THESIS PROPOSAL

Empowering Music Creation with Machine Learning

Hao-Wen Dong

Advisors: Julian McAuley and Taylor Berg-Kirkpatrick

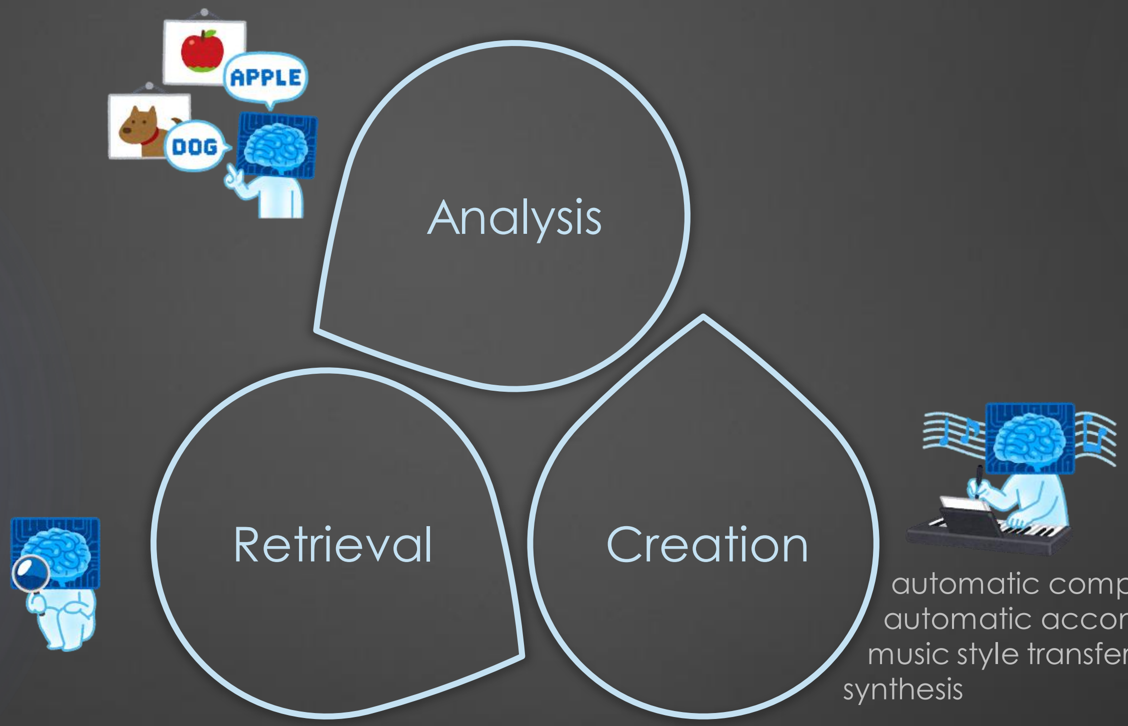
Outline

- Introduction
- Prior work
- Preliminary results
- Future work
- Timeline

Introduction

Music information research (MIR)

- Intelligent ways to analyze, retrieve and create music (Yang 2018)



automatic composition, automatic accompaniment, music style transfer, music synthesis

AI-powered music creation

- *Just to name a few…*

Sony CSL

AI.MUSIC

magenta

OpenAI

AI Labs.tw
台灣人工智慧實驗室

TikTok

Jukedeck

boomy

amper

EVOKE MUSIC

Aiva

How can ML empower music creation?

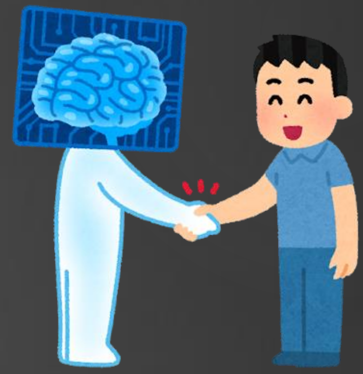
Big data

Learn patterns and rules from big data

Automation

Liberate humans from laborious tasks

Human-AI co-creation



Make the best of both worlds

Music creation workflow





Composition → Arrangement → Sound production → Mixing → Mastering

PQ, "5 Stages of the Music Production Process," *Icon Collective*, 2019. URL: https://iconcollective.edu/music-production-process/

Proposed research projects

Generate multitrack music from scratch or conditionally

Multitrack music generation

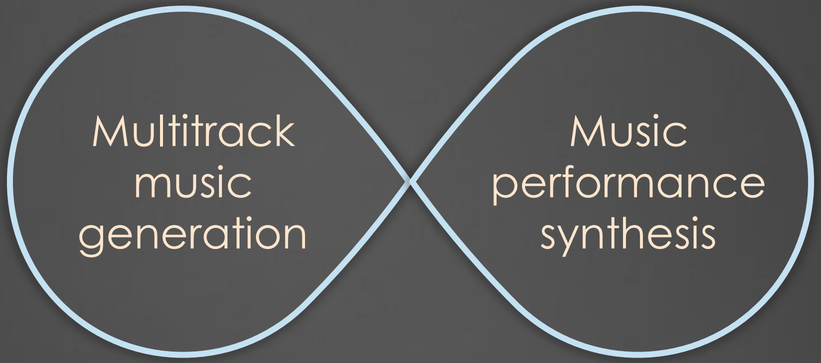
Music performance synthesis

Synthesize musical scores into expressive performance audio

Relevant topics

Generate multitrack music from scratch or conditionally

- Polyphonic music generation
- Automatic accompaniment
- Automatic arrangement



Synthesize musical scores into expressive performance audio

- Expressive music generation
- Performance rendering
- Audio synthesis

Challenges

Generate multitrack music from scratch or conditionally

- Inter-track dependency
- Music structure modeling
- Usability & controllability
- Human-AI interaction

Multitrack music generation

Music performance synthesis

Common challenges:

- Data source
- ML formulation
- Network architecture

Synthesize musical scores into expressive performance audio

- Expressiveness modeling
- Playing style modeling
- Interpretability & controllability

Prior Work

ML-based multitrack music generation

	Model	Representation	Symbolic
★ MuseGAN (Dong et al. 2018)	GAN	Piano roll	✓
MultitrackMusicVAE (Simon et al. 2018)	LSTM	MIDI-like	✓
BinaryMuseGAN (Dong et al. 2018)	GAN	Piano roll	✓
★ MuseNet (Payne 2018)	Transformer	MIDI-like	✓
LakhNES (Donahue et al. 2018)	Transformer	MIDI-like	✓
MMM (Ens & Posquier 2018)	Transformer	MIDI-like	✓
Jukebox (Dhariwal et al. 2018)	VQVAE	Waveform	

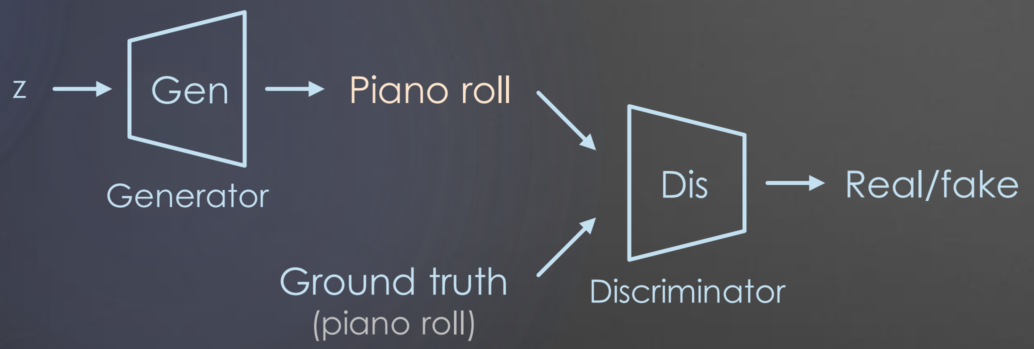
*This list excludes chorales and lead sheet generation systems.

1. Dong et al., "MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," in *AAAI*, 2018.
2. Simon et al., "Learning a Latent Space of Multitrack Measures," in *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2018.
3. Dong and Yi-Hsuan Yang, "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation," in *ISMIR*, 2018.
4. Payne, "MuseNet," OpenAI, Apr. 25, 2019. URL: https://openai.com/blog/musenet/.
5. Donahue et al., "LakhNES: Improving multi-instrumental music generation with cross-domain pre-training," in *ISMIR*, 2019.
6. Ens and Pasquier. "MMM: Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.
7. Dhariwal et al., "Jukebox: A Generative Model for Music," *arXiv preprint arXiv:2005.00341*, 2020.

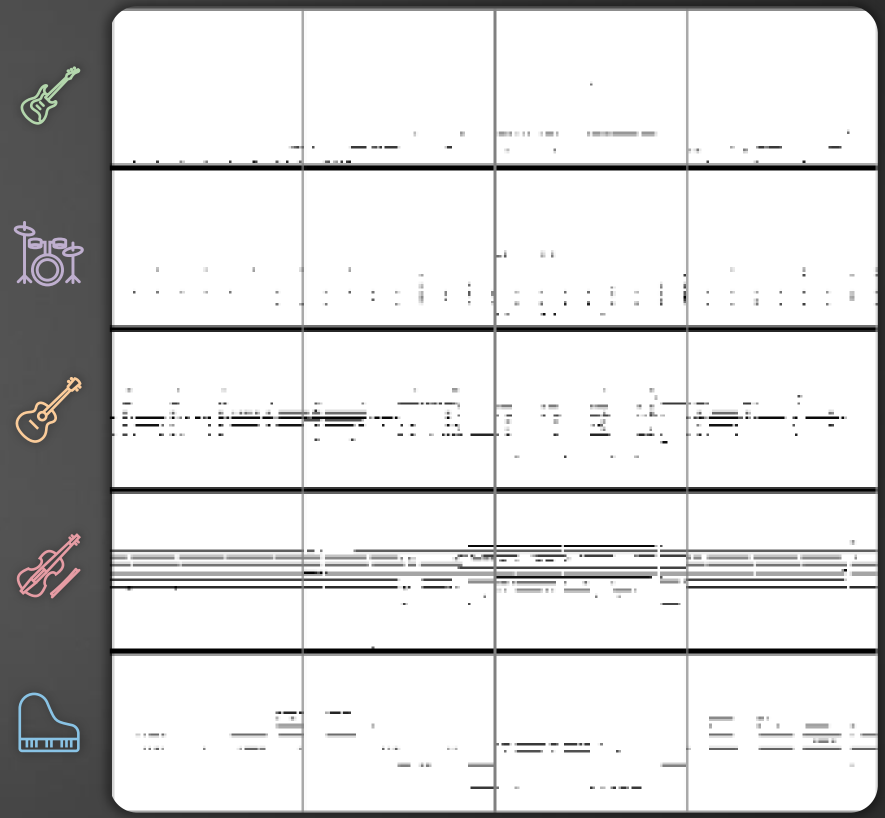
MuseGAN for multitrack music generation (Dong et al. 2018)

- Use convolutional GAN (Goodfellow et al. 2014; Radford et al. 2016)
- Use multitrack piano roll

Model



Generated sample

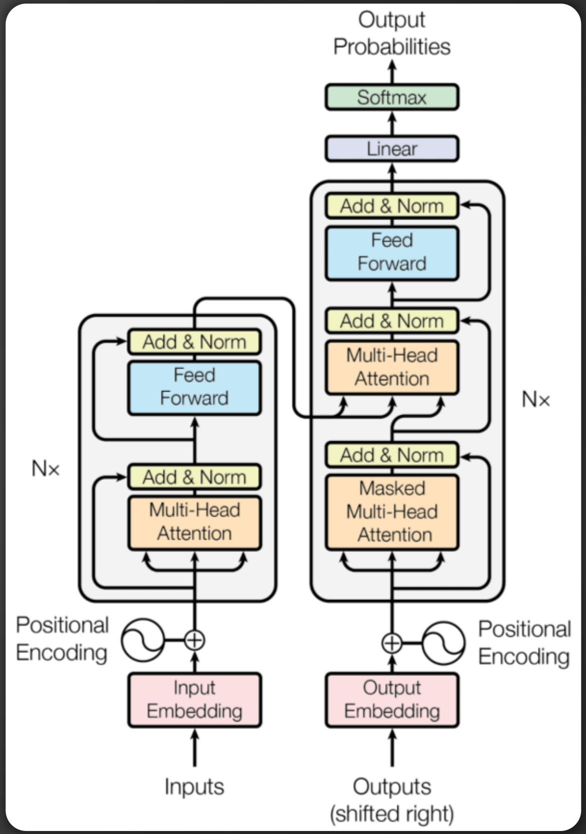


1. Dong et al., "MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," in *AAAI*, 2018.
2. Goodfelllow et al., "Generative Adversarial Nets," in *NeurIPS*, 2014.
3. Radford et al., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *ICLR*, 2016.

MuseNet for multitrack music generation (Payne 2019)

- Use transformer model (Vaswani et al. 2017)
- Use MIDI-like tokens

Model



Example representation

```
bach piano_strings start tempo90 piano:v72:G1
piano:v72:G2 piano:v72:B4 piano:v72:D4 violin:v80:G4
piano:v72:G4 piano:v72:B5 piano:v72:D5 wait:12
piano:v0:B5 wait:5 piano:v72:D5 wait:12 piano:v0:D5
wait:4 piano:v0:G1 piano:v0:G2 piano:v0:B4
piano:v0:D4 violin:v0:G4 piano:v0:G4 wait:1
piano:v72:G5 wait:12 piano:v0:G5 wait:5 piano:v72:D5
wait:12 piano:v0:D5 wait:5 piano:v72:B5 wait:12
```


1. Payne, "MuseNet," OpenAI, Apr. 25, 2019. URL: https://openai.com/blog/musenet/.
2. Vaswani et al., "Attention Is All You Need," in *NeurIPS*, 2017.

ML-based music performance synthesis

Score ⟶ Performance ⟶ Audio

- YQX (Widmer et al. 2009)
- Basis-Mixer (Cancino-Chacón et al. 2017)
- ★ VirtuosoNet (Jeong et al. 2019a)
- ISGN (Jeong et al. 2019b)
- CVRNN (Maezawa et al. 2019)

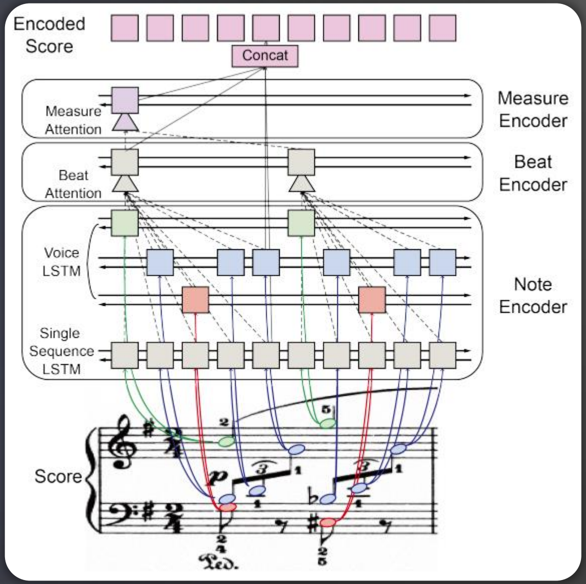
- PerformanceNet (Wang & Yang 2019)
- ★ Wave2Midi2Wave (Hawthorne et al. 2019)
- SynthNet (Schimbinschi et al. 2019)
- Mel2mel (Kim et al. 2019)
- ★ DDSP (Engel et al. 2020)
- DeepSinger (Ren et al. 2020)
- NWS (Hayes et al. 2021)
- MIDI-DDSP (Anonymous 2022)

1. Widmer et al., "YQX plays Chopin." AI magazine, 30(3):35, 2009.
2. Cancino-Chacón et al., "An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music," *Machine Learning*, 106(6), 887–909, 2017.
3. Jeong et al., "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," in *ISMIR*, 2019a.
4. Jeong et al., "Graph Neural Network for Music Score Data and Modeling Expressive Piano Performance," in *ICML*, 2019b.
5. Maezawa et al., "Rendering Music Performance With Interpretation Variations Using Conditional Variational RNN," in *ISMIR*, 2019.
6. Wang and Yang, "PerformanceNet: Score-to-Audio Music Generation with Multi-Band Convolutional Residual Network," in *AAAI*, 2019.
7. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *ICLR*, 2019.
8. Schimbinschi et al., "SynthNet: Learning to Synthesize Music End-to-End," in *IJCAI*, 2019.
9. Kim et al., "Neural Music Synthesis for Flexible Timbre Control," in *ICASSP*, 2019.
10. Engel et al., "DDSP: Differentiable Digital Signal Processing," in *ICLR*, 2020.
11. Ren et al., "DeepSinger: Singing Voice Synthesis with Data Mined From the Web," in *KDD*, 2020.
12. Hayes et al., "Neural Waveshaping Synthesis," in *ISMIR*, 2021.
13. Anonymous, "MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling," in *ICLR*, in press, 2022.

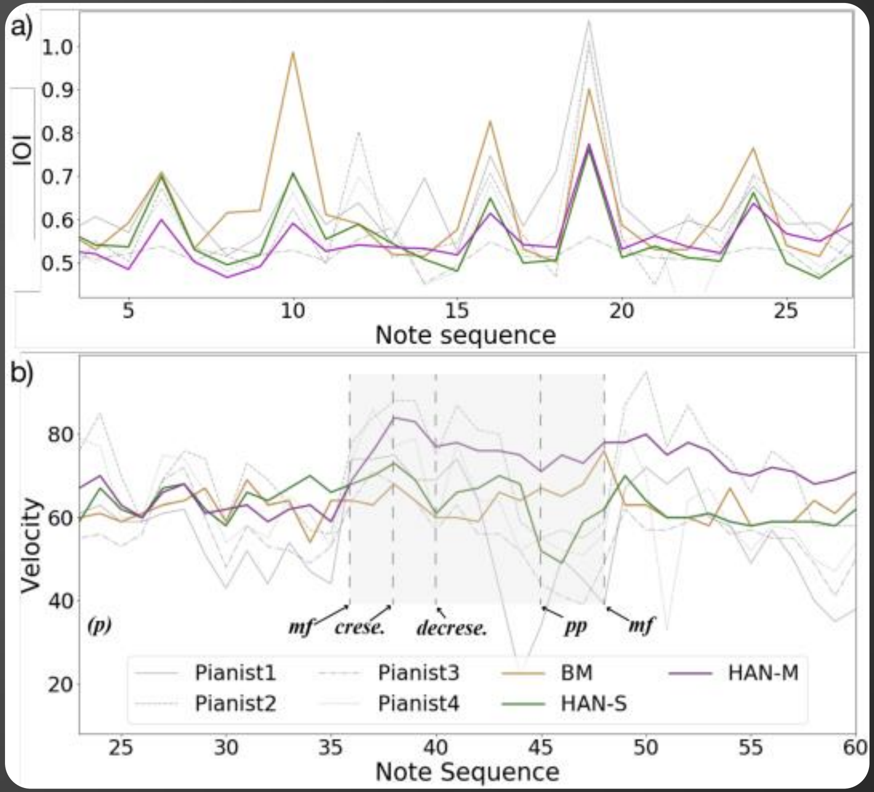
VirtuosoNet for performance rendering (Jeong et al. 2019)

- Use hierarchical RNN
- Use a musical score as the input

Model



Local tempo and dynamics

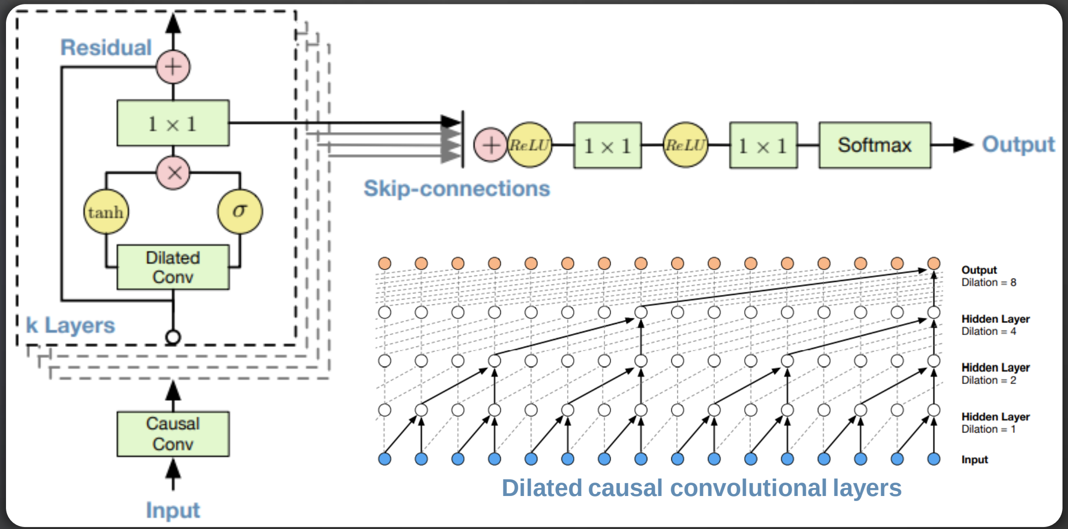


1. Jeong et al., "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," in *ISMIR*, 2019a.

Wave2Midi2Wave for music synthesis (Hawthorne et al. 2019)

- Use conditional WaveNet (van der Oord et al. 2016)
- Use piano rolls as the input

Model

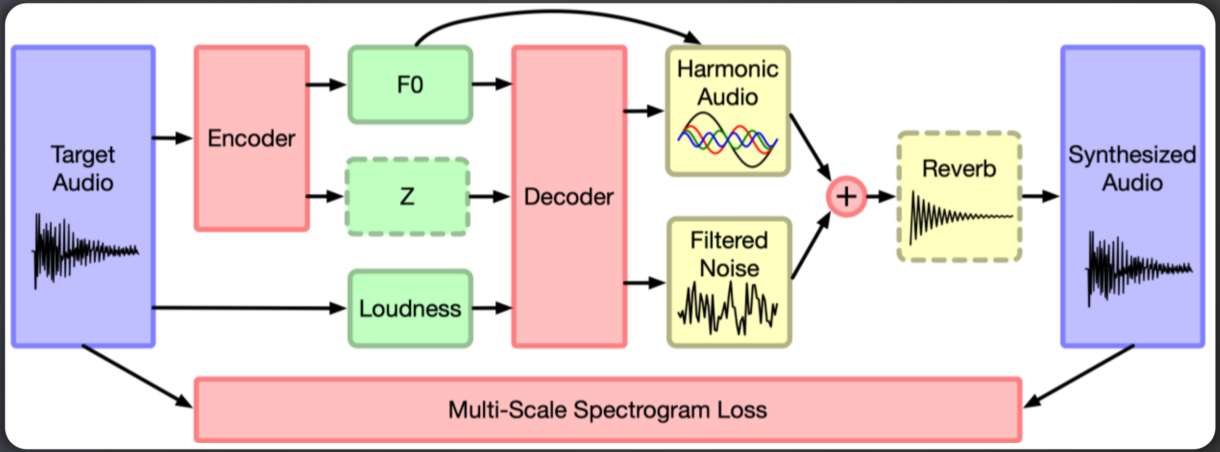


1. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *ICLR*, 2019.
2. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv preprint arXiv:1609.03499, 2016.

DDSP for music synthesis (Engel et al. 2020)

- Use autoencoder to reconstruct audio
- Use differentiable DSP modules as the decoder

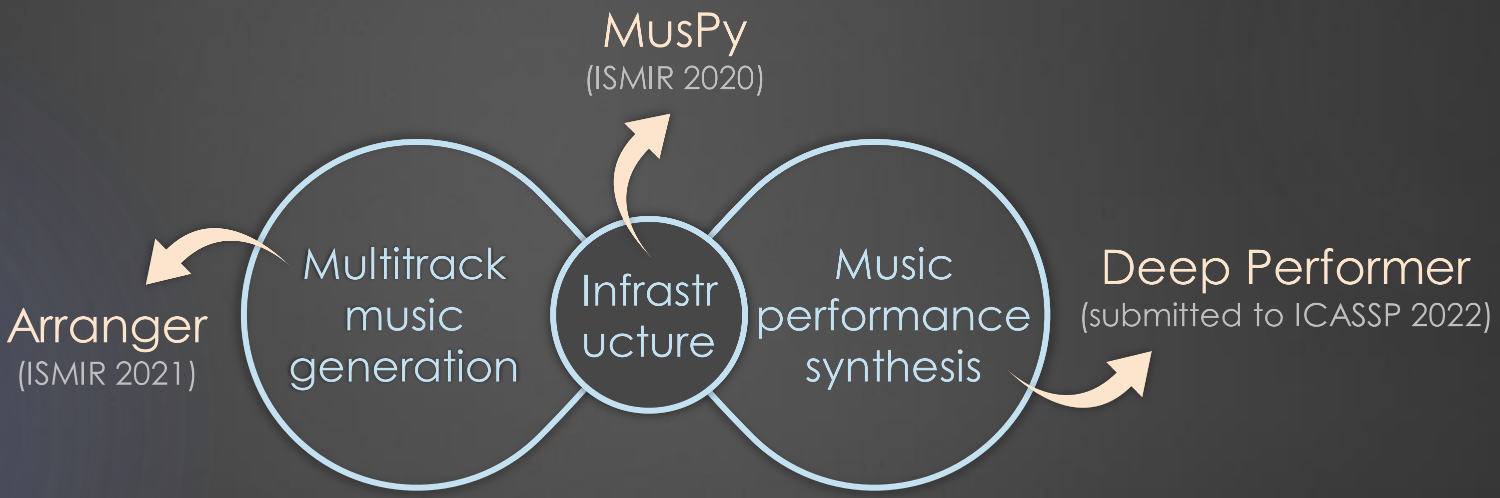
Model



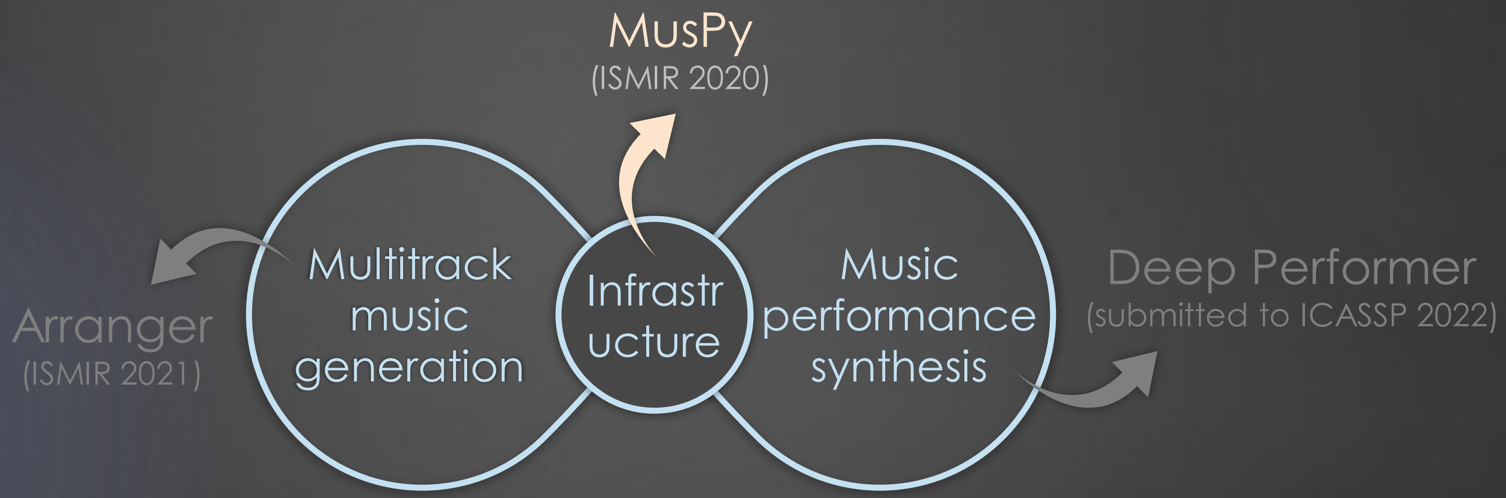
1. Engel et al., "DDSP: Differentiable Digital Signal Processing," in *ICLR*, 2020.

Preliminary Results

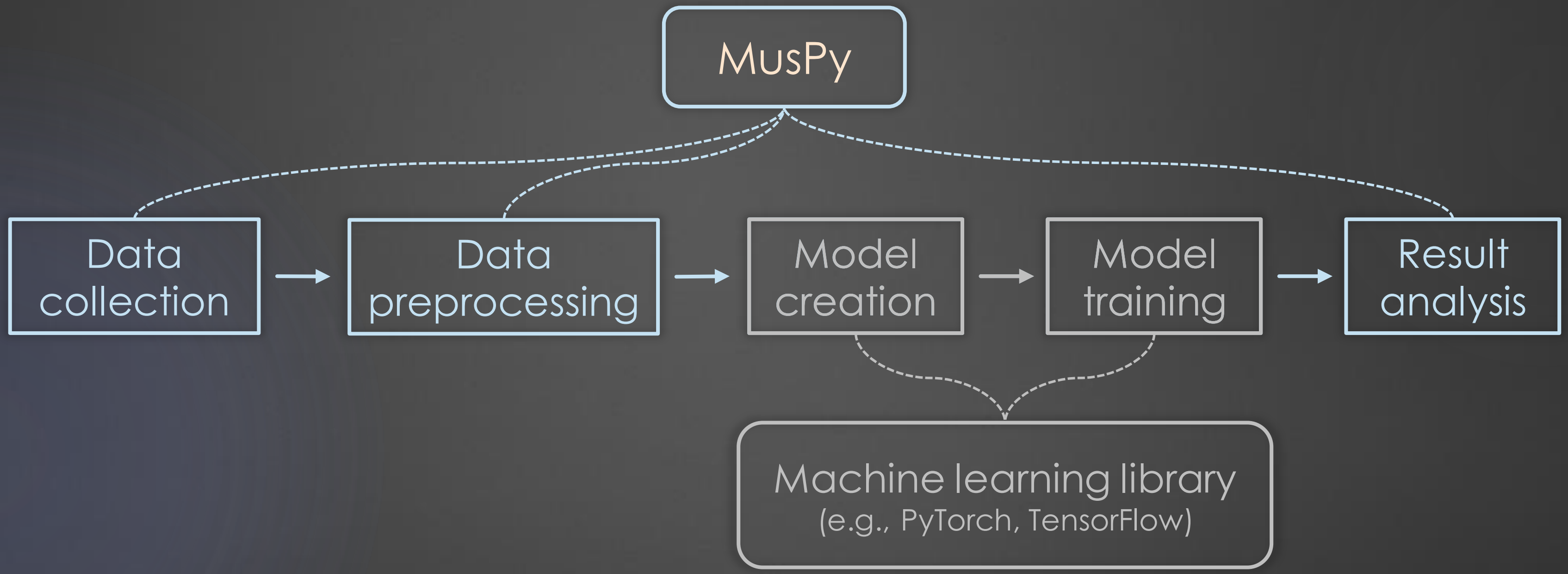
Preliminary results



Preliminary results

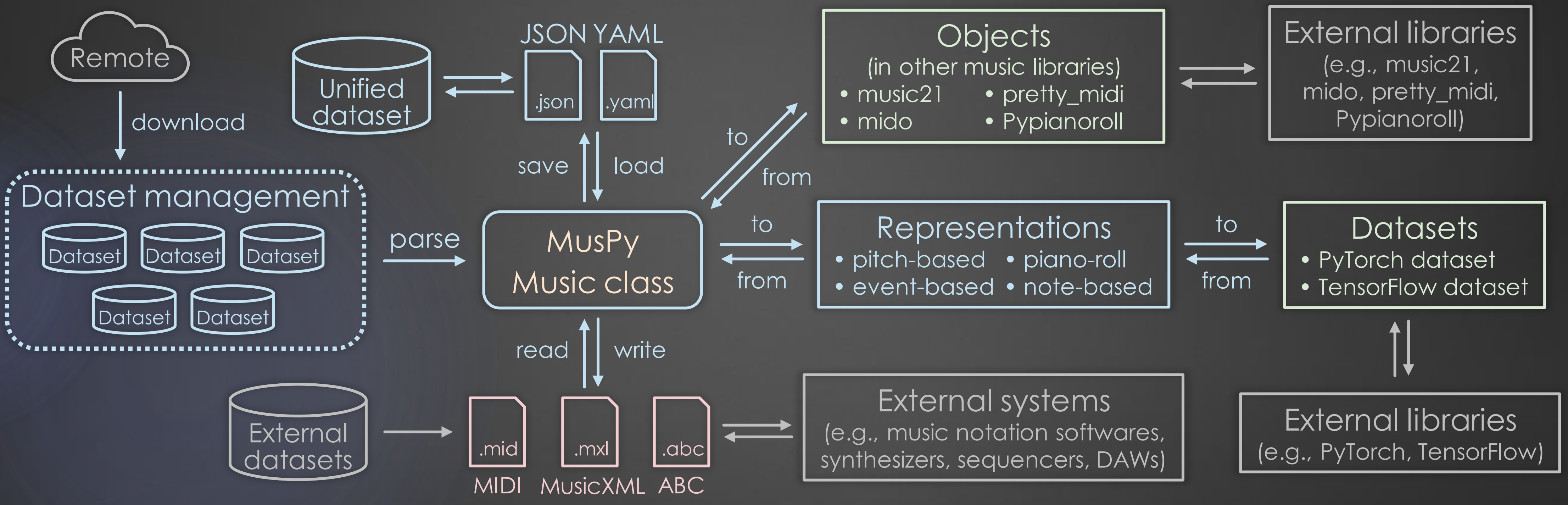


MusPy – Motivation



Dong et al., "MusPy: A Toolkit for Symbolic Music Generation," in *ISMIR*, 2020.

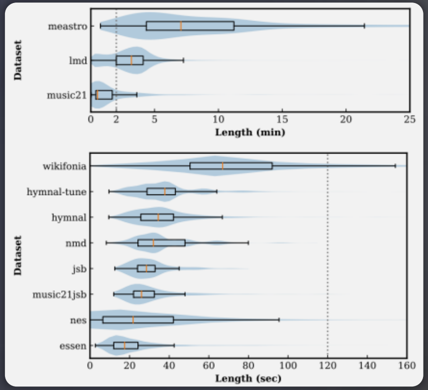
MusPy – Overview



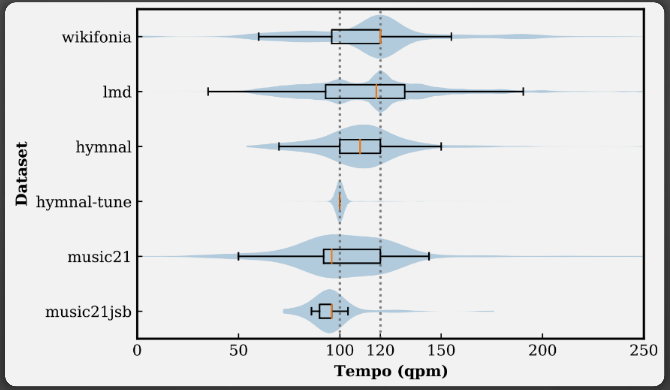
Dong et al., "MusPy: A Toolkit for Symbolic Music Generation," in *ISMIR*, 2020.

MusPy – Datasets

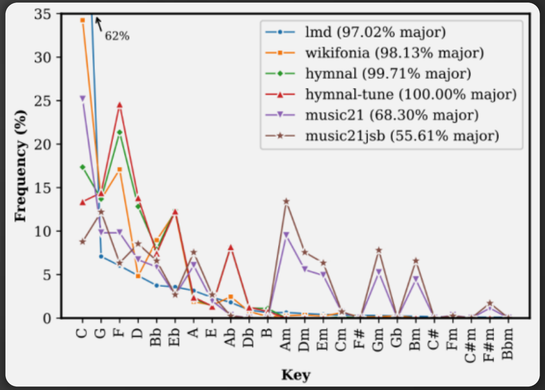
Dataset	Format	Hours	Songs	Genre	Melody	Chords	Multitrack
Lakh MIDI Dataset (LMD) [26]	MIDI	>9000	174,533	misc	△	△	△
MAESTRO Dataset [27]	MIDI	201.21	1,282	classical			
Wikifonia Lead Sheet Dataset [28]	MusicXML	198.40	6,405	misc	✓	✓	
Essen Folk Song Database [29]	ABC	56.62	9,034	folk	✓	✓	
NES Music Database [30]	MIDI	46.11	5,278	game	✓		✓
Hymnal Tune Dataset [31]	MIDI	18.74	1,756	hymn	✓		
Hymnal Dataset [31]	MIDI	17.50	1,723	hymn			
music21 Corpus [24]	misc	16.86	613	misc	△		△
Nottingham Database (NMD) [32]	ABC	10.54	1,036	folk	✓	✓	
music21 JSBach Corpus [24]	MusicXML	3.46	410	classical			✓
JSBach Chorale Dataset [11]	MIDI	3.21	382	classical			✓



Length distributions



Tempo distributions

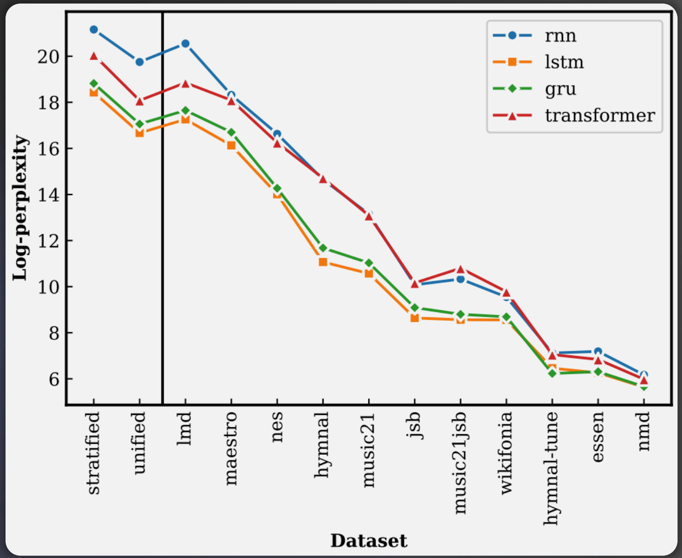


Key distributions

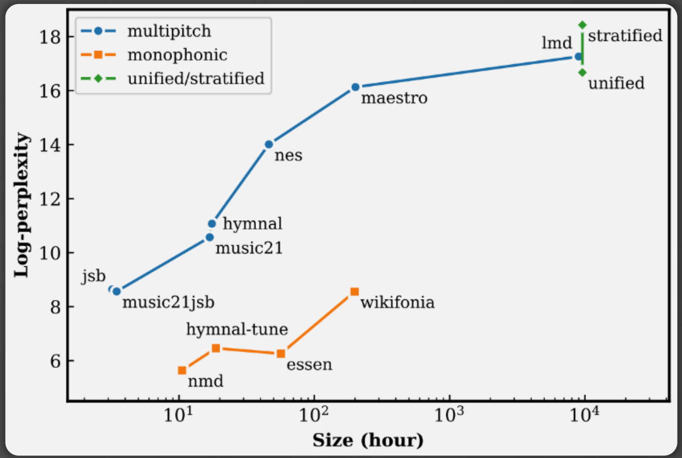
Dong et al., "MusPy: A Toolkit for Symbolic Music Generation," in *ISMIR*, 2020.

MusPy – Experiments

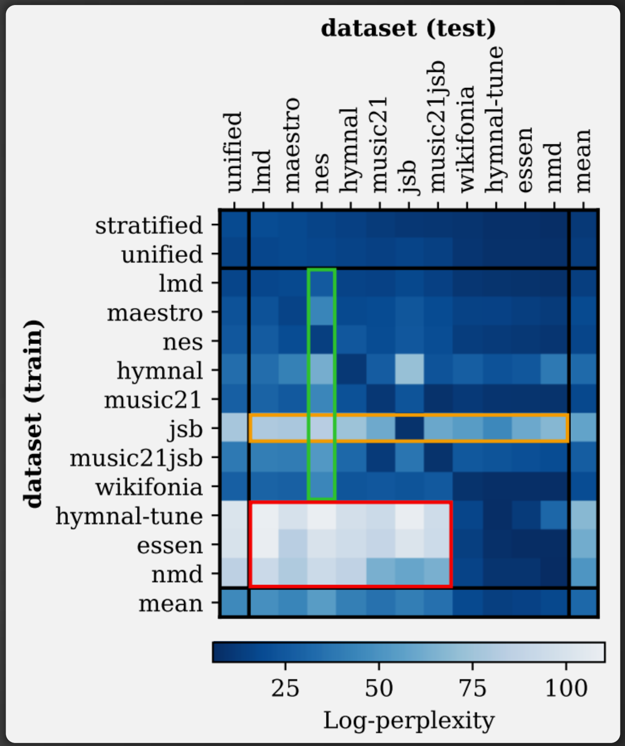
Perplexity



Perplexity vs dataset size

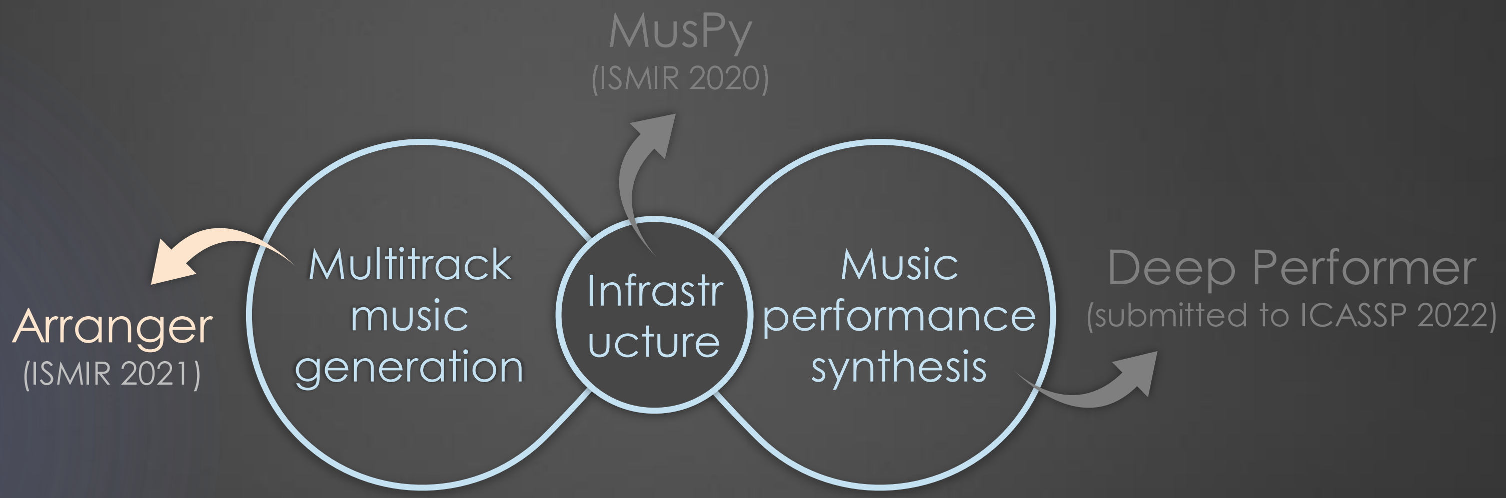


Cross-dataset generalizability



Dong et al., "MusPy: A Toolkit for Symbolic Music Generation," in *ISMIR*, 2020.

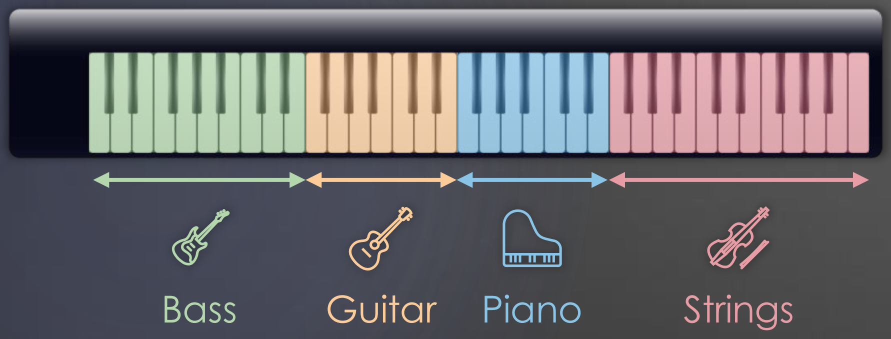
Preliminary results



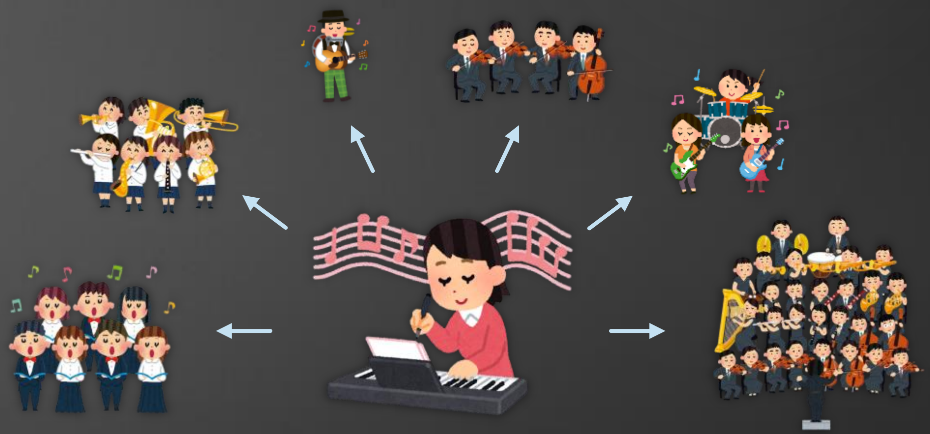
Arranger – Motivation

- Automatic instrumentation – Dynamically assign instruments to notes in solo music

Intelligent musical instruments



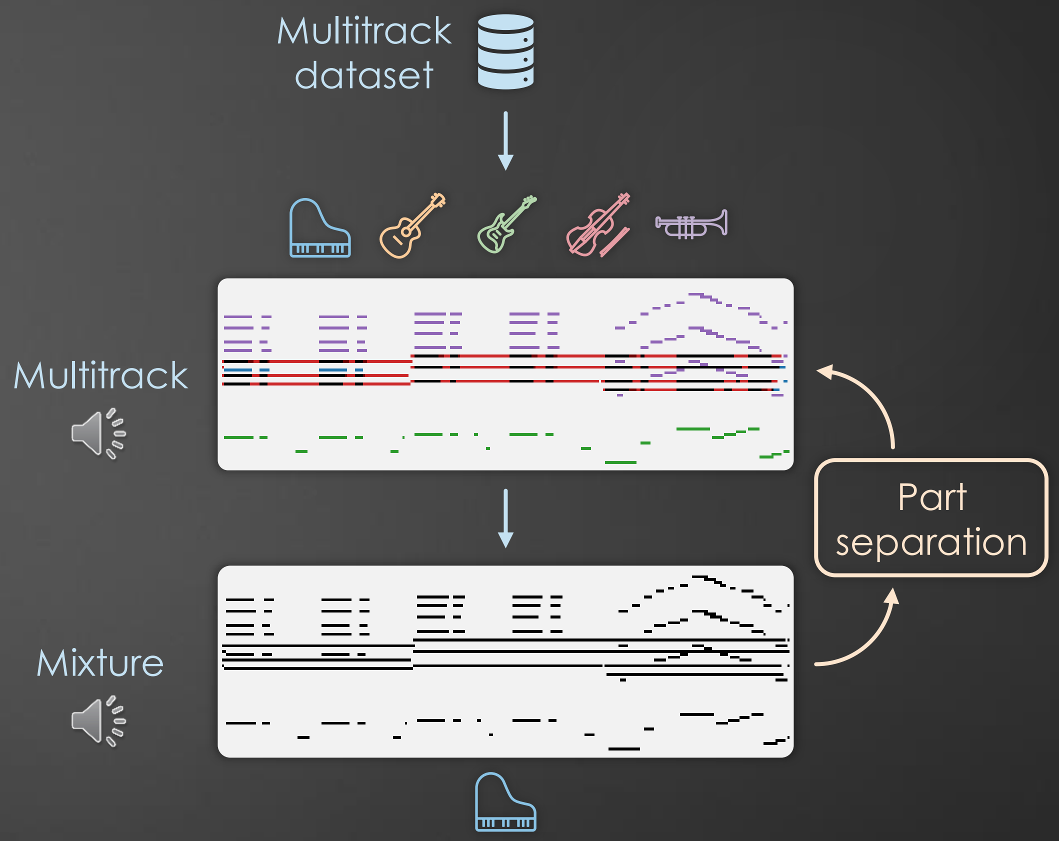
Assistive composing tools



Dong et al., "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," in *ISMIR*, 2021.

Arranger – Overview

1. Acquire paired data
2. Train a part separation model
3. Perform automatic instrumentation

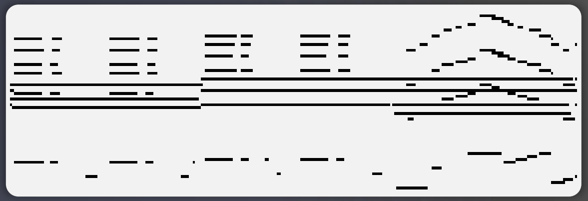


Dong et al., "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," in *ISMIR*, 2021.

Arranger – Problem formulation

- Part separation – Separate parts from their mixture in multitrack music
- Frame as a sequential multiclass classification problem

Mixture

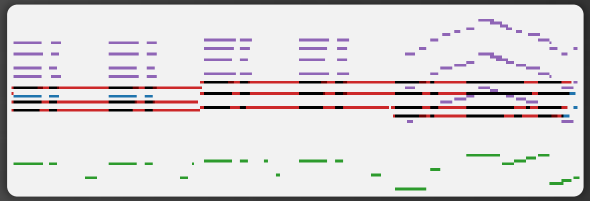


Notes

$(x_1, \ldots, x_N)$

→ Part separation model →

Multitrack



Part labels

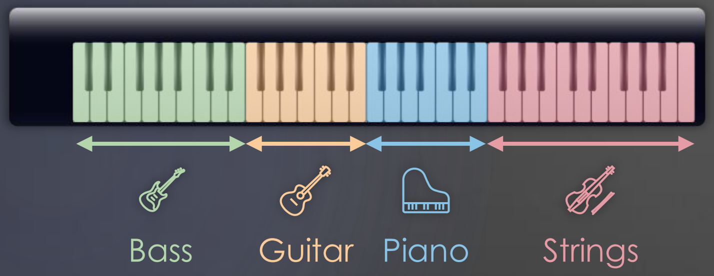
$(y_1, \ldots, y_N)$

Dong et al., "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," in *ISMIR*, 2021.

Arranger – Model

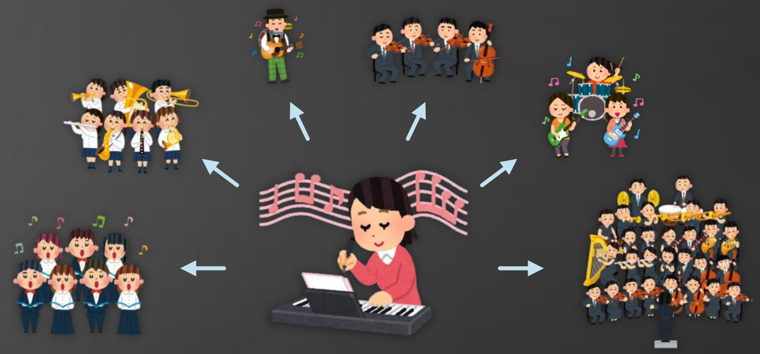
Online models

- LSTMs
- Transformer decoders



Offline models

- BiLSTMs
- Transformer encoders



Dong et al., "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," in *ISMIR*, 2021.

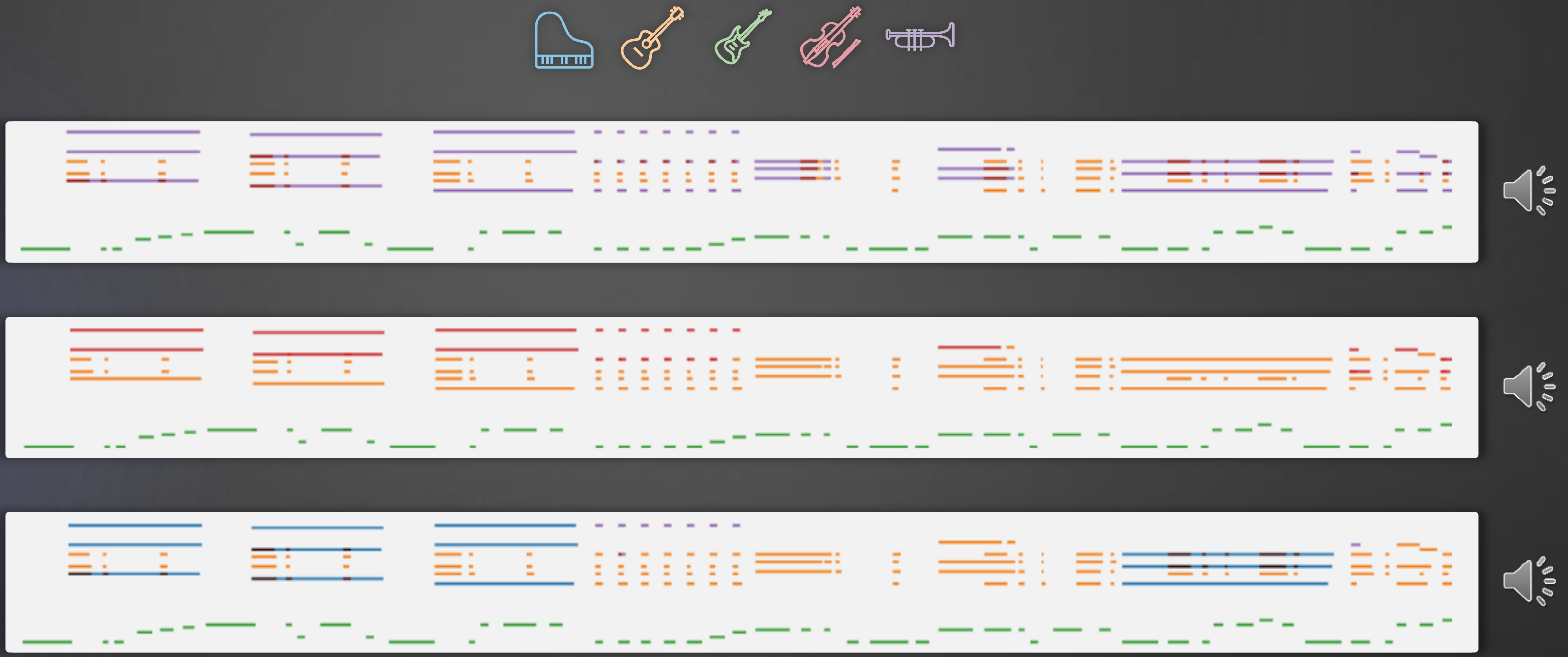
Arranger – Demo

- Produce convincing alternative instrumentations for an existing arrangement

Original

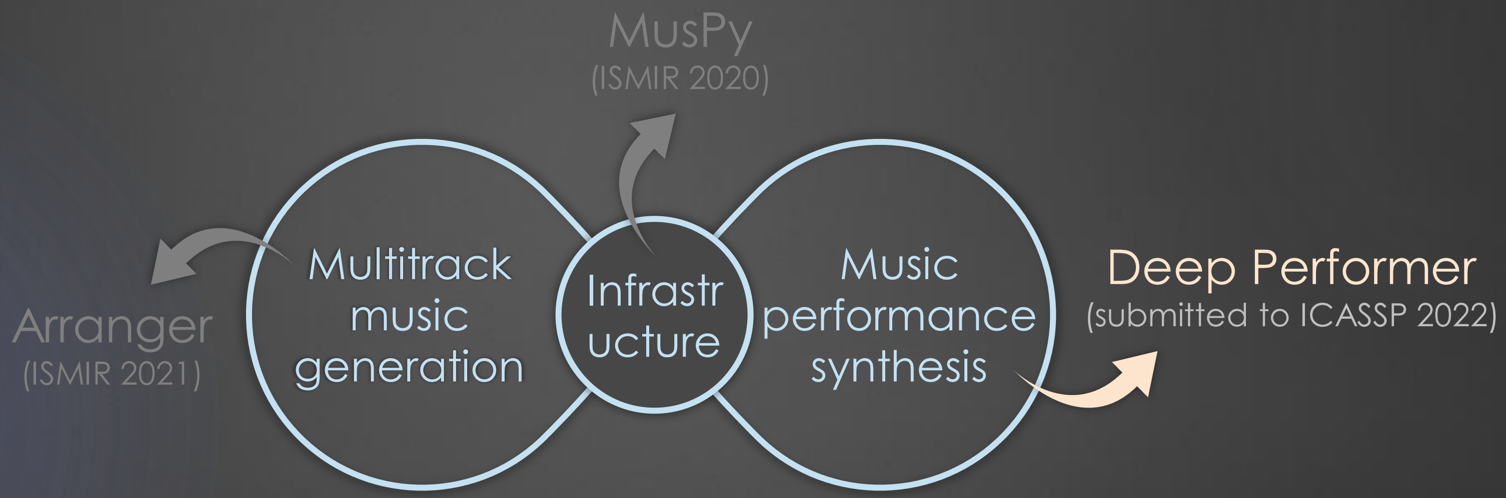
LSTM
(without entry hints)

BiLSTM
(with entry hints)

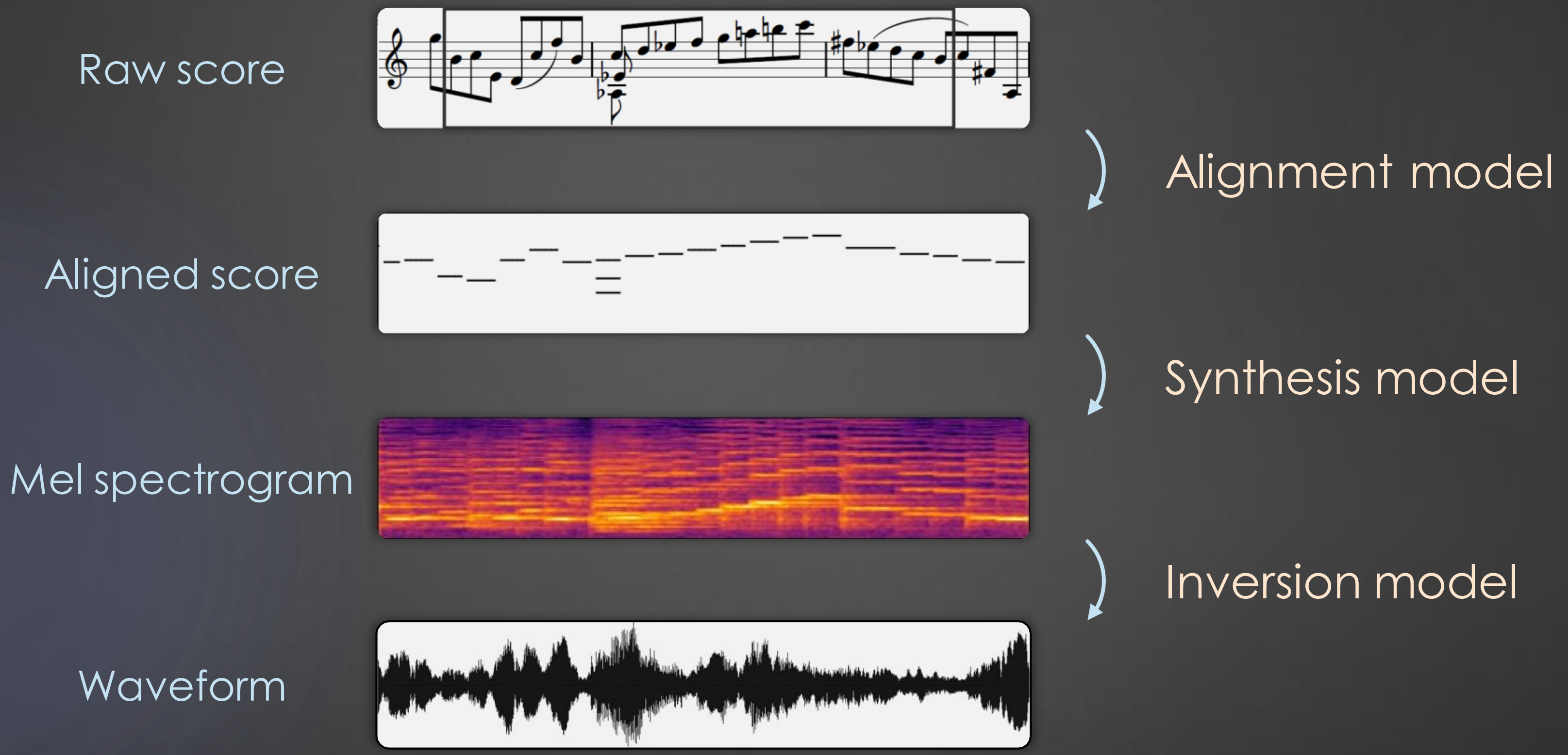


Dong et al., “Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music,” in *ISMIR*, 2021.

Preliminary results



Deep Performer – Overview



Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.

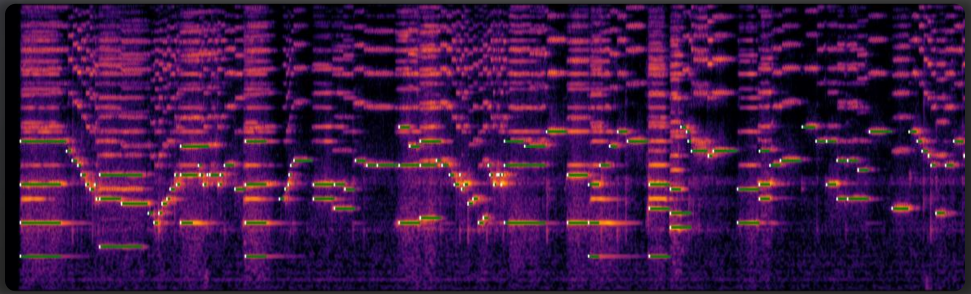
Deep Performer – Bach Violin Dataset

- Bach's sonatas and partitas for solo violin (BWV 1001–1006)
- 6.7 hours, 17 violinists

Data preparation

1. Synthesize the scores using FluidSynth (a free software synthesizer)
2. Run dynamic time warping on the spectrograms (synthesized audio & recording)

Alignment result

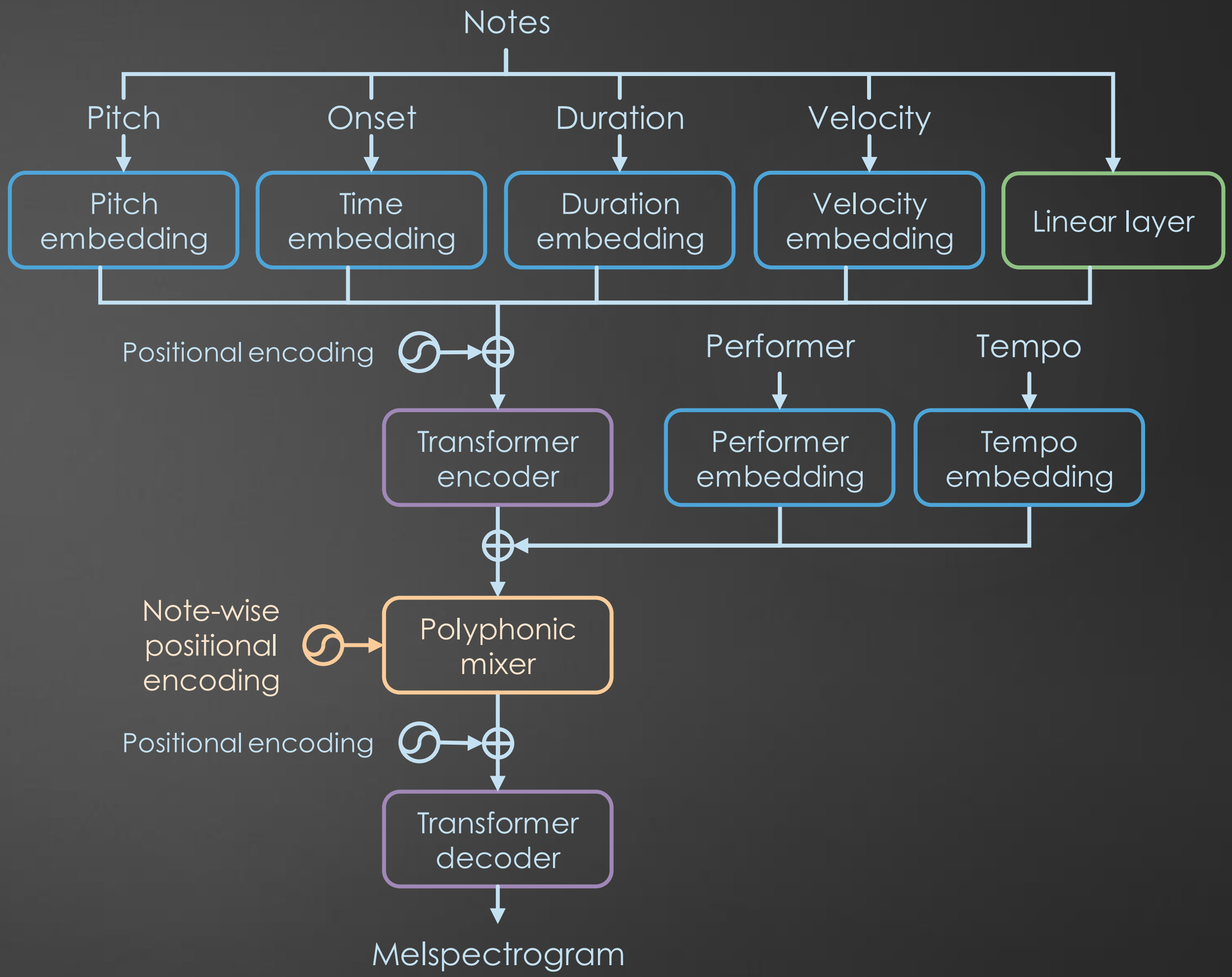


Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.

Deep Performer – Synthesis model

- A transformer network

  based on FastSpeech (Ren et al. 2019)



Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *NeurIPS*, 2019.
Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.

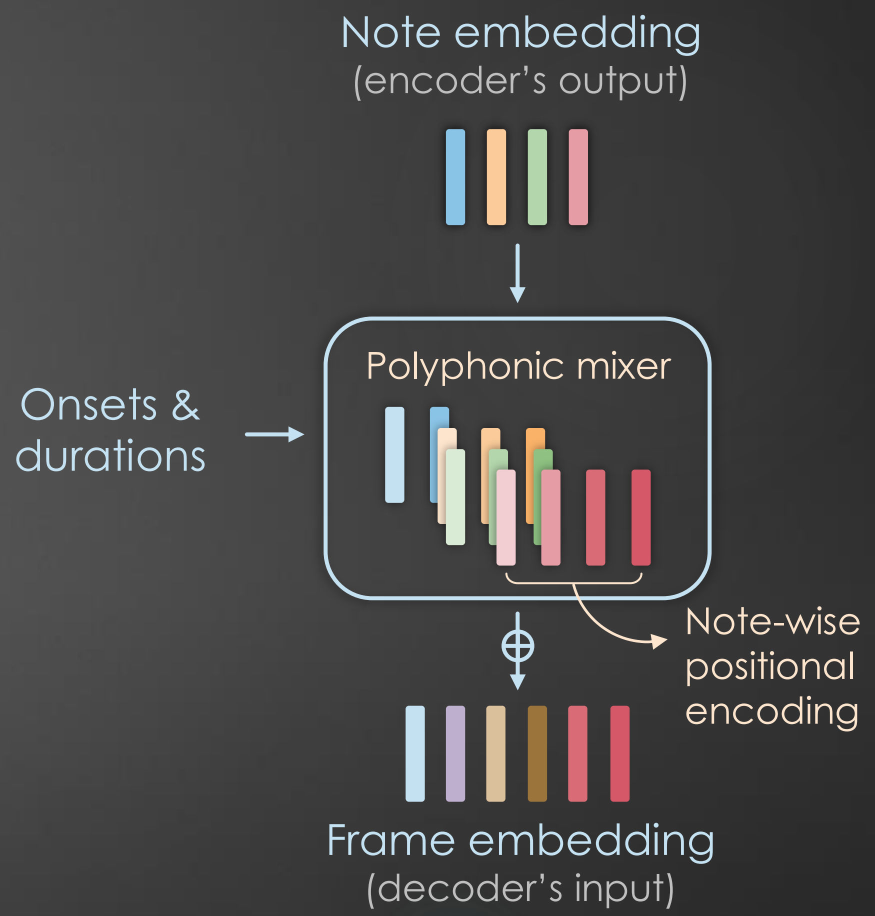
Deep Performer – Methods

- Polyphonic mixer

  Extend the state expansion mechanism to handle polyphonic inputs

- Note-wise positional encoding

  Provide positional information within each note for a fine-grained conditioning

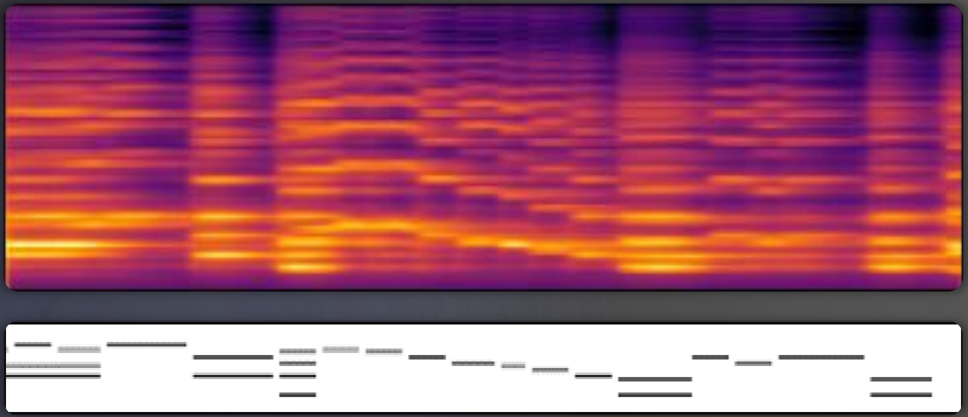


Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.

Deep Performer – Demo

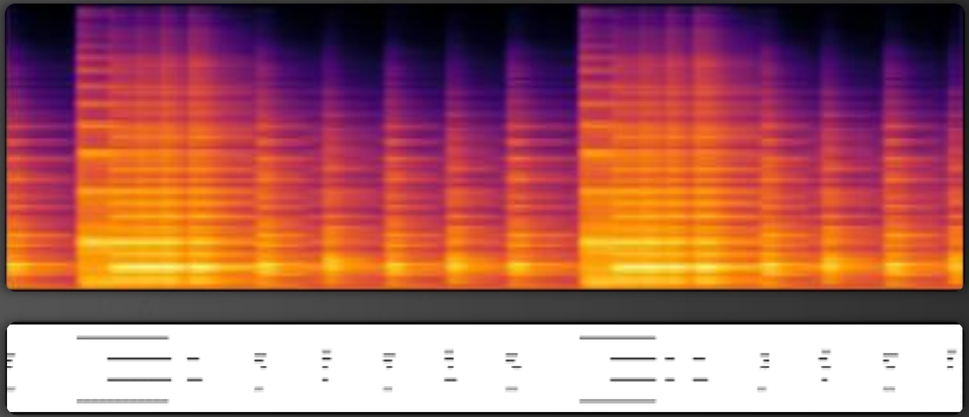
Violin

(trained on Bach Violin Dataset)



Piano

(trained on MAESTRO Dataset)

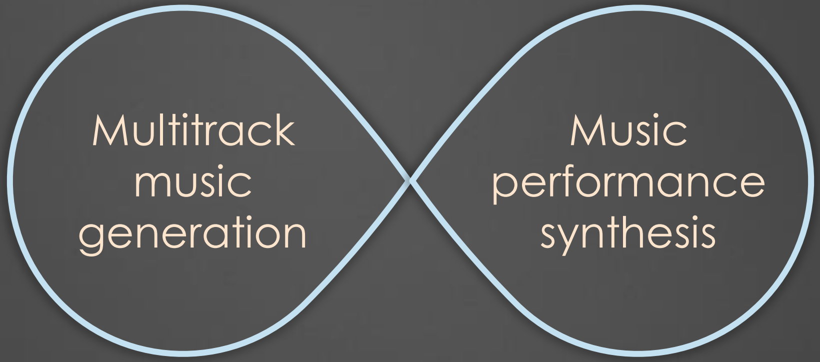


Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *ICLR*, 2019.
Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.

Future Work

Proposed research projects

Generate multitrack music from scratch or conditionally



Synthesize musical scores into expressive performance audio

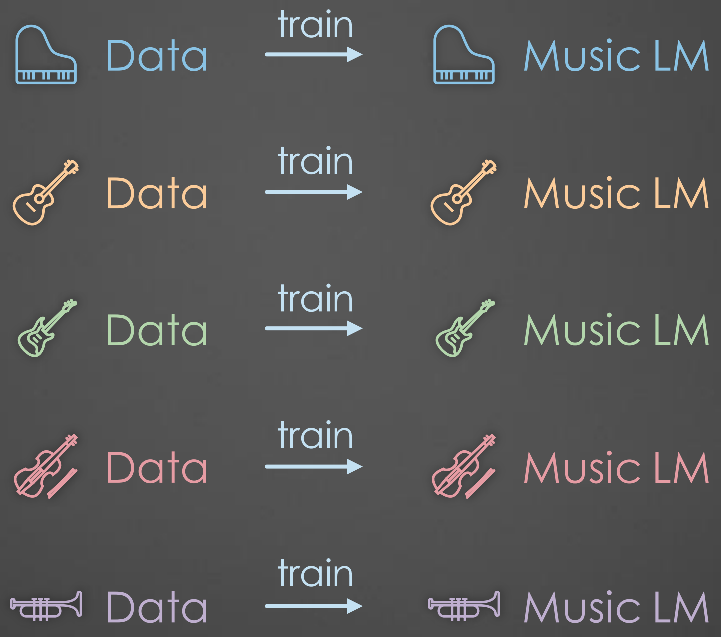
Multitrack music generation

- Data – MuseScore[1]
  - Over 1M musical scores in MusicXML format
  - Various ensembles and genres
  - Rich structural information (measures, repeats, jumps, etc.)

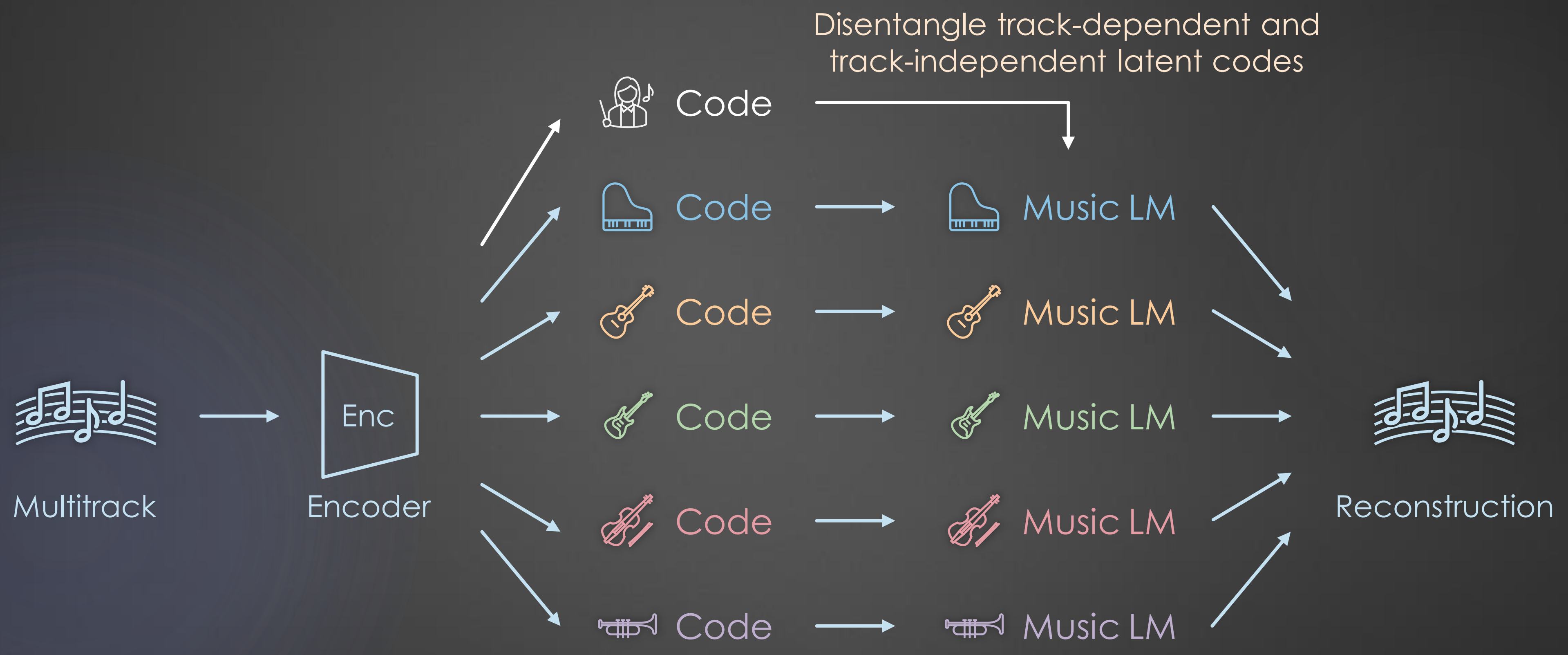
- Other relevant datasets
  - Lakh MIDI Dataset (Raffel et al. 2016)
  - MetaMIDI Dataset (Ens & Pasquier 2021)

1. MuseScore Forum, https://musescore.com/.
2. Raffel, "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching," *PhD Thesis*, 2016.
3. Ens and Pasquier, "Building the MetaMIDI Dataset: Linking Symbolic and Audio Musical Data," in *ISMIR*, 2021.

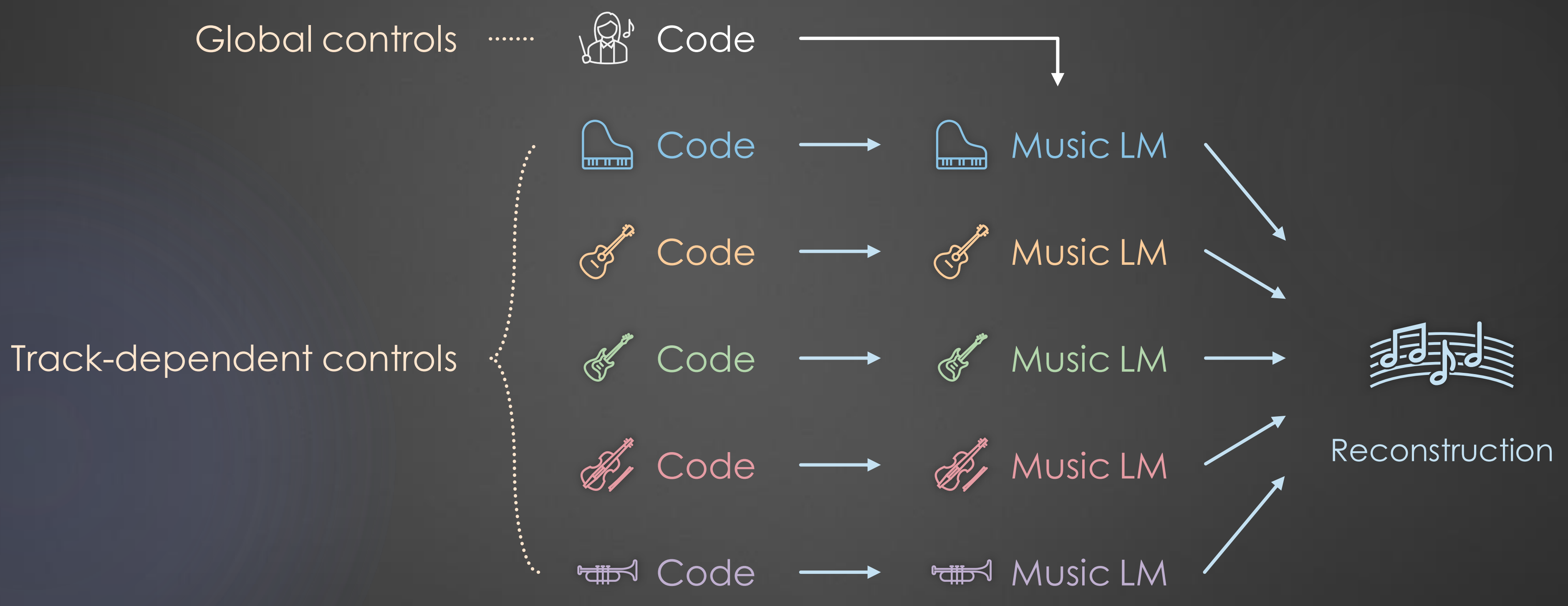
Instrument-specific music LMs



An autoencoder for multitrack music



Controllable multitrack generation



Music performance synthesis

- Data – Bach Violin Dataset (Dong et al. 2022)
  - 6.5 hours of recordings and musical scores
  - Bach's 6 *Sonatas and Partitas for Solo Violin*
  - Fine alignments (between the recordings and scores)
  - Various playing styles and recording setups (studios, recital halls, etc.)

- Other relevant datasets
  - URMP Dataset (Li et al. 2018)
  - MAESTRO Dataset (Hawthorne et al. 2018)

1. Dong et al., "Deep-Performer: Score-to-Audio Music Performance Synthesis," submitted to *ICASSP*, 2022.
2. Li et al., "Creating A Multi-track Classical Music Performance Dataset for Multi-modal Music Analysis: Challenges, Insights, and Applications," in *MM*, 2018.
3. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *ICLR*, 2019.

Modeling expressions & playing styles

Musical expressions

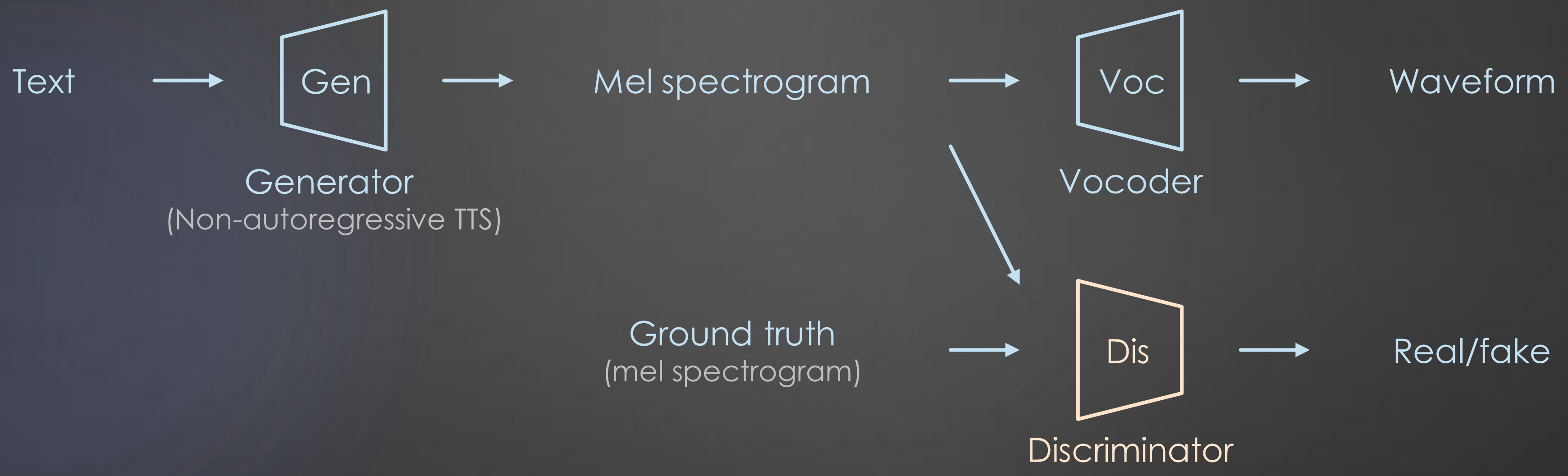
Dynamic, tempo, phrasing, articulation, etc.

Playing styles

Various musical interpretations of the same piece

Incorporating adversarial losses

- Improve the sharpness of the synthesized audio using adversarial losses
  - Promising results in speech synthesis (Yang et al. 2021)



1. Yang et al., "GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis," in *INTERSPEECH*, 2021.

Broader impacts

- Democratization of music creation
- Production of royalty-free music
- Applications in music education and therapy
- Insights into human-AI relationship

Timeline

Timeline

2022				2023				2024	
Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring

Thesis proposal

Multitrack music generation

Paper

Internship

Music performance synthesis

Paper

Internship

Integration

Paper

Thesis

Thesis defense

Thank you!