MambaKit: Towards Modular Intelligence - Cognitive Scaffolding in Next-Generation Drum Synthesis

Ófeigur Atli Steindórsson
University of Iceland
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
Reykjavík
oas14@hi.is & ofeigur95@gmail.com

Abstract

We present MambaKit, a hybrid neural-deterministic one-shot drum synthesizer that preserves human creative control while achieving high-quality sample generation through cognitive scaffolding, addressing the tension between AI creative capability and human controllability in percussive sound design. Current drum synthesis forces a false choice: accept black-box generation with limited control, or use traditional synthesis requiring years of specialized knowledge. Our system resolves this through structured priors, sine anchors derived from root notes, pitch envelopes, and ADSR parameters automatically extracted during training and exposed as interpretable controls during inference. A structured diffusion framework combines these harmonic anchors with learned noise injection, while the frequency-aware MAMBA2 architecture achieves up to 256x memory efficiency by matching temporal processing windows to signal characteristics, enabling 44.1kHz raw audio synthesis on a single A100 GPU with production-quality output emerging from the first training batch (size of 2), though with occasional instabilities that diminish with minimal additional training. The system preserves human creative agency through interpretable musical parameters while AI handles acoustic complexity. Complete drum arrangements created exclusively using MambaKit samples demonstrate real-world utility. Results suggest sustainable AI creativity emerges from structured human-AI partnerships rather than pure computational scaling.

1 Introduction

What happens to creative agency when AI handles sound design? Current approaches force musicians to choose between surrendering control to black-box systems or mastering complex traditional synthesis. MambaKit demonstrates a third path: **cognitive scaffolding** that preserves human creative intent while enabling AI acoustic synthesis.

This leads to our central research question: Can we develop a generative audio system that produces high-quality one-shot sounds with the real-time control and parameter interpretability that music producers require, while operating within practical computational constraints? Our frequency-aware MAMBA2 architecture achieves this on accessible hardware (single A100 40GB) with production-quality output emerging from the first training batch (size 2) (Figure 1), though with occasional instabilities that diminish with minimal additional training, while demonstrating up to 256x memory efficiency over naive approaches.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

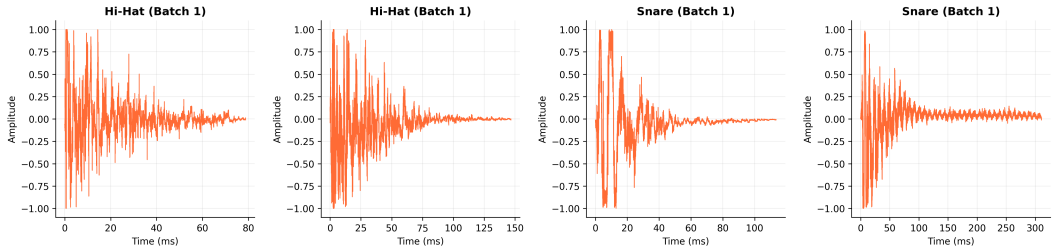

Figure 1: First-batch generation results showing immediate production-quality output. Hi-hat samples (left two panels) exhibit characteristic sharp attacks and rapid decay, while snare samples (right two panels) display complex attack transients with sustained noise content. Natural timbral variation within drum types demonstrates that cognitive scaffolding enables both musical competence and creative diversity from minimal training.

Human-AI Collaboration: Musicians control familiar parameters (MIDI notes, ADSR curves, and text prompts), while AI handles acoustic complexity. This division leverages human musical intentionality and AI's acoustic modeling capabilities. Unlike text prompts that capture broad characteristics ("punchy kick"), our approach provides precise temporal control through extracted envelope parameters along with text prompts.

Intentional Overfitting for Style Transfer: The system learns general synthesis principles from diverse data but can deliberately overfit to stylistic subsets, enabling generation within defined aesthetics using curated training data.

Efficiency Through Structure: Cognitive scaffolding accelerates training by providing harmonic relationships as structured priors rather than forcing the model to rediscover fundamental musical structure. This architectural innovation demonstrates that sustainable AI creativity emerges from human-AI partnerships rather than computational scaling.

2 Related Work

Neural Audio Synthesis Evolution. Early approaches used autoregressive models like WaveNet [1] for high-quality generation but required thousands of sequential steps, making inference prohibitively slow. GANs addressed speed through parallel generation, GANSynth [2] achieved superior quality and speed by operating in spectral domain, but struggled with temporal coherence in raw audio.

Large-Scale Approaches. Transformer-based systems like Jukebox [**Jukebox2020**] and MusicLM [**MusicLM**] generate long-form audio through hierarchical VQ-VAE compression but require massive computational resources (Jukebox used 512 V100s for training). Recent diffusion approaches like DiffWave [3] and Stable Audio [4] operate in latent spaces for efficiency but still require iterative denoising steps. Music ControlNet [5] extends diffusion-based generation with time-varying controls over melody, dynamics, and rhythm, demonstrating that structured musical priors improve controllability in spectral-domain synthesis.

State-Space Models. MAMBA2 [6] addresses Transformer limitations through selective state-space layers, achieving linear complexity and 5× throughput improvements while maintaining quality. Unlike iterative approaches, MAMBA2 generates sequences in a single forward pass, enabling real-time applications.

Positioning. While existing work focuses on either scaling (large models, massive compute) or speed (simplified architectures), none addresses the fundamental controllability problem for creative applications. Our approach uniquely combines MAMBA2 efficiency with structured musical priors, in a diffusion inspired approach, to enable both professional quality and intuitive control, demonstrating that cognitive scaffolding can achieve superior results compared to pure end-to-end scaling approaches.

3 Methodology

3.1 Cognitive Scaffolding Architecture

MambaKit employs a hybrid neural-deterministic approach where musical structure is provided as explicit priors rather than learned implicitly (Figure 2). The system automatically extracts ADSR envelopes and pitch trajectories from training data using Hilbert transform analysis and parametric optimization, then exposes these as interpretable controls during inference. This cognitive scaffolding enables the model to focus computational resources on acoustic texture synthesis while preserving human creative agency through familiar musical parameters.

The complete architecture (Figure 2) demonstrates the dual-timestep diffusion process with structured noise injection, channel-wise parameter conditioning, and parallel CNN/MAMBA2 processing paths that enable both rapid convergence and high-quality output.

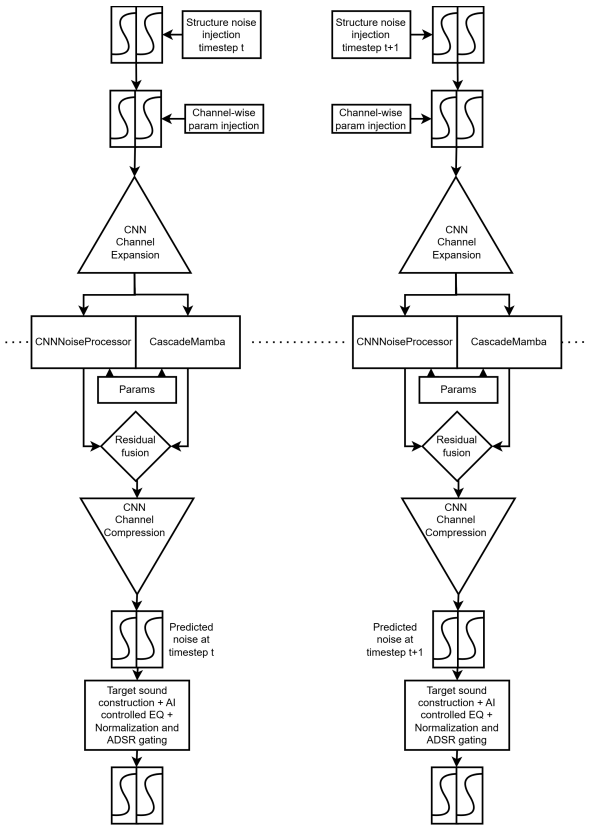

Figure 2: MambaKit architecture overview showing the complete diffusion pipeline with structured noise injection, parameter conditioning, and dual-path processing for cognitive scaffolding.

3.2 Frequency-Aware MAMBA2 Processing

The LogCascadeMambaAdvanced block (Figure 3) addresses a fundamental challenge in raw audio modeling: capturing a single 20Hz oscillations at 44.1kHz requires maintaining temporal state across 2,205 samples, while high-frequency transients demand precise localization. Our solution decomposes audio into logarithmic frequency bands (22050Hz down to 43Hz), with each band processed at optimal temporal resolution through dynamic downsampling.

Each frequency band operates multiple Mamba2 blocks with varying state space sizes (4, 8, 16, 32, 48, 64, 80, 160), where state space allocation matches frequency requirements: $s_{required} = \frac{4 \cdot f_s}{f_{target}}$ for capturing 4 complete oscillations. This approach achieves up to 256× memory efficiency for 20Hz

wave capture (assuming 4 oscillations provide adequate representation) compared to single large state space approaches while maintaining equivalent frequency resolution.

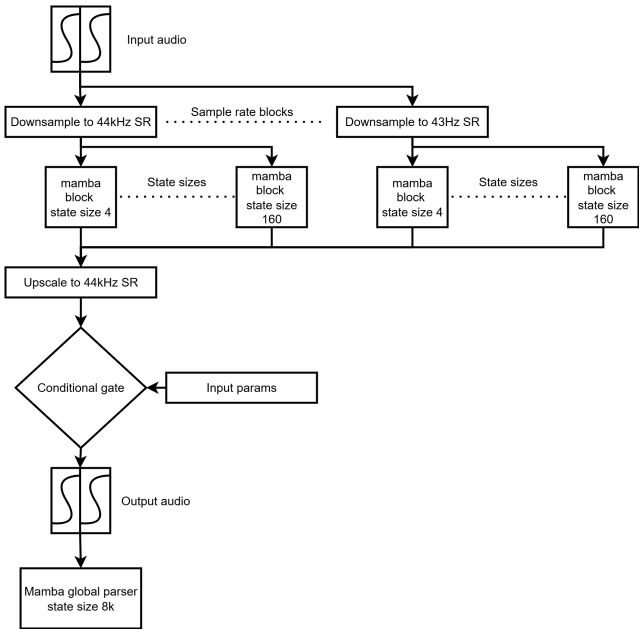

Figure 3: LogCascadeMambaAdvanced architecture showing frequency-decomposed processing paths with varying sample rates and state sizes, enabling up to 256x memory efficiency while preserving frequency resolution.

Outputs from all frequency paths are dynamically weighted via SkipGatesStack using musical parameter embeddings (ADSR, note, octave, text). Only paths exceeding a 0.5 threshold contribute to the final output, enabling adaptive processing based on musical context. A global Mamba2 parser with $d_{state} = 8192$ applies cross-frequency-band processing as a residual connection, as shown in the complete system overview (Figure 2).

3.3 Structured Diffusion Framework

Rather than starting from pure Gaussian noise, our structured diffusion combines parametrically generated sine anchors with learned noise injection. Sine anchors are synthesized using extracted pitch ADSR envelopes:

$$f(t) = 440 \cdot 2^{(p(t)+12 \cdot octave + note - 69)/12} \tag{1}$$

where $p(t)$ represents the time-varying pitch envelope in semitones. Phase coherence is maintained through incremental accumulation: $\phi(t) = 2\pi \sum_{i=0}^{t} f(i)/f_s$.

The diffusion process employs dual scheduling: cosine schedule β_t for noise injection and linear schedule α_t for sine anchor strength. At each timestep, the noisy signal combines both components:

$$x_{noisy} = x_{clean} + (noise \cdot \beta_t + sine \cdot \alpha_t \cdot mask) \tag{2}$$

ADSR conditioning is applied progressively, with envelope strength decreasing as noise injection reduces:

$$\text{ADSR}_t = \text{ADSR} \cdot \beta_t + (1.0 - \beta_t) \tag{3}$$

where β_t represents the noise schedule. This ensures that early diffusion steps (high noise, high β_t) apply strong ADSR conditioning to guide the structured generation process, while later steps (low noise, low β_t) reduce envelope influence to prevent over-shaping as the model refines acoustic details with minimal noise injection.

3.4 Training and Inference

The system employs dual supervision, training to both predict injected noise and reconstruct target signals. This combines the convergence benefits of noise prediction with the perceptual quality of direct reconstruction. Training converges rapidly with musically coherent outputs emerging within 50 batches of size 2, validating the efficiency gains from structured conditioning over pure end-to-end approaches.

During inference, users control generation through interpretable parameters (MIDI notes, ADSR curves, text descriptions) while the system handles acoustic complexity, enabling immediate creative productivity without requiring deep technical expertise in sound synthesis. By default, generation starts from silence, but the system also accepts input audio as a starting point, similar to img2img functionality in Stable Diffusion [7].

4 Results

Musical Validation. Six complete drum compositions created exclusively using MambaKit samples demonstrate practical creative utility **Audio examples:** `https://soundcloud.com/od-e-1/sets/mambakit-neural-drum-machine`. Tracks 1-2 demonstrate curated style transfer through intentional overfitting, where the model sees each drum multiple times, with minimal post-processing (low/high cuts and compression). Tracks 3-6 showcase generalization capabilities from ablation runs trained on 500 random samples from a 6k sample library, with 250 samples generated after each batch (size 2) and no drum seen twice. These tracks begin with completely raw model output, followed by silence, then demonstrate the same samples with minimal post-processing. Standard mixing processes were applied, but no AI artifact correction was required, validating production-ready quality across both training paradigms.

Architectural Component Analysis. For supporting evidence, frequency analysis done on generated data from systematic ablation studies across 6 model variants and training data baseline demonstrate the necessity of each component (Figure 4). Removing sine anchors (NoSine) severely degrades fundamental frequency preservation despite achieving superior numerical reconstruction loss, producing a more even low-end frequency spread that results in muddy kick drums lacking sub-bass clarity. ADSR conditioning proves critical for percussive behavior, its removal (NoADSR) produces sustained oscillations rather than transient drum characteristics. The high-frequency peaks visible in some variants result from the LogCascadeMamba frequency band resolution limits, which can be addressed with additional frequency bands at the cost of increased computational requirements. The FullRes architecture demonstrated perceptually superior quality, confirming that cognitive scaffolding enables better results than numerical optimization alone.

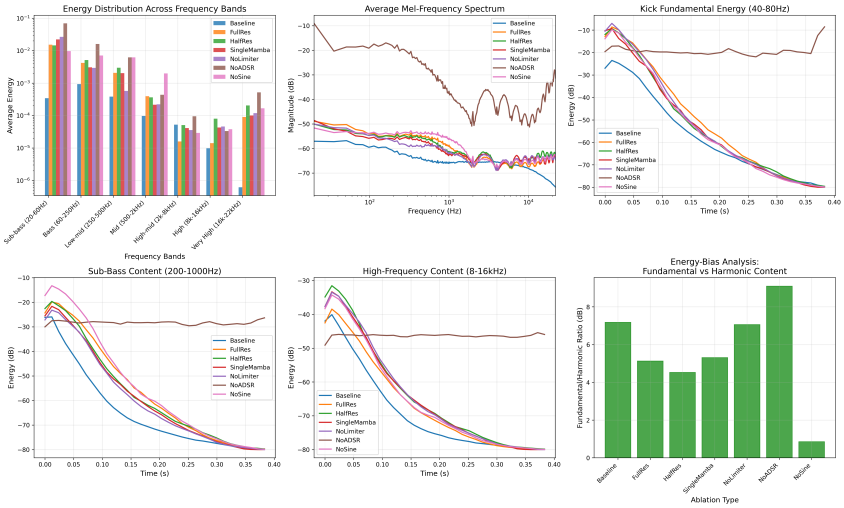

Figure 4: Comprehensive ablation analysis showing energy distribution, frequency spectrum characteristics, and fundamental/harmonic preservation across model variants.

5 Discussion

Architectural Insights. MambaKit demonstrates that domain-specific architectural innovations can be more effective than pure computational scaling for creative applications. The frequency-aware MAMBA2 processing and structured diffusion framework show how embedding musical knowledge as cognitive infrastructure enables more efficient learning than forcing models to rediscover fundamental relationships from scratch. This suggests that domain-specific priors may be essential for any task where optimization objectives misalign with human evaluation criteria.

Modular Intelligence Paradigm. MambaKit exemplifies a shift toward modular intelligence architectures. Rather than forcing neural networks to rediscover harmonic mathematics and temporal dynamics from scratch, we provide these as deterministic scaffolding. This principle is already emerging across AI domains, positional encodings in transformers provide spatial/temporal structure rather than requiring attention mechanisms to learn sequence ordering from scratch. Why wait for emergence when we can emerge it ourselves? This approach, delegating algorithmically solvable tasks to deterministic components while focusing neural resources on pattern recognition, offers a systematic approach to building more efficient and interpretable AI systems.

Complexity heuristic: This approach follows a systematic design principle: if a problem can be solved algorithmically in polynomial time, provide it as infrastructure; if it requires search, optimization, or high-dimensional pattern recognition, delegate to neural networks. Rather than forcing neural networks to rediscover harmonic mathematics and temporal dynamics from scratch, wasting billions of parameters on solved problems, we provide these as deterministic scaffolding, enabling focused learning on genuinely complex acoustic synthesis.

Human-AI Creative Partnership. The system preserves human creative agency through interpretable parameters while AI handles acoustic complexity. Musicians maintain control over musical intent (timing, envelope shapes, pitch relationships) while gaining access to sophisticated texture synthesis. This division of labor leverages human intentionality and neural pattern recognition rather than treating AI as either replacement or black-box tool.

Limitations and Future Work. The frequency decomposition approach requires careful tuning of band boundaries, and high-frequency resolution is limited by computational constraints. Complex-valued loss functions may offer improvements over current magnitude and spectral-based approaches. The cognitive scaffolding principle extends naturally to voice synthesis and melodic instruments through structured MIDI decomposition and harmonic priors, suggesting broader applications for modular intelligence architectures across creative domains.

6 Conclusion

MambaKit demonstrates that sustainable AI creativity emerges from structured human-AI partnerships rather than pure computational scaling. The system validates a fundamental architectural principle: already-defined systems do not require learning. Harmonic relationships, temporal envelopes, and frequency decompositions are mathematically solved domains that neural networks waste computational resources attempting to rediscover. By providing these as deterministic infrastructure, systems become more efficient, usable, and interpretable while enabling AI to focus on genuinely complex pattern recognition tasks.

This approach challenges the prevailing assumption that end-to-end learning is inherently superior. Instead of forcing neural networks to learn everything from scratch, we delegate algorithmically solvable problems to deterministic components while reserving neural capacity for high-dimensional synthesis tasks. This computational complexity-guided design principle aligns with recent calls for hybrid architectures that leverage structured reasoning [8]. The approach extends beyond audio: hybrid text embeddings could combine pre-defined language structures with learned representations, focusing neural resources on edge-cases rather than rediscovering fundamental linguistic relationships.

Complete experimental validation and comprehensive analysis are provided in the supplementary thesis. Rather than replacing human artistry, this work amplifies creative capability through interpretable control, suggesting new models for creative collaboration where authorship remains clearly human while technical possibilities expand dramatically. This modular intelligence paradigm offers a systematic framework for building more capable AI systems aligned with human creative intent.

References

[1] A. van den Oord, S. Dieleman, H. Zen, et al. *WaveNet: A Generative Model for Raw Audio*. 2016. arXiv: 1609.03499. URL: https://arxiv.org/abs/1609.03499.

[2] J. Engel, K. K. Agrawal, S. Chen, et al. "GANSynth: Adversarial Neural Audio Synthesis". In: *International Conference on Learning Representations (ICLR)*. 2019. arXiv: 1902.08710. URL: https://arxiv.org/abs/1902.08710.

[3] Z. Kong et al. "DiffWave: A Versatile Diffusion Model for Audio Synthesis". In: *9th International Conference on Learning Representations*. 2021. arXiv: 2009.09761. URL: https://arxiv.org/abs/2009.09761.

[4] Stability AI. *Stable Audio: Generative AI for Music and Sound*. Product announcement. 2023. URL: https://www.stableaudio.com/.

[5] Shih-Lun Wu et al. "Music ControlNet: Multiple Time-varying Controls for Music Generation". In: *arXiv preprint arXiv:2311.07069* (2023). URL: https://arxiv.org/abs/2311.07069.

[6] T. Dao and A. Gu. "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality (MAMBA-2)". In: *Proceedings of ICML 2024*. 2024. arXiv: 2405.21060. URL: https://arxiv.org/pdf/2405.21060.

[7] R. Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of IEEE/CVF CVPR 2022*. 2022, pp. 10684–10695. arXiv: 2112.10752. URL: https://arxiv.org/abs/2112.10752.

[8] Gary Marcus. "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". In: *arXiv preprint arXiv:2002.06177* (2020). URL: https://arxiv.org/abs/2002.06177.

A Technical Appendices and Supplementary Material

Complete experimental validation, mathematical derivations, and comprehensive analysis are provided in the supplementary M.Sc. thesis submitted as a separate PDF with this submission. The supplementary thesis is available at: https://drive.google.com/drive/folders/1Lo_WyHI9y5jnqDVJTqEp3U3JgtR2vVN_?usp=sharing

The thesis includes:

- Detailed architectural specifications and implementation details for the frequency-aware MAMBA2 processing system
- Comprehensive ablation studies across 6 model variants with statistical analysis
- Mathematical formulations for ADSR and pitch envelope extraction via Hilbert transform analysis
- Complete training procedures, hyperparameter specifications, and computational requirements
- Extended discussion of gradient trap phenomenon and energy-bias paradox in magnitude-based losses
- Broader implications for modular intelligence architectures and cognitive scaffolding principles
- Full experimental results, mel-spectrogram analysis, and perceptual validation studies
- Additional generated samples and audio examples demonstrating system capabilities

The supplementary material provides the complete technical foundation supporting all claims made in this paper, enabling full reproducibility of the experimental results.