

---

# Muse2Bach

李維釗 r12942089

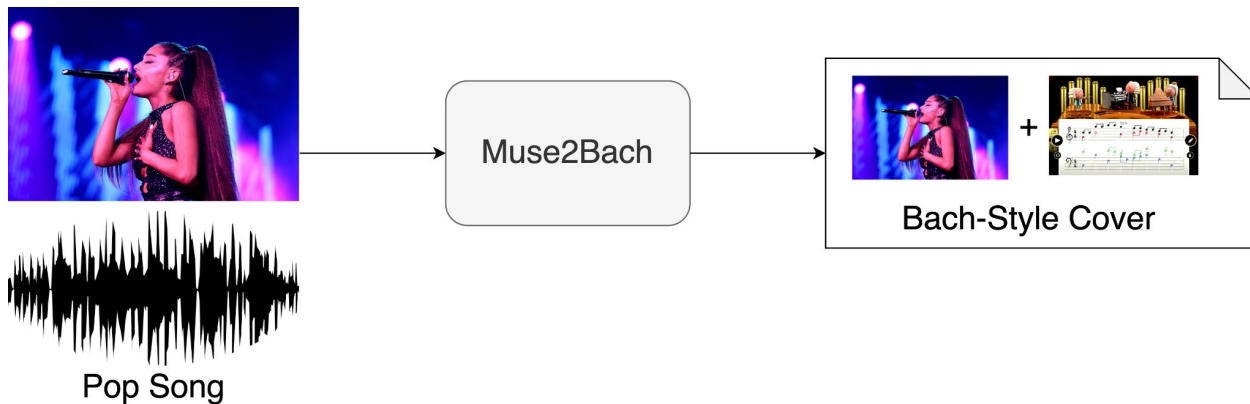
艾芯 r12942156

譚至斌 d12942015

---

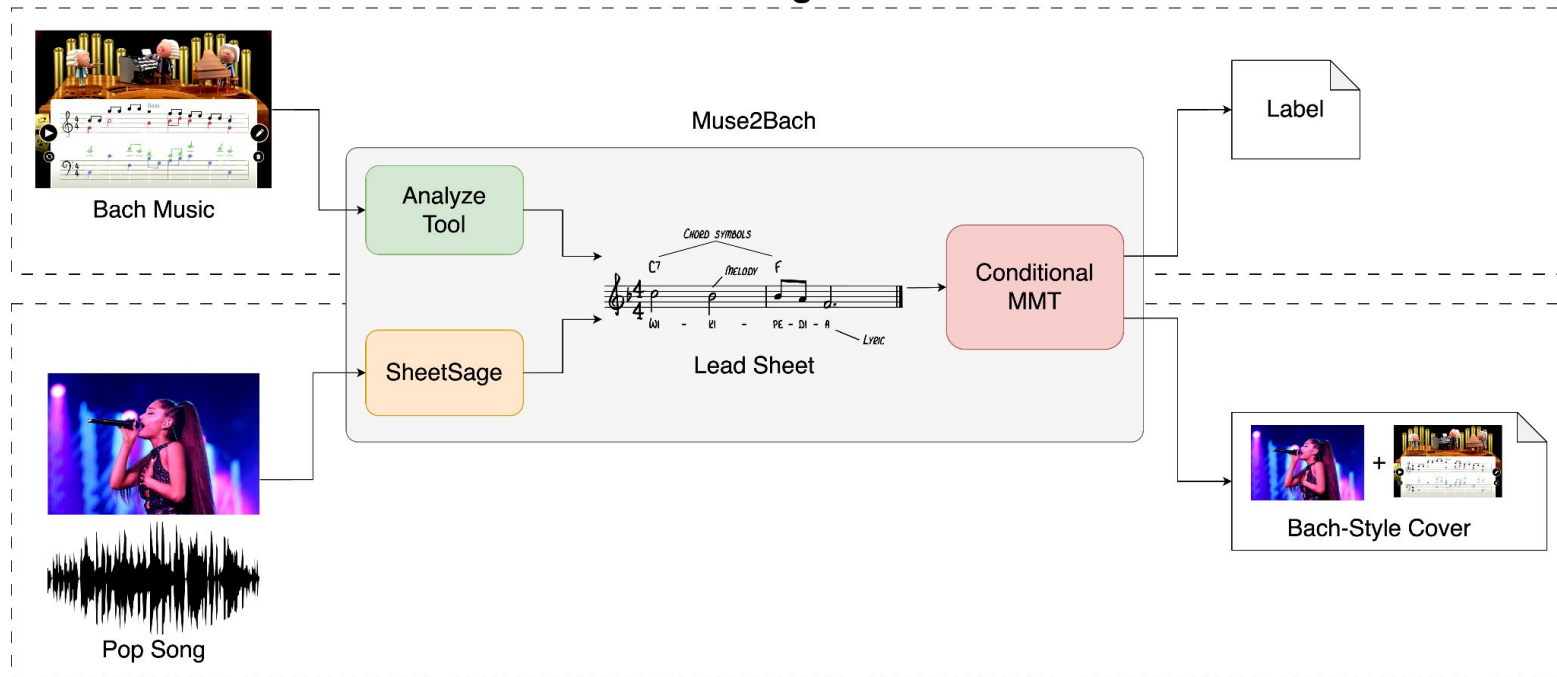
# Muse2Bach

- Objective
  - Convert pop song to Bach-style cover ([example](#))



# Mathodology

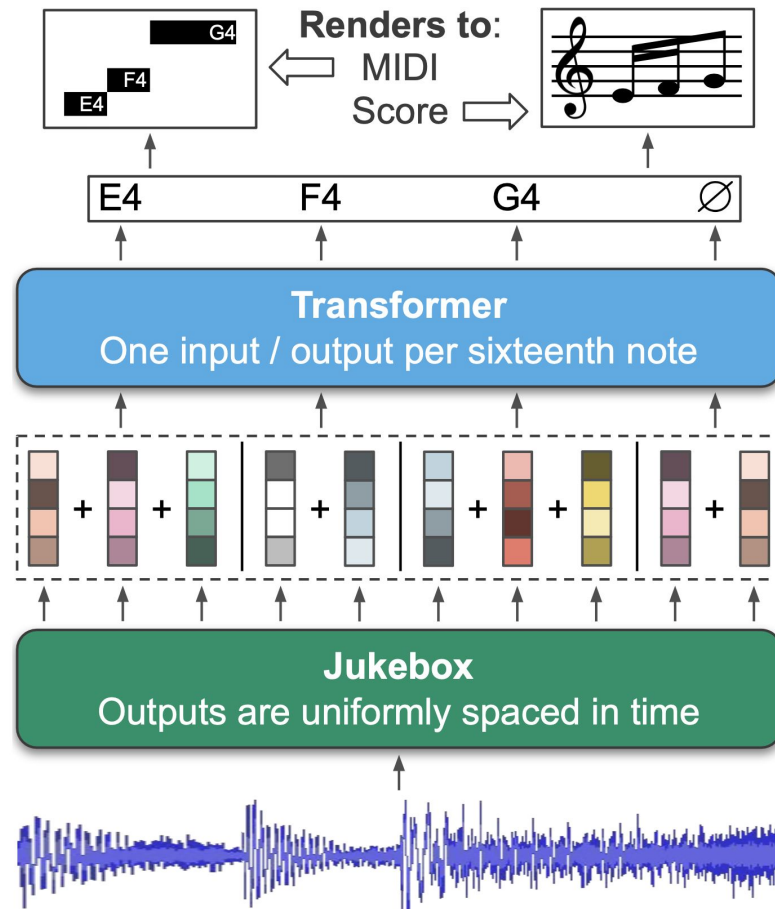
## Training



## Inference

# Model - SheetSage

- SheetSage
  - SOTA for lead sheet transcription
  - Take advantage of Jukebox prior

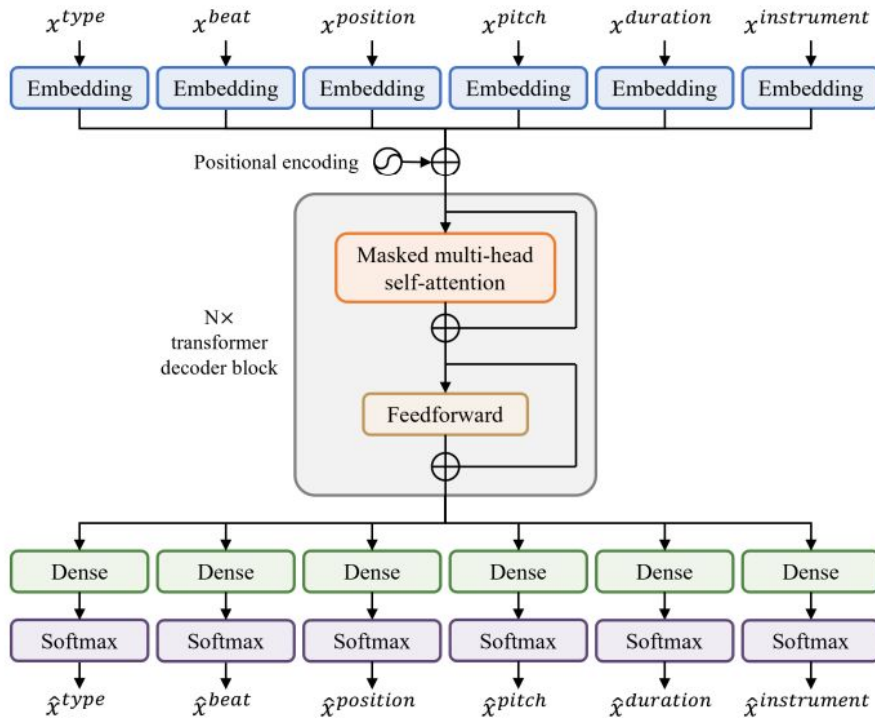


# Model - Multi-track Music Transformer

- MMT
  - Model structure
  - MMT representation

Model	Multitrack	Instrument control	Compound tokens	Generative modeling
REMI [5]				✓
MMM [10]	✓			✓
CP [6]			✓	✓
MusicBERT [15]	✓		✓	
FIGARO [11]	✓			✓
MMT (ours)	✓	✓	✓	✓

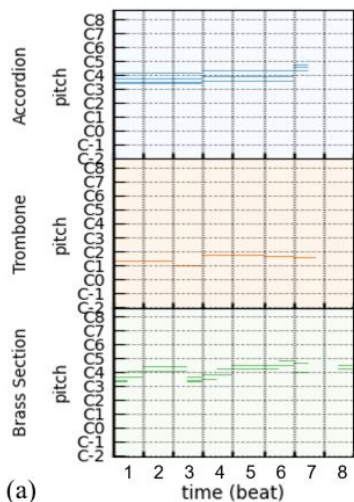
(Representation)



(Structure)

# Model - Multi-track Music Transformer (cont.)

- MMT
  - Unconditioned generation
  - Instrument-informed generation
  - N-beat continuation



(b)

```
(0, 0, 0, 0, 0, 0)
(1, 0, 0, 0, 0, 15)
(1, 0, 0, 0, 0, 36)
(1, 0, 0, 0, 0, 39)
(2, 0, 0, 0, 0, 0)
(3, 1, 1, 41, 15, 36)
(3, 1, 1, 65, 4, 39)
(3, 1, 1, 65, 17, 15)
(3, 1, 1, 68, 4, 39)
(3, 1, 1, 68, 17, 15)
(3, 1, 1, 73, 17, 15)
(3, 1, 13, 68, 4, 39)
(3, 1, 13, 73, 4, 39)
(3, 2, 1, 73, 12, 39)
(3, 2, 1, 77, 12, 39)
...
(4, 0, 0, 0, 0, 0)
```

(c)

Start of song ←

Instrument: accordion  
Instrument: trombone  
Instrument: brasses

Start of notes

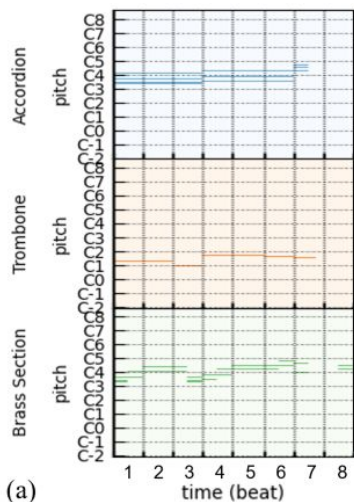
Note: beat=1, position=1, pitch=E2, duration=48, instrument=trombone  
Note: beat=1, position=1, pitch=E4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=E4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=G4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=C5, duration=72, instrument=accordion  
Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses  
Note: beat=2, position=1, pitch=C5, duration=36, instrument=brasses  
Note: beat=2, position=1, pitch=E5, duration=36, instrument=brasses

...  
End of song

(c)

# Model - Multi-track Music Transformer (cont.)

- MMT
  - Unconditioned generation
  - Instrument-informed generation
  - N-beat continuation



(b)

(0, 0, 0, 0, 0, 0)
(1, 0, 0, 0, 0, 15)
(1, 0, 0, 0, 0, 36)
(1, 0, 0, 0, 0, 39)
(2, 0, 0, 0, 0, 0)
(3, 1, 1, 41, 15, 36)
(3, 1, 1, 65, 4, 39)
(3, 1, 1, 65, 17, 15)
(3, 1, 1, 68, 4, 39)
(3, 1, 1, 68, 17, 15)
(3, 1, 1, 73, 17, 15)
(3, 1, 13, 68, 4, 39)
(3, 1, 13, 73, 4, 39)
(3, 2, 1, 73, 12, 39)
(3, 2, 1, 77, 12, 39)
...
(4, 0, 0, 0, 0, 0)

(b)

(c)

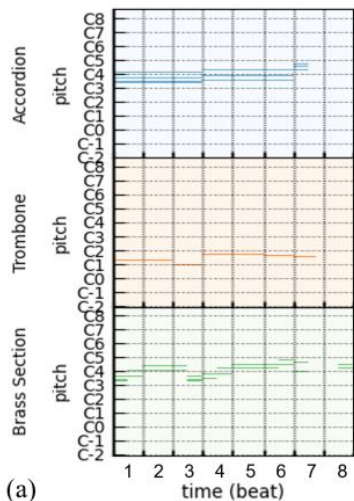
Start of song  
Instrument: accordion  
Instrument: trombone  
Instrument: brasses  
Start of notes

Note: beat=1, position=1, pitch=E2, duration=48, instrument=trombone  
Note: beat=1, position=1, pitch=E4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=E4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=G4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=C5, duration=72, instrument=accordion  
Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses  
Note: beat=2, position=1, pitch=C5, duration=36, instrument=brasses  
Note: beat=2, position=1, pitch=E5, duration=36, instrument=brasses  
...  
End of song

(c)

# Model - Multi-track Music Transformer (cont.)

- MMT
  - Unconditioned generation
  - Instrument-informed generation
  - N-beat continuation



(b)

```
(0, 0, 0, 0, 0, 0)
(1, 0, 0, 0, 0, 15)
(1, 0, 0, 0, 0, 36)
(1, 0, 0, 0, 0, 39)
(2, 0, 0, 0, 0, 0)
(3, 1, 1, 41, 15, 36)
(3, 1, 1, 65, 4, 39)
(3, 1, 1, 65, 17, 15)
(3, 1, 1, 68, 4, 39)
(3, 1, 1, 68, 17, 15)
(3, 1, 1, 73, 17, 15)
(3, 1, 13, 68, 4, 39)
(3, 1, 13, 73, 4, 39)
(3, 2, 1, 73, 12, 39)
(3, 2, 1, 77, 12, 39)
...
(4, 0, 0, 0, 0, 0)
```

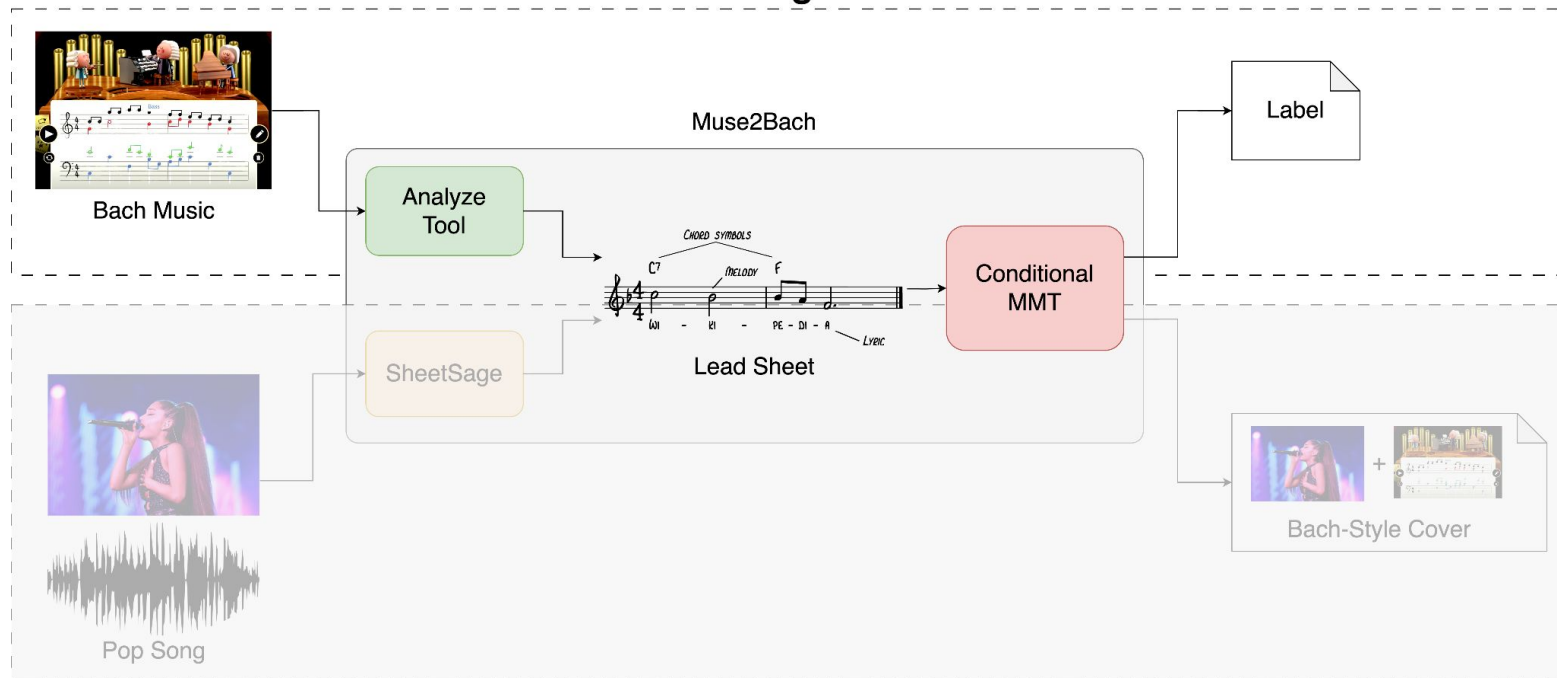
Start of song  
Instrument: accordion  
Instrument: trombone  
Instrument: brasses  
Start of notes  
Note: beat=1, position=1, pitch=E2, duration=48, instrument=trombone  
Note: beat=1, position=1, pitch=E4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=E4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=1, pitch=G4, duration=72, instrument=accordion  
Note: beat=1, position=1, pitch=C5, duration=72, instrument=accordion  
Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses  
Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses  
Note: beat=2, position=1, pitch=C5, duration=36, instrument=brasses  
Note: beat=2, position=1, pitch=E5, duration=36, instrument=brasses  
...  
End of song

(c)



# Tranining

## Training



## Inference

# Data

- Dataset: Chorale & Doodle
  - Doodle
    - user-entered melody + generated harmonization
    - random select 100000 segments with feedback = 2
    - 2 bars, four-part
  - Chorale
    - Bach four-part chorales
    - 189 full songs

# Data (cont.)

- Training details

- Chorale

- Chord

- early stop at 22000 steps

- No chord

- early stop at 23000 steps

- Doodle

- Chord

- No chord

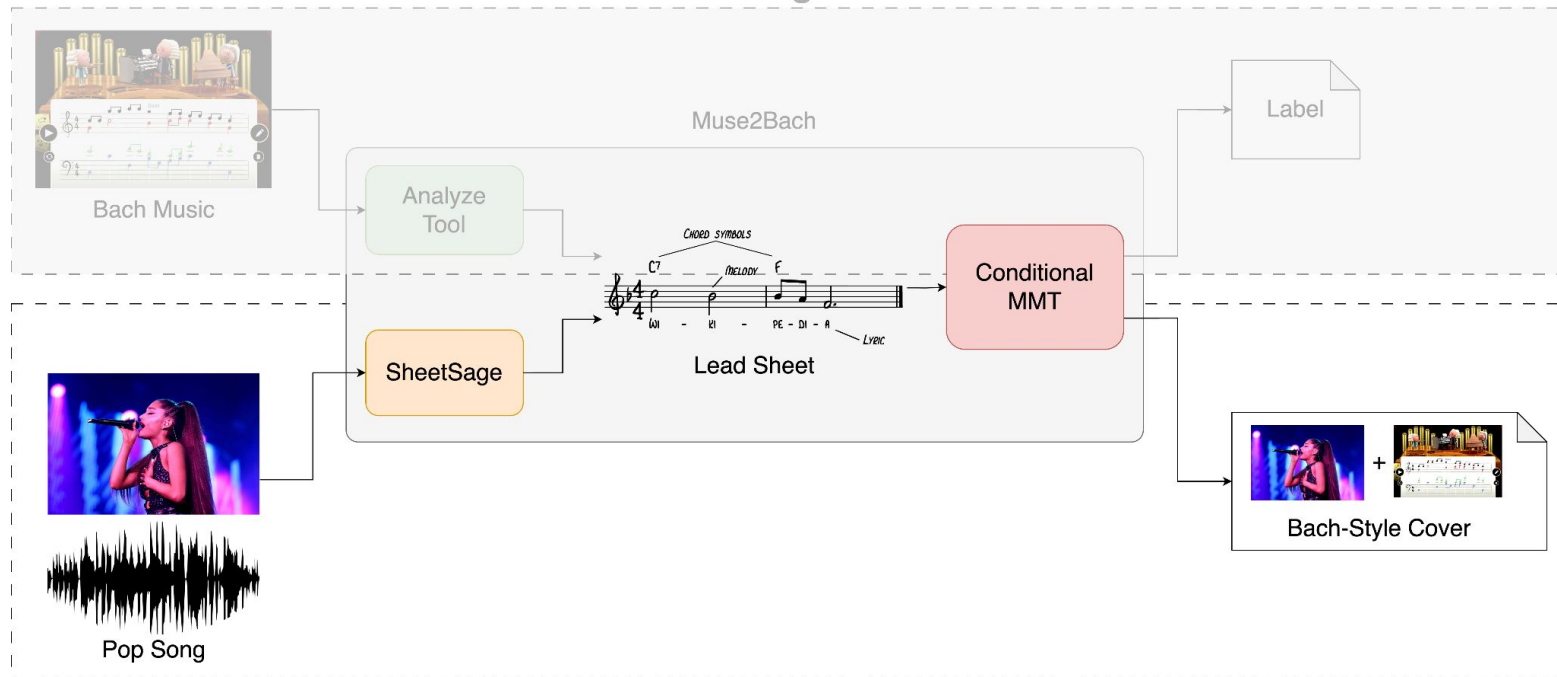
- MMT Representation

$(x^{\text{type}}, x^{\text{beat}}, x^{\text{position}}, x^{\text{pitch}}, x^{\text{duration}}, \mathbf{x^{\text{chord}}}, x^{\text{instrument}})$

- ❖ max input length = 1024
- ❖ embedding length = 512
- ❖ # of transformer layers = 6
- ❖ # of attention heads = 8
- ❖ batch size = 8
- ❖ drop out = 0.2
- ❖ training max step = 200000

# Inference

## Training



## Inference

# Demo

# Demo

OMG

Ditto

夜に駆ける

Chorale w/ chord



Chorale w/o chord



Doodle w/ chord



Doodle w/o chord

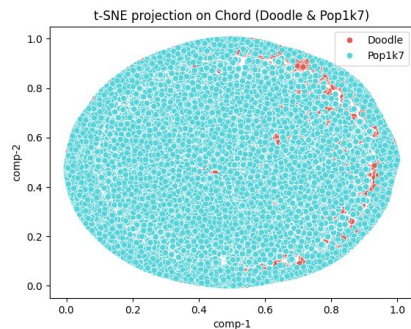
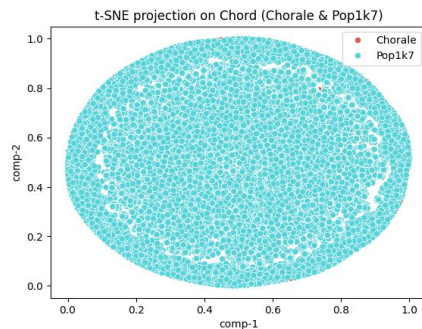
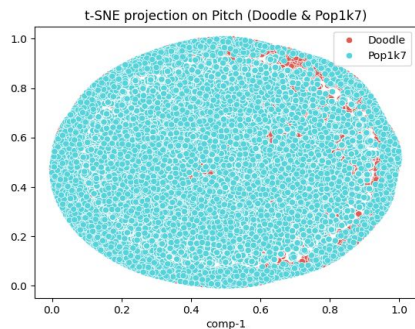
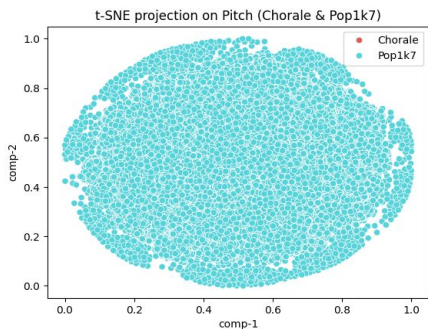


# Discussion

- 音樂差異
  - 巴洛克時代四部合聲 vs 現代流行樂

# Discussion

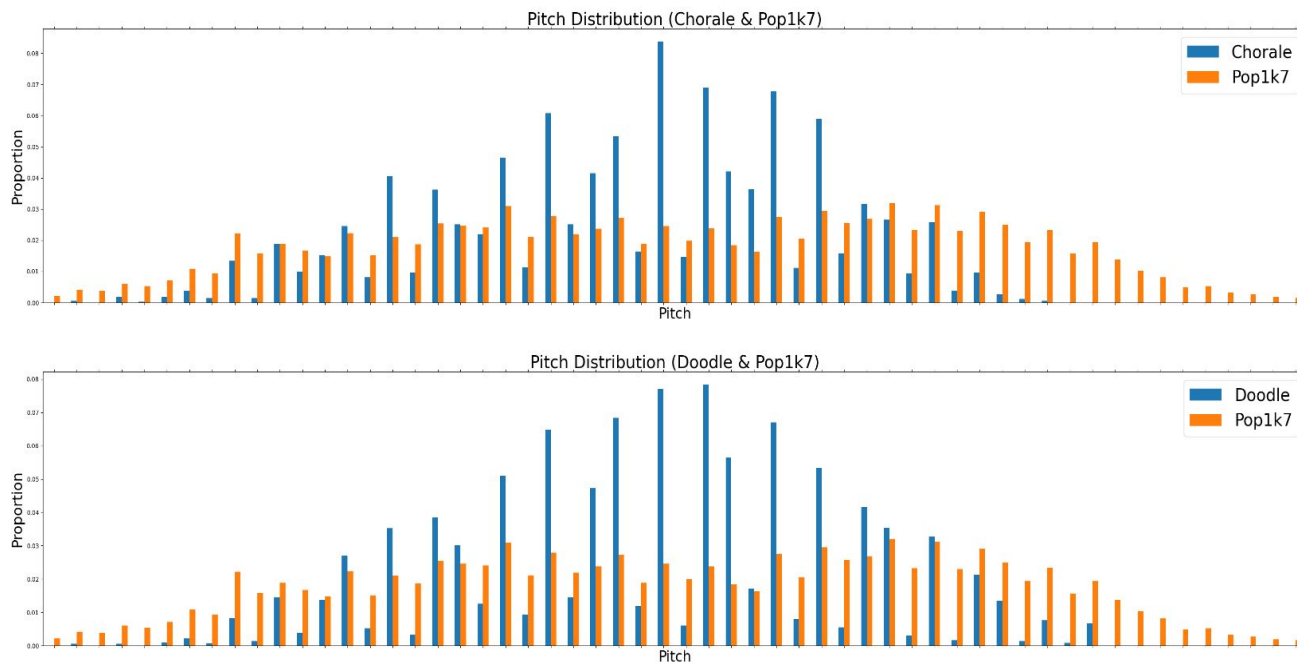
- Dataset bias (bach vs Pop1k7) - t-SNE for melody & chord





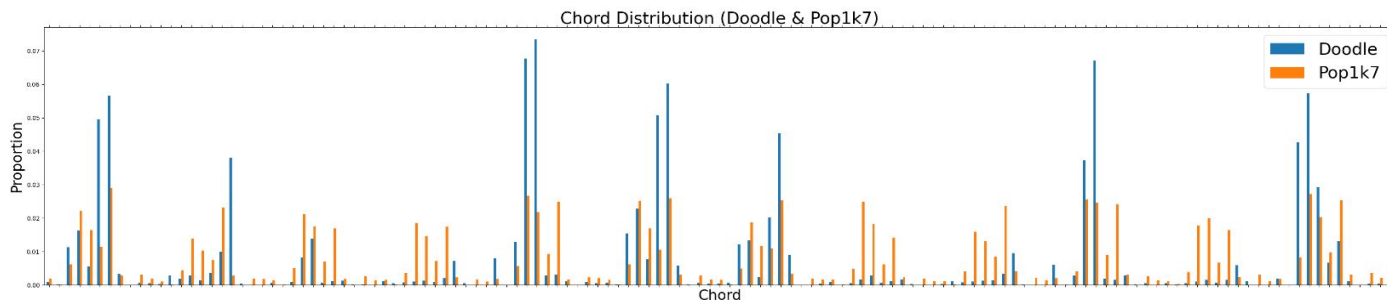
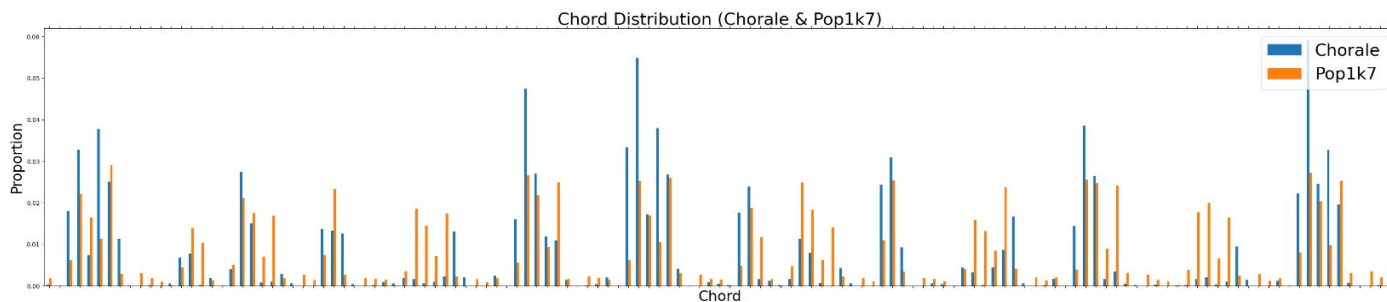
# Discussion

- Dataset bias (bach vs Pop1k7) - Pitch Distribution



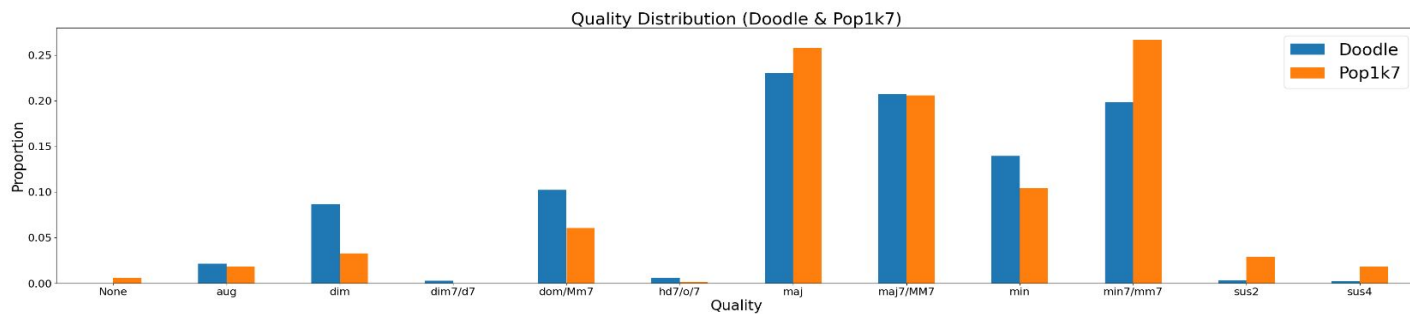
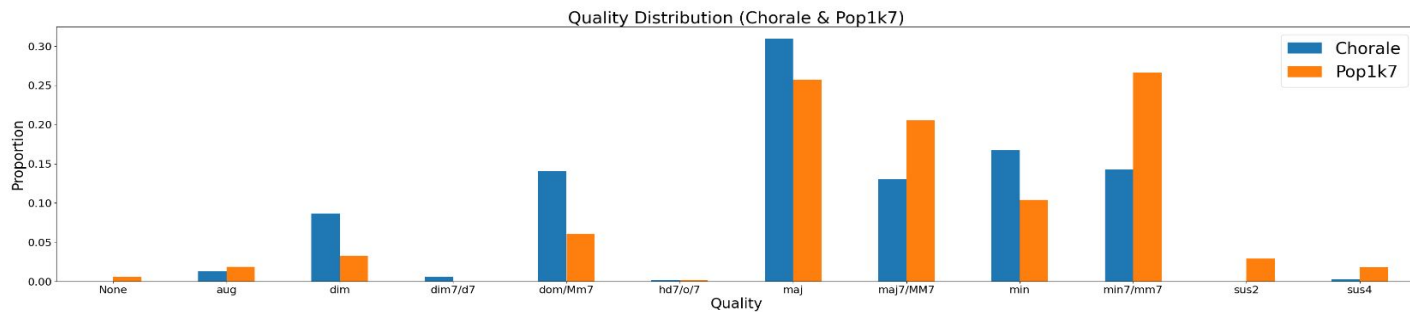
# Discussion

- Dataset bias (bach vs Pop1k7) - Chord Distribution (Root + Quality)



# Discussion

- Dataset bias (bach vs Pop1k7) - Chord Distribution (Quality)



# Conclusion

- Next step
  - Subjective evaluation
  - MMT's representation leads to empty bar in condition
  - Missing information of chord due to MMT's representation
  - Statistic is not clear
  - More condition
  - Lead-sheet free

The image displays a musical score in 4/4 time. The top staff is in treble clef, and the bottom staff is in bass clef. Above the treble staff, four chords are labeled: C, Am, Em, and G. The melody in the treble staff consists of the following notes: C4 (quarter), D4 (quarter), E4 (quarter), F4 (quarter), G4 (quarter), A4 (quarter), B4 (quarter), and C5 (quarter). The bass line consists of the following notes: C3 (half), F2 (half), G2 (half), and C3 (half). An orange box highlights the first three measures of the melody, which correspond to the C, Am, and Em chords. The fourth measure, corresponding to the G chord, is partially cut off by the right edge of the image.