PAT 498/598 (Fall 2024)

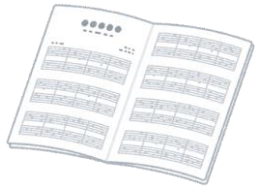
Special Topics: Generative AI for Music and Audio Creation

Lecture 13: Piano Roll-based Music Generation

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

(Recap) Four Paradigms



Symbolic music generation

- **Text-based**
  - MIDI

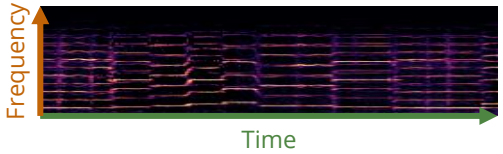
```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

- **Image-based**
  - Piano roll

Audio-domain music generation

- **Time series-based**
  - Waveform

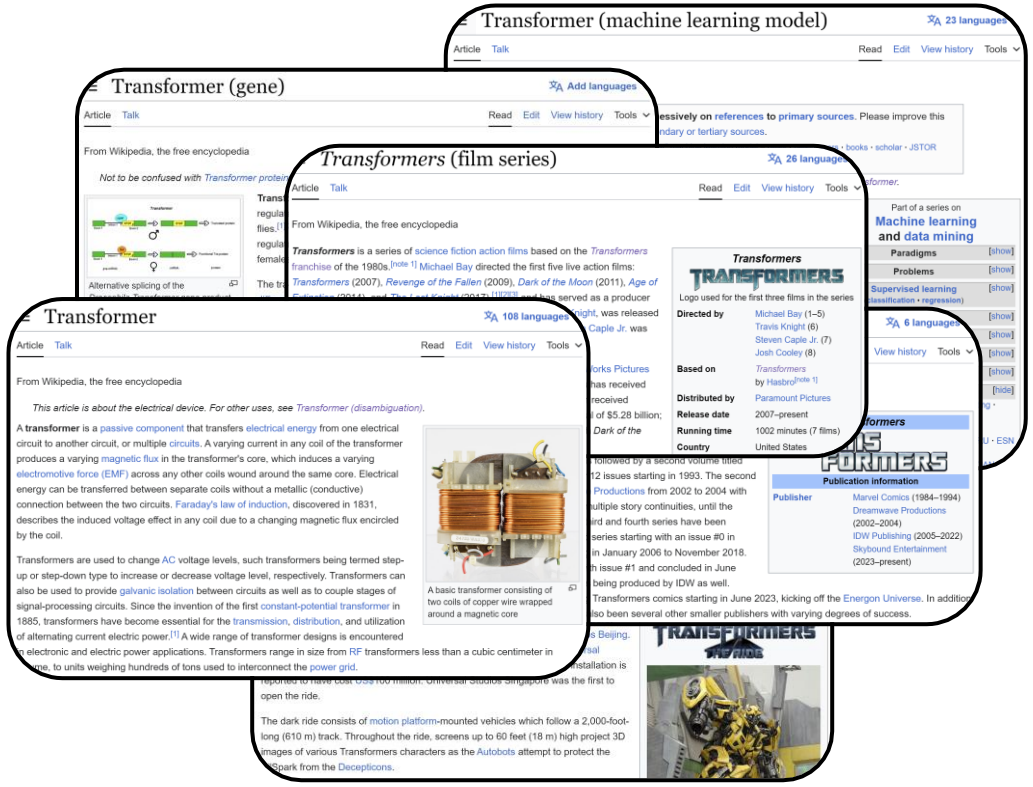
- **Image-based**
  - Spectrogram



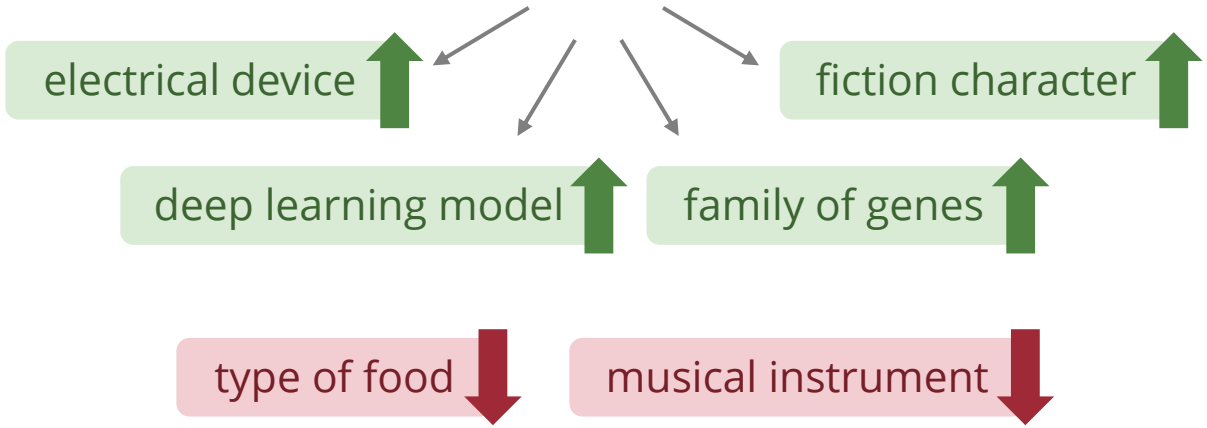
Today, we also have many latent-space based systems!

(Recap) Language Models

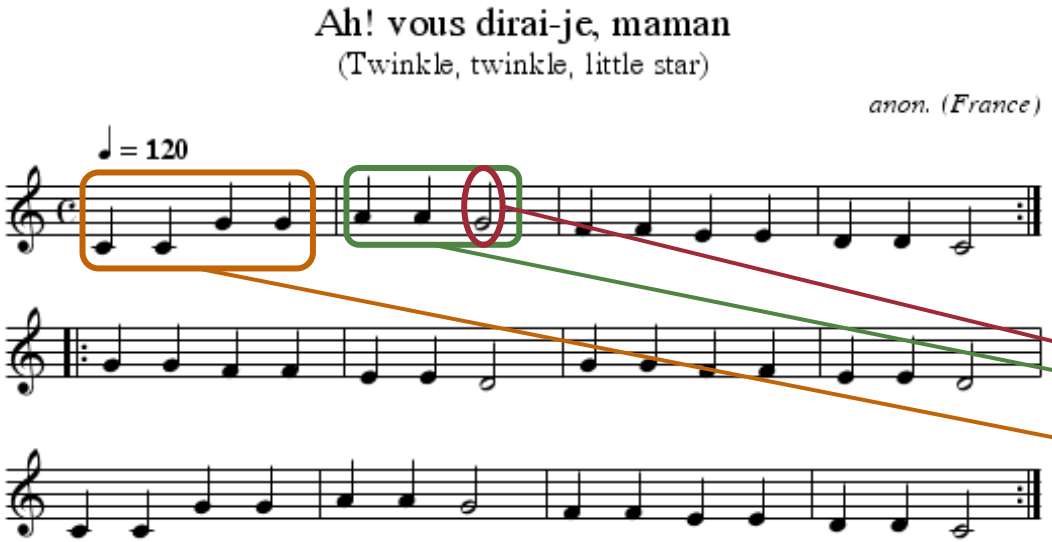
- Predicting the next word given the past sequence of words



A transformer is a _____



(Recap) An Example of ABC Notation



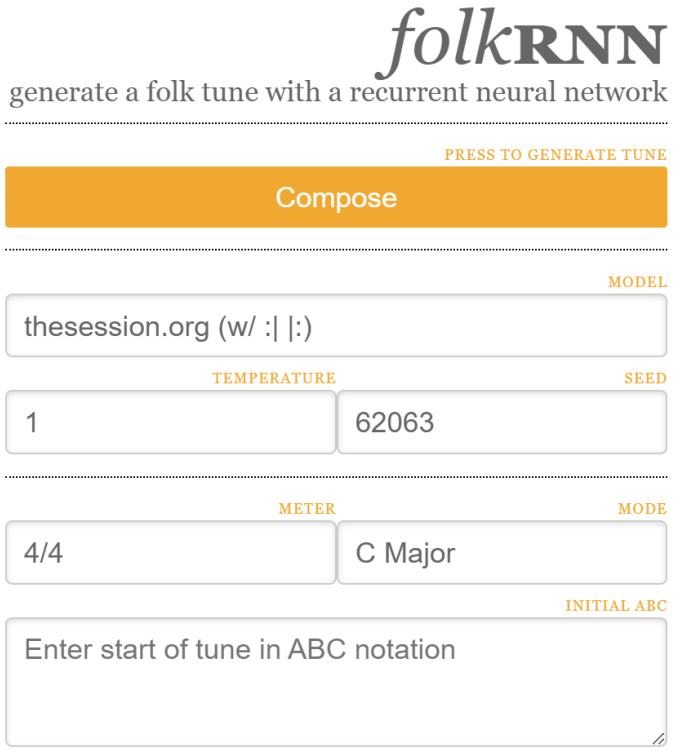
Metadata

```
X:571
T:Ah! vous dirai-je, maman
T:(Twinkle, twinkle, little star)
C:anon.
O:France
R:Nursery song
M:C  Meter
L:1/4  Unit note length (temporal resolution)
Q:120  Tempo
K:C  Key
CCGG|AAG2|FFEE|DDC2:|
|:GGFF|EED2|GGFF|EED2|
CCGG|AAG2|FFEE|DDC2:|
```

abcnotation.com/tunePage?a=trillian.mit.edu/~jc/music/abc/mirror/musicaviva.com/france/ah-vous-c/ah-vous-c-1/0000

(Recap) Example System: Folk RNN (Sturm et al., 2015)

- Data
  - Collections of folk tunes
- Representation
  - ABC notation without metadata
- Model
  - LSTM (long short-term memory)
  - Working on the character level

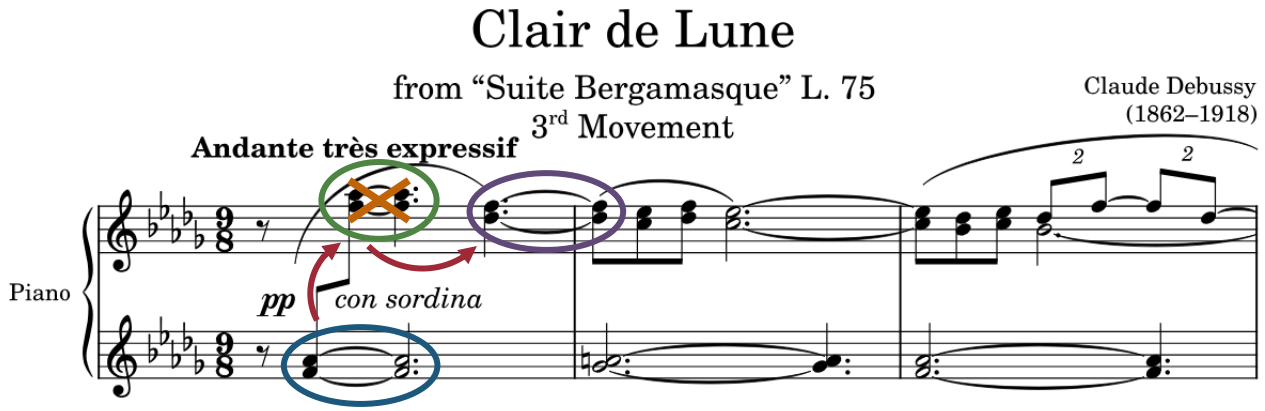


folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

(Recap) Representing Polyphonic Music

- We can now handle music with multi-pitch at the same time
  - In the literature, "polyphonic" & "multi-pitch" are often used interchangeably




```
Note_on_65, Note_on_68, Time_shift_eighth_note, Note_on_77, Note_on_80,
Time_shift_half_note, Note_off_77, Note_off_80, Note_on_73, Note_on_77,
Time_shift_dotted_quarter_note, Note_off_65, Note_off_68, ...
```

(Recap) Example: Performance RNN (Oore et al., 2020)

- Data
  - Yamaha e-Piano Competition dataset (MAESTRO)
- Representation
  - 128 Note-On events
  - 128 Note-Off events
  - 125 Time-Shift events (8ms–1s)
  - 32 Set-Velocity events — Handle dynamics
- Model
  - LSTM

Examples of generated music



Ian Simon and Sageev Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics," *Magenta Blog*, June 29, 2017.
Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan, "This Time with Feeling: Learning Expressive Musical Performance", *Neural Computing and Applications*, 32, 2020.

(Recap) Example: Music Transformer (Huang et al., 2019)

- Data
  - Yamaha e-Piano Competition dataset (MAESTRO)
- Representation
  - 128 Note-On events
  - 128 Note-Off events
  - 100 Time-Shift events (10ms–1s)
  - 32 Set-Velocity events
- Model
  - Transformer

Almost the same representation as PerformanceRNN

Handle dynamics

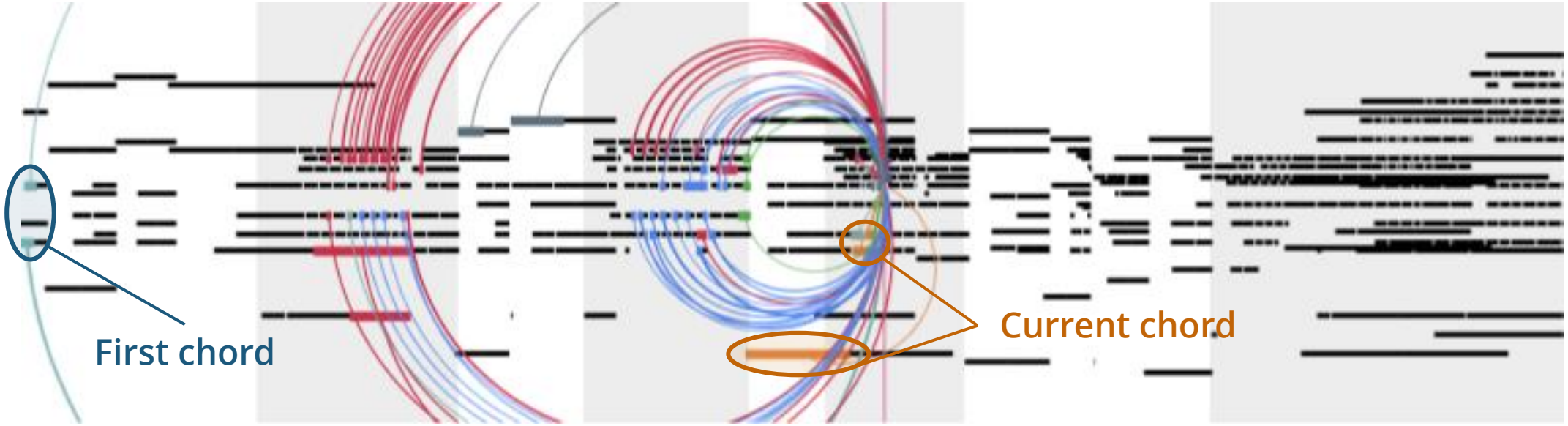
Examples of generated music



Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

(Recap) Visualizing Musical Self-attention

(Each color represents an attention head)



(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

(Recap) Example: MuseNet (Payne et al., 2019)

- Data
  - ClassicalArchives + BitMidi + MAESTRO
- Representation
  - Notes are represented as a compound word in the form of "instrument:velocity:pitch"
  - Time shifts in real time (sec)
- Model
  - Transformer

Example of generated music

```
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
...
```

Christine Payne, "MuseNet," *OpenAI*, 2019.

(Recap) Example: Multitrack Music Transformer (Dong et al., 2023)

- Data
  - Symbolic Orchestral Database (SOD)
- Representation
  - Notes are represented as a six-value tuple: (beat, position, pitch, duration, instrument)
  - No time shift events (Why?)
- Model
  - Multi-dimensional Transformer

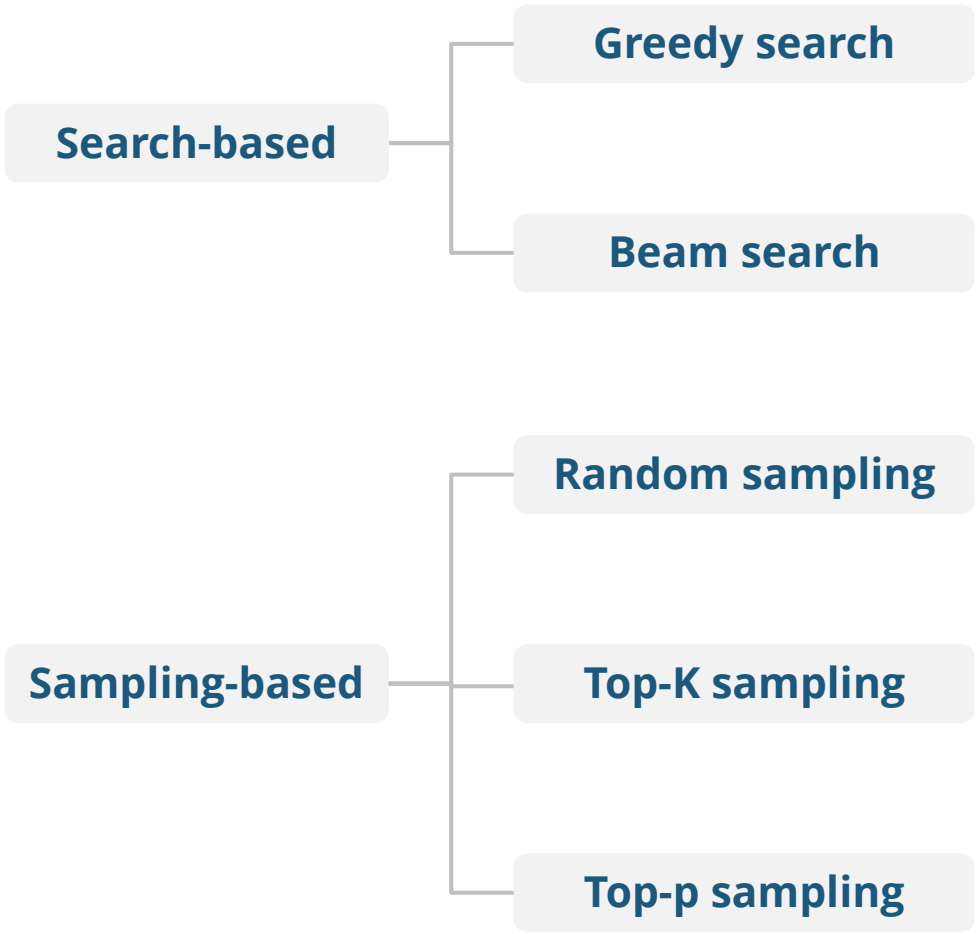
Example of generated music

```
(0,  0,  0,  0,  0,  0)   Start of song
(1,  0,  0,  0,  0, 15)   Instrument: accordion
(1,  0,  0,  0,  0, 36)   Instrument: trombone
(1,  0,  0,  0,  0, 39)   Instrument: brasses
(2,  0,  0,  0,  0,  0)   Start of notes
(3,  1,  1, 41, 15, 36)   Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3,  1,  1, 65,  4, 39)   Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3,  1,  1, 65, 17, 15)   Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3,  1,  1, 68,  4, 39)   Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3,  1,  1, 68, 17, 15)   Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3,  1,  1, 73, 17, 15)   Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3,  1, 13, 68,  4, 39)   Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3,  1, 13, 73,  4, 39)   Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3,  2,  1, 73, 12, 39)   Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3,  2,  1, 77, 12, 39)   Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
        ...               ...
(4,  0,  0,  0,  0,  0)   End of song
```

(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

(Recap) Decoding Strategies



Pick **the most probable word** at each step

Pick **the most probably word sequences** jointly considering the top-K probable sequences so far

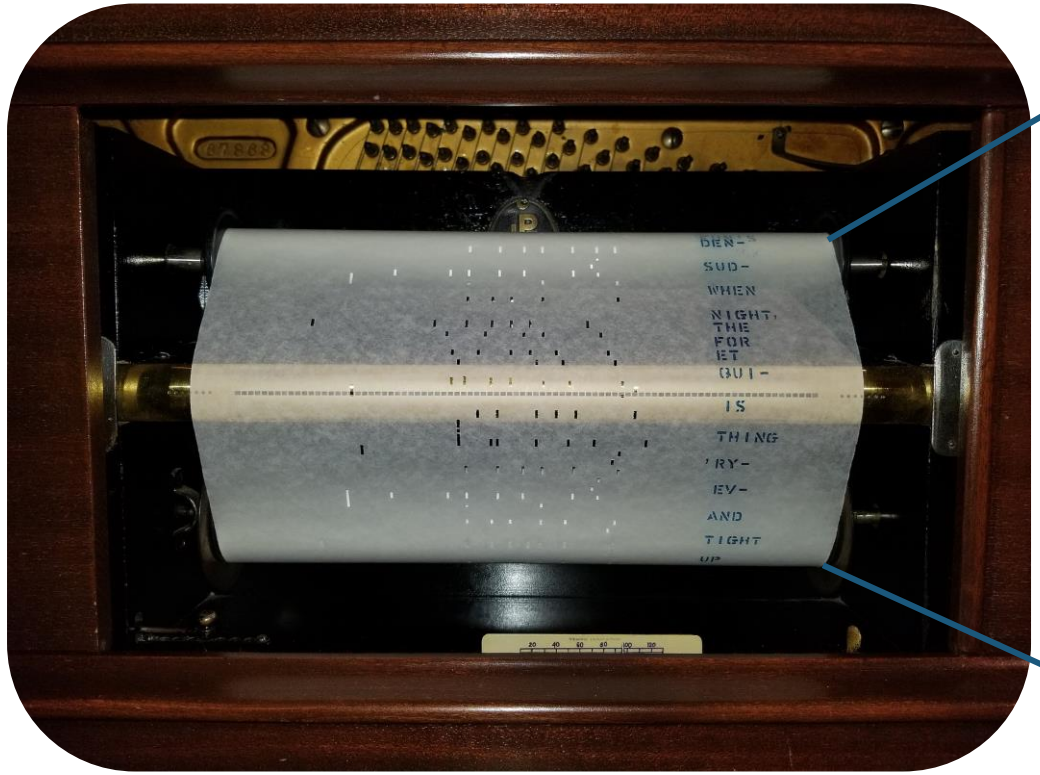
Sample randomly from the **raw output distribution**

Sample randomly from the output distribution considering **only the top-K probable words**

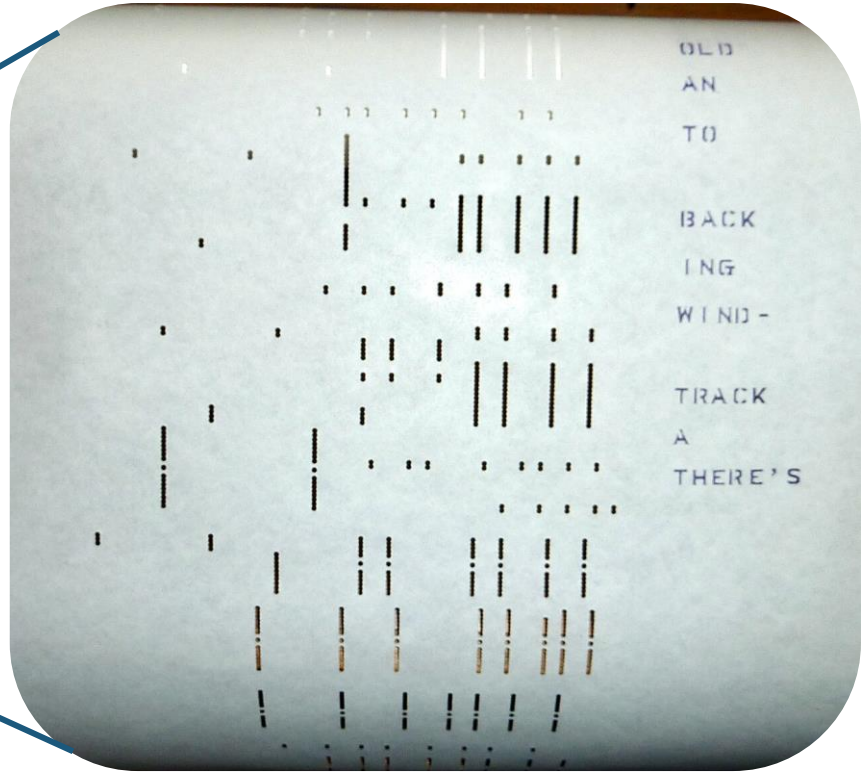
Sample randomly from the output distribution considering **only the top several probable words that make up the top p% of the distribution**

Piano Roll Representation

Piano Rolls



(Source: Draconichiaro)

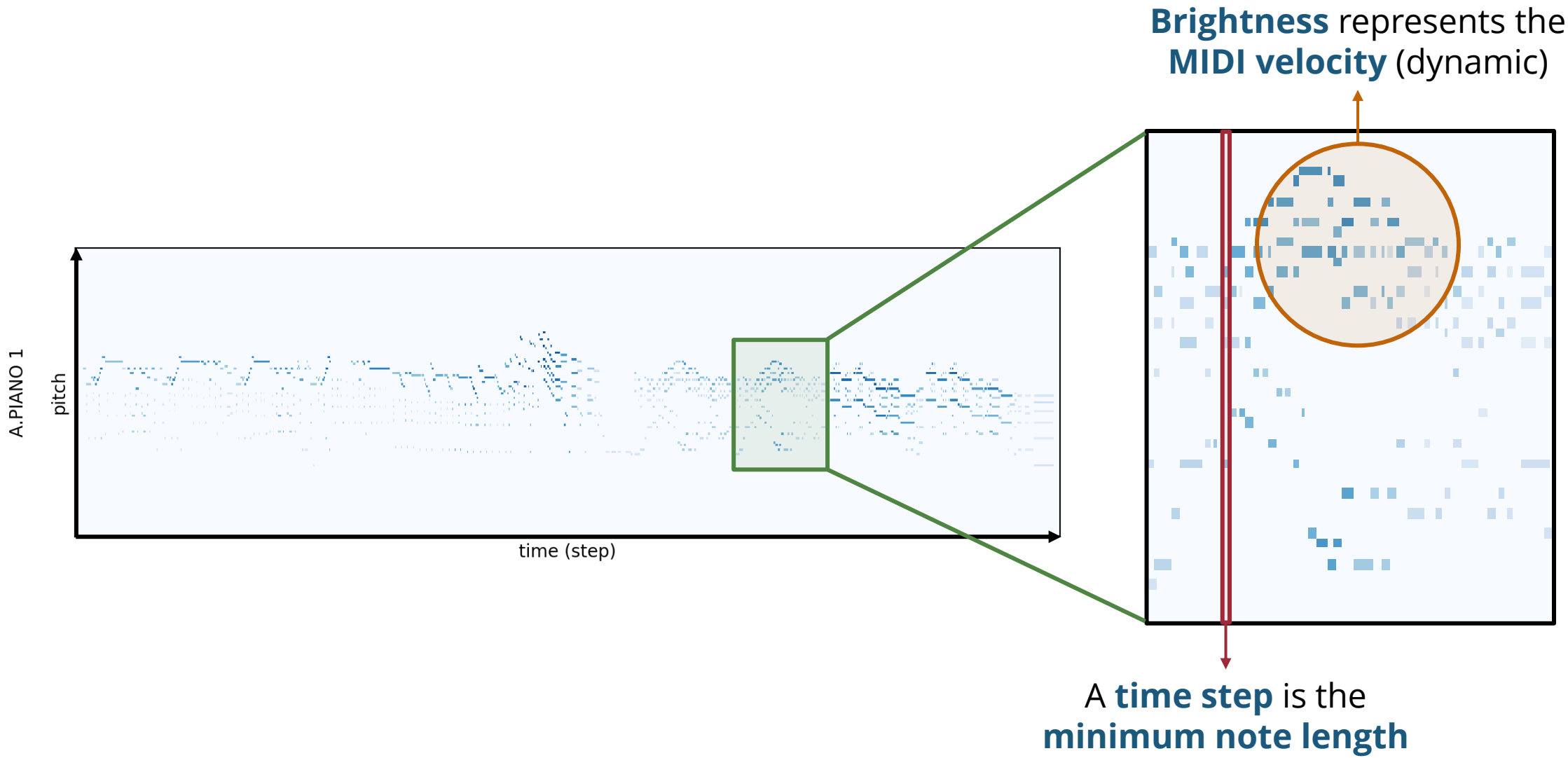


(Source: Tangerineduel)

Player Pianos

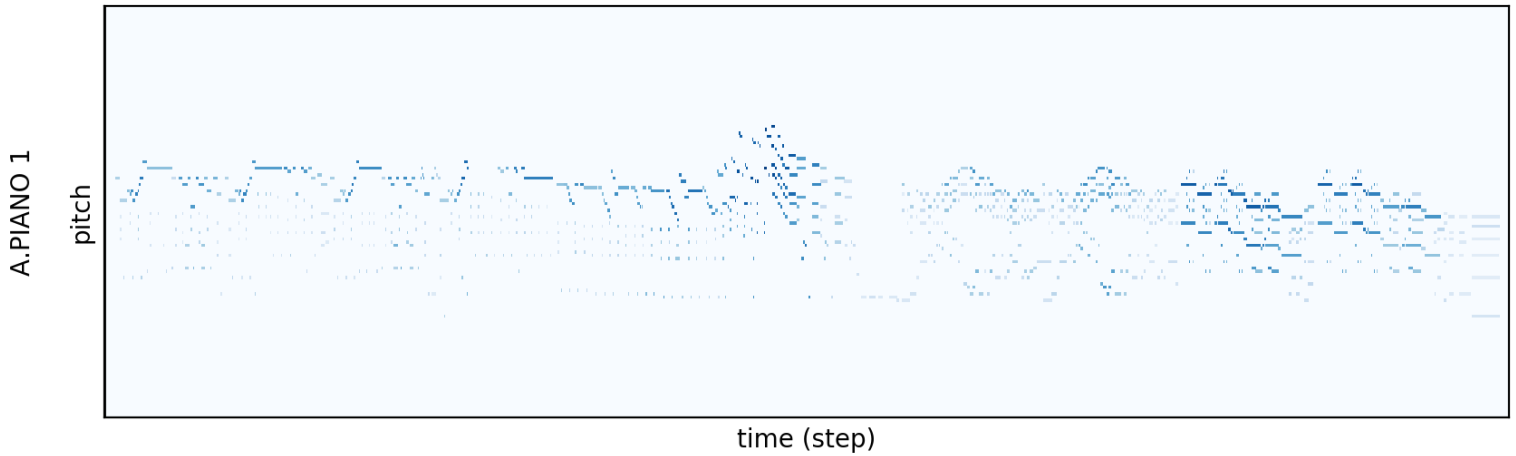
youtu.be/07krQ661fok

Piano Roll Representation

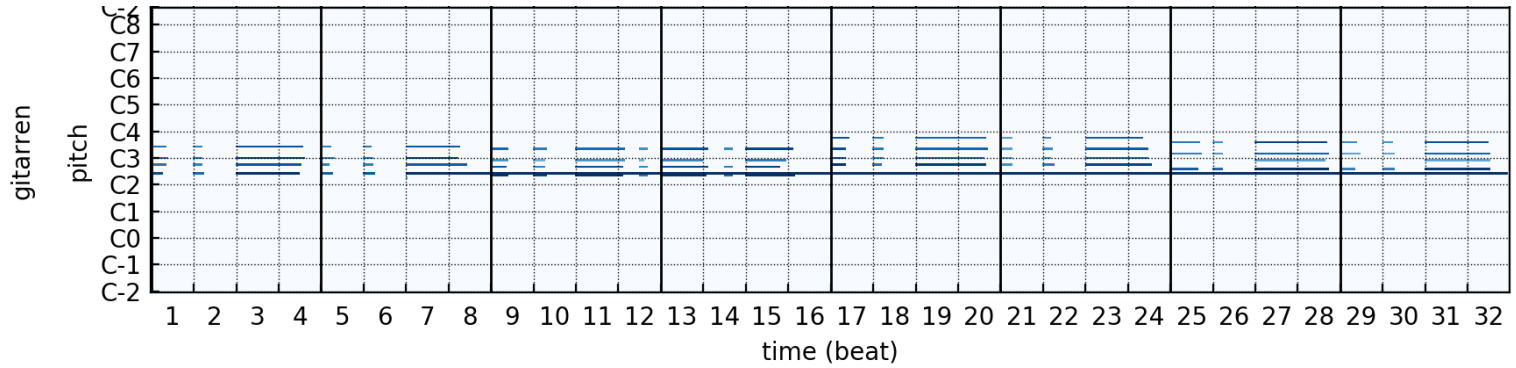


Piano Roll Representation

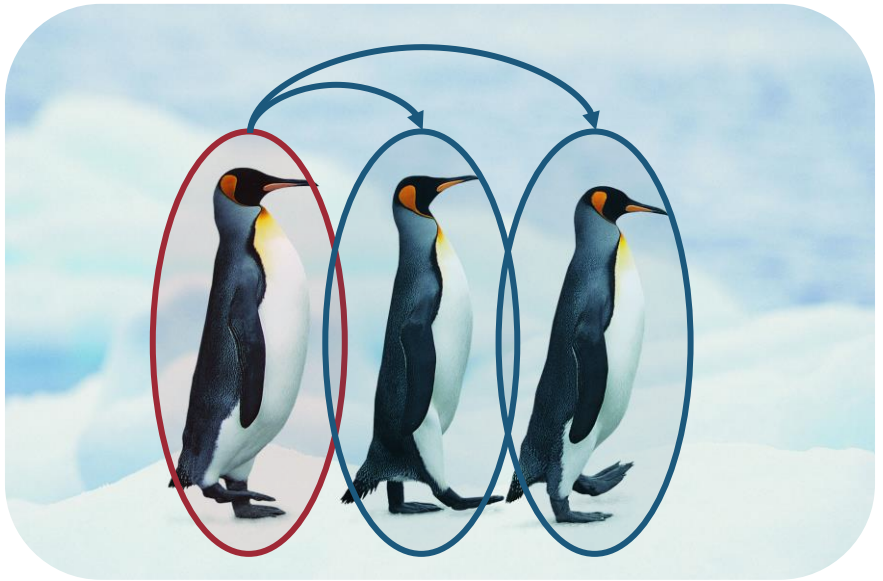
With expressive timing



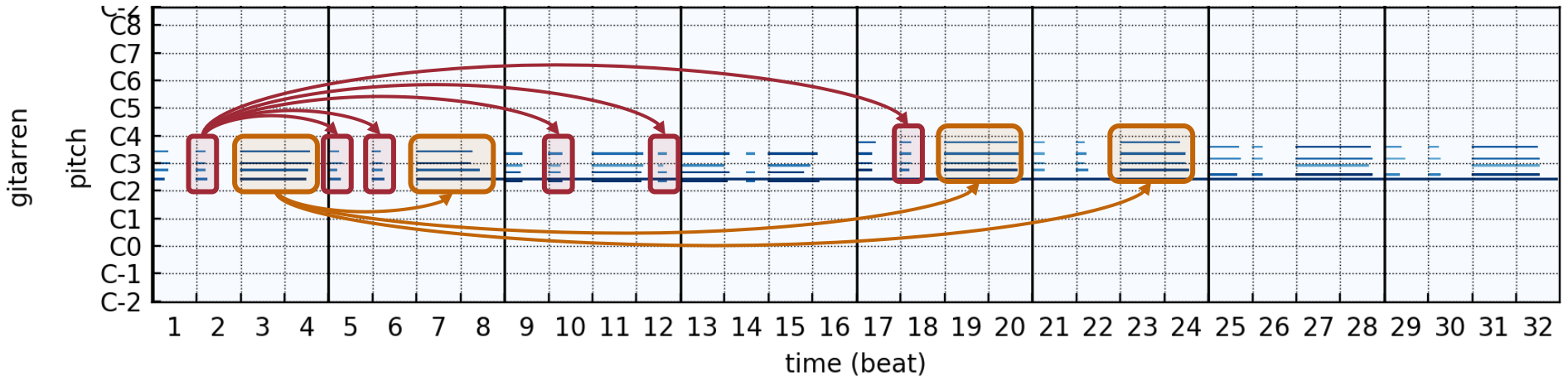
Without expressive timing



(Recap) Reusable Pattern Detectors



Why Piano Rolls?

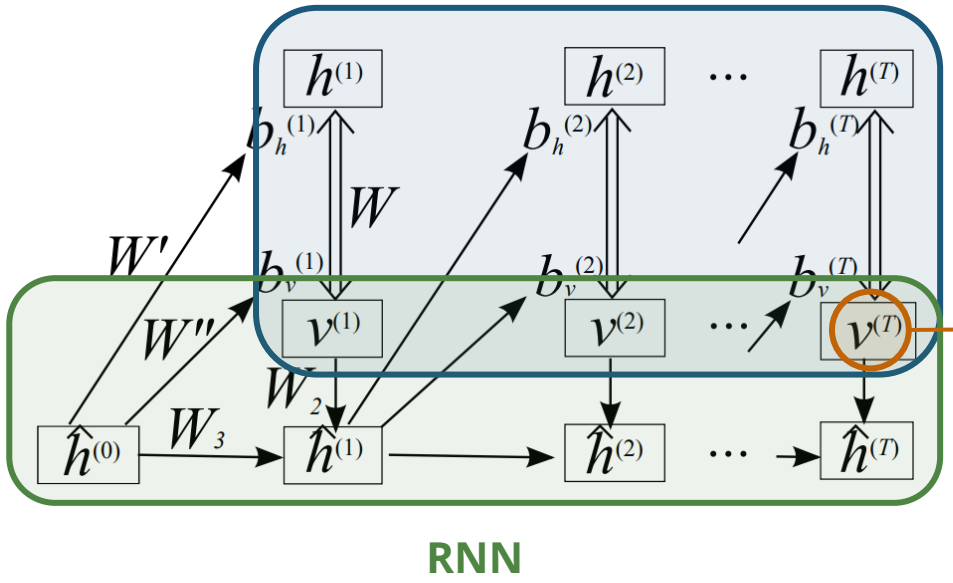


Many musical patterns like melodies, chords, scales and arpeggios are translational invariant in the temporal and pitch axes

Music Generation using GANs

RNN-RBM (Boulanger-Lewandowski et al., 2012)

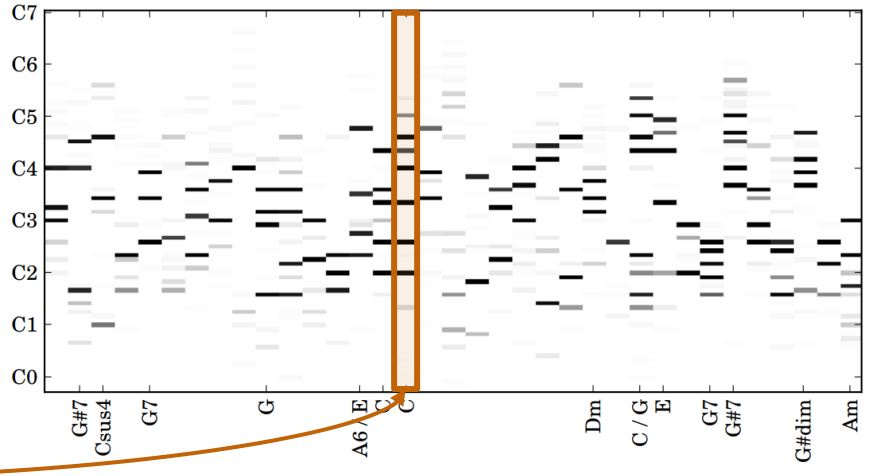
Restricted Boltzmann machine (RBM)



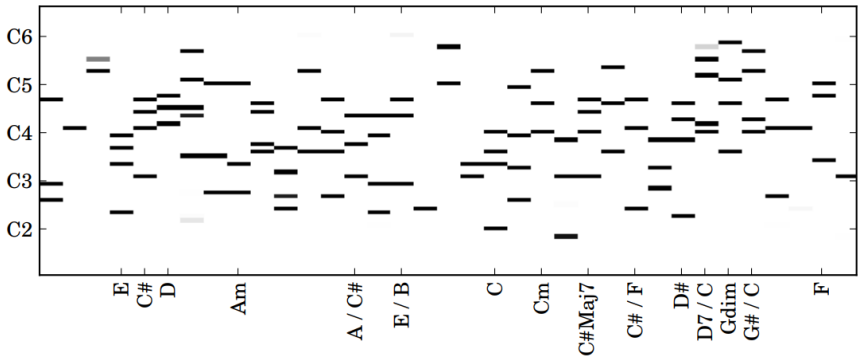
RNN

(Source: Boulanger-Lewandowski et al., 2012)

Piano-midi dataset

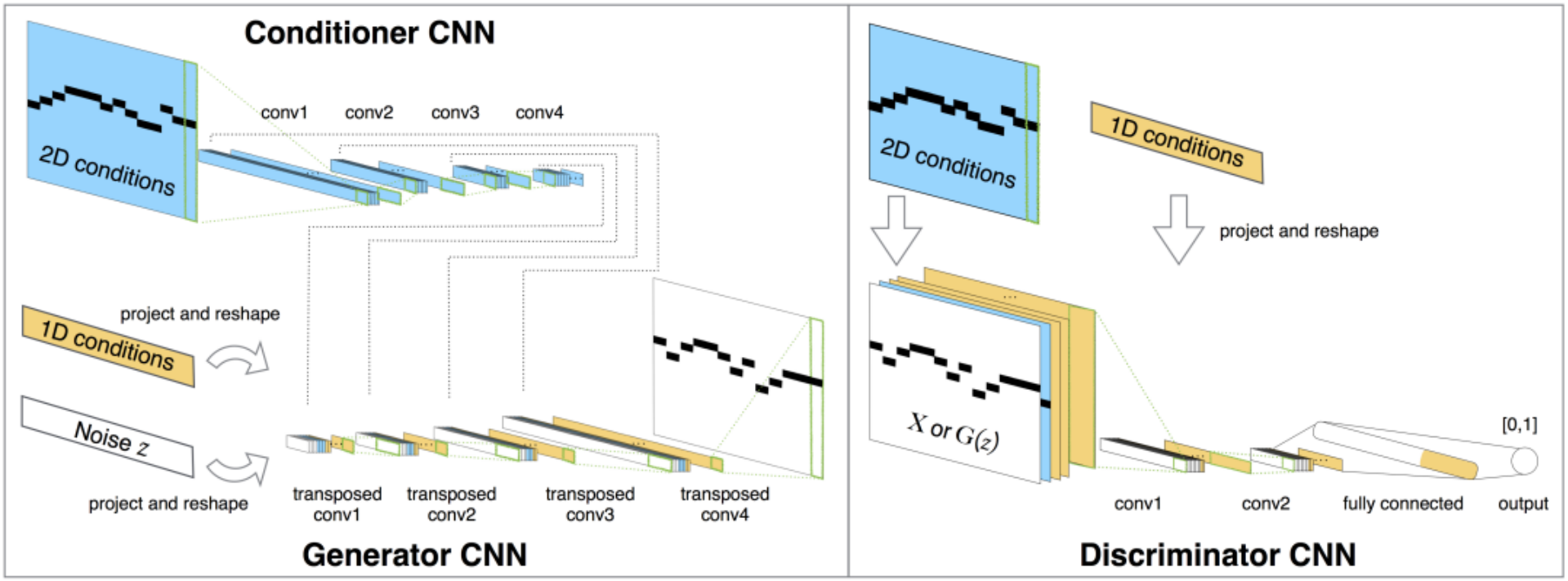


JS Bach chorales



Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription," *ICML*, 2022.

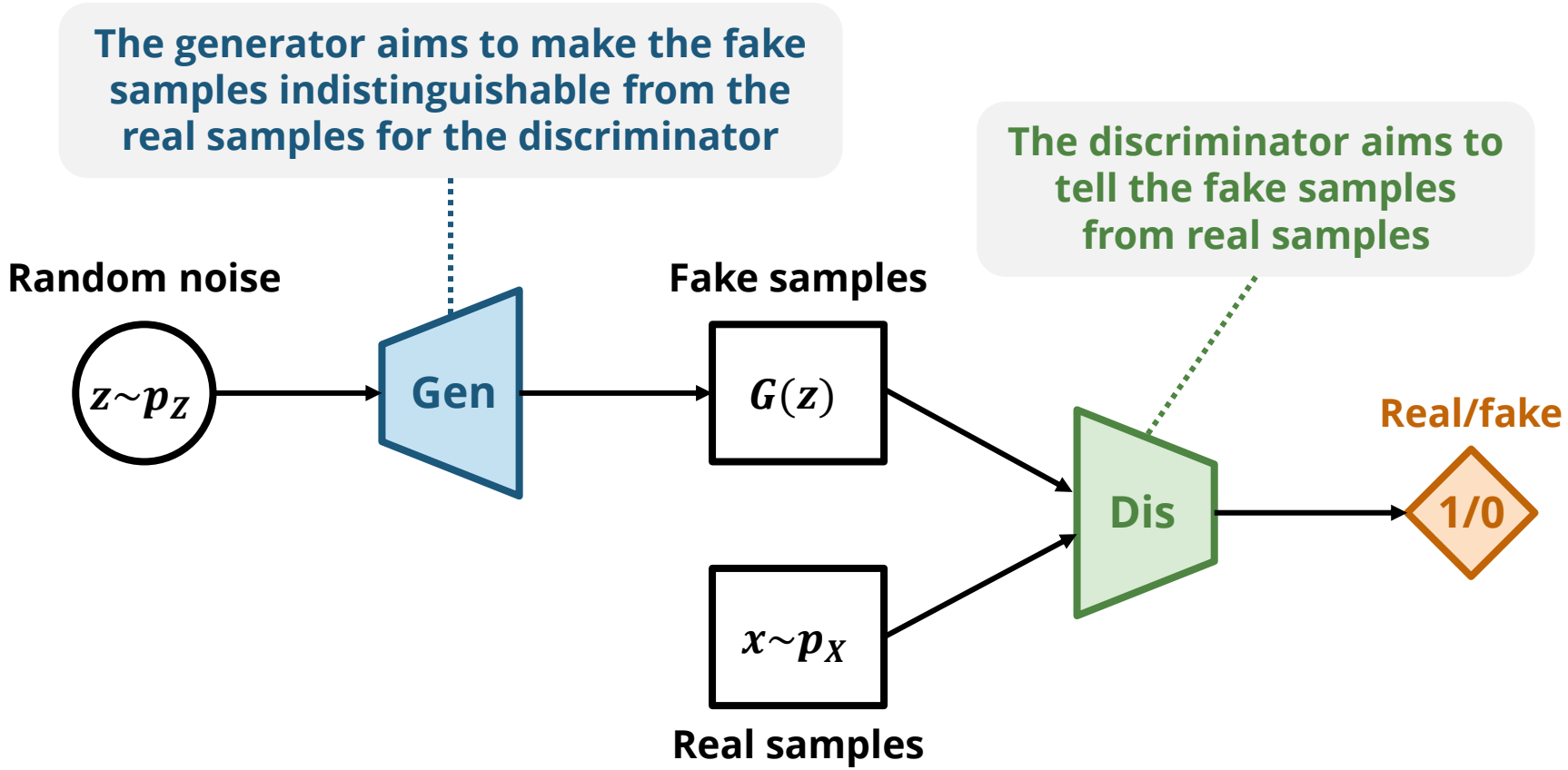
Example: MidiNet (Yang et al., 2017)



(Source: Yang et al., 2017)

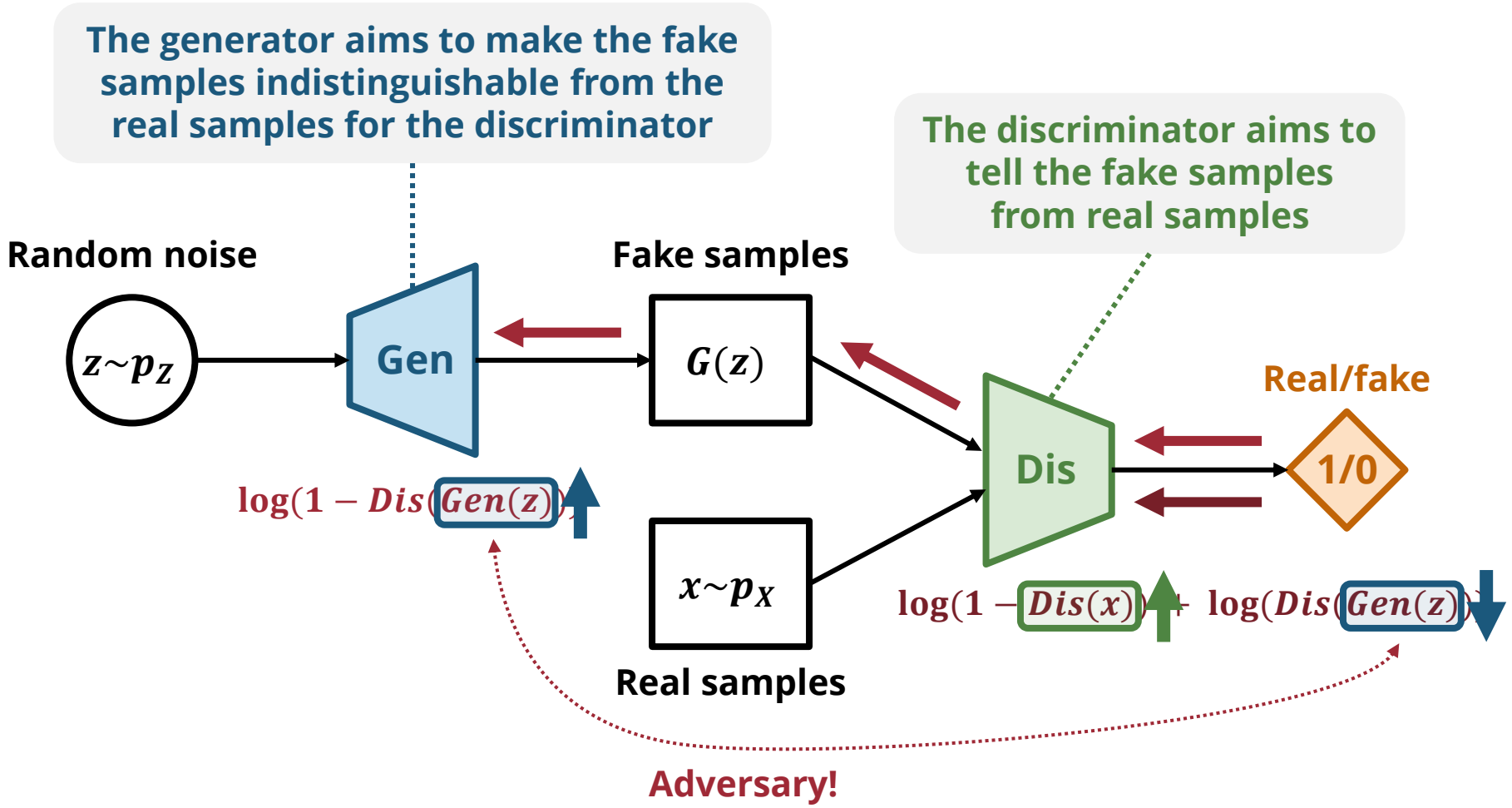
Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang, "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation," *ISMIR*, 2017.

(Recap) Generative Adversarial Nets (GANs)



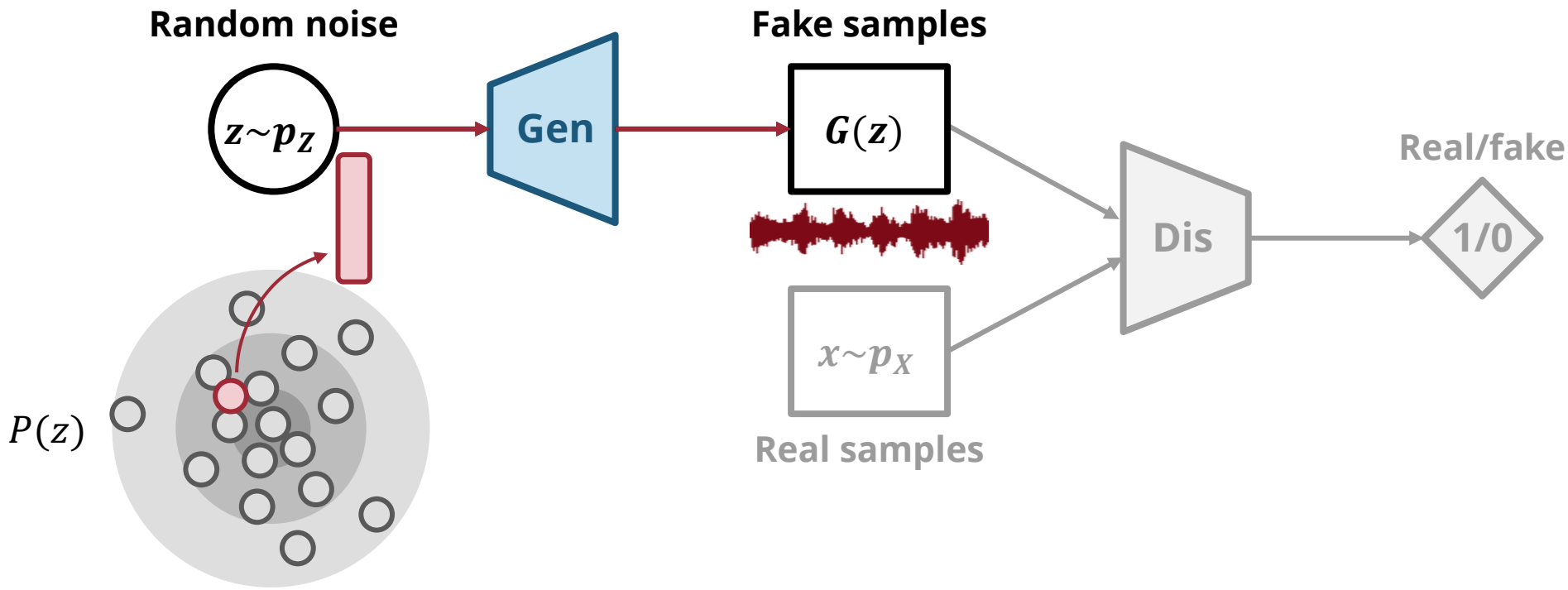
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

(Recap) Generative Adversarial Nets (GANs) – Training



Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

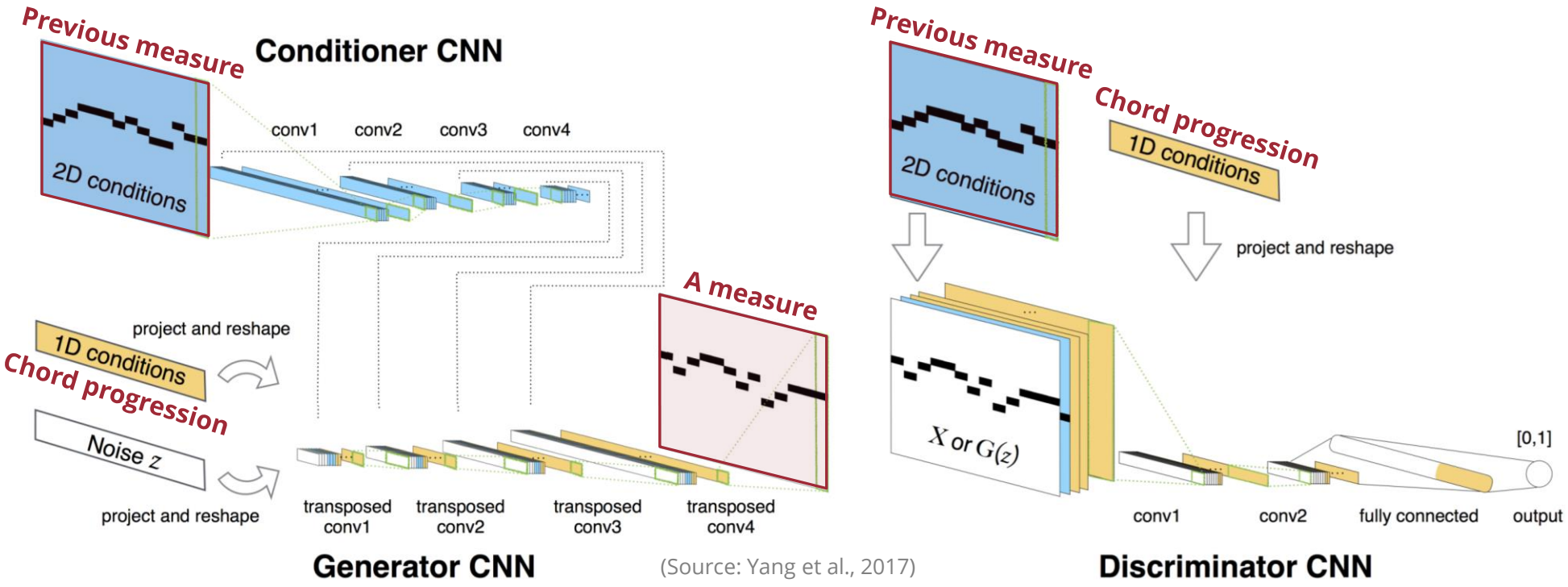
(Recap) Generative Adversarial Nets (GANs) – Generation



Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

Example: MidiNet (Yang et al., 2017)

Examples of generated music

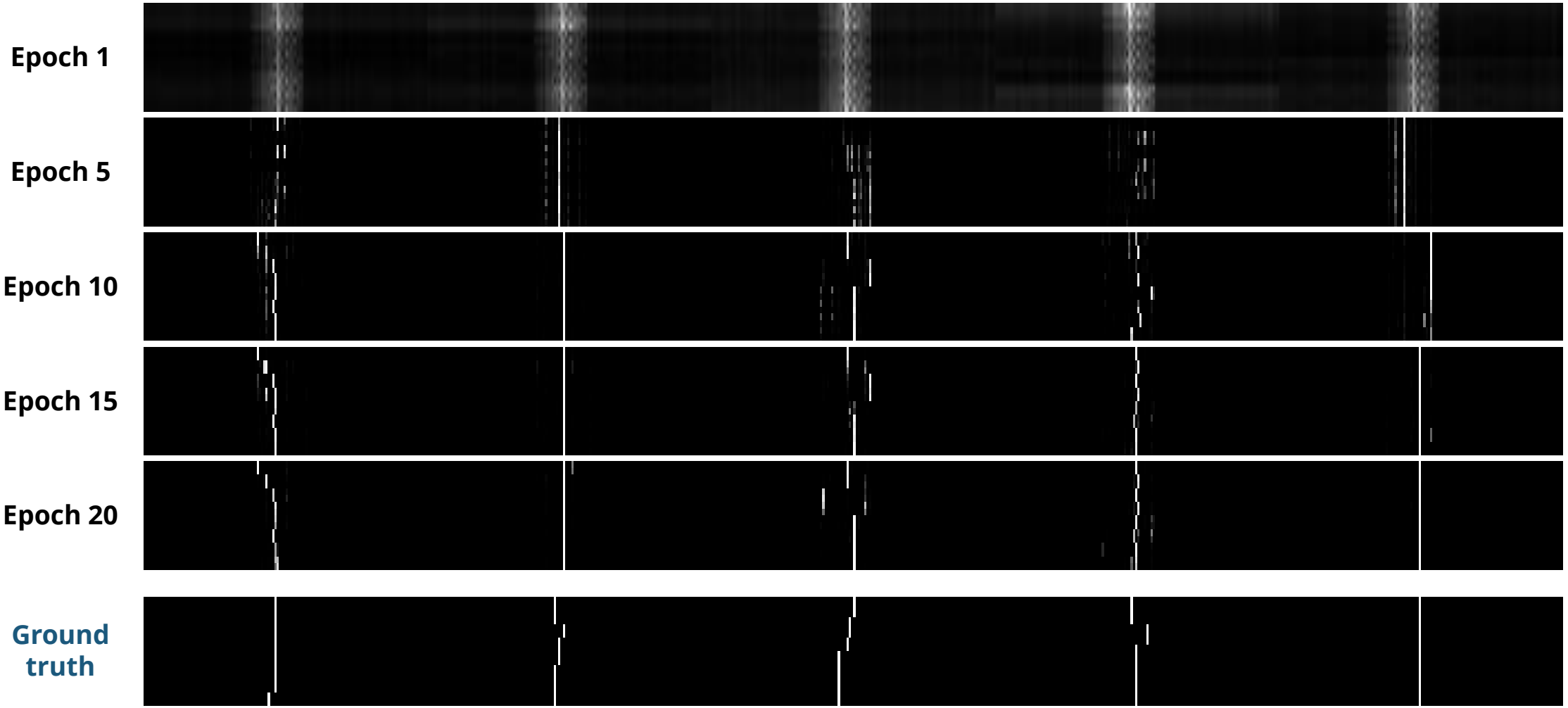


(Source: Yang et al., 2017)

MidiNet generates music measure-by-measure by conditioning on the last measure generated

Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang, "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation," *ISMIR*, 2017.

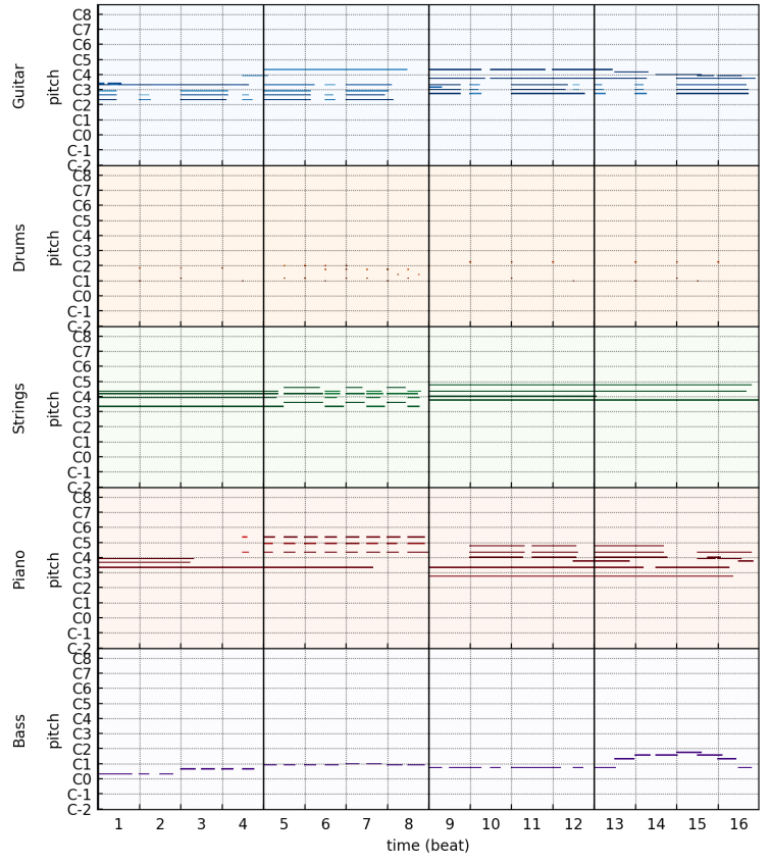
Example: MidiNet (Yang et al., 2017)

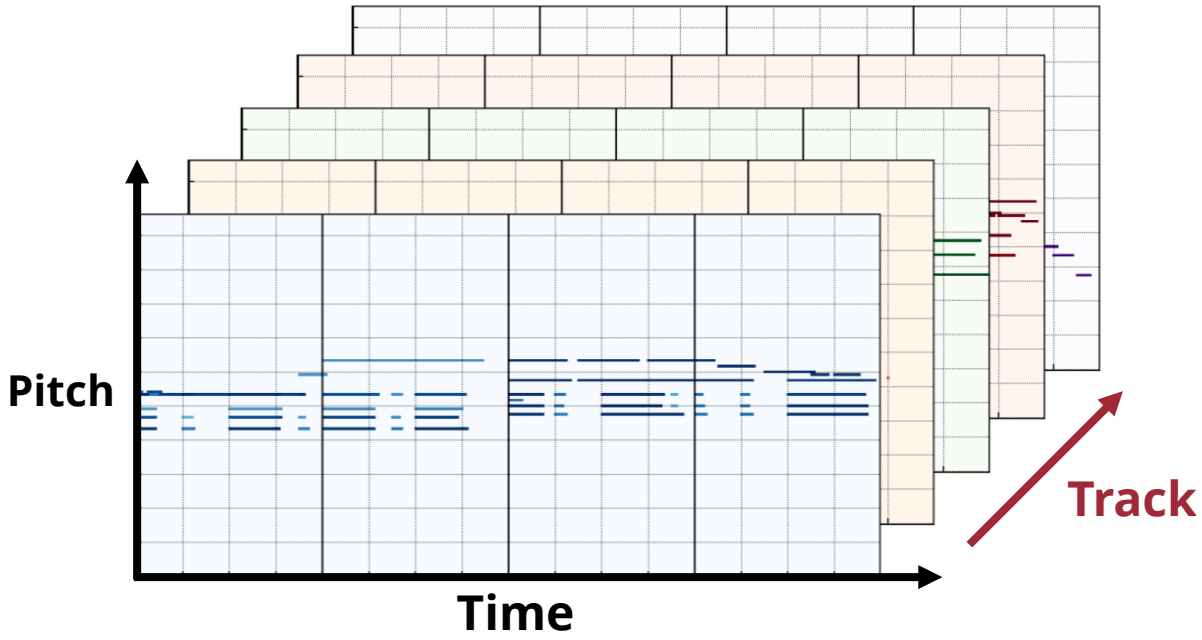


(Source: Yang et al., 2017)

Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang, "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation," *ISMIR*, 2017.

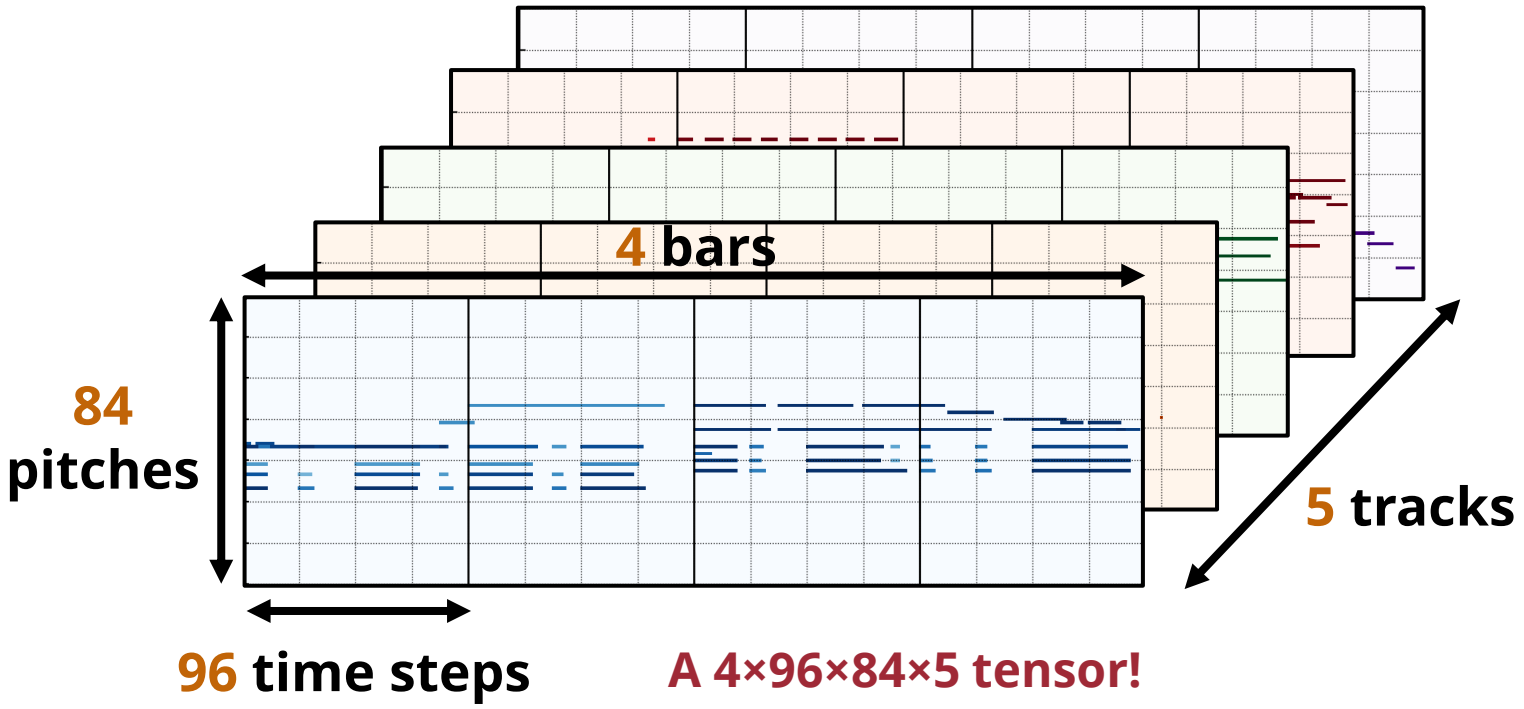
Multitrack Piano Rolls



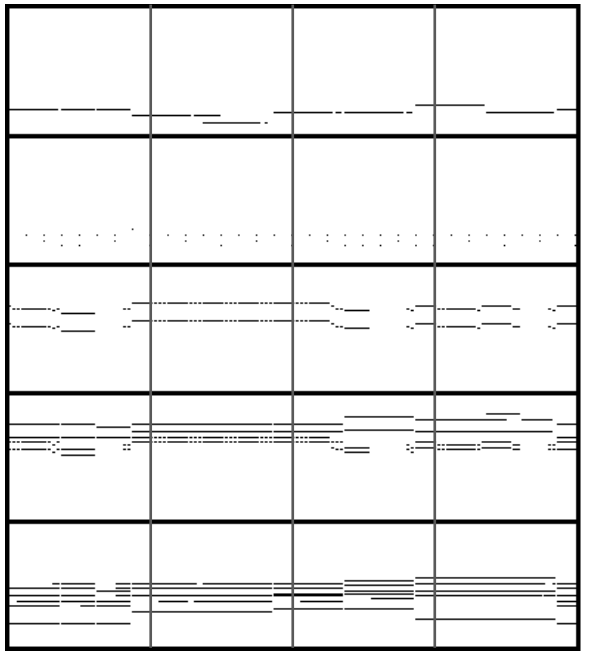


Example: MuseGAN (Dong et al., 2018)

Representation

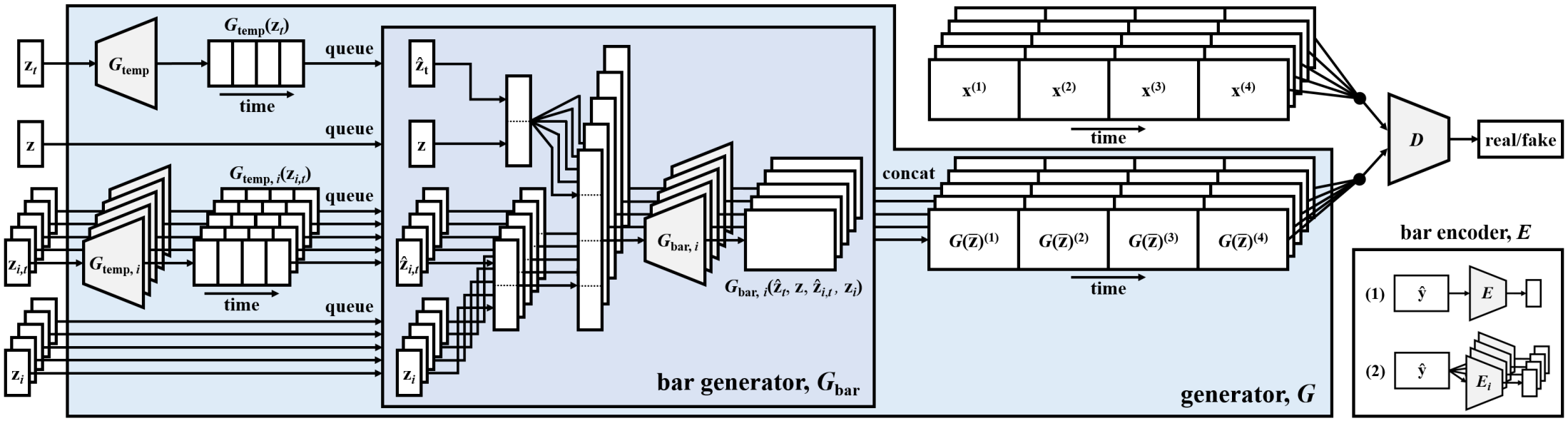


A training sample



Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

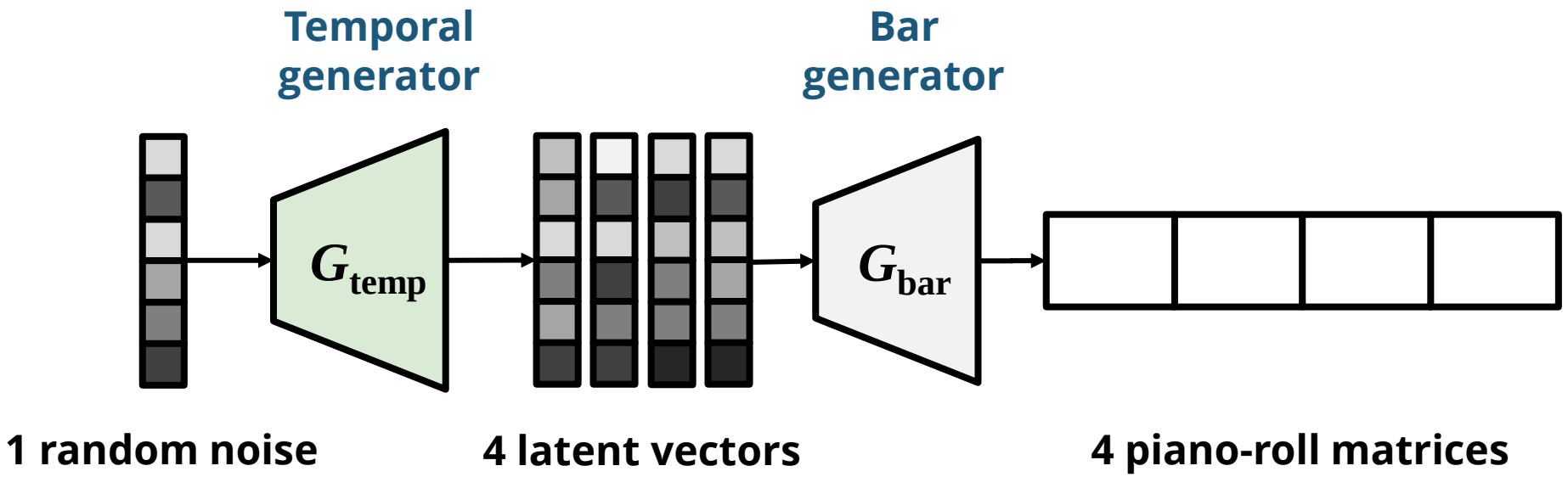
Example: MuseGAN (Dong et al., 2018)



(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

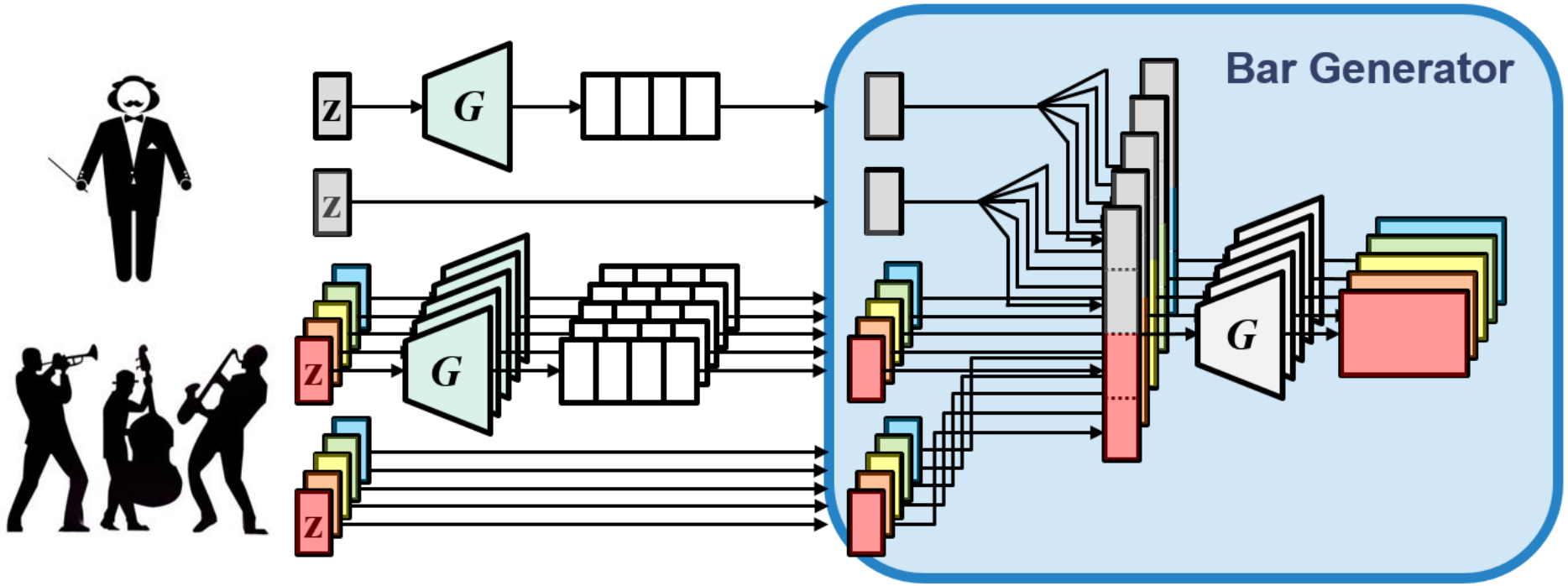
Example: MuseGAN (Dong et al., 2018)



(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

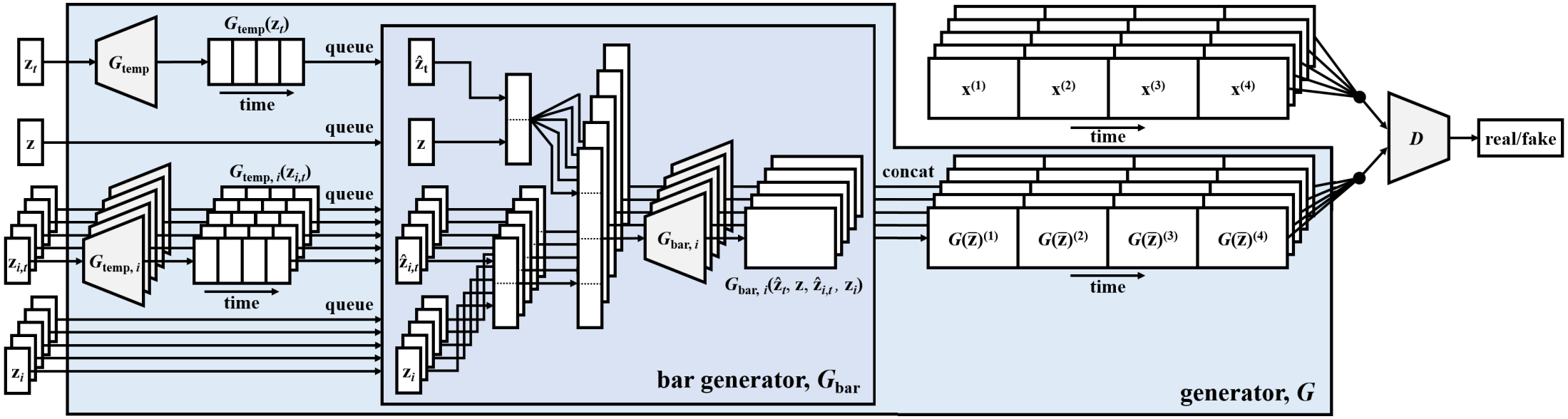
Example: MuseGAN (Dong et al., 2018)



(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

Example: MuseGAN (Dong et al., 2018)

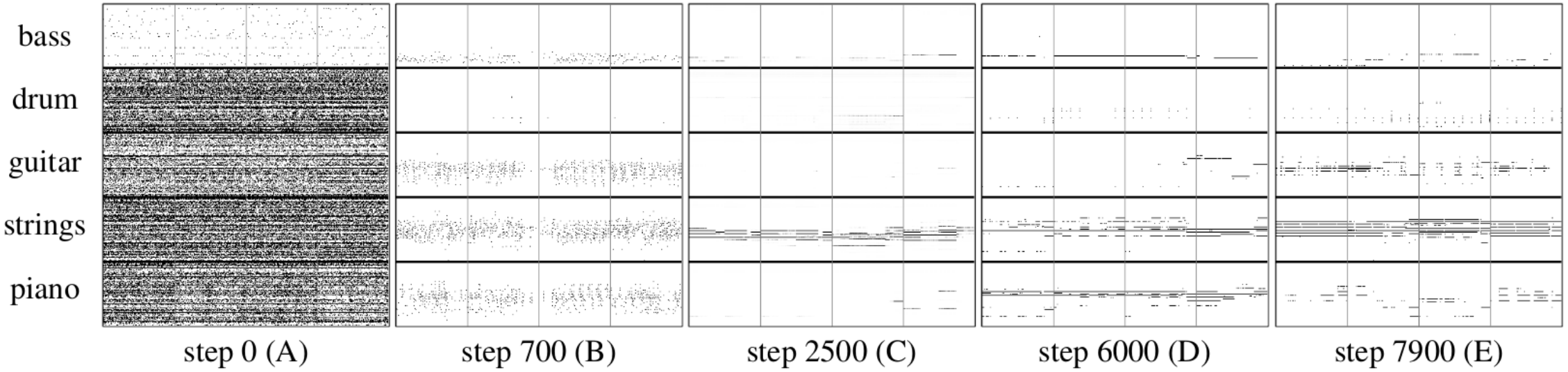


(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

Example: MuseGAN (Dong et al., 2018)

Examples of generated music

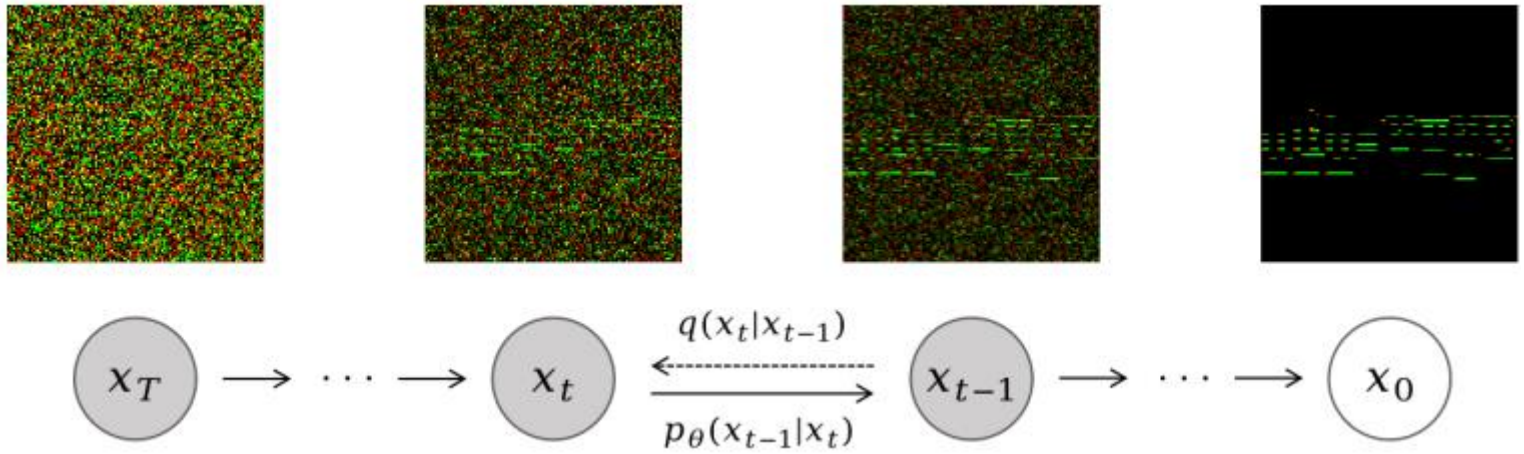


(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

Music Generation using Diffusion Models

Example: Polyffusion (Min et al., 2023)



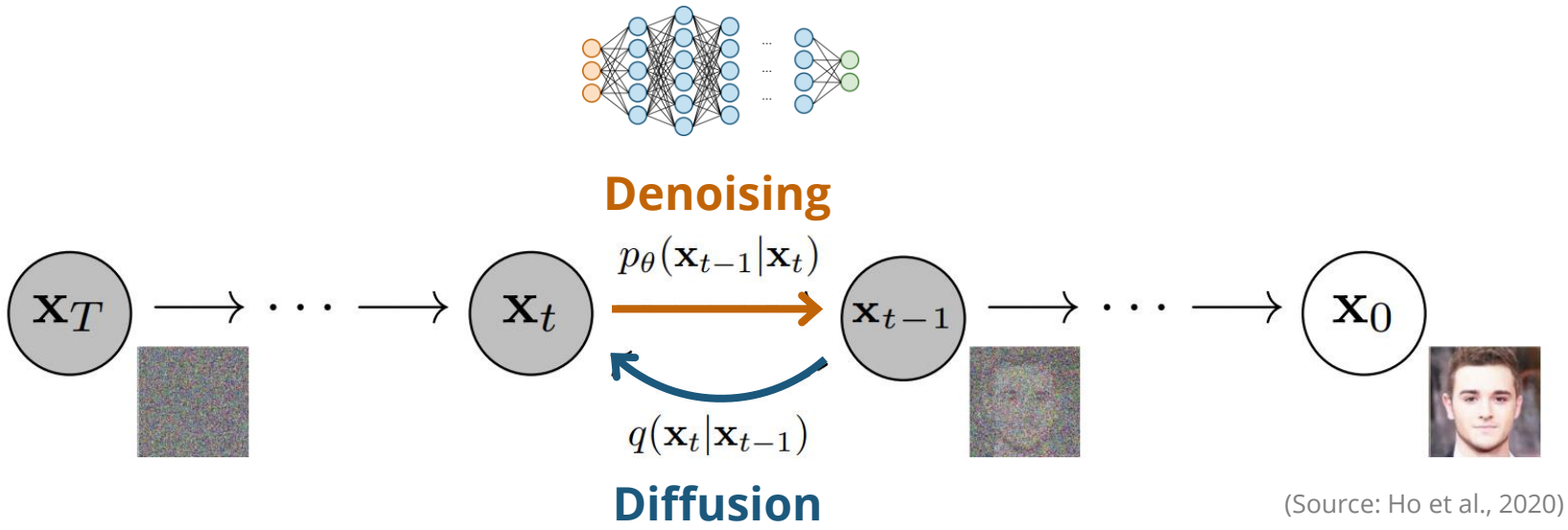
(Source: Min et al., 2023)

polyffusion.github.io

Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao, "Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls," *ISMIR*, 2023.

(Recap) Diffusion Models

- **Intuition**: Many denoising autoencoders stacked together

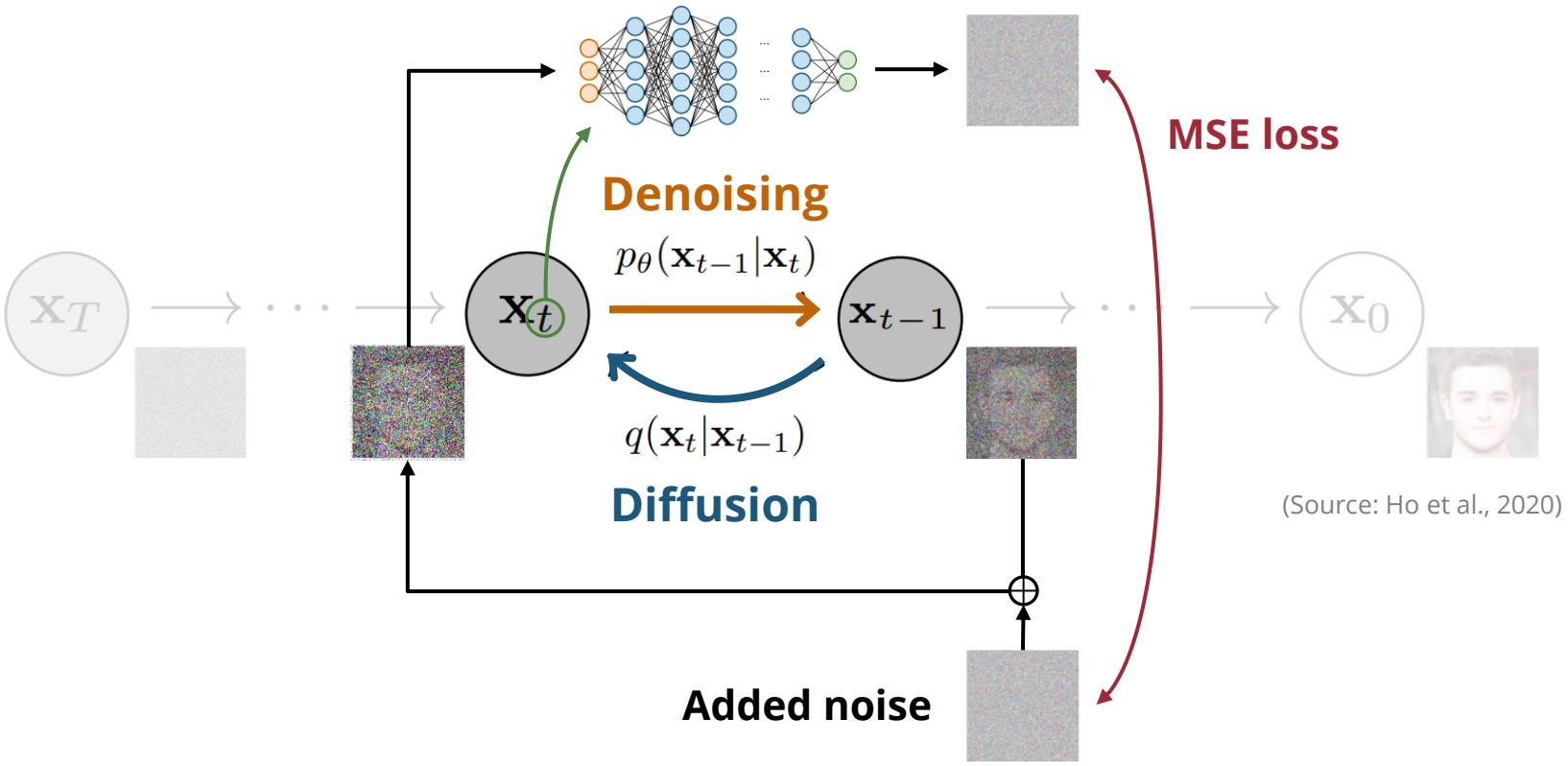


(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

(Recap) Diffusion Models – Training

- **Intuition**: Many denoising autoencoders stacked together

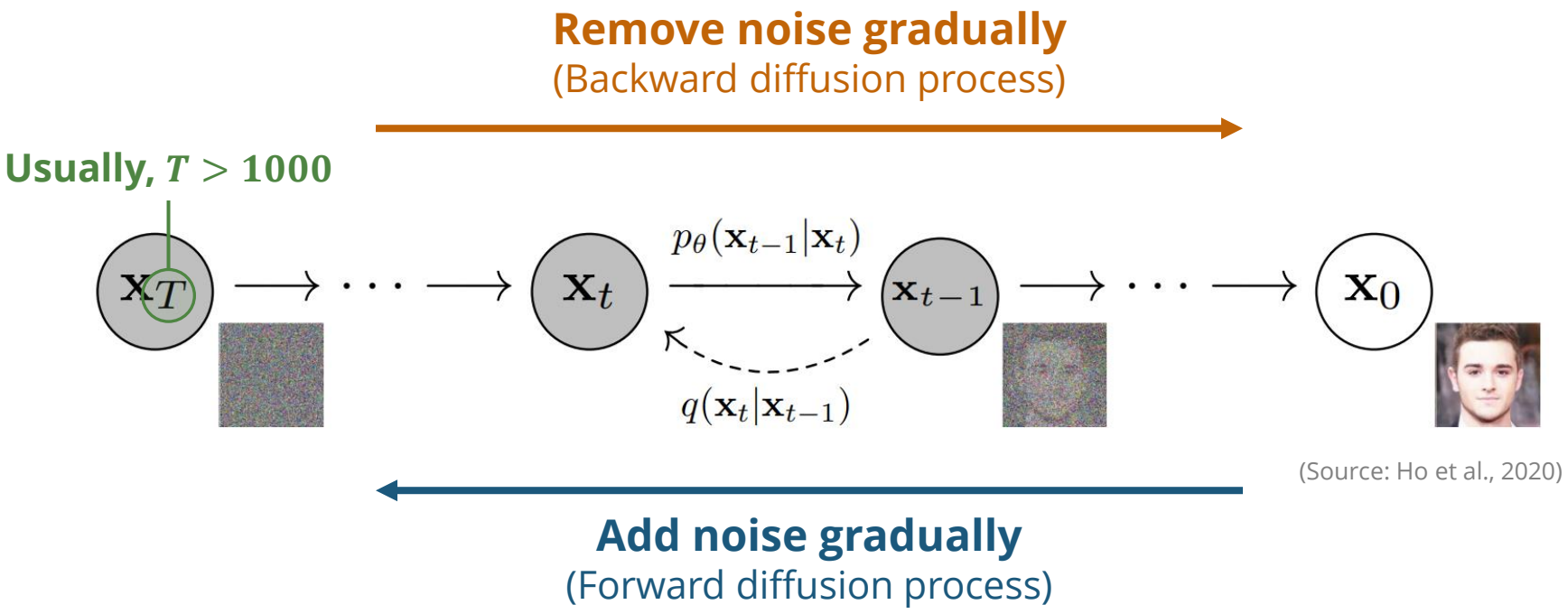


(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

(Recap) Diffusion Models

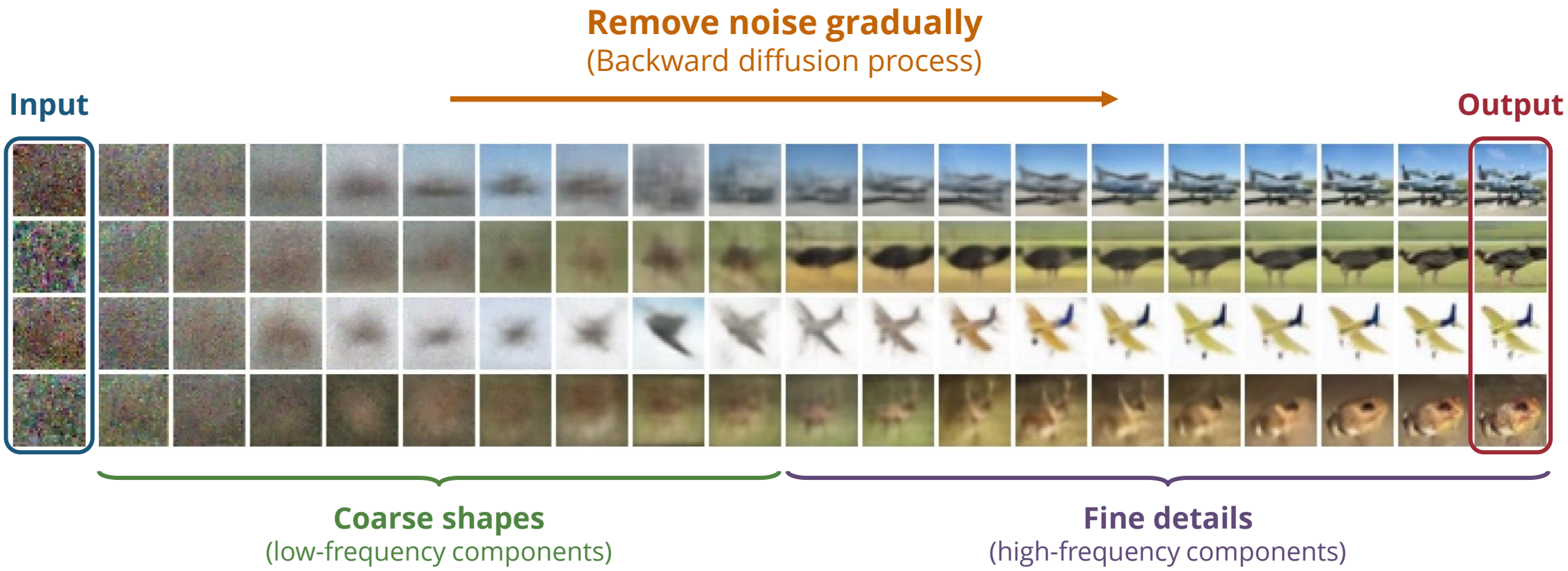
- **Intuition**: Many denoising autoencoders stacked together



(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

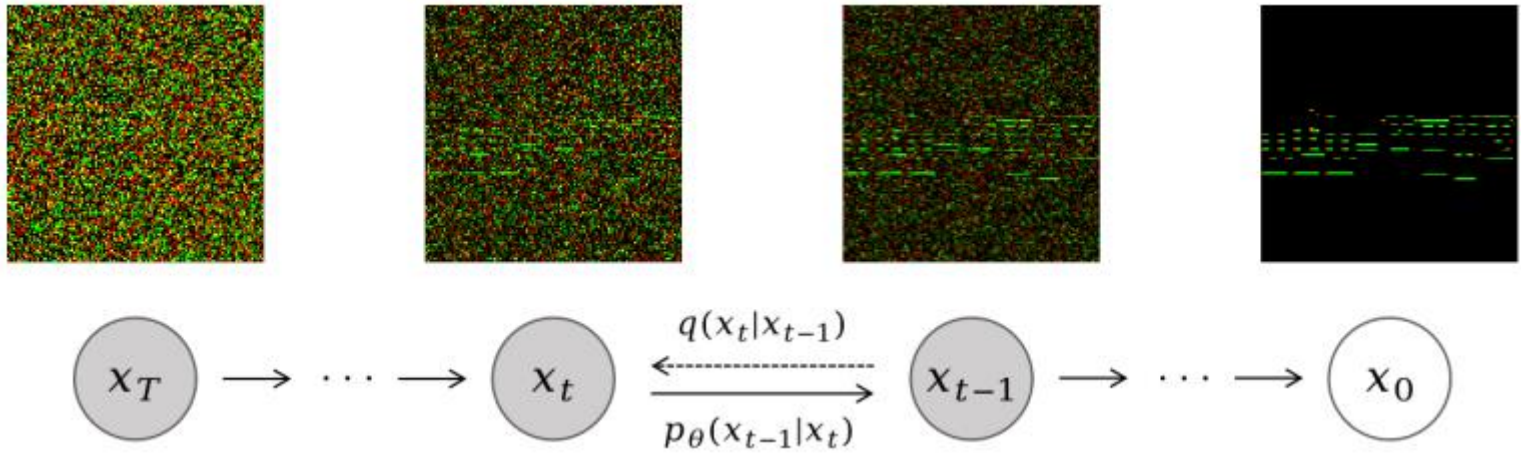
(Recap) Diffusion Models – Generation



(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Example: Polyffusion (Min et al., 2023)

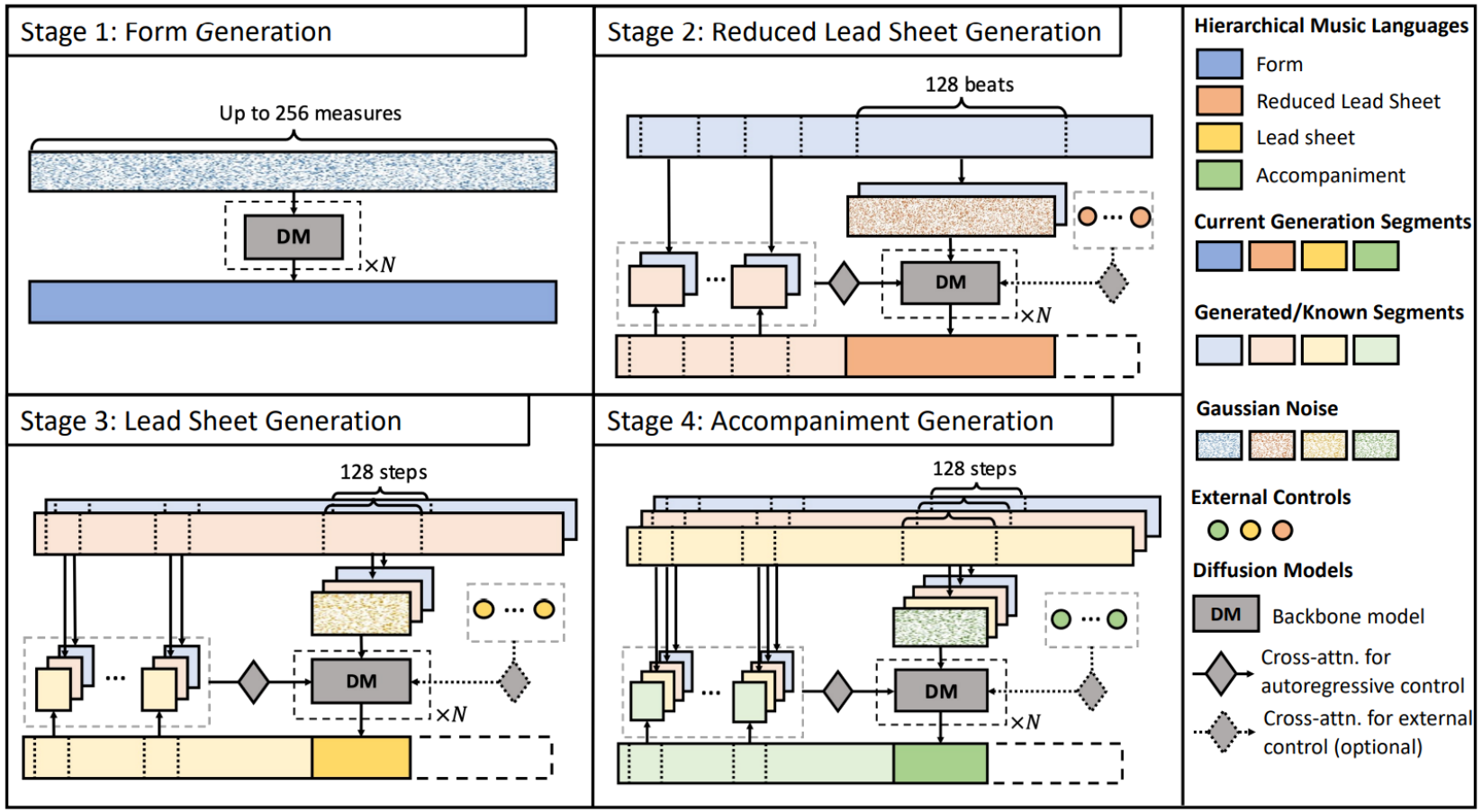


(Source: Min et al., 2023)

polyffusion.github.io

Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao, "Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls," *ISMIR*, 2023.

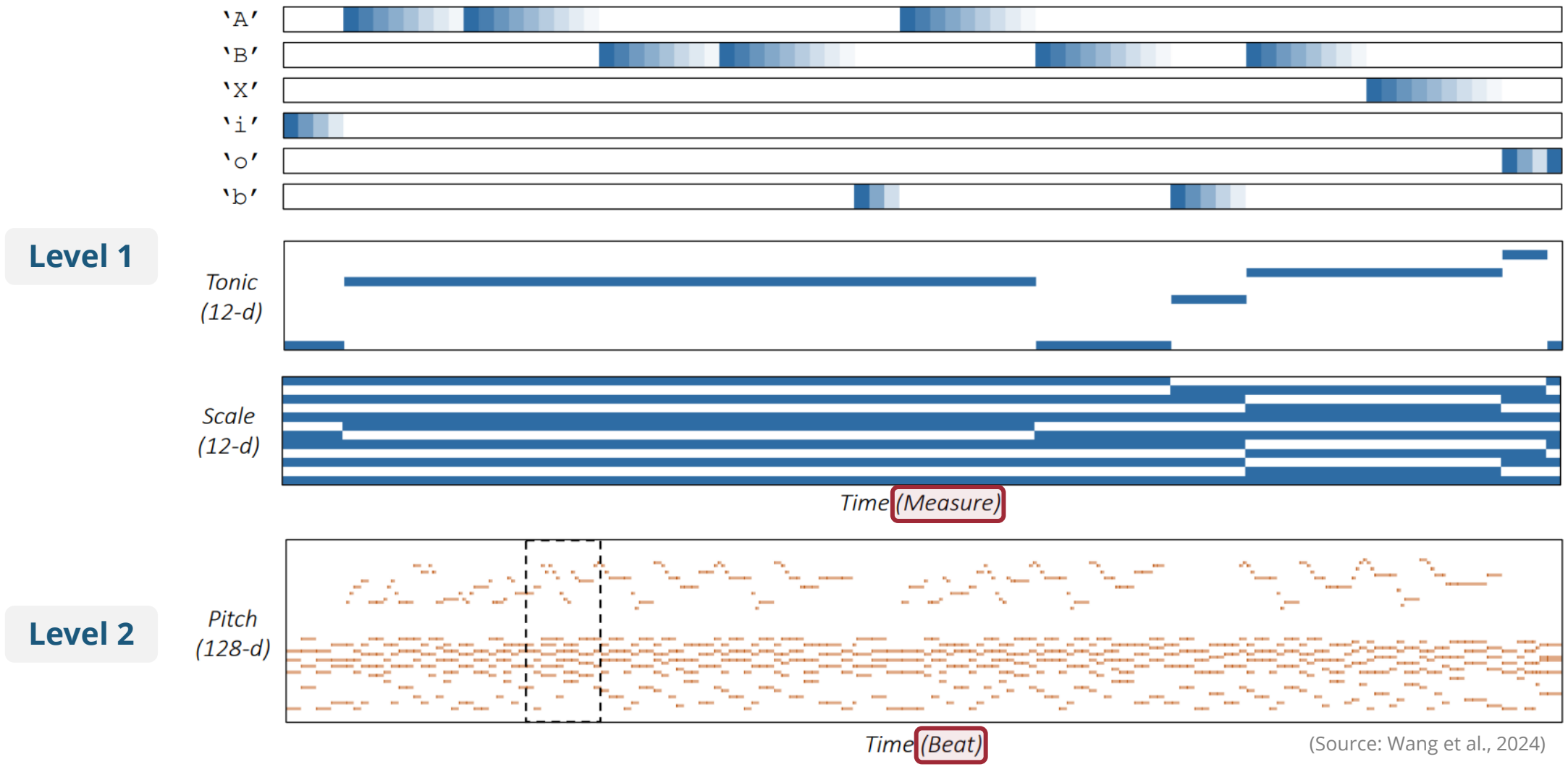
Example: Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

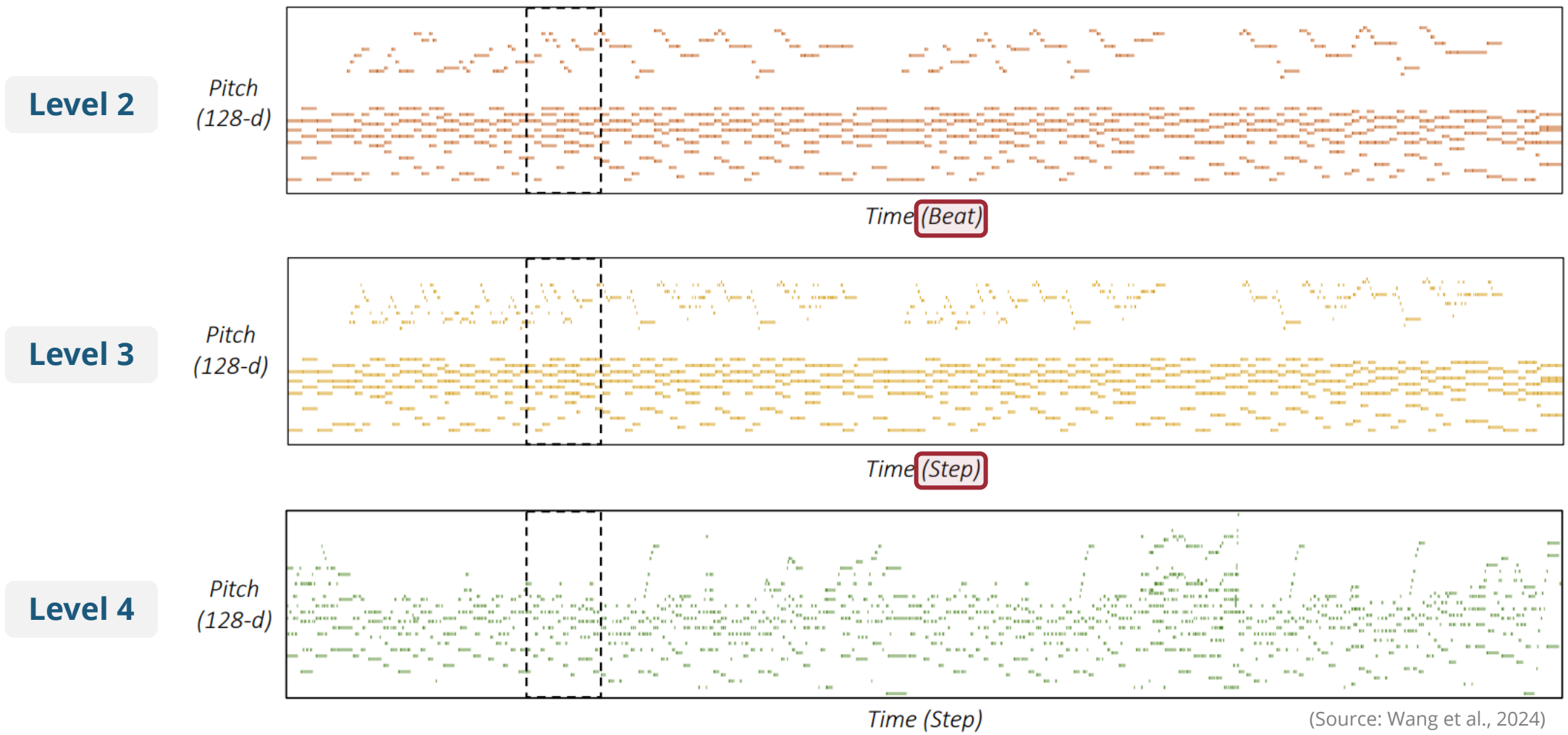
Example: Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

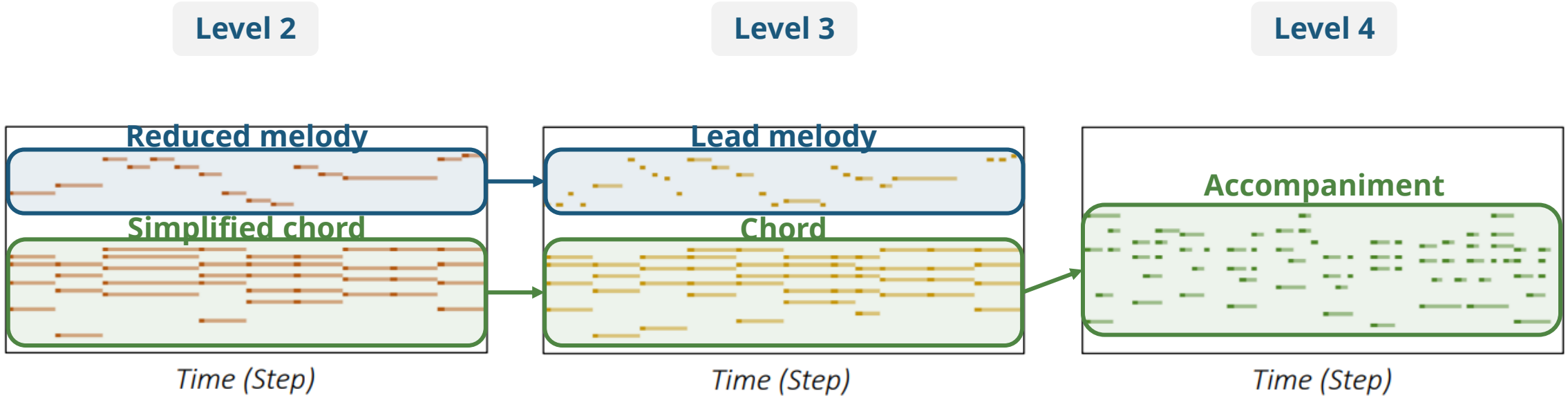
Example: Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

Example: Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

wholesonggen.github.io

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.