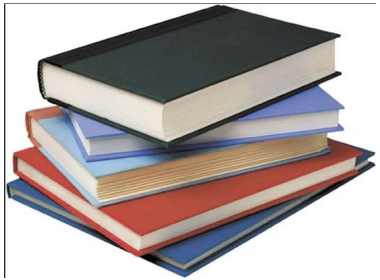
Using Compound Word Transformer for Classical Counterpoint Generation

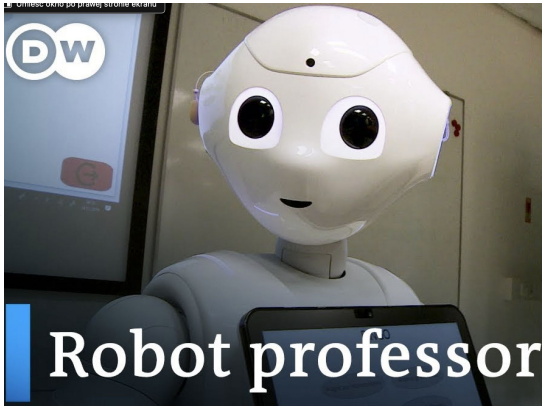
...or NLP-inspired stylized music generator

Artur Dobija & Eres Ferro Bastian | Machine Learning for NLP 2023/4

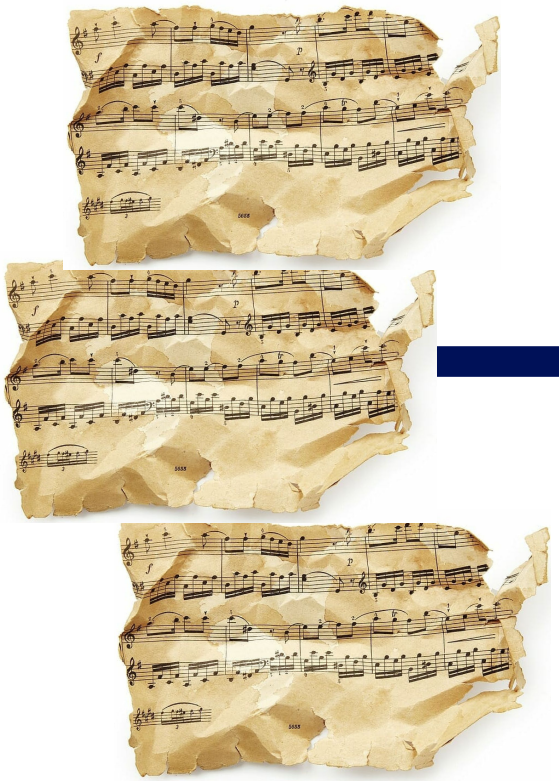
Universiteit
Leiden
The Netherlands

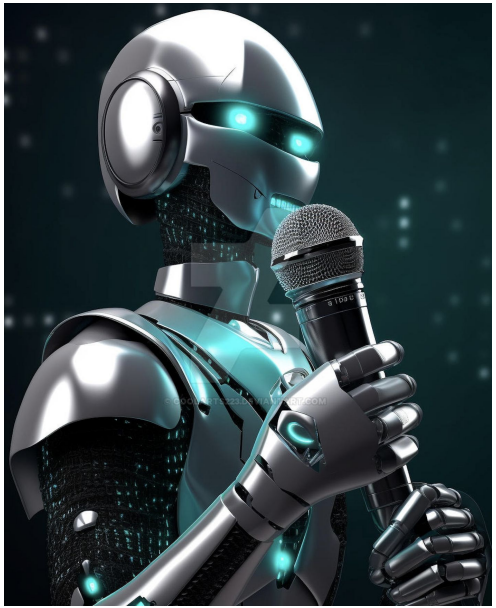
Natural… Language Processing?



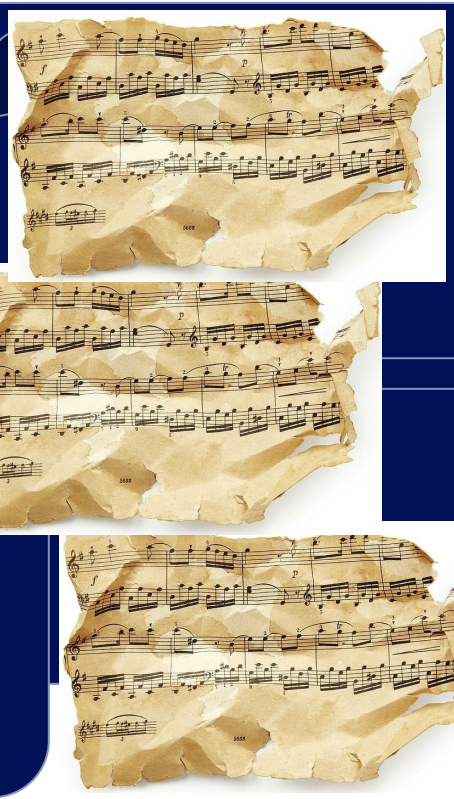


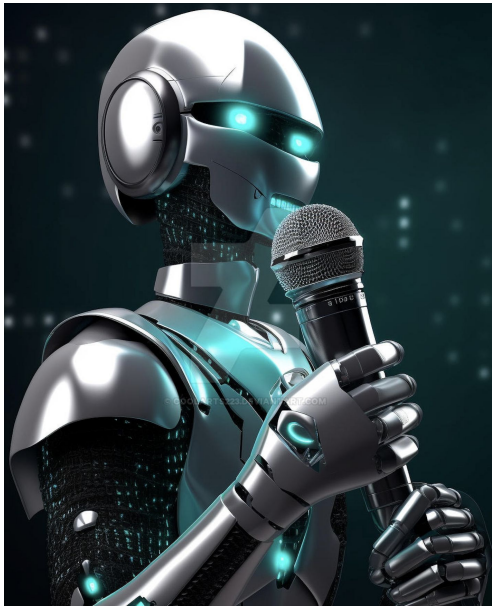
Natural... Music Processing?





Natural Music Processing





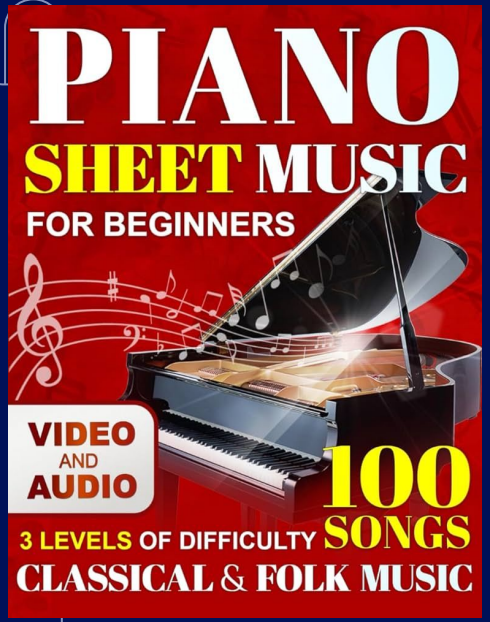
Using Compound Word Transformer for Classical Counterpoint Generation

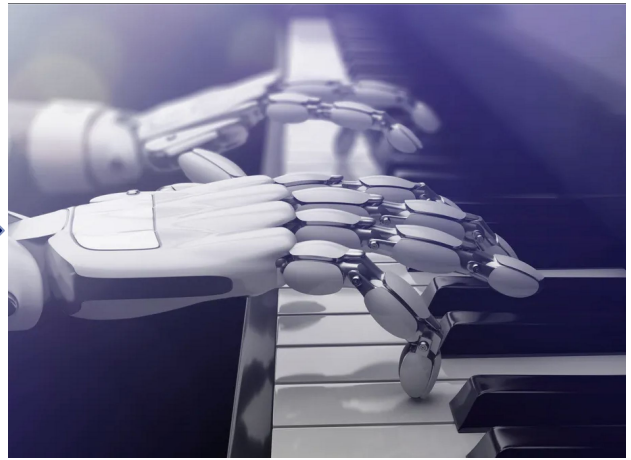
...or NLP-inspired stylized music generator

Artur Dobija & Eres Ferro Bastian | Machine Learning for NLP 2023/4

Universiteit Leiden
The Netherlands

Model we used.





Hsiao & al., **Compound Word Transformer**: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs

Hsiao & al., **Compound Word Transformer**

	Representation	Model	window	Voc. size	Data type
Music Transformer (Huang et al. 2019)	MIDI-like	Transformer	2,048	388	Classical performance
MuseNet (Payne 2019)	MIDI-like*	Transformer	4,096	N/A	Multi-track MIDI
LakhNES (Donahue et al. 2019)	MIDI-like*	Transformer-XL	512	630	Multi-track MIDI
TR autoencoder (Choi et al. 2020)	MIDI-like	Transformer	2,048	388	Classical performance
Pop Music TR (Huang and Yang 2020)	REMI	Transformer-XL	512	332	Pop piano performance
Transformer VAE (Jiang et al. 2020)	MIDI-like	Transformer	128	47	Pop lead sheets
Guitar Transformer (Chen et al. 2020)	REMI*	Transformer-XL	512	221	Guitar tabs
Jazz Transformer (Wu and Yang 2020)	REMI*	Transformer-XL	512	451	Jazz lead sheets
MMM (Ens and Pasquier 2020)	MIDI-like*	Transformer	2,048	>442	Multi-track MIDI
This work	CP	linear Transformer	5,120	350	Pop piano performance

Table 1: A comparison of existing Transformer-based models and the proposed one for automatic music composition. The representations marked with * are extensions of either MIDI-like (Oore et al. 2018) or REMI (Huang and Yang 2020).

- Transformer decoder
- Self-attention layers
- Feed-forward layers
- Adaptive sizes
- Using Compound Words
  - instead of individual tokens
  - derived from MIDI, but not MIDI

Midi is a *data-poor* symbolic music format,
containing most elementary note performance information.
Music score is *data-rich* (contains various metadata).

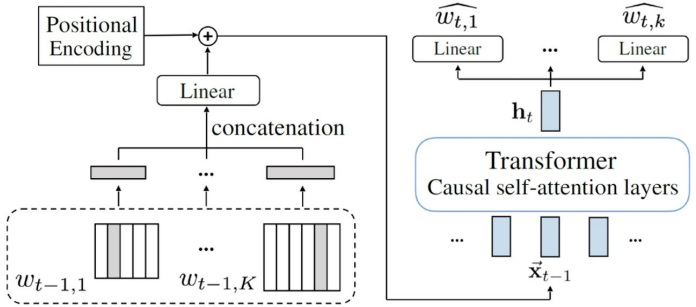
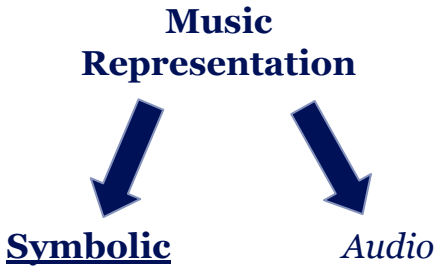


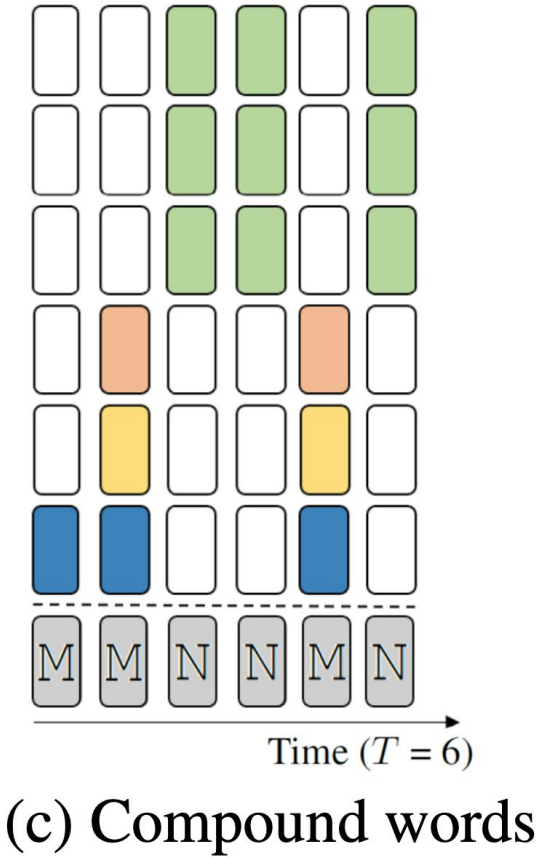
Figure 1: Illustration of the main ideas of the proposed compound word Transformer: (left) *compound word modeling* that combines the embeddings (colored gray) of multiple tokens $\{w_{t-1,k}\}_{k=1}^{K}$, one for each token type k, at each time step $t-1$ to form the input $\vec{\mathbf{x}}_{t-1}$ to the self-attention layers, and (right) *toke type-specific feed-forward heads* that predict the list of tokens for the next time step t at once at the output.

Hsiao & al., **Compound Word Transformer**

"To apply neural sequence models (...) to automatic music composition (...), one has to represent a piece of music as a sequence of tokens drawn from a predefined vocabulary (...).

Unlike the case in text, such a vocabulary usually involves tokens of various types."





(c) Compound words

Using Compound Word Transformer for Classical Counterpoint Generation

...or NLP-inspired stylized music generator

Artur Dobija & Eres Ferro Bastian | Machine Learning for NLP 2023/4

Universiteit Leiden
The Netherlands

Giovanni Pierluigi da Palestrina (1525-1594)

- High Renaissance Counterpoint style
  - polyphony, i.e. many independent voices
  - strict rhythmic movement
  - *simplification warning*
  - “only specific piano keys used

- Composed over 104 catholic masses, each containing:
  - Kyrie
  - Gloria
  - Credo
  - Sanctus
  - Agnus Dei

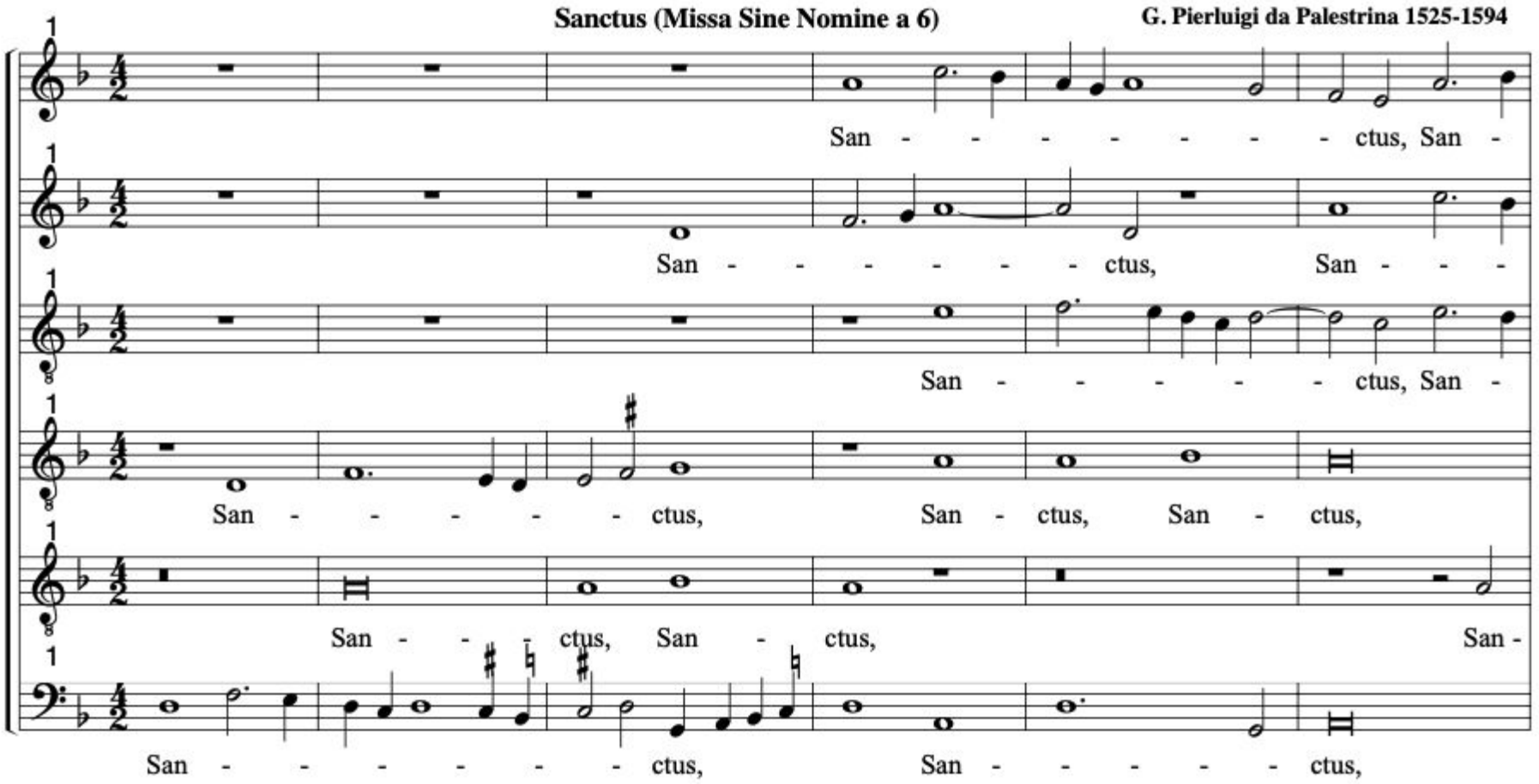
- Our database:
  - Encoded into **kern data format by John Miller*
  - over 1200 items of **kern files

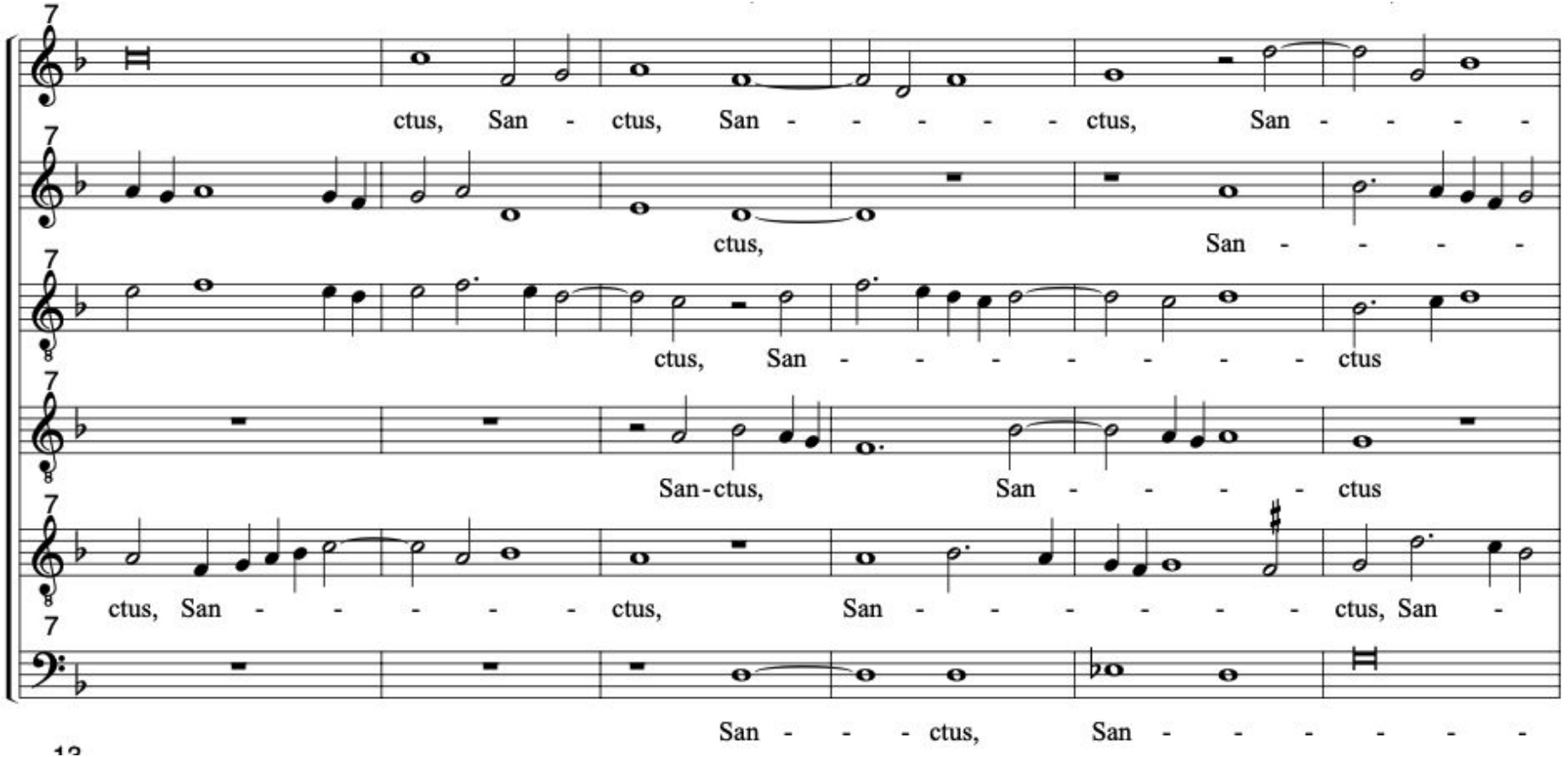
- Let’s listen together to his *Sanctus* from Missa Sine Nomine for 6 voices.

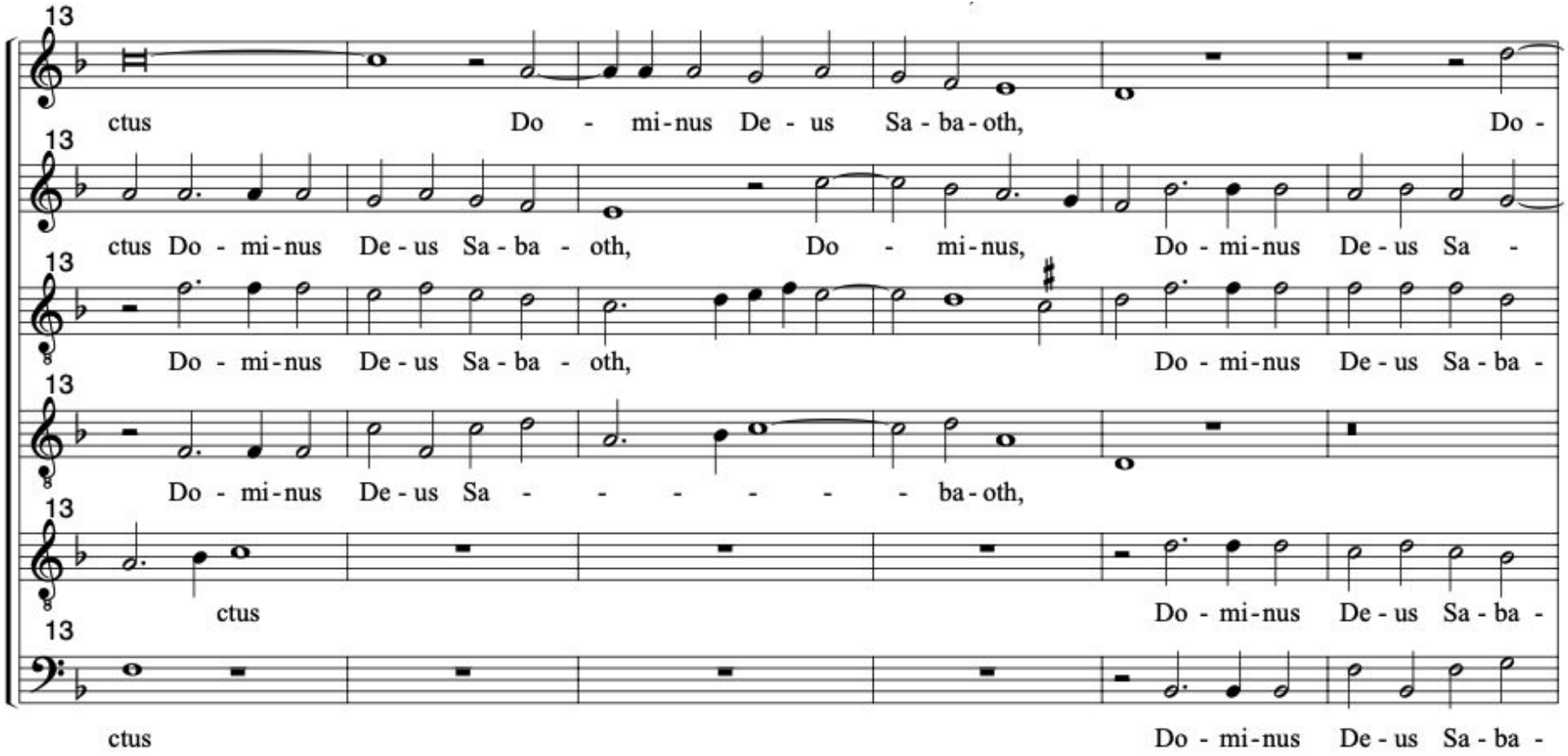
*Miller, J. E. (1992). Aspects of melodic construction in the masses of Palestrina: A computer-assisted study (Volumes I and II) (PhD dissertation). Northwestern University.

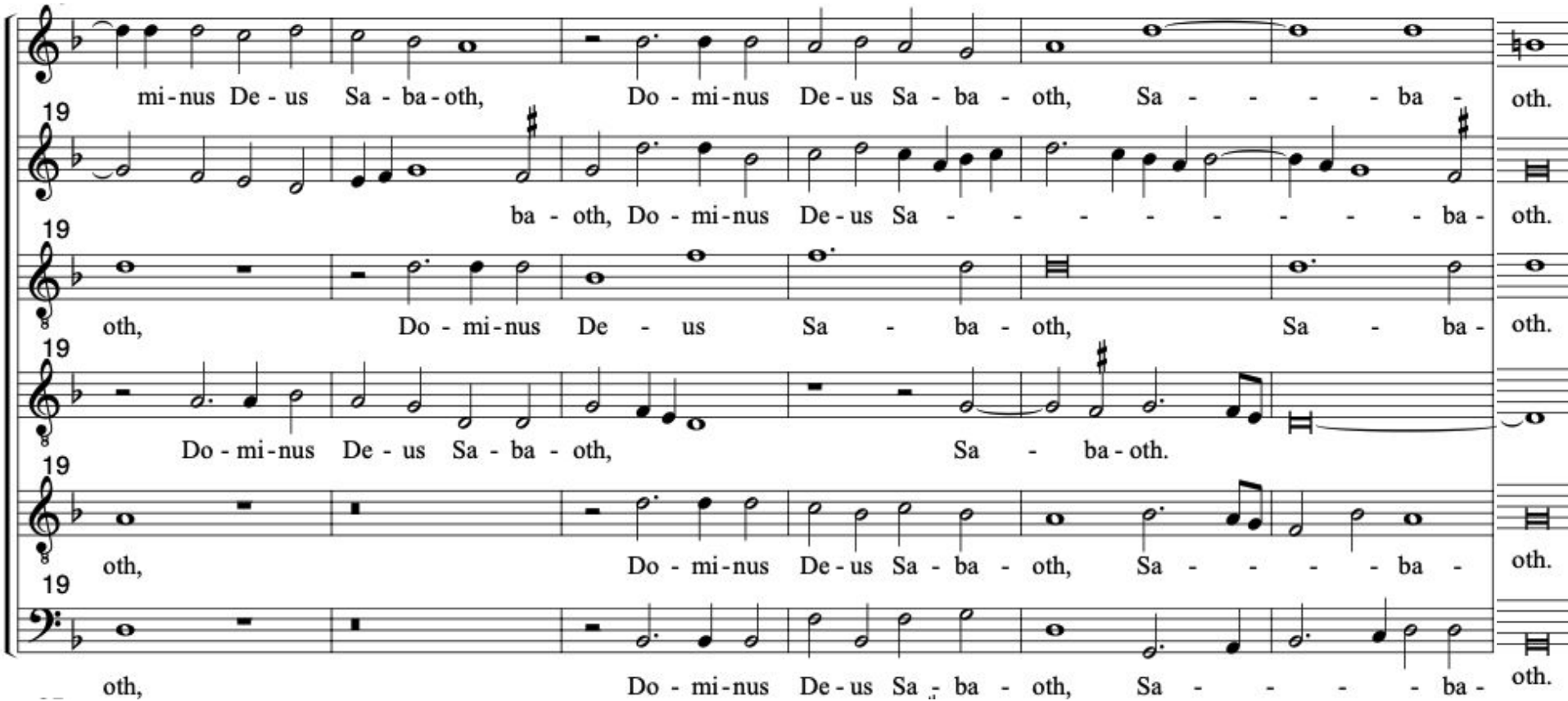
Let’s listen together!

https://open.spotify.com/track/7EODvGhxPboTYVgSwG3Eom?si=40a33a81773e44aa




ctus, San - ctus, San - - - - - ctus, San - - - -
ctus, San - - - -
ctus, San - - - - - - - - ctus
San-ctus, San - - - - ctus
ctus, San - - - - - ctus, San - - - - ctus, San -
San - - - ctus, San - - - - - -


13
ctus Do - mi-nus De - us Sa - ba - oth, Do -
13
ctus Do - mi-nus De - us Sa - ba - oth, Do - mi-nus, Do - mi-nus De - us Sa -
13
Do - mi-nus De - us Sa - ba - oth, Do - mi-nus De - us Sa - ba -
13
Do - mi-nus De - us Sa - - - - - - ba - oth,
13
ctus Do - mi-nus De - us Sa - ba -
13
ctus Do - mi-nus De - us Sa - ba -

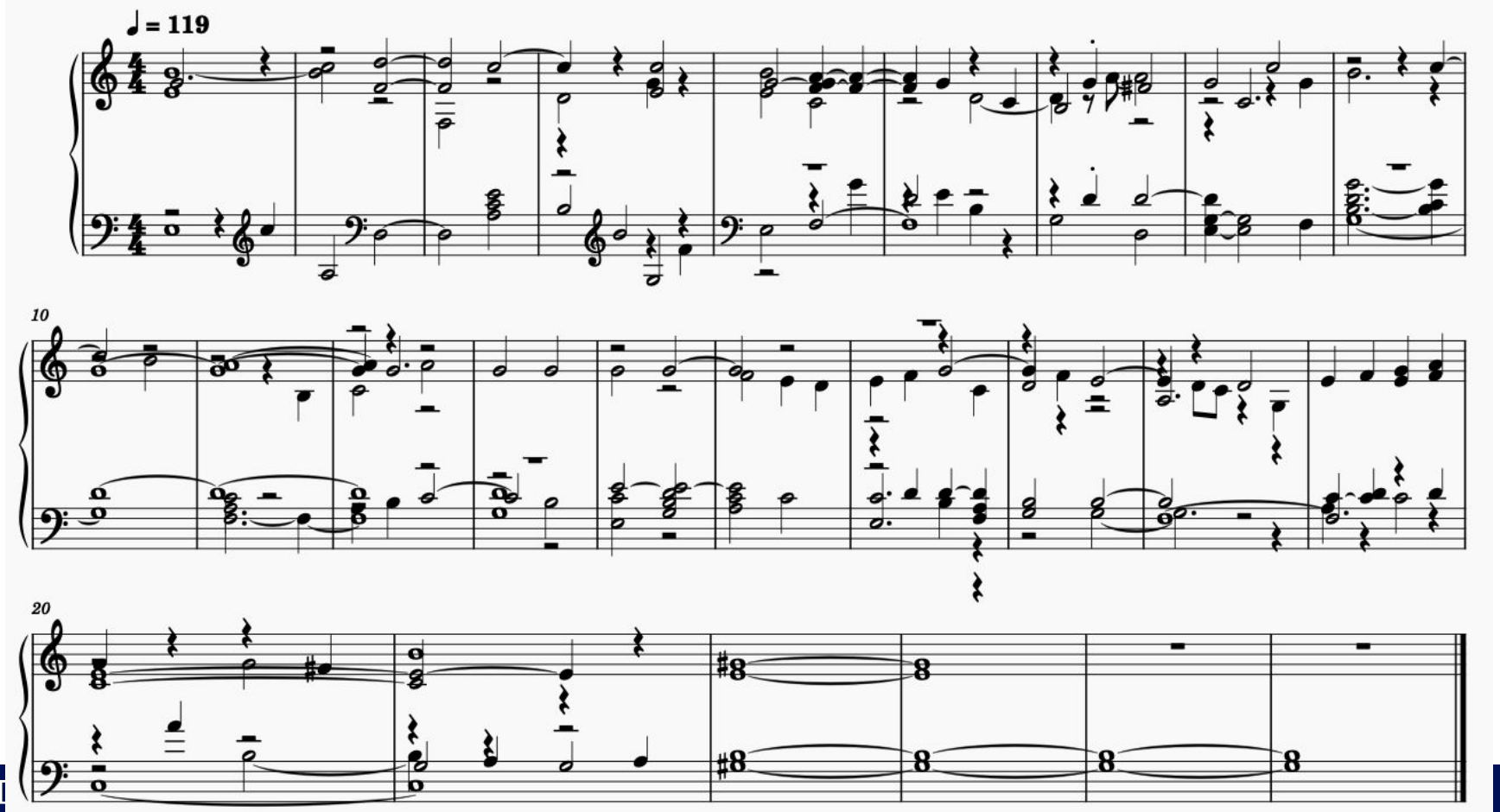


Discover the world at Leiden University

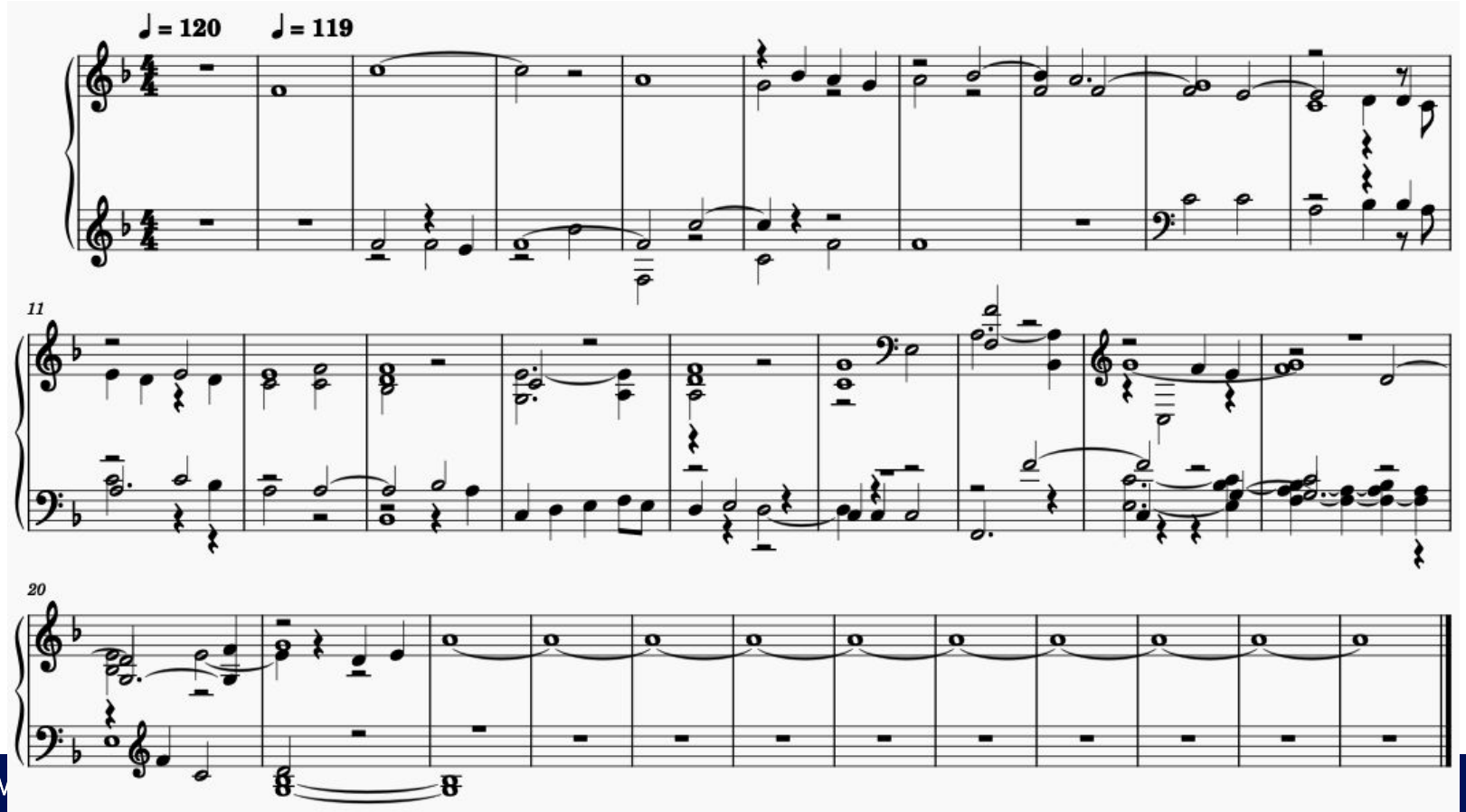
Model we all deserve?

Hsiao & al., **Compound Word Transformer**: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs

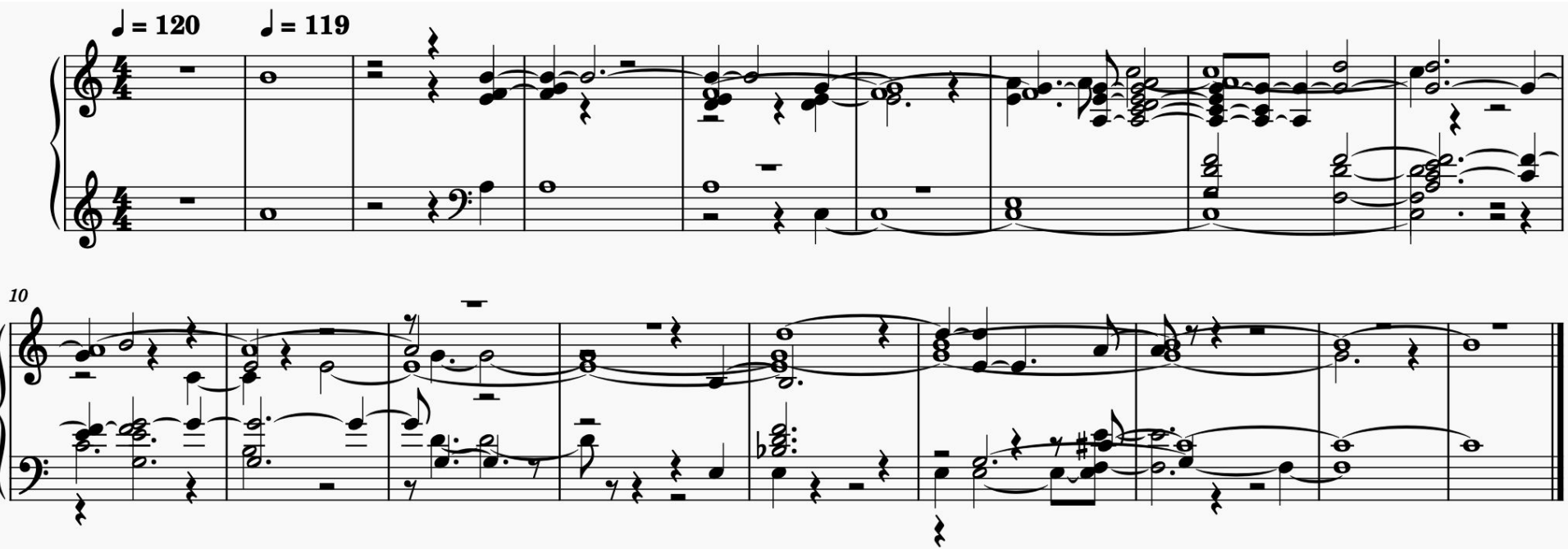
Example result (100 epochs, loss 0.19)



Example result – IMITATION!!!



Humble beginnings (epoch 1, loss 0.6)



Problem statement

How does Compound Word Transformer, serving as an example of NLP technique on music generation, performs with another type of repertoire on example of polyphonic music of G. P. da Palestrina?

Experiment

We conducted this experiment at the Digital Lab at PJ Veth with the following configuration:

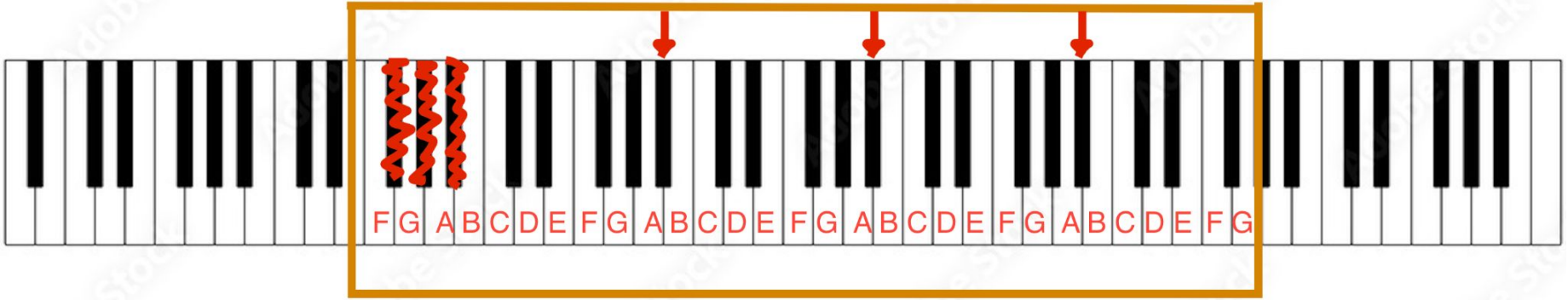
Devices:
Cuda: Nvidia RTX 3080 TI (1x)

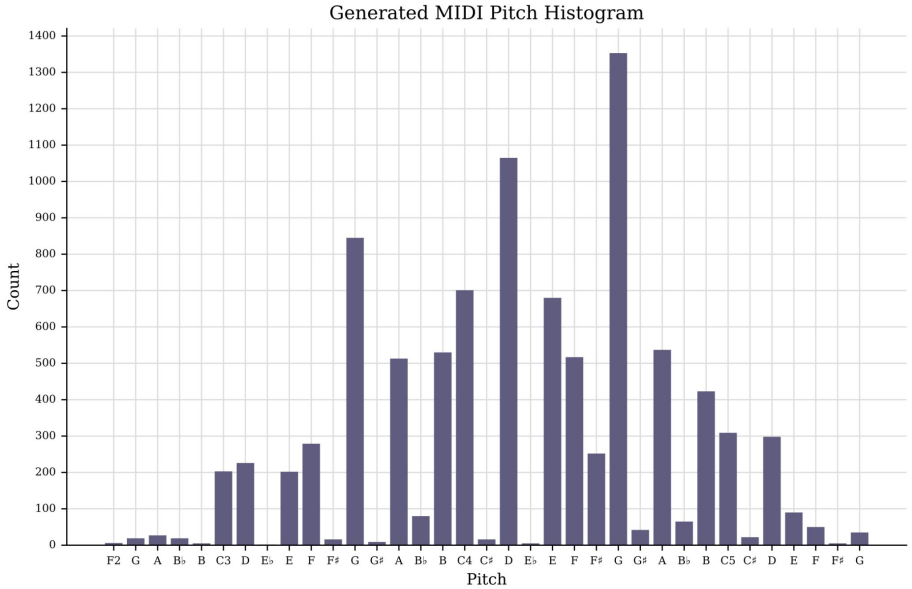
Training time:
2 hours for 100 epochs.

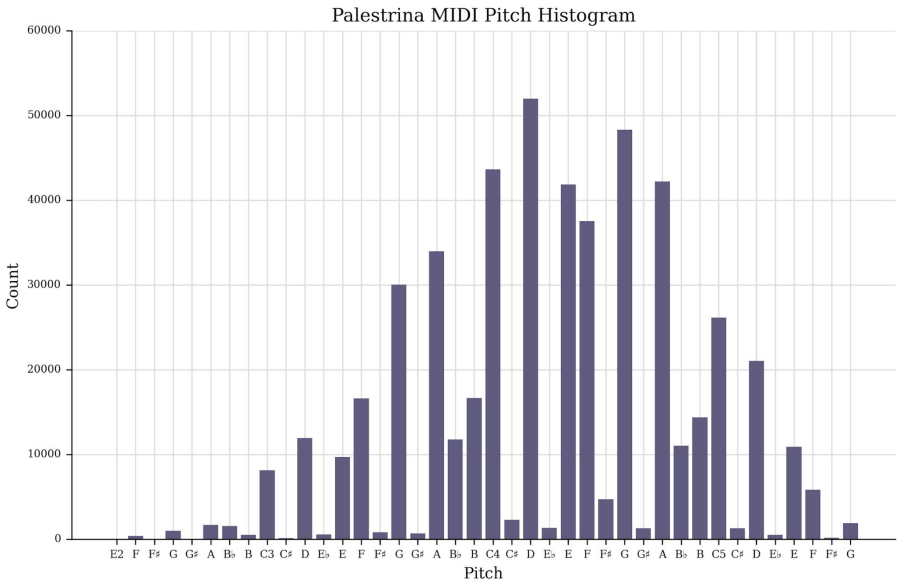
Parameters for training:
dimension size: 128
12 layers
8 attention heads
8 batch sizes
1e-4 learning rate

Metrics for evaluation:
Pitch histogram

Also Yang and Lerch (2020) for:
Cross-validation
Transition matrices



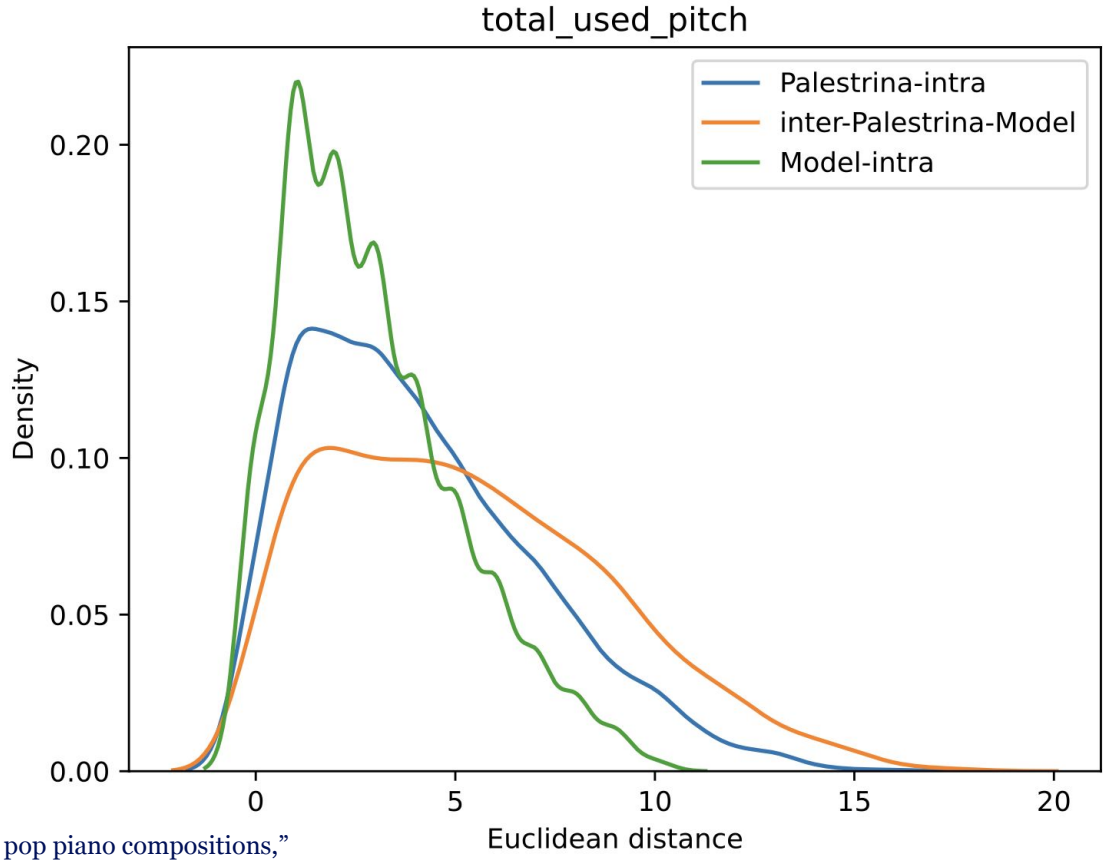




Y.-S. Huang and Y.-H. Yang. (2020) “Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,”

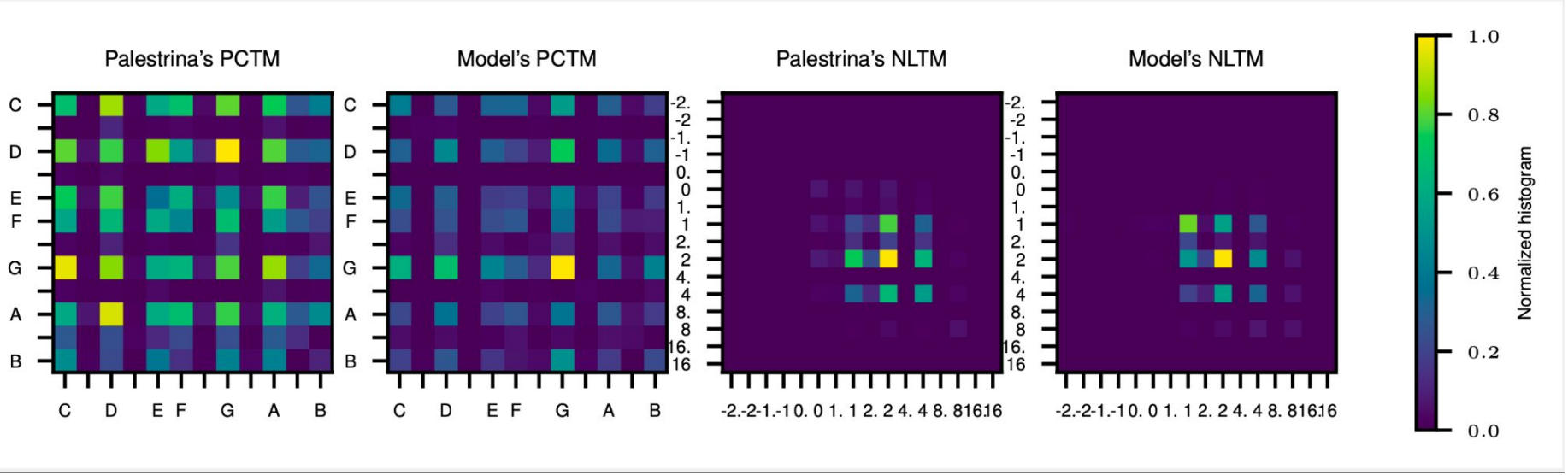
Pairwise Cross-validation Relative Metric

- Model-intra (green) has higher pitch density than inter-Palestrina (yellow) or model-Intra (blue).
- Model-intra also has sharp declines in density but gets back up again.
- Possible reason: long empty segments toward the end of the music.



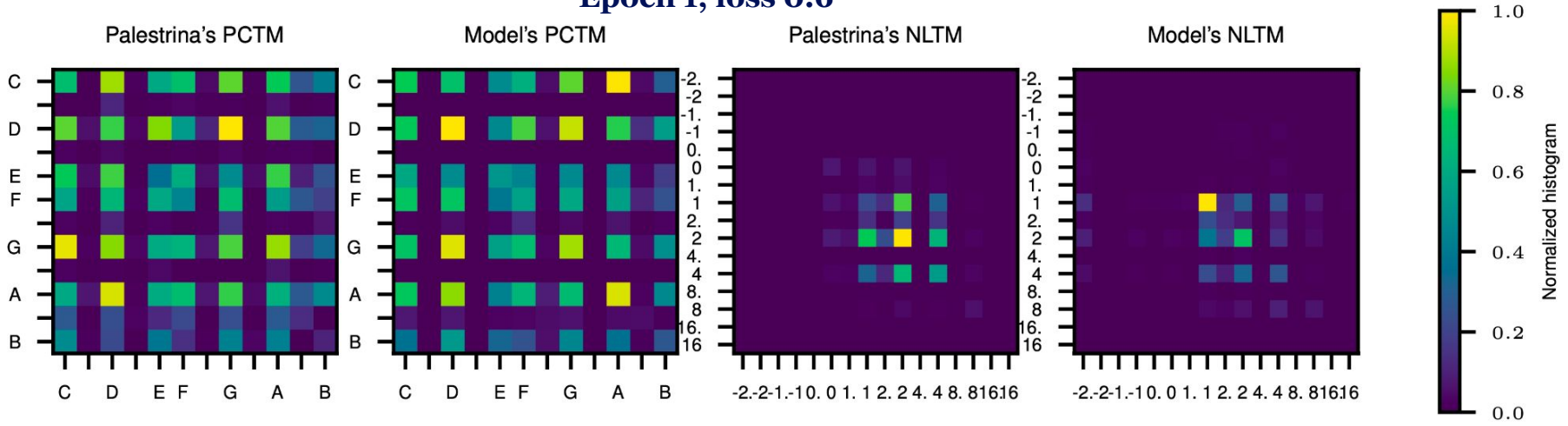
Y.-S. Huang and Y.-H. Yang. (2020) "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,"

Transition Matrix

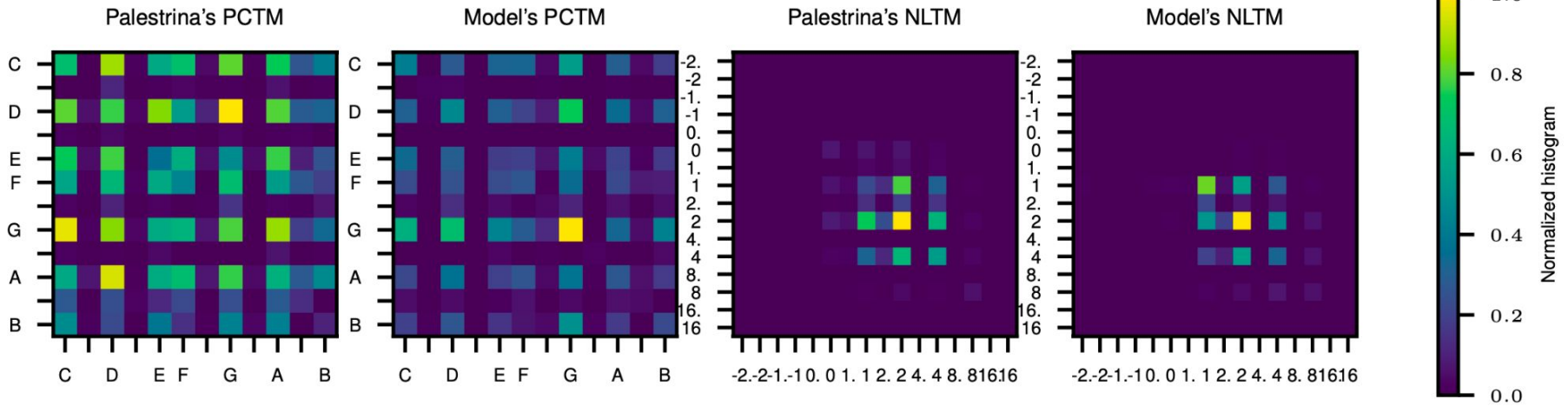


Y.-S. Huang and Y.-H. Yang. (2020) "Pop music
transformer: Beat-based modeling and generation of expressive pop piano compositions,"

Epoch 1, loss 0.6

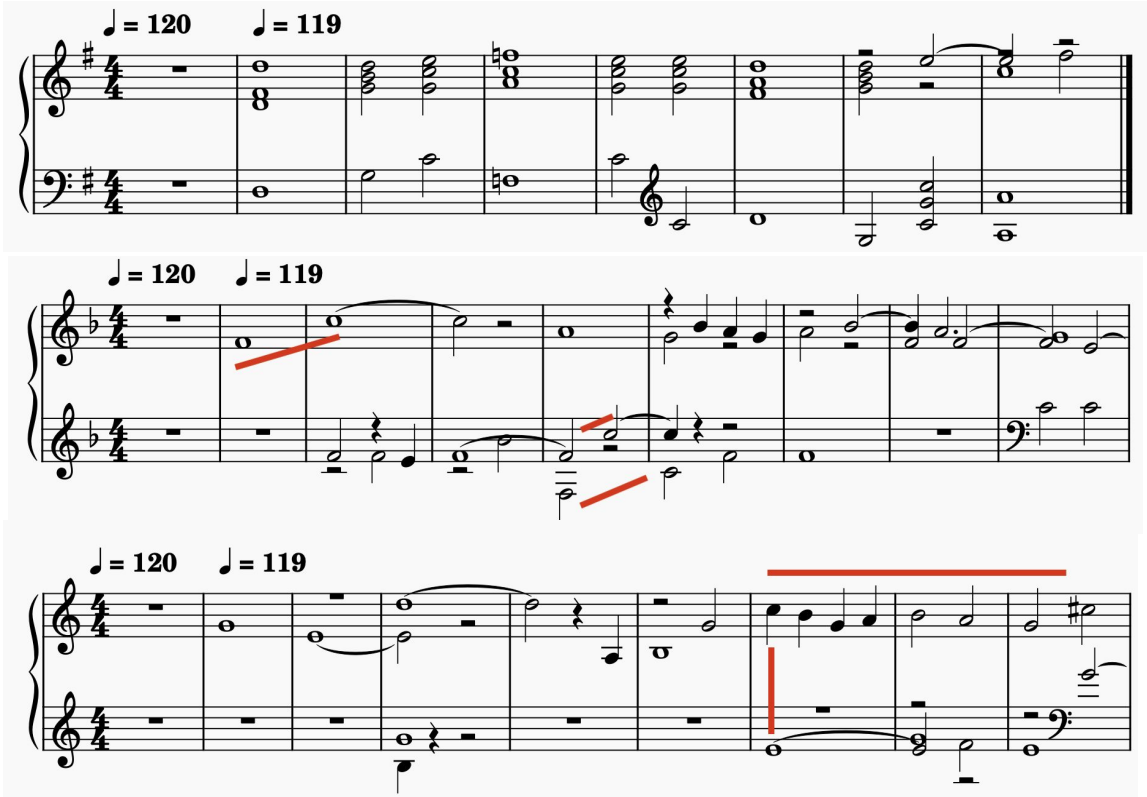


Epoch 100, loss 0.19



Musicological “moments”

- Correct & idiomatic harmonic progression
- Hints of initial imitation
- “Textbook” correct voice leading



Conclusion

- Compound Word Transformer is still highly capable tool despite the complexities of our data (e.g. another type of repertoire on example of polyphonic music).

- Very inexpensive to train! Yet, the inferences prove that the model learns Palestrina's musicological concept.

Future Work

- Rewriting implementation of the Compound Word to encode features more typical for Palestrina (e.g. voice tracking, melodic climaxes, motive importance in imitation, etc.)

- Future experiments that involve (i.e. measurements) the importance of attention layers on our result or how did the Transformer keep track of musicological concepts.
    - TCAV: Testing with Concept Activation Vectors*

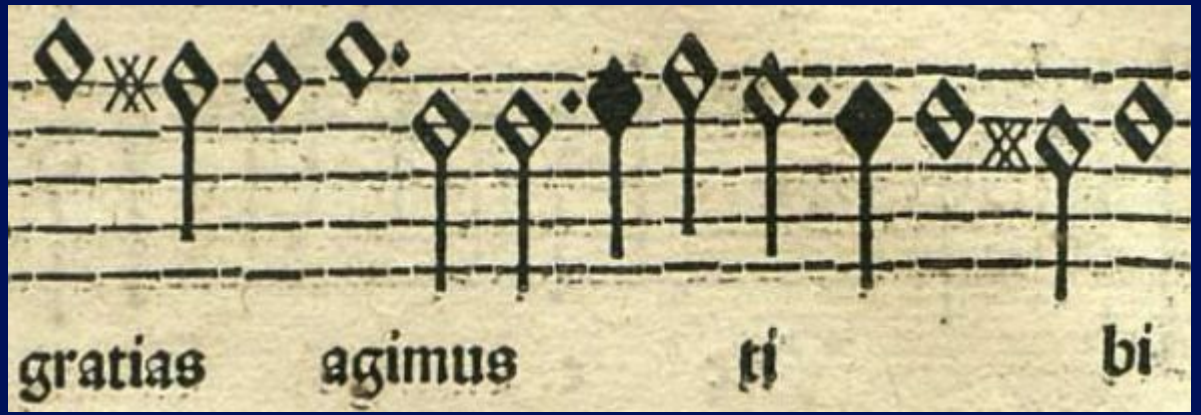
      Foscarin & al., **Concept-Based Techniques for "Musicologist-Friendly" Explanations in a Deep Music Classifier**

- Human-centered evaluations of the model's inferences:
    - Musical Turing test (Belgum et al., 1989)
    - HER metric (Kalonaris et al., 2020) – measuring vector space distances original (generated) score and its adjusted version by human evaluator.

Source code:

https://github.com/ArturJD96/Leiden_ML4NLP_2023-4

Thank you!



Universiteit Leiden
The Netherlands

References

Hsiao & al. (2021). "**Compound Word Transformer**: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs."

Miller, J. E. (1992). Aspects of melodic construction in the masses of Palestrina: A computer-assisted study (Volumes I and II) (PhD dissertation). Northwestern University.

E. Belgum, C. Roads, J. Chadabe, T. Toben- feld, and L. Spiegel. (1989) "A turing test for "musical intelligence"?"

S. Kalonaris, T. McLachlan, and A. Aljanaki. (2020) "Computational Linguistics Metrics for the Evaluation of Two-Part Counterpoint Generated with Neural Machine Translation."

Foscarin & al. (2022) "Concept-Based Techniques for "Musicologist-Friendly" Explanations in a Deep Music Classifier."

Y.-S. Huang and Y.-H. Yang. (2020) "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions."