary, shown in Table 2, Fourier features improve the mapping accuracy across all autoencoders, and the addition of the mapping MLP only minimally impacts the fidelity of reconstructed audio. Full details on the task, datasets, implementations,[3] results, and discussion can be found in Appendix A.1 and A.2. We also inspect the training error with respect to multiple frequency components to verify that the Fourier features are enabling the MLP to learn high-frequency details as claimed, results can be found in Appendix A.3.

[3]The source code of all experiments is released at: https://github.com/jasper-zheng/latent-terrain-pytorch

Table 2: Regressing trajectories with the drum collection.

(a) Peak Signal-to-Noise Ratio between original latent trajectories and regressed trajectories (higher is better).

	Peak Signal-to-Noise Ratio ↑			
Regression Config	RAVE	M2L	SAO	FlowDec
No Fourier features	27.516	29.774	21.697	23.530
Fourier features	**41.369**	**40.368**	**32.069**	**32.566**

(b) Fréchet Audio Distance of audio sets reconstructed from the original trajectories (encoder-decoder recon) and the regressed trajectories, comparing to the original audio set (lower is better).

	Fréchet Audio Distance ↓			
Regression Config	RAVE	M2L	SAO	FlowDec
Encoder-decoder recon.	7.664	6.659	2.324	2.009
No Fourier features	14.188	12.354	28.163	28.069
Fourier features	**8.139**	**7.511**	**2.387**	**2.473**

3.3 Configuring Fourier Features in Practice

Here we experiment with the mapping model with a 2D control plane with coordinates $(x, y) \in [0, 1]^2$, which is a common setting in the practice of sound space [63]. As shown in Figure 3, we train the mapping MLP with random control space trajectories (in practice these should be defined interactively by the user) paired with latent trajectories. We then visualise the sound space by sampling the MLP with coordinates in the $[0, 1]^2$ interval and display the first latent dimension as a greyscale matrix.

The Gaussian scale σ is a key parameter in configuring Fourier features when constructing the sound space. This mirrors the findings in [80] that σ qualitatively affects the balance between underfit and overfit. Visulisations in Figure 3 illustrate this point by showing the effect of σ on the overall flatness of the sound space. As described in the caption, a proper σ leads to a terrain with a steep and mountainous surface, which in effect leads to spectrally complex and varied sound when navigating the control space. We therefore expose this parameter to the API (see Section 4) to defer the practical parameter tuning to makers and designers who will be creating the sound space.

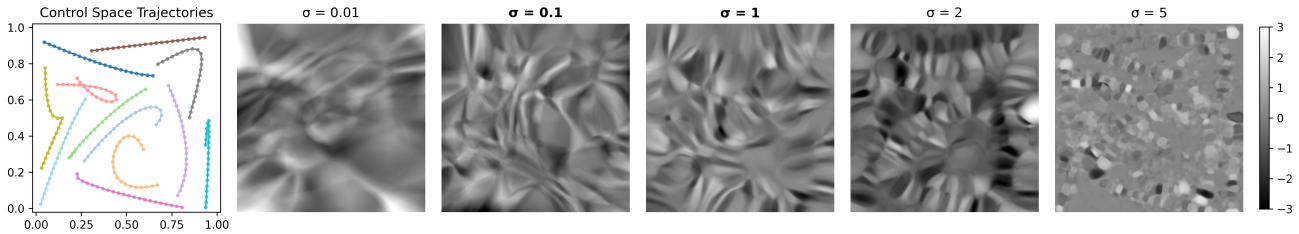


Figure 3: Latent terrains created by pairing controls space trajectories (randomly sampled) and latent trajectories (encoded by SAO using a set of 1-sec drum samples from MUSDB18 [56]). Visualisations of magnitude at the first latent dimension with various Gaussian scales σ. Lower σ causes overly flattened interpolation (underfit). Larger σ causes noisy sound space and poor generalisation to areas not covered by training data (overfit). This is subject to experimentation, but we found $\sigma \in [0.1, 1]$ generally can lead to a musically interesting result.

3.4 On-the-fly Training and Inference

Relevant works on neural audio typically evaluate model efficiency at inference [18]. However, in our setting the interactive training also plays an important role in adapting autoencoders. We therefore also evaluate the on-device training speed.

We measured the time taken to train the mapping MLP for 1000 steps across various training batch sizes. A training step is defined as one optimisation cycle in which one batch of training data is processed. Results in Figure 4 (Left) were tested on the CPU of an Apple Silicon M4 Max laptop with 48GB RAM. Training in all conditions can be done within 2 seconds. To put this into context, encoding a 10-sec audio sample results in 215 latent codes. With a batch size of 16, one training epoch can be done by 14 steps. In our experience, 500 epochs can ensure a converged training, therefore yields roughly 1.2 minutes training time. This matches works such as [71] that involve training MLPs as mapping models, claiming an under 2 minutes training time.

For real-time sound synthesis, we measured the four autoencoders across various streaming buffer sizes by Real-Time Factor (RTF), calculated as the ratio between the time taken to process a buffer and the temporal duration of that buffer. We compared RTFs under two conditions: inference of the decoder together with the mapping MLP, and inference of only the decoder with randomly sampled latent vectors. Results in Figure 4 (Right) were tested on the GPU of the same M4 Max laptop, showing that the addition of the MLP only introduces a minimum increase in the RTF. Full results across more devices are shown in Appendix A.4.

4 Situating Latent Terrain in DMI Design

We developed `nn.terrain~`,[4] a package of Max/MSP objects for NIME makers, encapsulating the sound space method described in Section 3. We refer to the sound space as *terrain* to draw an analogy to wave terrain synthesis [33]. Objects in the package provide Max-style APIs that allow users to build, adapt, and use terrains with pre-trained autoencoders. The package also includes utility objects for interactive trajectory plotting, real-time latent recording, trajectory playback, and more.

The `nn.terrain~` online project page provides installation guide, API documentation, instructions, tutorials, and Max help files. It also includes artist-contributed patch examples and demos. We hope to use this collectively built repository to respond to the conference theme **Communities**. In particular, the musical and technical communities that are exploring AI-enhanced DMIs such as neural audio autoencoders [16, 43, 49].

[4] https://jasper-zheng.github.io/nn_terrain/

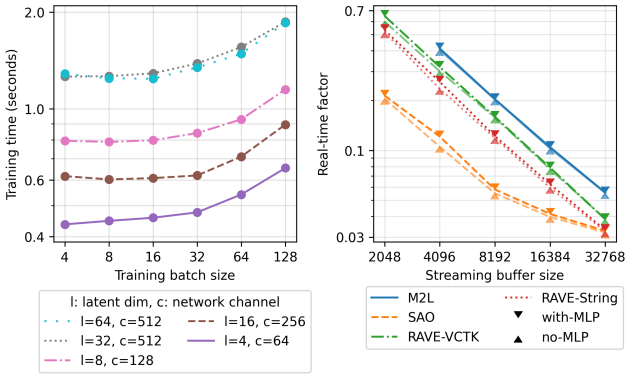


Figure 4: Left: Average time needed to train the MLP for 1000 steps against training batch size, across MLP configurations. Right: Mean real-time factor against buffer size, across all decoders, comparing with and without the MLP. Mean values were computed over 100 runs per condition. Both metrics are lower the better.

4.1 Origin Story

To situate the algorithm into the environment of DMI design, the development of `nn.terrain~` is informed by the first author's practices of music compositions with neural audio autoencoders. The API implementation is shaped by the author's intended ways of building sound spaces and using them for music compositions in practice. At the time of writing, the author had been using `nn.terrain~` for DMI design for 12 months, from the initial prototype to the publicly released package. This continuous technical practice [53] is documented by write-ups and presentations at a series of public engagement events across the start, middle, and end of the project, combined with major technical change logs (shown in Appendix B). Here we report three design milestones that focus on the minutiae and rationale of appropriating the technical artefact into creative practices. The aim is to demonstrate how musical interactions can emerge from, and be shaped by, the material-oriented resistances and constraints.

4.2 Milestones and Design Artefacts

We describe three Graphical User Interface (GUI) designed along with the package development. These interfaces mark the milestones that lead to the latest release of `nn.terrain~`. All GUIs (screenshots in Figure 5) explore ways of using `nn.terrain~` for music compositions and performances.

Interface A: Embodied Sound Space Navigation. The first GUI design with `nn.terrain~` focuses on exploring the embodied movement-sound coupling in the sound space. It offers a drawing pad as the control space, in which the stylus (x, y) location is used to navigate in the sound space and produce real-time sonic feedback. The artistic intent behind the interface is to use the drawing canvas as a space to create graphical music scores, and to trace the musical gestures in the scores. One of the emerging strategies of using it is annotating sonically interesting zones on the drawing pad and experimenting with ways of re-enacting sounds in these zones.

Interface B: Interactive Sound Space Construction. In the second design and development phase, we explore the use of terrains in a wider range of models and audio collections. Therefore the API design focuses on supporting the interactive terrain construction within the Max/MSP environment. As a result, the cycle of training pairs gathering, model training, monitoring, and inference was distributed to three interdependent Max objects. One emerging way of using the terrain for sound synthesis is looking for ways of programming the terrain navigation. For instance, we added the right-most panel that allows modulating the stylus (x, y), mimicking existing practices of composing with sound spaces, such as the "oscillations and orbits" control in [61].

Interface C: Programming Trajectory Playback. In the third phase, we explore more ways of navigating the terrains in sample-based music composition. The input control space coordinates were exposed to the API to allow for customised input devices, rather than constrained to the stylus drawing pad to open the design space of input devices. An `nn.terrain.gui` object is refined as a utility object to control the playback of trajectories. As shown in Figure 5C, the user first loads audio samples and plots trajectories in the "train mode" the same way as the previous interface. Then in the "playback mode" they define a new set of trajectories from the 2D canvas as "latent samples". Trajectories can be programmed by the pattern in the step sequencer on the left.

5 Practice-Based Exploration

After the initial release of the `nn.terrain~` package, we collaborated with four artist-researchers to explore NIME design with autoencoder and latent terrain. The connection with each artist-researcher is motivated by a mutual interest in exploring neural audio synthesis in NIME. We aimed to situate `nn.terrain~` within the practice of each artist-researcher [74], in particular, audio programming, music performance, DMI, and sound installation design.[5] The collaborations started in the summer of 2025 and lasted over three to five months. During the period, we had regular meetings (either bi-weekly or monthly, with a combination of remote and in-person modes) with each artist-researcher to catch up on the progress. The ideation and conceptualisation of each work are done by the artist-researcher after their first meeting with the first author to grasp an initial understanding of the package. Whereas the role of the first author is a facilitator who briefs and explains the use of `nn.terrain~`, signposts relevant resources and documentation, and provides technical support throughout the process. The artist-researchers join this paper as the second to fifth co-authors to bring their voices into our discussion and acknowledge their project contribution.

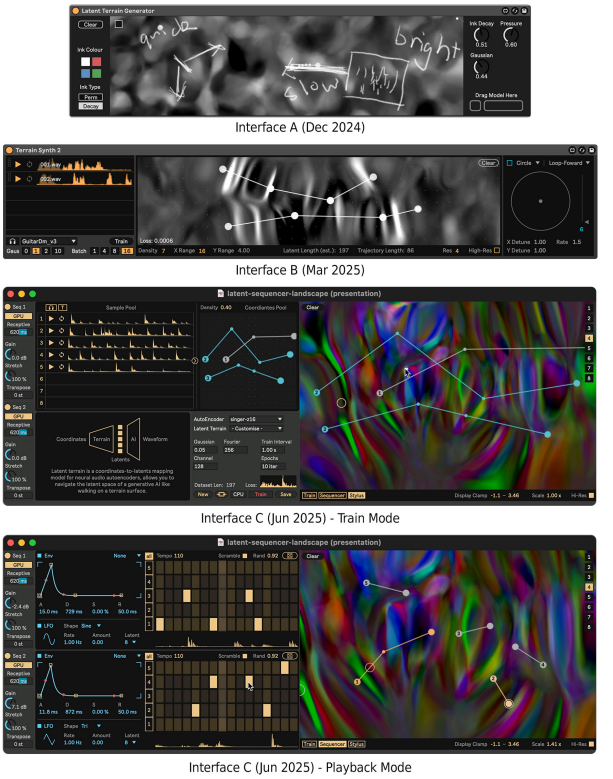


Figure 5: Three GUI designs with Latent Terrain, span from Dec 2024 to Jun 2025. Recorded demos of each interface can be found in the online project page.

5.1 Portfolio of Works

To collectively showcase artist-researchers' ways of integrating `nn.terrain~` into NIME design, we create an annotated portfolio. An annotated portfolio [26] brings together design artefacts into a systematic body of work, to share the design thinking and experience behind them, and show the broader design issues the collection addresses. The following sections and Table 3 present a summary of each project. The full portfolio is also showcased on the project website.[6]

Project 1: Repressive Terrain by Keigo Yoshida. A data sonification patch (Figure 6) that aims to create a meditative listening experience, in which the audience's time-varying Electroencephalography (EEG) data is sonified to a real-time synthesised soundscape. The artistic theme is the active manipulation of consciousness in a focused meditation, and an adversarial tension between the audience's striving for calm and the algorithm's impulse toward arousal sonic responses. It exploits the interactive sound space adaptation aspect of the Latent Terrain toolkit to create a real-time evolving soundscape.

Project 2: nn/mémoire by Jiatong Liu. An audio-visual virtual environment (Figure 7) that aims to create an embodied listening experience that reimagines the sound of cultural heritage in Hutongs, a type of traditional northern Chinese courtyard house in Beijing. The environment is a virtual gallery that allows audiences to wander through and hear the real-time generated soundscape that captures the sound of Hutongs. The sound space

[5] Detailed positionality statement of each artist can be found in Appendix C.

[6] https://jasper-zheng.github.io/nn_terrain/posts/annotated-portfolio/

Table 3: Summary of the four projects in our portfolio.

	Type of work	Autoencoder	Audio collection	Key aspect explored by the work
Project 1	A data sonification patch	RAVE	Samples of vintage music	Interactive sound space adaptation by biofeedback
Project 2	An audio-visual environment	Stable Audio Open	Archival and field recordings in Hutongs	Crafting a sound space for an virtual gallery
Project 3	A live performance patch	RAVE	Samples of organ instrument recordings	Autonomous micro-variations in latent space navigation
Project 4	A soundscape composition	Stable Audio Open	Various (drum loops, vocals, and piano)	Neural synthesis as a resampling tool for composition

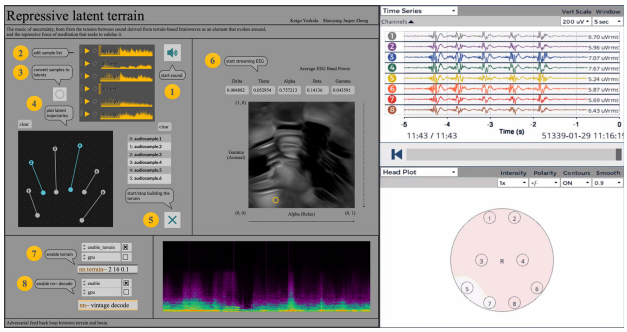


Figure 6: Screenshots of the customised patch created by Keigo Yoshida and the first author, using with signals from an OpenBCI EEG headset.

Figure 7: A screenshot of nn/mémoire by Jiatong Liu.

is constructed with archival recordings from the Sound Art Museum Beijing, aiming to preserve the vanishing living culture that is now vanishing due to urban changes. It explores methods to craft latent spaces into an immersive sound environment and how latent spaces open a new approach to sound design.

Project 3: Trek by Nico García-Peguinho and Nikhil Bullock. A collaborative composition and live performance system that aims for controlled timbral materials creation from minimal gestural input. It explores sampling and interpolation strategies for the latent space. Specifically, coordinates in the latent space are indexed into zones based on their acoustic characteristics. It then uses a boids algorithm[7] to create micro-variations to navigate these zones, to generate naturalistic sonic drift. It uses post-synthesis signal processing techniques to sculpt the sound. The autoencoder is in duet with a generative drum patch.

Project 4: ambient_terrain_1 by Dan Hearn. A soundscape composition produced by resampling recorded audio and field recordings. Dan explores ways of adapting autoencoders to his own existing music production practice. He created a customised

[7]https://cs.stanford.edu/people/eroberts/courses/soco/projects/2008-09/modeling-natural-systems/boids.html

Figure 8: Nico García-Peguinho and Nikhil Bullock performing their piece at the Interactive Digital Multimedia Techniques Concert at Queen Mary University of London.

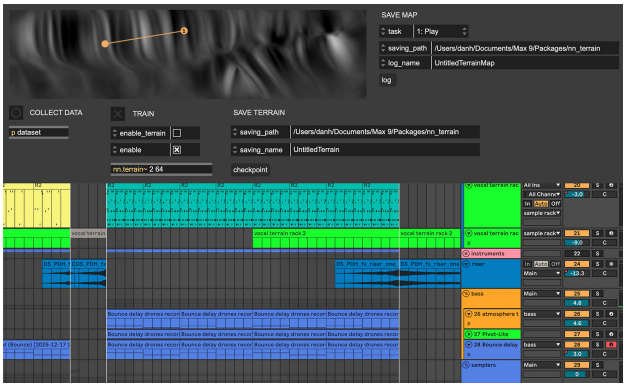


Figure 9: Screenshots showing the cutomised Max/MSP patch by Dan Hearn.

patch to generate short audio loops by sampling in the latent space. Audio materials derived from the patch are used throughout the track with non-AI composition tools. This composition process explores the combination of AI and non-AI materials in ways that the use of an AI tool is not immediately apparent to listeners, but seamlessly integrated in the sample-based music production workflow.

5.2 Annotated Themes

To elicit the design considerations and artistic choices raised by each project, we conducted a semi-structured interview with each artist-researcher after the completion of their project. The interviews are either 60-min live or asynchronous, focusing on artist-researchers describing the concept and matter of concern of their work. We then transcribed the interview audio recordings to analyse emerging themes that bridge between their work and broader research opportunities in NIME, Human-Computer Interaction (HCI), and AI, which could inform future works. Here we present these themes.

Uncertainty and ambiguity as characteristics. Across NIME and HCI, uncertainty and ambiguity have been considered as constitutive qualities that can be engaged by artistic practice [14, 17], requiring artists to deliberately work and cultivate their practices to respond to the unpredictability [77]. In the making process of all four pieces, the uncertainty and ambiguity features of neural audio autoencoders have been discussed repeatedly. "Learning to deal with the unpredictability" (Jiatong) has become a central aspect to be considered in the interface design. For instance, Dan characterised his workflow as "explorative and serendipitous", Nico referred to an "active listening" [49] way of composing in which ambiguity remained but became legible through embodied exploring and listening. In Keigo's piece, the constantly changing sound space is seen as an adversarial entity that actively resisted the performer's intentions, turning uncertainty into a compositional driver. In this respect, the complex and opaque aspects of latent spaces is acknowledged, and captured as a source for uncertainty and ambiguity within design processes.

Entangled agencies in material-centred design. In a material-centred view, the computational materiality of AI is considered as an active medium [64], in which creative practice is shaped and constrained by materials rather than solely directed by human intention imposed upon tools [92]. In musical interaction, such perspectives locate agencies in the engagement and negotiation between the designer and their materials [44]. In our portfolio, Jiatong describes that "[the design of the virtual gallery space] reveals itself" through the movement and listening in the latent space, whereas Keigo describes "let it traces and accesses past states (training data) in real-time, not to establish full control over the outcome". Taken together, design intent arise not from transparent control over the materials, but from learning how the material's capacities act, respond, and transform the creative process.

Designing with a material assemblage. Across all four pieces in our portfolio, the autoencoder itself is not an isolated instrument. Instead, makers worked with a material constellation in which the autoencoders interact with other materials in the workspace, forming an ecology of processes and environment [60]. Such as Jiatong's virtual gallery space, which was built according to the latent sound space. In this respect, design intents arise not from the autoencoder alone, but from its entanglement with tools, bodies, and environments. From an *assemblage* view [27] of AI, autoencoders belong to a broader constellation of human and non-human actors, including other synthesis algorithms, sensors, other coding environments, and the listening and composing practices. This assemblage is not a static configuration of materials, but an ongoing process in which materials are continuously constructed and de(con)structed [12]. For instance, Dan described that "[autoencoder] gradually finds its place within my sample-based music production workflow"–the role of autoencoders in the creative practice became clear through repeated trials and errors, rather than a predetermined identity.

6 Discussion

The material-oriented exploration presented in this paper is not about inventing new neural audio autoencoders, nor proposing new model steering strategies that surpass works such as [21, 46, 87]. Instead, it aims to use the Latent Terrain package to position neural audio autoencoders within a space of NIME crafting, where they are engaged as materials to be probed and learned through practice. Here we discuss our response to the research question: *What are the challenges and opportunities of using neural audio autoencoders as design materials for NIME makers?*

6.1 Latent Spaces as Tailorable Sound Spaces

We highlight two aspects of neural audio autoencoders that were explored by the makers to discuss their opportunities in NIME making.

First, makers explore ways of constructing the latent sound space that align with their creative intent. Latent Terrain can be seen as a way of adapting autoencoders with a curated audio collection. In the broader space of AI as material for artworks, building and curating training datasets has been an approach for artists to intervene in generative models and bring agency into their work [6, 17]. Latent Terrain offers an alternative way for artists to integrate their curated audio collection: tailoring the latent spaces of large generative models into a smaller and customised subset, in the form of a sound space. The use of curated datasets was explored in Jiatong's work, in which the choice of using archival recordings as the audio corpus aligns with the themes of the vanishing living culture. In addition, Keigo's work explored the interactive adaptation of the sound space, in which the mapping model is constantly being trained and updated, forming the theme of uncertainty in the listening experience.

Second, makers explore ways of navigating the latent sound space to balance control and uncertainty. Latent Terrain can be seen as an approach of tailoring the high-dimensional latent space into a low-dimensional control space. In our portfolio, ways of navigating the latent sound space are entry points to make clear the connection between the autoencoder and a range of non-AI materials in the makers' existing workspace or practices. For instance, Dan's practice of sample-based music composition informed his way of sampling in the latent space to derive materials for composition, whereas Nico's use of the boids algorithm to navigate the space has become a unique theme in his work.

In summary, we suggest that future work explore the use of curated audio collections, the interactive adaptation of latent space, and the plural ways of navigating the latent sound space to discover its opportunities in NIME.

6.2 Toward a Shared View on Autoencoders

In Section 1 we highlighted several practical challenges of enabling hands-on engagement with neural audio autoencoders such as technical barriers, lack of resources and support, and the complex nature of latent spaces. Reflecting on the development of `nn.terrain~` and the artistic exploration, we argue that the lack of a space to share practice, knowledge, and techniques can be another challenge in unpacking the creative possibilities of autoencoders in NIME making.

When collecting the four pieces into the portfolio, each maker uses domain-specific techniques to work with autoencoders, such as the "active and embodied listening" approach by Jiatong and Nico. They also invent unique vocabularies to articulate their technical practices, such as Dan's way of describing his latent resampling approach and Keigo's way of describing the synthesis process as "tracing and accessing past state". We suggest that these multifaceted approaches can benefit from a sustained community that establishes shareable practices [91], technical know-how [84], and folk theories [67] to explore AI materials in NIME making. In a broader DMI design view, cultural exchanges with and within artists' communities play an important

role in supporting digital luthiers' practice-based engagement with technical artefacts [4]. An example can be drawn from how the turntable was transformed into a musical instrument by the development of its own musical culture, techniques, and virtuosi [35, p. 326]. Neural audio autoencoders, on a similar trajectory, become design materials in NIME through sustained, practice-based engagement within a crafting space that is shared by both music communities and technical communities.

7 Conclusion

We contribute a technical method, a Max/MSP package, and a practice-based account that foregrounds crafting as a site to explore emerging AI-enhanced NIME designs. We focused on the decoder-only latent space walk technique with neural audio autoencoders. Our technical experimentation with autoencoders resulted in Latent Terrain, a Max/MSP package that constructs low-dimensional sound spaces from high-dimensional latent trajectories. By positioning neural audio autoencoders within a crafting space, we explored the challenges and opportunities of using them as design materials in NIME. Through a collaborative exploration with four artist-researchers, we delivered an annotated portfolio to showcase a domain of design opened by neural audio autoencoders. We highlighted that makers explore ways of constructing the latent sound space that align with their creative intent, and ways of navigating them to establish control over the sound synthesis. We called for a community view of autoencoders in which artists and makers are empowered to share technical know-how and practices with AI materials to unpack their creative possibilities in NIME making.

8 Ethical Standards

All autoencoder models used in the papers are pre-trained models trained on audio files licensed under CC0, CC BY, or CC Sampling Plus. Data used for interactive machine learning is either under CC0 or CC BY licences, or created by the authors themselves.

The practice-based exploration described in the paper is conducted within the authors and did not involve experiments with other human participants. The project was approved by the Queen Mary University of London ethics committee, reference number: `QMERC20.565.DSEECS25.026`.

Acknowledgments

Shuoyang Zheng is a research student supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (grant number EP/S022694/1). We would like to thank all the artists for their time and engagement. Thank you to Joseph Meyer and all colleagues who have gave thoughtful feedback to help improving the quality of this paper.

References

[1] Sabina Hyoju Ahn, Ryan Millett, and Seyeon Park. 2025. Eco-Sonic Interfaces for Embodied AI Sound Exploration. *Proceedings of the International Conference on New Interfaces for Musical Expression* (June 2025), 1–5. https://doi.org/10.5281/zenodo.15698772

[2] Memo Akten, Rebecca Fiebrink, and Mick Grierson. 2018. Deep Meditations: Controlled navigation of latent space. In *NeurIPS 2018 Workshop on Machine Learning for Creativity and Design*. https://nips2018creativity.github.io/doc/Deep_Meditations.pdf

[3] Erik Arisholm, Lionel C. Briand, and Audun Foyen. 2004. Dynamic coupling measurement for object-oriented software. *IEEE Transactions on Software Engineering* 30, 8 (2004), 491 – 506. https://doi.org/10.1109/TSE.2004.41

[4] Jack Armitage, Thor Magnusson, and Andrew McPherson. 2023. Studying Subtle and Detailed Digital Lutherie: Motivational Contexts and Technical Needs. *Proceedings of the International Conference on New Interfaces for Musical Expression* (May 2023), 1–9.

[5] Jack Armitage, Victor Shepardson, and Thor Magnusson. 2024. Tölvera: Composing With Basal Agencies. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 282–291. https://doi.org/10.5281/zenodo.13904854

[6] Anne Arzberger, Maria Luce Lupetti, and Elisa Giaccardi. 2024. Reflexive Data Curation: Opportunities and Challenges for Embracing Uncertainty in Human–AI Collaboration. *ACM Trans. Comput.-Hum. Interact.* 31, 6 (Dec. 2024). https://doi.org/10.1145/3689042

[7] Misagh Azimi and Mo H. Zareei. 2025. Live Improvisation with Fine-Tuned Generative AI: A Musical Metacreation Approach. *Proceedings of the International Conference on New Interfaces for Musical Expression* (June 2025), 389–393. https://doi.org/10.5281/zenodo.15698902

[8] Benjamin Keith Bacon, Leonardo Auri, and Holly Ambrozic McKee. 2025. Approaches to notation for embodied engagement with a novel neural network-based musical instrument. *Frontiers in Computer Science* Volume 7 - 2025 (2025). https://doi.org/10.3389/fcomp.2025.1597806

[9] Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

[10] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445481

[11] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc.

[12] John Bowers and Annika Haas. 2014. Hybrid Resonant Assemblages: Rethinking Instruments, Touch and Performance in New Interfaces for Musical Expression. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Goldsmiths, University of London, London, United Kingdom, 7–12. https://doi.org/10.5281/zenodo.1178718

[13] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. 2021. Exploring XAI for the Arts: Explaining Latent Space in Generative Music. In *1st Workshop on eXplainable AI approaches for debugging and diagnosis*.

[14] Nick Bryan-Kinns, Ashley Noel-Hirst, and Corey Ford. 2024. Using Incongruous Genres to Explore Music Making with AI Generated Content. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*. Association for Computing Machinery, New York, NY, USA, 229–240. https://doi.org/10.1145/3635636.3656198

[15] Nick Bryan-Kinns, Shuoyang Jasper Zheng, Francisco Castro, Makayla Lewis, Jia-Rey Chang, Gabriel Vigliensoni, Terence Broad, Michael Paul Clemens, and Elizabeth Wilson. 2025. XAIxArts Manifesto: Explainable AI for the Arts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3706599.3716227

[16] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. https://doi.org/10.48550/arXiv.2111.05011

[17] Baptiste Caramiaux and Sarah Fdili Alaoui. 2022. "Explorers of Unknown Planets": Practices and Politics of Artificial Intelligence in Visual Arts. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022). https://doi.org/10.1145/3555578

[18] Franco Caspe, Andrew McPherson, and Mark Sandler. 2025. Waveform Autoencoding at the Edge of Perceivable Latency. *Proceedings of the International Conference on New Interfaces for Musical Expression* (June 2025), 73–76. https://doi.org/10.5281/zenodo.15699550

[19] Isaac Clarke, Francesco Ardan Dal Rí, and Raul Masu. 2025. Longevity of Deep Generative Models in NIME: Challenges and Practices for Reactivation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, 224–230. https://doi.org/10.5281/zenodo.15735662

[20] Nick Collins. 2025. Unstable audio: code bending text-to-music generation. *Journal of the Audio Engineering Society* 15 (Oct. 2025). https://aes2.org/publications/elibrary-page/?id=23004

[21] Nils Demerlé, Philippe Esling, Guillaume Doras, and David Genova. 2024. Combining Audio Control and Style Transfer Using Latent Diffusion. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*. ISMIR, 721–728. https://doi.org/10.5281/zenodo.14877437

[22] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 1068–1077.

[23] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable Audio Open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10888461

[24] Rebecca Fiebrink, Dan Trueman, and Perry R. Cook. 2009. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Pittsburgh, PA, United States, 280–285. https://doi.org/10.5281/zenodo.1177513

[25] Leandro Garber, Tomás Ciccola, and Juan Cruz Amusategui. 2020. AudioStellar, an open source corpus-based musical instrument for latent sound structure

discovery and sonic experimentation. In *Proceedings of the 2021 International Computer Music Conference*. hdl.handle.net/2027/fulcrum.t435gg568

[26] Bill Gaver and John Bowers. 2012. Annotated portfolios. *Interactions* 19, 4 (July 2012), 40–49. https://doi.org/10.1145/2212877.2212889

[27] Artemi-Maria Gioti, Aaron Einbond, and Georgina Born. 2022. Composing the Assemblage: Probing Aesthetic and Technical Dimensions of Artistic Creation with Machine Learning. *Computer Music Journal* 46, 4 (Dec. 2022), 62–80. https://doi.org/10.1162/comj_a_00658

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[29] Jeff Gregorio and Youngmoo Kim. 2019. Augmenting Parametric Synthesis with Learned Timbral Controllers. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Marcelo Queiroz and Anna Xambó Sedó (Eds.). UFRGS, Porto Alegre, Brazil, 431–436. https://doi.org/10.5281/zenodo.3673025

[30] Nicola Leonard Hein and Viola Yip. 2024. Transsonic | Sonic Fluidity. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 57–61. https://doi.org/10.5281/zenodo.15028017

[31] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[32] Intelligent Instruments Lab. 2023. rave-models (Revision ad15daf). https://doi.org/10.57967/hf/1235

[33] Stuart James. 2005. *Developing a flexible and expressive realtime polyphonic wave terrain synthesis instrument based on a visual and multidimensional methodology*. Ph. D. Dissertation. Edith Cowan University. https://ro.ecu.edu.au/theses/107

[34] Hyeon Jeon, Hyunwook Lee, Yun-Hsin Kuo, Taehyun Yang, Daniel Archambault, Sungahn Ko, Takanori Fujiwara, Kwan-Liu Ma, and Jinwook Seo. 2025. Unveiling High-dimensional Backstage: A Survey for Reliable Visual Analytics with Dimensionality Reduction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3706598.3713551

[35] Sergi Jordà. 2004. Instruments and players: Some thoughts on digital lutherie. *Journal of New Music Research* 33, 3 (Sept. 2004), 321–341. https://doi.org/10.1080/0929821042000317886

[36] Trisha Khallaghi, Pete Bennett, and Atau Tanaka. 2025. SquishySonics: A Deformable Interface for the Physical Control of Real-time AI Sound Generation Tools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3706599.3721175

[37] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*. 2350–2354. https://doi.org/10.21437/Interspeech.2019-2219

[38] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA.

[39] Giacomo Lepri, Nicola Privato, and Thor Magnusson. 2024. Embodied Sketching for Neural Synthesis. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures (AM '24)*. Association for Computing Machinery, New York, NY, USA, 549–551. https://doi.org/10.1145/3678299.3678358

[40] Daniel J. Levitin, Stephen McAdams, and Robert L. Adams. 2002. Control parameters for musical instruments: a foundation for new mappings of gesture to sound. *Organised Sound* 7, 2 (2002), 171–189. https://doi.org/10.1017/S135577180200208X

[41] Fangzheng Liu, Lancelot Blanchard, Don D. Haddad, and Joseph Paradiso. 2025. Two Sonification Methods for the MindCube. *Proceedings of the International Conference on New Interfaces for Musical Expression* (June 2025), 511–515. https://doi.org/10.5281/zenodo.15698944

[42] Joseph Meyer, Nick Bryan-Kinns, Sarah Fdili Alaoui, Mick Grierson, and Rebecca Fiebrink. 2025. Interactive Movement-to-Audio with Pre-Trained Neural Networks. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. Association for Computing Machinery, New York, NY, USA, 491–493. https://doi.org/10.1145/3698061.3734415

[43] Christopher Mitcheltree, Bogdan Teleaga, Andrew Fyfe, Naotake Masuda, Matthias Schäfer, Alfie Bradic, and Nao Tokui. 2025. Neutone SDK: An Open Source Framework for Neural Audio Processing. In *AES International Conference on Artificial Intelligence and Machine Learning for Audio*. https://doi.org/10.48550/arXiv.2508.09126

[44] Tom Mudd. 2019. Material-Oriented Musical Interactions. In *New Directions in Music and Human-Computer Interaction*, Simon Holland, Tom Mudd, Katie Wilkie-McKenna, Andrew McPherson, and Marcelo M. Wanderley (Eds.). Springer International Publishing, Cham, 123–133. https://doi.org/10.1007/978-3-319-92069-6_8

[45] Tim Murray-Browne and Panagiotis Tigas. 2021. Latent Mappings: Generating Open-Ended Expressive Mappings Using Variational Autoencoders. In *International Conference on New Interfaces for Musical Expression*. https://doi.org/10.21428/92fbeb44.9d4bcd4b

[46] Sarah Nabi, Nils Demerlé, Geoffroy Peeters, Frederic Bevilacqua, and Philippe Esling. 2025. Adding temporal musical controls on top of pretrained generative models. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*. ISMIR, 793–800. https://doi.org/10.5281/zenodo.17811482

[47] Sarah Nabi, Philippe Esling, Geoffroy Peeters, and Frédéric Bevilacqua. 2024. Embodied exploration of deep latent spaces in interactive dance-music performance. In *Proceedings of the 9th International Conference on Movement and Computing (MOCO '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3658852.3659072

[48] Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner. 2024. Diff-a-Riff: Musical Accompaniment Co-Creation via Latent Diffusion Models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*. ISMIR, 272–280. https://doi.org/10.5281/zenodo.14877327

[49] Ashley Noel-Hirst, Charalampos Saitis, and Nick Bryan-Kinns. 2025. Sampling the Latent Space: Exploring the Creative Potential of Generative AI Through the Lens of Sample-Based Music Making. https://doi.org/10.5281/zenodo.16946825

[50] Fabian Ostermann and Igor Vatolkin. 2022. AAM: Artificial Audio Multitracks Dataset. https://doi.org/10.5281/zenodo.5794629

[51] Marco Pasini, Stefan Lattner, and George Fazekas. 2024. Music2Latent: Consistency Autoencoders for Latent Audio Compression. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*. ISMIR, 111–119. https://doi.org/10.5281/zenodo.14877289

[52] Marco Pasini, Stefan Lattner, and George Fazekas. 2025. CoDiCodec: Unifying Continuous and Discrete Compressed Representations of Audio. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*. ISMIR, 447–455. https://doi.org/10.5281/zenodo.17811403

[53] Teresa Pelinski, Andrew McPherson, and Rebecca Fiebrink. 2025. Ways of knowing, ways of writing: technical practice research in new musical instrument design. *Journal of New Music Research* 53, 0 (2025), 79–92. https://doi.org/10.1080/09298215.2024.2442348

[54] Nicola Privato, Thor Magnusson, and Einar Torfi Einarsson. 2023. Magnetic Interactions as a Somatosensory Interface. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, 387–393. https://doi.org/10.5281/zenodo.11189218

[55] Nicola Privato, Victor Shepardson, Giacomo Lepri, and Thor Magnusson. 2024. Stacco: Exploring the Embodied Perception of Latent Representations in Neural Synthesis. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 424–431. https://doi.org/10.5281/zenodo.13904899

[56] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. MUSDB18-HQ - an uncompressed version of MUSDB18. https://doi.org/10.5281/zenodo.3338373

[57] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, 5301–5310.

[58] Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS'07)*. Curran Associates Inc., 1177–1184.

[59] Nicole Robson, Andrew McPherson, and Nick Bryan-Kinns. 2024. Thinking with Sound: Exploring the Experience of Listening to an Ultrasonic Art Installation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642616

[60] Matthew Rodger, Paul Stapleton, Maarten van Walstijn, Miguel Ortiz, and Laurel S Pardue. 2020. What makes a good musical instrument? A matter of processes, ecologies and specificities. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 405–410. https://doi.org/10.5281/zenodo.4813438

[61] Gerard Roma. 2023. Agent-Based Music Live Coding: Sonic adventures in 2D. *Organised Sound* 28, 2 (2023), 231–240. https://doi.org/10.1017/S1355771823000274

[62] Gerard Roma, Owen Green, and Pierre Alexandre Tremblay. 2019. Adaptive mapping of sound collections for data-driven musical interfaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Anna Xambó Sedó (Ed.). UFRGS, Porto Alegre, Brazil, 313–318. https://doi.org/10.5281/zenodo.3672976

[63] Diemo Schwarz. 2012. The Sound Space as Musical Instrument: Playing Corpus-Based Concatenative Synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. University of Michigan, Ann Arbor, Michigan. https://doi.org/10.5281/zenodo.1180593

[64] Hugo Scurto, Baptiste Caramiaux, and Frederic Bevilacqua. 2021. Prototyping Machine Learning Through Diffractive Art Practice. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 2013–2025. https://doi.org/10.1145/3461778.3462163

[65] Hugo Scurto and Ludmila Postel. 2023. Soundwalking Deep Latent Spaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, 232–235. https://doi.org/10.5281/zenodo.11189166

[66] Nicholas Shaheed and Ge Wang. 2024. I Am Sitting in a (Latent) Room. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 333–338. https://doi.org/10.5281/zenodo.13904872
[67] Renee Shelby, Shalaleh Rismani, and Negar Rostamzadeh. 2024. Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642461
[68] Victor Shepardson and Thor Magnusson. 2023. The Living Looper: Rethinking the Musical Loop as a Machine Action-Perception Loop. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, 224–231. https://doi.org/10.5281/zenodo.11189164
[69] Victor Shepardson, Jonathan Reus, and Thor Magnusson. 2024. Tungnaá: a Hyper-realistic Voice Synthesis Instrument for Real-Time Exploration of Extended Vocal Expressions. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, S. M. Astrid Bin and Courtney N. Reed (Eds.). Utrecht, Netherlands, 536–540. https://doi.org/10.5281/zenodo.13904943
[70] Jordie Shier, Rodrigo Constanzo, Charalampos Saitis, Andrew Robertson, and Andrew McPherson. 2025. Designing Percussive Timbre Remappings: Negotiating Audio Representations and Evolving Parameter Spaces. *Proceedings of the International Conference on New Interfaces for Musical Expression* (June 2025), 452–461. https://doi.org/10.5281/zenodo.15698926
[71] Jordie Shier, Charalampos Saitis, Andrew Robertson, and Andrew McPherson. 2024. Real-time Timbre Remapping with Differentiable DSP. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, S. M. Astrid Bin and Courtney N. Reed (Eds.). Utrecht, Netherlands, 377–385. https://doi.org/10.5281/zenodo.13904884
[72] Kenneth O. Stanley. 2007. Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines* 8, 2 (June 2007), 131–162. https://doi.org/10.1007/s10710-007-9028-8
[73] Domenico Stefani, Matteo Tomasetti, Filiippo Angeloni, and Luca Turchet. 2024. Esteso: Interactive AI Music Duet Based on Player-Idiosyncratic Extended Double Bass Techniques. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 490–498. https://doi.org/10.5281/zenodo.13904929
[74] Miriam Sturdee, Makayla Lewis, Angelika Strohmayer, Katta Spiel, Nantia Koulidou, Sarah Fdili Alaoui, and Josh Urban Davis. 2021. A Plurality of Practices: Artistic Narratives in HCI Research. In *Proceedings of the 13th Conference on Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3450741.3466771
[75] Ben Swift. 2025. PANIC!. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Zenodo, 1–4. https://doi.org/10.5281/zenodo.17800975
[76] Steve Symons. 2024. The Perceptron: A Multi-player Entangled Instrument based on Interpretive Mapping and Intra-action.. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures (AM '24)*. Association for Computing Machinery, New York, NY, USA, 385–391. https://doi.org/10.1145/3678299.3678338
[77] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2020. Al-terity: Non-Rigid Musical Instrument with Artificial Intelligence Applied to Real-Time Audio Synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 337–342. https://doi.org/10.5281/zenodo.4813402
[78] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2021. AI-terity 2.0: An Autonomous NIME Featuring GANSpaceSynth Deep Learning Model. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Shanghai, China. https://doi.org/10.21428/92fbeb44.3d0e9e12
[79] Koray Tahiroğlu and Lonce Wyse. 2024. Latent Spaces as Platforms for Sonic Creativity. In *Proceedings of the 15th International Conference on Computational Creativity*. Sweden.
[80] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA.
[81] Mattie Tesfaldet, Xavier Snelgrove, and David Vazquez. 2019. Fourier-CPPNs for Image Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 3173–3176. https://doi.org/10.1109/ICCVW.2019.00392
[82] Nao Tokui and Tom Baker. 2025. Latent Granular Resynthesis using Neural Audio Codecs. In *Extended Abstracts for the Late-Breaking Demo Session of the 26th Int. Society for Music Information Retrieval Conf.* Daejeon, South Korea. https://doi.org/10.48550/arXiv.2507.19202
[83] Christopher J. Tralie and Ben Cantil. 2024. The Concatenator: A Bayesian Approach to Real Time Concatenative Musaicing. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*. ISMIR, 882–889. https://doi.org/10.5281/zenodo.14877471
[84] Pierre Alexandre Tremblay, Gerard Roma, and Owen Green. 2021. Enabling Programmatic Data Mining as Musicking: The Fluid Corpus Manipulation Toolkit. *Computer Music Journal* 45, 2 (June 2021), 9–23. https://doi.org/10.1162/comj_a_00600
[85] Robert Tubb and Simon Dixon. 2014. A Zoomable Mapping of a Musical Parameter Space Using Hilbert Curves. *Computer Music Journal* 38, 3 (2014), 23–33. https://doi.org/10.1162/COMJ_a_00254
[86] Gabriel Vigliensoni, , and Rebecca Fiebrink. 2024. Data- and interaction-driven approaches for sustained musical practices with machine learning. *Journal of New Music Research* 53, 1-2 (March 2024), 19–32. https://doi.org/10.1080/09298215.2024.2442361
[87] Gabriel Vigliensoni and Rebecca Fiebrink. 2023. Steering latent audio models through interactive machine learning. In *In Proceedings of the 14th International Conference on Computational Creativity*. Ontario, Canada. https://doi.org/10.5281/zenodo.8087978
[88] Federico Visi. 2024. The Sophtar: a networkable feedback string instrument with embedded machine learning. *Proceedings of the International Conference on New Interfaces for Musical Expression* (Sept. 2024), 142–148. https://doi.org/10.5281/zenodo.13904810
[89] Aline Weber, Lucas Nunes Alegre, Jim Torresen, and Bruno C. da Silva. 2019. Parameterized Melody Generation with Autoencoders and Temporally-Consistent Noise. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Marcelo Queiroz and Anna Xambó Sedó (Eds.). UFRGS, Porto Alegre, Brazil, 174–179. https://doi.org/10.5281/zenodo.3672914
[90] Simon Welker, Matthew Le, Ricky T. Q. Chen, Wei-Ning Hsu, Timo Gerkmann, Alexander Richard, and Yi-Chiao Wu. 2025. FlowDec: A flow-based full-band general audio codec with high perceptual quality. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/pdf?id=uxDFlPGRLX
[91] Etienne Wenger. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.
[92] Mikael Wiberg. 2018. Material-Centered Interaction Design. In *The Materiality of Interaction: Notes on the Materials of Interaction Design*. The MIT Press. https://doi.org/10.7551/mitpress/10427.003.0006
[93] Elizabeth Wilson, Deva Schubert, Mika Satomi, Alex McLean, and Juan Felipe Amaya Gozalez. 2023. MosAIck: Staging Contemporary AI Performance - Connecting Live Coding, E-Textiles and Movement. In *Proceedings of the 7th International Conference on Live Coding*. Zenodo, Utrecht, Netherlands. https://doi.org/10.5281/zenodo.7843540
[94] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. https://doi.org/10.7488/ds/2645
[95] Matthew Yee-King. 2022. Latent Spaces: A Creative Approach. In *The Language of Creative AI: Practices, Aesthetics and Structures*, Craig Vear and Fabrizio Poltronieri (Eds.). Springer International Publishing, Cham, 137–154. https://doi.org/10.1007/978-3-031-10960-7_8
[96] Victor Zappi and Andrew McPherson. 2018. Hackable Instruments: Supporting Appropriation and Modification in Digital Musical Interaction. *Frontiers in ICT* 5 (Oct. 2018). https://doi.org/10.3389/fict.2018.00026
[97] Jianing Zheng, Nick Bryan-Kinns, and Andrew P. McPherson. 2022. Material Matters: Exploring Materiality in Digital Musical Instruments Design. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 976–986. https://doi.org/10.1145/3532106.3533523
[98] Shuoyang Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2024. A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials. In *Proceedings of The second international workshop on eXplainable AI for the Arts (XAIxArts 2)*. https://doi.org/10.48550/arXiv.2407.04379
[99] Shuoyang Jasper Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2025. Exploring Gestural Affordances in Audio Latent Space Navigation. *Frontiers in Computer Science* 7 (Sept. 2025). https://doi.org/10.3389/fcomp.2025.1575202

A Experiments Details

We present full details for each task in Section 3.2.2.

A.1 Task Details

Our initial effort was implemented on a 1D sound space. A 5-sec audio segment is arranged on a timeline, and the mapping model described in Section 3.1 is trained with time coordinates paired with the latent trajectory encoded from the audio segment. A timeline $V = \{v_1, v_2, \ldots, v_t\} \in [0, 1]^{1\times t}$ is used for training and an alternative timeline V' offset by $\phi \sim \mathcal{N}(0, 1e^{-6})$ for testing.

Three audio datasets are used as sources for audio segments to cover a range of timbre variance: (i) all unmixed drum tracks in MUSDB18 [56], (ii) all unmixed tracks in AAM [50] with instruments that are in the "Bowed" and "Guitars" category, and (iii) speech recording in VCTK [94]. For each audio segment in a collection, an MLP is trained to recreate the latent trajectory of that segment given the training timeline V. For each audio collection, a hyperparameter search [11] is performed to optimise network and training configuration. For M2L, SAO, and FlowDec,

pre-trained model provided by the original works is used across all audio collections. For RAVE, three pretrained models from online repositories [16, 32] are used for three audio collections, respectively. This is due to the common practice of training RAVEs on domain-specific datasets.

A.2 Additional Results

Standard MLPs without Fourier features are used as benchmarks to validate the benefit of applying Fourier feature mappings. Peak Signal-to-Noise Ratio (PSNR) is used to compare the regressed latent trajectories with the original latent trajectories. Table 4a shows that integrating Fourier features offers better latent PSNRs. The better PSNR on RAVE is suspected to be due to its smaller latent dimensionality than the others. The better PSNR on M2L is suspected to be due to its larger compression ratio [51], so the same length audio segment is encoded to fewer latent vectors, resulting in less dense input coordinates.

Fréchet Audio Distance (FAD) [37] is used to evaluate the overall similarity between original audio collections and reconstructed audio collections, which contain those reconstructed from the original latent trajectories (i.e., directly forward passed by encoder-decoder), and the regressed latent trajectories by MLPs. We use the VGGish [31] embeddings to compute the audio distance. Table 4b shows that with Fourier features mapping, the regression of latent trajectories only minimally degrades FAD compared to the encoder-decoder reconstruction, which largely depends on the capacity of the autoencoder itself.

A.3 Training loss frequency components

We inspect the training error with respect to multiple frequency components to verify that the Fourier features are enabling the MLP to learn high-frequency details as claimed. We separate the training error with respect to different frequency components. Figure 10 shows a typical record of training errors: the model without Fourier features converges better in the low-frequency band but fails to converge further on the high-frequency band, whereas the two models that have Fourier features converge faster on all bands. This verifies that the introduction of Fourier features has allowed the model to resemble the high-dynamic nature of latent trajectories, and shows that the Gaussian scale σ is a key configuration that affects the model behaviour, which we elaborate on in Section 3.3.

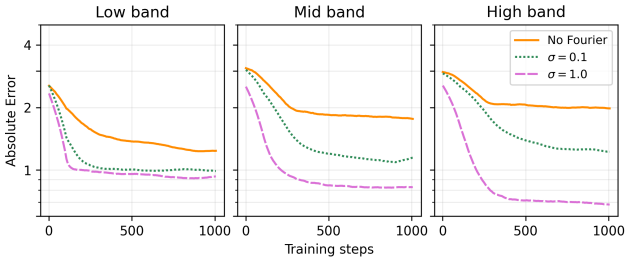


Figure 10: Absolute errors in each training step with respect to three frequency bands from low to high.

A.4 On-the-fly Training and Inference

We present full results for the on-the-fly training and inference evaluation from Section 3.4 in the main text. Results shown in Figure 12 were tested on an Apple Silicon M4 Max MacBook Pro with 48GB RAM, and a Dell G16 7630 with NVIDIA RTX4070 GPU.

B Detailed Milestones and System Design

We present full details for each interface and milestone from Section 4 in the main text, in combination with major change logs and the details of the public engagement events shown in Table 5.

Interface A: Embodied Sound Space Navigation

The first GUI design using `nn.terrain~` focuses on exploring the embodied movement-sound coupling in the sound space. It offers a drawing pad as the control space, in which the stylus (x, y) location is used to navigate in the sound space. The GUI is used with a stylus as the input device to explore real-time sonic feedback in response to stylus movements. It renders the sound space using the visualisation method described in Section 3.3. The interface is presented in a series of "Soundwalking Workshops" in which musicians are tasked to explore the GUI as an instrument and create musical scores on the drawing pad. It was also used in a guest performance that premiered at the Interactive Digital Media Technology concert at Queen Mary University of London.

At this stage, it only marks a minimal viable prototype of using pre-built terrains at the inference stage. Consequently, it relies heavily on programming the MLP mapping model in Python and then exporting it to TorchScript[8] to load in Max. This offline manner of integrating the sound space led to a long waiting time between programming it and actually seeing and hearing it. In this way, the time for musical exploration was spent mostly in navigating the sound space, annotating sonically interesting zones on the drawing pad, and experimenting with expressive ways of re-enacting sounds in these zones.

Interface B: Interactive Sound Space Construction

The second development phase targeted an interactive talk and demo at the IRCAM Forum Workshop 2025, and a guest performance that premiered at Queen Mary Music Festival 2025. The API development in this phase focused on the interactive construction of terrains. We attempted to match the workflow of building terrains in Max/MSP with the workflow in Python. As a result, the cycle of training pairs gathering, model training, monitoring, and inference was distributed to three interdependent Max objects. We considered concepts of coupling and cohesion [3] in software engineering when defining the objects' API. The GUI design focuses on presenting the terrain building workflow in a straightforward way. In particular, loading customised audio samples, plotting trajectories on the 2D plane, and iterative training while monitoring the terrain visualisation.

We considered potential ways of programming the stylus input. Therefore, the drawing pad allows for modulating the stylus coordinates with an external source. In addition, the autoencoder that was previously encapsulated in the system was removed. In this way, `nn.terrain~` only fulfils the coordinate-to-latent mapping, and the autoencoder real-time inference relies on the existing `nn~`[9] object from ACIDS-IRCAM, which has already been widely used by practitioners. Given this flexibility, in Interface B we created the right-most panel that allows modulating the stylus (x, y) as a way of programming the trajectories in a terrain, mimicking the "oscillations and orbits" control in [61].

[8] https://docs.pytorch.org/docs/2.5/jit.html
[9] https://github.com/acids-ircam/nn_tilde

Table 4: Regressing trajectories with the three audio collection across four autoencoders.

(a) Peak Signal-to-Noise Ratio between original latent trajectories and regressed trajectories (higher is better).

	Peak Signal-to-Noise Ratio ↑											
	Drum				String				VCTK			
Regression Config	RAVE	M2L	SAO	FlowDec	RAVE	M2L	SAO	FlowDec	RAVE	M2L	SAO	FlowDec
No Fourier features	27.516	29.774	21.697	23.530	19.463	29.003	20.666	22.837	18.345	29.075	19.459	20.768
Fourier features	**41.369**	**40.368**	**32.069**	**32.566**	**28.328**	**38.966**	**28.024**	**31.997**	**30.591**	**38.104**	**30.786**	**30.052**

(b) Fréchet Audio Distance of audio sets reconstructed from the original trajectories (encoder-decoder reconstruction) and two sets of regressed trajectories, comparing to the original audio samples (lower is better).

	Fréchet Audio Distance ↓											
	Drum				String				VCTK			
Regression Config	RAVE	M2L	SAO	FlowDec	RAVE	M2L	SAO	FlowDec	RAVE	M2L	SAO	FlowDec
Encoder-decoder recon.	7.664	6.659	2.324	2.009	2.718	2.859	2.642	2.102	3.313	5.439	1.546	1.039
No Fourier features	14.188	12.354	28.163	28.069	17.665	9.267	10.542	10.322	25.801	8.451	50.474	19.592
Fourier features	**8.139**	**7.511**	**2.387**	**2.473**	**4.783**	**5.092**	**4.023**	**2.498**	**3.689**	**7.562**	**2.180**	**1.259**

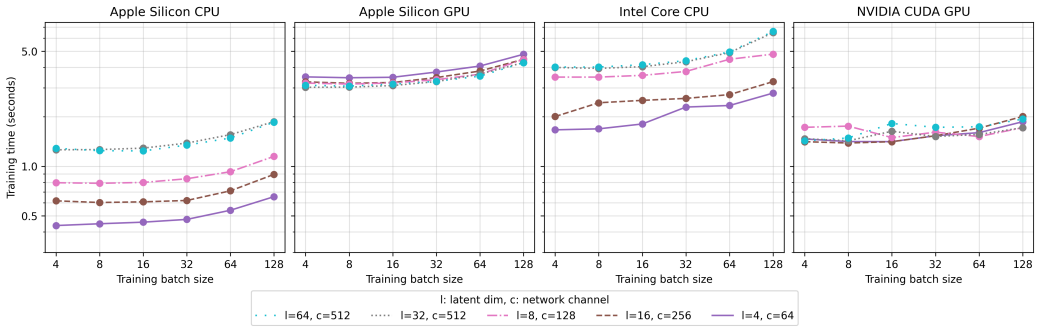


Figure 11: Average time needed to train the MLP for 1000 steps against training batch size, across various MLP configurations and devices.

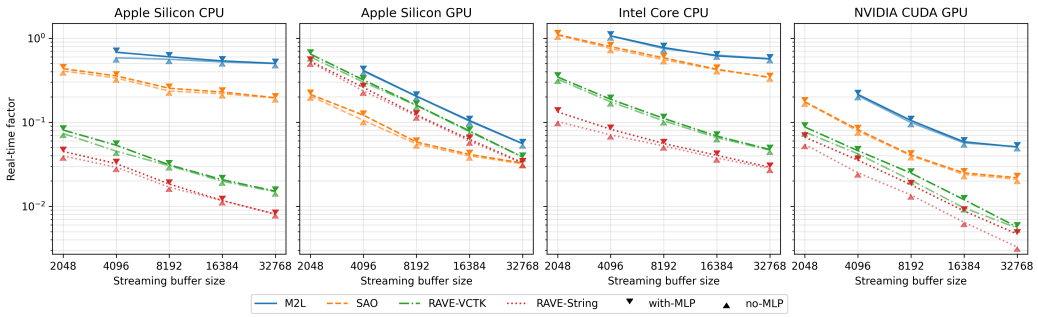


Figure 12: Mean real-time factor (lower the better) against buffer size, across all decoders and devices, comparing with and without the MLP. In all plots, mean values were computed over 100 runs per condition.

Interface C: Programming Trajectory Playback

The third development phase targeted an interactive project exhibition at Sónar+D 2025. We considered potential ways of using the terrains other than the stylus. Therefore, the input (control space coordinates) and output (latent vectors) were exposed to the API to allow for appropriation for other modes of interaction. The initial project idea was to create a "2D Granular Synth" or a "2D Sampler", in which the standard audio sample arranged on a 1D timeline is replaced by a 2D sample canvas. This is practically impossible when sampling in the waveform domain because the sound playback cannot interpolate between spaces that are not filled with audio samples, unless using a discrete corpus-based sound space, as in concatenative synthesis [25, 83]. However, when sampling in the latent domain, the learned MLP mapping model can cover the entire 2D canvas with continuous coordinate input, and also generalise to spaces that are not initially arranged with samples, by interpolation.

Table 5: Milestones of practice-led explorations that have guided the technical development.

Change Logs	`++ Exporting trained mapping model from PyTorch, loading it in MaxMSP`
	`++ Creating mapping visualisation in MaxMSP`
	`++ Tracking the stylus position as input`
	`++ Logging stylus behaviours`
	`++ Encapsulating nn~ for simplicity`				

Mile-stones	**Time**	**Event**	**Type**	**Input Devices**	**Autoencoder**
	Oct 2024	Soundwalking Workshop	Interactive demo	Stylus and tablet interface	RAVE
	Dec 2024	Interactive Digital Multimedia Techniques Concert	Composition and performance	Stylus and tablet interface	RAVE

Change Logs	`++ Defined independent objects nn.terrain~, nn.terrain.encode, and nn.terrain.gui for cohesion`
	`++ Interactive building sound space in MaxMSP: Encoding audio buffers to latents and interactive plotting spatial trajectories`
	`-- Removed nn~ to expose latent signals for appropriation`
	`++ Modulating stylus coordinates to program the trajectory playback`
	`-- Logging stylus behaviours`				

Mile-stones	**Time**	**Event**	**Type**	**Input Devices**	**Autoencoder**
	Mar 2025	Queen Mary Music Festival	Composition and performance	Mixing desk knobs	RAVE
	Mar 2025	IRCAM Forum Workshop	Demonstration	Stylus and tablet interface	RAVE

Change Logs	`++ Refined nn.terrain.gui for trajectory playback`
	`++ Added nn.terrain.record to record trajectory, to record trajectories from higher dimensional control space (e.g., 3D)`
	`++ Exposed more MLP and Fourier feature's hyperparameters`
	`++ Added the checkpoint method to save trained terrain`				

Initial release of the MaxMSP objects package					
Mile-stones	**Time**	**Event**	**Type**	**Input Devices**	**Autoencoder**
	Jun 2025	Sónar+D Project Area	Interactive demo	Touchscreen	RAVE, Music2Latent
Change Logs	`++ Multi-threaded nn.terrain~ for on-the-fly machine learning` `++ Added support for stereo channels to support the autoencoder in Stable Audio Open`				

To achieve this, the *nn.terrain.gui* object was refined. As shown in Figure 5C, the user first loads audio samples and plots trajectories in the "train mode" the same way as the previous interface. Then in the "playback mode" they define a new set of trajectories from the 2D canvas as "latent samples". Trajectories are triggered by the pattern in the step sequencer on the left. In addition, it integrates the M2L [51] model as the underlying autoencoder. As a result, more MLP hyperparameters were exposed to the API to allow for hyperparameter tuning given the versatility of M2L.

C Artists Statements

Here we include the bio of the four artist-researchers (group) who participanted in the practice-based exploration in Section 5.

Keigo Yoshida is a Tokyo-based artist and scientist who integrates neuroscience and computer science to explore novel forms of artistic expression. His notable works include the A/V performances "Propagation" and "Mineral Neurons", both presented at Sónar+D 2025, and "liberated frequencies" (Installation and A/V performance). The latter work was supported by Flying Tokyo 2024, involved collaboration with METI and Rhizomatiks, and was also presented at IRCAM Forum Workshops 2025 Hors-les-Murs.

Nico García-Peguinho is an artist and researcher investigating machine listening and algorithmic composition. His work has been exhibited and performed at venues including Creative Coding Utrecht (live-coding AV performance), Manchester Histories Festival Opening Ceremony (8-channel installation), and MANTIS festival (multichannel fixed compositions). He is a first-year PhD student at the Centre for Digital Music (C4DM), Queen Mary University of London, and a member of the Sensing the Forest project. **Nikhil Bullock** is a London-based musician and researcher whose music is influenced by UK club culture and algorithmic composition. His practice explores generative processes as a tool for sonic exploration. He is currently pursuing an MSc in Sound and Music Computing at Queen Mary University of London.

Jiatong Liu is a Creative Technologist studying at the Creative Coding Institute, University of the Arts London. Their practice and research revolve around Machine Learning, Composition and Film - combining narrative design, spatial design and technology.

Dan Hearn is a creative technologist and musician working with interaction, sound, and machine learning. His work centres on building tools and interfaces that invite exploration rather than automation. He works across software, hardware, and machine learning to develop prototypes, instruments, and installations.