Probing Latent Space Interactions with Real-time Generative Audio Models Through a Physical Controller

Domenico Stefani*
Francesco Ardan Dal Rì*
Luca Turchet
domenico.stefani@unitn.it
francesco.dalri-2@unitn.it
luca.turchet@unitn.it
DISI - Department of Information Engineering and Computer Science
Trento, Italy

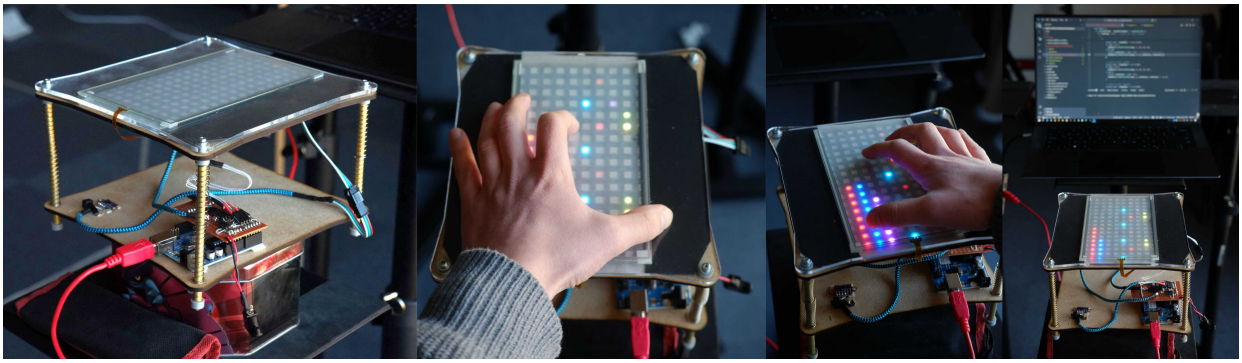

Figure 1: Proposed physical controller for latent space traversal.

Abstract

Recent advances have made deep generative models practical for live music performance. Moreover, latent spaces are increasingly exposed as sites of musical interaction. However, their use as control substrates remains overall under-explored as a first-class design problem. We propose a tangible, reconfigurable controller, conceived as a design probe for investigating the tradeoff between exploration and reproducibility affordances across different mapping and visual feedback strategies. The system is demonstrated through three use cases for as many representative models for real-time synthesis of audio or symbolic music: RAVE, for neural synthesis with high-dimensional opaque latents; MT-GEN_DDSP, exploiting a known latent space with labeled timbre cluster; and GrooveTransformer, for semantically-structured rhythm pattern generation. Each use-case employs distinct sensor-to-latent mappings and visual feedback approaches tailored to the model's characteristics, ranging from 3D latent traversal with interaction heatmaps, to semantic dimension mapping with timbre cluster visualization, to descriptor-guided space navigation. By treating latent spaces as explicit interaction surfaces, rather than implementation details, this work contributes to ongoing discussions about controllability, legibility, and appropriation in machine learning-based musical instruments.

*Both authors contributed equally to this research.

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

Keywords

Neural Audio Synthesis, Latent Space, Embodied Interaction, Real-time Control, Tangible interfaces

1 Introduction

While early deep generative models for music largely operated in offline rendering workflows [17, 51], recent advances in deep learning and computational power have enabled their use as practical, instrument-like systems, capable of real-time audio synthesis, timbre transfer, and symbolic generation [8, 20, 28]. Additionally, the NIME community has a long tradition of using machine learning to build new instruments and performance relationships, which constitutes a natural playground for investigating modes of interaction with such systems [10, 36].

In particular, latent spaces - the internal representations through which deep models encode and organize learned musical concepts - offer a distinctive interaction opportunity, especially when exposed as control parameters. Exposing and navigating latent spaces as an input control can yield learned non-linear transformations [40] that are difficult to access through conventional mappings.

Although several works have begun to address interaction with latent spaces [29, 44, 48, 57], performance-oriented interaction with these remains comparatively under-explored relative to other aspects of generative music systems, especially as a first-class design problem [1, 5]. Latent spaces are often treated as implementation details rather than performance surfaces, as interaction layers frequently default to fader/knob interfaces and simple mappings [3, 6]. In addition, established NIME and HCI perspectives argue that the design focus should be on *interaction*, intended as what people can do, perceive, and learn, rather than on the interface artifact itself [2]. From this viewpoint, latent

spaces are not merely control domains: they are interaction substrates whose navigability, legibility, and appropriability can be designed, tested, and iterated [18].

However, latent spaces as interfaces to sound synthesis display idiosyncratic characteristics such as high-dimensionality, no guaranteed alignment of dimensions to human concepts (i.e., semantic-opaqueness) [7, 49], and entangled / non-orthogonal dimensions [44]. Due to these properties, the use of learned latent spaces as control interfaces may not introduce a fundamentally new interaction challenge, but rather, it recontextualizes within a qualitatively different type of control space the well-established tradeoff between *variability* and *reproducibility* affordances of musical interfaces (concepts close to those of explorability and virtuosity [35]).

Supported by recent findings from other voices in NIME and real-time deep generative audio research [29, 44, 57], we argue that some established interaction design strategies must be reconsidered for the new learned and emergent structure of latent interfaces.

In this paper, we present a physical interface for performance with real-time deep generative audio models, made to explore tradeoffs between explorability and reproducibility afforded by different sensor and visual-feedback mappings - Fig. 1. As such, in this work, we intend the device as a design probe [47], which we apply to three representative use-cases among real-time deep generative models:

- RAVE [8];
- MT-GEN_DDSP [15]
- GrooveTransformer [27, 28]

The 3D interface presented here is composed of a 2D sensing surface suspended over springs, with vertical excursion as the third axis. A matrix of RGB addressable LEDs is used as visual feedback, either displaying a slice of a 3D space or information about the model's output. The source code for the interface, models, and manufacturing designs are provided in the project's repository[1].

2 Background

2.1 Machine Learning and Generative Models in NIME

The use of machine learning as a material for musical instrument design has a long tradition within the NIME community, from early work on gesture recognition and mapping strategies [22], to recent approaches based on deep generative models [36, 48]. In this context, learned models have been repeatedly framed as a means to establish novel performance relationships [23, 53].

Early work demonstrated that autoencoder-based architectures can learn meaningful low-dimensional embeddings of instrument sounds, enabling interpolation and timbre morphing between instruments. For example, Engel et al. [21] showed that a WaveNet-style autoencoder can synthesize realistic musical notes while providing a latent manifold suitable for musical control and timbral interpolation. More recently, advances in model architecture and computational efficiency have significantly improved the feasibility of deploying neural generative models in real-time musical contexts. In particular, RAVE [8] represents a key milestone in this direction, enabling fast, high-quality neural audio synthesis suitable for live performance. In parallel, hybrid approaches such as DDSP [20, 26] combine learned representations with structured synthesis models, offering new perspectives on controllability and interpretability in neural audio systems. Together, these developments reinforce the view of deep generative models as practical, instrument-like systems within NIME practice.

Despite this growing body of work, surveys and meta-analyses of machine learning in NIME point out that interaction design is often treated as secondary to model performance or technical novelty [36]. This motivates a closer examination of how musicians can meaningfully interact with learned representations, particularly latent spaces, in real-time performance settings.

2.2 Latent Spaces Interaction in Musical Systems

The idea of using learned latent spaces from real-time deep generative models as a control surface for musical expression has gained traction in recent years. While early systems often exposed latent variables through simple parameter mappings [30], more recent work has explicitly focused on how performers navigate, explore, and appropriate these spaces during musical interaction. A central challenge lies in providing intuitive, human-centered control over latent spaces, often high-dimensional and semantically opaque. The proof-of-concept system by Vigliensoni and Fiebrink [52] maps a low-dimensional human performance space to the latent space of a neural audio model by training a regression model on a set of demonstrative actions, enabling real-time steering. *Ai-Terity 2.0* by Tahiroğlu et al. [48] explores the latent space of a GANSynth [19] model through a tangible, deformable interface designed to probe control directions through physical interaction. *Nebula*, introduced by Horta Valenzuela and Tomás [29], employs PCA to reorganize RAVE latent spaces into performable timbral dimensions, supported by a graphical interface and sensor-based instruments that enable visualized trajectories and gestural control.

At the same time, intuitivity may not always be the primary goal of interface designers for interfaces to deep generative audio models. *Stacco*, by Privato et al. [44], provides control over latent spaces through a set of magnetic attractors with inherently intertwined magnetic fields, deliberately leaning into the entangled and opaque nature of latent dimensions.

Alternative metaphors for latent space exploration have also been proposed. Scurto and Postel [46] introduced an approach in the form of a sound walk, where users move through a 3D virtual environment to explore a latent sound space generated by deep learning. Complementarily, Zheng et al. [57] investigate the subjective, embodied experience of navigating audio latent spaces using gestural interfaces and graphic scores, providing insights on how musicians develop performance techniques and gestural vocabularies through repeated exploration of latent terrains.

These works demonstrate a growing interest in latent spaces as sites of musical interaction.

2.3 Interaction Design, Tangibility, and Control Intimacy

Beyond model architecture and control mappings, many studies report on the importance of interaction design principles in shaping how musicians engage with complex computational systems [12, 14, 24]. Within HCI and NIME, interaction is often framed not as the manipulation of parameters, but as what users can do, perceive, and learn through sustained engagement [2, 18].

[1] https://github.com/domenicostefani/latent-space-audio-controller

From this perspective, latent spaces can be intended as *interaction substrates* whose navigability and legibility are central design concerns.

A foundational tension in the design of digital musical instruments concerns the balance between explorability and reproducibility. The most comprehensive account of this relationship is given by Jordà [35], who frames it through the lens of expressivity versus virtuosity, declining these into more granular concepts like variability and reproducibility of an instrument's musical output. Moore [41] first introduced *control intimacy* as a performer's perception of the match between the behaviour of an instrument and their capabilities when controlling it. Wessel and Wright [54] expanded on the concept, relating it to virtuosity. Similarly, Hunt et al. [31, 32] often mention instrument-effectiveness and mapping-learnability in their research on instrumental mapping, which are tightly linked to reproducibility. Magnusson [38] introduces an 8D epistemic space for musical devices that provides a framework for analyzing these tensions. Several of its dimensions relate closely to explorability and reproducibility (i.e., *Improvisation+Generality* and Magnusson's *Explorability+Expressive Constraints*). Furthermore, Boden [4] highlights how full explorability without reproducibility can be futile, as some locations of a large search space may be irrelevant to the task at hand.

Research on tangible and physical interfaces for musical interaction provides further grounding for embedding latent space control into physical artifacts. Early work on tangible user interfaces demonstrated how physical affordances can support expressive and exploratory interaction with digital systems [33]. In the musical domain, matrix-based and spatial controllers have long been explored as alternatives to linear parameter mappings, supporting multidimensional navigation and embodied interaction [43].

Recent work on explainable AI in creative practice has motivated designs where visual feedback, spatial metaphors, and interaction traces function as situated, performative explanations that emerge through interaction itself [1, 5]. Across these perspectives, a recurring theme is the need for interaction designs that remain understandable, transferable, and reusable beyond their original technical context. This issue has gained increasing attention in discussions of documentation, reproducibility, and longevity within the NIME community [9, 13, 36].

Where prior work has focused on synthesis quality, algorithmic control strategies, or software-based exploration of latent spaces, *physical, tangible, high-dimensional control surfaces* explicitly designed to traverse and sculpt latent spaces in performance remain relatively underexplored. The present work contributes to this emerging area.

3 Physical Interface

The proposed physical controller is a 3D interface revolving around a 2D resistive touch surface (`GT070E`), a distance sensor, and a matrix of addressable RGB LEDs (`WS2812B` 16x16) for visual feedback. The interface (Fig. 2 left) consists of a flat upper surface (containing the matrix and transparent touch surface) which is mounted on four spring-loaded brass rods. Pressing on the upper surface causes the rods to go through the lower plate and the entire surface to lower, while return is granted by the springs (Fig. 2 right). The touch surface provides 2D data while a time-of-flight distance sensor (`VL53L0X`) measures the third dimension. An Arduino Uno is used to interface with a computer running the

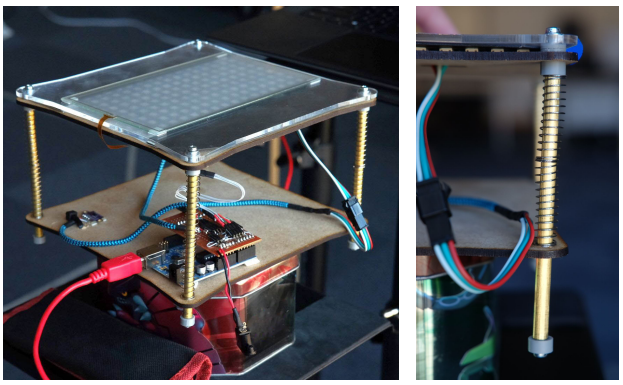

Figure 2: Overall view of the device (left) and detail of the spring suspension system for the moving surface (right).

models. The collapsing surface is inspired by the Trautoniumn's pressure/depth-sensing rails [45].

The interface is composed of low-cost components and material offcuts (MDF wood, acrylic), which were laser-cut at a local FabLab. Additionally, small 3D printed anchors were made to connect rods to the upper surface through screws. Designs are open-source and accessible in the project's repository[1].

4 Three Use Cases

After the development of the hardware controller, we identified three existing real-time capable generative models in the literature, each with peculiar capabilities and musical behaviors. For each, we devised a distinct mapping strategy for the control and visual feedback (Table 1). Such ideas did not arise from a strict design process: they instead emerged spontaneously through hands-on interaction with the chosen models. Previous exploratory use revealed distinctive affordances and limitations, which, in turn, suggested how control parameters and visual feedback could be meaningfully exposed through the interface.

For the sake of conciseness, this section provides an overview of the overall process for adapting models to the interface. For additional technical details of architectures and training pipelines, please refer to the repository[1].

Table 1: Model comparison across mapping strategy and visual feedback

Model	Mapping Strategy	Visual Feedback
RAVE [8]	3D direct	Traversal Heatmap
MT-GEN_DDSP [15]	Dimension upscaling	3D Landmarks
GrooveTransformer [28]	Semantic dimensions	Model informed

4.1 RAVE

A RAVE decoder [8] was chosen as one of the use cases, as a prime example of real-time deep audio synthesis methods still widely used in literature (e.g., [29, 44]. RAVE is a variational autoencoder operating directly on audio signals, capable of shrinking audio frames down to 8D or 4D latent vectors while maintaining good reconstruction metrics. As a result of the neural compression, RAVE's latent space is defined by relevant dimensions that have no semantic meaning or clear mapping to the synthesized sound (e.g., no pitch slider, attack control). While a full RAVE architecture can be used for timbre transfer, a RAVE decoder

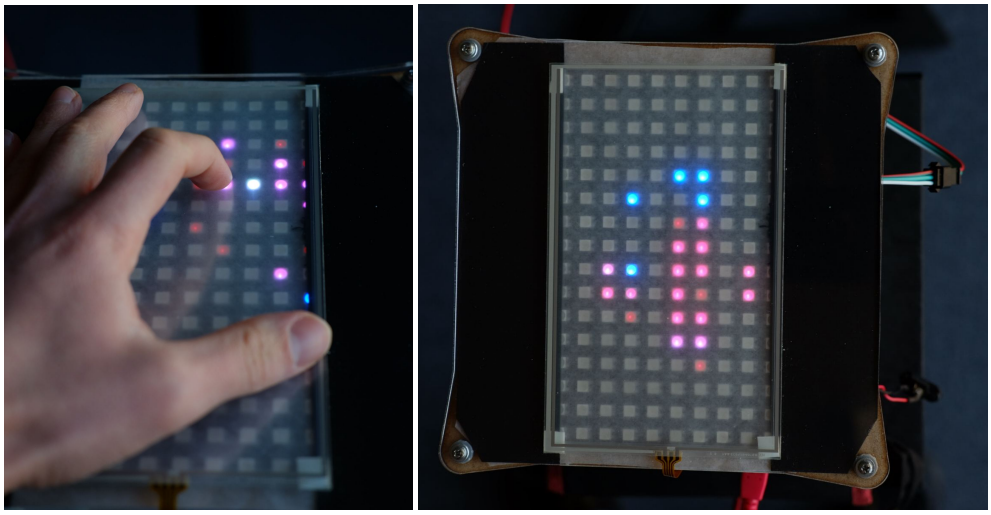

Figure 3: Heatmap visualization method, dynamically showing most visited areas of the 3D space, with a color scale that shows decay of "heat" through time.

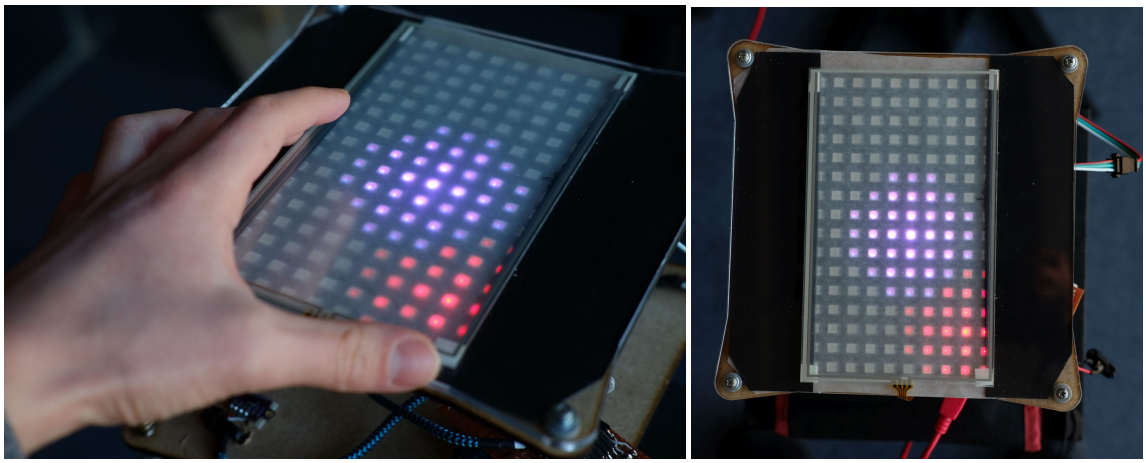

Figure 4: Cluster visualization method showing colored instrument clusters at a specific slice of the 3D space, controlled by the depth motion.

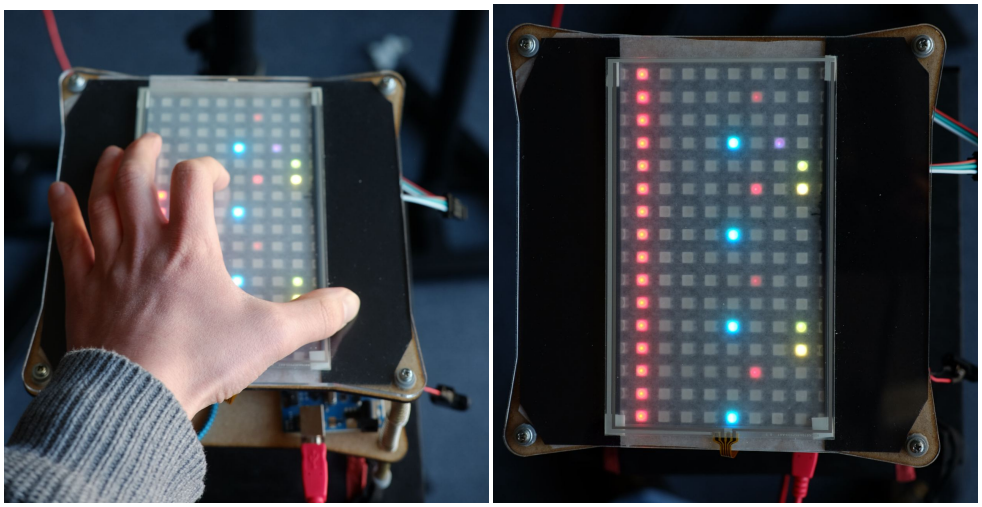

Figure 5: Semantic feedback method for the GrooveTransformer, showing on the left three columns with live feedback from x, y, and depth sensor input, and on the right four rhythm sequence patterns outputted by the model.

alone can be used for synthesis. We chose a pretrained model (i.e., `Sol_ordinario`, from the Acids group[2]). The model is run in a Pure Data (Pd) patch (Plugdata port) through the nn_tilde[3] external. The interface communicates through serial messages with the Pd patch directly (Fig. 6).

4.1.1 Mapping Strategy. RAVE exposes a low-dimensional latent representation (4D in the selected configuration) that can be directly manipulated at inference time, fostering immediate and reactive interaction. Therefore, we adopt a *3D direct latent mapping*, where we map the 3D controller axes to the first three latent dimensions of the RAVE decoder. The 4th dimension can be left fixed or set as a linear combination of the three dimensions in the Pd patch.

4.1.2 Visual Feedback. RAVE's latent dimensions are not explicitly associated with semantic musical attributes; therefore, we choose a visual feedback that could balance the exploration-learning mapping strategy. We implement a heatmap visualization to keep track of the recent and most visited loci (Fig. 3). During performance, the Arduino monitors the current position at 20ms intervals and adds heat to corresponding cells of a downsampled representation of the 3D sensor space. Heat decays in a non-linear fashion in time. This is rendered on the LED matrix to show the slice of the heatmap at the current depth/vertical axis value. Heat is rendered with a color scale, resulting in a colored map that emphasizes regions of sustained activity. Heatmap visual feedback is meant to support orientation and path retracing.

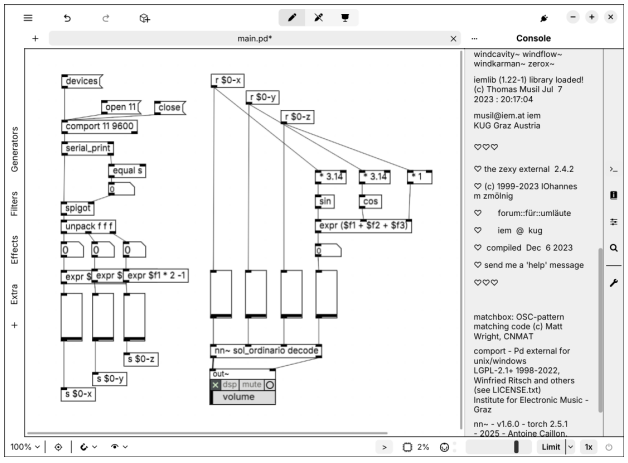

Figure 6: Plugdata patch with a serial message receiver, feeding controls from the physical interface to a XXX RAVE decoder running in nn_tilde.

4.2 MT-GEN_DDSP

MT-GEN_DDSP [15] is a generative architecture in which a DDSP [20] decoder is trained directly from latent representations encoded, offline, by a variational autoencoder (VAE). The VAE learns a disentangled representation that separates timbre and dynamics: this information, along with pitch and envelopes, is jointly used by the decoder to generate audio signals.

The VAE model is explicitly trained to separate the latent space into meaningful timbre clusters and, therefore, already provides an ideal access point for latent control. We first train a small version of an MT-GEN_DDSP model in its 4-timbre configuration, adding a phase-preserving mechanism for continuous audio streaming. Training data consists of instrumental samples from the *TinySOL* [11] dataset, i.e., *horn*, *accordion*, *violin*, and *bass clarinet*. Therefore, the disentangled representation was achieved by exploiting the instrumental labels from the dataset as separate timbre classes. Due to its small size, the model can efficiently be inferenced in real-time (in our case, through a Python script).

4.2.1 Mapping Strategy. We decided to use the controller to navigate the timbral latent space as direct access to this space represents the most peculiar property of the model, leaving pitch, dynamics, and envelope parameters to be controlled by external systems, such as MIDI keyboards or sequencers. Unlike RAVE, where a compact subspace is exposed directly for interaction, the 16-dimensional latent space of MT-GEN_DDSP is trained to encode semantically meaningful timbral factors but does not provide a low-dimensional control space. Consequently, a projection between the latent space and the controller's 3D interaction domain is required.

For each training sample, we extract the corresponding 16-dimensional latent vector and project the full latent dataset into three dimensions using t-SNE. The resulting 3D coordinates, paired with their respective 16-dimensional vectors, are used to train a small MultiLayer Perceptron (MLP) that learns the inverse transformation (i.e., *dimensionality upscaling*). During inference, the controller provides 3D coordinates, which are upscaled by the MLP into 16-dimensional latent vectors and passed to the MT-GEN_DDSP (see Fig. 7).

4.2.2 Visual Feedback. The 3D space generated through t-SNE dimensionality reduction for control mapping is repurposed to visualize the aforementioned meaningful data clusters. Instrumental timbre clusters are color-coded by their label for visual separation in the 3D space. The LED matrix is set to display depth slices of the resulting space, corresponding to vertical-excursion sensor readings.

To obtain cluster positions, we normalize [-1,1] the t-SNE space and compute the center of mass and diameter for each timbre cluster. When the model is loaded, this information, along with color relative to each cluster, is sent to the controller for direct visualization - Fig. 4.

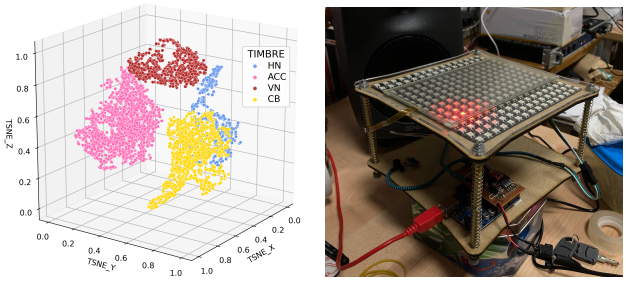

Figure 7: MT-GEN_DDSP mu_t t-SNE latent representation and the respective rendering on the controller. With the surface at the upper position (i.e., not depressed), the top of the sole upper cluster is visible.

4.3 GrooveTransformer

GrooveTransformer [28] is a transformer-based drum pattern generator which infers a bar-level HVO representation, i.e., a

[2]https://acids-ircam.github.io/rave_models_download
[3]https://github.com/acids-ircam/nn_tilde

discrete-time matrix encoding hits (H), velocities (V), and microtiming offsets (O), for multiple drum voices. We use the *light-version* model from the original repository, pretrained on the *Groove MIDI Dataset* (GMD) [25], and segment performances into individual bars (empty bars and outliers are filtered out).

4.3.1 Mapping Strategy. The mapping strategy selected for GrooveTransformer consists of forcing semantic meaning on latent control dimensions. To do so, we extract the encoder memory for each bar and compute three continuous descriptors (density, intensity, timing looseness). An AutoEncoder (AE) is then trained to reconstruct the encoder memory while constraining its first three latent dimensions to match the standardized descriptors, an approach inspired by Fader Networks [37] (see Fig. 8).

During inference, 3D coordinates are mapped to nearby points in this guided space via k-nearest neighbors; one neighbor is sampled, decoded back into encoder memory, and passed through the GrooveTransformer output stage to recover an HVO matrix, which is finally converted to OSC events. OSC messages are used to trigger drum samples.

4.3.2 Visual Feedback. For visual feedback, we chose to provide model-specific information by displaying the generated HVO matrix directly onto the controller surface. Since the model operates in discrete time at the bar level, we realized that it was important to display the upcoming pattern before it is played. We therefore used four columns of LEDs as step sequencers, where colors encode different drum voices, and luminosity reflects velocity. Microtiming offsets are ignored in the visualization. Remaining empty columns were used to display density, intensity, and timing looseness sensor input, helping support proprioception when interacting with the interface (see Fig. 5).

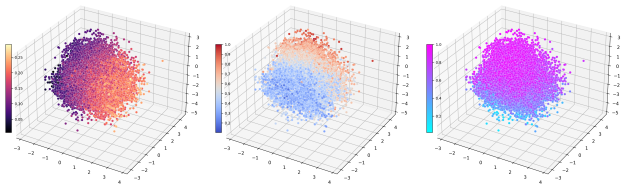

Figure 8: Latent representation of the AE injected into the GrooveTransformer, color-coded by (a) density; (b) intensity; and (c) timing.

5 Design Reflections

Here, we report on our reflections on the design and first interactions with the three distinct approaches to physical control and visual feedback for real-time generative sound models.

5.1 Latent Spaces and Explorability - Riproducibility

The three approaches to latent space control occupy different points along the explorability-reproducibility spectrum [34].

The direct mapping approach used with the RAVE model can be positioned towards the explorability end of the spectrum, as it leans into the semantic opaqueness and entangled nature of latent space dimensions [7, 49]. Despite the deterministic nature of the mapping from 3D control space to latent dimensions, the intertwined nature of the space itself makes it hard to predict exact musical output resulting from a given control input, and was found to encourage exploration. This resonates with *Stacco*'s approach [44], though the entanglement here is confined to the latent dimensions rather than extending to the control interface.

The GrooveTransformer-based approach, on the other hand, is at the reproducibility end of the spectrum. The control mapping is based on forcing structure on the latent space around known and understandable musical dimensions, inspired by descriptor-guided approaches [28, 37]. This is desirable in terms of control and reproducibility, as it allows the performer to have a clear understanding of how their control input will affect the sound output, therefore favoring learnability and intentional repetitions [56].

Finally, the MT-GEN_DDSP-based approach can be positioned in the middle of the spectrum, as the latent space, although high-dimensional, is explicitly structured around timbral semantics through training [15]. The upscaling network offers full control over latent dimensions and supports a certain degree of exploration. In this sense, it relates to the RAVE case, preserving continuous latent traversals, yet it is also close to the GrooveTransformer case, in that semantic, timbre-related landmarks constrain the space and support reproducibility. This strategy is conceptually close to *Nebula* [29], which reorganizes latents towards performability; here, however, landmarks provide orientation without restructuring the space into orthogonal control axes.

5.2 Visual Feedback as Legibility

In the three systems, the visual feedback has been designed to be not solely ancillary to latent control, but as an additional interaction layer that supports how the space is perceived and appropriated [50]. Indeed, each display enacts peculiar forms of legibility, resulting from different assumptions about what musicians need to see to produce meaningful actions. These assumptions were derived from practical considerations that emerged spontaneously during exploratory sessions with the models and mappings.

Moreover, visual feedback strategies resulted also from the relative sensor mapping strategies, either by following a similar concept, or by counterbalancing characteristics of the sensor mapping. Compounded with sensor mapping modalities, these promoted different modes of interaction.

In the case of the GrooveTransformer, mirroring the model output anticipates a preview of the upcoming generated bar, reinforcing predictability and intentionality. With RAVE, the heatmap reflecting history supports self-guided traversal explorations without imposing semantic interpretation. Finally, the fixed landmarks for the MT-GEN_DDSP provide orientation but also leave room for discovery. This resonates with recent arguments that explainability for creative practices can be enacted through situated performative feedback [1, 5].

Finally, visual feedback was found to be more immediate than the sole audio component, thereby allowing for partial offloading of cognitive processes into it while performing, in line with e.g., [16, 39].

5.3 Transversal Considerations

Beyond control mapping reflections, other practical aspects emerged during development. Notably, implementing the RAVE-based approach entirely within a Pd patch enabled rapid prototyping and iteration.

More complex mapping strategies were easily implemented and tested (e.g., for example creating combinations of control axes to control additional latent dimensions). Additionally, RAVE models were easily swapped in and out. This supports practices of appropriation of the instrument and mappings [55]. On the contrary, MT-GEN_DDSP and GrooveTransformer-based approaches were more reluctant to change. For MT-GEN_DDSP, control mapping and visualization are based on a known latent space and require both a reduction and an upscaling step. This makes it more difficult to test different sounds and latent spaces. For GrooveTransformer, the control mapping is intrinsically less flexible.

5.4 Limitations and Future Directions

This paper uses three models as situated probes across the explorability/reproducibility spectrum, consistent with interaction perspectives in NIME and HCI [2, 18]. We reported on development considerations arising from the design process and first experiences with the use cases as performers. Accordingly, we acknowledge the limited scope of preliminary findings on the proposed design probe. Future work will include longitudinal user studies of performer learning, as virtuosity and instrumental identity emerge through sustained practice [34, 42].

Moreover, the system was intentionally designed using relatively low-cost components, offcuts, and other material ready-at-hand, which introduced hardware limitations: the reduced microcontroller memory constrained both the spatial and color resolution of most visual renderings, and the spring mechanism worked but did not provide consistent pressure at all corners. A more refined future iteration should be based on a more apt microcontroller and explore gear-matched lowering mechanisms, striving for robustness and longevity [9, 13].

6 Conclusion

This paper presented a physical controller designed as a probe for exploring latent space interaction in real-time generative audio models. Through three distinct use cases, RAVE, MT-GEN_DDSP, and GrooveTransformer, we demonstrated how different mapping strategies and visual feedback approaches can address the explorability-reproducibility tradeoff inherent in latent space navigation. Each use case occupied a different position along this spectrum: RAVE prioritized exploration through direct 3D mapping with heatmap traces; MT-GEN_DDSP balanced both affordances via dimensionality upscaling and cluster landmarks; and GrooveTransformer emphasized reproducibility through semantically-guided dimensions with anticipatory feedback.

By treating latent spaces as explicit interaction surfaces, this work contributes to ongoing discussions about controllability, legibility, and appropriation in machine learning-based musical instruments.

Ethical Standards

This research did not involve human participants, animal subjects, or any procedures requiring institutional ethical approval. No user studies, interviews, or experiments involving personal data were conducted. The authors declare no conflicts of interest, financial or non-financial, related to the development or evaluation of the physical controller or the system presented in this paper. This work was carried out without external funding. All materials, software, and hardware used in the research were obtained through standard academic or personal resources. The authors affirm that the research adheres to accepted principles of ethical and professional conduct within the NIME community, and no ethical issues were encountered in the conception, design, or reporting of this work.

Acknowledgments

We thank Scarab Extension™ and Garage Manenti for the 3D printing knowledge and tree supports.

References

[1] Jack Armitage, Nicola Privato, Victor Shepardson, and Celeste Betancur Gutierrez. 2023. Explainable AI in music performance: Case studies from live coding and sound spatialisation. In *XAI in Action: Past, Present, and Future Applications*. 13 pages.

[2] Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *Proceedings of the working conference on Advanced visual interfaces*. 15–22.

[3] Frédéric Bevilacqua, Rémy Müller, and Norbert Schnell. 2005. MnM: a Max/MSP mapping toolbox. In *New Interfaces for Musical Expression*. 85–88.

[4] Margaret A. Boden. 2004. *The Creative Mind: Myths and Mechanisms* (2 ed.). Routledge, London.

[5] 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 *NeurIPS 2021 Workshop on Explainable AI for Debugging (XAI4Debugging)*. 14 pages. https://openreview.net/forum?id=GLhY_0xMLZr

[6] Jamie Bullock and Ali Momeni. 2015. Ml. lib: robust, cross-platform, open-source machine learning for max and pure data.. In *NIME*. 265–270.

[7] Antoine Caillon, Adrien Bitton, Brice Gatinet, and Philippe Esling. 2020. Timbre latent space: exploration and creative aspects. In *Proceedings of the 2nd International Conference on Timbre (Timbre 2020)*.

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

[9] Filipe Calegario, João Tragtenberg, Christian Frisson, Eduardo Meneses, Joseph Malloch, Vincent Cusson, and Marcelo M. Wanderley. 2021. Documentation and Replicability in the NIME Community. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 4, 24 pages. https://doi.org/10.21428/92fbeb44.dc50e34d

[10] Baptiste Caramiaux and Atau Tanaka. 2013. Machine learning of musical gestures: Principles and review. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Graduate School of Culture Technology, KAIST, 513–518.

[11] Carmine Emanuele Cella, Daniele Ghisi, Vincent Lostanlen, Fabien Lévy, Joshua Fineberg, and Yan Maresz. 2020. OrchideaSOL: A Dataset of Extended Instrumental Techniques for Computer-Aided Orchestration. In *Proceedings of the International Computer Music Conference*. 48–55.

[12] Alan Chamberlain, Adrian Hazzard, Elizabeth Kelly, Mads Bødker, and Maria Kallionpää. 2021. From AI, creativity and music to IoT, HCI, musical instrument design and audio interaction: a journey in sound. *Personal and Ubiquitous Computing* 25, 4 (2021), 617–620.

[13] 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*. 224–230.

[14] António Correia. 2024. On the Human-AI Metaphorical Interplay for Culturally Sensitive Generative AI Design in Music Co-Creation.. In *IUI Workshops*. 1–6.

[15] Francesco Ardan Dal Rì and Nicola Conci. 2026. Improving Interpretability in Generative Multitimbral DDSP Frameworks via Semantically-Disentangled Musical Attributes. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[16] Francesco Ardan Dal Rí, Francesca Zanghellini, and Raul Masu. 2023. Sharing the Same Sound: Reflecting on Interactions between a Live Coder and a Violinist. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 20, 8 pages. https://doi.org/10.5281/zenodo.11189135

[17] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. arXiv:2005.00341 [eess.AS] https://arxiv.org/abs/2005.00341

[18] Alan Dix. 2007. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. BCS Learning & Development, 4 pages.

[19] Jesse Engel, Karan Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1xT0teAG

[20] Jesse Engel, Chenjie Gu, Adam Roberts, et al. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*. 19 pages. https://openreview.net/forum?id=B1x1ma4tDr

[21] 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*. PMLR, 1068–1077.

[22] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, Vol. 3. 2–1.

[23] 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*. International Conference on New Interfaces for Musical Expression, 280–285.

[24] Anders Gammelgård-Larsen, Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. 2024. Designing for human-AI interaction: Comparing intermittent, continuous, and proactive interactions for a music application. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[25] Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. 2019. Learning to Groove with Inverse Sequence Transformations. In *International Conference on Machine Learning (ICML)*. 10 pages.

[26] Gregorio Andrea Giudici, Franco Caspe, Leonardo Gabrielli, Stefano Squartini, and Luca Turchet. 2025. Distilling DDSP: Exploring Real-Time Audio Generation on Embedded Systems. *Journal of the Audio Engineering Society* 73, 6 (2025), 331–343.

[27] Behzad Haki, Nicholas Evans, and Sergi Jordà. 2024. GrooveTransformer: A Generative Drum Sequencer Eurorack Module. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 39, 5 pages. https://doi.org/10.5281/zenodo.13904848

[28] Behzad Haki, Marina Nieto, Teresa Pelinski, and Sergi Jordà. 2022. Real-Time Drum Accompaniment Using Transformer Architecture. In *Proceedings of the 3rd International Conference on on AI and Musical Creativity*. 10 pages. https://doi.org/10.5281/zenodo.7088343

[29] Moisés Horta Valenzuela and Enrique Tomás. 2025. Nebula: A PCA-Based Method To Explore Rave-Encoded Audio Representations. In *Proceedings of the 22nd Sound and Music Computing Conference (SMC2025), Graz, July 2025*. 49–56. https://doi.org/10.5281/zenodo.15839719

[30] Andy Hunt and Marcelo M Wanderley. 2002. Mapping performer parameters to synthesis engines. *Organised sound* 7, 2 (2002), 97–108.

[31] Andy Hunt, Marcelo M. Wanderley, and Ross Kirk. 2000. Towards a Model for Instrumental Mapping in Expert Musical Interaction. In *Proceedings of the 2000 International Computer Music Conference (ICMC)*. 209–212.

[32] Andy D. Hunt, Marcelo M. Wanderley, and Matthew Paradis. 2002. The importance of Parameter Mapping in Electronic Instrument Design. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (24-26 May, 2002). 88–93. https://doi.org/10.5281/zenodo.1176424

[33] Hiroshi Ishii. 2008. Tangible Bits: Beyond Pixels. In *Proceedings of the 2nd International Conference on Tangible and Embedded Interaction* (Bonn, Germany) *(TEI '08)*. xv–xxv. https://doi.org/10.1145/1347390.1347392

[34] Sergi Jordà. 2004. Digital Instruments and Players: Part I – Efficiency and Apprenticeship. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 59–63. https://doi.org/10.5281/zenodo.1176619

[35] Sergi Jordà. 2004. Digital Instruments and Players: Part II – Diversity, Freedom and Control. In *Proceedings of the 2004 International Computer Music Conference*. International Computer Music Association, 706–710.

[36] Théo Jourdan and Baptiste Caramiaux. 2023. Machine Learning for Musical Expression: A Systematic Literature Review. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 46, 13 pages. https://doi.org/10.5281/zenodo.11189198

[37] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: manipulating images by sliding attributes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. 5969–5978.

[38] Thor Magnusson. 2010. An Epistemic Dimension Space for Musical Devices. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 43–46. https://doi.org/10.5281/zenodo.1177837

[39] Raul Masu and Francesco Ardan Dal Rì. 2023. Visual Representations to Stimulate New Musicking Strategies in Live Coding. *Organised Sound* 28, 2 (2023), 218–230.

[40] Lu Mi, Tianxing He, Core Francisco Park, Hao Wang, Yue Wang, and Nir Shavit. 2021. Revisiting Latent-Space Interpolation via a Quantitative Evaluation Framework. arXiv:2110.06421 [cs.LG] https://arxiv.org/abs/2110.06421

[41] F. Richard Moore. 1988. The Dysfunctions of MIDI. *Computer Music Journal* 12, 1 (1988), 19–28.

[42] Fabio Morreale, Andrew P. McPherson, and Marcelo Wanderley. 2018. NIME Identity from the Performer's Perspective. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 168–173. https://doi.org/10.5281/zenodo.1302533

[43] Dan Overholt. 2001. The MATRIX : A Novel Controller for Musical Expression. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 38–41. https://doi.org/10.5281/zenodo.1176372

[44] Nicola Privato, Victor Shepardson, Giacomo Lepri, and Thor Magnusson. 2024. Stacco: Exploring the Embodied Perception of Latent Representations in Neural Synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 62, 8 pages. https://doi.org/10.5281/zenodo.13904899

[45] Oskar Sala. 1950. Das Mixtur-Trautonium. *Physikalische Blätter* 6, 9 (1950), 390–398. https://doi.org/10.1002/phbl.19500060902

[46] Hugo Scurto and Ludmila Postel. 2023. Soundwalking Deep Latent Spaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 33, 4 pages. https://doi.org/10.5281/zenodo.11189166

[47] Koray Tahıroğlu, Thor Magnusson, Adam Parkinson, Iris Garrelfs, and Atau Tanaka. 2020. Digital Musical Instruments as Probes: How computation changes the mode-of-being of musical instruments. *Organised Sound* 25, 1 (2020), 64–74.

[48] 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*. Article 80, 20 pages. https://doi.org/10.21428/92fbeb44.3d0e9e12

[49] Kıvanç Tatar, Kelsey Cotton, and Daniel Bisig. 2023. Sound design strategies for latent audio space explorations using deep learning architectures. In *Proceedings of the Sound and Music Computing Conference (SMC 2023)*.

[50] Sam Trolland, Alon Ilsar, and Jon McCormack. 2025. Visually-Led Design for Gestural Audiovisual Instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 328–336.

[51] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *Proc. SSW 2016*. 125–125.

[52] Gabriel Vigliensoni and Rebecca Fiebrink. 2023. Interacting with Neural Audio Synthesis Models Through Interactive Machine Learning. In *Proceedings of the 1st International Workshop on Explainable AI for the Arts (XAIxArts) at C&C 2023*. 1 pages. arXiv:2310.06428 [cs.AI] https://arxiv.org/abs/2310.06428 ACM.

[53] Federico Ghelli Visi and Atau Tanaka. 2021. Interactive machine learning of musical gesture. In *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity*. Springer, Cham, Switzerland, 771–798.

[54] David Wessel and Matthew Wright. 2002. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal* 26, 3 (2002), 11–14.

[55] Victor Zappi and Andrew McPherson. 2014. Dimensionality and Appropriation in Digital Musical Instrument Design. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 455–460. https://doi.org/10.5281/zenodo.1178993

[56] Shuoyang Jasper Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2025. Exploring gestural affordances in audio latent space navigation. *Frontiers in Computer Science* 7 (2025), 1575202.

[57] Shuoyang Jasper Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2025. Exploring gestural affordances in audio latent space navigation. *Frontiers in Computer Science* Volume 7 (2025), 24 pages. https://doi.org/10.3389/fcomp.2025.1575202