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



# Rotary Informational Embeddings for 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**

In this paper, we present preliminary results on *rotary informational embeddings* (RotIE), an extension of rotary positional embeddings (RoPE) for Transformer-based symbolic music generation. With RotIE, we adapt the rotary mechanism to encode arbitrary integer-valued information such as pitch, absolute time, or intra-bar positions directly into the attention computation, allowing the model to depend on relative differences in musical attributes rather than on sequential position only. We focus on one representative per-head strategy and evaluate it on the Lakh MIDI and POP909 datasets. The presented results show improved perplexity over a regular Transformer, the Music Transformer, and a RoPE baseline, particularly on longer unseen sequences.

**Keywords:** Attention, RoPE, Symbolic Music Generation.

## 1. Introduction

Over the recent years, Transformer-based approaches have shown particular promise for symbolic music generation [1–6]. One challenge in this context is how to inject relative structure into attention. Additive relative attention has been effective for music [1, 7], and rotary positional embeddings (RoPE) [8] provide an efficient multiplicative alternative whose inner products depend on relative token positions. However, RoPE has primarily been used to encode sequential position, even though music tokens also carry attributes such as pitch, absolute time, and position within a bar.

In this paper, we discuss how this restriction can be addressed by extending RoPE to handle arbitrary integer-valued information. We refer to this extension as *rotary informational embeddings* (RotIE), which allow the encoding of multiple types of information, such as pitch values, temporal positions, or intra-bar timings, directly into the attention computation. Here, RotIE preserves the efficiency of RoPE by encoding information through rotations rather than learned embeddings. As a result, the attention mechanism becomes dependent not only on relative token positions but also on relative differences in the embedded musical attributes.

We investigated several ways of assigning such information within rotary attention. In this paper, we focus on one of these methods, viz. a *per-head strategy*, together with a compact predictive evaluation on two symbolic music datasets. The remaining strategies will be discussed in a subsequent paper.

A key advantage of RotIE over RoPE is the ability to include multiple types of information in the same attention step, enabling the model to jointly consider multiple musical dimensions. The per-head variant discussed here provides a simple illustration of the main idea in a compact setting.

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

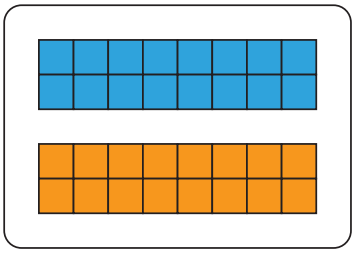

*Figure 1.* Simplified view of the representative per-head RotIE variant. Different heads receive different information sources, here illustrated by alternating colours.

## 2. Background on RoPE

With the RoPE approach [8], consecutive pairs of query and key values are rotated by angles determined by token positions. Let $v = (v_1, v_2)$ be a vector, $s$ a position to encode, and $\theta$ a pre-defined rotation parameter. Then, the rotary positional embedding is defined as follows:

$$\mathrm{RoPE}(v, s, \theta) = \begin{pmatrix} \cos(s\theta) & -\sin(s\theta) \\ \sin(s\theta) & \cos(s\theta) \end{pmatrix} v.$$

This construction yields query-key inner products that depend on relative position differences. With RotIE, we make use of the same mechanism to inject information based on integer-valued token attributes rather than positions only.

In symbolic music generation, this is of particular interest due to the fact that musical tokens carry structure beyond their sequential position. Attributes such as pitch, absolute time, and intra-bar position encode relationships that are important for rhythmic and harmonic coherence. With RotIE we extend the relative mechanism of RoPE from sequence indices to musically meaningful attributes.

Earlier additive relative-attention approaches in music [1, 7] enrich attention scores by means of learned terms derived from relative distances, whereas RoPE offers a multiplicative mechanism through rotations. With RotIE, we adapt this multiplicative mechanism to encode multiple types of musical information.

## 3. Rotary Informational Embeddings

Let an input sequence have embeddings $t_1^{emb}, \ldots, t_{l_{seq}}^{emb}$ and an associated integer information vector

$$S_{nfo} = (n_1, \ldots, n_{l_{seq}}).$$

For a consecutive query pair

$$q_{i,k} = (Q_{[i,2k-1]}, Q_{[i,2k]})$$

and key pair

$$k_{j,k} = (K_{[j,2k-1]}, K_{[j,2k]}),$$

RotIE makes use of the RoPE rotation using information values rather than token positions:

$$\mathrm{RotIE}(q_{i,k}, n_i, \theta_k) = \begin{pmatrix} \cos(n_i\theta_k) & -\sin(n_i\theta_k) \\ \sin(n_i\theta_k) & \cos(n_i\theta_k) \end{pmatrix} q_{i,k}.$$

The corresponding query-key inner product then depends on the relative difference between the two information values, $n_j - n_i$, exactly as RoPE depends on relative position. RotIE therefore generalises RoPE from positions to arbitrary discrete attributes.

To combine multiple information sources

$$S_{nfo}^1, \ldots, S_{nfo}^{n_{nfo}},$$

one can distribute them across different parts of the attention computation. In the representative variant considered here, each attention head receives one information source, cycling across the available sources when needed. In our experiments, these sources cover complementary positional and timing-related attributes.

Formally, each head is assigned one information source in a round-robin fashion and rotates all consecutive pairs in that head using the corresponding values. This produces heads specialised to different relative musical dimensions while leaving the rest of the Transformer architecture unchanged. More explicitly, let $h$ denote the head index and $n_{nfo}$ the number of available information sources. Then, the *per-head assignment* is given by

$$y = (h - 1) \bmod n_{nfo} + 1,$$

where the information sources are reused cyclically if the number of heads exceeds the number of information types. In this way, the per-head design preserves a simple correspondence between heads and musical dimensions while requiring no changes to the remaining Transformer components. Figure 1 illustrates this head-wise assignment schematically.

## 4. Experimental Setup

We conducted our experiments on both the widely used Lakh MIDI dataset [9] and the POP909 dataset [10]. MIDI files are deduplicated [11], quantised, augmented by random transposition, and segmented into 16-bar excerpts using our large-vocabulary note-like tokenisation approach [12], which is comparable to REMI [2]. Here, MIDI data is represented using a combination of *play*, *duration*, and *pause* tokens, indicating the start, duration, and temporal position of notes within a bar.

The resulting training material contains 729,062 16-bar sequences, with an average of 910 $\pm$ 520 tokens per sequence, 57 $\pm$ 36 tokens per bar, and 3.15 $\pm$ 1.56 instruments per piece.

All models are trained on 16-bar excerpts. In order to evaluate the ability of the models to generalise to longer sequences, we report the evaluation on 32-bar test sequences in the present comparison. We compare a regular Transformer [13], the Music Transformer [1], a RoPE baseline [8], and the representative per-head RotIE variant.

For all models, we use the same compact setting with model dimensionality 512, four Transformer layers, and four attention heads. For the representative per-head RotIE model, we settled on a configuration including token positions, absolute time, and intra-bar positions, assigning these information sources across the heads.

This choice is based on previous experiments with information selection and on preliminary assessments of output quality, where this configuration produced the most musically coherent results. Position retains the sequential signal used by conventional RoPE, while absolute-time and intra-bar information provide additional cues about temporal relationships in music.

Our main objective metric in the present experiment is test-set perplexity, a standard measure of next-token prediction quality on held-out data. Lower values indicate a model that makes more correct predictions in which it is more certain.

## 5. Results

Table 1 shows the results of this evaluation. Here, we focus on musical pieces of 32 bars in length, twice as long compared to the sequences seen during training. The representative per-head RotIE model outperforms all three baselines on both datasets and shows particularly strong gains over the Music Transformer and the RoPE baseline.

The contrast is particularly notable for the regular Transformer baseline, whose perplexity becomes prohibitively high on both 32-bar settings, and for the RoPE baseline, which is consistently outperformed by the representative RotIE model. These results suggest that

| Model | 32 Bars Lakh | 32 Bars POP909 |
|---|---|---|
| Regular Transformer | $> 999$ | $> 999$ |
| Music Transformer | 2.77 | 6.28 |
| RoFormer | 2.94 | 7.01 |
| RotIE (per-head) | **2.47** | **5.30** |

*Table 1.* Perplexity on 32-bar evaluation sequences, i.e., longer sequences than seen during training (lower is better).

enriching rotary attention with structured musical information can improve next-token prediction under length extrapolation.

The per-head design is particularly attractive because different heads can learn to specialise to different relative musical dimensions while the remainder of the Transformer stays unchanged. This yields a simple and interpretable representative variant.

In this paper, we restrict ourselves to one representative variant and one compact objective evaluation. Our aim is to isolate the central idea and show that it already yields promising gains in a concise setting.

## 6. Discussion

Our results suggest that RotIE is a promising approach for symbolic music generation. In particular, they indicate that extending the rotary mechanism to encode additional musical information can improve next-token prediction, especially when generalising to longer sequences than seen during training.

At the same time, the present results suggest that the selection of which information to encode is an important modelling decision. In the representative per-head setting considered here, combining token position with absolute-time and intra-bar information appears to provide a useful compromise between preserving the sequential signal of RoPE and incorporating additional temporal structure relevant for music.

Moreover, our approach is not tied specifically to music. Any structured domain in which tokens carry meaningful integer-valued attributes may benefit from the same principle, namely replacing pure sequence indices by task-relevant relative quantities inside rotary attention.

## 7. Conclusion

In this paper, we presented preliminary results on RotIE, a generalisation of rotary embeddings that incorporates integer-valued musical information directly into the attention mechanism. We considered one representative per-head strategy and evaluated it on the task of symbolic music generation. The presented results indicate improved perplexity over the considered baselines, particularly when generalising to longer sequences than seen during training. Overall, these findings suggest that RotIE is a promising and efficient way of enriching rotary attention with musically meaningful information. Furthermore, while we demonstrated its effectiveness for symbolic music generation, we conjecture that the approach could be generalised to other structured domains where tokens carry meaningful integer-valued attributes.

## References


[1] 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*). OpenReview.net, 2019.

[2] 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.

[3] J. Liu, Y. Dong, Z. Cheng, X. Zhang, X. Li, F. Yu, and M. Sun. "Symphony Generation with Permutation Invariant Language Model". In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference* (*ISMIR 2022*). 2022, pp. 551–558.

[4] 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). arXiv: 2306.00110.

[5] S. Wu and Y. Yang. "MuseMorphose: Full-Song and Fine-Grained Piano Music Style Transfer With One Transformer VAE". In: *IEEE ACM Transactions on Audio Speech and Language Processing* 31 (2023), pp. 1953–1967.

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

[7] 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*). Association for Computational Linguistics, 2018, pp. 464–468.

[8] 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), p. 127063.

[9] C. Raffel. "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching". PhD thesis. Columbia University, USA, 2016.

[10] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia. "POP909: A Pop-song Dataset for Music Arrangement Generation". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference* (*ISMIR 2020*). 2020.

[11] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T. Liu. "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (*ACL 2021*). Association for Computational Linguistics, 2021, pp. 791–800.

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

[13] 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.