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

Vishakh Begari   VIBEDJENTLE@GMAIL.COM

Independent Researcher
Bavaria, Germany
ORCID: 0009-0008-9399-8581

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, Transformer, 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 standard E–B–G–D–A–E. 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 and their tablature is provided. These are deemed essential for capturing the expressive nuances of guitar performance.

2.1 Expressions

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

A bend involves pushing or pulling or pushing a string laterally across the fretboard to raise its pitch.
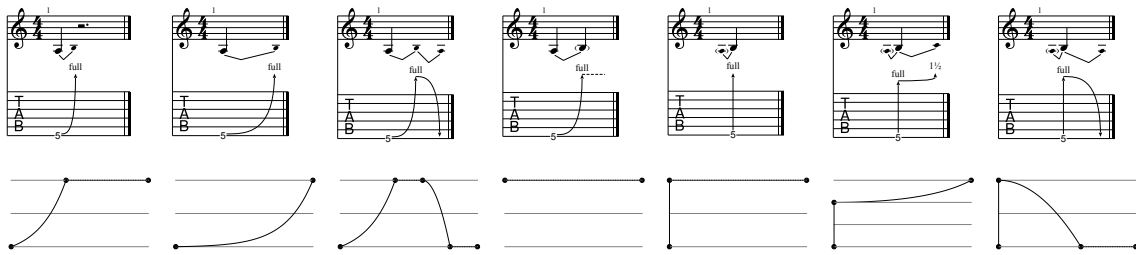

Figure 1: Bend types and their pitch variations. Horizontal line represents a semitone.

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.
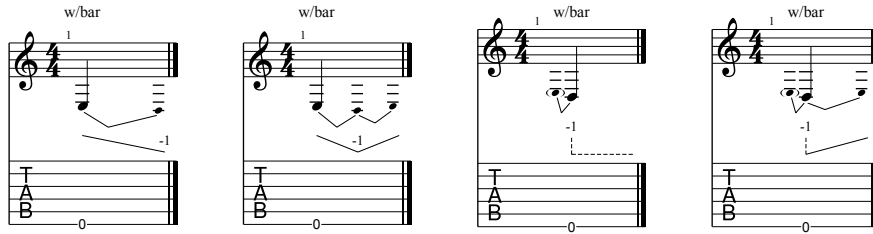

Figure 2: Tremlo types and their pitch variations.

2.1.4 SLIDE

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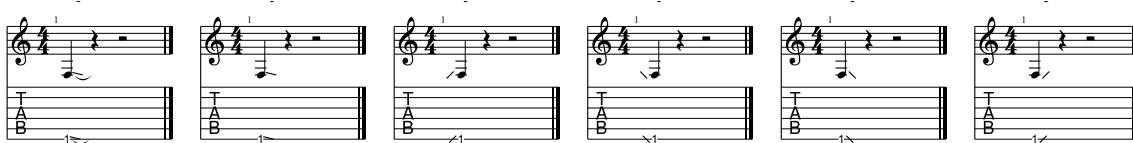

Figure 3: Slide types

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

Involves bending and releasing the string rapidly to oscillate the pitch slightly above and below the target note.

2.1.8 HARMONIC - NATURAL

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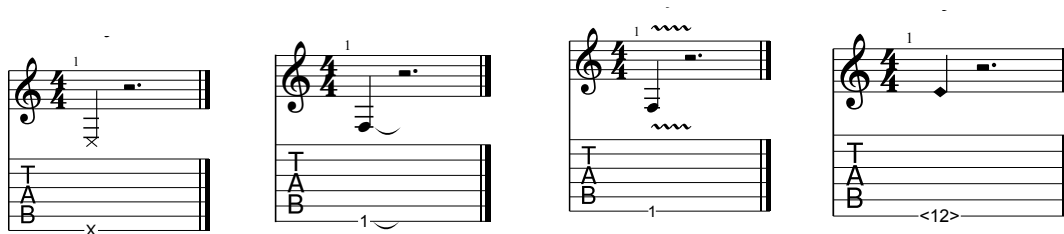

Figure 4: From left to right - Dead note, Hammer On, Vibrato, Natural Harmonic

2.2 Quantization

MIDI formats represent time signatures using ticks, requiring quantization techniques to accurately derive rhythmic patterns. For simplicity, the minimum note duration was limited to a sixteenth note. Off-beat notes were quantized to the nearest beat within the bar, and

all notes were aligned to a 4/4 time signature. In cases where notes extended beyond the standard measure length, the measure was dynamically adjusted to accommodate them.

2.3 String and fret

Since the original datasets consist of piano-based transcriptions, mapping them to guitar fret positions poses a challenge. For chord voicings, the root note was first fixed, and subsequent notes were assigned to consecutive strings. If multiple notes fell on the same string, their pitch was adjusted—by adding or subtracting octaves—until a playable fret position was found on different strings.

3 Statistics

For this study, a dataset of guitar tablatures was manually created through transcription of publicly available audio recordings of modern metal songs. The transcriptions were created by ear and focused specifically on expressive guitar techniques. Existing symbolic music datasets often lack detailed guitar-specific expressions such as palm muting, tremolo picking, or slides—elements that are crucial to the stylistic integrity of modern metal music. Since these articulations are rarely annotated in standard MIDI or tablature datasets, manual transcription was necessary to ensure expressive accuracy and genre fidelity. The resulting dataset offers a more nuanced representation of guitar performance suited for expression-aware generative modeling. The expression statistics were calculated over a total of 198656 notes, with approximately 25.39 % of them featuring accents.
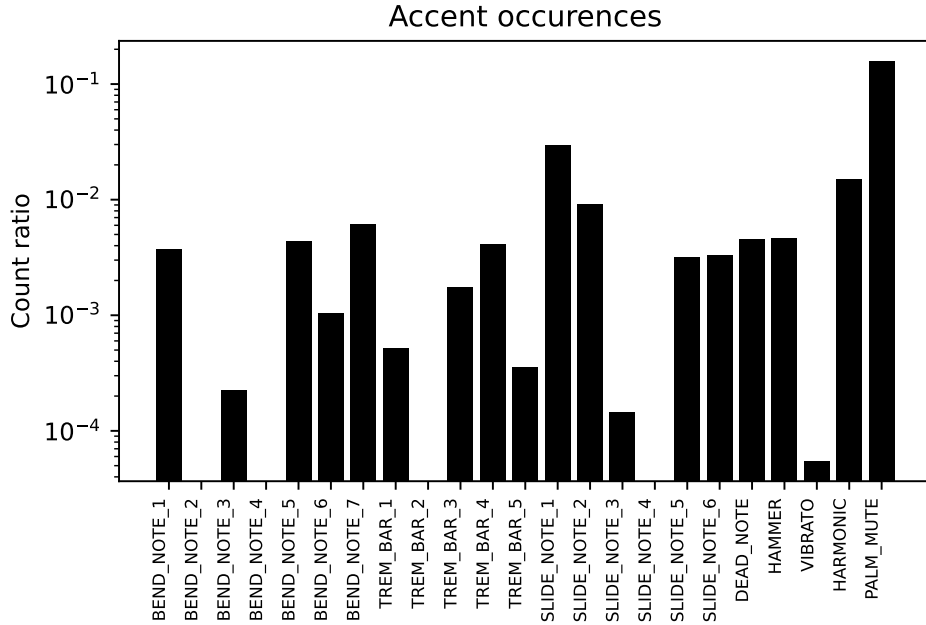

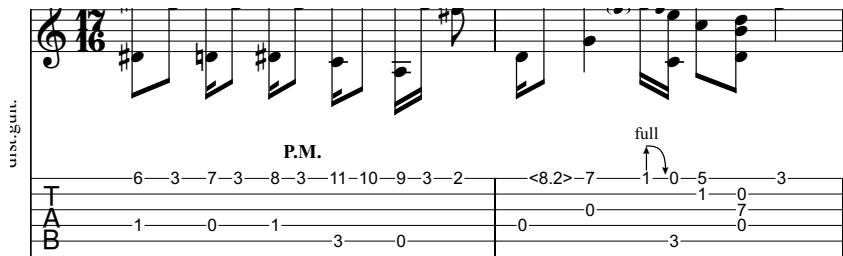
Figure 5: Percentages of techniques.

Figure 6: Tablature showing palm mute, bend, harmonic and custom measure.

Accents were added to the original dataset at random, following the expression ratios derived from Figure 5. Although these percentages were based on individual musical notes, tremolo is an exception, as it is applied to an entire beat, which may contain multiple notes.

After expression augmentation, the sequences were exported to .gp5 files using the Py-GuitarPro Abakumov (2014) library, preserving all note-level attributes and guitar-specific articulations. This allowed for compatibility with standard tablature software and ensured that the data could be both inspected and rendered in a musically meaningful format. The dataset is made available at Begari (2025).

Acknowledgments and Disclosure of Funding

References

Sviatoslav Abakumov. Pyguitarpro. https://pyguitarpro.readthedocs.io/en/stable/, 2014. Revision c88d0350.

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`.