

# Modeling Symbolic Music with Natural Language Processing Approaches

Analyzing and Experimenting with Multiple Levels of Representations

---

**Dinh-Viet-Toan LE**

PhD Thesis Defense

November 3, 2025

Xavier HINAUT	Chargé de recherche Inria, Inria Bordeaux	Reviewer
Cheng-Zhi Anna HUANG	Associate Professor, MIT	Reviewer
Emmanouil BENETOS	Reader & Director of research, QMUL	Examiner
Chloé BRAUD	Chargée de recherche CNRS, IRIT	Examiner
Marius BILASCO	Professeur des universités, CRISTAL	Examiner
Patrick BAS	Directeur de recherche CNRS, CRISTAL	Jury president
Louis BIGO	Professeur des universités, LaBRI	Co-director
Marc TOMMASI	Professeur des universités, CRISTAL	Co-director
Mikaela KELLER	Maîtresse de conférence, CRISTAL	Co-supervisor

# Music & language

A musical score for Beethoven's Symphony No. 5, Movement 1, Bars 323-330. The score is in 2/4 time with a key signature of two flats. It features six instruments: Flute (Fl.), Clarinet (Cl.), Bassoon (Fg.), Violin 1 (Vln.1), Violin 2 (Vln.2), and Cello/Bass (Cb.). The score is annotated with colored boxes highlighting specific melodic or harmonic patterns. Yellow boxes appear above the Flute and Clarinet staves. Red boxes highlight patterns in the Bassoon and Violin 2 staves. Blue boxes highlight patterns in the Cello/Bass staff. The notes are primarily eighth and sixteenth notes, with some rests and grace notes.

beethoven5

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1. Bars 323–330.

# Music & language

“A musical phrase is passed through the instruments like a dialog.”

A musical score for Beethoven's Symphony No. 5, Movement 1, Bars 323-330. The score is in 2/4 time with a key signature of four flats. It features eight staves: Flute (Fl.), Clarinet (Cl.), Bassoon (Fg.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (C. Cb.), Double Bass (Vcl.), and Bassoon (Vla.). The music is divided into three sections by color-coded boxes: yellow boxes highlight the Flute and Clarinet parts, red boxes highlight the Bassoon and Double Bass parts, and blue boxes highlight the Violin and Cello parts. The score shows various musical patterns, including eighth-note chords and sixteenth-note figures.

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1. Bars 323–330.

- “Call-and-response”
- “Musical discourse”
- “A composer’s vocabulary”
- ...
- “*Musical language*”

---

Jackendoff, *Parallels and nonparallels between language and music*, 2009.

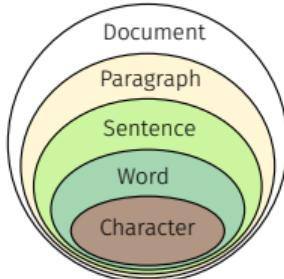
# Music & language: some fundamental similarities...

## Modalities

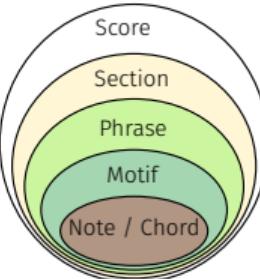
	Auditory	Written
Language	Speech Audio	Text Score
Music		

## Segmentation

### Language



### Music



## Sequential representation

```
<Model_><_ing> <music_><_al> <content> <with> <NLP>
```



## “Grammar”

- Grammatical trees
- Tonal theory (GTTM [Lerdahl & Jackendoff 1987])

## Expectancy

- The cat is eating the [X]
- Functional harmony

no-cadence

tonic

# Music & language: ...and major differences

- Temporal dimension in language & music
- Polyphony in music
- Polymorphism of musical symbols

A musical score consisting of two staves: Treble (top) and Bass (bottom). The Treble staff has a key signature of one sharp (F#) and a time signature of common time (indicated by 'C'). The Bass staff has a key signature of one sharp (F#) and a time signature of common time (indicated by 'C').

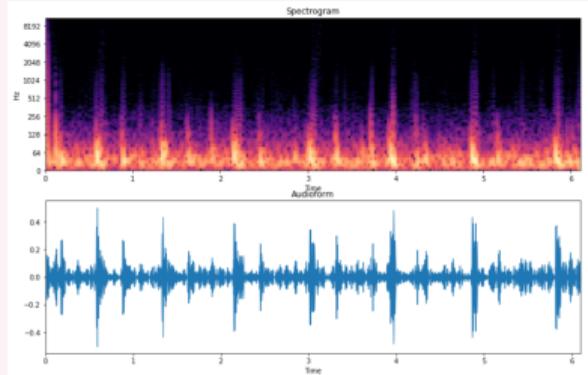
An orange box highlights a specific note in the Treble staff, which is a eighth-note with a sharp accidental. An orange arrow points from this note to a blue oval highlighting a measure in the Bass staff. A legend box provides the following information:

Pitch:	D4
Accidental:	#
Duration:	♪
Dynamics:	<b>p</b>

Below the staves, there are eight green vertical bars of varying heights, representing the temporal dimension or duration of the notes.

# Symbolic music

## Audio



- Waveforms, spectrograms...
- Timbre, performance expressivity...

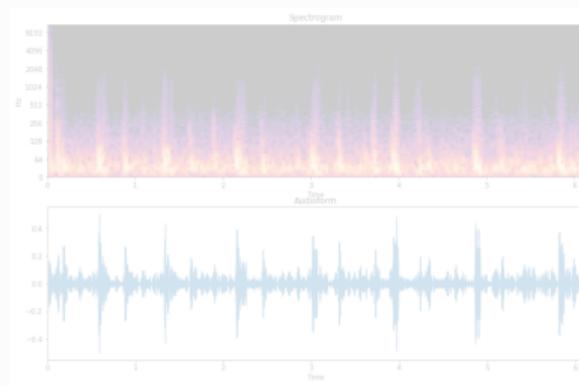
## Symbolic music



- Score (sheet music):
  - Notes, chords, instruments, dynamics, articulations...

# Symbolic music

## Audio



- Waveforms, spectrograms...
- Timbre, performance expressivity...

## Symbolic music



- Score (sheet music):
  - Notes, chords, instruments, dynamics, articulations...

## **Natural Language Processing (NLP)**

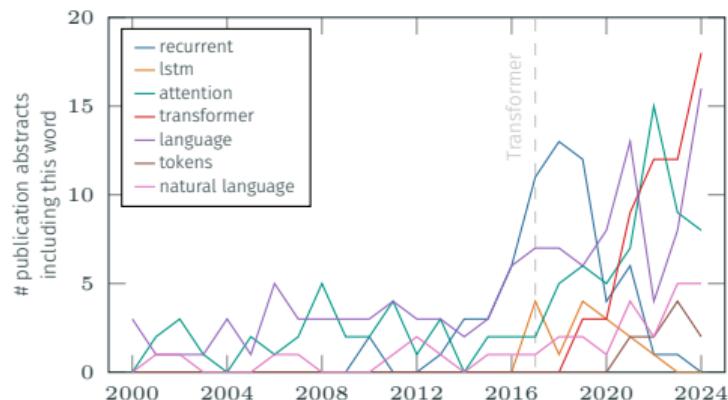
- Computer science applied to linguistics
- Interaction between computer and *written* human natural language

## **Music Information Retrieval (MIR)**

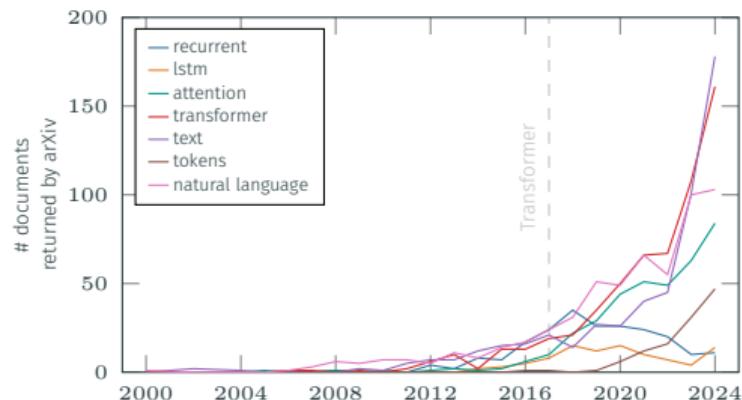
- Computer science applied to music
- Interaction between computer and (symbolic) music

# NLP tools in symbolic MIR research

Number of ISMIR papers which include NLP-related terms in their abstract

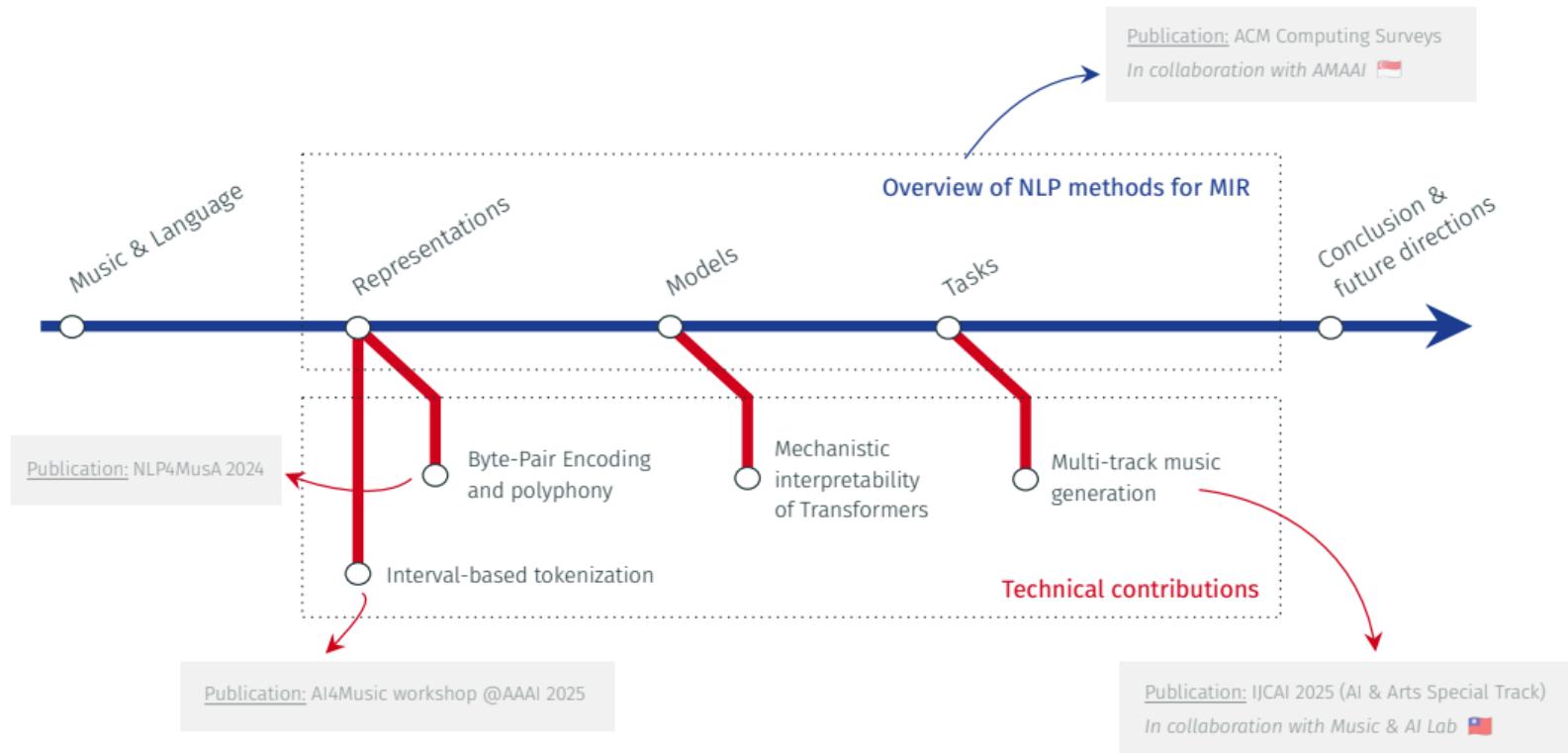


Number of MIR arXiv papers involving NLP-related terms

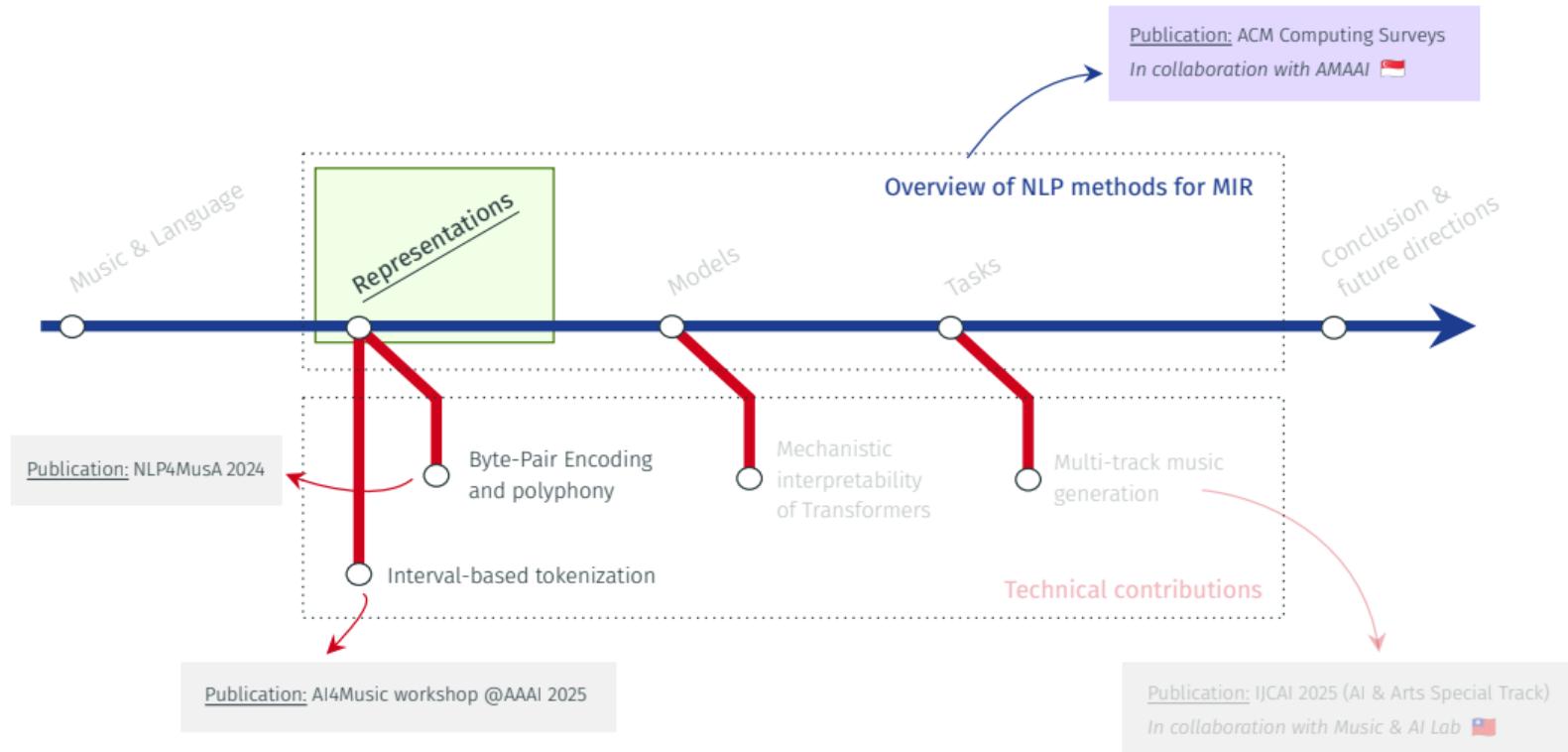


Le & al., *Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey*, ACM Computing Surveys 2025. In collaboration with AMAAI (Singapore University of Technology and Design).

# Outline



# Outline – Sequential representations



# Tokenization in text

## Tokenization

**Representing** a complex content (e.g. text, music) into a **sequence** of elements for computational processing.

# Tokenization in text

## Tokenization

Representing a complex content (e.g. text, music) into a sequence of elements for computational processing.

I love singing but it is exhausting

### Word-level tokenization

→ <I>, <love>, <singing>, <but>, <it>, <is>, <exhausting>

### Character-level tokenization

→ <I>, <\_>, <l>, <o>, <v>, <e>, <\_>, <s>, <i>, <n>, ...

### Subword-level tokenization

→ <I>, <love>, <sing>, <\_ing>, <but>, <it>, <is>, <exhaust>, <\_ing>

# What does a musical score encode?

A musical score excerpt for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (Cb.), and Bassoon (Vla.). The score is labeled with various musical parameters:

- Tempo:**  $\text{♩} = 80$
- Rhythm:** Indicated by vertical dashed lines separating measures.
- Dynamics:** Dynamics are shown above the staves, such as  $p$  (piano) for Vln.1 and Vln.2, and  $p$  (pianissimo) for Cb. Articulations like "pizz." (pizzicato) and "arco." (arco) are also indicated.
- Pitch:** The pitch is indicated by the letter  $e$  above the staves.
- Polyphony:** The vertical arrangement of the staves represents polyphony.
- Structure:** The structure is indicated by measure numbers and repeat signs, such as "a 2." and a repeat sign with a "2." above it.
- Articulations:** Articulations are indicated by specific markings on the stems of the notes, such as dots and dashes.

# What does a musical score encode?

The image shows a musical score for six instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (Cb.), Double Bass (Vc.), and Bassoon (Vla.). The score is labeled with measure number 101 and tempo  $\text{♩} = 80$ . The music consists of two systems. The first system starts with a rest for the Oboe, followed by eighth-note patterns for Vln.1, Vln.2, and Vla. The second system begins with a bassoon solo (labeled 'a 2.' above the staff) followed by eighth-note patterns for Vln.1, Vln.2, and Vla. The Vln.1 and Vln.2 parts in the second system are highlighted with blue boxes.

Annotations on the left side of the score include:

- Dynamics**: Labels  $p$  (piano) under Vln.1 and Vln.2, and  $p$  (pianissimo) under Vla.
- Pitch**: Labels  $e$  (pitch) above the staves of Ob., Vln.1, Vln.2, Vla., and Vc./Cb.
- Rhythm**: Labels  $-$  (rest) above the staves of Ob., Vln.1, Vln.2, Vla., and Vc./Cb.
- Articulations**: Labels *pizz.* (pizzicato) above Vla. and *arco.* (arco) above Vc./Cb.
- Polyphony**: Labels  $\times$  (crosses) below the staves of Vln.1, Vln.2, Vla., and Vc./Cb. in the first system, indicating multiple voices or parts.
- Structure**: Labels  $a 2.$  (a 2.) above the bassoon staff in the second system, indicating a repeat or section label.

# What does a musical score encode?

The image shows a musical score for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (Cb.), and Double Bass (Vc.). The score is labeled "101" and includes a tempo marking of "♩ = 80". A pink box highlights the word "Rhythm" above the score. Annotations include:

- Dynamics:** Dynamics like "p" (piano) and "pizz." (pizzicato) are marked with blue boxes.
- Pitch:** The pitch "e" is marked with a blue box.
- Rhythm:** Rhythmic patterns are highlighted with red boxes, including a grace note and a sixteenth-note pattern.
- Articulations:** Articulation marks like "pizz." and "arco." are marked with red boxes.
- Polyphony:** The term "Polyphony" is written below the score.
- Structure:** The term "Structure" is written to the right of the score.

# What does a musical score encode?

The image shows a musical score excerpt with five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (Cb.), and Bassoon (Vla.). The score includes several annotations:

- Tempo:** A green box labeled "Tempo" contains the marking  $\text{♩} = 80$ .
- Rhythm:** A red box labeled "Rhythm" highlights a sixteenth-note pattern in the Vln.1 staff.
- Dynamics:** The word "Dynamics" is positioned to the left of the score, with specific dynamic markings like  $p$  (piano) and  $f$  (fortissimo) highlighted by blue boxes.
- Pitch:** A blue box labeled "Pitch" highlights the pitch  $e$  on the Ob. staff.
- Articulations:** The word "Articulations" is positioned to the right of the score, with articulation marks like "pizz." (pizzicato) and "arco." (arco) highlighted by red boxes.
- Polyphony:** The word "Polyphony" is positioned below the score, highlighting the simultaneous multiple voices of the instruments.
- Structure:** The word "Structure" is positioned to the right of the score, highlighting the overall form or section, indicated by the marking "a 2."

# What does a musical score encode?

The image shows a musical score with five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello (Vcl.), and Double Bass (Cb.). The score includes several annotations:

- Tempo:**  $\text{♩} = 80$
- Dynamics:** Dynamics like  $p$  (piano) and **pizz.** (pizzicato) are indicated.
- Pitch:** The pitch  $e$  is marked on the first staff.
- Rhythm:** Rhythmic values like eighth and sixteenth notes are shown.
- Articulations:** Articulations like **pizz.** and **arco.** are marked.
- Polyphony:** The concept of multiple voices or parts playing simultaneously is illustrated.
- Structure:** A vertical pink bar highlights a structural element, with smaller pink marks below it.

# What does a musical score encode?

The image shows a musical score for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello/Violoncello (Vla./Cb.), and Double Bass (Vc. Cb.). The score is annotated with several colored boxes highlighting different musical features:

- Tempo:** A green box contains the tempo marking  $\text{♩} = 80$ .
- Rhythm:** A pink box highlights a measure where the Vln.1 part has a sixteenth-note pattern.
- Dynamics:** An orange box highlights the dynamic  $p$  (pianissimo) for the Vln.1 and Vln.2 parts.
- Pitch:** A blue box highlights the pitch  $e$  on the treble clef staff.
- Polyphony:** An orange box highlights the simultaneous notes played by the Vln.1, Vln.2, and Vla. parts.
- Articulations:** A pink box highlights the articulation *pizz.* (pizzicato) for the Vln.1 and Vla. parts.
- Structure:** A pink box highlights a vertical bar line and the repeat sign  $a 2.$  at the end of a section.

# What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Polyphony

Structure

Articulations

101

♩ = 80

a 2.

p

pizz.

arco.

Ob.

Vln.1

Vln.2

Vla.

Vcl. Cb.

Vcl. Cb.

pizz.

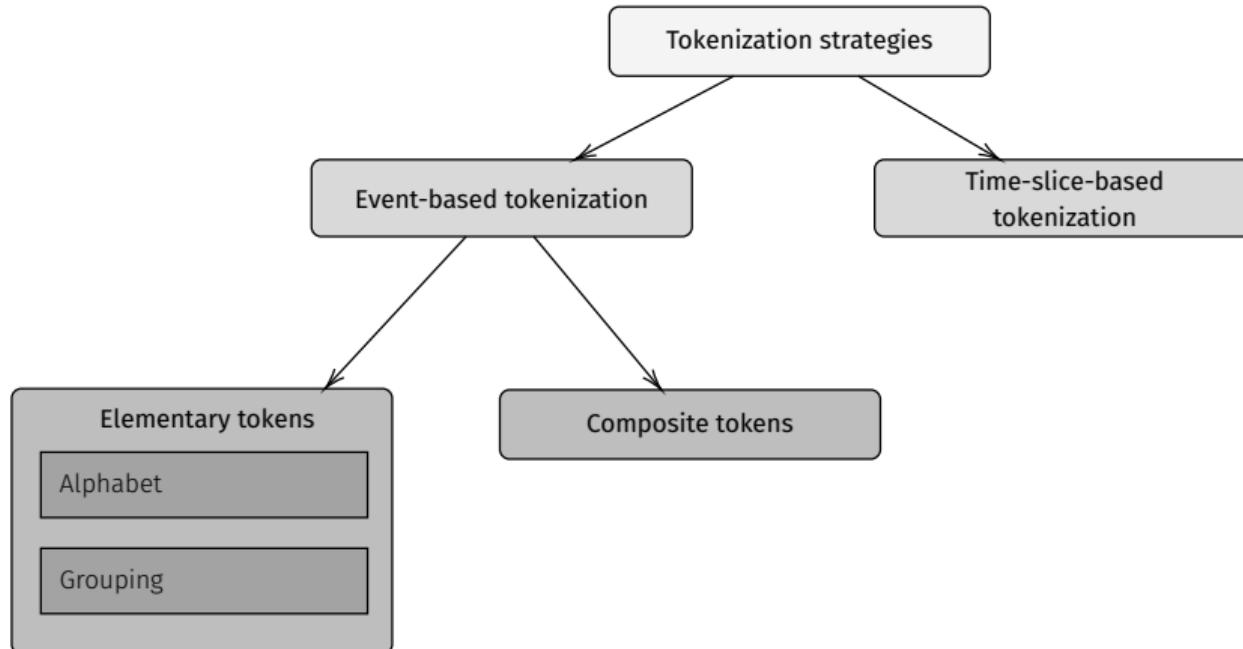
Structure

# What does a musical score encode?

The image shows a musical score with five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Cello/Bass (Vla.), and Double Bass/Cello (Vc./Cb.). The score includes several annotations:

- Tempo:**  $\text{♩} = 80$
- Rhythm:** Various rhythmic patterns are shown, with some notes highlighted in yellow.
- Dynamics:** Dynamics like  $p$  (piano) and  $f$  (fortissimo) are indicated by yellow boxes.
- Pitch:** The pitch is labeled with  $e$  on the treble clef staves and  $e$  on the bass clef staff.
- Polyphony:** The vertical arrangement of multiple voices per staff is highlighted.
- Articulations:** Articulation marks like "pizz." (pizzicato) and "arco." (arco) are shown.
- Structure:** A vertical pink bar highlights a section of the score, with small pink arrows pointing downwards along the staff.

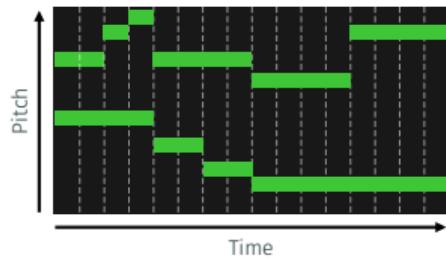
# A taxonomy of tokenization strategies for symbolic music



# How to tokenize music?



Pianoroll slices



MIDI-VAE [Brunner & al. 2018]

Harmony Transformer [Chen & al. 2019]

RNBert [Sailor 2024]



D5, \_\_, E5, F5, D5, \_\_, \_\_, \_\_, C5, \_\_, \_\_, \_\_, E5  
A4, \_\_, \_\_, G4, \_\_, F4, \_\_, E4, \_\_, \_\_, \_\_, E4  
C4, \_\_, \_\_, B3, \_\_, \_\_, \_\_, G3, \_\_, \_\_, \_\_, A3  
F3, \_\_, D3, \_\_, G3, \_\_, \_\_, \_\_, C2, \_\_, \_\_, \_\_, C#2  
1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1  
0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0

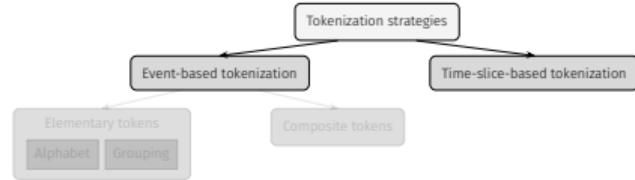
DeepBach [Hadjeres & al. 2017]



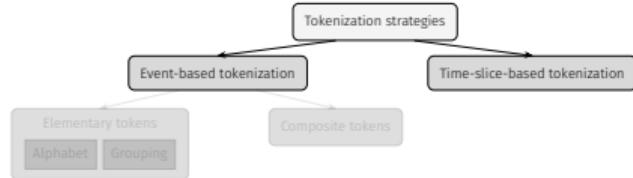
1	o--o--	17	---o-	33	---o-	49	o-o---
2	-----	18	-----	34	-----	50	-----
3	---o-	19	---oo-	35	---oo-	51	---oo-
4	-----	20	-----	36	-----	52	-----
5	--o-o-	21	--o-o-	37	--o-o-	53	--o-o-
6	-----	22	-----	38	-----	54	-----
7	---o-	23	---o-	39	---o-	55	---o-
8	-----	24	--o---	40	---o-	56	---o-
9	--oo-	25	--oo-	41	--o-o-	57	--o-o-
10	-----	26	-----	42	-----	58	-----
11	---oo-	27	---oo-	43	---oo-	59	---oo-
12	-----	28	-----	44	--o-o-	60	--o-o-
13	--o-o-	29	--o-o-	45	--o-o-	61	--o-o-
14	-----	30	-----	46	--o-o-	62	--o-o-
15	--oo-	31	--oo-	47	--o-	63	--oo-
16	-----	32	-----	48	---o-	64	-----

Drums tokenization [Zhang & al. 2023]

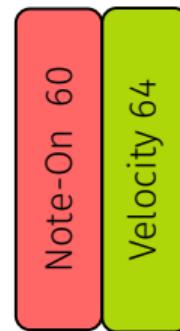
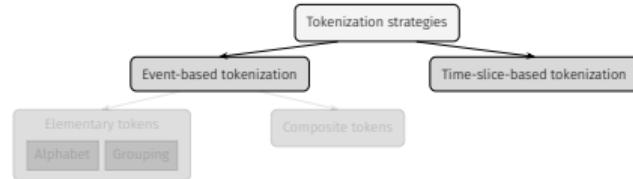
# How to tokenize music?



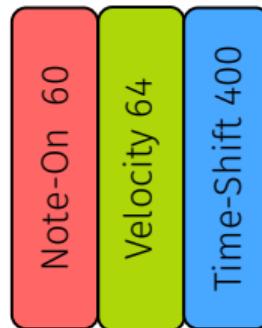
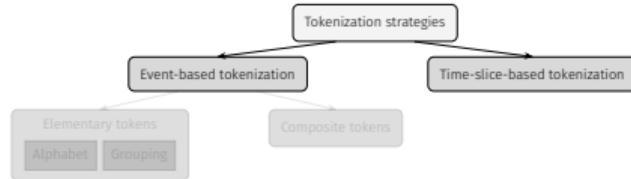
# How to tokenize music?



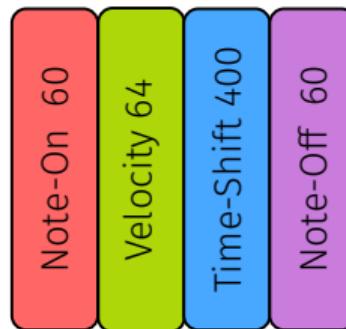
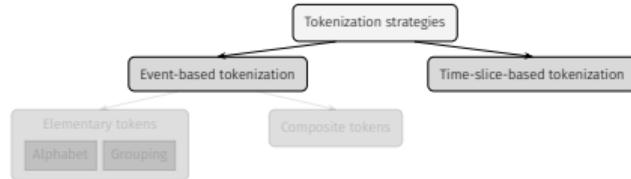
# How to tokenize music?



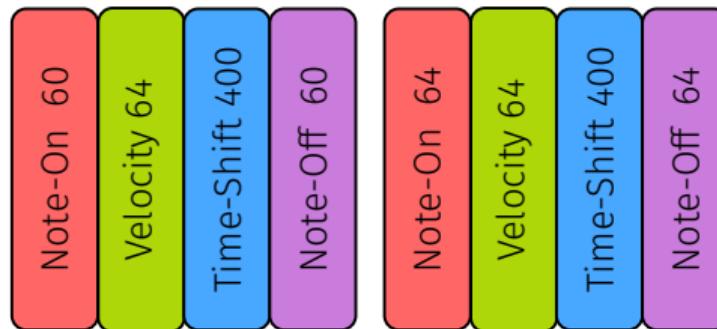
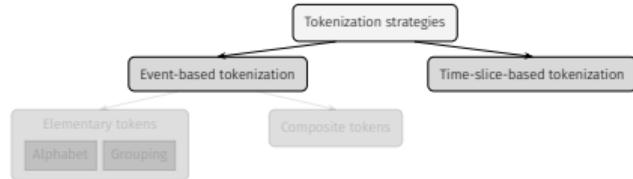
# How to tokenize music?



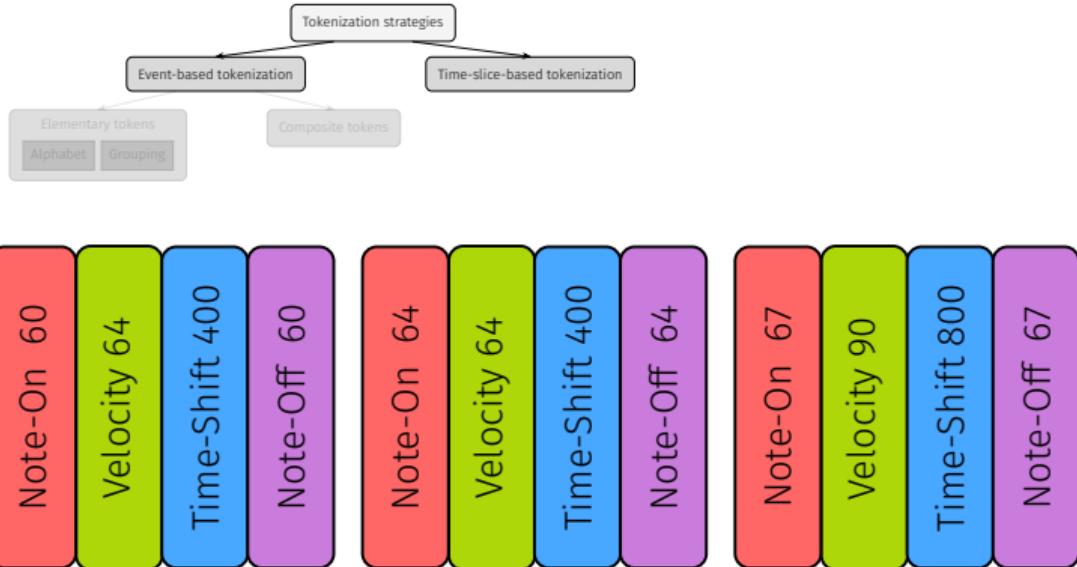
# How to tokenize music?



# How to tokenize music?

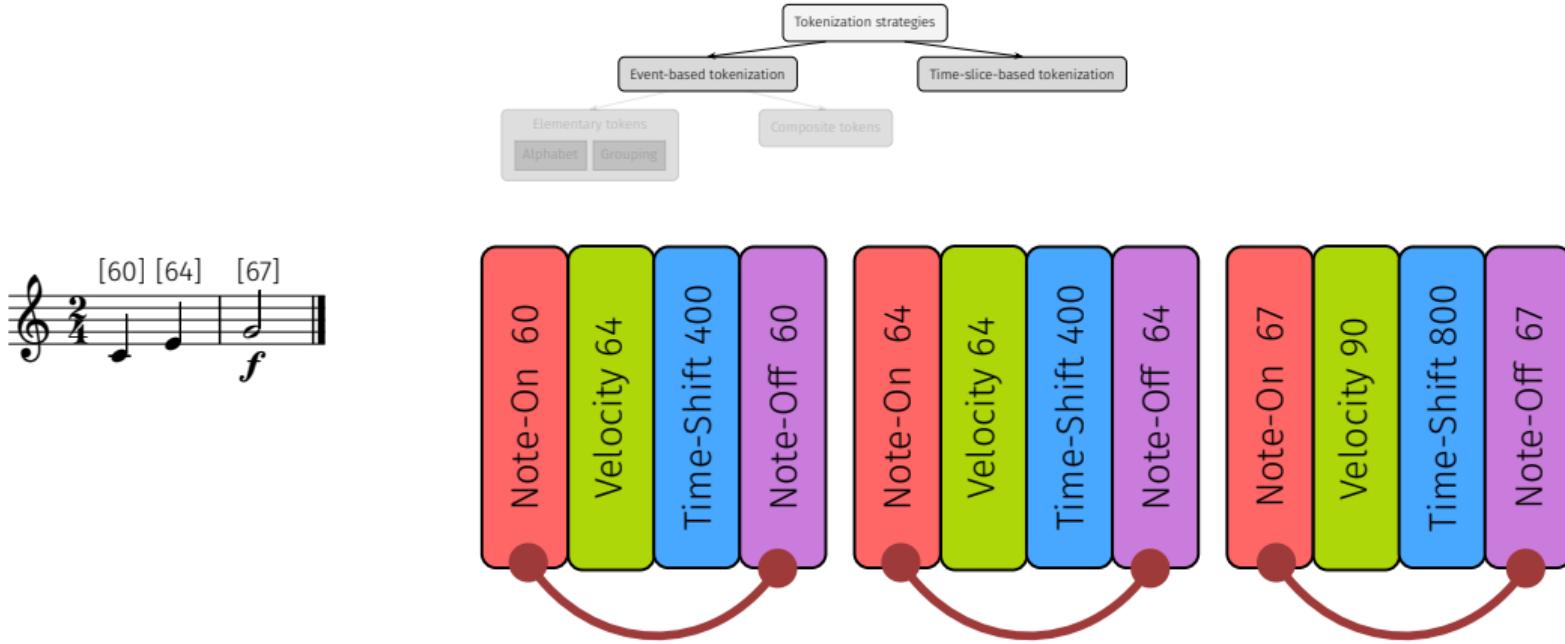


# How to tokenize music?

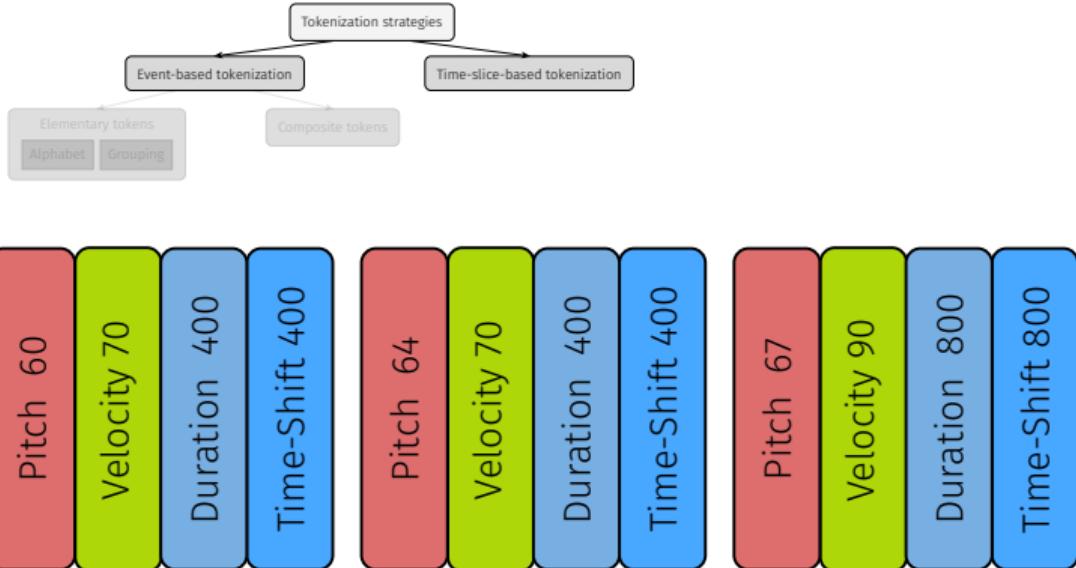


MIDI-like: Huang & al., *Music Transformer: Generating Music with Long-term Structure*, ICLR 2019

# How to tokenize music?

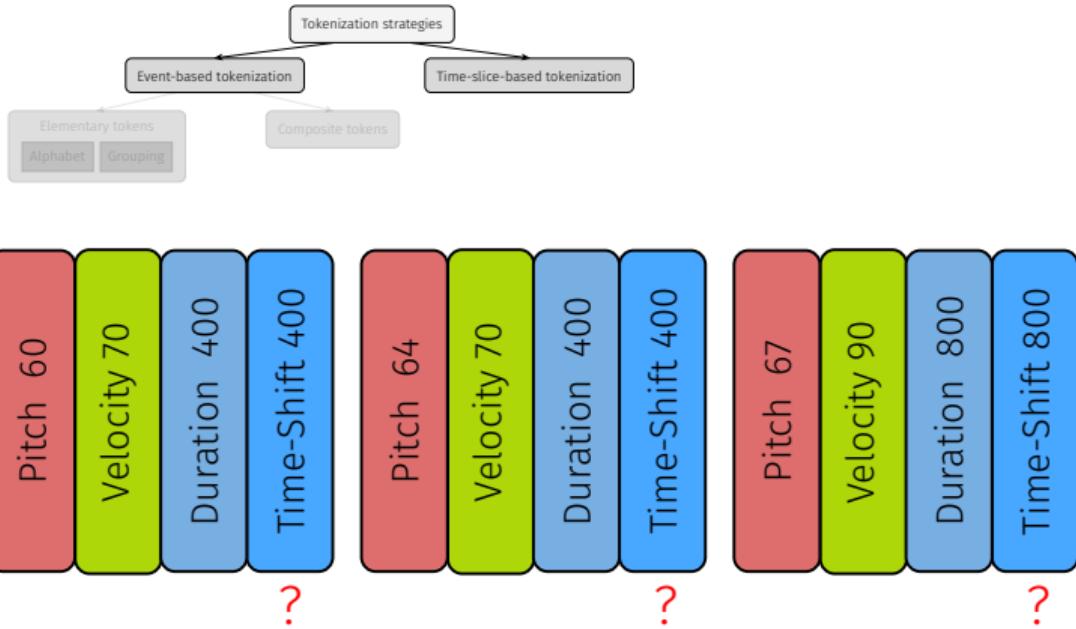


# How to tokenize music?

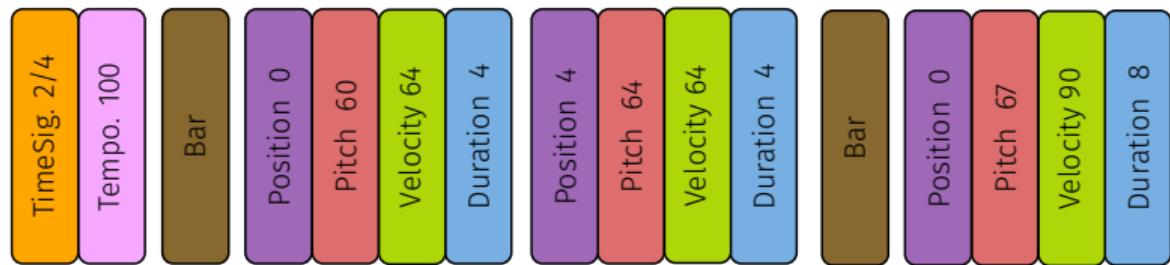
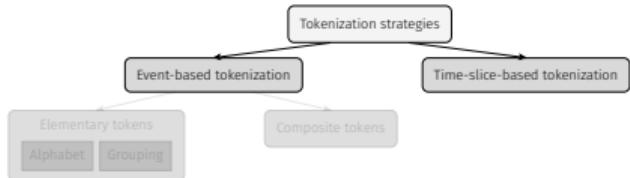


Structured: Hadjares & al., *The Piano Inpainting Application*, 2021

# How to tokenize music?

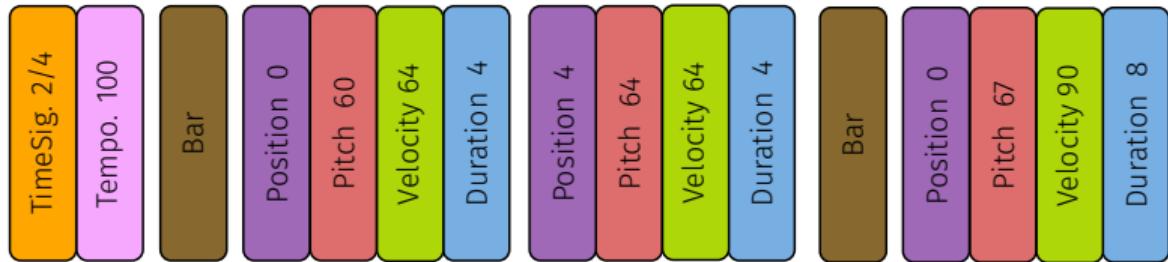
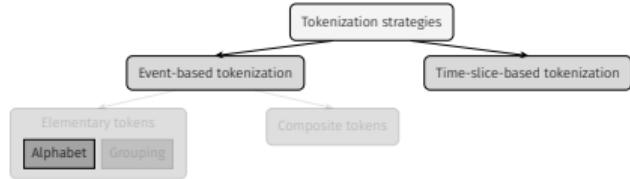


# How to tokenize music?

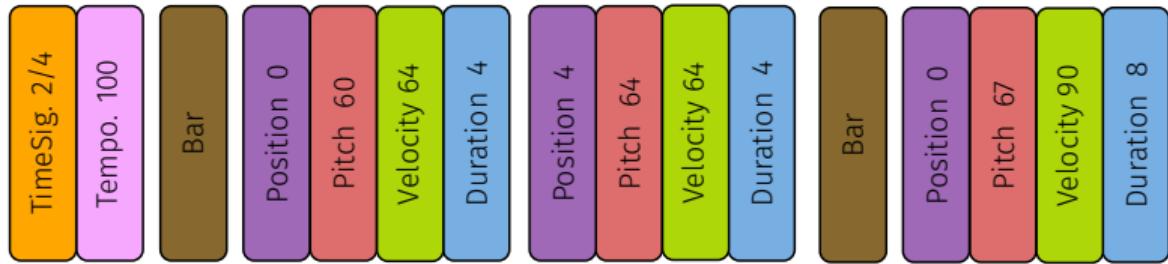
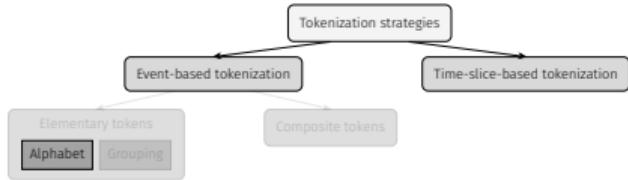


REMI: Huang & al., *Pop Music Transformer*, MM 2020

# How to tokenize music?

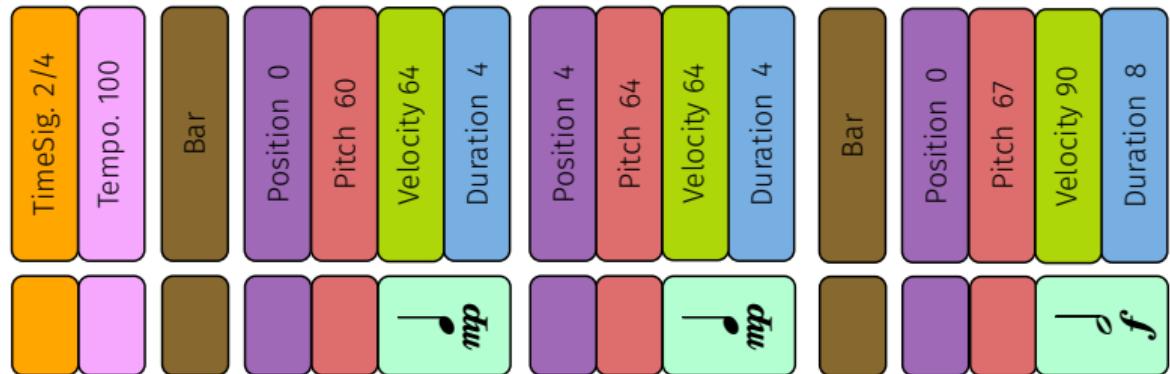
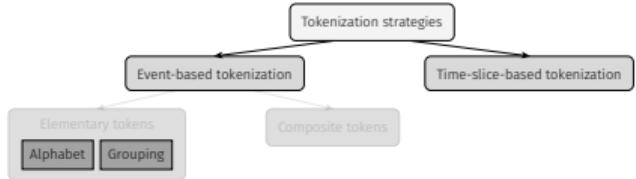


# How to tokenize music?



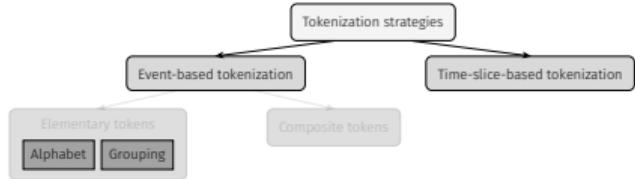
Grouping tokens together?

# How to tokenize music?

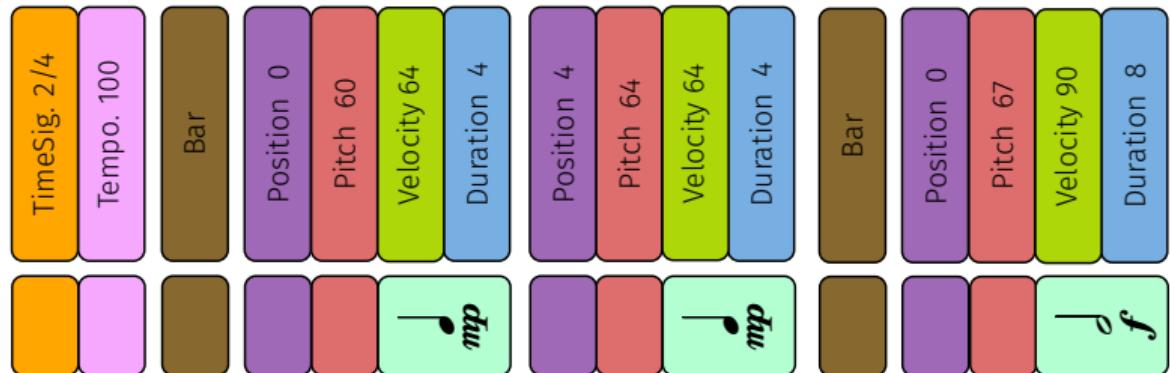


Rule-based ; n-gram: Conklin & al., 1995; Byte-Pair Encoding: Fradet & al., 2023

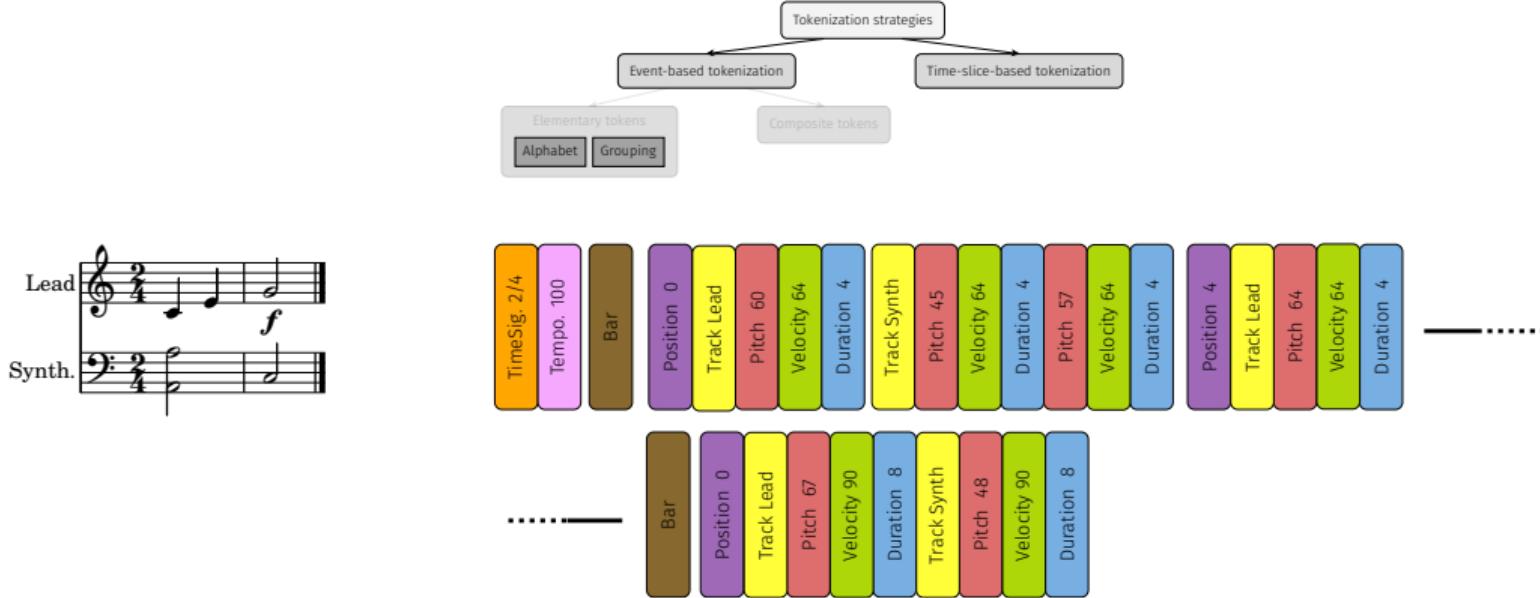
# How to tokenize music?



Multi-track?

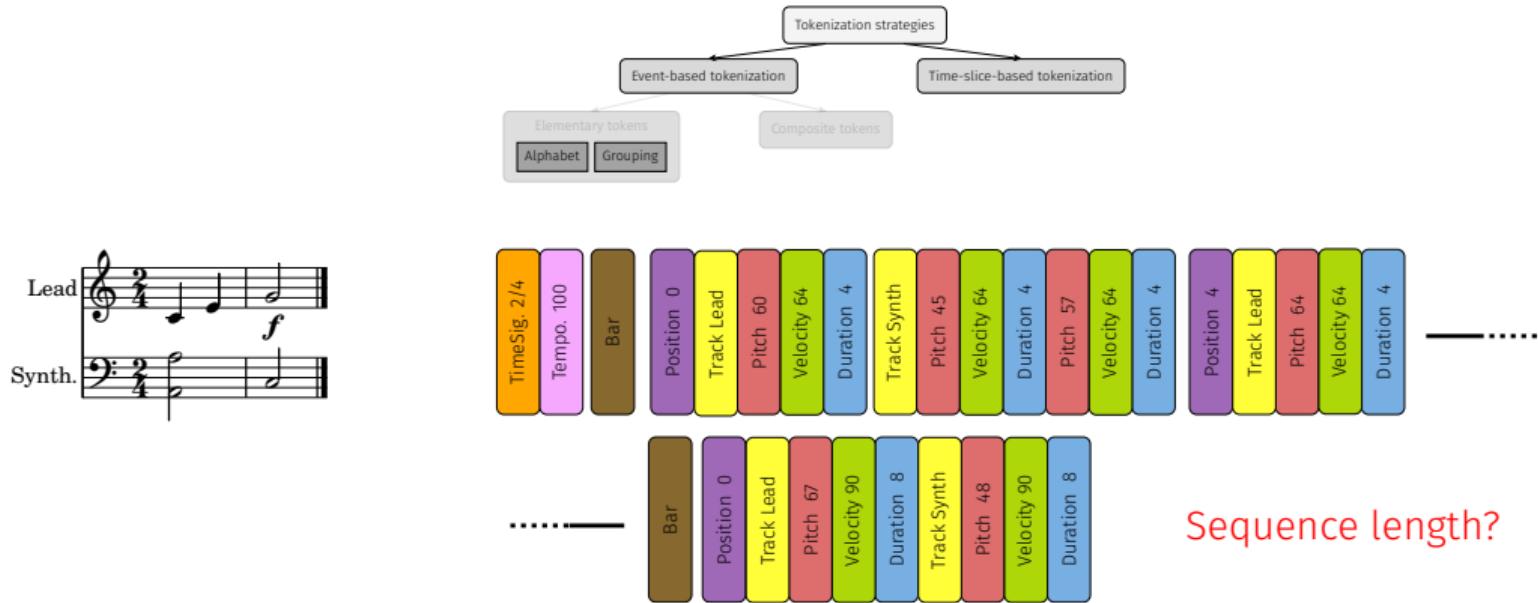


# How to tokenize music?

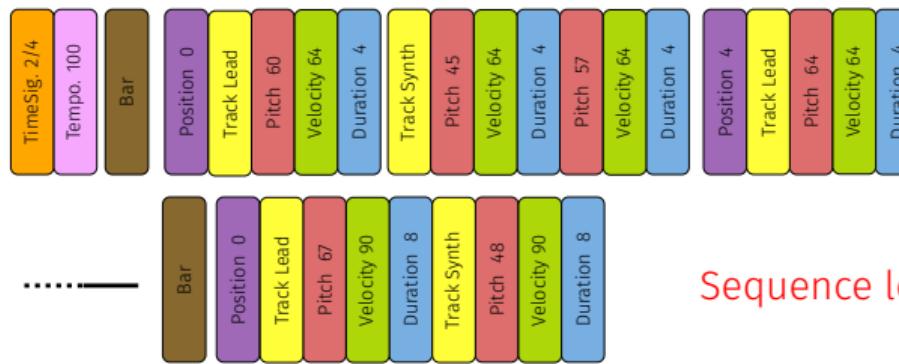
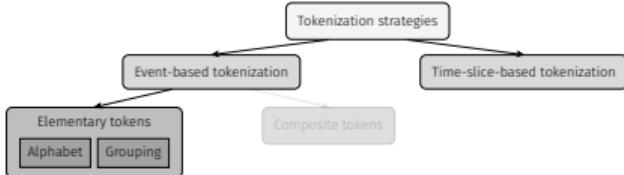


REMI+: von Rutte & al., FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control, ICLR 2023

# How to tokenize music?

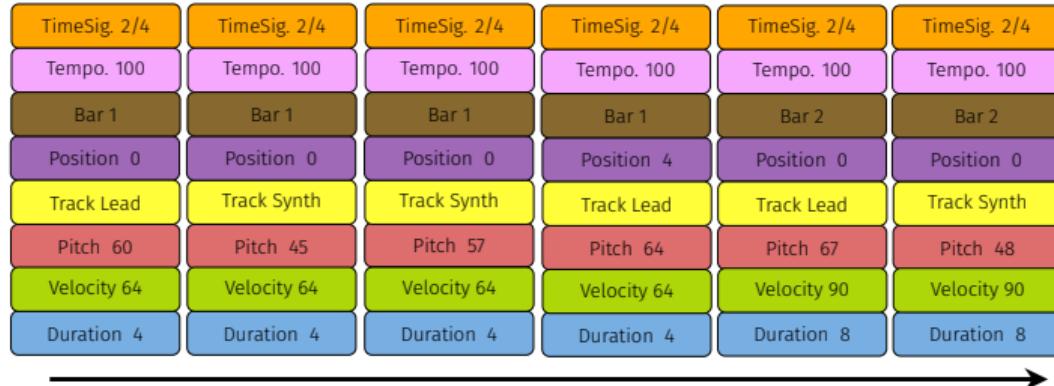
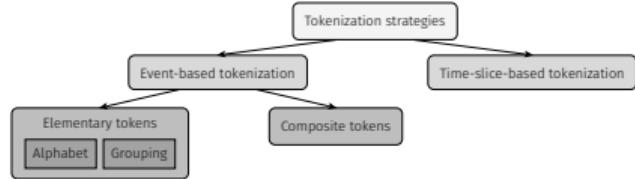


# How to tokenize music?



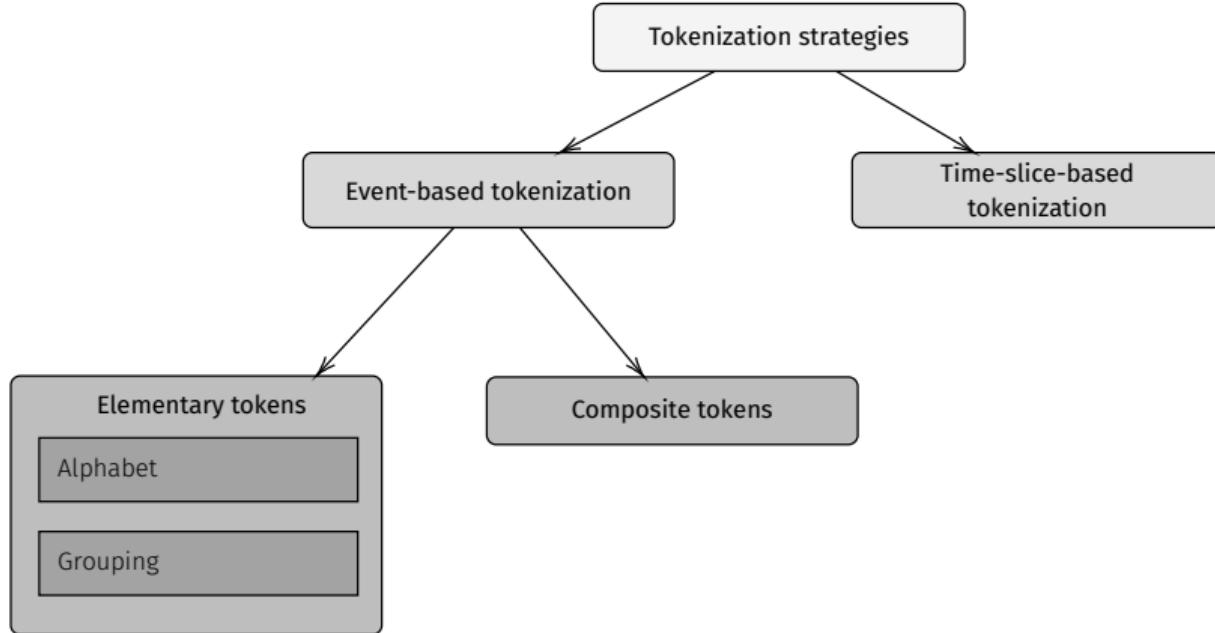
Sequence length?

# How to tokenize music?



Octuple: Zeng & al., *MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training*, ACL 2021

# A taxonomy of tokenization strategies for symbolic music

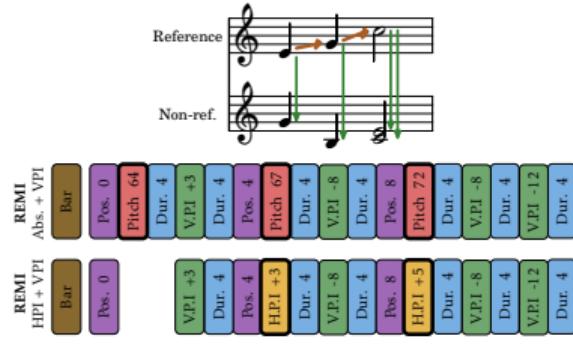


---

Le & al., *Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey*, ACM Computing Surveys 2025. In collaboration with AMAAI (Singapore University of Technology and Design).

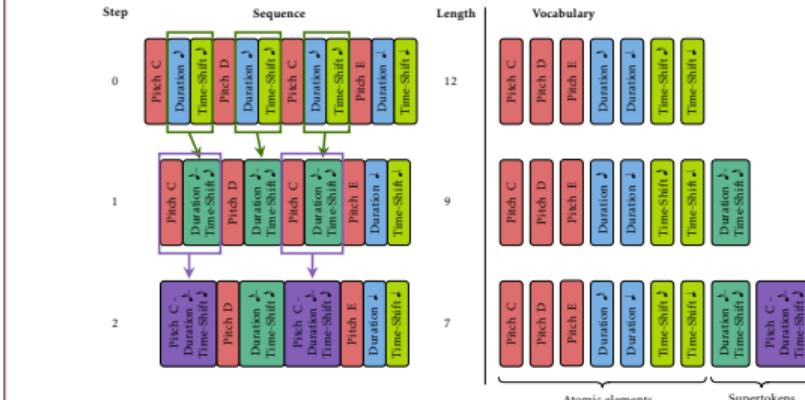
# Technical contributions – Tokenization expressiveness

## Alphabet: Interval-based tokenization



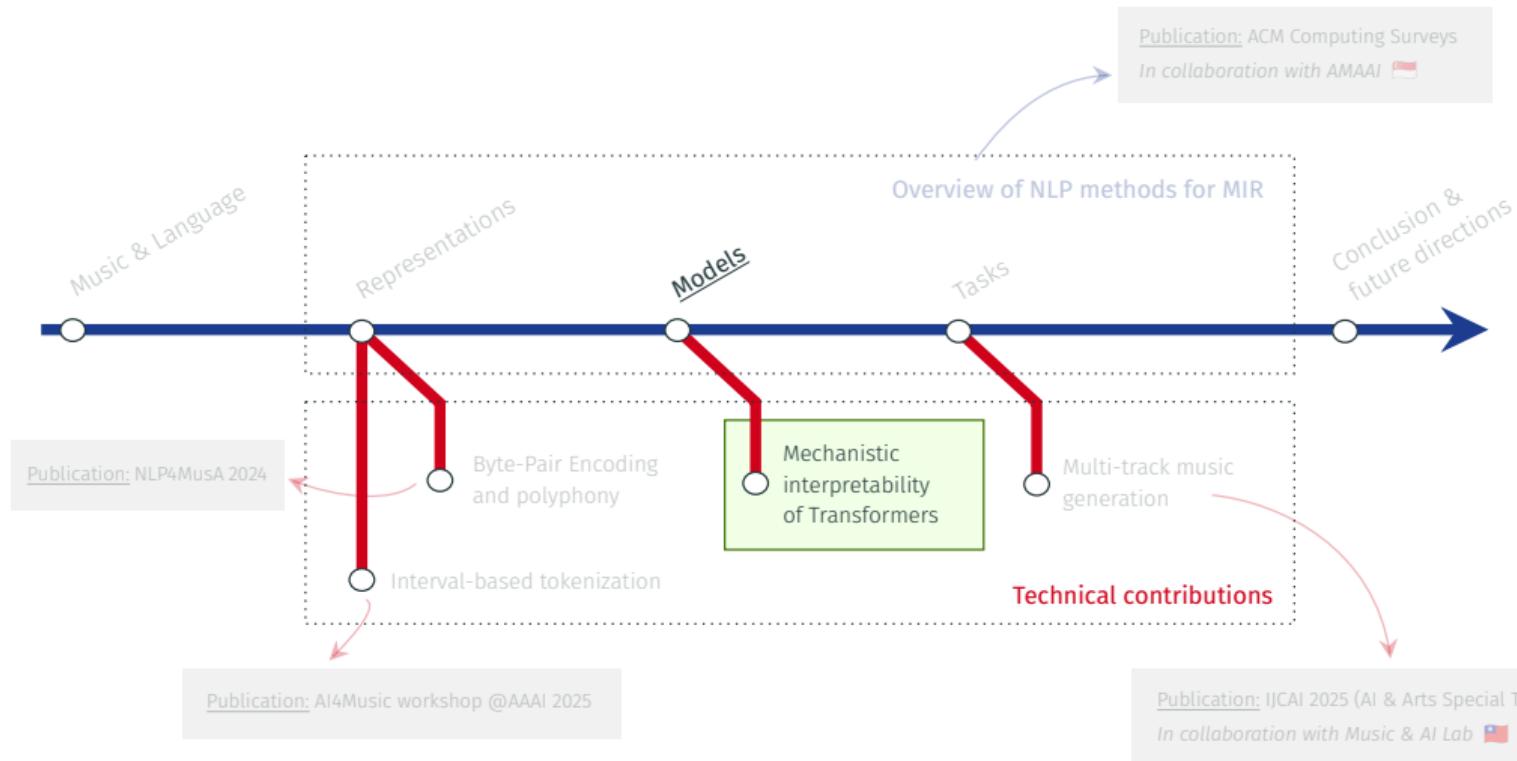
Le & al., Evaluating Interval-based Tokenization for Pitch Representation in Symbolic Music Analysis, AI4Music workshop at AAAI 2025

## Grouping: Byte-pair encoding & polyphony

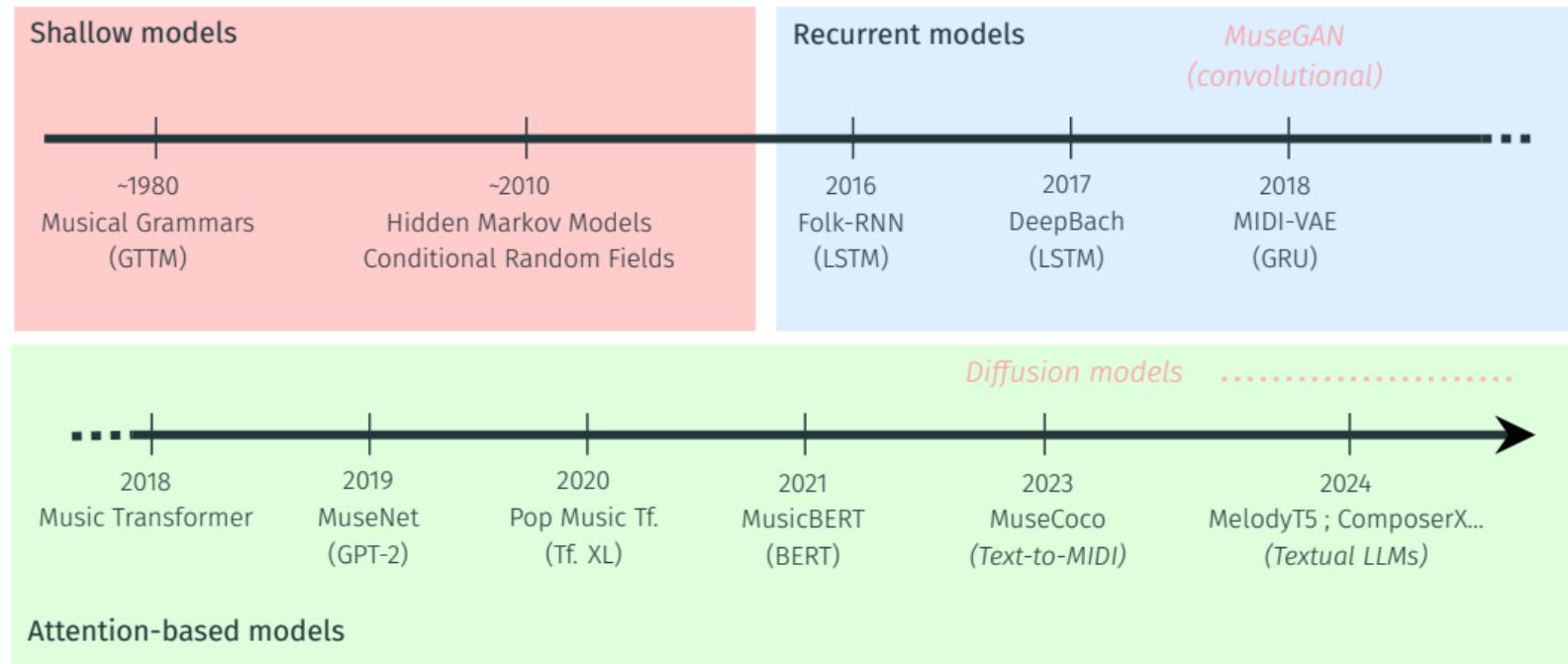


Le & al., Analyzing Byte-Pair Encoding on Monophonic and Polyphonic Symbolic Music: A Focus on Musical Phrase Segmentation, NLP4MusA 2024

# Outline – Models



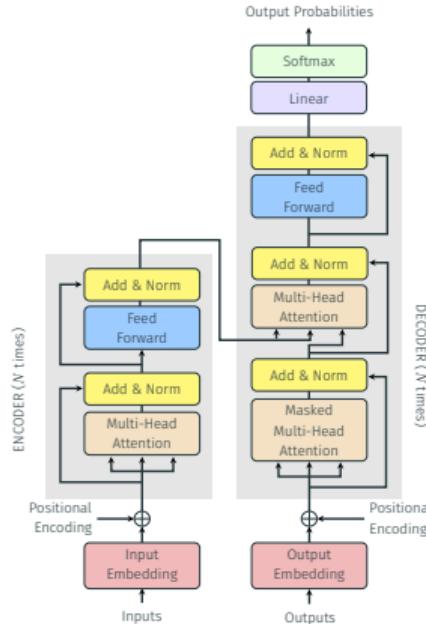
# NLP models for symbolic music processing



# Attention mechanism and Transformers

## Attention mechanism

- Build new representations of an input token sequence
- Weight the importance of each token according to its relevance to the others



Vaswani & al., *Attention is All You Need*, 2017

## Model interpretability (or *explainability*)

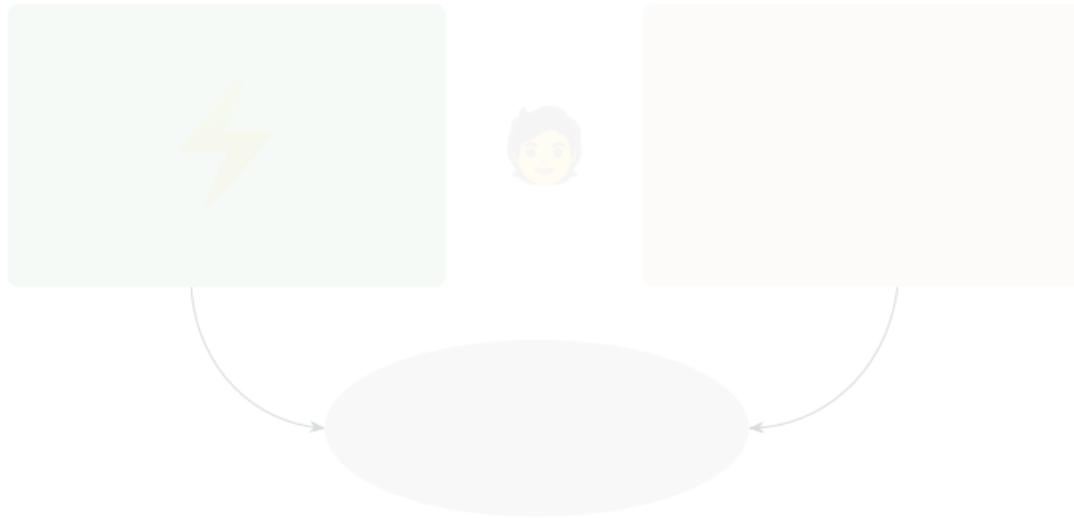
- Make a model's decision process transparent, faithful, and interpretable
- *Human-centered* explanations

In Music Information Retrieval:

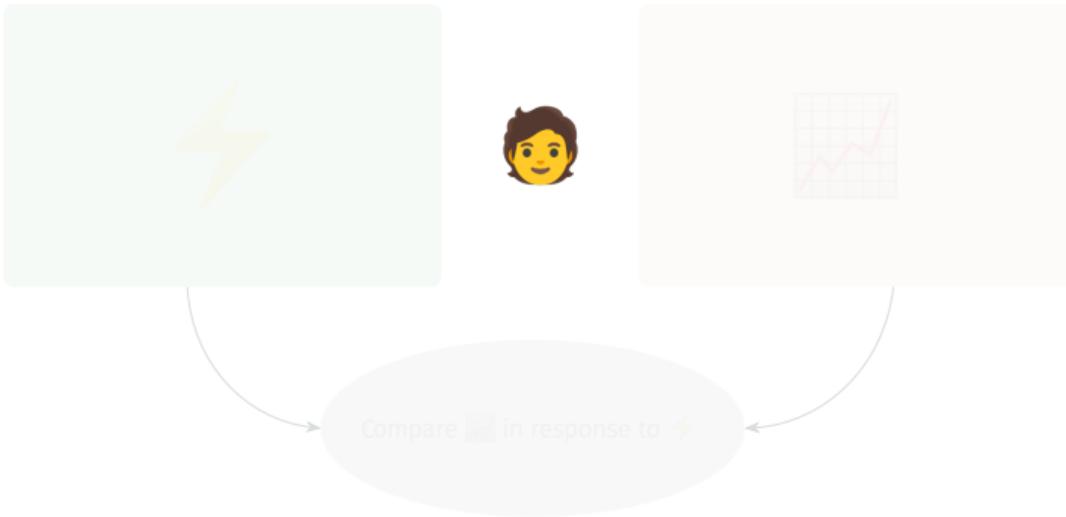
- **Chain-of-thoughts** for ABC Notation-based models [Zhou & al. 2024]
- Attention mechanism **visualization** in music generation [Huang & al. 2018]

⇒ Our approach: **Analyzing** the attention mechanism in the context of functional harmonic analysis

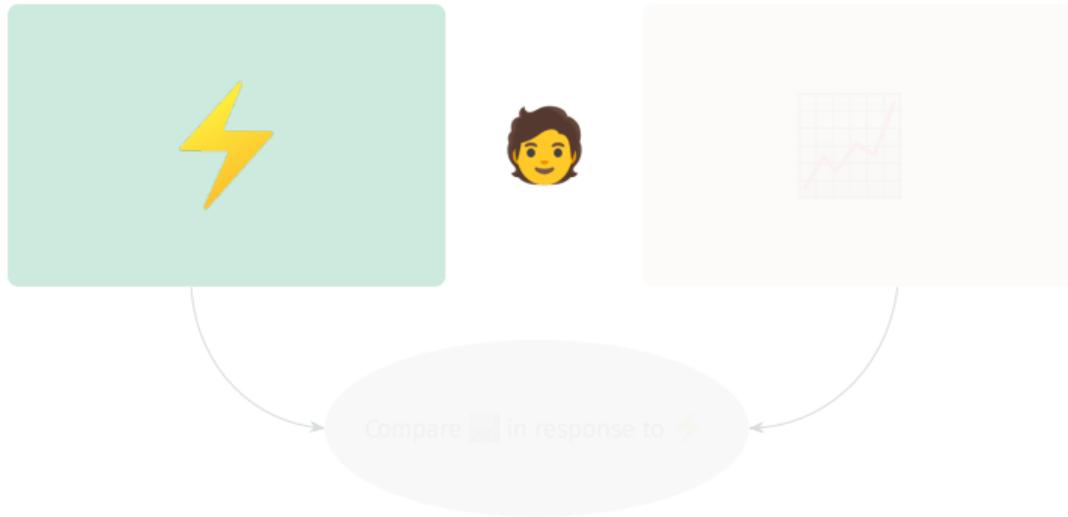
# A methodology to interpret attention with regards to musical characteristics



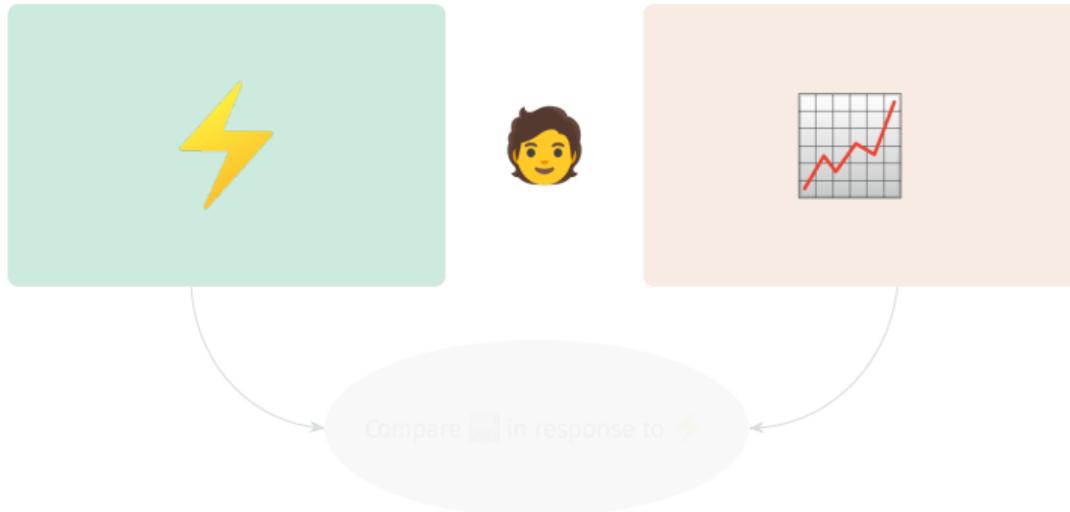
# A methodology to interpret attention with regards to musical characteristics



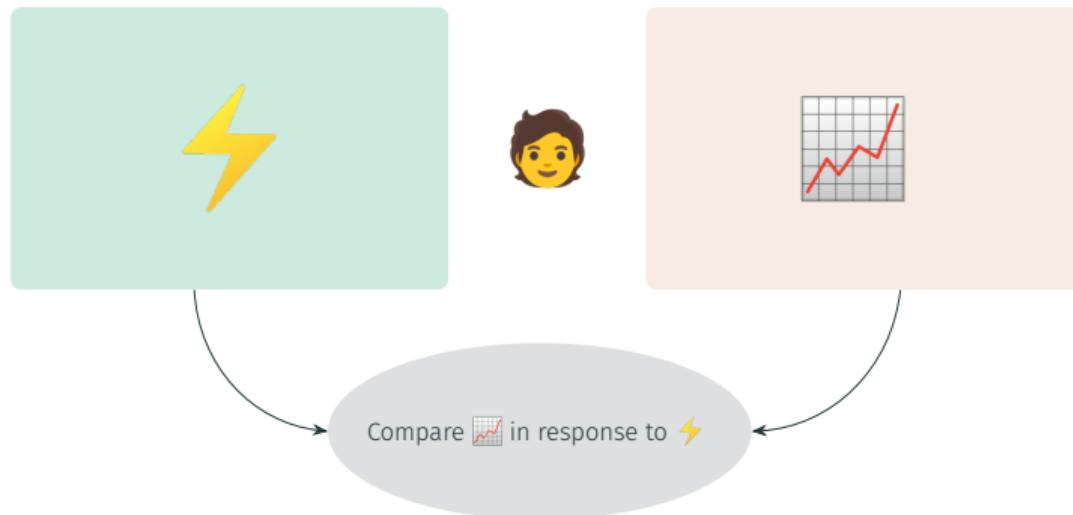
# A methodology to interpret attention with regards to musical characteristics



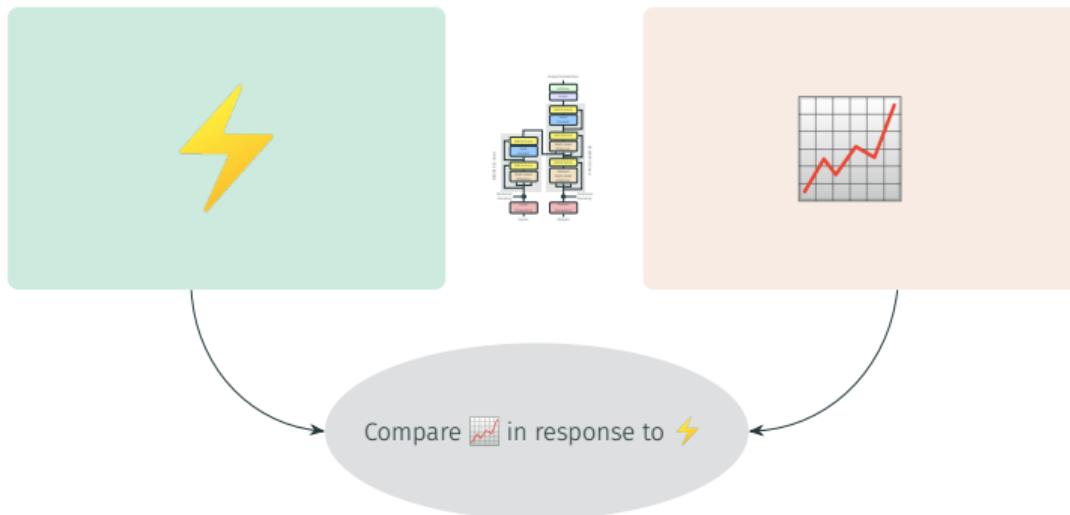
# A methodology to interpret attention with regards to musical characteristics



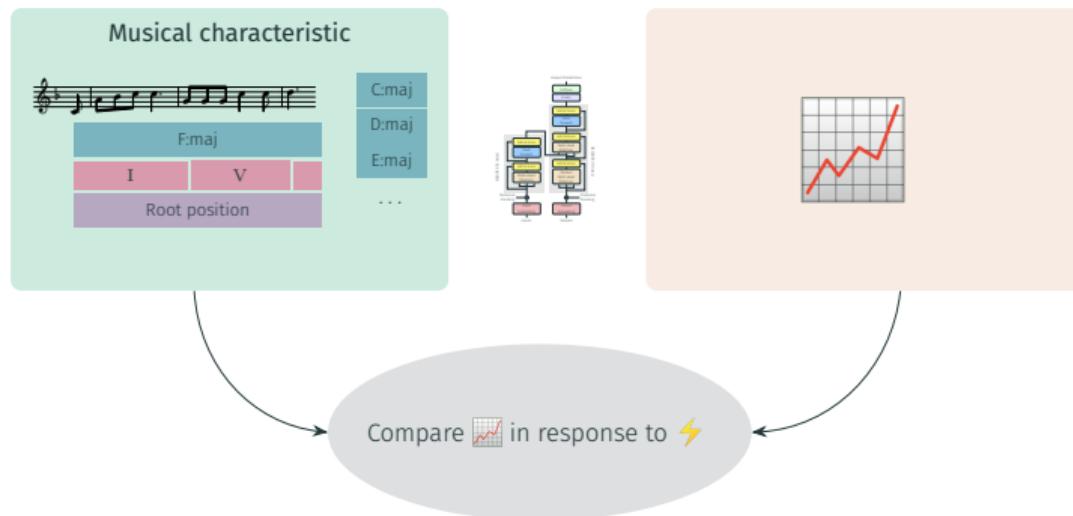
# A methodology to interpret attention with regards to musical characteristics



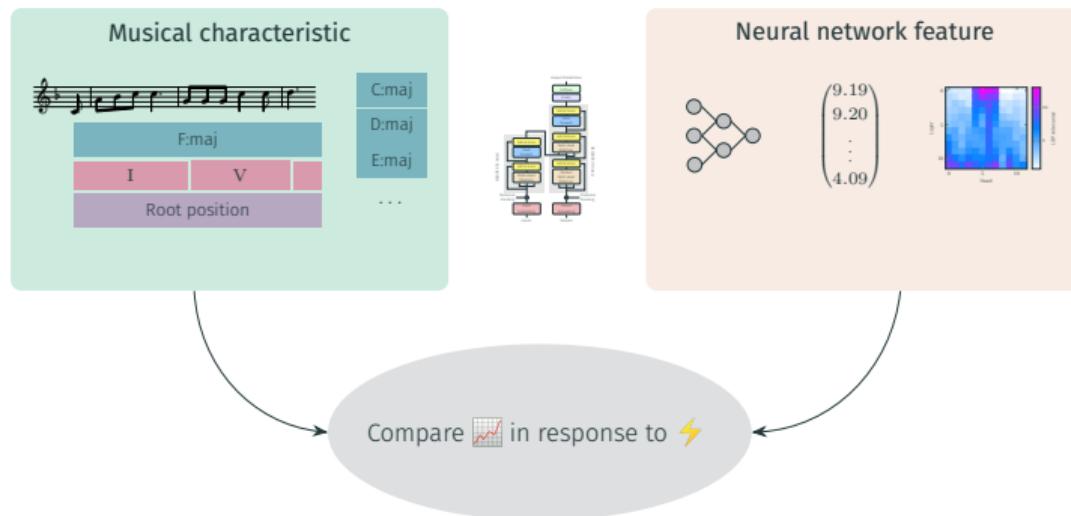
# A methodology to interpret attention with regards to musical characteristics



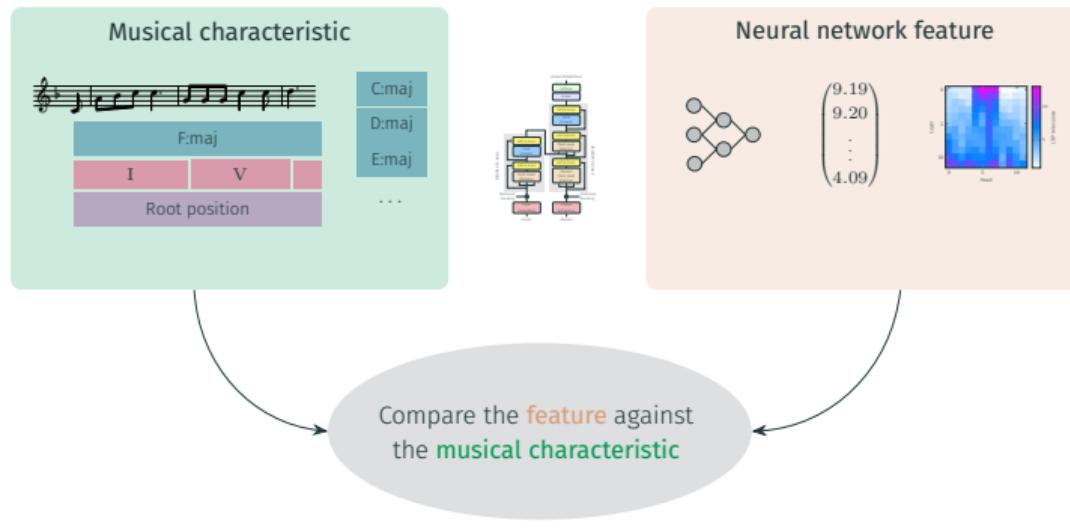
# A methodology to interpret attention with regards to musical characteristics



# A methodology to interpret attention with regards to musical characteristics



# A methodology to interpret attention with regards to musical characteristics



# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang



# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in common time, key signature C major. The top staff shows a melodic line with various note heads and stems. The bottom staff shows a harmonic bass line with eighth-note patterns. A vertical blue bar highlights a section of the top staff between measures 4 and 6. A green bracket labeled "C maj" spans the entire duration of the highlighted section. A green bracket labeled "Local key" points to the bass staff, indicating the harmonic function of the bass line. A green curved arrow at the end of the bass staff indicates a change in harmonic function.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in C major (G clef). The top staff has eighth-note patterns, and the bottom staff has sixteenth-note patterns. A vertical blue bar highlights a section of the music. An orange bracket labeled "Local key" points to the beginning of the piece. Below the blue bar, the letter "G" is shown with an arrow pointing to "V", labeled "Degree". A green bracket covers the end of the piece.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in C major. The top staff shows a melodic line with various note heads and stems. The bottom staff shows a harmonic bass line with eighth-note chords. A vertical blue bar highlights a specific measure. Below the score, three concepts are defined:

- Local key:** Indicated by a green bracket under the first measure.
- Quality:** Indicated by a blue bracket under the blue-highlighted measure, with an orange arrow pointing from the text to the G<sup>7</sup> chord symbol.
- Degree:** Indicated by an orange bracket under the blue-highlighted measure, with an orange arrow pointing from the text to the V symbol.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in C major. The top staff shows a melodic line with various note heads and stems. The bottom staff shows a harmonic bass line with eighth-note patterns. A vertical blue bar highlights a specific measure. Below the score, four concepts are defined:

- Local key:** C maj
- Quality:** G<sup>7</sup>
- Degree:** V
- Inversion:** 7 or 6 or 5 or 4 or 2

Annotations include:

- A green bracket under the first measure is labeled "Local key".
- An orange dashed arrow points from "Quality" to the G<sup>7</sup> chord.
- An orange dashed arrow points from "Degree" to the V (VII) degree.
- A pink dashed arrow points from "Inversion" to the Roman numerals below.
- A green curved bracket spans the measures from the local key annotation to the inversion annotation.
- A green bracket under the last measure groups the inversion options.
- A trill symbol is shown above the final measure.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in common time, key C major. The top staff shows a melodic line with eighth and sixteenth notes, and the bottom staff shows a harmonic bass line with eighth notes. A vertical blue bar highlights a specific chordal moment. Below the score, four concepts are defined:

- Local key:** C maj
- Quality:**  $G^7/D$
- Degree:**  $V_3^4$
- Inversion:** 7 or 6 or 5 or 4 or 3 or 2

Annotations show arrows pointing from the highlighted chord to the corresponding labels. A green bracket groups the Local key and Quality, and another green bracket groups Degree and Inversion. A large green curved arrow points from the Quality/Degree group towards the Inversion group.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score for two staves in common time, key of C major. The top staff shows a melodic line with various note heads and stems. The bottom staff shows a harmonic bass line with quarter notes. A blue rectangular box highlights a section of the top staff between measures 5 and 6. An orange curved arrow points from the bass note D at the beginning of this section to the top staff. Below this, a pink curved arrow points from the bass note G at the start of the section to the top staff. A green bracket below the bass staff indicates the section from measure 5 to measure 6. Below the top staff, the text "C maj" is written. In the center of the highlighted section, the harmonic analysis is shown as  $G^7/D \rightsquigarrow V_3^4$ . To the right of this, several Roman numerals are listed:  $7$  or  $\frac{6}{5}$  or  $\frac{4}{3}$  or  $\frac{4}{2}$ . A green bracket and a green curved arrow are positioned to the right of the analysis, spanning from the end of the highlighted section to the end of the score.

# Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

A musical score in C major (two staves, treble and bass) with various harmonic annotations. The top staff has a blue shaded region over the second measure, with orange arrows pointing from the bass staff below it. Below the bass staff, a green bracket spans the first two measures, labeled "C maj". Below the blue-shaded area, the annotation "G<sup>7</sup>/D ~> V<sub>3</sub><sup>4</sup>" is shown. To the right, another green bracket spans the last three measures, with the Roman numerals "7 or 6 or 4 or 2" written below it. The top staff concludes with a trill symbol.

≈ “attention arrows” in text?

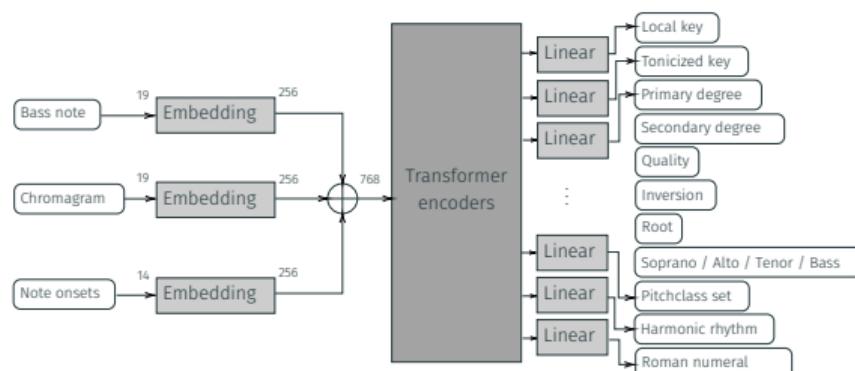


# Automatic Functional Harmonic Analysis

## Representation

- Time-slice-based tokenization: quantization at the ♩

## Model architecture



### Transformer encoders

- Stacked layers of multi-head self-attention<sup>1</sup>
  - Bi-directional
  - 12 layers
  - 12 attention heads per layer

<sup>1</sup>Vaswani & al., *Attention is All You Need*, 2017

# Automatic Functional Harmonic Analysis

## Training

- End-to-end training
- 596 pieces, 28M tokens [Nápoles López & al., 2021]
- Era: baroque → early romantic
- Instruments: solo piano → string quartet

## Performances (accuracy)

Model	Key	Degree	Quality	Inversion	Root	Global	# params
AugmentedNet v1.9.1	82.2	67.0	79.7	78.8	83.0	–	<b>105k</b>
RNBert	<b>82.5</b>	<b>85.9</b>	86.5	87.2	–	–	109M
Our model (mono-task)	79.5	81.7	90.0	89.7	92.0	87.5	–
Our model (multi-task)	81.7	84.3	<b>90.9</b>	<b>90.3</b>	<b>92.8</b>	<b>88.6</b>	94.9M

# Understanding inner mechanisms in local key detection

modulation

John Rutter, *For the beauty of the earth.* (Performance: Taipei Male Choir)

A musical score excerpt for three voices (SATB) in common time. The key signature changes from two flats (C minor) to one sharp (G major). The vocal parts are: Soprano (S), Alto (A), and Bass (B). The piano accompaniment is also shown. The score consists of four staves of music, with the vocal parts on the top three staves and the piano on the bottom staff. The piano part includes bass notes and harmonic indications.

A musical score excerpt for three voices (SATB) in common time. The key signature changes from one sharp (G major) to no sharps or flats (D major). The vocal parts are: Soprano (S), Alto (A), and Bass (B). The piano accompaniment is also shown. The score consists of four staves of music, with the vocal parts on the top three staves and the piano on the bottom staff. The piano part includes bass notes and harmonic indications.

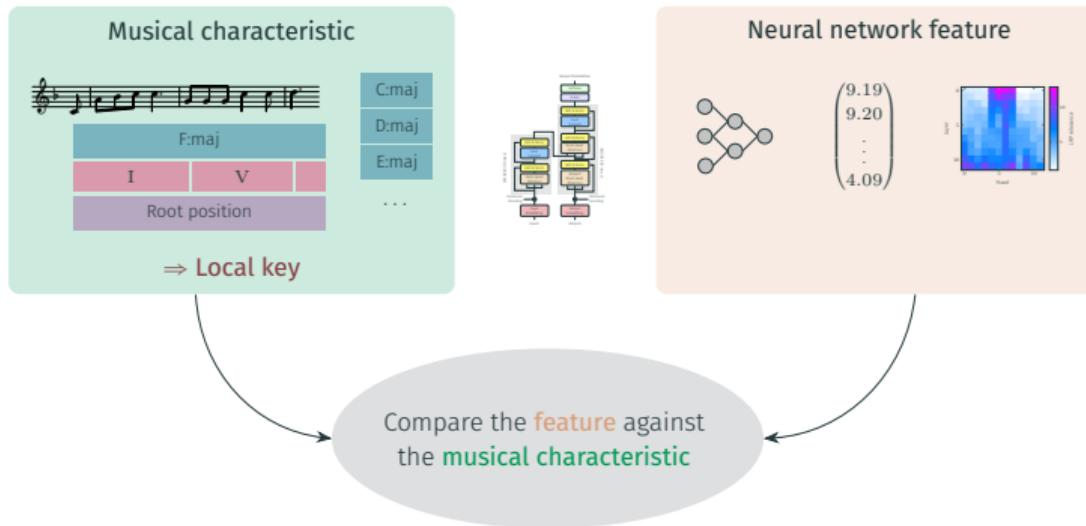
# Understanding inner mechanisms in local key detection



Do the attention heads relevant for the annotation remain the same before and after the key change?

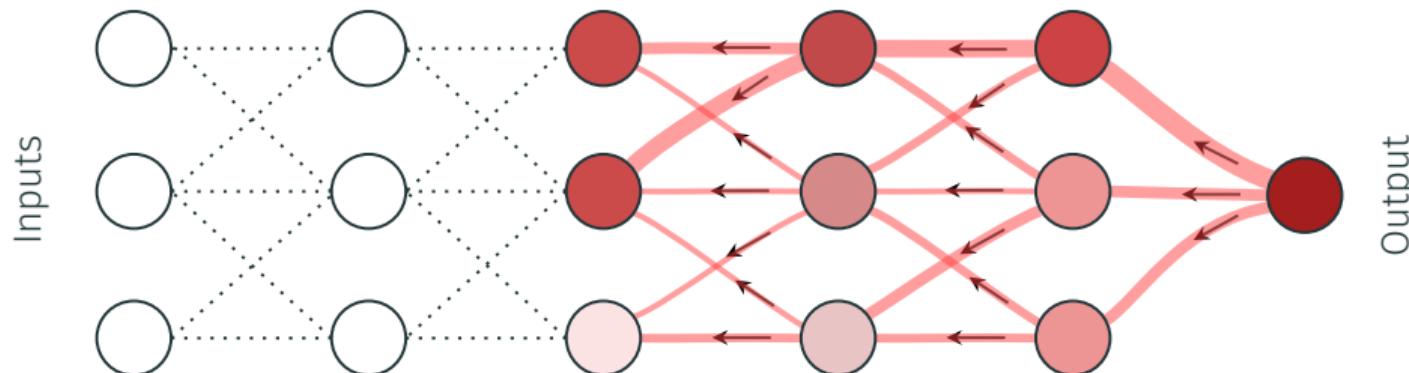
In the following, we will consider the model *only* trained on local key detection.

# A methodology to interpret attention with regards to musical characteristics



# Layer-wise relevance propagation (LRP)

1. Consider a trained model on a 1-class classification task
2. Perform a forward pass and keep track of logits
3. LRP backward pass:



---

Montavon & al., *Layer-Wise Relevance Propagation: An Overview*, 2019

# Layer-wise relevance propagation (LRP)

(In NLP) Model: Llama-3.2 (next word prediction)

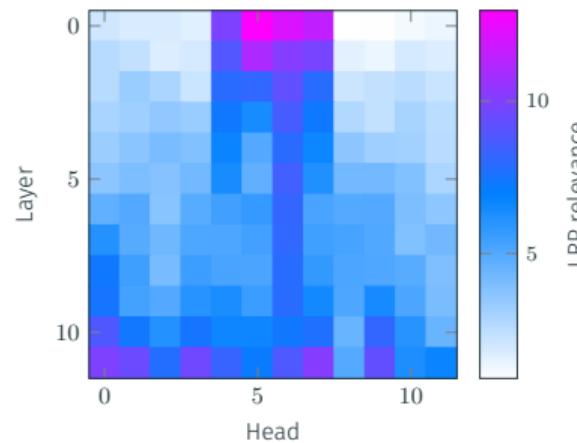
<|BOS|> NASA's Artemis program aims to return humans to the Moon by 2026, with the Artemis II mission to orbit around the moon. The Artemis III mission explores further goals by landing a spacecraft near the lunar south pole. They will be the first humans to walk on the moon since the Apollo missions. What is the purpose of Artemis II? Its goal is to

<|BOS|> NASA's Artemis program aims to return humans to the Moon by 2026, with the Artemis II mission to orbit around the moon. The Artemis III mission explores further goals by landing a spacecraft near the lunar south pole. They will be the first humans to walk on the moon since the Apollo missions. What is the purpose of Artemis III? Its goal is to

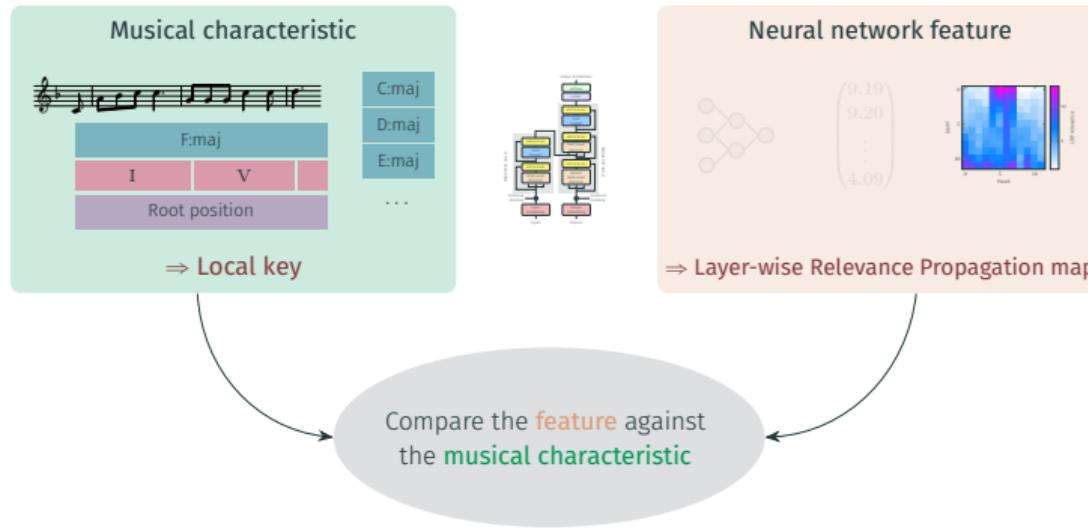
# Layer-wise relevance propagation (LRP)

*What if back-propagation is not performed up to the initial tokens?*

**LRP map:** contribution of each attention head to a prediction



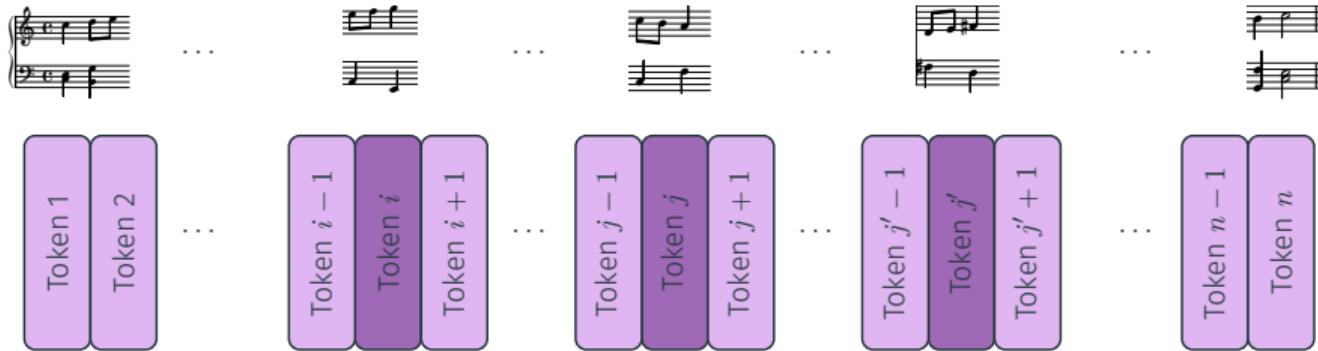
# A methodology to interpret attention with regards to musical characteristics



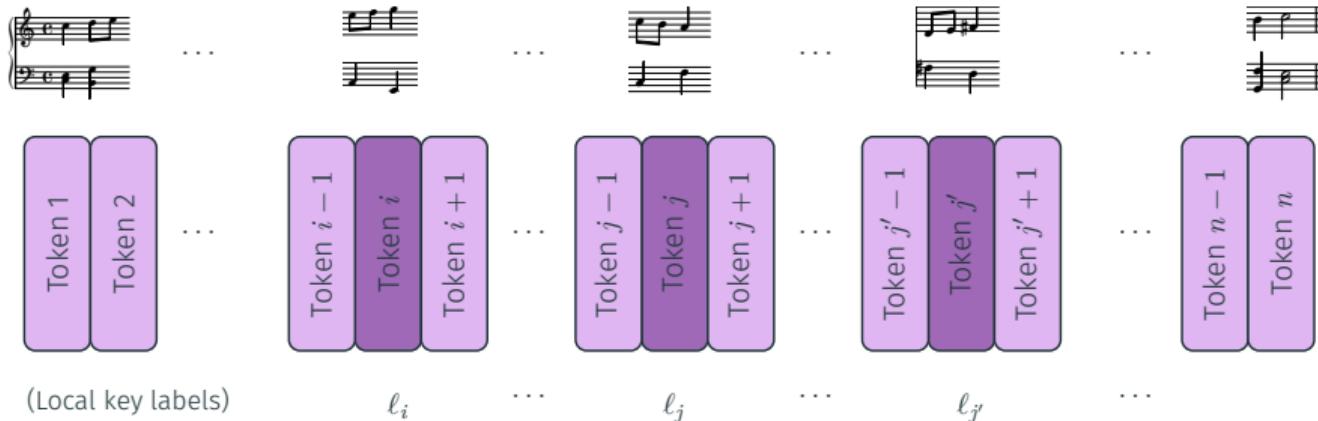
## Attention head relevance for local key detection



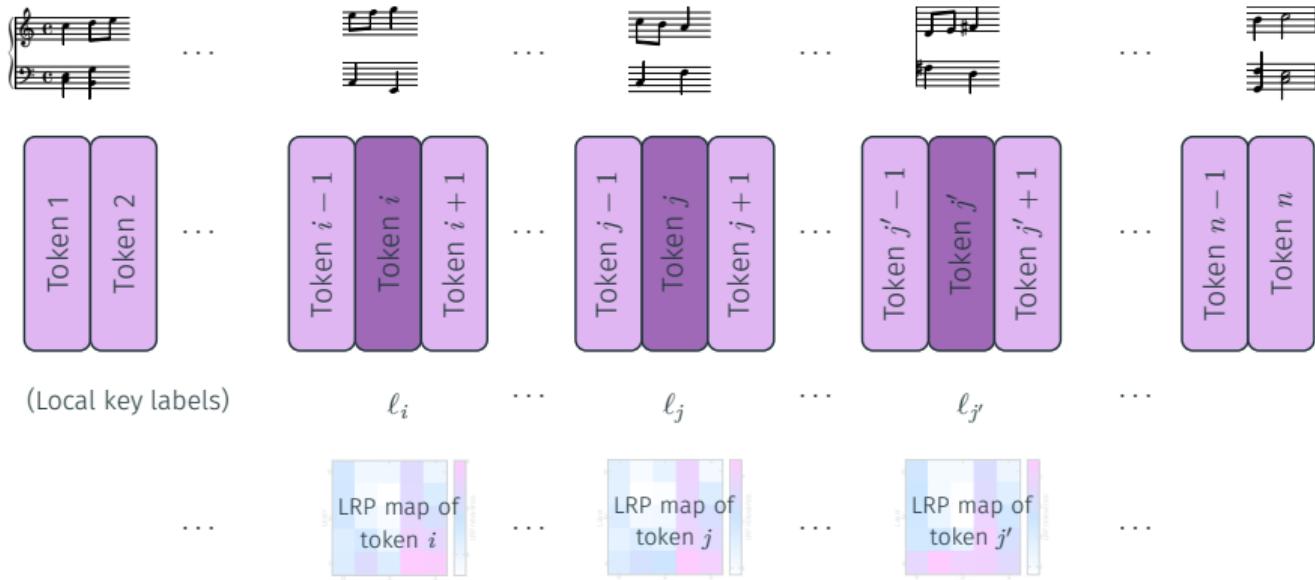
# Attention head relevance for local key detection



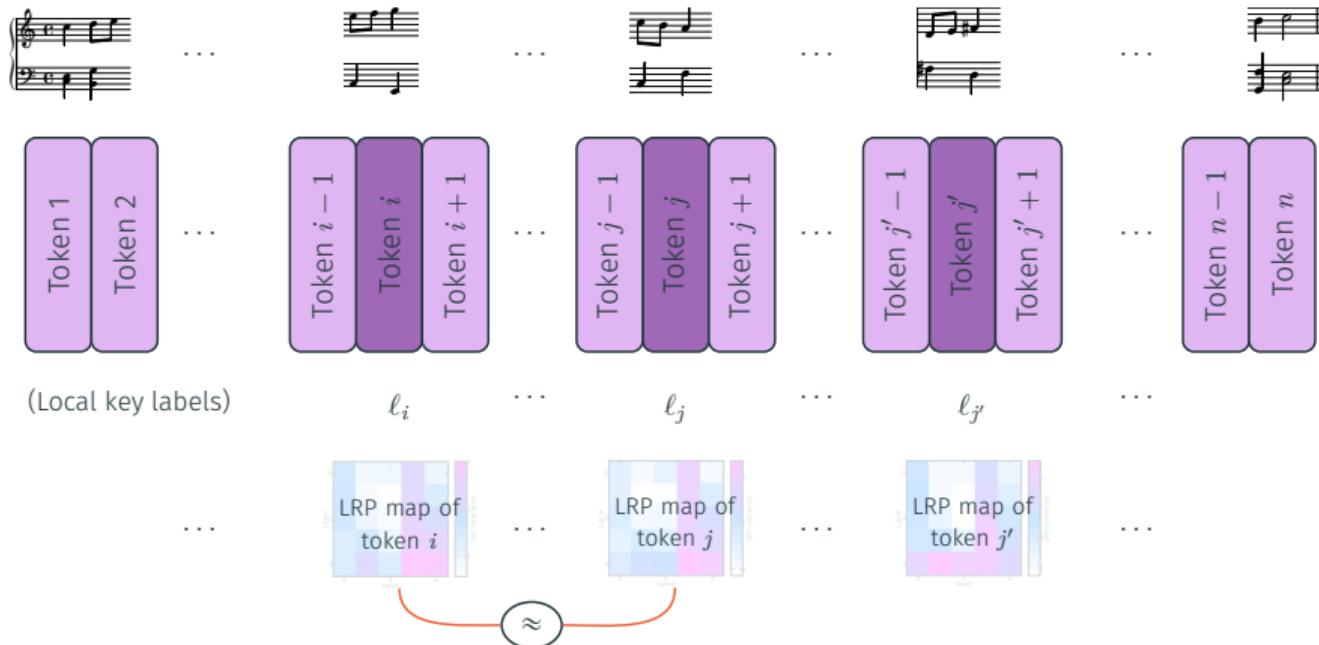
## Attention head relevance for local key detection



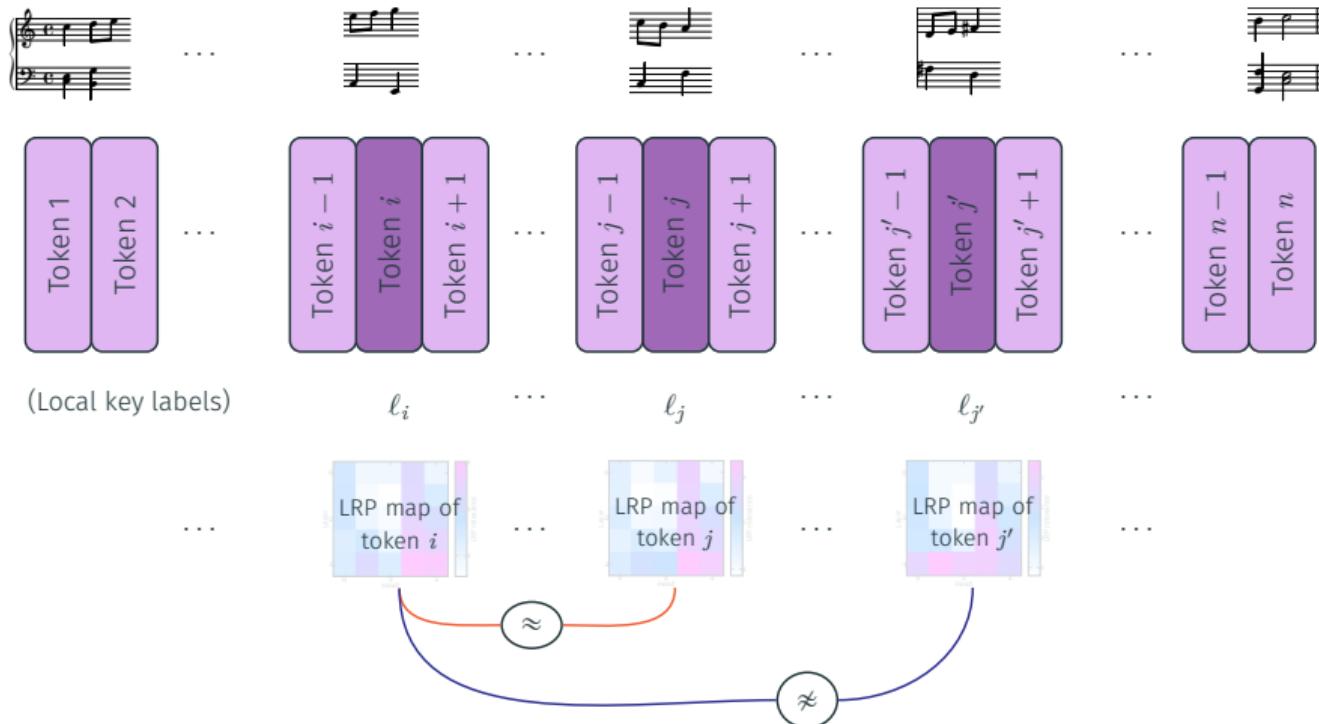
# Attention head relevance for local key detection



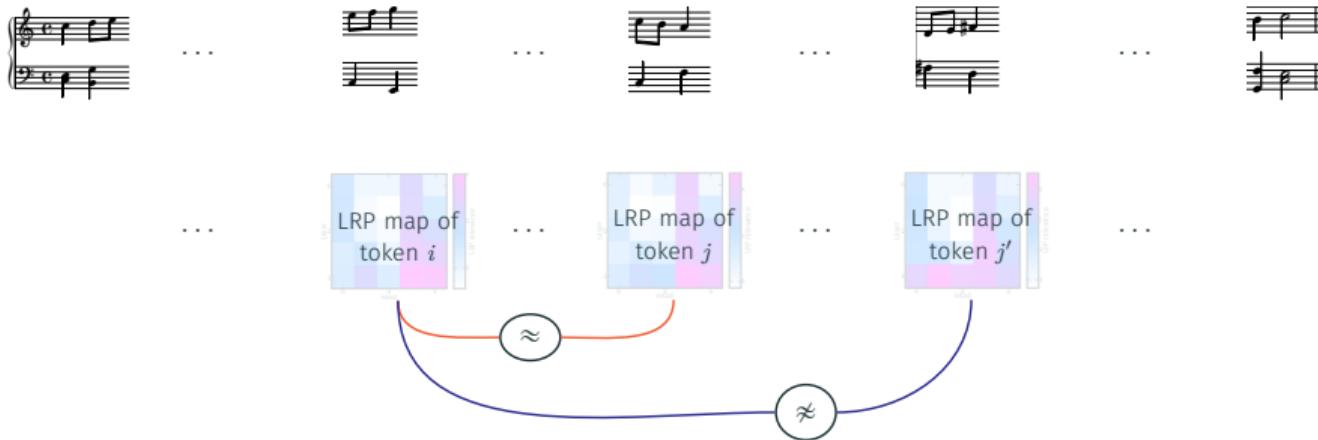
# Attention head relevance for local key detection



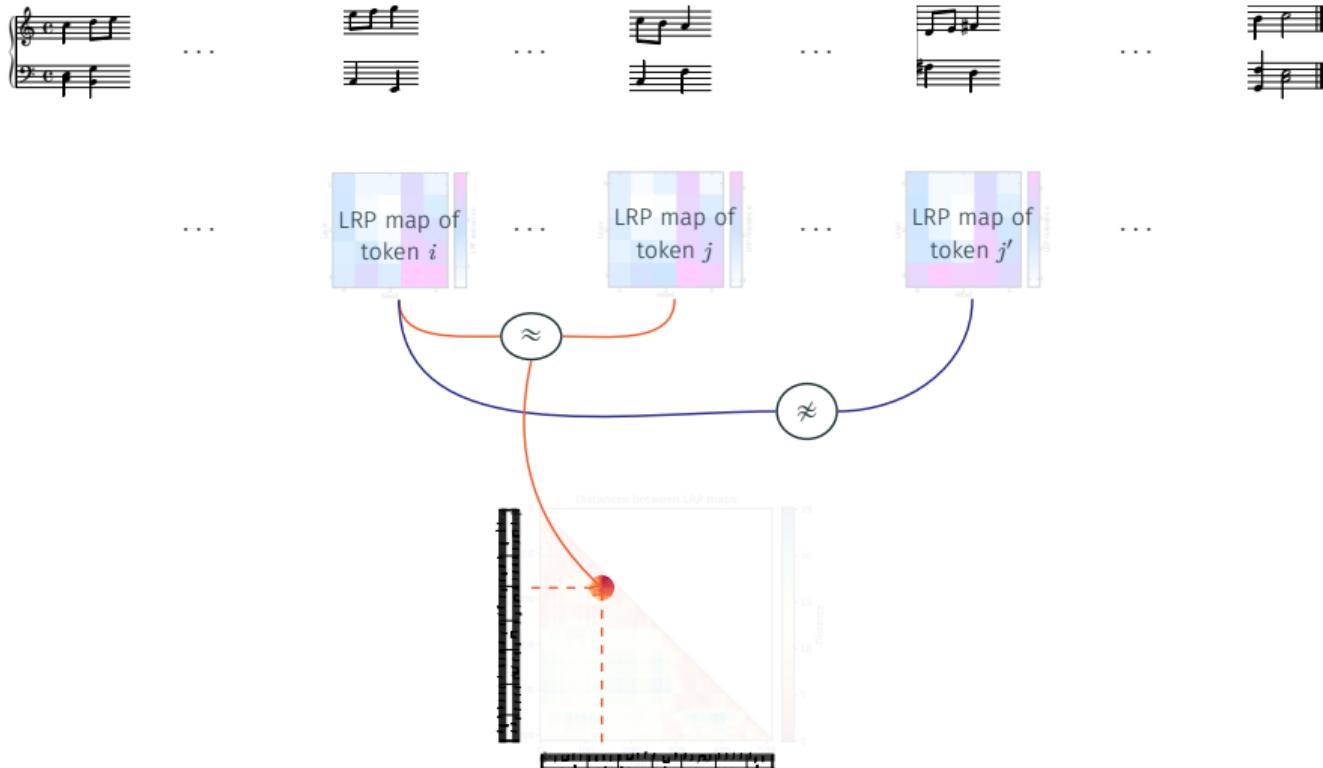
# Attention head relevance for local key detection



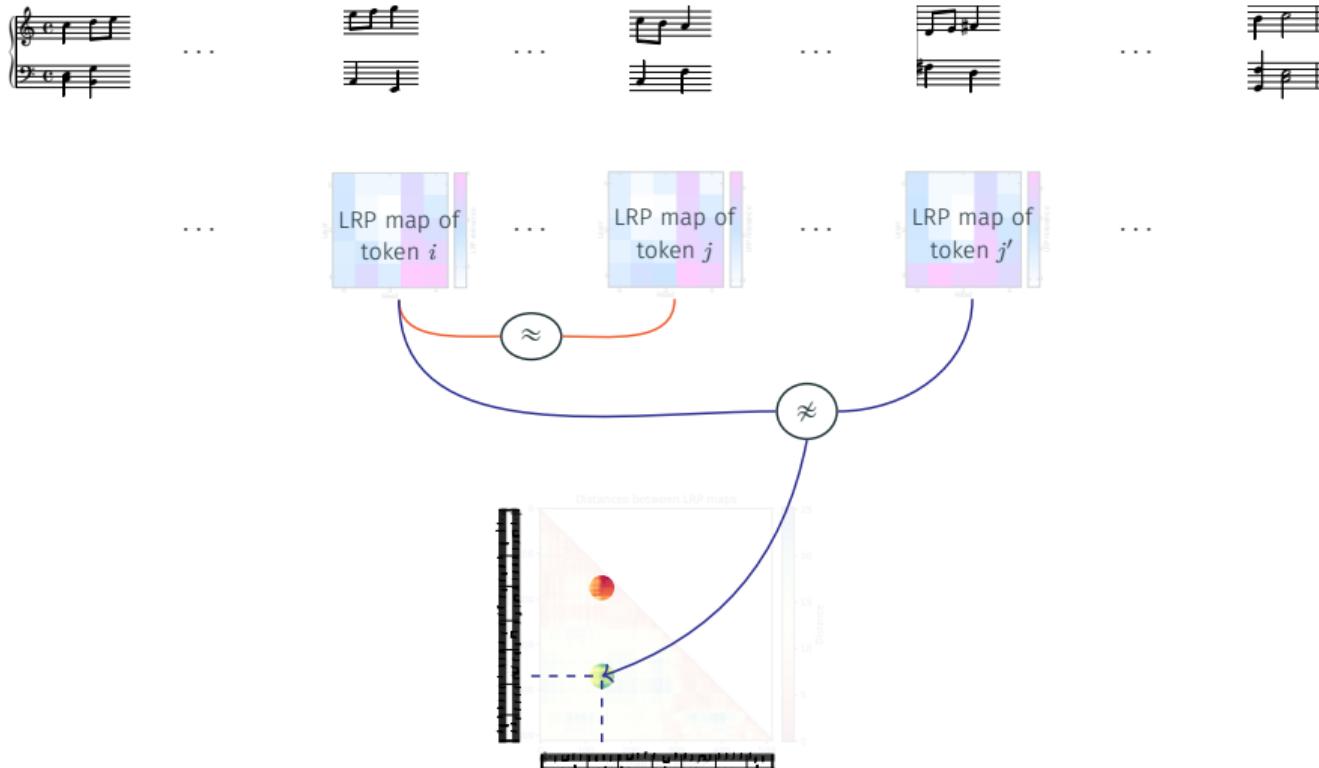
# Attention head relevance for local key detection



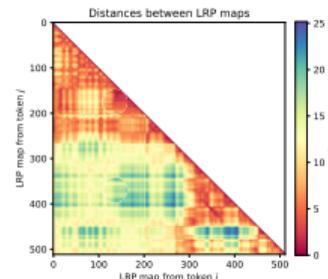
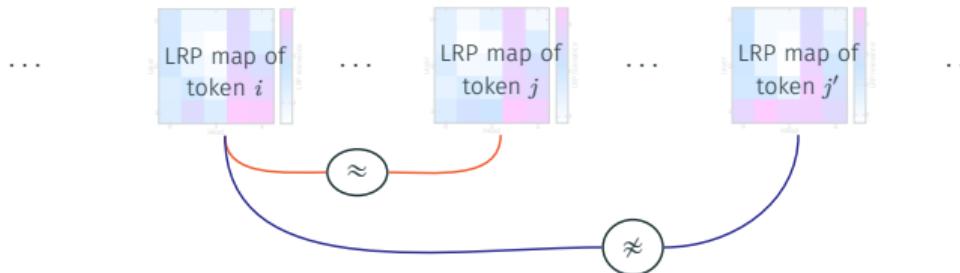
# Attention head relevance for local key detection



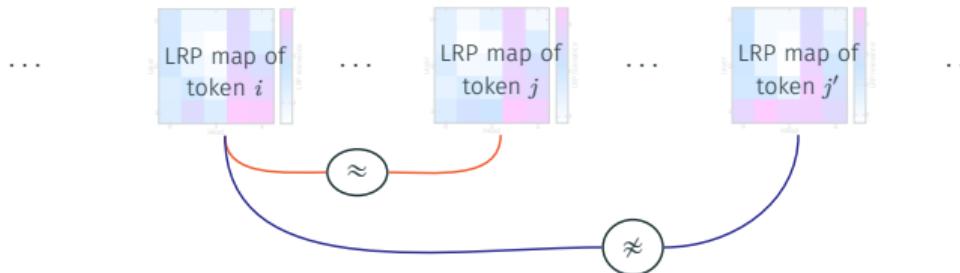
# Attention head relevance for local key detection



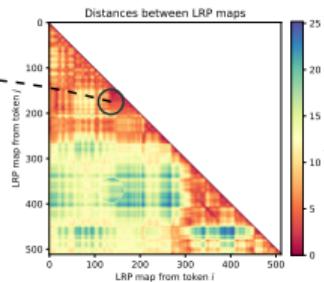
# Attention head relevance for local key detection



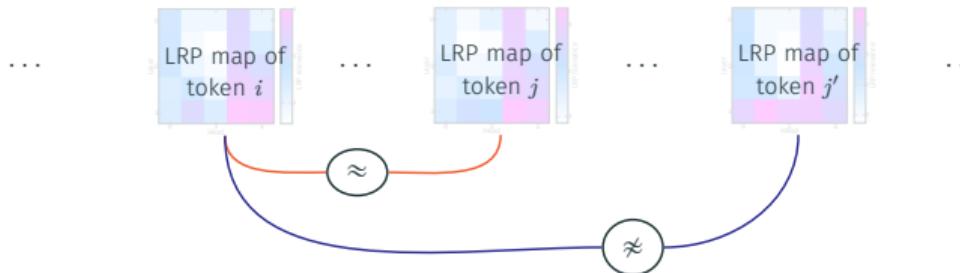
# Attention head relevance for local key detection



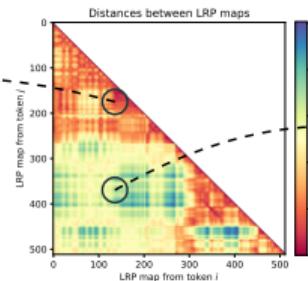
"Similar attention heads  
were used to label these  
two tokens."



# Attention head relevance for local key detection



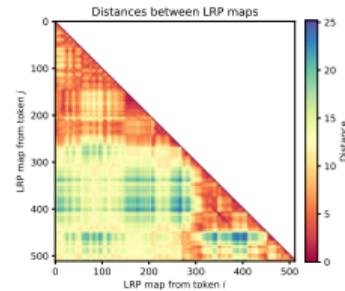
"Similar attention heads  
were used to label these  
two tokens."



"Different attention heads  
were used to label these  
two tokens."

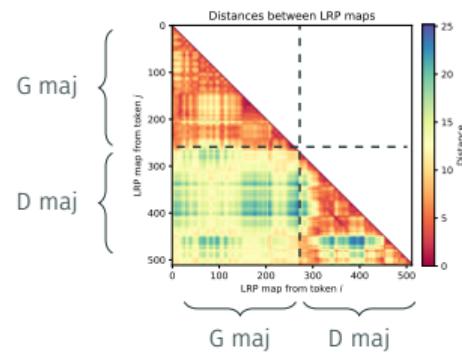
# Attention head relevance for local key detection

## Relating distances to labels



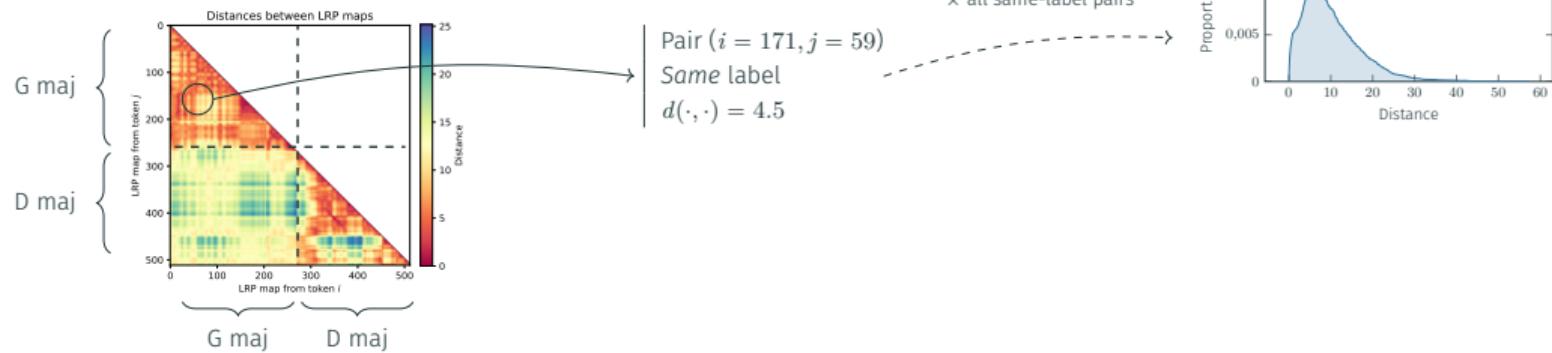
# Attention head relevance for local key detection

## Relating distances to labels



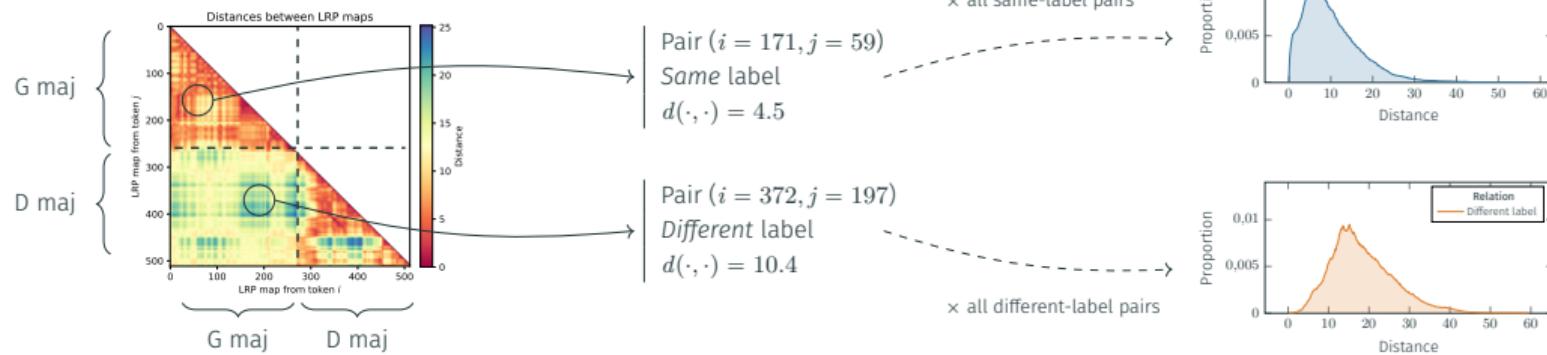
# Attention head relevance for local key detection

## Relating distances to labels



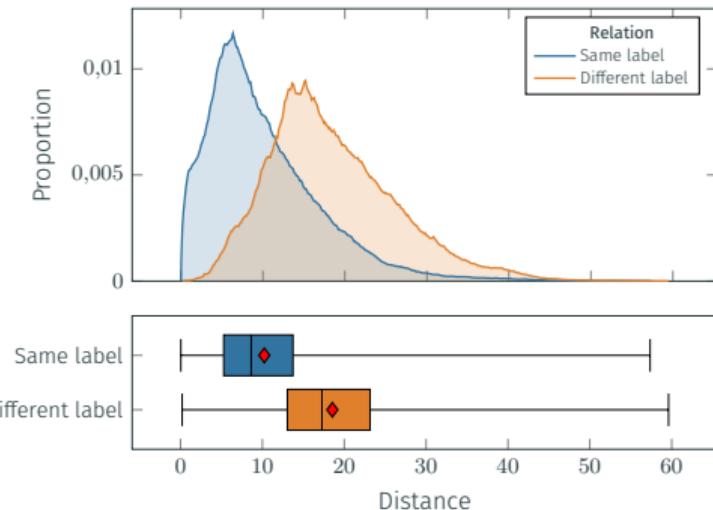
# Attention head relevance for local key detection

## Relating distances to labels



⇒ Comparing two distributions

# Attention head relevance for local key detection



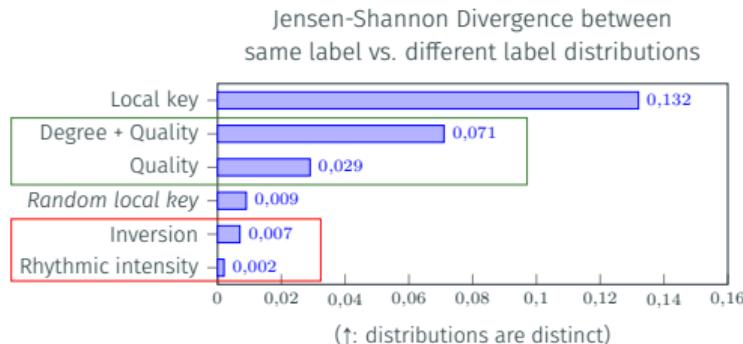
- The distributions between same vs. different labels are distinct ( $\blacktriangle | \blacktriangle$ )  
⇒ The relevant attention heads are different when changing keys.
- Distances for same label pairs are lower than different label pairs ( $\blacklozenge$ )

The attention heads relevant for the annotation of the *same local key* are *more similar* than for the annotation of different local keys.

# Attention head relevance for local key detection

## Comparing with other features

What if we consider other labels beyond local key?



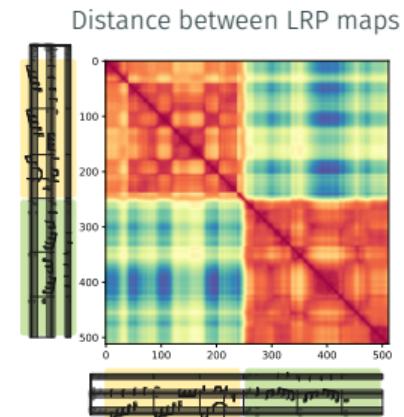
- Other harmonic features are also present in the model.
- Less relevant features for local key detection show less signal.

A model *only* trained on local key detection can capture  
*human-relevant* harmonic features it has not been trained on.

# Attention head relevance for local key detection

What leads the model to “think” that a modulation has occurred?

Original modulation ( $B\flat$  maj  $\rightarrow$  G maj)

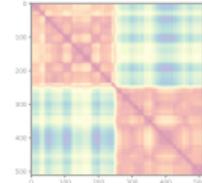


---

John Rutter, *For the beauty of the earth.*

# Attention head relevance for local key detection

Initial piece

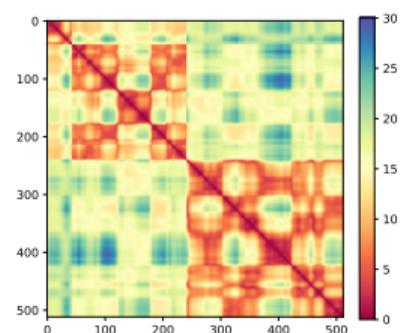


Keep the key signature

( $\Rightarrow$  Modulation  
 $B\flat$  maj  $\rightarrow$  G min)

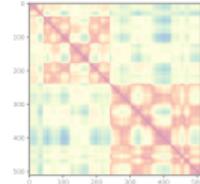


keep-key-signature



# Attention head relevance for local key detection

Last step:  
Key signature kept



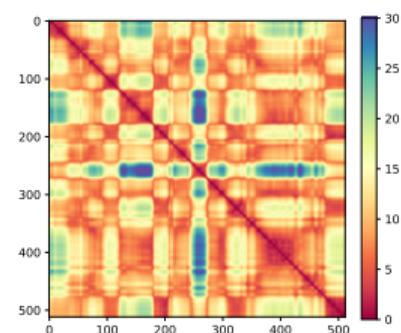
---

Transpose the modulated part to the initial key

(Bb maj)

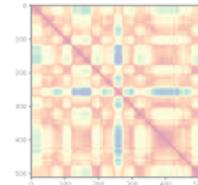


transpose



# Attention head relevance for local key detection

Last step:  
Key signature kept  
+ Transposition

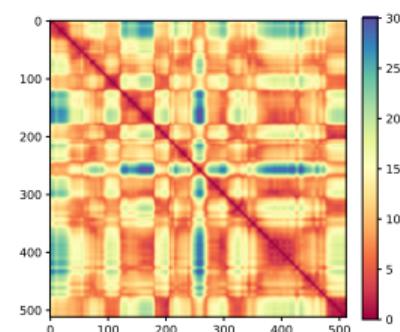


(F♯ in G maj)

Remove the leading tone of the initial modulation



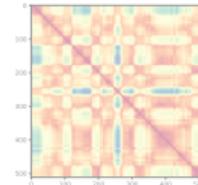
no-fis



# Attention head relevance for local key detection

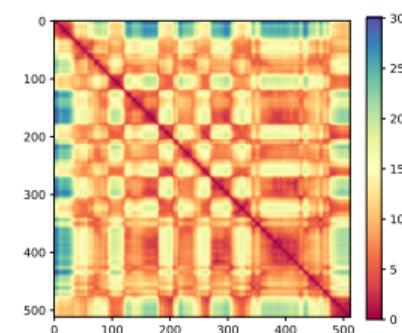
Last step:

- Key signature kept
- + Transposition
- + No leading tone

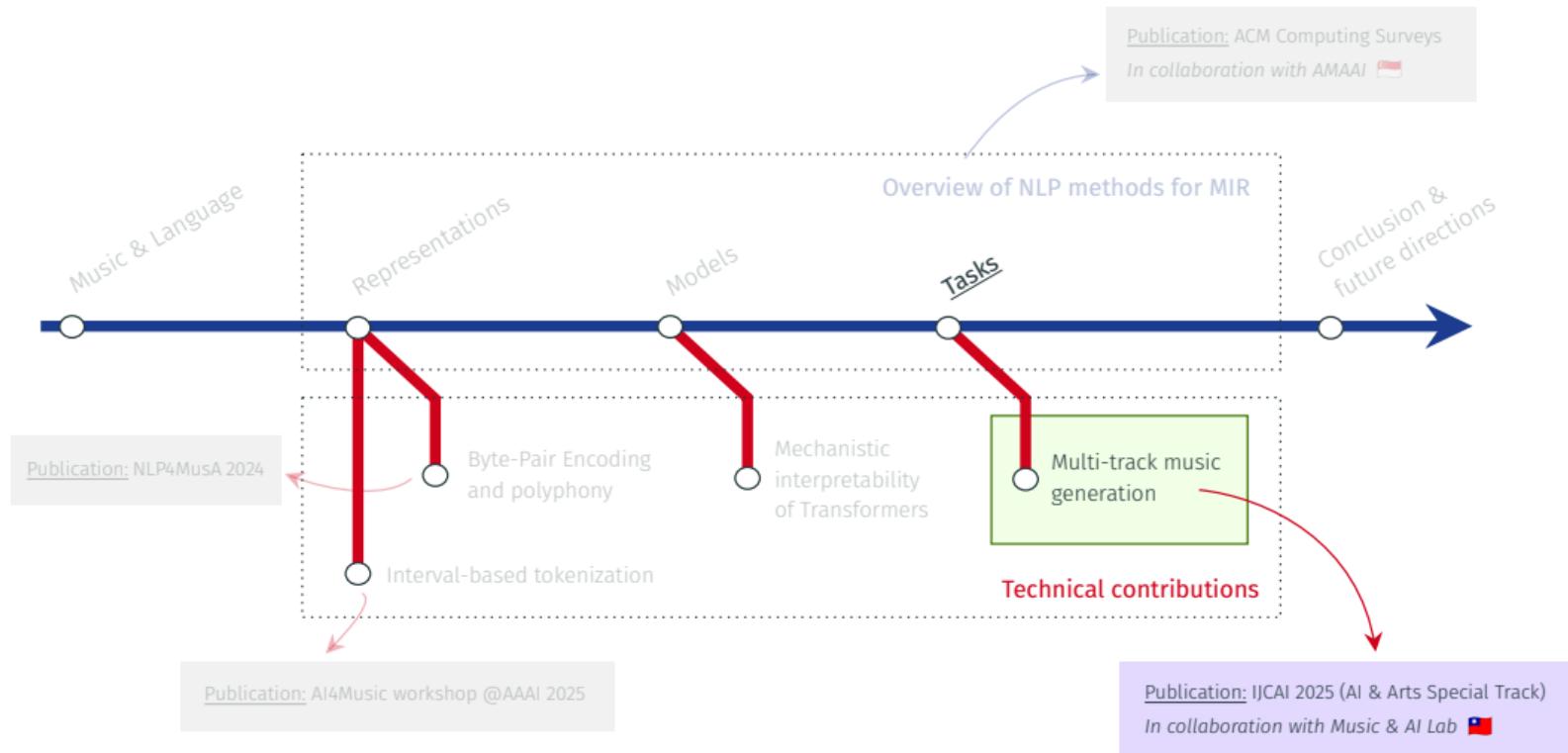


Re-write the initial modulation preparation

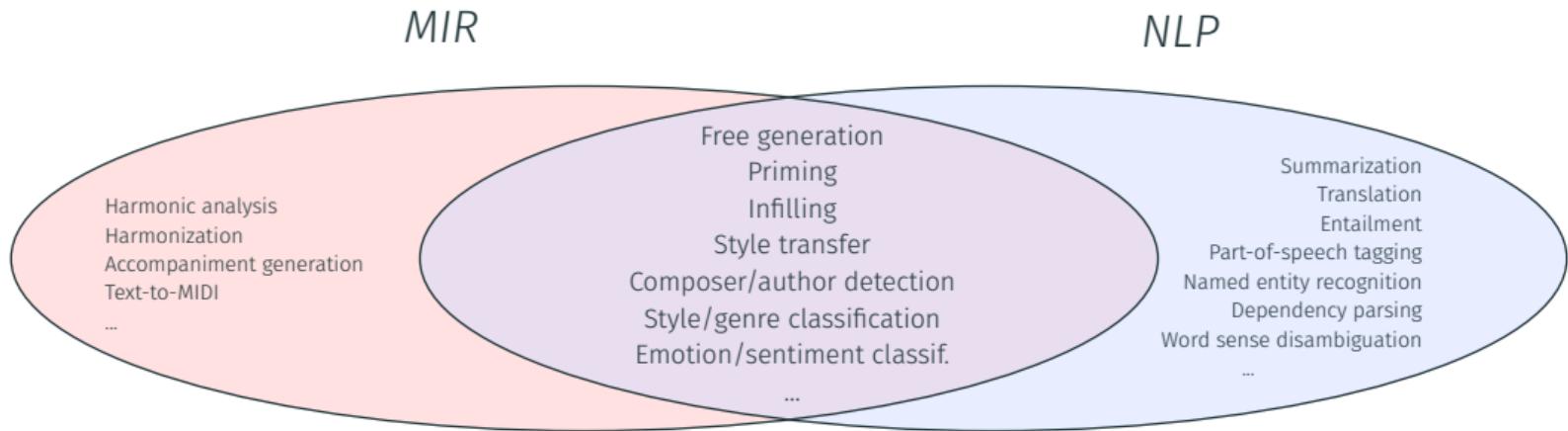
$\text{ii} \rightarrow \text{V} \rightarrow \text{I}$   
 $G \text{ major } (\text{Am} \rightarrow \text{D} \rightarrow \text{G})$   
to  $B\flat \text{ major } (\text{Cm} \rightarrow \text{F} \rightarrow \text{B}\flat)$



# Outline – Tasks



# Tasks in NLP and MIR



# Re-orchestration: instrumentation, texture, melody

## Instrumentation

### Symphonic orchestra

Musical score for the Symphonic orchestra section of Beethoven's Symphony No. 5, Mvnt. 1, Bars 1-5. The score includes parts for Flauti, Oboi, Clarinetti in B., Fagotti, Corni in Es., Trombe in C., Timpani in C.G., Violino I, Violino II, Viola, Violoncello, and Basso. The tempo is Allegro con brio, dynamic ff.

### String orchestra

Musical score for the String orchestra section of Beethoven's Symphony No. 5, Mvnt. 1, Bars 1-5. The score includes parts for Violin I, Violin II, Viola, Cello, and Contrabass. The tempo is Allegro con brio, dynamic ff.

### Piano - violin duet

Musical score for the Piano - violin duet section of Beethoven's Symphony No. 5, Mvnt. 1, Bars 1-5. The score includes parts for Violino and Pianoforte. The tempo is Allegro con brio, dynamic ff.

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1, Bars 1-5.

### Piano solo

Musical score for the Piano solo section of Beethoven's Symphony No. 5, Mvnt. 1, Bars 1-5. The score includes parts for Instruments (Violins and Clarinet), Manuals, and Pedal. The tempo is Allegro con brio, dynamic ff.

### Organ solo

Musical score for the Organ solo section of Beethoven's Symphony No. 5, Mvnt. 1, Bars 1-5. The score includes parts for Manuals and Pedal. The tempo is Allegro con brio, dynamic ff.

Orchestral texture

Melodic fidelity

# Re-orchestration: instrumentation, texture, melody

## Orchestral texture



dvorak9-10

A musical score excerpt in 2/4 time with a key signature of one sharp. It shows multiple staves: Horn and Trombone (top), Bassoon (second), Trombones (third), Trombone and Cello (fourth), and Trombones and Cello (fifth). The Trombones play eighth-note patterns throughout the measures. The Cello joins in on the fourth measure.

dvorak9-279

A. Dvořák, *Symphony No. 9 "From the New World"*, Mvnt. 4.

Le & al., *A Corpus Describing Orchestral Texture in First Movements of Classical and Early-Romantic Symphonies*, DLfM 2022

# Re-orchestration: instrumentation, texture, melody

Instrumentation

Orchestral texture

## Melodic fidelity

朋友 (Wakin Chau)



いつも何度も (from Spirited Away - Joe Hisaishi)



Ritournelle

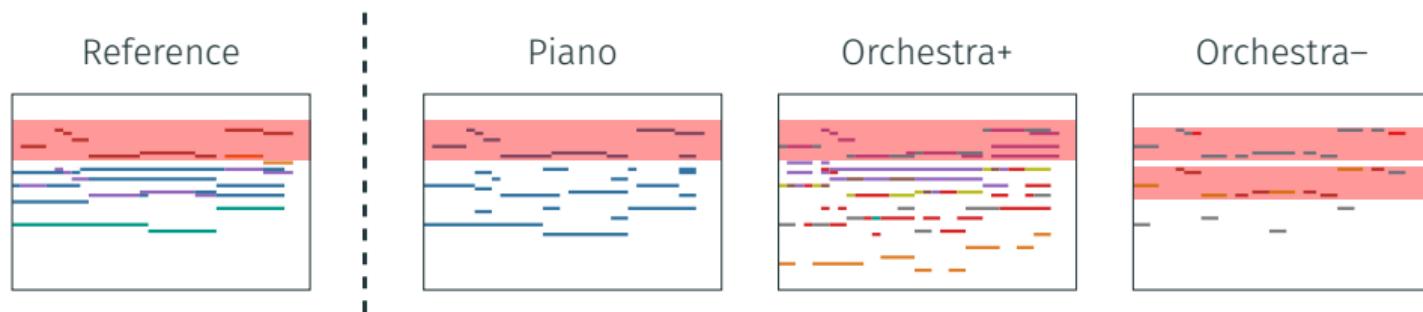


你被寫在我的歌裡 (sodagreen)



("Pachelbel canon"-like chord progression: I – V – vi – iii – IV – I – (ii) – V)

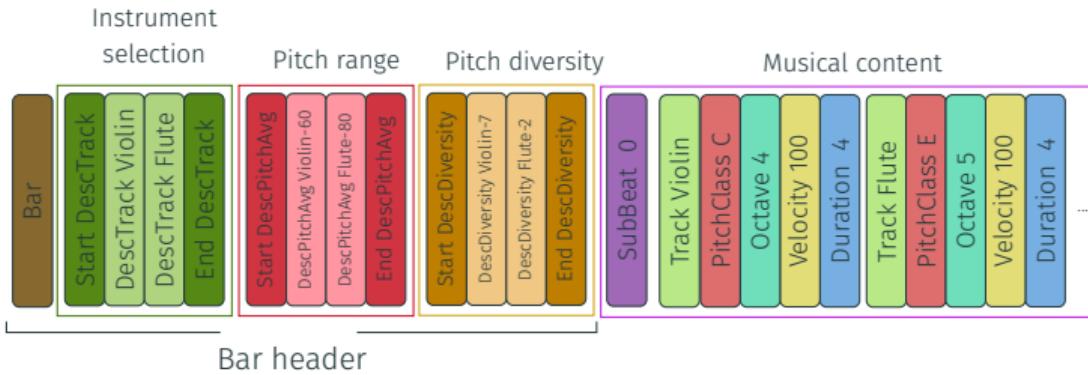
## METEOR: Melody-aware Texture-controllable Symbolic Music Reorchestration via Transformer VAE



---

Le & Yang, METEOR: Melody-aware Texture-controllable Symbolic Orchestral Music Generation via Transformer VAE, IJCAI 2025 (AI & Arts Special Track). In collaboration with Music & AI Lab (National Taiwan University).

# Tokenization & Textural controls



## Bar-level

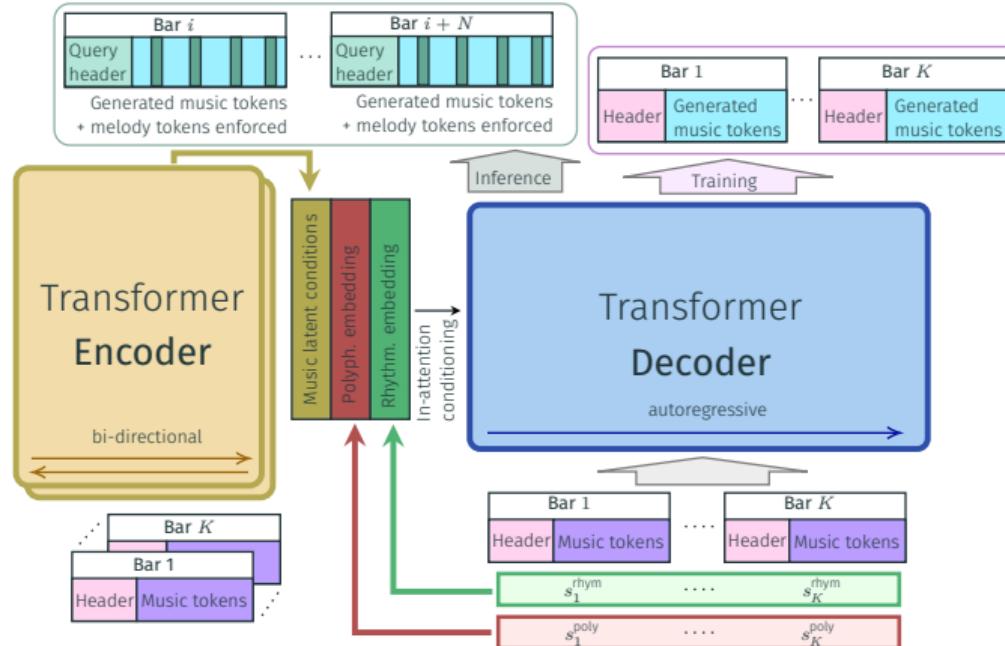
- Rhythmic intensity
- Polyphony

## Bar+track-level

- Average pitch
- Pitch diversity

⇒ Derived statistically from the dataset

# Model architecture



Wu & Yang, *MuseMorphose: Full-Song and Fine-Grained Piano Music Style Transfer with One Transformer VAE*, 2022

## Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference

# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)

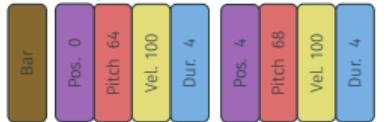


2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference

# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



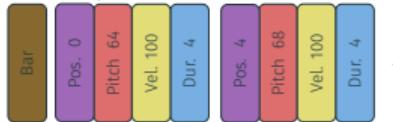
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



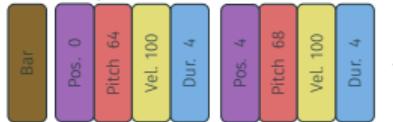
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



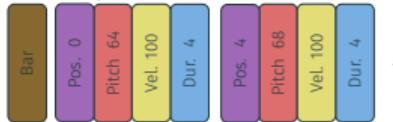
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



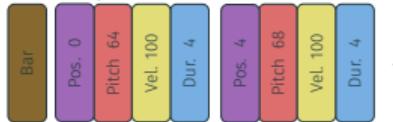
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



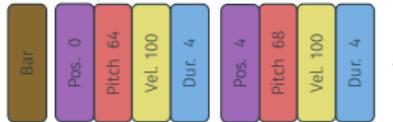
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



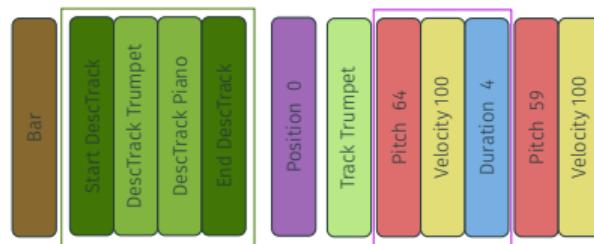
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



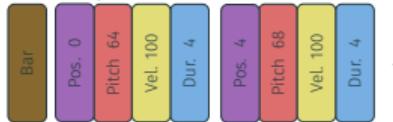
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



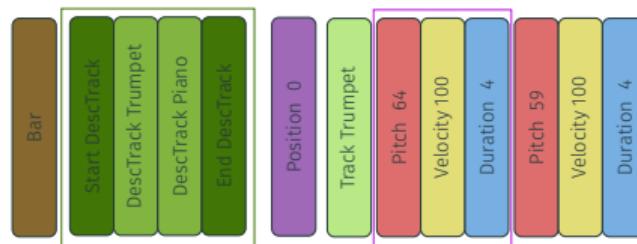
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



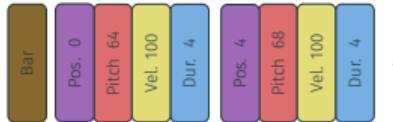
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



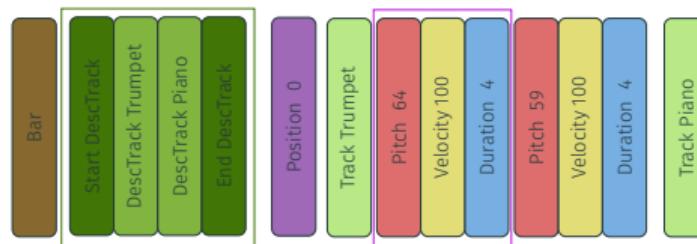
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



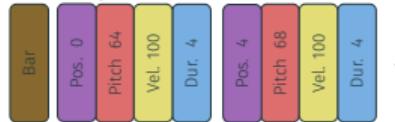
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



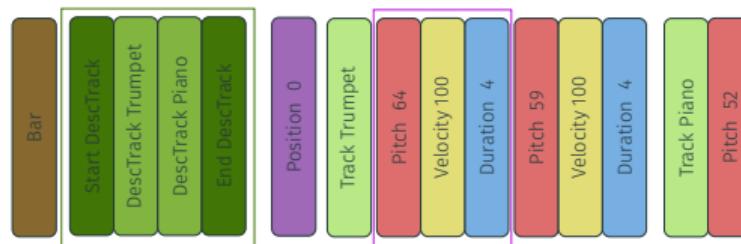
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



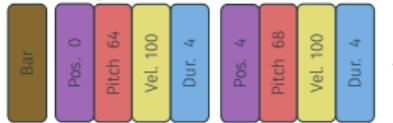
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



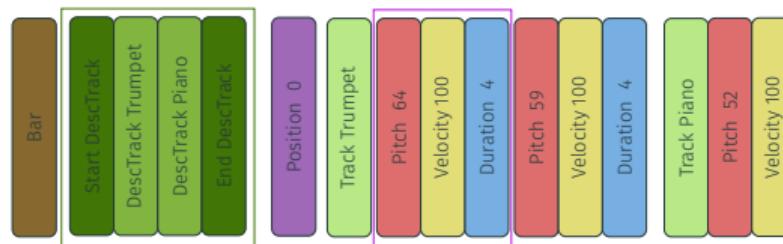
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



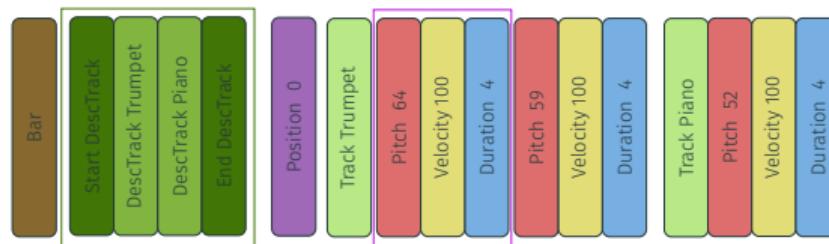
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



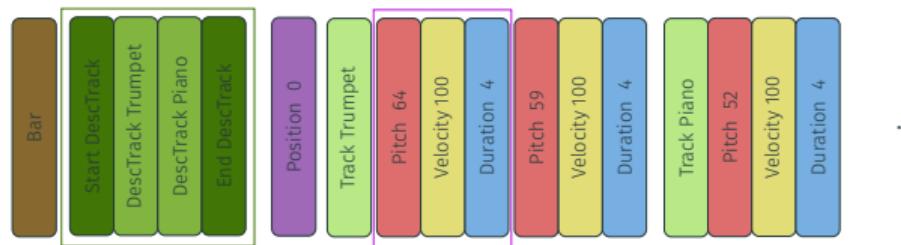
# Melodic fidelity: inference guidance

Melody enforcing during inference.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



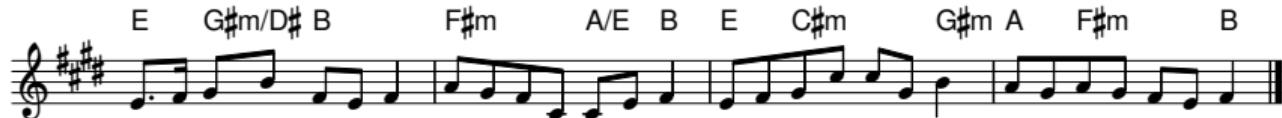
# Lead sheet orchestration... for free?

Lead sheet: melody + chords

A musical lead sheet featuring a single-line melody on a treble clef staff. Above the staff, a sequence of chords is listed with their corresponding time signatures: E (2/4), G#m/D# (2/4), B (2/4), F#m (2/4), A/E (2/4), B (2/4), E (2/4), C#m (2/4), G#m (2/4), A (2/4), F#m (2/4), and B (2/4). The melody consists of eighth-note patterns, and the chords are indicated by vertical stems extending from the notes.

# Lead sheet orchestration... for free?

Lead sheet: melody + chords



⇒ It is simply a low rhythmicity multi-track sample

A musical staff in G major (two sharps) with a common time signature. The melody consists of eighth-note patterns. Below the staff, harmonic chords are shown as vertical stacks of notes. The chords correspond to the labels above: E, G#m/D# B, F#m, A/E, B, E, C#m, G#m A, F#m, and B.

⇒ Lead sheet orchestration = re-orchestration with higher rhythmicity  
(and higher polyphony)

oboe-cello-harp-trio

## Baseline models

- FIGARO<sup>2</sup>: Multi-track music style transfer
- AccoMontage-band<sup>3</sup>: Lead sheet band arrangement

**Objective** (fidelity + controllability metrics) and **subjective** evaluation.

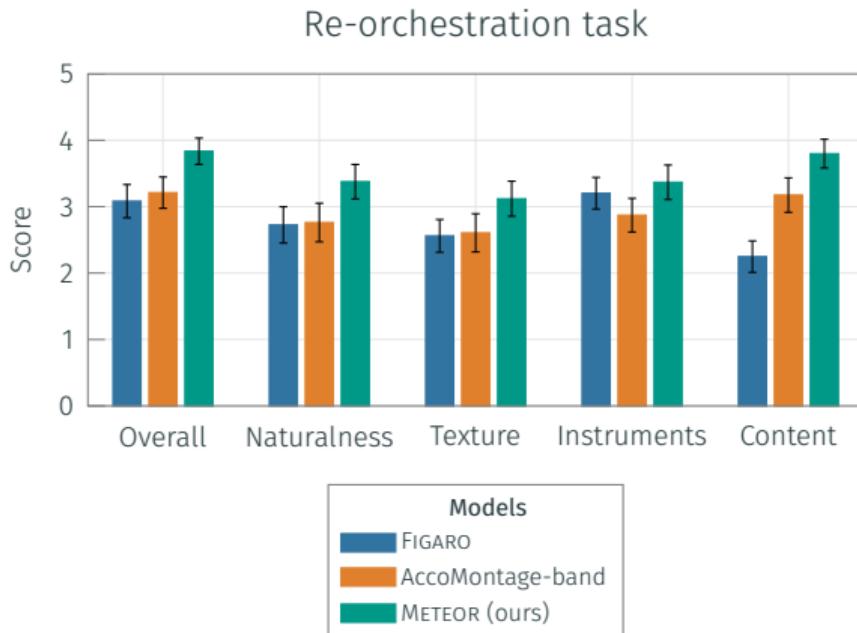
**Criteria:** Overall musicality ; Naturalness ; Texture fidelity ; Convincing use of instruments ; Content coherency.

---

<sup>2</sup>Rütte & al., FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control, 2023

<sup>3</sup>Zhao & al., Structured multi-track accompaniment arrangement via style prior modelling, 2024

# Subjective evaluation



Piano solo → orchestra

Same texture

(Rachmaninoff - Prelude in G minor)

reference

meteor

- **Content coherency:** FIGARO does not keep the melody figaro
- **Texture fidelity:** AccoMontage relies on a lead sheet transcription accomontage

# Towards symbolic music generation for *human* use?

piano-trio-reorchestration 4



<sup>4</sup>Reference: lead sheet of Auld Lang Syne (*Ce n'est qu'un au revoir*).

# Towards symbolic music generation for *human* use?

piano-trio-reorchestration 4

Cello

4

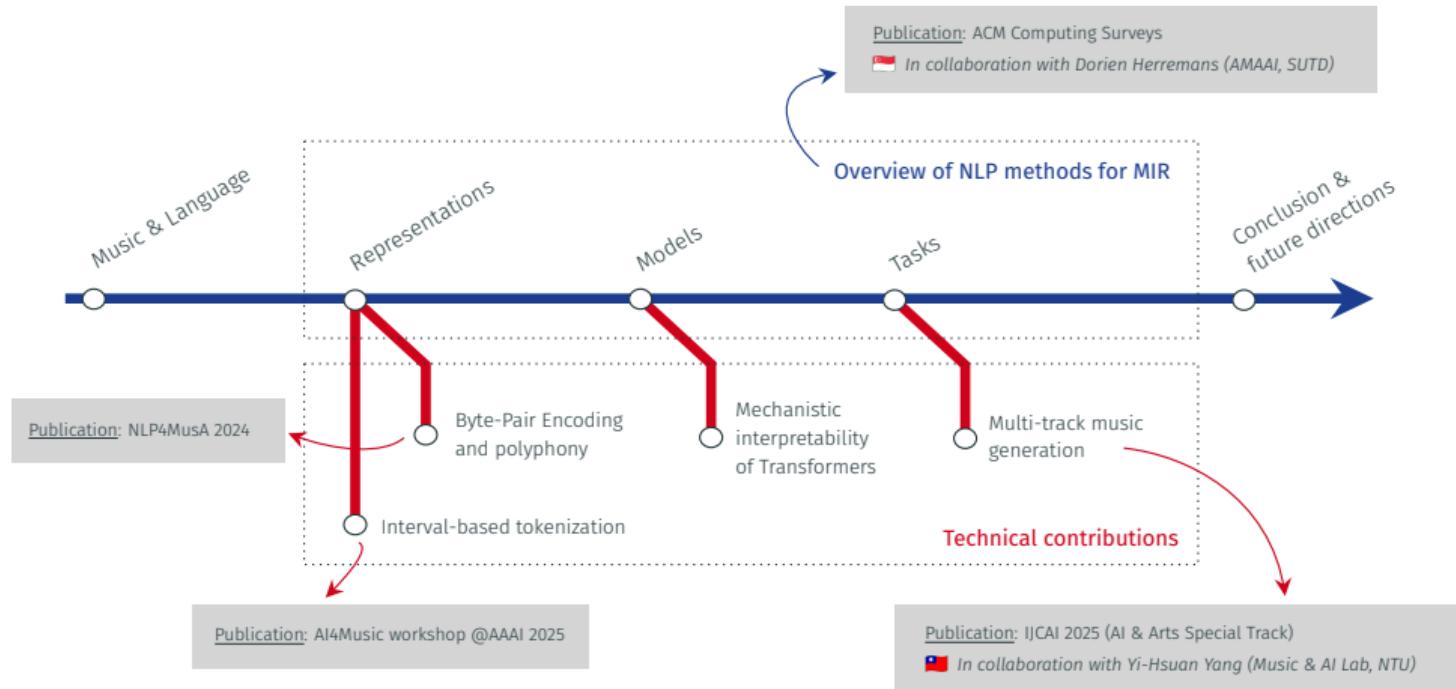
- Enharmonics, voice separation,...
  - Awkward fingerings, articulations, breathings,...
- How to ensure **humanly understandable** and **playable** generated music?

<sup>4</sup>Reference: lead sheet of Auld Lang Syne (*Ce n'est qu'un au revoir*).

## Future directions & conclusion

---

# Contributions, collaborations & publications



Other publications: Computational musicology (DLfM 2022) ; Music corpora (TISMIR 2025) ; AI for general public (Culture & Recherche 2024).

## Musical alphabet vs. text alphabet

- Is the Latin alphabet really the closest to musical alphabet?
- Can one still rely on MIDI to make *playable* music generation?

## Data availability: text vs. symbolic music data

- Text data is released much more quickly than symbolic music
- Are larger models really the solution?

# Future directions

## Practices in the NLP field for MIR

- **Model explainability:** understanding the behavior of a model is essential, particularly in creative contexts.
  - *In NLP:* model-agnostic tools ; mechanistic interpretability
  - ISMIR 2025 Tutorial “Explainable AI for MIR”
- **Benchmarks:** how to compare models not developed for the same tasks?
  - *In NLP:* common metrics (BLEU for translation, ...) ; standardized benchmarks (GLUE, ...)
  - MIREX competition ; SMC Benchmark [Wang & al. ISMIR 2025]

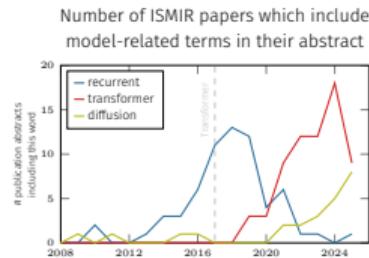
## Towards further models for MIR

## Practices in the NLP field for MIR

### Towards further models for MIR

- Transformers are still used, but have become “standard”
- Adapting models from other fields
  - e.g. diffusion models from image processing

~~ What about a model tailored for music, beyond “simple” adaptations of other models?



## How can one structure a MIR project?

- **Task:** how precisely is it musically defined?
- **Representation:** what motivates a particular choice?
- **Model:** why is it suitable for the chosen task and representation?

## NLP as a *toolbox* for MIR

- NLP tools can be efficient to process symbolic music.
- Applying NLP tools for symbolic music should be first driven by *musical* questions.

## How can one structure a MIR project?

- **Task:** how precisely is it musically defined?
- **Representation:** what motivates a particular choice?
- **Model:** why is it suitable for the chosen task and representation?

## NLP as a *toolbox* for MIR

- NLP tools can be efficient to process symbolic music.
- Applying NLP tools for symbolic music should be first driven by *musical* questions.

Thanks for listening!

# Modeling Symbolic Music with Natural Language Processing Approaches

