Relative Positional Embeddings for Track Level Representations in Masked Contrastive Learning

Jordan Phillips JPHILL39@GMU.EDU

Junyoung Koh SOLBON1212@YONSEI.AC.KR

Abstract

We introduce Myna-RPE, an extension of the Myna framework that directly produces track-level music embeddings from full-length mel-spectrograms in an end-to-end fashion. Existing methods typically embed fixed-length chunks and aggregate to get track-level features. This approach often fails to capture long-range dependencies and may suppress salient information present in standout segments across the full track. Our method adapts relative positional encodings into the Myna framework to accept much larger inputs, extending to entire tracks. We train the model using a contrastive learning objective with patch-out augmentation on MTG-Jamendo, and evaluate it via linear probes on several MIR benchmarks (GTZAN, GiantSteps, EmoMUSIC).

Keywords: music information retrieval, transformer autoencoder, relative positional embeddings, patchout, contrastive learning, track-level embeddings

1. Introduction

The field of Music Information Retrieval has traditionally been dominated by supervised learning (Pons and Serra, 2019; Koutini et al., 2022; Koh et al., 2025; Baumann, 2021). These frameworks use supervised tasks such as key and genre classification, and music auto-tagging on labeled datasets in order to learn representations about the underlying audio (Gong et al., 2021; Koutini et al., 2022; Kim et al., 2019; Pons and Serra, 2019). However, interest has recently been shifting towards self-supervised learning because of the abundance of raw unlabeled data to learn from. Contrastive frameworks such as SimCLR, COLA, CLMR, and MULE use a contrastive loss between positive and negative pairs to learn musical representations from audio (Chen et al., 2020; Saeed et al., 2020; Spijkervet and Burgoyne, 2021; Kim et al., 2019). Recently, Myna combined the PaSST patchout approach (Koutini et al., 2022) with this contrastive paradigm and achieved state-of-the-art performance among models trained in public datasets in downstream tasks, specifically mood detection, key detection, and genre classification (Yonay et al., 2025).

Musical recordings inherently vary in length, which poses a challenge for models that assume fixed-size inputs. Most current Transformer-based architectures, including ViT-derived models, are constrained to process inputs up to a predefined maximum length due to their positional embedding structure. Specifically, they use absolute positional embeddings like sinusoidal (Yonay et al., 2025) or learned (Koutini et al., 2022) positional embeddings to inject positional information into the patches of the underlying ViT networks. To handle longer recordings, prior methods typically segment each track into fixed-length chunks, compute embeddings for each chunk, and then aggregate the results at inference time. (Yonay

et al., 2025; Kim et al., 2019) This chunk-based strategy enables evaluation on variable-length data but introduces a fundamental mismatch: the model is trained and evaluated on short excerpts, yet the provided labels correspond to the entire track. Consequently, aggregation at the chunk level overlooks long-range dependencies and global structure across the full track, degrading the quality of track-level representations. Moreover, this aggregation approach does not translate naturally to unsupervised settings such as clustering or music recommendation, where the objective is to obtain a single latent representation per track.

We adapt the popular Relative Positional Embedding algorithm, ALiBi, to the 2D audio spectrogram space, enabling the ViT encoder to process variable-length sequences up to the full track length without segmentation or aggregation, enabling the ViT encoder to process variable-length spectrogram sequences. Our approach can inherently extrapolate beyond typical training window lengths up to the full length of a track without segmentation or aggregation. We evaluate these positional encoding techniques on downstream MIR tasks on the segment and track level, and explore it's impact on downstream MIR tasks.

Our primary contributions are as follows

1. A practical adaptation of the Myna framework with 2D and 1D ALIBI implementations
2. An empirical comparison between the Myna-RPE approach and previous works across downstream MIR tasks

2. Related Work

2.1. Contrastive Learning Frameworks

Contrastive learning has shown itself to be a strong self-supervised learning paradigm for the music domain, enabling models to learn meaningful structure from unlabeled data. In vision and audio, early frameworks such as SimCLR (Chen et al., 2020) established the principle of learning by distinguishing between positive and negative pairs, driving large gains in generalization. Building on this foundation, CLMR (Spijkervet and Burgoyne, 2021) introduced contrastive learning to music by adapting SimCLR's augmentations, such as temporal cropping, pitch shifting, and time masking, to audio waveforms and spectrograms. While effective, these early frameworks relied heavily on augmentations designed for images rather than music, limiting their ability to capture long-range harmonic and rhythmic dependencies.

Later approaches sought to inject stronger inductive biases about musical structure. MERT (Li et al., 2024) adopted a dual teacher–student architecture where an acoustic teacher and a Constant-Q Transform (CQT) teacher guide a residual vector quantization VAE (RVQ-VAE) (Défossez et al., 2022; Zeghidour et al., 2021). This design improved robustness to timbral variation and enabled stronger downstream performance on MIR tasks. Such multi-teacher systems are computationally intensive and depend on pre-trained models that restrict adaptability to new domains.

More recently, Myna (Yonay et al., 2025) demonstrated that competitive representations can be learned using a simpler, fully contrastive objective. Myna extends the SimCLR formulation by introducing aggressive token patchout inspired by PaSST (Koutini et al., 2022). This serves two purposes. Firstly, it reduces the computational burden of the

Transformer by shortening the sequence length and secondly it acts as a regularizer that prevents the model from overfitting to specific temporal or frequency positions (Yonay et al., 2025). Importantly, Patchout preserves the overall structure of the spectrogram, allowing the model to learn robust global features.

Unfortunately, all of these approaches have the same key limitation: they operate on fixed-length segments, discarding long range temporal dependencies and structural cues that span an entire track and are unable to inherently process long inputs. Aggregating chunk-level embeddings (e.g., via averaging) can mitigate this to some extent but may destructively interfere with important information and fail to produce a single holistic representation suitable for clustering or recommendation.

In contrast to these works, Myna-RPE aims to *remove* the need to evaluate on fixed length segments, eliminating any possible destructive interference between chunk latent space aggregation while simultaneously allowing the model to learn long form structural patterns

2.2. Relative Positional Embeddings

The underlying attention operation transformers compute is position-invariant (Vaswani et al., 2017), and so frequently a positional embedding strategy is used to inject positional information into the input. The simplest type of positional embeddings are absolute embeddings, which add some distinct vector to each input token. Absolute Positional Embeddings (APE) perform well for a multitude of tasks, but have very limited extrapolation capabilities to contexts larger than those seen in training (Press et al., 2021). A second and more complex type of positional embeddings are Relative Positional Embeddings (RPE). These embeddings seek to inject the distances of token pairs into the attention framework. ALIBI was the first and most prominent RPE. It operates by adding a negative bias to the attention matrix which scales linearly with token pair distance. ViT networks traditionally use APEs (Wu et al., 2020) however recent work has begun to explore the adaption of RPEs to ViT architectures. (Veisi et al., 2025) adapted the ALIBI algorithm from 1D to 2D to solve grid based visual reasoning problems. Another RPE called RoPE has seen wide adoption in LLM frameworks (Su et al., 2023). It operates by rotating query and key combinations along the complex plane according to their position in the input sequence. Similarly to ALIBI, RoPE has recently been extrapolated from 1D to 2D by (Heo et al., 2024) in order to work within ViT based frameworks. RoPE does not generalize as well as additive bias based methods (Press et al., 2021) however it's generalization capabilities outperform absolute positional embeddings on long context lengths.

3. Method

3.1. Preliminaries

Our work builds on the Myna framework (Yonay et al., 2025), which itself is an extension of the CLMR contrastive learning framework (Spijkervet and Burgoyne, 2021). In CLMR and Myna, the goal is to learn meaningful and discriminative representations of audio segments without requiring explicit labels. To do this, our pre-training dataset of audio recordings is first transformed into mel-spectrograms. For each batch, positive pairs are

formed by randomly sampling two fixed-length chunks from the same parent spectrograms, while negative pairs are constructed from chunks belonging to different tracks. The input is patchified, and the set of patches is masked with patchout and fed into the encoder, which transforms the patches into a latent representation. The encoder is trained to maximize the similarity between representations of positive pairs and minimize similarity between negative pairs using the InfoNCE loss (Rusak et al., 2025). This encourages the model to learn features that are invariant to small transformations and time shifts within a track while remaining discriminative across different tracks.

3.2. Model Architecture

We use a modified version of the Vision Transformer (ViT) (Wu et al., 2020), Myna, with a prepended CLS token instead of global average pooling. We train this encoder using InfoNCE loss (Rusak et al., 2025). For patching we break inputs into 16x16 non-overlapping patches similar to (Dosovitskiy et al., 2021). Our key modification is to replace absolute positional encodings with relative positional encodings. Specifically, we use 2D ALiBi (Veisi et al., 2025) and a 1D-ALiBi over time with learned embeddings for the frequency axis.

3.3. Positional Embeddings

Sinusoidal and learned positional embeddings do not extrapolate well to longer context lengths than seen in training. In order to circumvent this we adapt the 2D ALiBi variant introduced in (Veisi et al., 2025). In its original 1D form, ALiBi defines an additive positional bias as:

$$A^n_{i,j} = \frac{q^n_i \cdot k^n_j}{\sqrt{d}} + \mathbf{B}_{P_{i,j}}, \qquad \mathbf{B}_{P_{i,j}} = r \cdot |i - j|, \tag{1}$$

where $\mathbf{P}_{i,j}$ represents the relative positional offset between tokens i and j, and r is a predefined slope that penalizes tokens based on their distance (Press et al., 2021).

This can naturally be extended by extrapolating the distance formula to 2 dimensions. Hence, the 2D-RPE bias is computed as:

$$\mathbf{B}_{P_{i,j}} = d\big((x_i, y_i), (x_j, y_j)\big) \tag{2}$$

where $d\big((x_i, y_i), (x_j, y_j)\big)$ represents the 2D Manhattan distance between coordinates (x_i, y_i) and (x_j, y_j). (Veisi et al., 2025) use a right and left slope in order to add information about the raster order, as that is important for generation. However, as our task is not generative in nature, we will only be using this simplified version with no right and left slopes. Similarly we can define a time-only ALIBI bias as:

$$\mathbf{B}_{P_{i,j}} = |x_i - x_j| \tag{3}$$

This mirrors the original 1D ALiBi bias with the distance changed from the raster order to the distance along the x-axis. However, this bias encodes no vertical information across patches. Because of this, we include a set of d learned position embeddings across the y-axis, where d is the number of vertical patches in a spectrogram, so that each patch retains both vertical (frequency) and horizontal (time) information relative to other patches.

This formulation enables the integration of RPE schemes into our 2D patch-based network. Because we are mainly concerned about the relative position along the temporal axis, we test both RPE achemes, ALIBI and RoPE, along both temporal and frequency axis, and then just along the temporal axis with learned positional embeddings for the frqeuency axis.

3.4. Training Objective

The loss algorithm InfoNCE, as given in (Rusak et al., 2025) is defined for a positive pair (i, j) drawn from the joint distribution $p_{\text{pos}}(x_i, x_j)$ and a set of negatives (k, l) drawn from the product of marginals $p_{\text{neg}}(x_k)p_{\text{neg}}(x_l)$, the InfoNCE objective encourages high similarity for positive pairs and low similarity for negatives.

Let $s_\theta(x_a, x_b)$ denote the cosine similarity function and $\tau > 0$ a temperature parameter. Then, for a positive pair (i, j), the loss contribution is:

$$\mathcal{L}^+_{i,j} = -\log \frac{\exp\big(s_\theta(x_i, x_j)/\tau\big)}{\sum\limits_{(k,l)\in\mathcal{P}_{i,j}} \exp\big(s_\theta(x_k, x_l)/\tau\big)} \tag{4}$$

where $\mathcal{P}_{i,j}$ denotes the set containing one positive pair (i, j) and all corresponding negative pairs.

Overall, the full InfoNCE loss over all positive pairs in a batch is:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \mathcal{L}^+_{i,j}. \tag{5}$$

3.5. Implementation Details

In practice, applying any relative bias scheme with patchout presents a technical challenge. The primary challenge arises from the removal of tokens from the sequence. Traditional methods fail in this context as they assume a contiguous sequence of tokens in order to calculate relative distances. In any implementation of a patchout aware relative bias we must store the original coordinates of tokens, incurring a minor additional memory overhead. This coordinate memoization method is used to generate 2D-ALiBi biases via manhattan distance Equation (2) from a masked input, and Temporal ALiBi biases via distance in the time axis Equation (3). We use the AdamW optimizer with a learning rate of 2e-4 and weight decay of 1e-4, batch size of 1024, and train for 512 epochs.

3.6. Pre-Training Dataset

We pre-train our models on the top-50-tags subset of the MTG-Jamendo dataset (Bogdanov et al., 2019), containing 55k track recordings each of variable length. In previous works AudioSet is frequently used (Yonay et al., 2025) because of it's size (Gemmeke et al., 2017). However, this dataset is frequently in flux, and so for reproducibility we chose MTG-Jamendo. Additionally, many clips in AudioSet have overlapping sounds, background noise, unrelated voices, etc, which may hamper the quality of our training.

4. Experiments

4.1. Downstream Datasets

First we evaluate our approaches on standard MIR tasks. We use the GTZAN (Tzanetakis and Cook, 2002) dataset for genre classification, the EmoMUSIC (Soleymani et al., 2013) dataset for emotion detection, and the GiantSteps (Knees et al., 2015) dataset for key detection. We use the same evaluation procedure used in (Castellon et al., 2021) for a fair evaluation, training shallow MLP and Linear models on latent representations with a grid search over hyperparameters.

4.2. Evaluation Metrics

We evaluate our embeddings using both supervised and unsupervised metrics. For supervised evaluation, we train lightweight linear classifiers on top of frozen contrastive embeddings models to perform standard MIR tasks such as genre classification, and key or mood prediction. Following common practice, we report Accuracy, F1-score, and Mean Average Precision (mAP) on downstream datasets. These metrics reflect the ability of the learned representation to linearly separate semantic information relevant to musical perception. To get a final aggregated score, we take the average score for each task. Tasks with multiple metrics are averaged across them to get the score for that task.

4.3. Results

Model	**Size**	**GTZAN**	**GiantSteps**	**EmoMusic**	
		Acc	**Acc**	**A**	**V**
Myna Hybrid	22M	77.9	68.0	70.8	55.2
MusiCNN	32M	75.5	53.6	69.7	50.2
MERT-330M	330M	79.3	65.6	74.7	61.2
Jukebox	5B	79.7	66.7	72.1	61.7
1D ALiBi + Frequency Embeddings	22M	74.87	82.57	59.37	43.45
2D ALiBi	22M	78.39	76.5	68.06	44.05

Table 1: Linear probe performance on downstream MIR tasks. Metrics include GTZAN (Accuracy), GiantSteps (Accuracy), and EmoMusic (Arousal, Valence). The final column shows the average across all metrics. Our evaluation process is identical to (Yonay et al., 2025) and (Castellon et al., 2021)

5. Discussion

Firstly, our baseline model with sinusoidal positional embeddings outperforms Myna on the GTZAN dataset but underperforms on EmoMusic. We believe this to be a direct result of the input size. Our model operates on larger segments of mel-spectrograms and therefore

has the opportunity to focus more on long term relations within the spectrogram, which leads to better performance on genre related tasks, but worse performance on harmonics related tasks.

Secondly, our 2D ALiBi model outperforms the 1D ALiBi model across all metrics except key detection. The extrapolated dimension is time and not frequency, and so the ALiBi representations are very meaningful across time. The results show there is also a storng benefit from having them across the frequency axis as well, although having learned frequency embeddings increased the performence on key detection significantly.

6. Future Work

6.1. RoPE and Other RPEs

Our proposed method shows how RPE's can improve performance on downstream tasks in this modality. We anticipate that other additive RPEs could have similar performance. Similarly, we believe that Rotary Positional Embeddings are another RPE worth investigating, especially as the cyclical nature of rotary embeddings naturally compliment the cyclical nature of music.

6.2. Training Length

In this work, we test the effect of ALIBI on the Myna framework. Because ALIBI allows inference on any length of input, we believe that training on a wide variety of segment lengths is a promising method to further reinforce extrapolation on long contexts. Future work should explore the effect of different distributions of input lengths and training on longer contexts.

7. Conclusion

In this work we introduced Myna-RPE, an end to end approach for track-level music representation that uses relative positional biases and patch-out contrastive training to produce embeddings that generalize to large inputs. Our approach shows that relative positional embeddings can be extrapolated to model 2D audio representations, allowing models generalize beyond their normal capabilities. This finding highlights a promising direction for future research.

References

Stefan A Baumann. Deeper convolutional neural networks and broad augmentation policies improve performance in musical key estimation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pages 42–49. ISMIR, November 2021. doi: 10.5281/zenodo.5624477. URL https://doi.org/10.5281/zenodo.5624477.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. URL http://hdl.handle.net/10230/42015.

Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval, 2021. URL https://arxiv.org/abs/2107.05677.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL https://arxiv.org/abs/2210.13438.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. URL https://arxiv.org/abs/2104.01778.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer, 2024. URL https://arxiv.org/abs/2403.13298.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Mule: Multimodal universal language embedding, 2019. URL https://arxiv.org/abs/1909.03493.

Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *International Society for Music Information Retrieval Conference*, 2015. URL https://api.semanticscholar.org/CorpusID:15836728.

Junyoung Koh, Soo Yong Kim, Gyu Hyeong Choi, and Yongwon Choi. Aiba: Attention-based instrument band alignment for text-to-audio diffusion, 2025. URL https://arxiv.org/abs/2509.20891.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2753–2757. ISCA, 2022. doi: 10.21437/Interspeech.2022-227. URL https://doi.org/10.21437/Interspeech.2022-227.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg,

Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training, 2024. URL https://arxiv.org/abs/2306.00107.

Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging, 2019. URL https://arxiv.org/abs/1909.06654.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2021.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice, 2025. URL https://arxiv.org/abs/2407.00143.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations, 2020. URL https://arxiv.org/abs/2010.10915.

Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, page 1–6, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323963. doi: 10.1145/2506364.2506365. URL https://doi.org/10.1145/2506364.2506365.

Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations, 2021. URL https://arxiv.org/abs/2103.09410.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ali Veisi, Hamidreza Amirzadeh, and Amir Mansourian. Context-aware biases for length extrapolation. *arXiv preprint arXiv:2503.08067*, 2025. URL https://arxiv.org/abs/2503.08067.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Ori Yonay, Tracy Hammond, and Tianbao Yang. Myna: Masking-based contrastive learning of musical representations, 2025. URL https://arxiv.org/abs/2502.12511.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec, 2021. URL https://arxiv.org/abs/2107.03312.