Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

On Efficient Computational Methods for Transformer-Based Symbolic Music Generation

Felix Schön* and Hans Tompits**
Institute of Logic and Computation E192-03,
Technische Universität Wien,
Favoritenstraße 9-11, 1040 Vienna, Austria

Abstract

Although Transformer models have shown particular promise for symbolic music generation, their quadratic computational complexity with respect to sequence length presents significant challenges for longer musical pieces. In this paper, we describe the goals and progress of an ongoing dissertation addressing these challenges through three interconnected research directions, aiming at the development of (i) novel tokenisation strategies that significantly reduce sequence lengths while maintaining generation quality, (ii) efficient methods for incorporating arbitrary musical information into attention mechanisms through both additive and multiplicative approaches, yielding statistically significant improvements over strong baselines, and (iii) a hierarchical attention architecture that explicitly models the multi-level structure of music across beats, bars, and larger segments using specialised block-sparse attention patterns. Results achieved so far support our central hypothesis that domain-aware architectural choices, informed by music theory, can yield significant improvements over generic sequence-modelling approaches.

Keywords: Symbolic Music Generation, Transformer, Attention, Block-Sparsity.

1. Introduction

Symbolic music generation refers to the process of composing music based on a formalisable process. In this discipline, music is represented using discrete elements called *tokens* which can, e.g., mark the beginning and end of notes, their pitch, duration, or other musical attributes. As music can thus be viewed as a type of "language", it lends itself naturally to the application of large language models. Although a variety of different approaches have been applied to symbolic music generation [1–4], Transformer models [5] have proven particularly promising in this regard [6–9].

At the heart of the Transformer architecture lies the *attention* mechanism [5, 10], which computes similarity scores between pairs of token vector representations, enabling the capture of long-range dependencies crucial for modelling complex musical structures such as recurring motifs, harmonic progressions, and large-scale patterns. However, the computational complexity of attention scales quadratically with respect to sequence length, presenting significant challenges for processing longer musical pieces. This limitation has motivated research into more efficient attention mechanisms, including approximation methods [11, 12], hardware-aware implementations [13], and structured sparsity patterns [14, 15].

In this paper, we give an overview about the goals and the progress of an ongoing dissertation addressing these challenges through three interconnected research directions, aiming at the development of

(i) the impact of tokenisation strategies on model performance and efficiency;
(ii) efficient methods for incorporating arbitrary relative musical information into attention mechanisms through both additive and multiplicative approaches; and
(iii) hierarchical attention mechanisms that explicitly model the multi-level structure inherent to music.

*schoen@kr.tuwien.ac.at **tompits@kr.tuwien.ac.at

Our central hypothesis is that by deliberately designing how musical information is represented and processed at each architectural level, we can create models that are not only more computationally efficient but also better aligned with the hierarchical and relational nature of musical structure. This undertaking builds upon our previous work on a Transformer-based composition system [16], with several results already published [17–20].

2. Research Directions and Progress

2.1. Research Question RQ1: Musical Data Representation

RQ1: *How can we improve the representation of musical data for symbolic music generation?*

The representation of musical data as a sequence of discrete tokens significantly influences both the ability of a model to learn musical structure and the computational resources required [21]. Existing tokenisation strategies, such as the MIDI-like representations used in the Music Transformer [22], or REMI [23], often produce long sequences that incur high computational costs due to the quadratic complexity of self-attention. To address this, we introduced seven novel tokenisation approaches, categorised as either *MIDI-like* or *note-like* representations [17]. The note-like approaches are inspired by traditional music notation and incorporate a *running value* concept, where a specified duration applies to all subsequent notes until superseded, thereby avoiding redundant per-note duration tokens. We further proposed *large-vocabulary* representations, which encode multiple musical attributes (e.g., pitch and duration) into a single token, allowing use with unmodified model architectures. On a combination of the *piano-midi.de*,[1] *ADL Piano MIDI* [24], and *ASAP* [25] datasets, our large-vocabulary note-like tokeniser achieves an average sequence length of 190 tokens. This represents a reduction of over 50% compared to a REMI-like representation (386 tokens on average), while also reducing GPU memory requirements by over 75% (1.95 GB vs. 8.09 GB). A small-scale user study with 11 participants (272 pairwise comparisons) confirmed that these efficiency gains do not come at the cost of perceived musical quality.

2.2. Research Question RQ2: Relative Information Attention

RQ2: *How can musical information beyond token positions be efficiently integrated into the attention mechanism?*

While relative attention mechanisms [22, 26] have proven effective for music generation, they remain limited to encoding positional relationships only. In musical contexts, however, attributes like pitch intervals, rhythmic patterns, and harmonic relationships provide essential structural information that positional encodings alone cannot capture. We address this gap with two alternative approaches that pursue similar objectives via different mechanisms, outlined in what follows.

2.2.1. Additive Approach: Sparse Relative Information Injection

We introduced *sparse pre-calculated relative information injection* (SPRII) [18], an additive method that extends relative positional attention to support arbitrary integer-valued information such as pitch values, temporal positions, or intra-bar timings. SPRII precomputes a matrix of relative information values and efficiently integrates these into the attention scores using block-sparse matrix operations. To that end, we introduced `blksprs`, a Triton-based [27] PyTorch library [20] that provides comprehensive support for block-sparse matrix operations, including multiplication, softmax, gather, scatter, and transposition. In a practical evaluation for the Transformer, employing block-sparse operations reduced

[1] http://piano-midi.de

Model	16 Bars	32 Bars
Regular Transformer	2.53	>999
Music Transformer	2.46	2.77
RoFormer	2.51	2.94
SPRII (pos. + time + time bar)	2.42	2.64
RotIE (interleaved pairwise)	**2.40**	3.04
RotIE (per-head)	2.42	**2.47**

Table 1. Perplexity (lower is better) on the Lakh MIDI test set. SPRII and most RotIE variants significantly outperform baselines ($p < 0.01$).

training time by up to 35% and memory consumption by up to 45% compared to regular, sparsity-agnostic baselines.

2.2.2. **Multiplicative Approach: Rotary Informational Embeddings**

As an alternative method, we introduced *rotary informational embeddings* (RotIE) [19], a generalisation of rotary positional embeddings (RoPE) [28] to arbitrary integer-valued information. RotIE encodes information through rotations of query and key vectors such that their dot product becomes a function of the relative differences of the encoded information. Notably, this incurs no additional parameters or runtime overhead. In our work, we proposed four distinct strategies for integrating multiple information types: *per-layer* (different information per Transformer layer), *per-head* (different information per attention head), and two *pairwise* approaches (consecutive and interleaved) encoding all information within each head.

2.2.3. **Results**

Table 1 reports the perplexity results for the best SPRII and RotIE configurations on the Lakh MIDI dataset. We compare them to a regular Transformer, the Music Transformer [22], and a RoFormer [28] baseline. Both approaches yield statistically significant improvements over all baselines (with p-values less than 0.01). The two methods exhibit different strengths: while SPRII delivers consistent improvements across both sequence lengths, the per-head RotIE strategy excels on longer sequences (32 bars), achieving a perplexity of 2.47. This corresponds to an improvement of approximately 11% over the Music Transformer (2.77). The interleaved pairwise strategy achieves the best 16-bar result (2.40). A distributional analysis comparing 250 generated pieces per model to ground truth data shows that RotIE models better capture the distributions of relative pitch distances and temporal relationships, with improvements of up to 52% in Wasserstein distance over baselines. User studies (14 participants each) confirm that these improvements do not come at the cost of perceived musical quality, with SPRII and RotIE models rating competitively with ground truth on harmony and rhythm.

2.3. **Research Question RQ3: Hierarchical Attention**

RQ3: *How can attention mechanisms explicitly and efficiently model the multi-level hierarchical structure of music?*

Music is inherently hierarchical: notes form beats, beats form bars, and bars form larger phrases, yet current Transformer models typically process sequences on a flat level, disregarding these structures. While the Museformer [29] introduced summary tokens to capture information about bars, we aim to generalise and extend this concept to support a fine-grained, multi-level hierarchy.

To this end, we are developing a novel Transformer architecture that replaces standard causal attention with structure-aware hierarchical attention. Here, special *hierarchy tokens* are inserted at structural boundaries, e.g., beat-, bar-, and multi-bar block onsets, while specialised sparse attention masks govern the information flow across these levels. For each layer, two sparse attention stages are performed: First, internal hierarchy representations are built through two *hierarchy construction* approaches. In the *local hierarchy construction* step, hierarchy tokens attend to segment-local content tokens. In the *cross-level hierarchy construction* step (cascaded or propagated), hierarchy tokens of the lowest level attend to segment-local content tokens, while higher levels attend to lower-level hierarchy tokens. Second, *hierarchy-to-content integration* injects these hierarchy-encoding representations back into the sequence using structural links across configurable segment distances (e.g., 1–2 bars back). We rearrange the sequence prior to the attention computation to group hierarchy tokens, yielding denser sparse blocks and improving GPU utilisation. Furthermore, we will explore three distinct hierarchy-construction variants: The *direct* approach uses local hierarchy construction at all hierarchy levels, while with the *cascaded* approach, we perform cross-level hierarchy construction bottom-up within each layer. Finally, the *propagated* approach uses previous-layer hierarchy outputs for cross-level hierarchy construction from the second layer onward, with a configurable direct or cascaded first layer.

We are currently implementing these variants, along with the custom Triton [27] GPU kernels that compute block-sparse attention masks from a given hierarchy structure. As a next step, we plan experiments against the Music Transformer and Museformer, focusing on long-form generation (32+ bars).

3. Evaluation Methodology

Across all research directions, we employ a consistent mixed-methods evaluation approach, combining *quantitative metrics* with *qualitative assessments*. Our primary objective metric is perplexity on held-out test sets from multiple datasets, with statistical significance assessed using non-parametric hypothesis tests. For RQ2, we additionally conduct a distributional analysis comparing statistical properties of generated music (relative pitch, temporal, and harmonic distances) to ground truth sequences using Wasserstein distances. Subjective quality is assessed through small-scale user studies, where participants rate generated pieces on harmony, rhythm, variedness, authenticity, and overall quality using a 5-point Likert scale. It is important to note that these studies are primarily designed to confirm that technical improvements do not come at the cost of perceived musical quality. For RQ3, we plan to follow the same methodology, comparing the three HierarTune variants against the Music Transformer and Museformer, with particular emphasis on long-form generation (pieces of 32 bars in length or more).

4. Conclusion and Outlook

In this paper, we provided an overview about the research goals and progress of an ongoing dissertation investigating efficient computational methods for Transformer-based symbolic music generation through three interconnected research questions, RQ1, RQ2, and RQ3. Concerning RQ1, our novel tokenisation strategies achieve significant sequence length reduction without compromising quality. As for RQ2, SPRII and RotIE enable the efficient integration of arbitrary musical information into the attention mechanism, yielding statistically significant perplexity improvements over strong baselines. Finally, for RQ3, we have designed and started implementing a hierarchical attention architecture using custom Triton GPU kernels. The results obtained so far provide good promise towards our central

hypothesis that domain-aware architectural choices, informed by music theory, can yield significant improvements over generic sequence modelling approaches. For the remaining work, we will focus on the completion and experimental evaluation of our hierarchical attention architecture against the Music Transformer and Museformer.

References

[1] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription". In: *Proceedings of the 29th International Conference on Machine Learning* (*ICML 2012*). 2012.

[2] G. Hadjeres, F. Pachet, and F. Nielsen. "DeepBach: A Steerable Model for Bach Chorales Generation". In: *Proceedings of the 34th International Conference on Machine Learning* (*ICML 2017*). Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1362–1371.

[3] H. Dong, W. Hsiao, L. Yang, and Y. Yang. "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment". In: *Proceedings of the 32nd Conference on Artificial Intelligence* (*AAAI 2018*). AAAI Press, 2018, pp. 34–41.

[4] G. Mittal, J. H. Engel, C. Hawthorne, and I. Simon. "Symbolic Music Generation with Diffusion Models". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference* (*ISMIR 2021*). 2021, pp. 468–475.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems* (*NIPS 2017*). 2017.

[6] C. Payne. *MuseNet*. https://openai.com/blog/musenet. 2019.

[7] W. Hsiao, J. Liu, Y. Yeh, and Y. Yang. "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs". In: *Proceedings of the 35th Conference on Artificial Intelligence* (*AAAI 2021*). AAAI Press, 2021, pp. 178–186.

[8] P. Lu, X. Xu, C. Kang, B. Yu, C. Xing, X. Tan, and J. Bian. "MuseCoco: Generating Symbolic Music from Text". In: *CoRR* abs/2306.00110 (2023).

[9] J. Thickstun, D. L. W. Hall, C. Donahue, and P. Liang. "Anticipatory Music Transformer". In: *Transactions on Machine Learning Research* 2024 (2024).

[10] D. Bahdanau, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *Proceedings of the 3rd International Conference on Learning Representations* (*ICLR 2015*). 2015.

[11] N. Kitaev, L. Kaiser, and A. Levskaya. "Reformer: The Efficient Transformer". In: *Proceedings of the 8th International Conference on Learning Representations* (*ICLR 2020*). 2020.

[12] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. "Linformer: Self-Attention with Linear Complexity". In: *CoRR* abs/2006.04768 (2020).

[13] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness". In: *Proceedings of the 35th Annual Conference on Neural Information Processing Systems* (*NeurIPS 2022*). 2022.

[14] R. Child, S. Gray, A. Radford, and I. Sutskever. "Generating Long Sequences with Sparse Transformers". In: *CoRR* abs/1904.10509 (2019).

[15] I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The Long-Document Transformer". In: *CoRR* abs/2004.05150 (2020).

[16] F. Schön and H. Tompits. "PAUL-2: An Upgraded Transformer-Based Redesign of the Algorithmic Composer PAUL". In: *Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence* (*AIxIA 2023*). Vol. 14318. Lecture Notes in Computer Science. Springer, 2023, pp. 278–291.

[17] F. Schön and H. Tompits. "On Different Symbolic Music Representations for Algorithmic Composition Approaches Based on Neural Sequence Models". In: *Proceedings of the 23rd International Conference of the Italian Association for Artificial Intelligence* (*AIxIA 2024*). Vol. 15450. Lecture Notes in Computer Science. Springer, 2024, pp. 274–287.

[18] F. Schön and H. Tompits. "Efficient Additive Relative Information Attention for Transformer-based Symbolic Music Composition". In: *Proceedings of the 39th Canadian Conference on Artificial Intelligence* (*Canadian AI 2026*). Vol. 318. Proceedings of Machine Learning Research. 2026.

[19] F. Schön and H. Tompits. "Rotary Informational Embeddings for Symbolic Music Generation". In: *Proceedings of the 39th Canadian Conference on Artificial Intelligence* (*Canadian AI 2026*). Vol. 318. Proceedings of Machine Learning Research. 2026.

[20] F. Schön and H. Tompits. "`blksprs`: A Triton Library for Block-Sparse Matrix Operations". In: *Proceedings of the 39th Canadian Conference on Artificial Intelligence* (*Canadian AI 2026*). Vol. 318. Proceedings of Machine Learning Research. 2026.

[21] I. J. Goodfellow, Y. Bengio, and A. C. Courville. *Deep Learning*. MIT Press, 2016.

[22] C. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. "Music Transformer: Generating Music with Long-Term Structure". In: *Proceedings of the 7th International Conference on Learning Representations* (*ICLR 2019*). 2019.

[23] Y. Huang and Y. Yang. "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions". In: *Proceedings of the 28th International Conference on Multimedia* (*ACM 2020*). ACM, 2020, pp. 1180–1188.

[24] L. N. Ferreira, L. H. S. Lelis, and J. Whitehead. "Computer-Generated Music for Tabletop Role-Playing Games". In: *Proceedings of the 16th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (*AIIDE 2020*). AAAI Press, 2020, pp. 59–65.

[25] F. Foscarin, A. McLeod, P. Rigaux, F. Jacquemard, and I. Sakai. "ASAP: A Dataset of Aligned Scores and Performances for Piano Transcription". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference* (*ISMIR 2020*). 2020, pp. 534–541.

[26] P. Shaw, J. Uszkoreit, and A. Vaswani. "Self-Attention with Relative Position Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (*NAACL-HLT 2018*). 2018.

[27] P. Tillet, H. Kung, and D. D. Cox. "Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations". In: *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages* (*MAPL@PLDI 2019*). ACM, 2019.

[28] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. "RoFormer: Enhanced Transformer with Rotary Position Embedding". In: *Neurocomputing* 568 (2024).

[29] B. Yu, P. Lu, R. Wang, W. Hu, X. Tan, W. Ye, S. Zhang, T. Qin, and T. Liu. "Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation". In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems* (*NeurIPS 2022*). 2022.