

# MIDI Transformer Tokenization

Jonas Veit, NAMEN EINFÜGEN

January 8, 2025

## Current workflow

1. Specify the MIDI data
2. Select a suitable Tokenizer. Right now we use REMI
3. train the tokenizer with BPE
4. save the tokenizer and vocabulary as json
5. generate training, validation and test sets via `generate_train_val_test_sets.py`  
we already augment data within this process if we want to.
6. Our GoePT expects integers as tokens and `np.memmap` expects binary files. We use `np.to_file` in order to create the right binaries

## Different tokenizers

- For simple melodies: REMI or MelodyTokenizer.
- For more complex, polyphonic data: Consider PolyphonyTokenizer or NoteTokenizer.
- For structured MIDI data: Use Structured or TSD.
- For music generation: REMI or MIDI-Like might work well.

## Things to keep in mind

- We need to specify the sequence length
- We need to specify the embedding dimension
- We need to specify the *vocab<sub>size</sub>*