Journal of Machine Learning Research 23 (2025) 1-5 Submitted 1/21; Revised 5/25; Published 9/25

SCORE-SET: A dataset of GuitarPro files for Music Phrase Generation and Sequence Learning

Vishakh Begari DJENTLEVIBE32168@GMAIL.COM
Independent Researcher
Gauting, 82131, Germany

Editor: Adam Smasher

Abstract

A curated dataset of Guitar Pro tablature files (.gp5 format), tailored for tasks involving guitar music generation, sequence modeling, and performance-aware learning is provided. The dataset is derived from MIDI notes in Hawthorne et al. (2019) and Kong et al. (2022) which have been adapted into rhythm guitar tracks. These tracks are further processed to include a variety of expression settings typical of guitar performance, such as bends, slides, vibrato, and palm muting, to better reflect the nuances of real-world guitar playing. Dataset available at Begari (2025).

Keywords: Dataset, Guitar, Tablature, Transformers, Sequence Learning

1 Introduction

Advancements in machine learning have led to significant progress in the field of automatic music generation, particularly with symbolic representations such as MIDI. While datasets like Hawthorne et al. (2019) Gemmeke et al. (2017) Thickstun et al. (2017) Bertin-Mahieux et al. (2011) Peracha (2022) Bradshaw and Colton (2025) Kong et al. (2022) have enabled research in mostly piano music generation, there remains a lack of large-scale, high-quality resources tailored specifically to the guitar—a highly expressive and technically diverse instrument.

Guitar music presents unique challenges for modeling due to its polyphonic nature, alternate tunings, and rich expressive techniques (e.g., bends, slides, palm muting). Existing symbolic music datasets often lack this level of nuance, limiting the development of models capable of learning and generating realistic guitar performance.

To address this gap, curated dataset of Guitar Pro tablature files (.gp5 format) is provided designed for guitar music generation, sequence modeling, and performance-aware learning. The dataset is derived from the MIDI information found in Hawthorne et al. (2019) and Kong et al. (2022), with melodies adapted innto rhythm guitar tracks and enriched with expressive elements common in guitar playing.

2 SCORE-SET Dataset

MIDI notes provide information about both pitch and timing, specifying when a note is played, its duration, and its musical pitch. In the context of guitar tablature, the pitch is mapped to an open string and fret position, while the duration is quantized to musical

beats. The guitar used is a 6-string instrument tuned to FCG#D#A#D#. Both single notes and chords are automatically encoded along with their corresponding beat durations.

To begin with, an overview of articulations to be used in the dataset is provided. These are deemed essential for capturing the expressive nuances of guitar performance.

2.1 Accents

Accentuation in playing refer to emphasising specific notes or rhythms to create dynamics and expression in music.

2.1.1 Palm mute

Palm mute - A technique of lightly resting the edge of palm on the strings near bridge while plucking or strumming.

2.1.2 Bends

Pushing or pulling a string sideways across the fretboard, raising its pitch.

2.1.3 Tremolo bar

A tremolo bar or whammy bar is a device attached to the bridge of a guitar that allows bending the pitch of notes by pushing or pulling on the bar.

2.1.4 Slide

Smoothly moving up or down the fretboard without lifting off on the string, creating a seamless transition between notes.

2.1.5 Dead note

(muted note or ghost note) Muting the string to produce a percussive sound.

2.1.6 Hammer on / pull off

Allows playing 2 notes in succession without picking the second note. Pressing down onto a higher fret to play a note without picking it, producing smooth legato transition between notes. Pulling off a higher fretted note while the lower fret note is still pressed.

2.1.7 Vibrato

Repeated bending and releasing of a note to create subtle pitch variation.

2.1.8 Harmonic - Natural

Producing bell-like, chime sound by lightly touching a string at specific points along the fretboard.

3 Statistics

Acknowledgments and Disclosure of Funding

All acknowledgements go at the end of the paper before appendices and references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found on the JMLR website.

Appendix A.

In this appendix we prove the following theorem from Section 6.2:
Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \;\Rightarrow\; v = w = 0$ in a given dataset $\mathcal{D}$). Let N_{v0}, N_{w0} be the number of data points for which $v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then*

$$N_{v0} \;>\; N_{w0} \;\;\Rightarrow\;\; I_{uv} \;\leq\; I_{uw}$$

with equality only if u is identically 0. ■

Appendix B.

Proof. We use the notation:

$$P_v(i) \;=\; \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \;\equiv\; P_v(0) \;=\; 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of v taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by H. We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0$....

Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.

References

Vishakh Begari. Score-set. `https://github.com/DjentleViBe/SCORE-SET`, 2025. Accessed: 2025-07-11.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In Anssi Klapuri and Colby Leider, editors, *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, pages 591–596. University of Miami, 2011. URL `http://ismir2011.ismir.net/papers/OS6-1.pdf`.

Louis Bradshaw and Simon Colton. Aria-midi: A dataset of piano midi files for symbolic music modeling, 2025. URL `https://arxiv.org/abs/2504.15071`.

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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1lYRjC9F7`.

Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. Giantmidi-piano: A large-scale midi dataset for classical piano music, 2022. URL https://arxiv.org/abs/2010.07061.

Omar Peracha. Js fake chorales: a synthetic dataset of polyphonic music with human annotation, 2022. URL https://arxiv.org/abs/2107.10388.

John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch, 2017. URL https://arxiv.org/abs/1611.09827.