PAT 498/598 (Fall 2024)

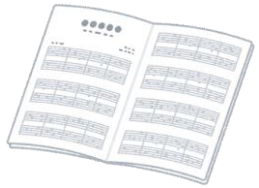
Special Topics: Generative AI for Music and Audio Creation

Lecture 14: Controllable 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

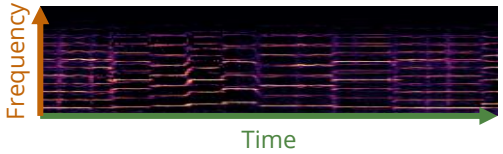
```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

- **Image-based**
  - Piano roll

Audio-domain music generation

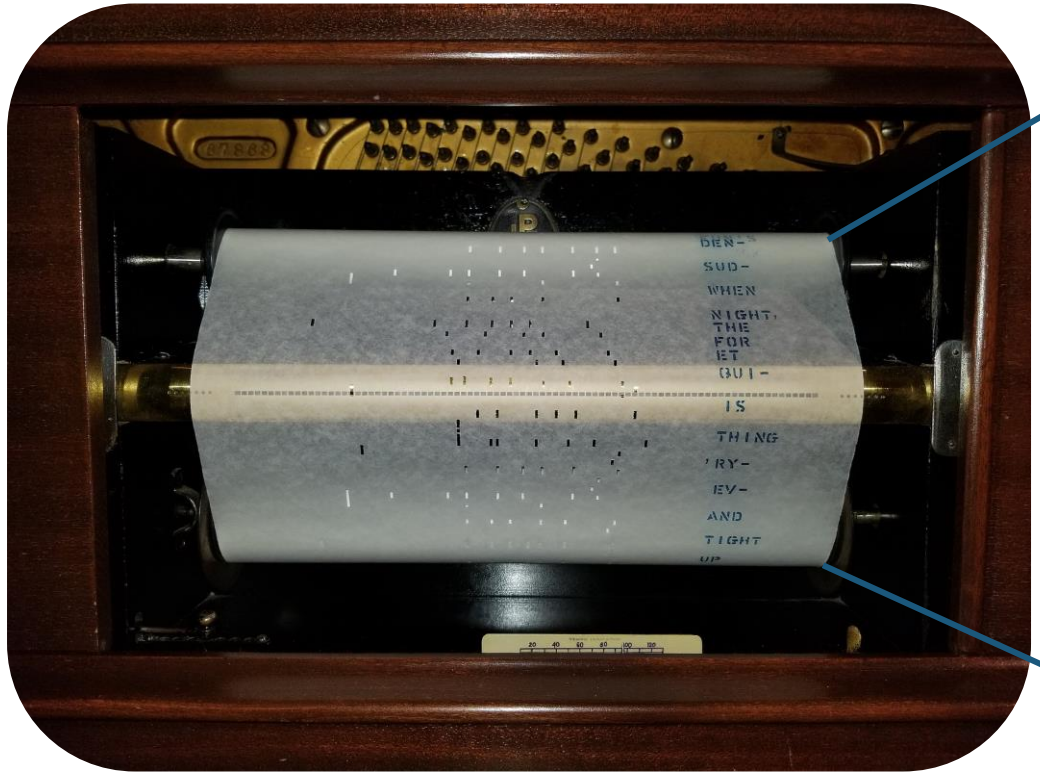
- **Time series-based**
  - Waveform

- **Image-based**
  - Spectrogram

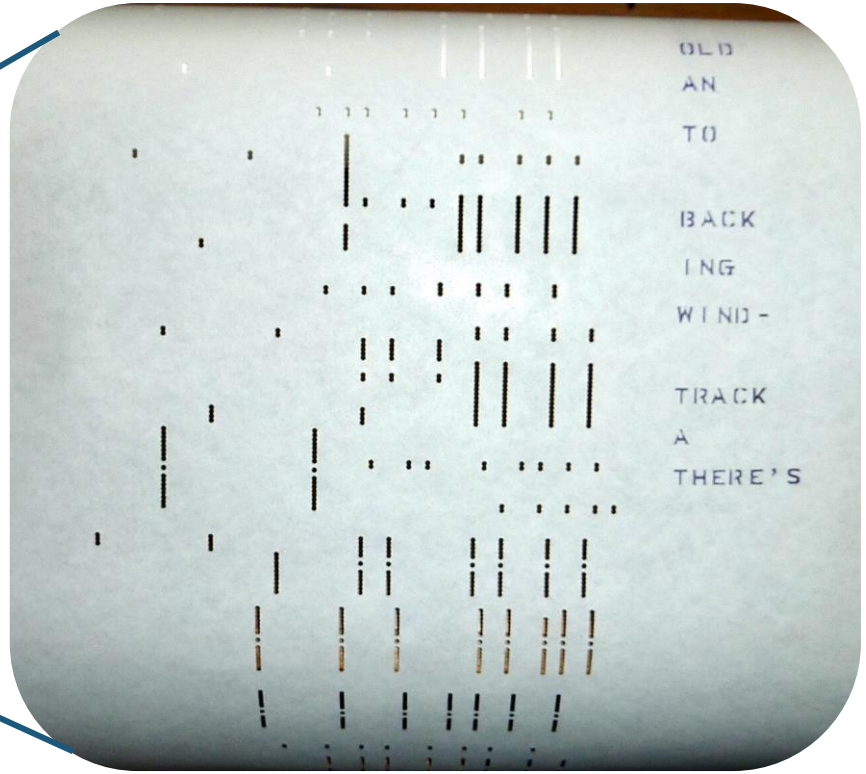


Today, we also have many latent-space based systems!

(Recap) Piano Rolls

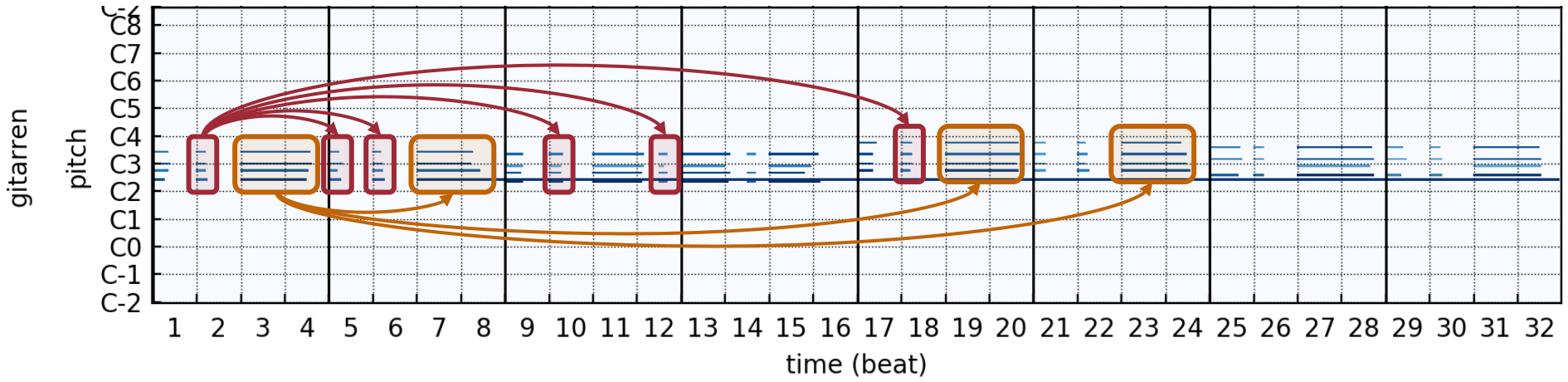


(Source: Draconichiaro)



(Source: Tangerineduel)

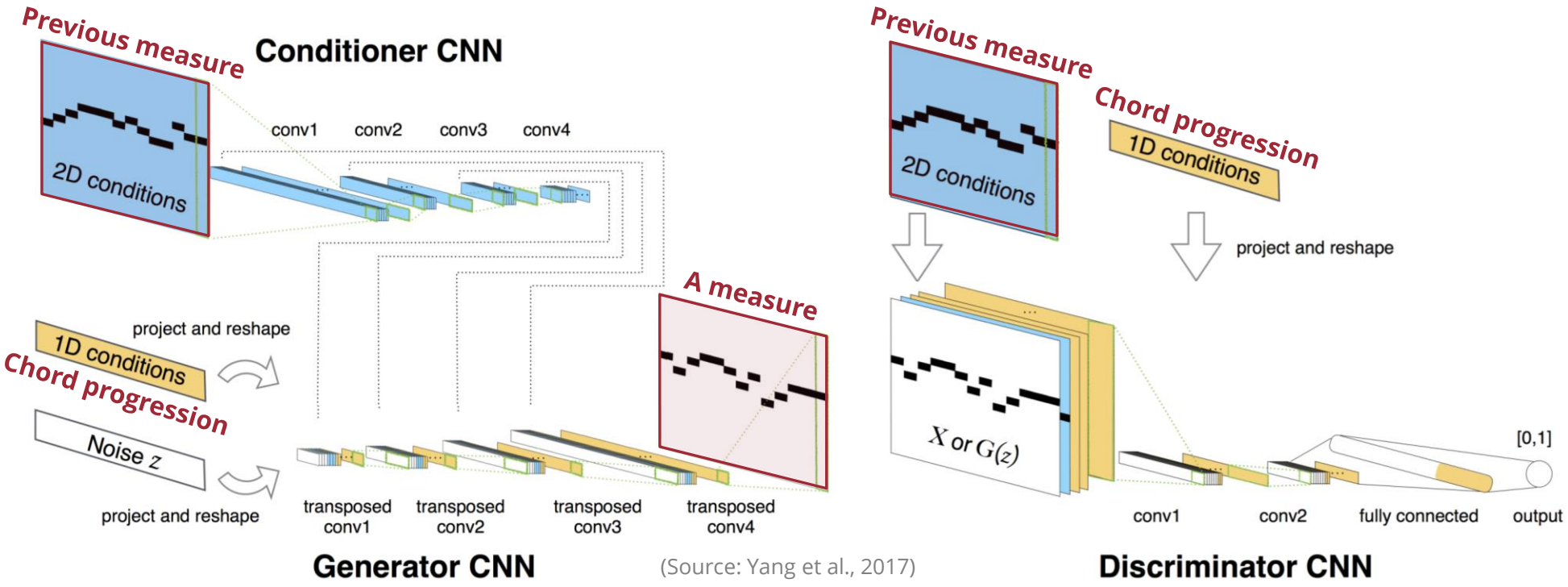
(Recap) Why Piano Rolls?



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

(Recap) Example: MidiNet (Yang et al., 2017)

Examples of generated music

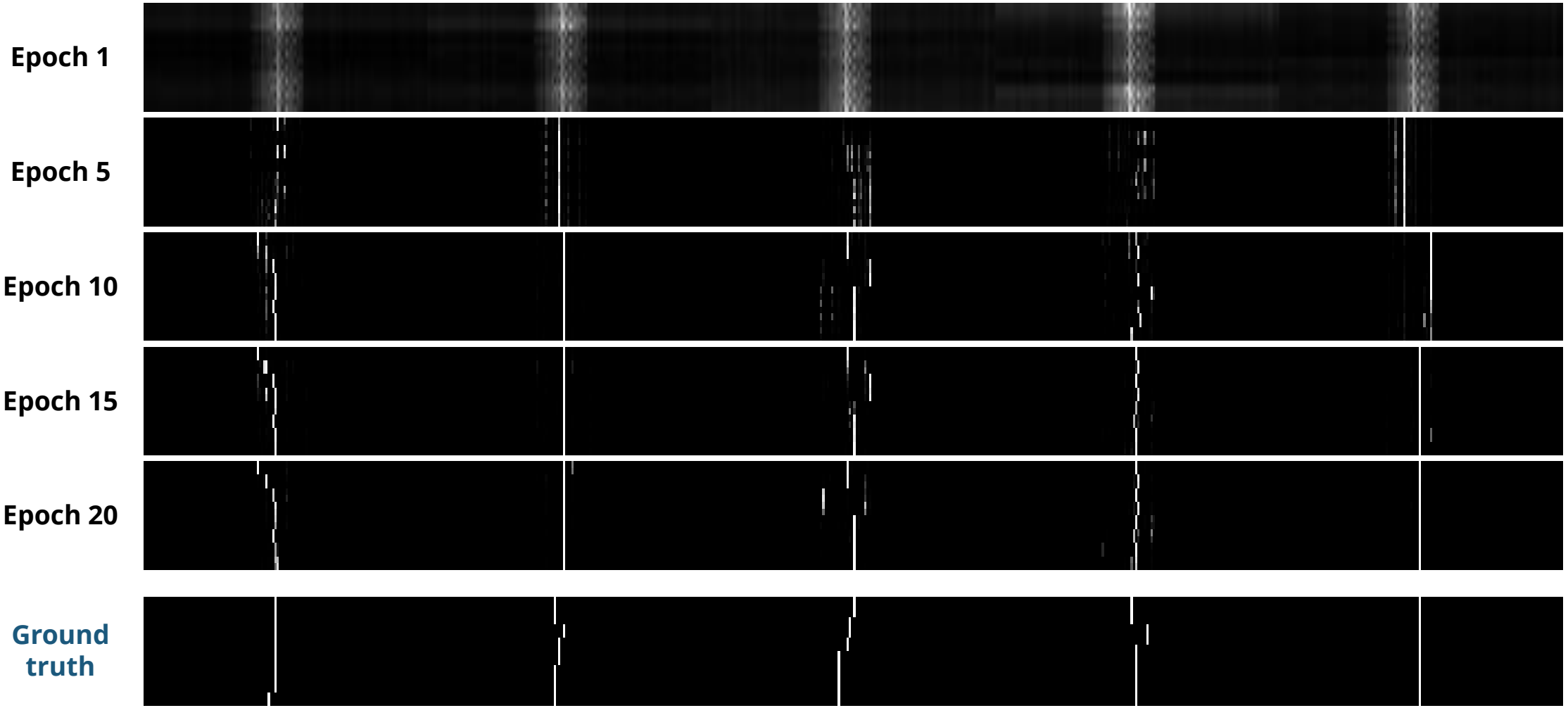


(Source: Yang et al., 2017)

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.

(Recap) Example: MidiNet (Yang et al., 2017)

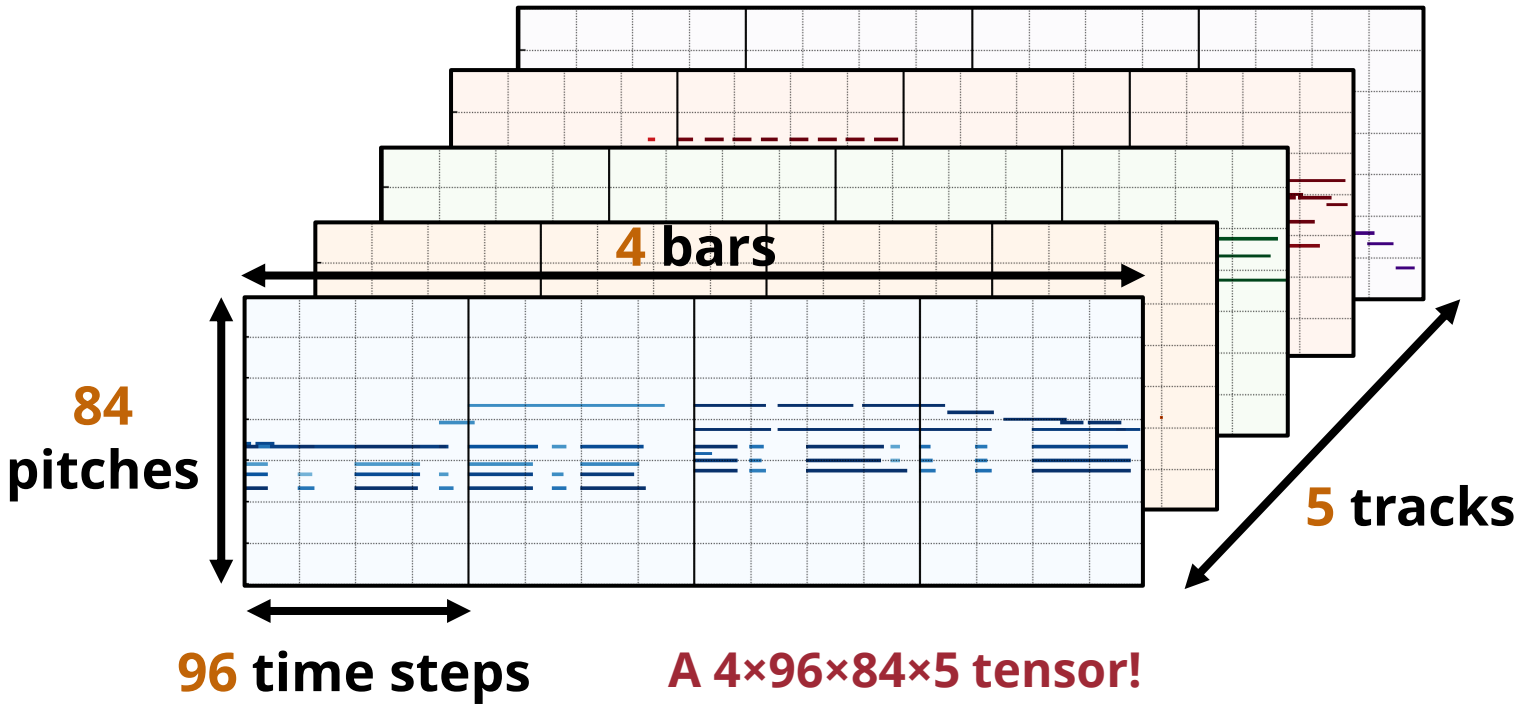


(Source: Yang et al., 2017)

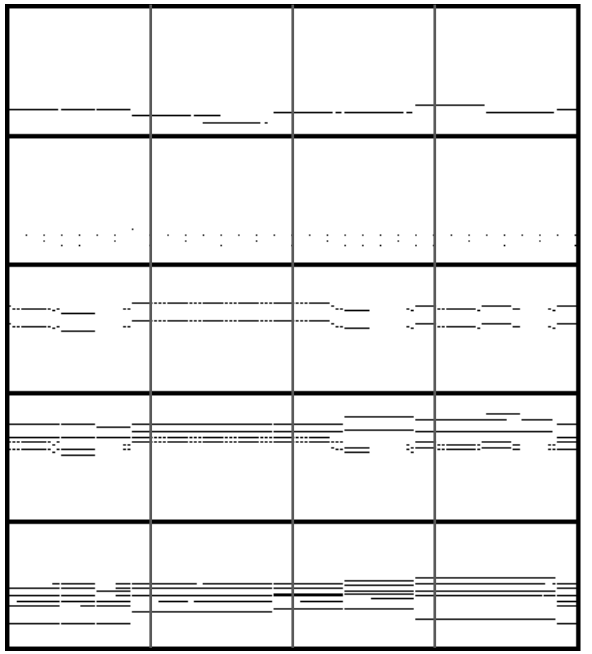
Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang, "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation," *ISMIR*, 2017.

(Recap) Example: MuseGAN (Dong et al., 2018)

Representation

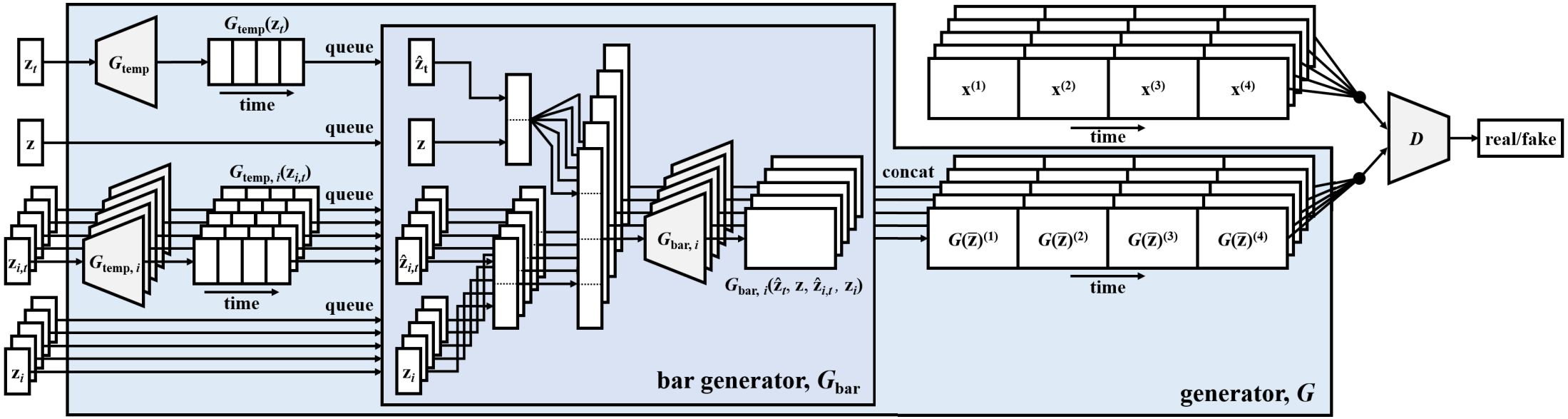


A training sample



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.

(Recap) Example: MuseGAN (Dong et al., 2018)

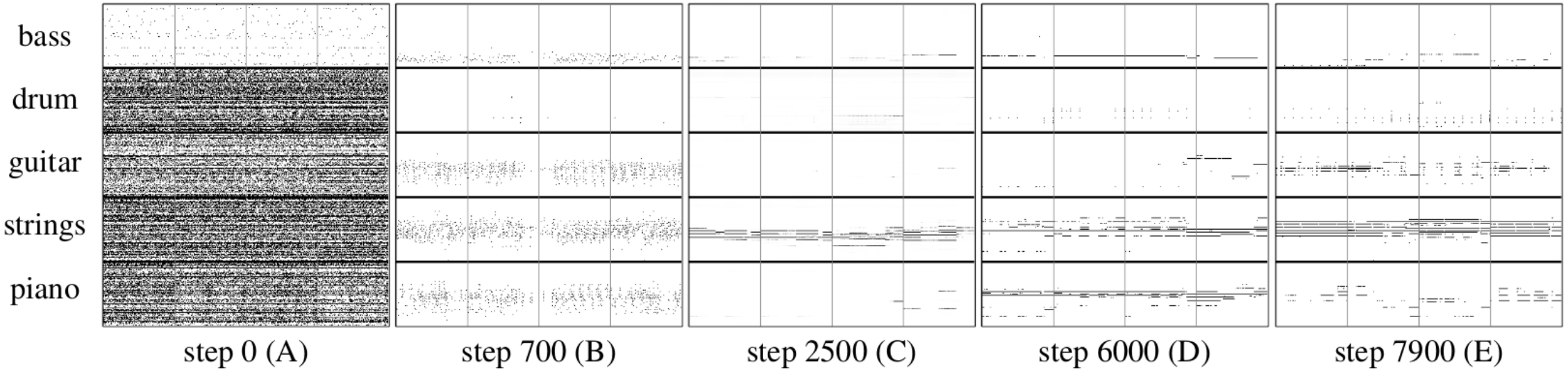


(Source: Dong et al., 2018)

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.

(Recap) Example: MuseGAN (Dong et al., 2018)

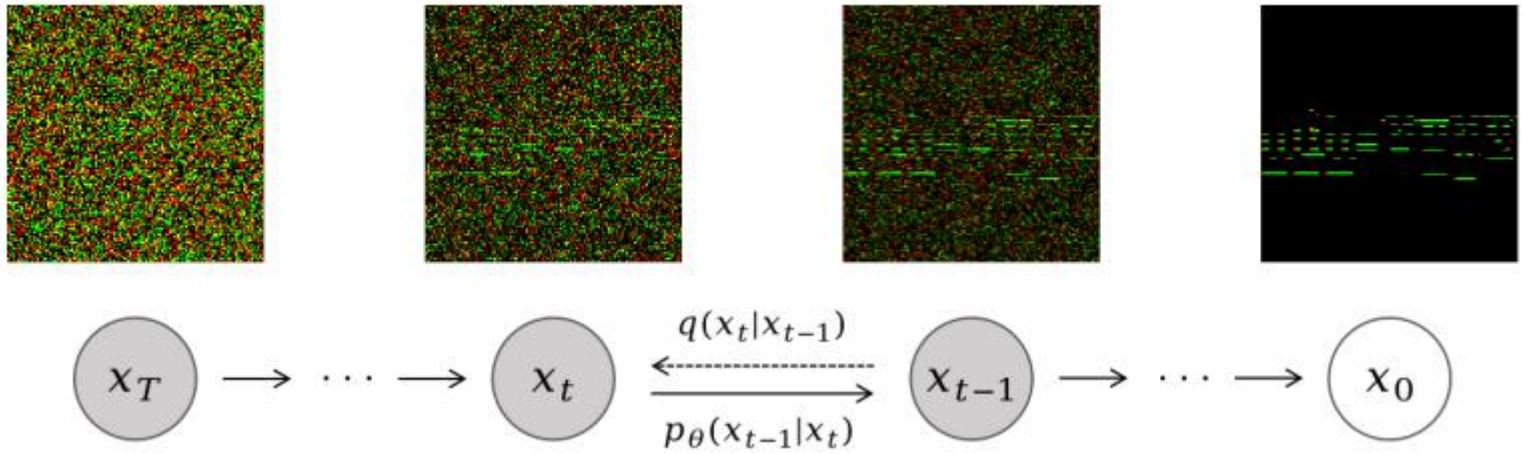
Examples of generated music



(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.

(Recap) Example: Polyffusion (Min et al., 2023)

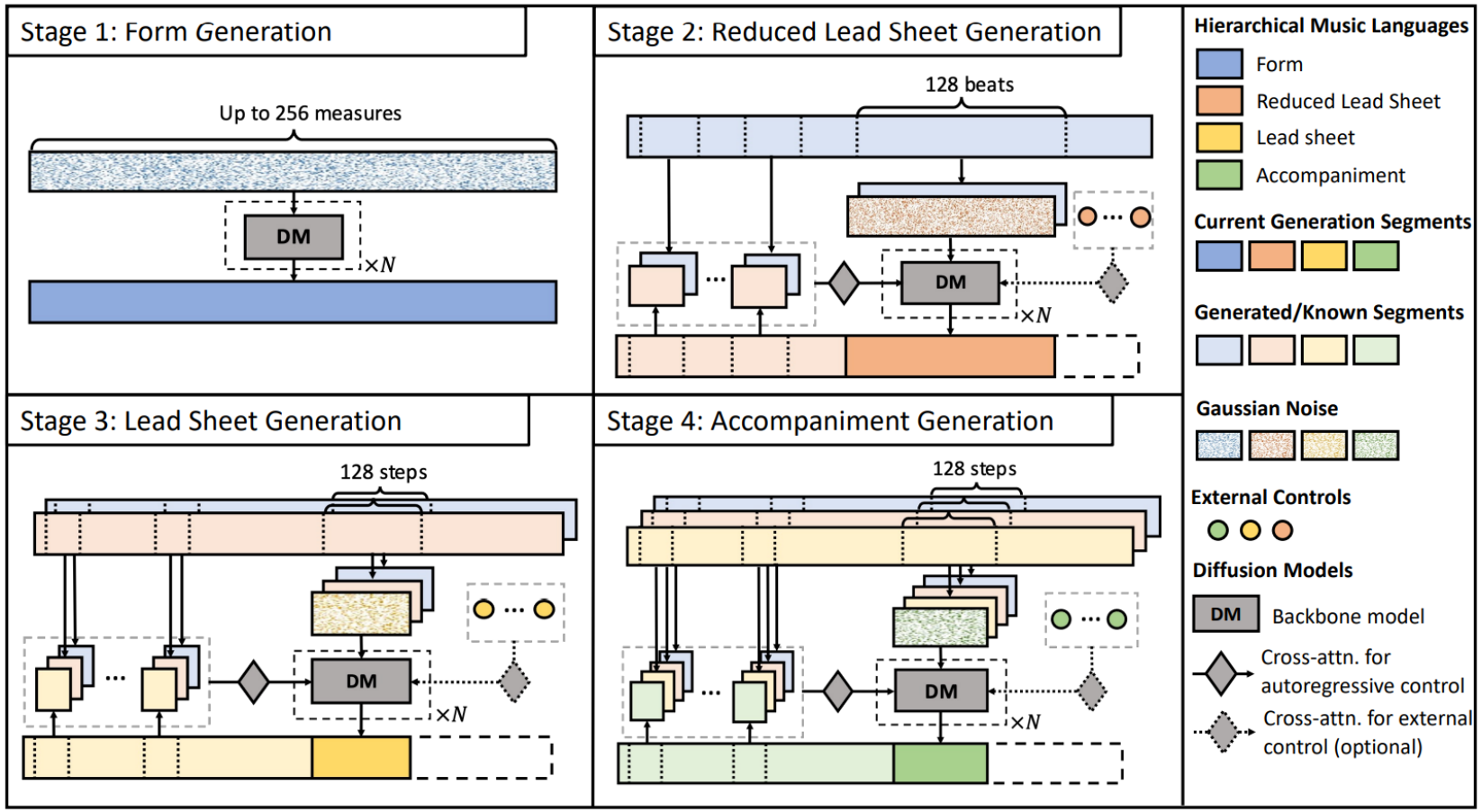


(Source: Min et al., 2023)

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.

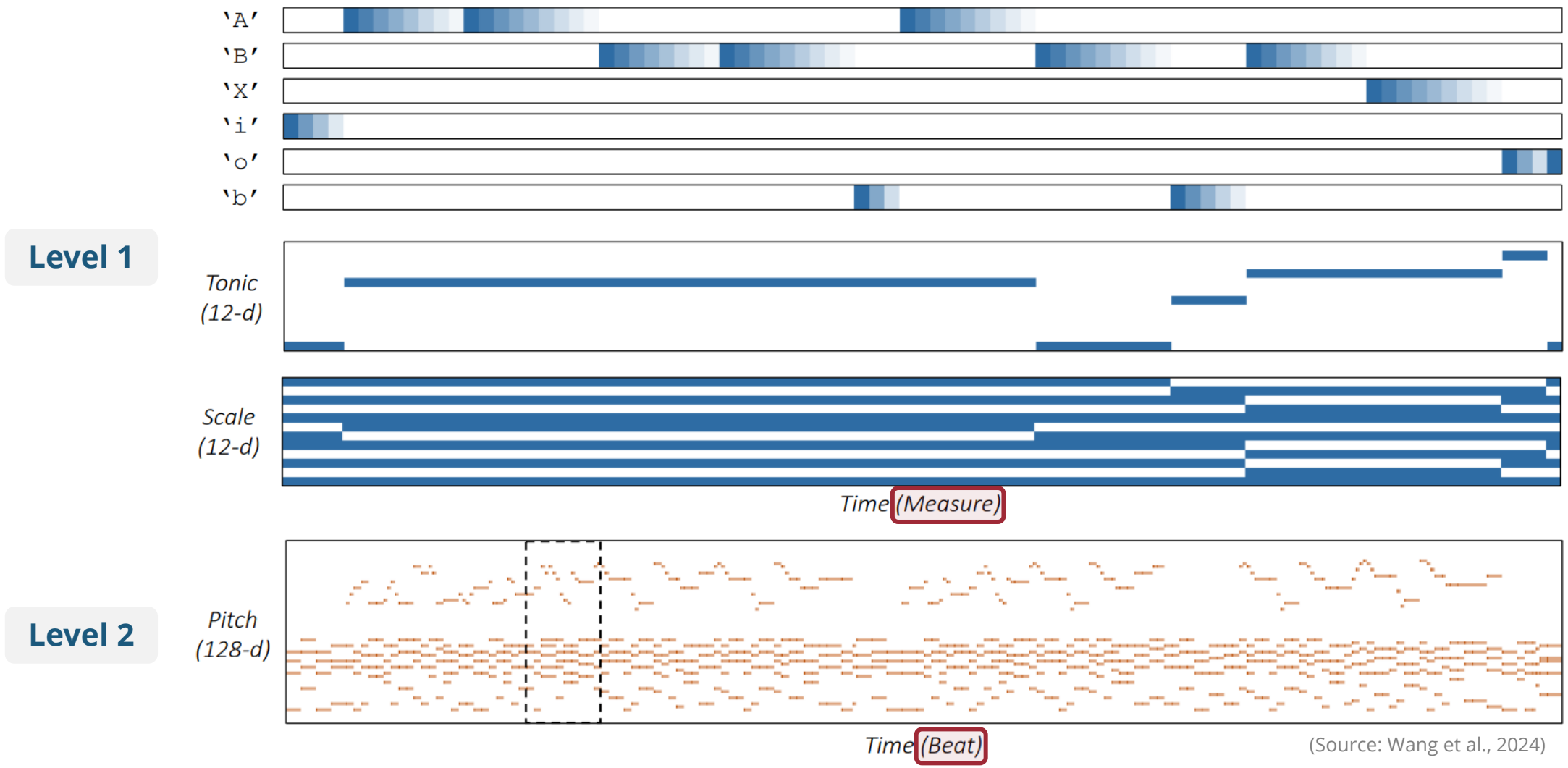
(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



(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.

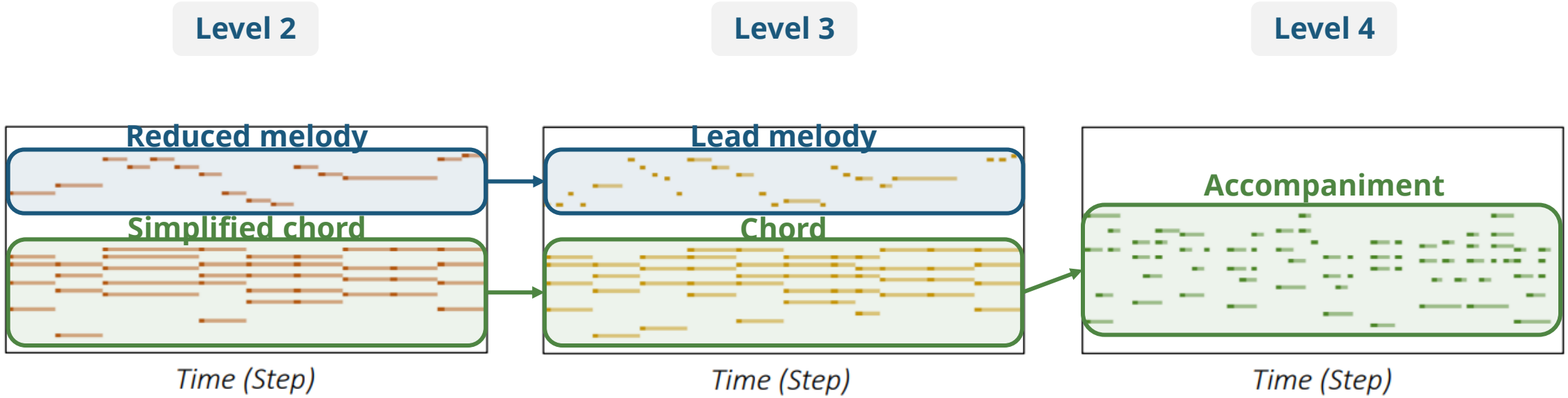
(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



(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.

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



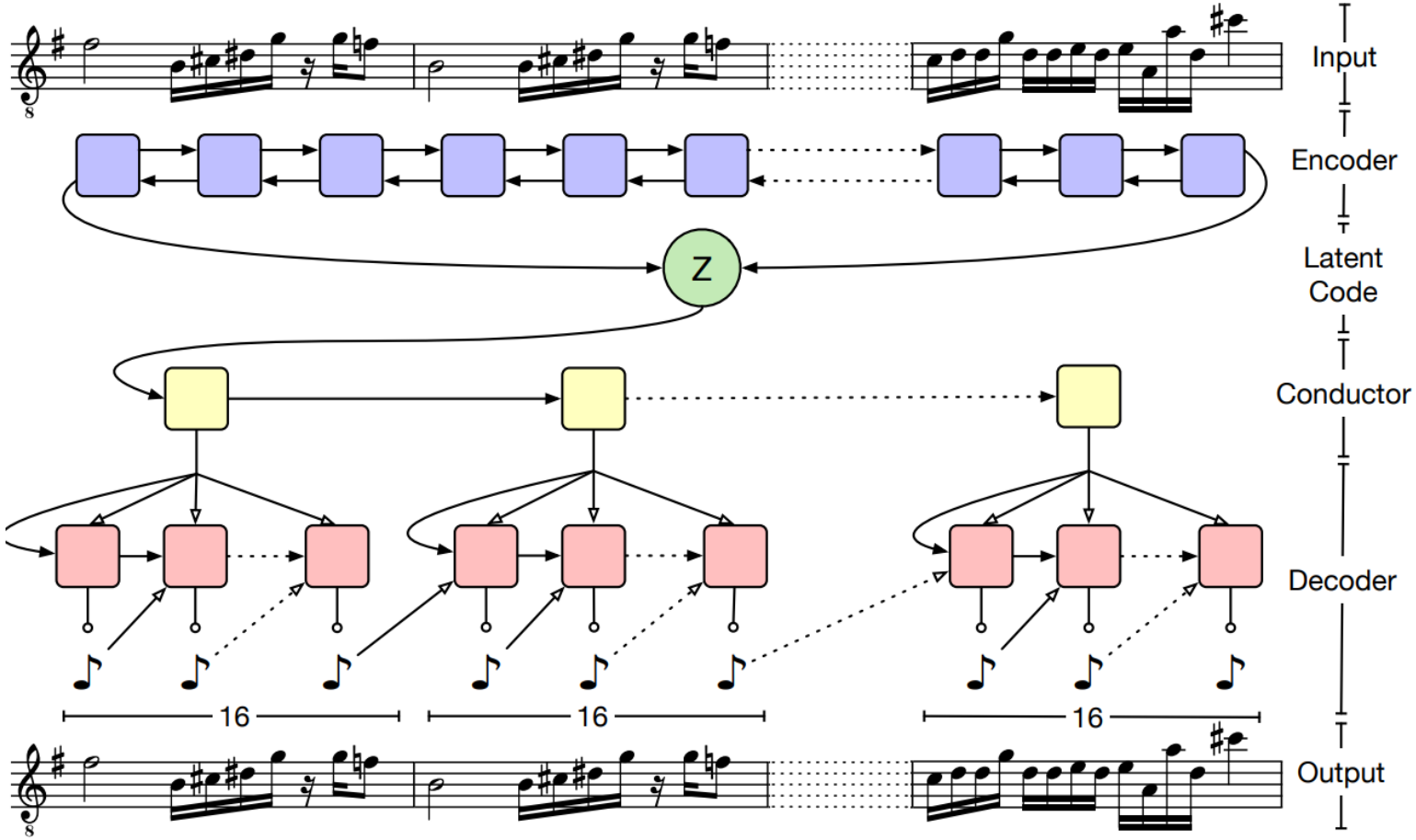
(Source: Wang et al., 2024)

wholesonggen.github.io

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

Latent Space-based Music Generation

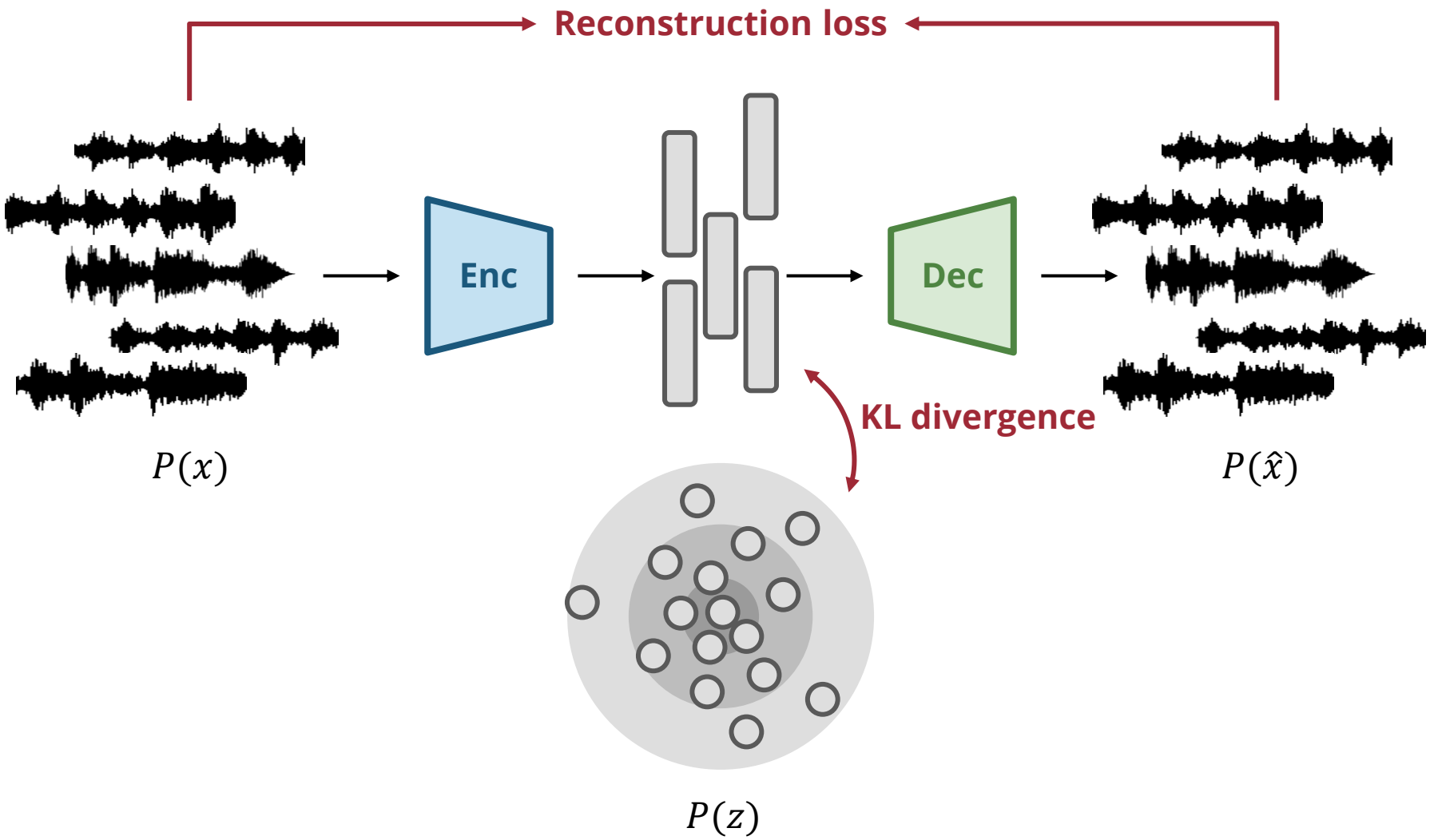
Example: MusicVAE (Roberts et al., 2018)



(Source: Roberts et al., 2018)

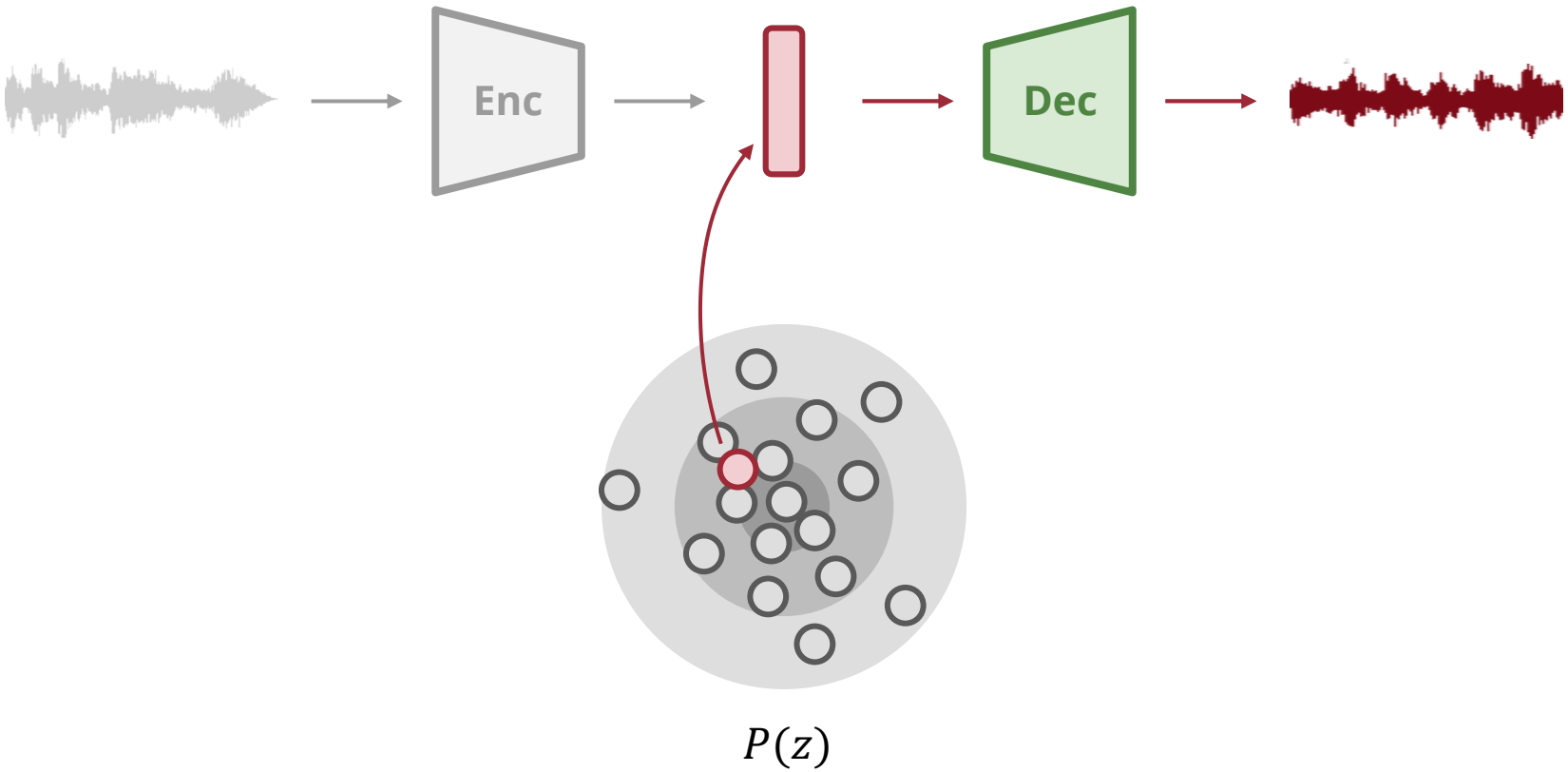
Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

(Recap) Variational Autoencoders (VAEs) – Training



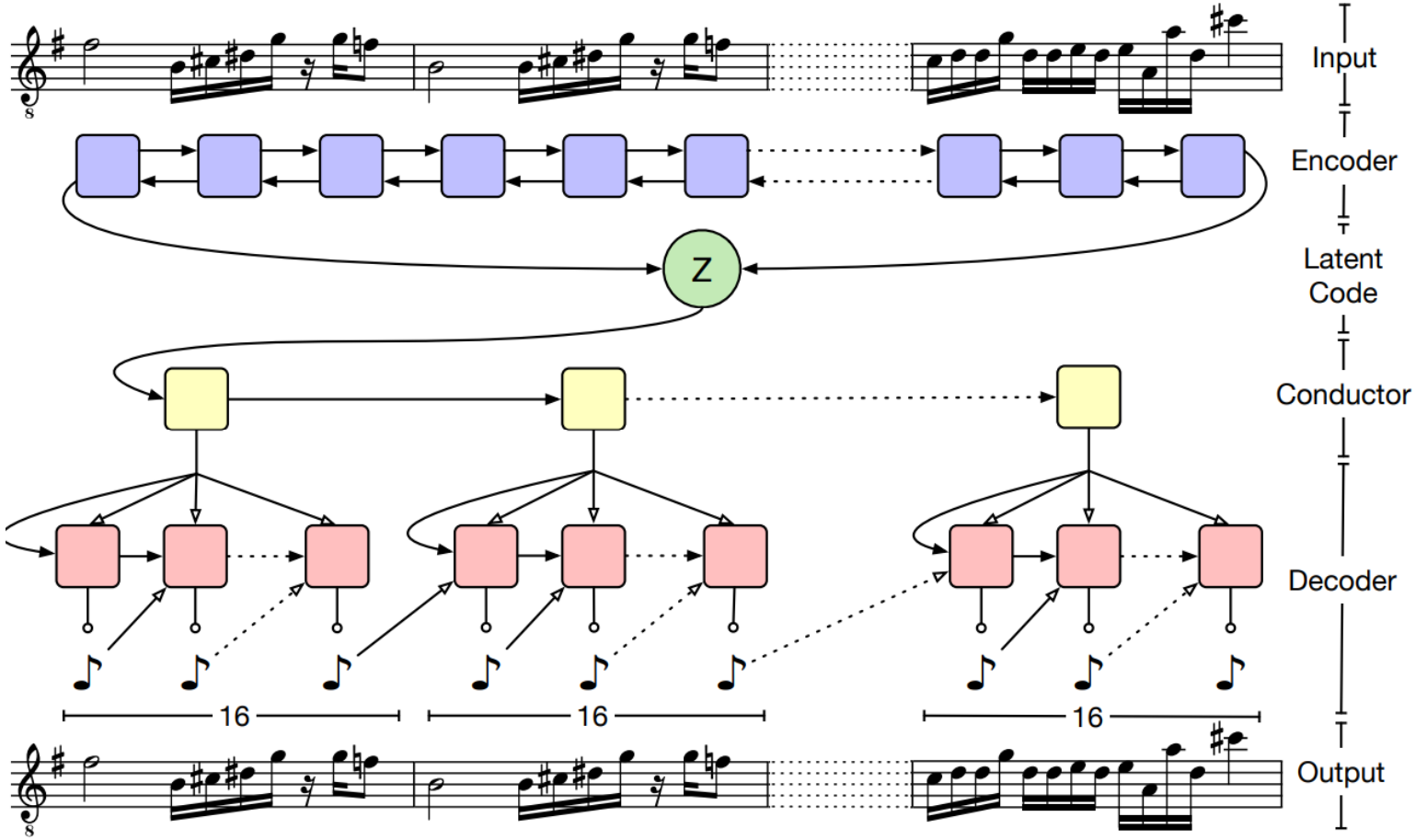
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

(Recap) Variational Autoencoders (VAEs) – Generation



Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

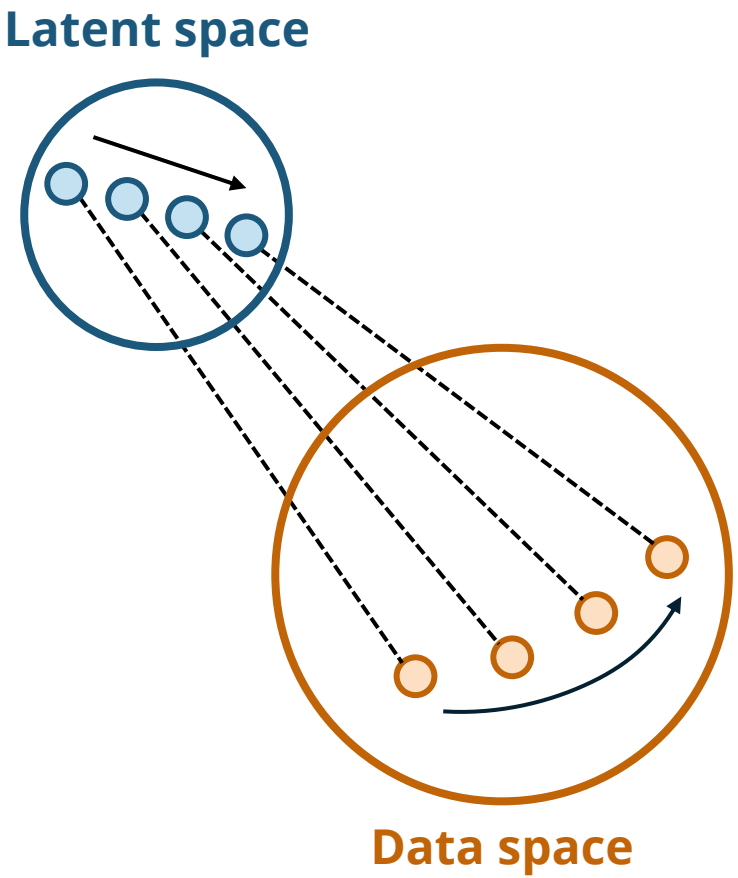
Example: MusicVAE (Roberts et al., 2018)

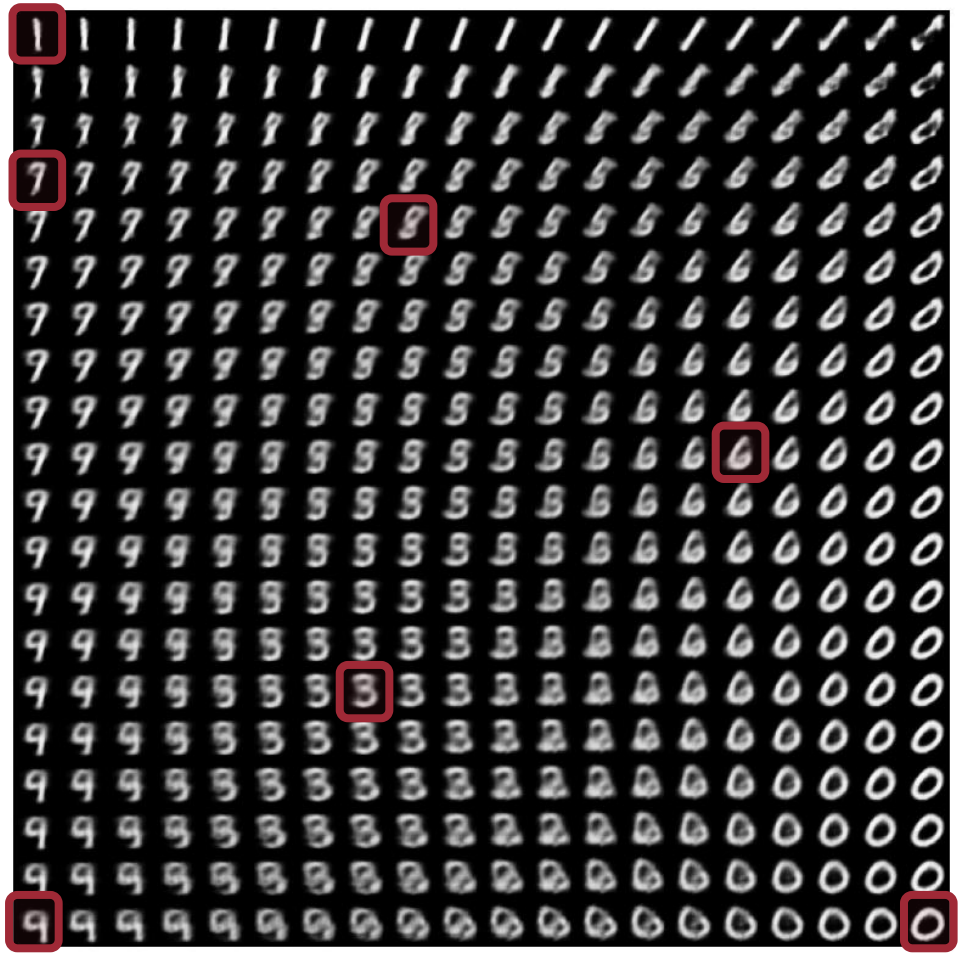


(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

(Recap) Decoding the Latent Space of a VAE

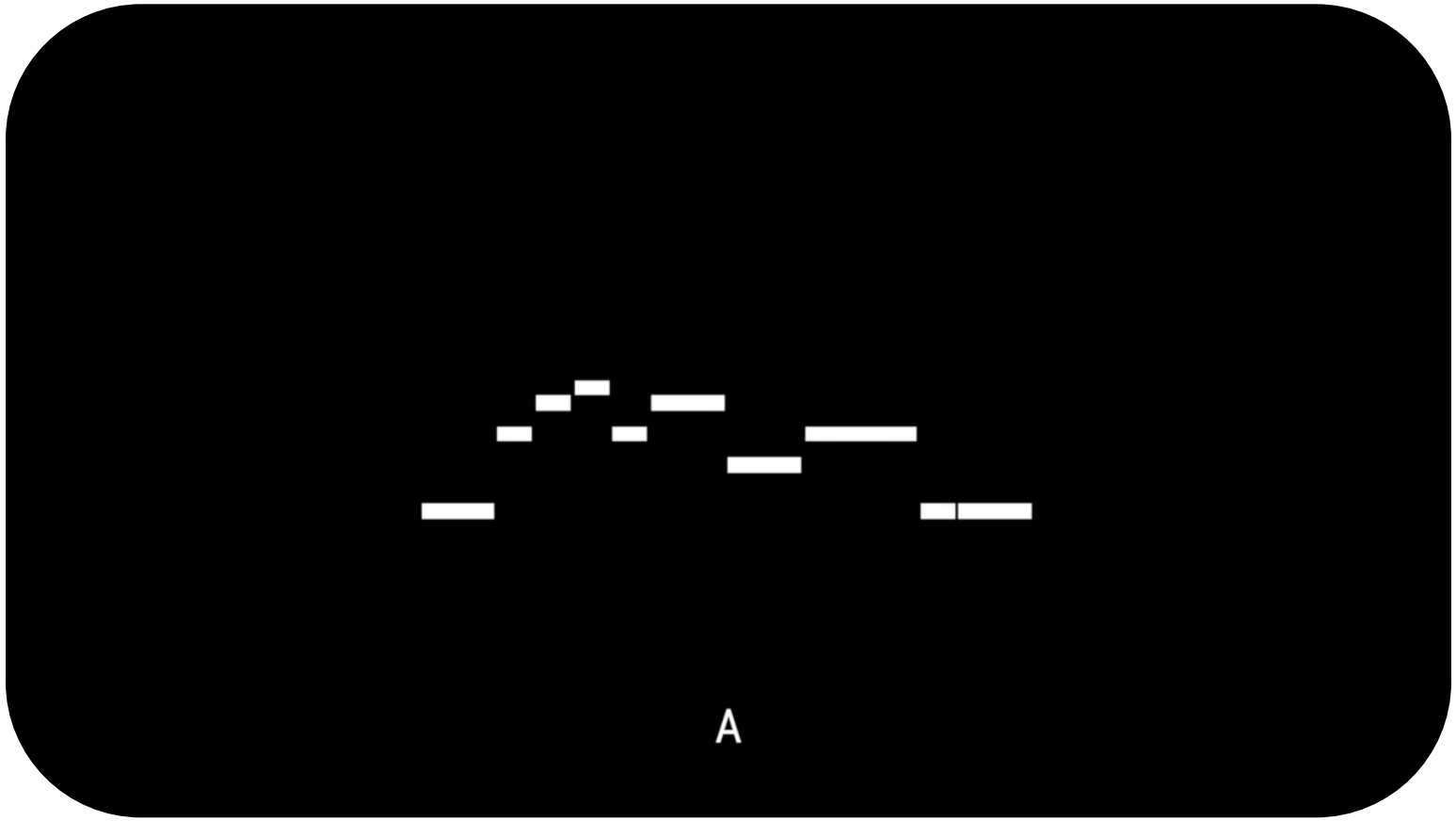




(Source: tensorflow.org)

tensorflow.org/tutorials/generative/autoencoder

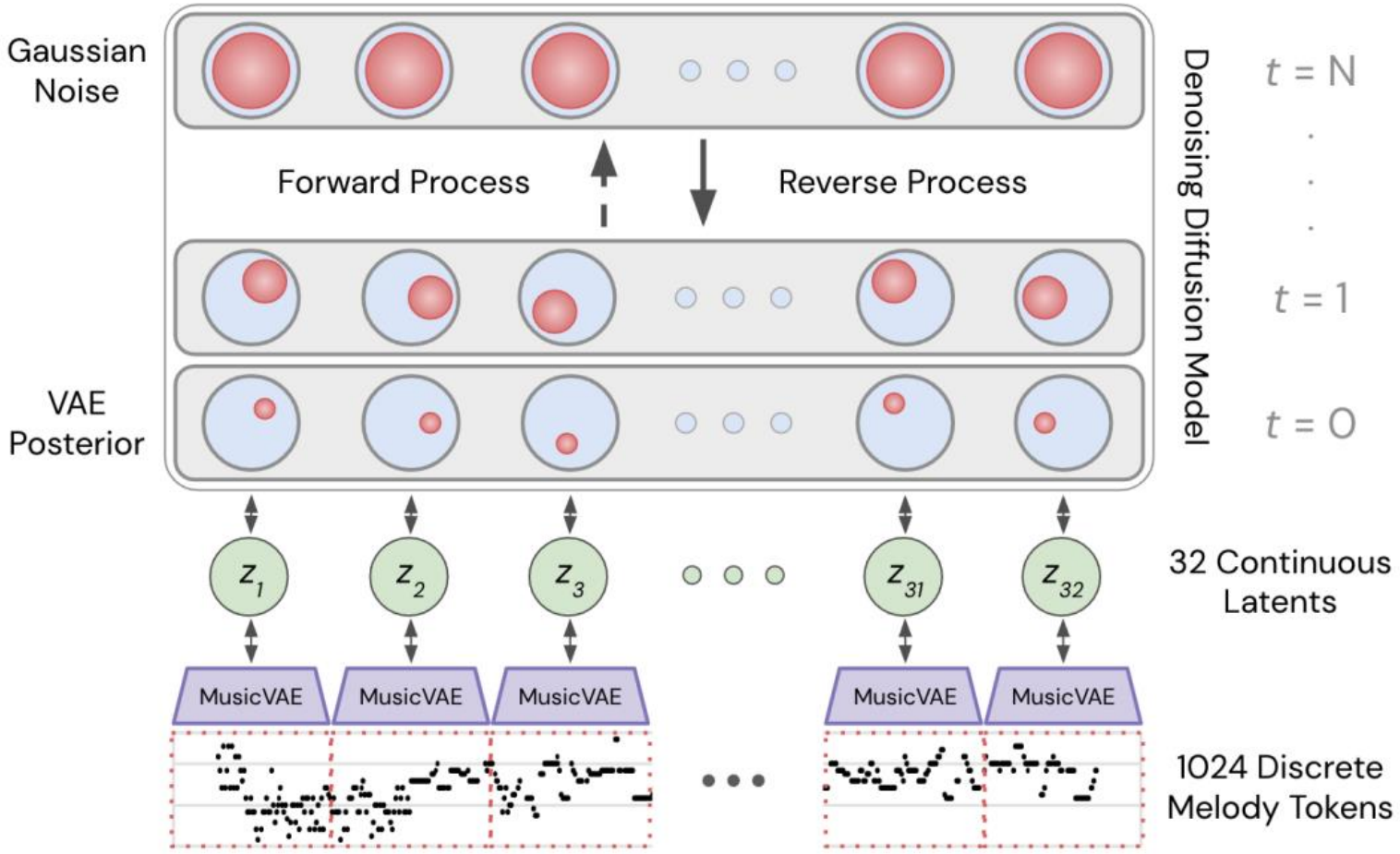
Example: MusicVAE (Roberts et al., 2018)



(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

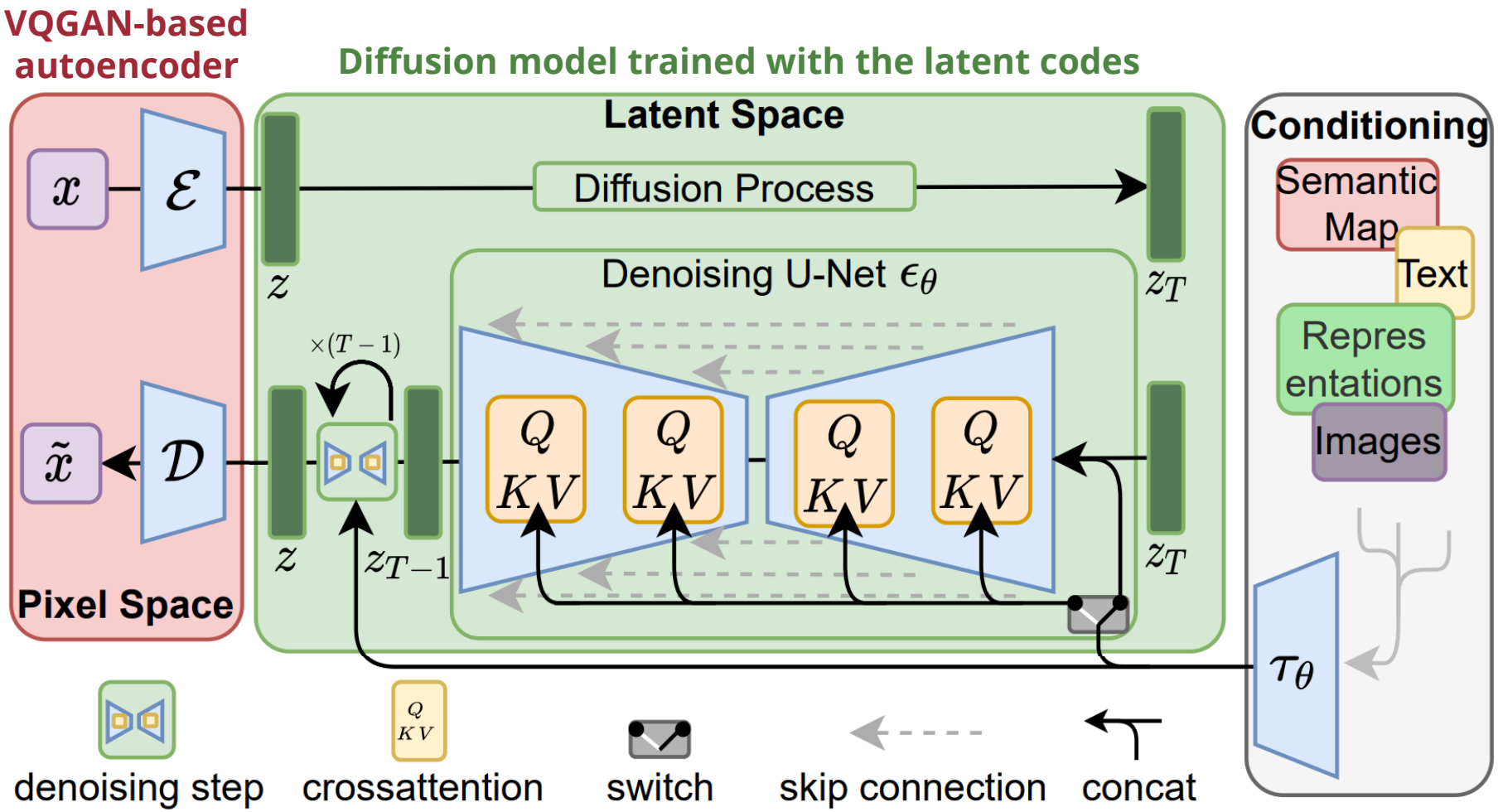
Example: Latent Diffusion (Mittal et al., 2021)



(Source: Mittal et al., 2021)

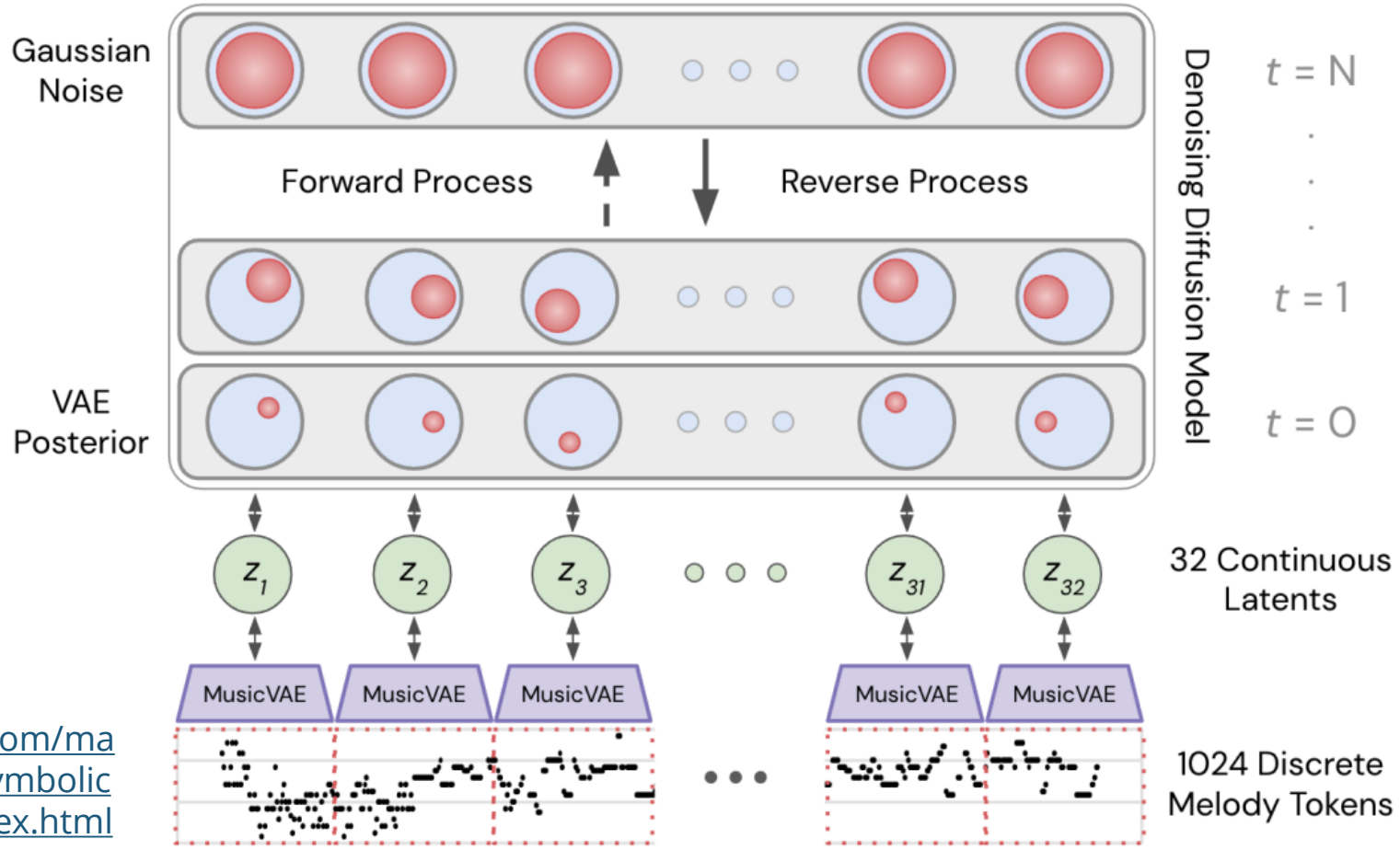
Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon, "Symbolic Music Generation with Diffusion Models," *ISMIR*, 2021.

(Recap) Latent Diffusion Models (LDMs)



(Source: Rombach et al., 2022)

Example: Latent Diffusion (Mittal et al., 2021)



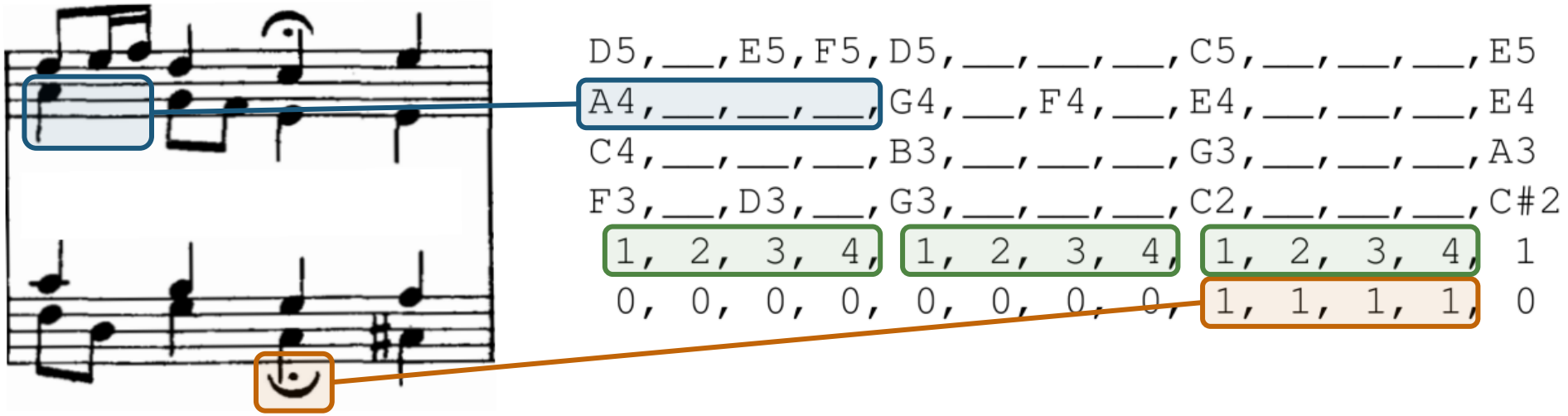
storage.googleapis.com/magentadata/papers/symbolic-music-diffusion/index.html

(Source: Mittal et al., 2021)

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon, "Symbolic Music Generation with Diffusion Models," *ISMIR*, 2021.

Music Infilling Models

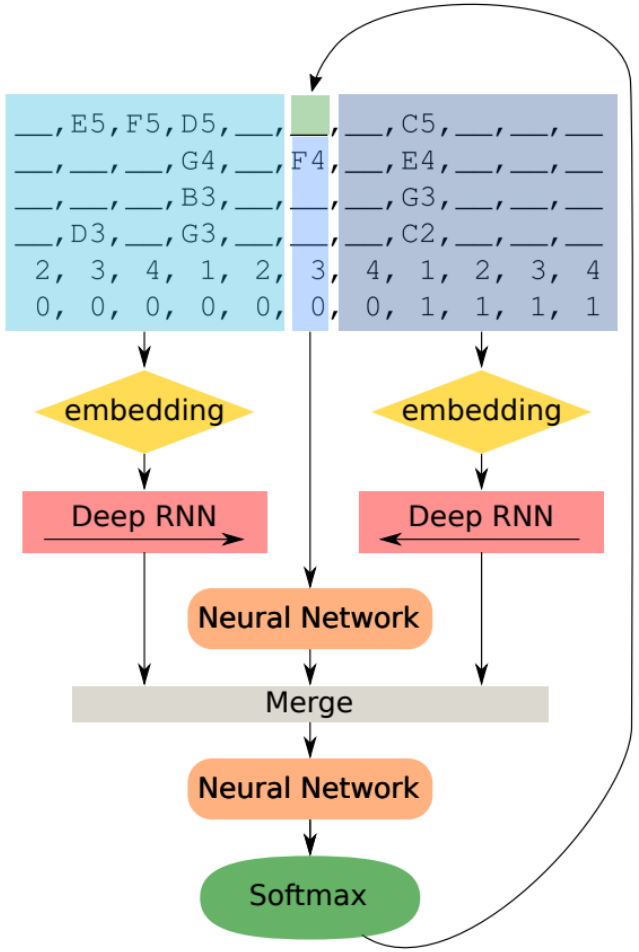
Example: DeepBach (Hadjeres et al., 2017)



(Source: Hadjeres et al., 2017)

Gaëtan Hadjeres, François Pachet, and Frank Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," *ICML*, 2017.

Example: DeepBach (Hadjeres et al., 2017)



(Source: Hadjeres et al., 2017)

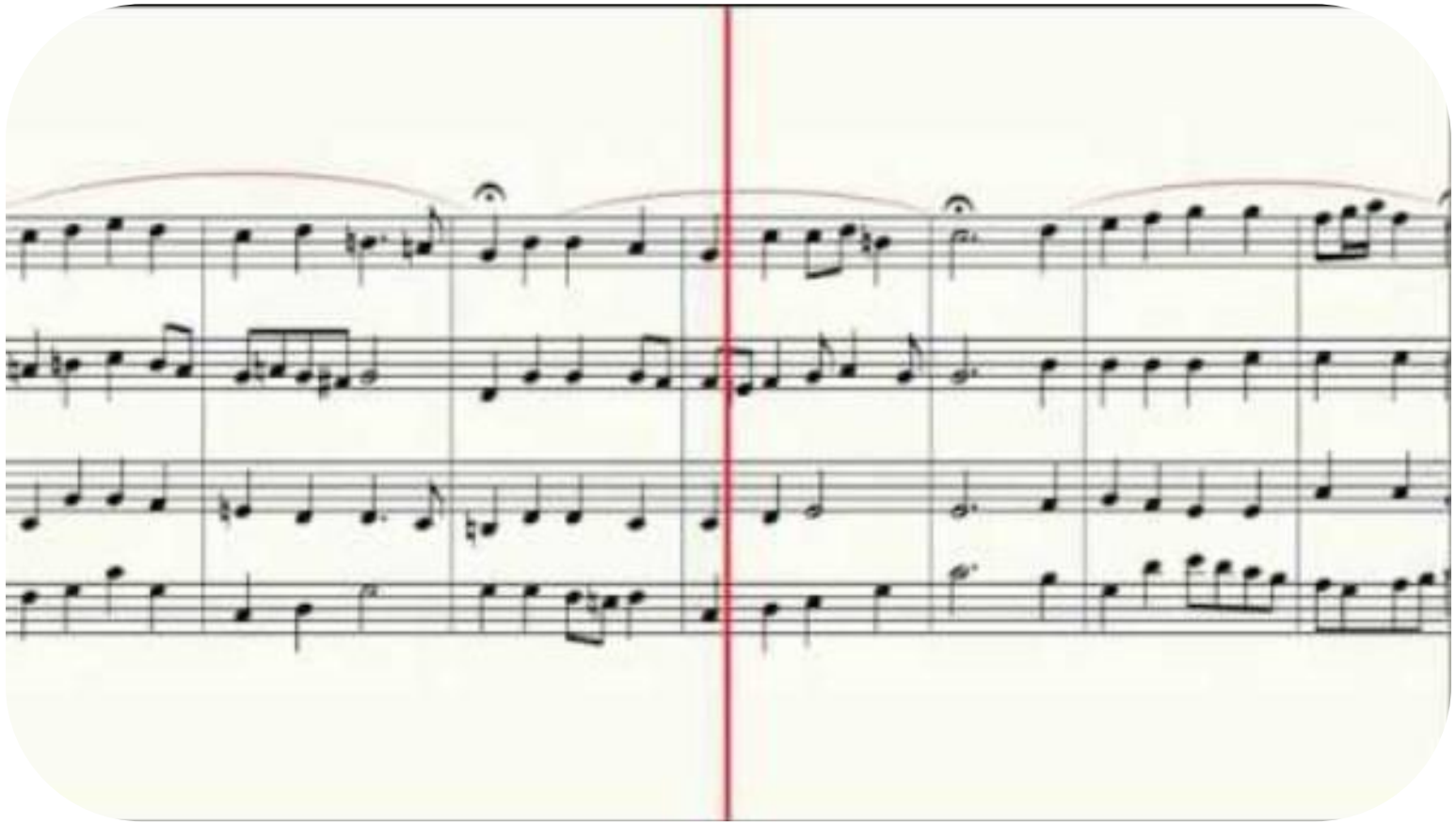
Algorithm 1 Pseudo-Gibbs sampling

Input: Chorale length L, metadata $\mathcal{M}$ containing lists of length L, probability distributions (p_1, p_2, p_3, p_4), maximum number of iterations M
Create four lists $\mathcal{V} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \mathcal{V}_4)$ of length L
{The lists are initialized with random notes drawn from the ranges of the corresponding voices (sampled uniformly or from the marginal distributions of the notes)}
for m from 1 to M **do**
    Choose voice i uniformly between 1 and 4
    Choose time t uniformly between 1 and L
    Re-sample $\mathcal{V}_i^t$ from $p_i(\mathcal{V}_i^t | \mathcal{V}_{\setminus i,t}, \mathcal{M}, \theta_i)$
end for
Output: $\mathcal{V} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \mathcal{V}_4)$

Gaëtan Hadjeres, François Pachet, and Frank Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," *ICML*, 2017.

Example: DeepBach (Hadjeres et al., 2017)

Reharmonization example

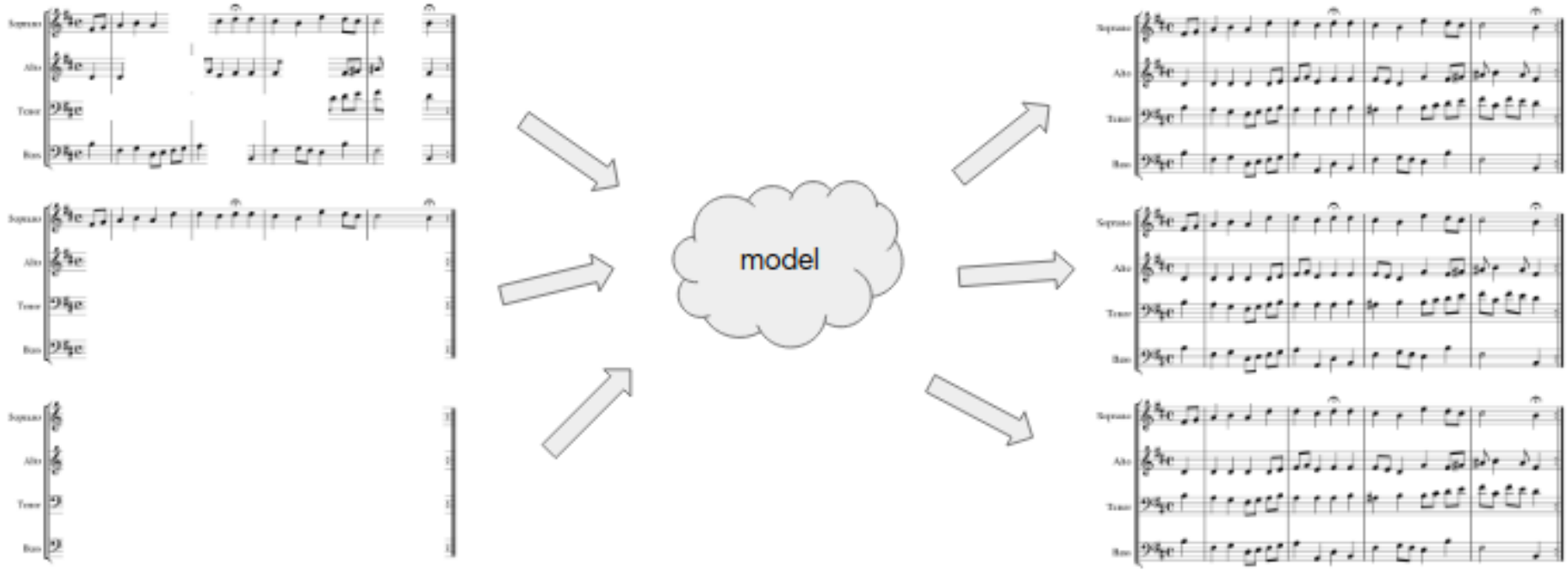


youtu.be/QiBM7-5hA6o

Gaëtan Hadjeres, François Pachet, and Frank Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," *ICML*, 2017.

Example: Coconet (Huang et al., 2017)

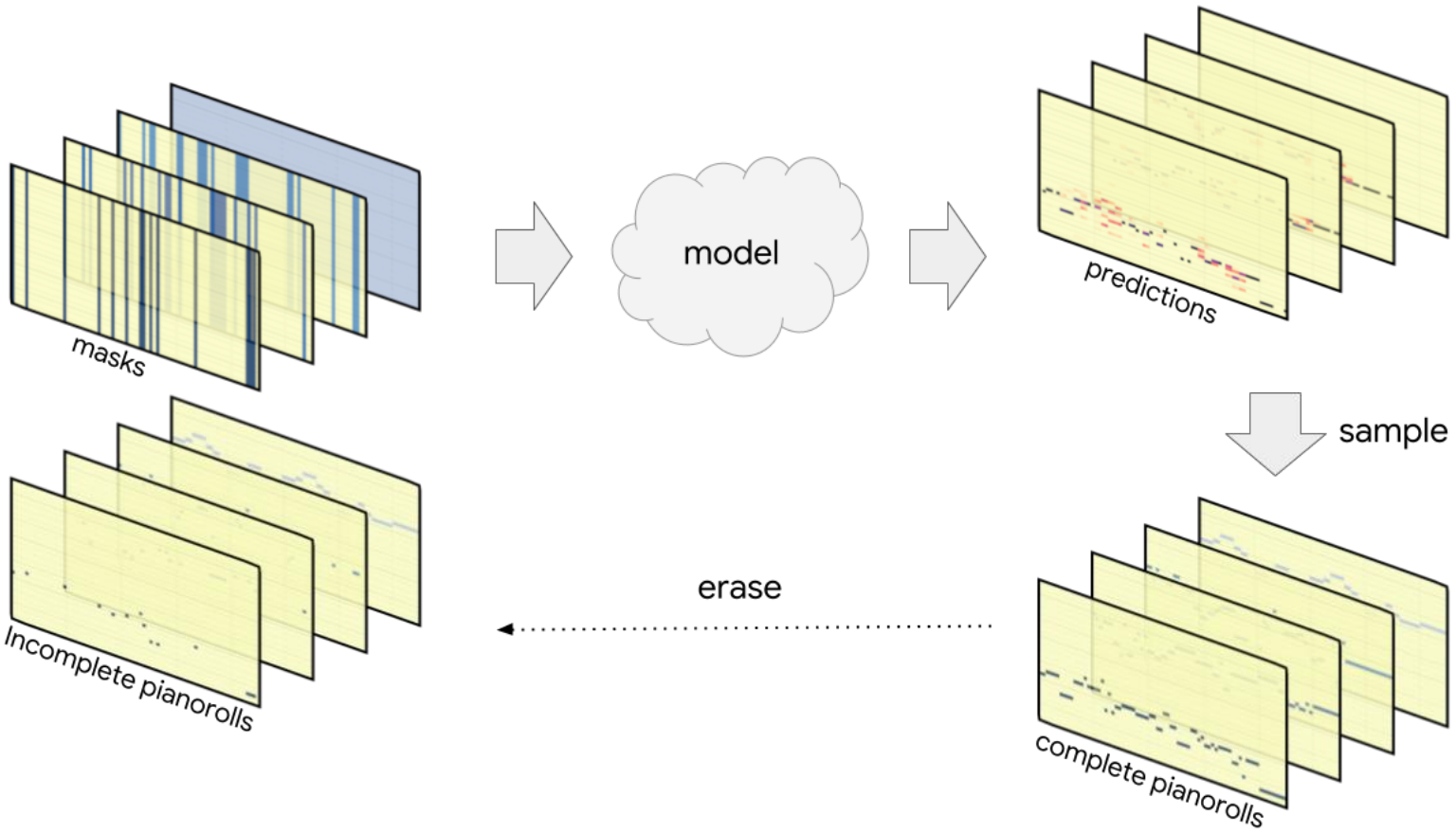
- Based on Orderless NADE (Uria et al, 2014)



(Source: Huang et al., 2019)

Benigno Uria, Iain Murray, and Hugo Larochelle, "A Deep and Tractable Density Estimator," *ICML*, 2014.
Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.
Cheng-Zhi Anna Huang, Tim Cooijmans, Monica Dinculescu, Adam Roberts, and Curtis Hawthorne, "Coconet: the ML model behind today's Bach Doodle," *Magenta Blog*, 2019.

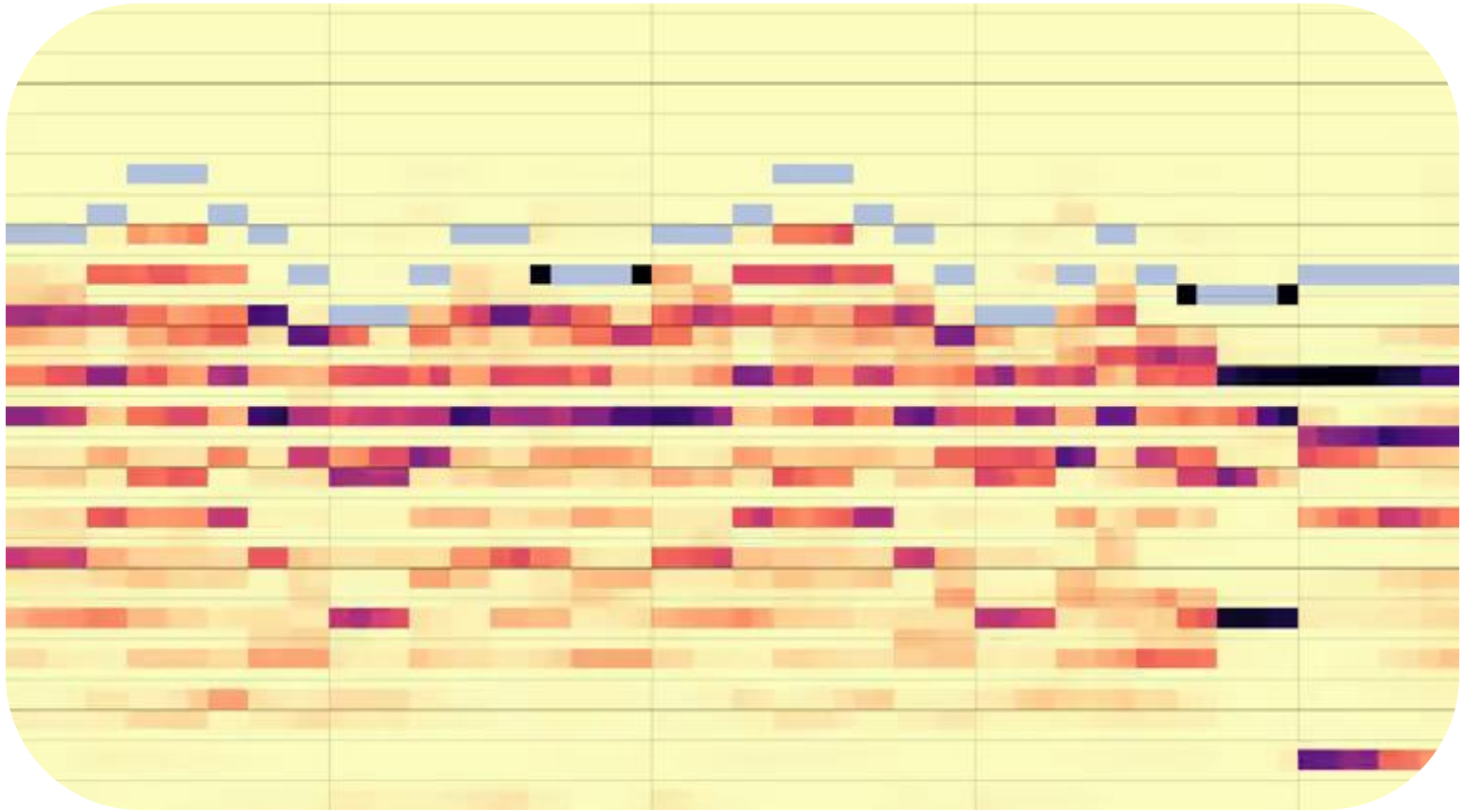
Example: Coconet (Huang et al., 2017)



(Source: Huang et al., 2019)

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.
Cheng-Zhi Anna Huang, Tim Cooijmans, Monica Dinculescu, Adam Roberts, and Curtis Hawthorne, "Coconet: the ML model behind today's Bach Doodle," *Magenta Blog*, 2019.

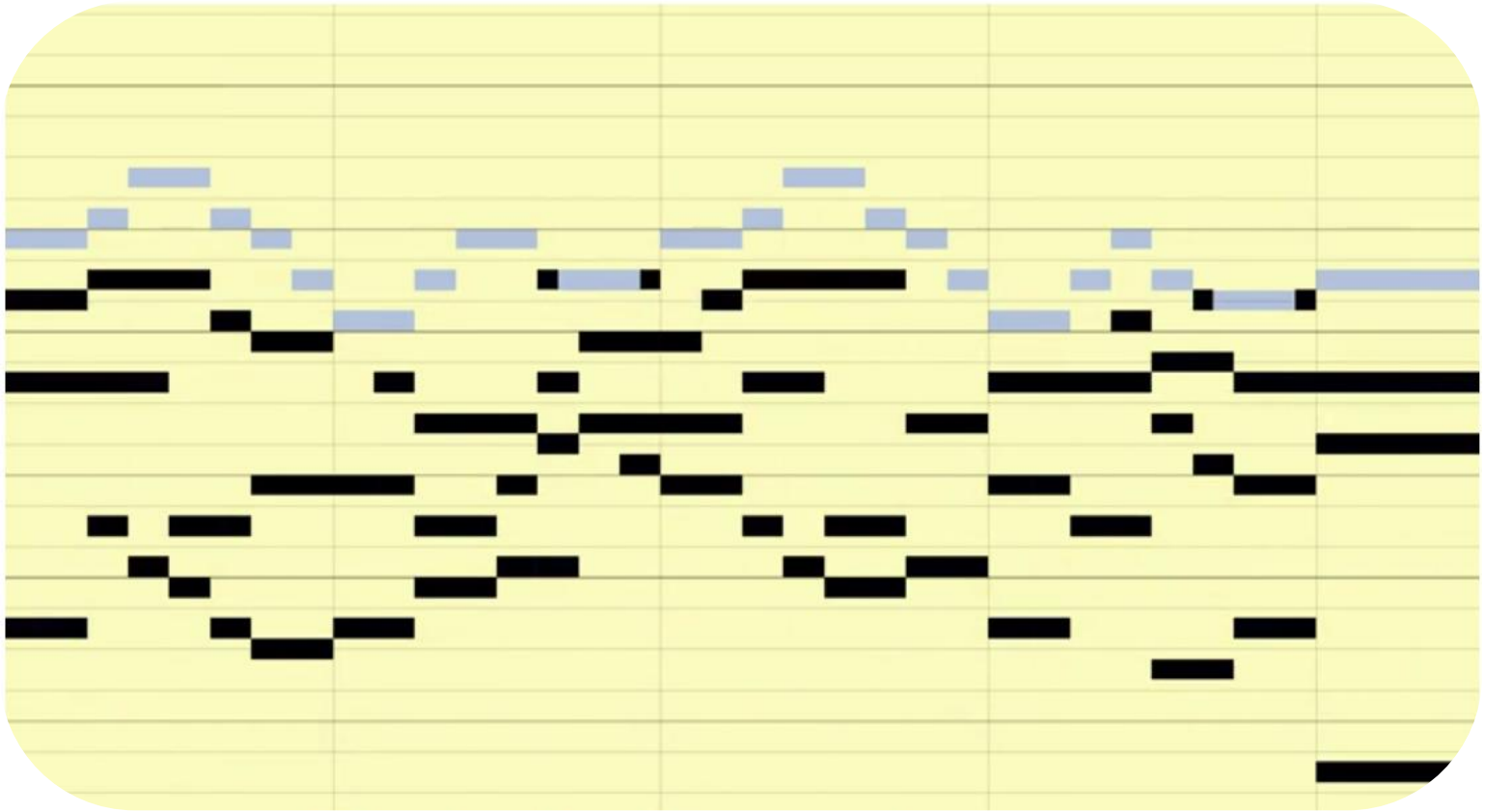
Example: Coconet (Huang et al., 2017)



(Source: Huang et al., 2017)

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.

Example: Coconet (Huang et al., 2017)



(Source: Huang et al., 2017)

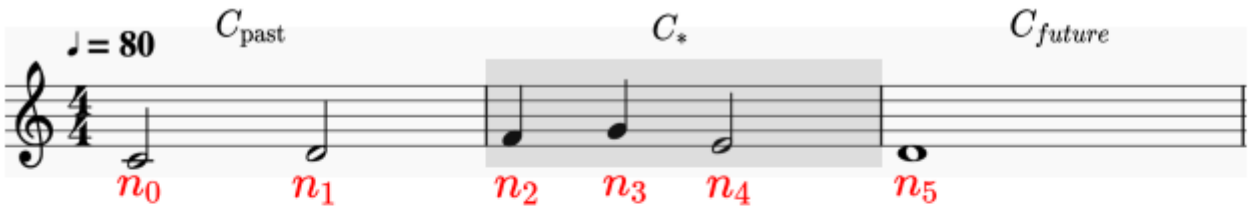
Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.

Example: JS Bach Doodle (2019)

doodles.google/doodle/celebrating-johann-sebastian-bach/

youtu.be/XBfYPp6KF2g & magenta.tensorflow.org/coconet

Example: Variable-Length Infilling (VLI) (Chang et al., 2021)



ILM

Input:	n_0	n_1	[BLANK]	n_5	[SEP]
target:	n_2	n_3	n_4	[EOS]	

FELIX

Input:	n_0	n_1	[MASK]	[MASK]	[MASK]	[MASK]	[MASK]	n_5
target:			n_2	n_3	n_4	[PAD]	[PAD]	

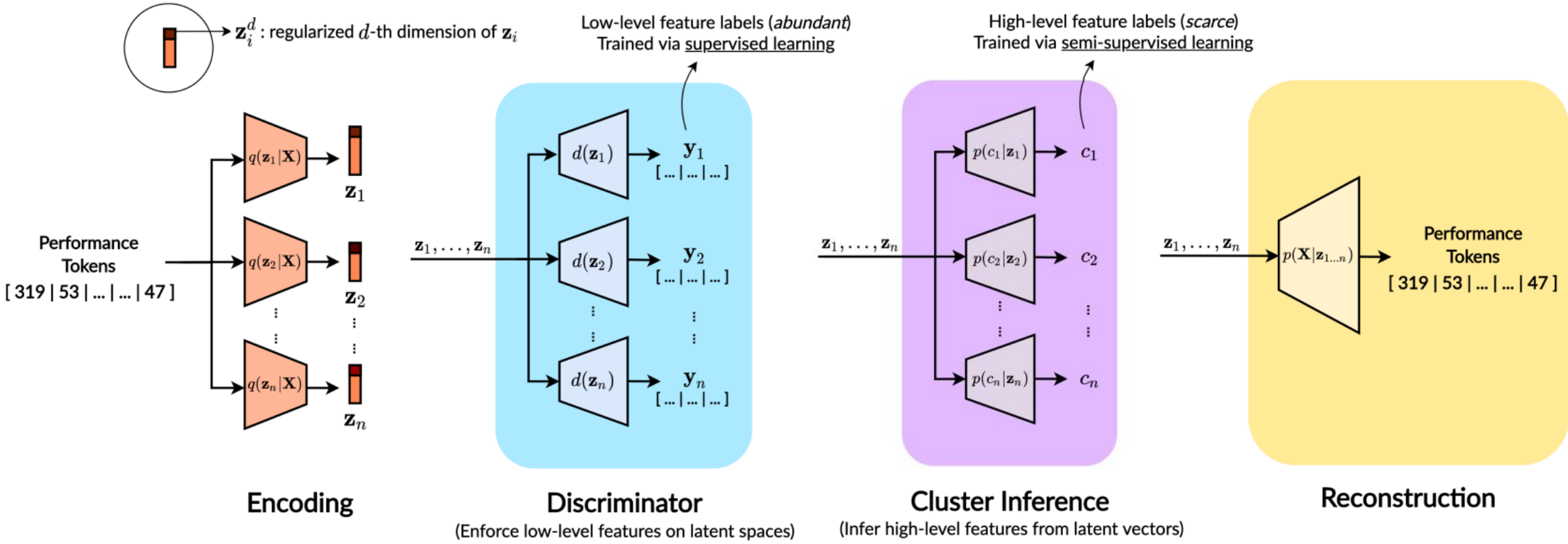
(Source: Chang et al., 2021)

jackyhsiung.github.io/piano-infilling-demo

Chin-Jui Chang, Chun-Yi Lee, and Yi-Hsuan Yang, "Variable-Length Music Score Infilling via XLNet and Musically Specialized Positional Encoding," *ISMIR*, 2021.
Chris Donahue, Mina Lee, and Percy Liang, "Enabling Language Models to Fill in the Blanks," *ACL*, 2020.
Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido, "Felix: Flexible Text Editing Through Tagging and Insertion," *Findings of ACL*, 2020.

Controllable Music Generation

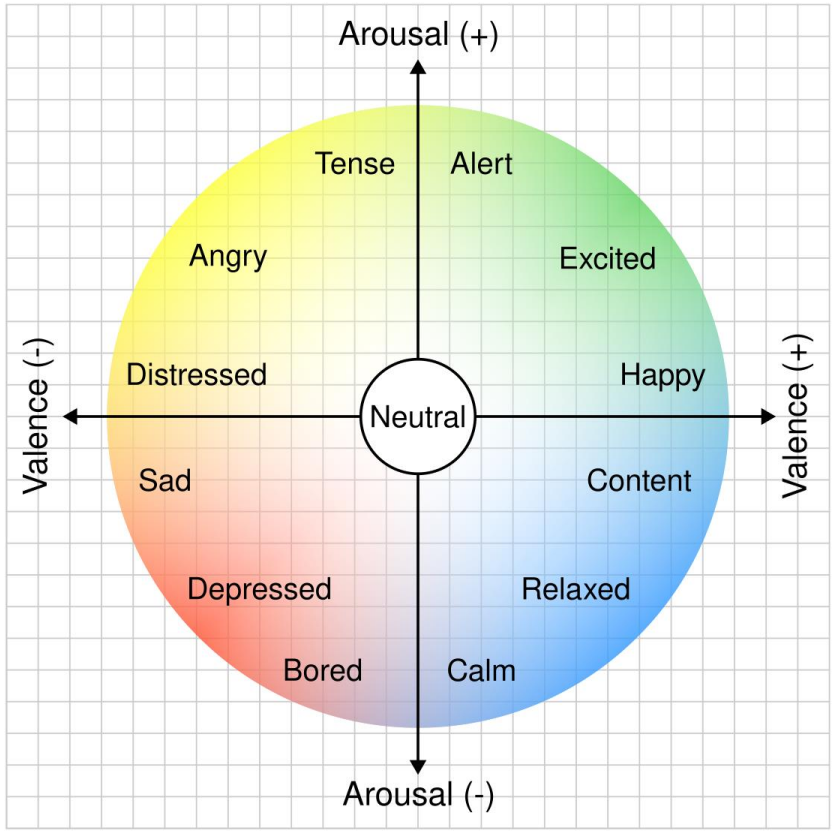
Example: Music FaderNet (Tan & Herremans, 2020)



(Source: Tan & Herremeans, 2020)

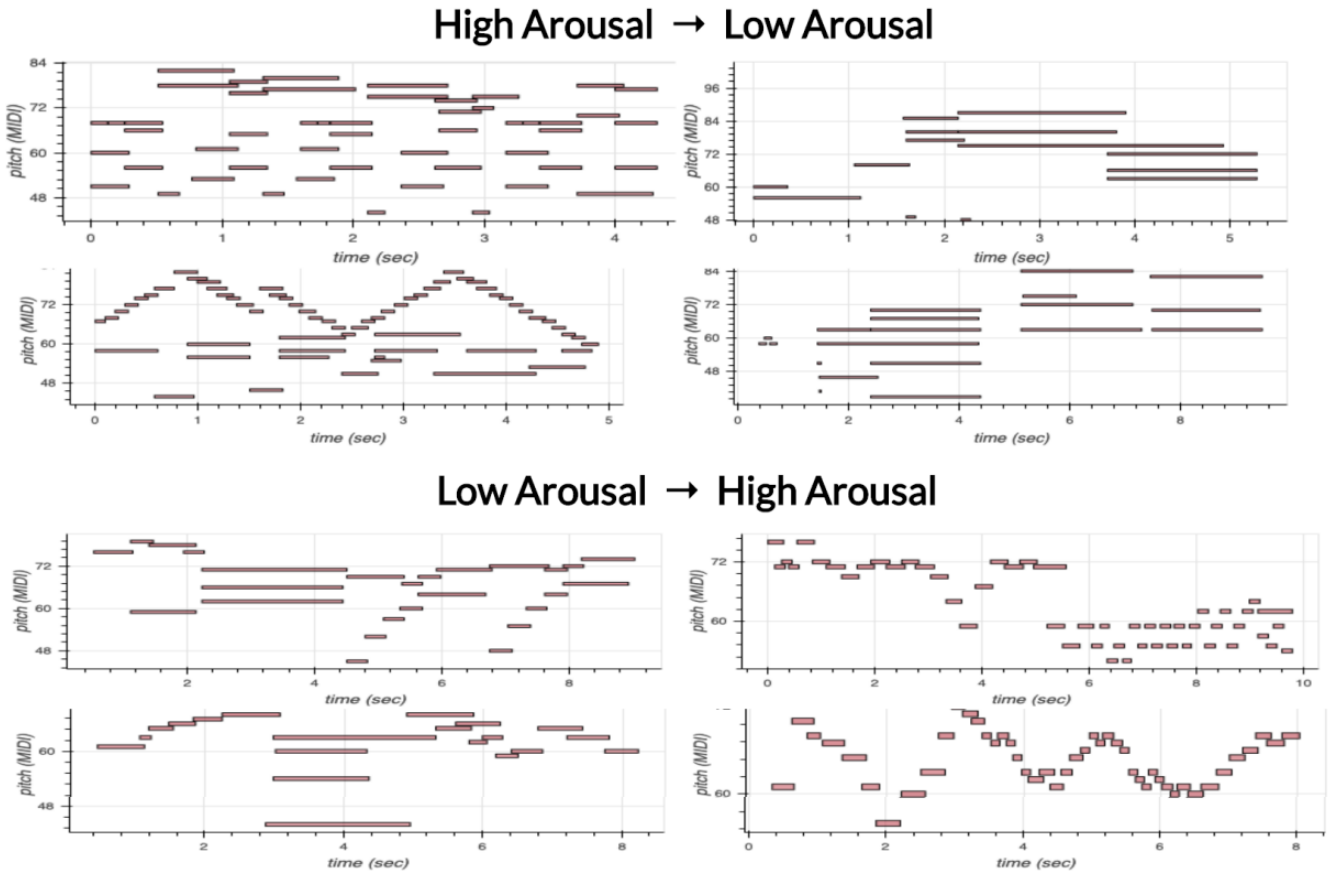
Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Valence-Arousal Model for Emotion



(Source: mrAnmol)

Example: Music FaderNet (Tan & Herremans, 2020)

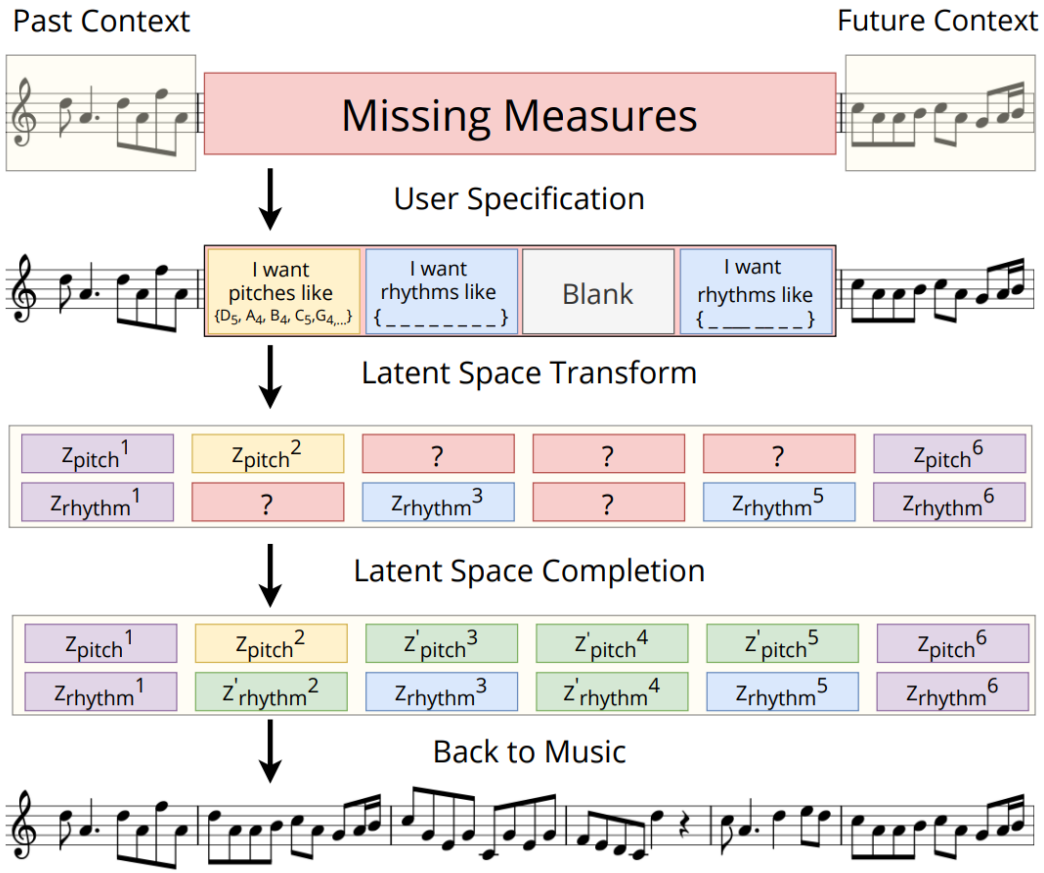


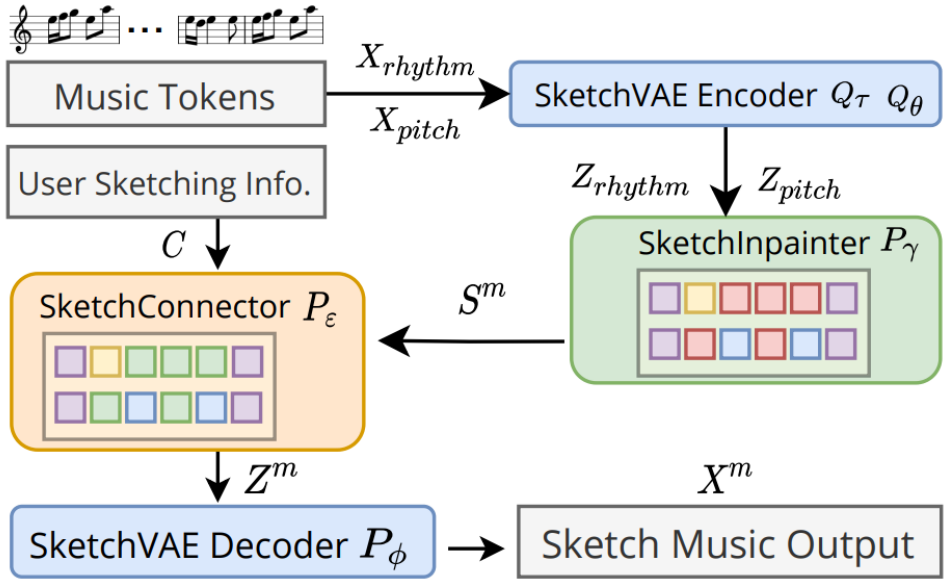
(Source: Tan & Herremeans, 2020)

music-fadernets.github.io

Hao Hao Tan and Dorien Herremans, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," ISMIR, 2020.

Example: Music SketchNet (Chen et al., 2020)

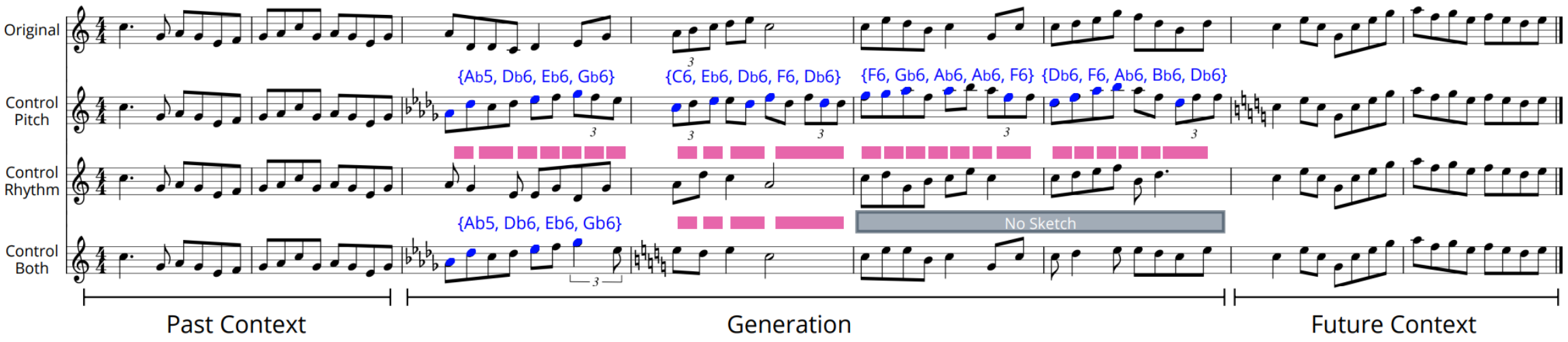




(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

Example: Music SketchNet (Chen et al., 2020)

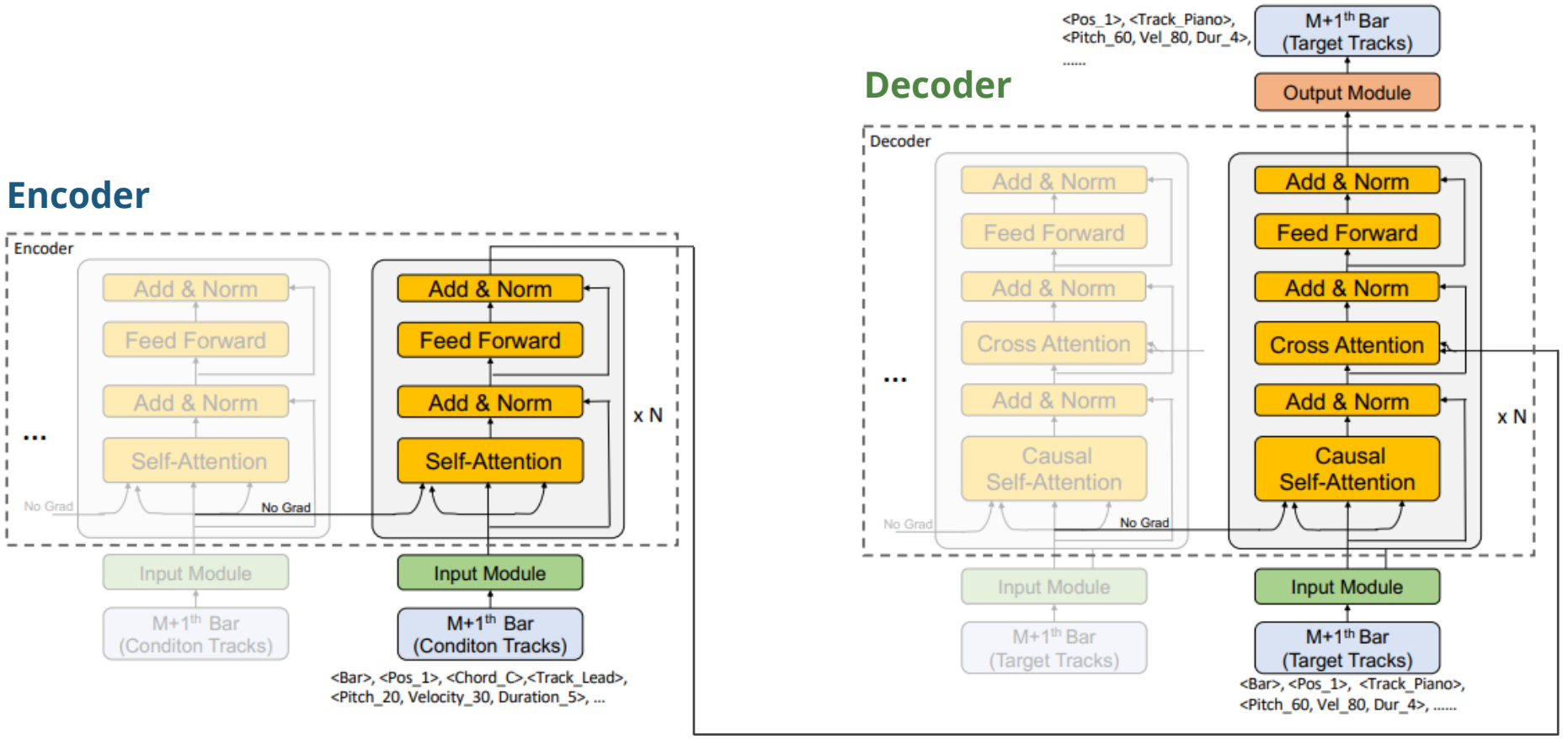


(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

Music Accompaniment

Example: PopMAG (Ren et al., 2020)



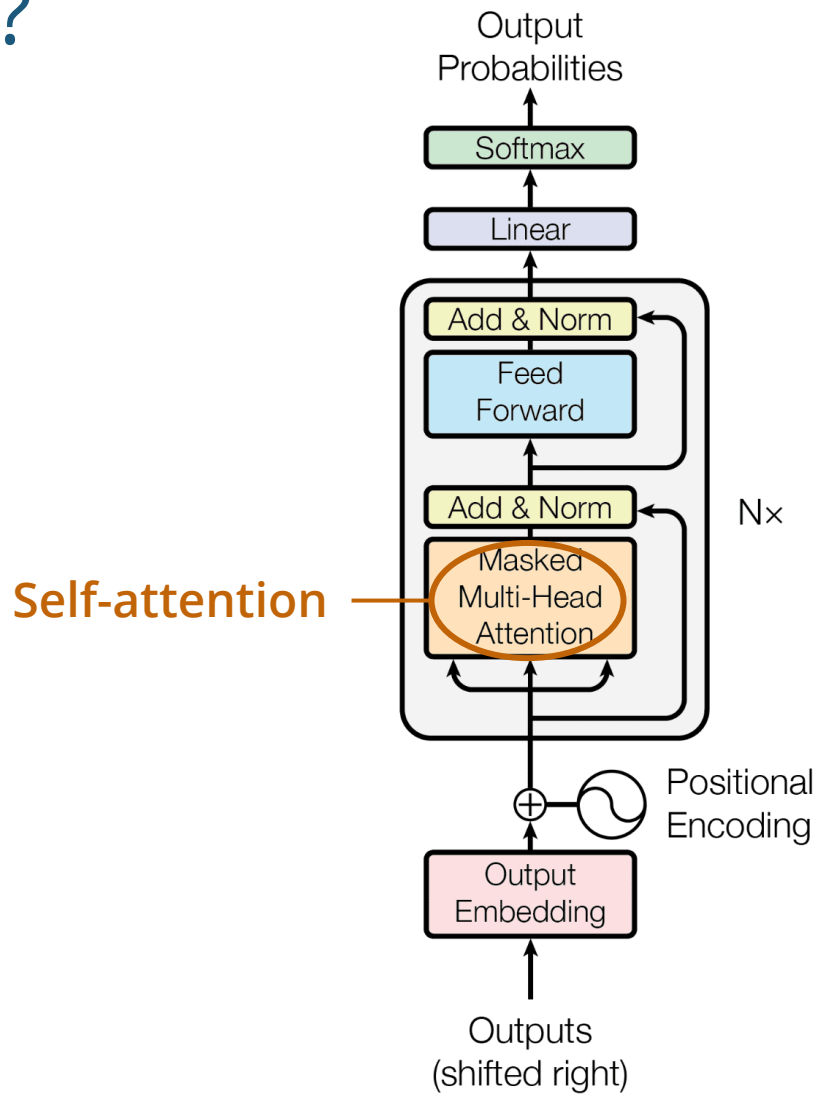
(Source: Ren et al., 2020)

ai-muzic.github.io/popmag

Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, "PopMAG: Pop Music Accompaniment Generation," *MM*, 2020.

(Recap) What is a Transformer?

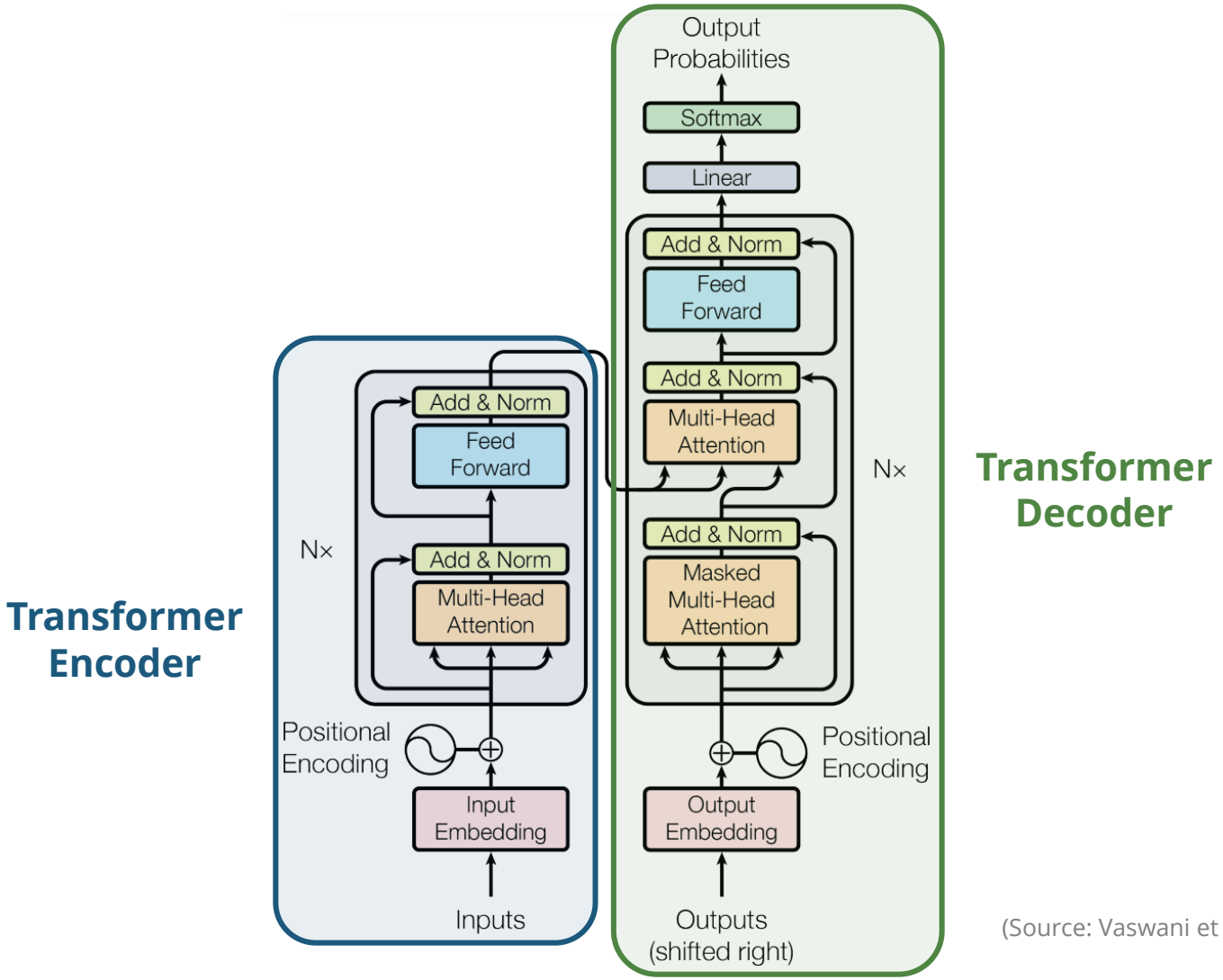
- A type of neural network that use the **self-attention mechanism**



(Source: Vaswani et al., 2017; adapted)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

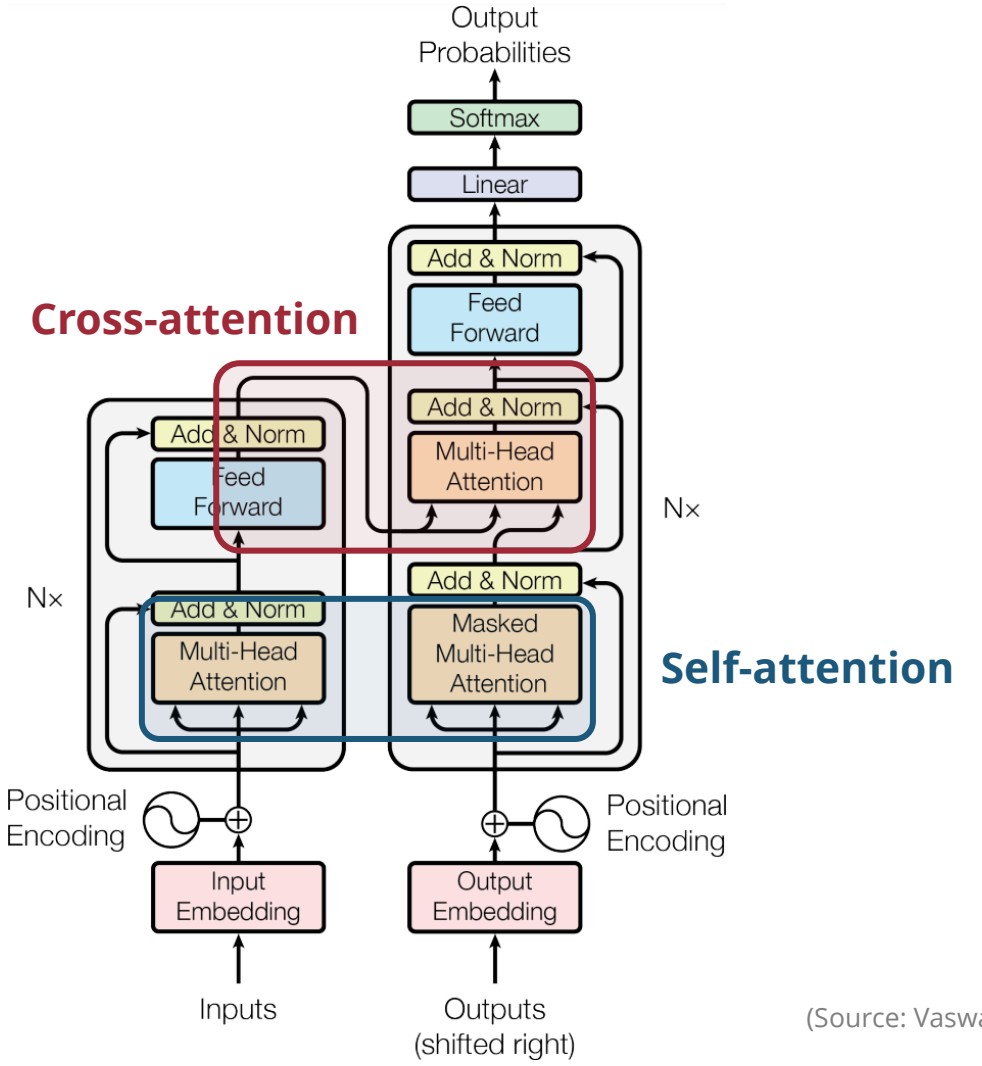
The Original Transformer – Encoder & Decoder



(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

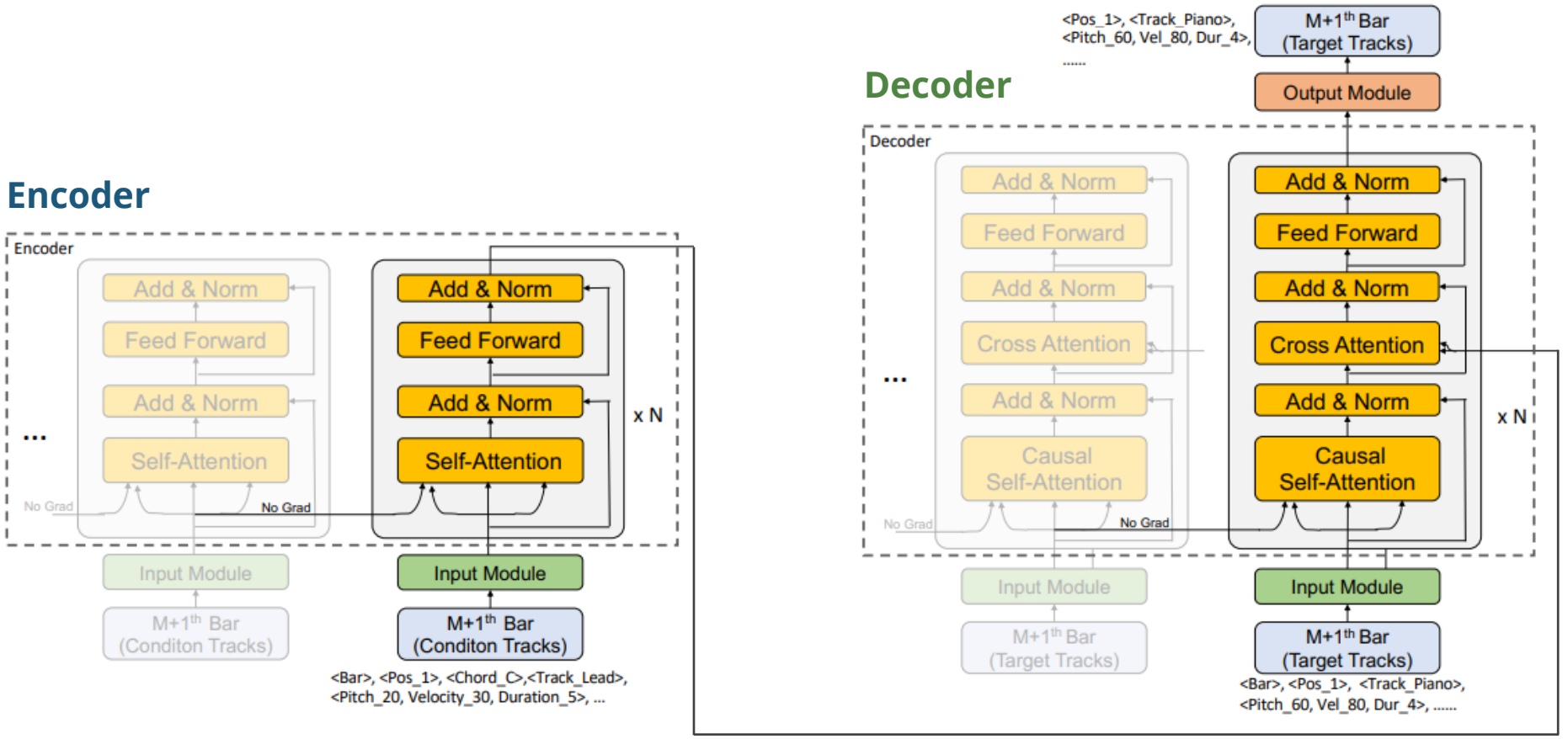
The Original Transformer – Cross-attention



(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Example: PopMAG (Ren et al., 2020)



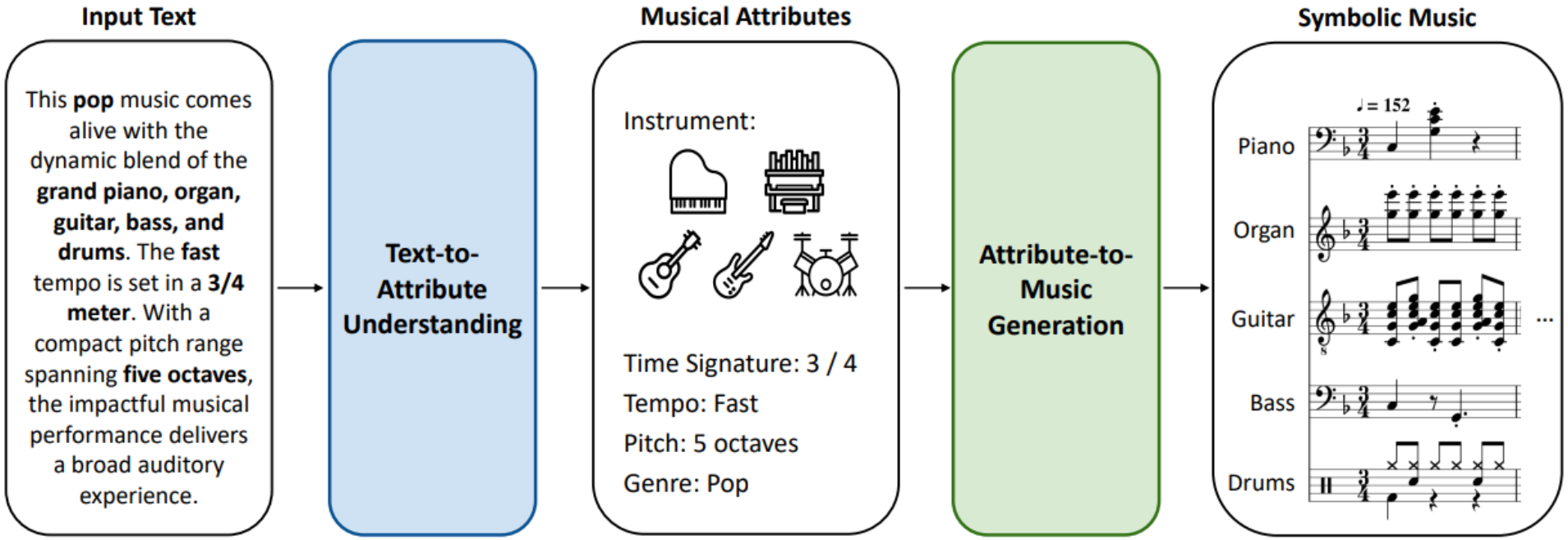
(Source: Ren et al., 2020)

ai-muzic.github.io/popmag

Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, "PopMAG: Pop Music Accompaniment Generation," *MM*, 2020.

Multimodal Music Generation

Example: MuseCoco (Lu et al., 2023)



(Source: Lu et al., 2023)

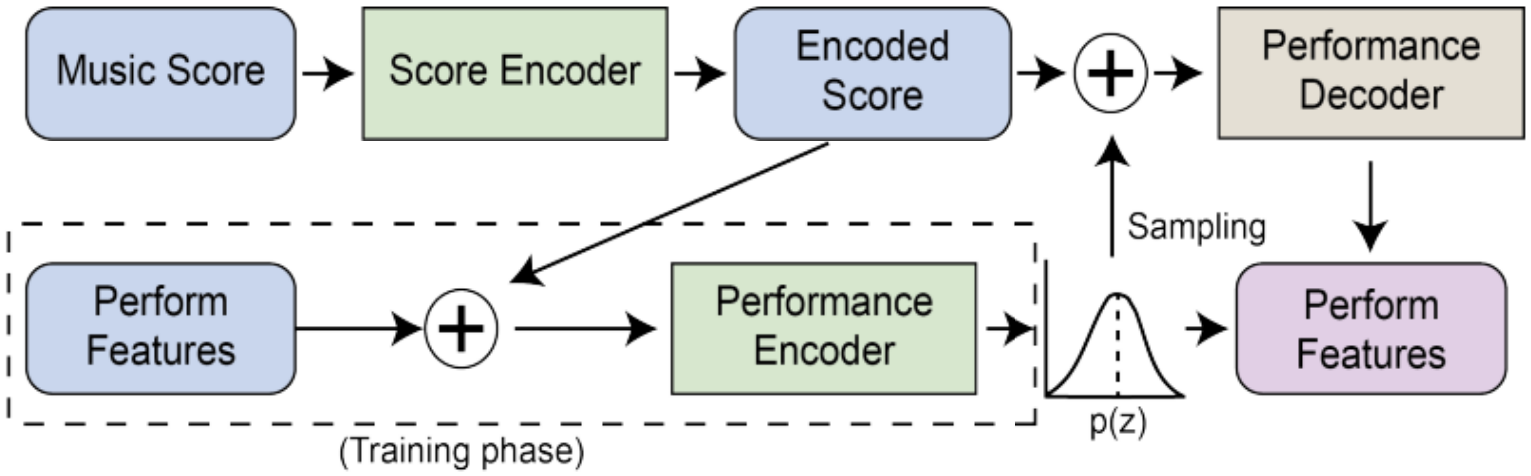
ai-muzic.github.io/musecoco

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian, "MuseCoco: Generating Symbolic Music from Text," *arXiv preprint arXiv:2306.00110*, 2023.

Performance Rendering

Example: VirtuosoNet (Jeong et al., 2019)

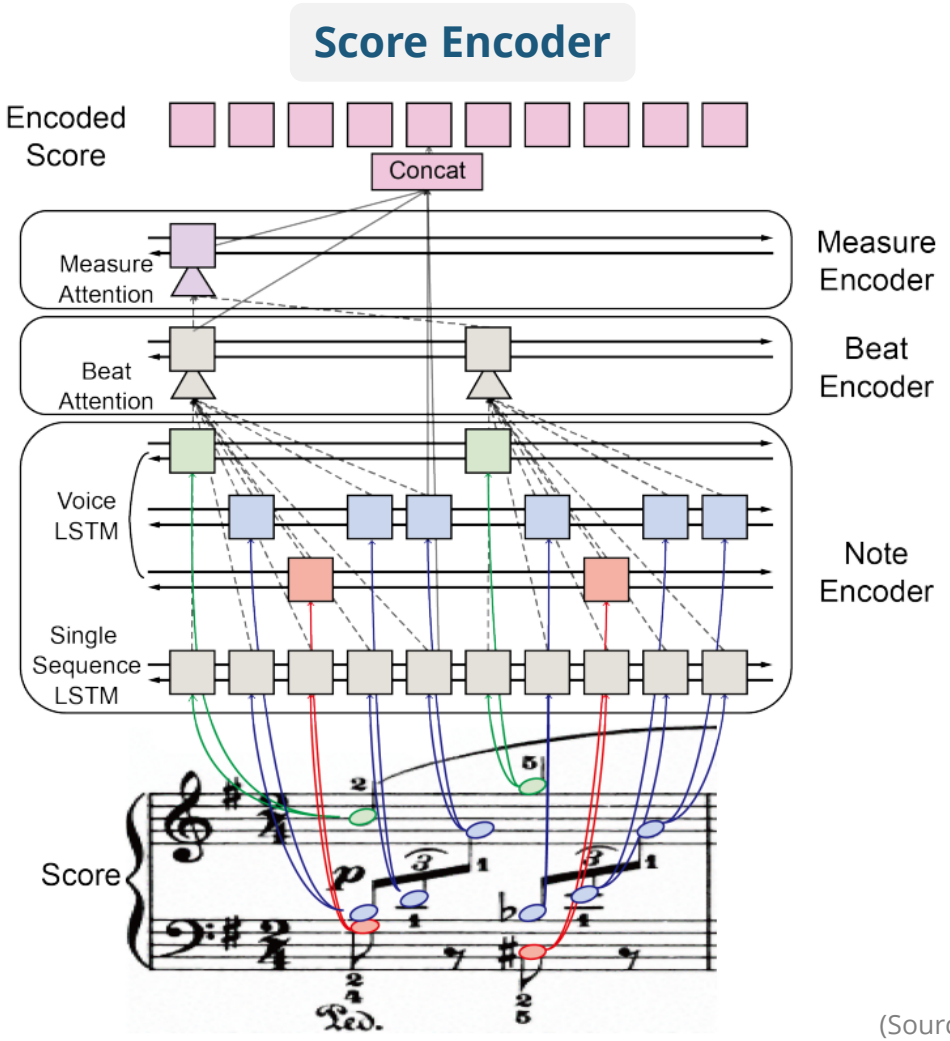
- **Input**: pitch, duration, articulation marking, slur and beam status, tempo marking, and dynamic marking, etc.
- **Output**: absolute tempo, velocity, onset deviation, articulation, pedal usages

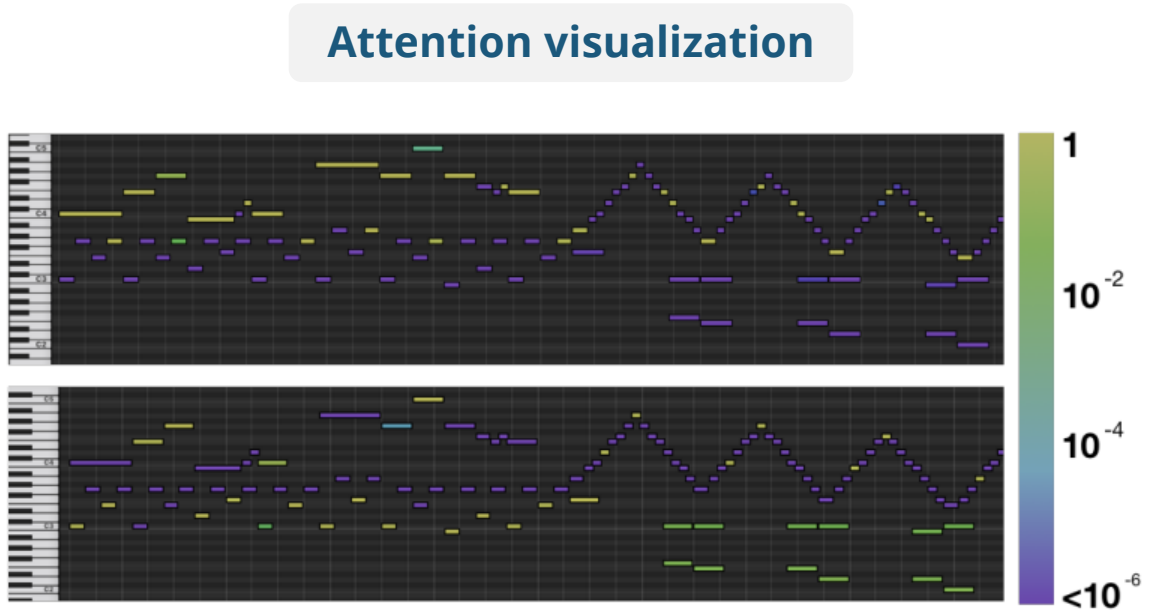


(Source: Jeong et al., 2019)

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

Example: VirtuosoNet (Jeong et al., 2019)





(Source: Jeong et al., 2019)

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

Example: VirtuosoNet (Jeong et al., 2019)

youtu.be/6HeFJQf2h2o

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

Example: VirtuosoNet (Jeong et al., 2019)

youtu.be/BN0ZCBS9q0Y

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

Example: VirtuosoNet (Jeong et al., 2019)

youtu.be/hPBR2Rxu3-s

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

Open Questions

Open Questions

- How to generate **long-term structure**?
- How to enable more **intuitive controls** for music generation systems?
- How to adopt these models for **improvisation**?
- We are **running out of symbolic music data**
  - Can we learn symbolic music composition from **listening to raw audio**?
- **Is symbolic reasoning a must** for true AI music generation?
  - Can an AI play perfect music **without processing it into symbolic music internally**?
- How do humans **learn to create music**?

(Recap) A Simplified Music Production Workflow