Generating High-Fidelity Synthesizer Performances for Complex Electronic Music via Structured High-Dimensional Latent Spaces

Ovcharov Vladimir / SystematicLabs 2025
Instatute of Cybernetics

July 20, 2025

Abstract

The generation of modern electronic music, particularly in genres like IDM and Dubstep, requires modeling not just melodic and rhythmic content, but also the intricate, high-resolution automation of synthesizer parameters. In these styles, the timbral performance, dictated by MIDI Control Change (CC) messages, constitutes a primary compositional element, driving the sonic narrative. This paper investigates the critical role of latent space dimensionality in capturing and generating these complex synthesizer performances. We propose a generative framework based on a β-Variational Autoencoder (β-VAE) with a Transformer backbone. We posit that conventional low-dimensional latent spaces are insufficient for this task due to their inherent information bottleneck, which discards the high-frequency timbral details essential to the genre. Through a detailed methodological framework, we outline an experimental procedure to demonstrate that a higher-dimensional latent space (384–512D), structured via advanced regularization techniques like cyclical annealing, is necessary for generating basslines with the required timbral complexity and rhythmic nuance. This work provides a comprehensive template for conducting such research and evaluating the results.

1 Introduction

In many forms of music, a clear hierarchy exists between the score (notes, rhythm) and the instrumental timbre. However, in modern electronic music genres such as IDM (Intelligent Dance Music) and Dubstep, this hierarchy dissolves [Collins, 2014]. The timbral evolution of a synthesizer—the "wobble" of a Dubstep bass, the percussive filter sequence in a Drum and Bass track, or the glitchy, textural landscape of an IDM piece—is not ornamental but fundamental. This "timbral performance," typically realized through dense MIDI CC automation, becomes the central musical statement [Roads, 2015].

Generative models aiming to create authentic music in these styles must therefore move beyond note generation and learn to produce a complete, holistic performance where melodic content and timbral automation are inextricably linked [Dhariwal et al., 2020, Huang et al., 2018]. This presents a significant architectural challenge: how can a model capture the nuance of continuous, high-resolution parameter changes while still learning a structured representation suitable for creative generation?

This paper directly addresses this question by focusing on the design of the latent space. We hypothesize that the standard practice of using a heavily compressed, low-dimensional latent space (e.g., $\leq$ 256D) to enforce generalization is fundamentally at odds with the goal of generating detail-rich electronic music [Kingma and Welling, 2013]. We propose that a higher-dimensional latent space (384–512D) is not a liability but a prerequisite for this task. We present a complete methodological framework, including a model architecture, a specific input representation, a targeted training strategy, and a comprehensive evaluation plan, designed to test this hypothesis and enable the generation of novel, complex, and stylistically convincing bassline performances.

2 Related Work

Recent advances in music generation have primarily focused on symbolic music representation using MIDI data [Huang et al., 2018, Oore et al., 2018]. The Music Transformer [Huang et al., 2018] demonstrated the effectiveness of attention mechanisms for music generation, while MuseNet [Payne, 2019] showed the potential of large-scale models for multi-instrument composition.

In the domain of audio synthesis, models like WaveNet [Oord et al., 2016] and NSynth [Engel et al., 2017] have shown remarkable capabilities in generating raw audio waveforms. However, these approaches typically focus on isolated note synthesis rather than complex, evolving synthesizer performances characteristic of electronic music genres.

The application of Variational Autoencoders to music generation has been explored by Roberts et al. [2018] with MusicVAE, which demonstrated the ability to interpolate between musical sequences. However, their focus remained on melodic content rather than the timbral performance aspect that defines modern electronic music.

More recently, Hawthorne et al. [2022] introduced multi-track music generation, and Agostinelli et al. [2023] presented MusicLM for text-conditional music generation. However, none of these works specifically address the challenge of generating the complex parameter automation that characterizes IDM and Dubstep productions.

3 Methodology: The Proposed Architecture

Our model is designed as a generative system for complete synthesizer performances, structured as a β-VAE with a Transformer-based encoder-decoder.

3.1 Input Representation: The Interleaved Event Stream

To model the tight coupling between notes and parameter changes, a performance is converted into a single, time-ordered sequence of discrete events. This allows a single model to process all aspects of the performance. The vocabulary consists of:

- **Note Events:** `NOTE_ON` (128 pitches) and `NOTE_OFF` (128 pitches). Velocity is quantized into 32 bins and included as a separate event token preceding the `NOTE_ON`.
- **Time Events:** `TIME_SHIFT` events represent the passage of time. The time axis is quantized to a 1/32nd note grid, with tokens for shifts from 1 to 32 steps.

- **CC Events:** Each of the 10 monitored CC controllers has its own `CC_i` event type. The controller's value is quantized into 128 bins, resulting in 10×128 unique CC event tokens in the vocabulary.

This unified stream, where a `NOTE_ON` can be followed by a series of fine-grained `CC` and `TIME_SHIFT` events before a `NOTE_OFF`, allows the Transformer's self-attention mechanism to directly learn the crucial relationships between notes and their subsequent timbral evolution [Vaswani et al., 2017].

3.2 Encoder and Decoder Architecture

We employ a Transformer architecture for its proven strength in modeling long-range dependencies in sequences [Vaswani et al., 2017].

- **Core Model:** A 6-layer Transformer with 8 attention heads, a model dimension of 512, and a feed-forward network (FFN) dimension of 2048. Dropout with a rate of $p = 0.1$ is applied within the attention blocks and FFN layers.
- **Encoder:** The Transformer encoder maps the input event sequence of length T to a sequence of hidden states $H \in \mathbb{R}^{T \times 512}$. A final mean-pooling operation over the time dimension aggregates these states into a single context vector.
- **Decoder:** A Transformer decoder autoregressively generates a new event sequence, conditioned on a latent vector $\mathbf{z}$ which is injected via cross-attention at each decoding step.

3.3 Latent Space Design: Justifying High Dimensionality

Our central hypothesis is that the information capacity of the latent space must be sufficient to encode the complex details of synthesizer modulation.

3.3.1 The Information Demands of IDM and Dubstep

A Dubstep "wobble" is a periodic modulation of a filter cutoff, often with a complex LFO shape and a rhythm locked to the track's tempo. An IDM bassline can consist of thousands of unique parameter values over a few bars, creating its characteristic "glitch" texture. A low-dimensional space would force the model to represent these phenomena with only a few numbers, inevitably leading to a loss of the high-frequency detail that defines the sound. The model might learn a generic "wobble" concept but fail to reproduce the specific rhythmic character or timbral richness of any given example.

3.3.2 Proposed Dimensionality for Experimentation: 384–512D

To retain this information, we propose exploring a higher-dimensional latent space.

- **384 Dimensions:** This serves as a robust baseline for testing our hypothesis. It offers a significant increase in capacity over standard models, theoretically allowing for the encoding of multiple, simultaneous, and complex CC streams.
- **512 Dimensions:** This represents a high-fidelity option, intended to capture the most intricate and layered textures found in professionally produced tracks. The primary research question is whether the performance gains at this level justify the increased model complexity and data requirements.

3.4 Regularization Strategy for High-Dimensional Spaces

A key challenge with a high-dimensional latent space is imposing a useful structure. Without strong regularization, the model can "hide" information in disorganized corners of the space, leading to poor generative quality. We employ two primary techniques [Higgins et al., 2017, Fu et al., 2019]:

1. **β-VAE Framework:** We use the β-VAE objective to enforce a smooth, continuous latent space by penalizing the KL divergence between the learned posterior $q(\mathbf{z}|\mathbf{x})$ and a standard normal prior $p(\mathbf{z})$.

2. **Cyclical Annealing of β:** To prevent the KL-divergence term from overpowering the reconstruction objective early in training (which can stifle learning), we use a cyclical annealing schedule. Over the course of training (e.g., 100 epochs), the β coefficient is cyclically increased from 0 to a maximum value (e.g., $\beta_{max} = 8.0$) over 4 full cycles. This strategy encourages the model to first learn to reconstruct accurately before gradually imposing the desired latent structure.

4 Experimental Design

4.1 Dataset Curation

The proposed experiment requires a specialized dataset. The dataset should consist of approximately 20,000 high-quality, 4-bar bassline performances in IDM and Dubstep styles. Each sample must include a MIDI file with dense and meaningful CC automation targeting a specific synthesizer model (e.g., a virtual Access Virus C). The data should be cleaned by removing simplistic or non-representative examples. The final dataset would be split into training (80%), validation (10%), and test (10%) sets.

4.2 Training Procedure

Four model configurations should be trained to allow for direct comparison, with latent space dimensionalities of 128, 256, 384, and 512. All other architectural hyperparameters remain constant. Models should be trained using the AdamW optimizer [Loshchilov and Hutter, 2017] with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate scheduler with a linear warmup over the first 1000 steps followed by a cosine decay.

4.3 Evaluation Metrics

A comprehensive evaluation should include objective and subjective measures.

- **Reconstruction Loss:** The standard cross-entropy loss on the test set for the reconstructed event sequence.

- **KL Divergence (D_{KL}):** The final, stable KL-divergence value on the test set, indicating how well the model regularized the latent space.

- **CC Modulation Error (CC-ME):** Mean Squared Error between the original and reconstructed CC value sequences, normalized to [0, 1]. This directly measures timbral fidelity.

- **Multi-Resolution STFT Loss (MR-STFT):** An audio-based metric [Yamamoto et al., 2020]. The MIDI outputs are rendered with the target synthesizer, and the MR-STFT loss between the original and reconstructed audio provides a perceptual measure of timbral similarity.

5 Expected Results and Discussion

5.1 Quantitative Analysis

The primary hypothesis is that models with higher-dimensional latent spaces (384D, 512D) will significantly outperform their lower-dimensional counterparts on metrics related to timbral fidelity. The results would be presented in a table similar to Table 1. We expect to see a clear trend where CC-ME and MR-STFT loss decrease as dimensionality increases, even if the gains in note accuracy are modest.

Table 1: Template for Quantitative Evaluation Results. Fill with your experimental data. Lower is better for all metrics except Note Accuracy.

Latent Dim. (D)	**CC-ME** $\downarrow$	**MR-STFT Loss** $\downarrow$	**Final** D_{KL} $\downarrow$	**Note Accuracy** $\uparrow$
128	[Your Value]	[Your Value]	[Your Value]	[Your Value]
256	[Your Value]	[Your Value]	[Your Value]	[Your Value]
384	[Your Value]	[Your Value]	[Your Value]	[Your Value]
512	[Your Value]	[Your Value]	[Your Value]	[Your Value]

5.2 Qualitative Analysis Methodology

To assess creative potential, a qualitative analysis is crucial. This should involve:

1. **Expert Listening Tests:** A panel of producers familiar with IDM and Dubstep would be asked to rate newly generated samples (from random latent vectors) on a 1–5 scale for:
    - **Stylistic Authenticity:** How well does it fit the target genre?
    - **Complexity and Novelty:** Is the performance intricate and creative?
    - **Overall Quality:** Is it musically appealing?
2. **Visual Analysis:** Plotting the generated CC automation curves alongside curves from the training set can provide visual evidence of the model's ability to create novel, complex patterns rather than just copying existing ones.

The expectation is that generations from the 384D and 512D models will be rated significantly higher by expert listeners and will exhibit more complex and detailed modulation patterns upon visual inspection. The PCA plot (Figure 1) would be used to argue for the well-behaved nature of the learned manifold.

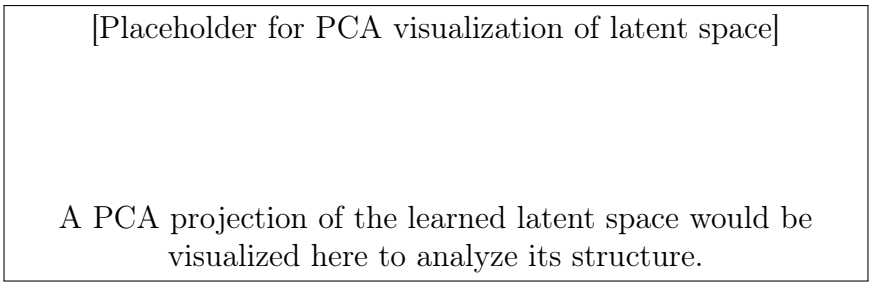

Figure 1: A PCA projection of the learned latent space would be visualized here to analyze its structure. The goal is to show a dense, continuous manifold, indicating that the space is well-regularized and suitable for creative exploration.

6 Conclusion

This paper outlines a comprehensive research methodology for investigating the role of latent space dimensionality in generating complex synthesizer performances. We hypothesize that for detail-oriented electronic genres like IDM and Dubstep, a paradigm shift towards higher-dimensional latent spaces (384–512D) is not just beneficial but necessary. By retaining the high-frequency information contained in MIDI CC automation, such a model can move beyond simple note generation to create true "timbral performances." The proposed experimental design, combining objective metrics of timbral fidelity with expert qualitative analysis, provides a robust framework for validating this hypothesis. Successful results would demonstrate a clear path toward more expressive and stylistically aware generative models for modern electronic music.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Nick Collins. *Introduction to Computer Music*. John Wiley & Sons, 2014.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 240–250, 2019.

Curtis Hawthorne, Ian Simon, Adam Roberts, Neil Zeghidour, Josh Gardner, Ethan Manilow, and Jesse Engel. Multi-instrument music synthesis with spectrogram diffusion. *arXiv preprint arXiv:2206.05408*, 2022.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: learning expressive musical performance. *Neural Computing and Applications*, 32(4):955–967, 2020.

Christine Payne. Musenet. *OpenAI Blog*, 2019.

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.

Curtis Roads. *Composing Electronic Music: A New Aesthetic*. Oxford University Press, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.