Automatic Music Generation

More Effective
More Structural
More Controllable

CS4347 · CS5647

Oct
23

Stan Ma

Ph.D. Graduated from NUS
Research Fellow @ SMC

Hobbies:

Clarinet, Flute
Japanese, Graphic Design

Research Interests:

Music Generation
Lyrics Generation
Music Analysis

For Human Health and Potential

Why Automatic Music Generation (AMG)?

Entertainment & Art

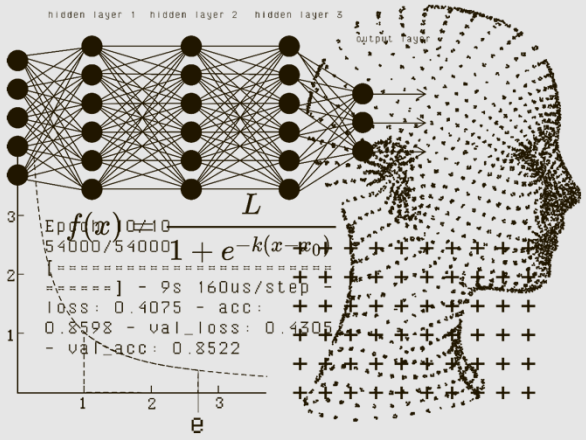
Inspire Musician Composition

Enhancing Human Activities

Paradigm

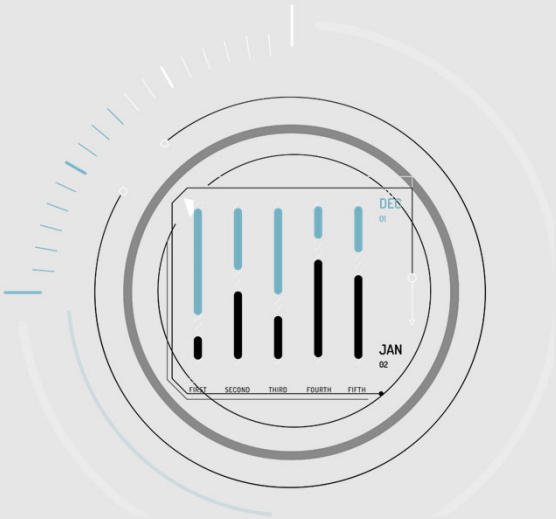
Music Representation

+



Generation Model

+



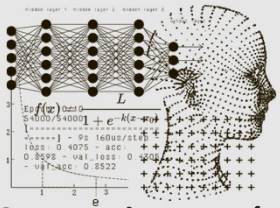
Control

=

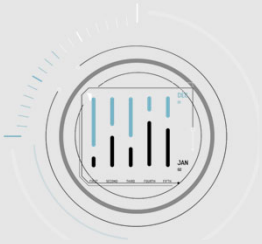
Pleasant and Useful Music

Paradigm

Music Representation



Generation Model



Control

- How music is read and represented in computer?
- How computer generates pleasant and creative music like human?
- How to control music generation according to our requirements?
- How to generate personalized music that a user prefers?
- How to start with your own music generation model?

Music Representations in AMG

From Human Perceptual to Computation Effective

Human Perception

How human perceive music?

Listening

Audio Domain
MP3, WAV, etc.

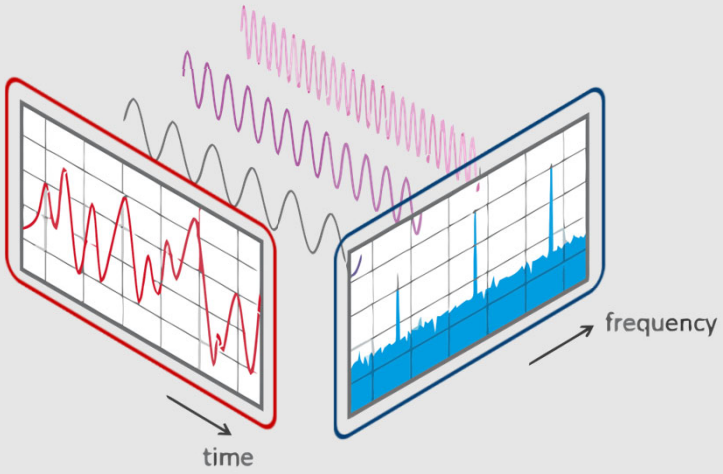
Reading

Symbolic Domain
Scores, MIDI files

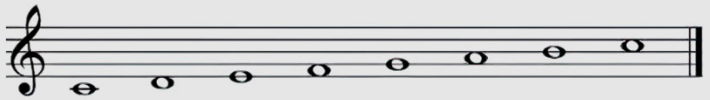
Human Perception

Audio Domain

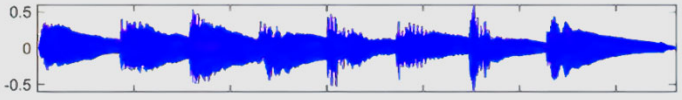
Framing → Windowing → Fourier Transform → Chroma Features



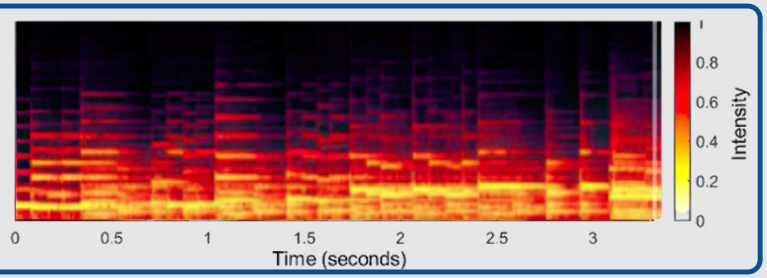
(a) Music



(b) Waveform

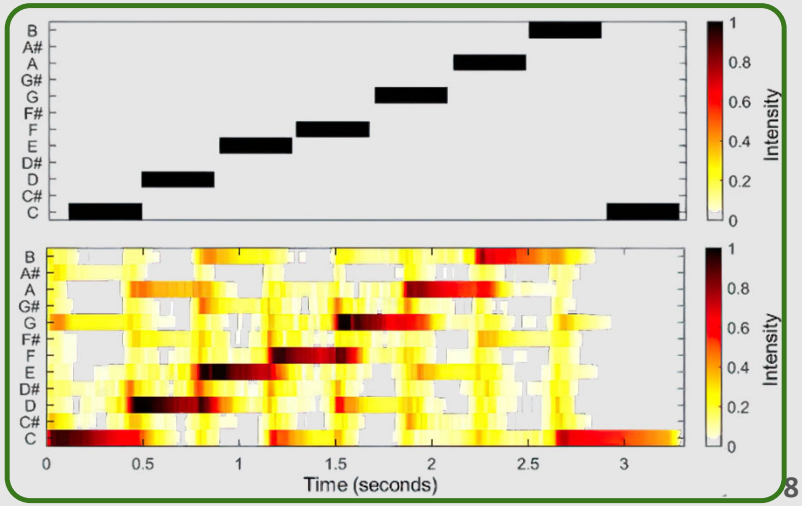


(c) Spectrogram



(d) Chromatogram

(d) Chromatogram

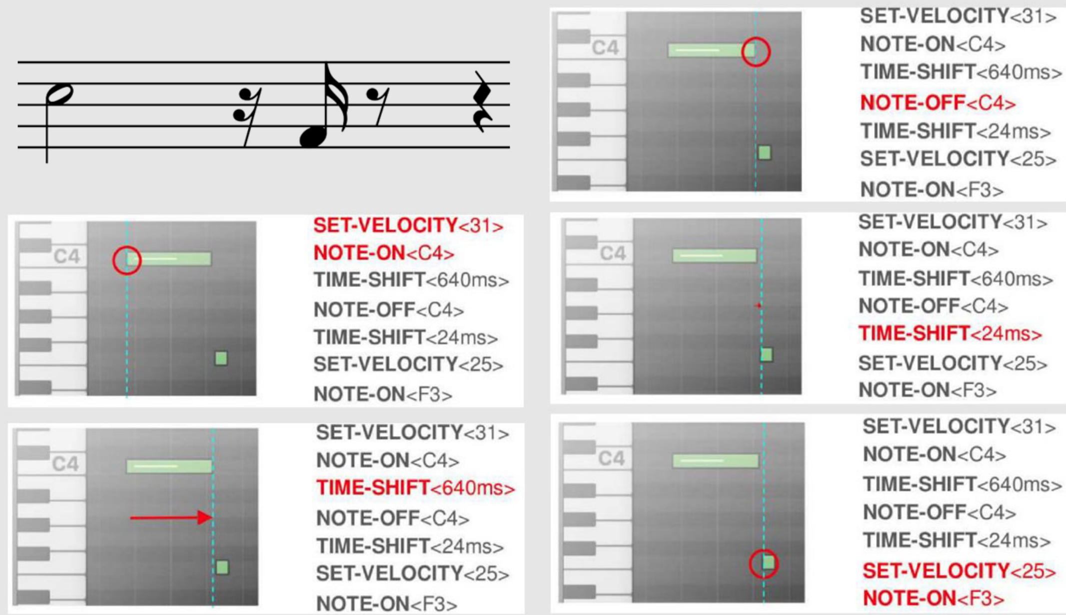


Human Perception

Symbolic Domain (Small file size, Instrument independent)

(a) MIDI: Musical Instrument Digital Interface (Event-based) → 1D

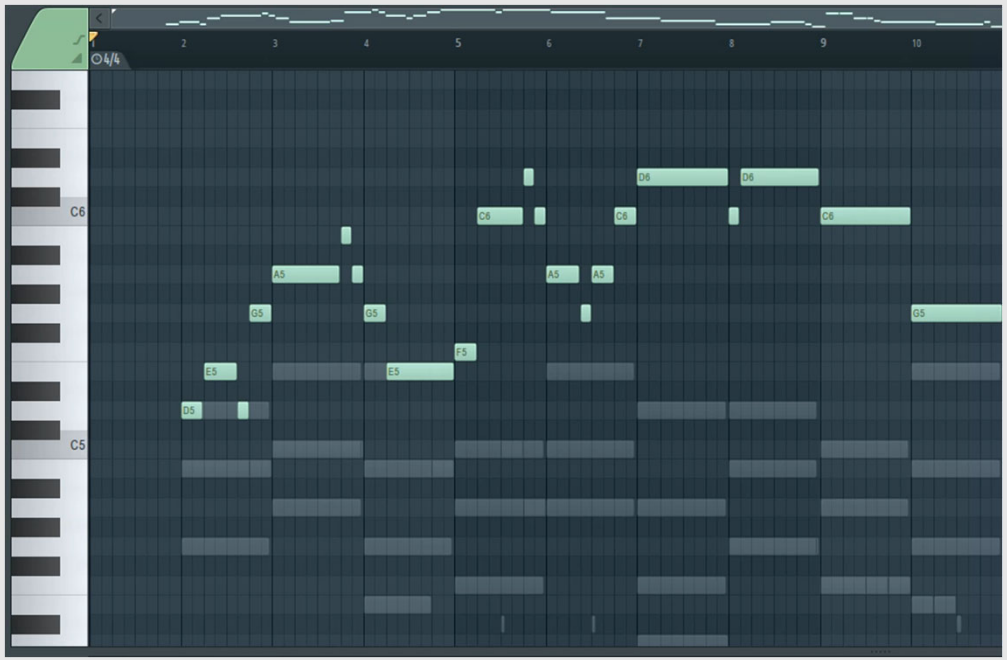
Task	Event	Description
Score	Note On	One for each pitch
	Note Off	One for each pitch
	Note Duration	Inferred from the time gap between a Note On and the corresponding Note Off,by accumulating the Time Shift events in between
	Time Shift	Shift the currenttime forward by the corresponding number of quantized time steps
	Position	Points to different discrete locations in a bar
	Bar	Marks the bar lines
	Piece Start	Marks the start of a piece
	Chord	One for each chord
	Program Se-lect/Instrument	Setthe MIDI program number atthe beginning of each track
	Track	Used to mark each track.
Performance	Note Velocity	MIDI velocity quantized into m bins. This event sets the velocity for subse-quent note-on events
	Tempo	Accountfor local changes in tempo(BPM)

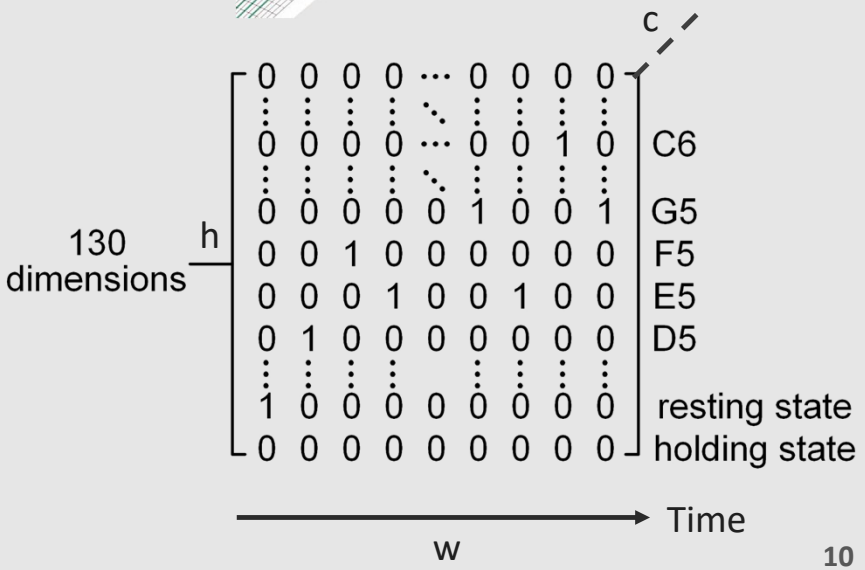


Human Perception to Machine-readable

Symbolic Domain

(b) Pianoroll: 1D → 2D, readable for human and machines





Machine-readable

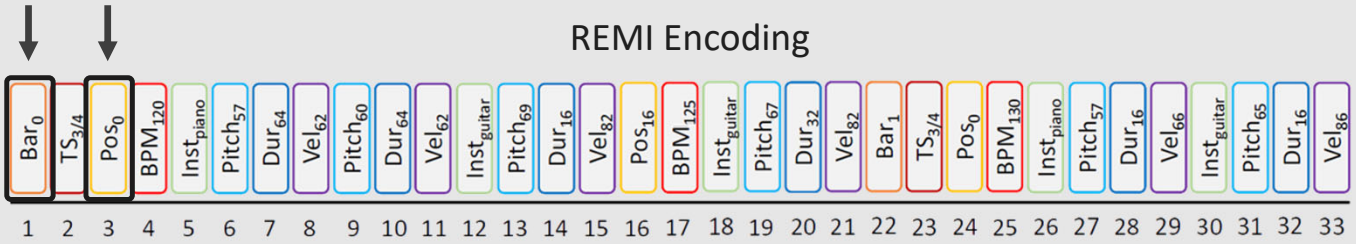
(c) Compound Event Sequence

Encoding	Token number
REMI [1]	15K
CP [2]	6.9K
Octuple [3]	3.6K

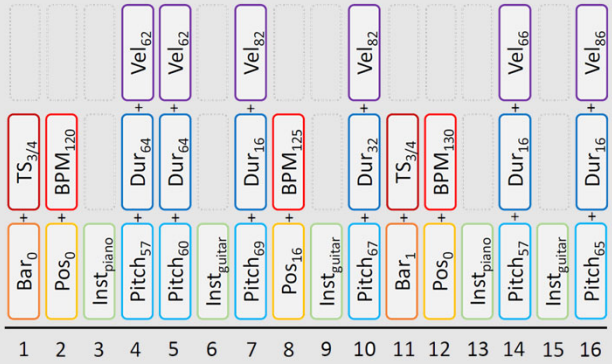
Less tokens means:

Shorter inputs
Denser gradient matrix

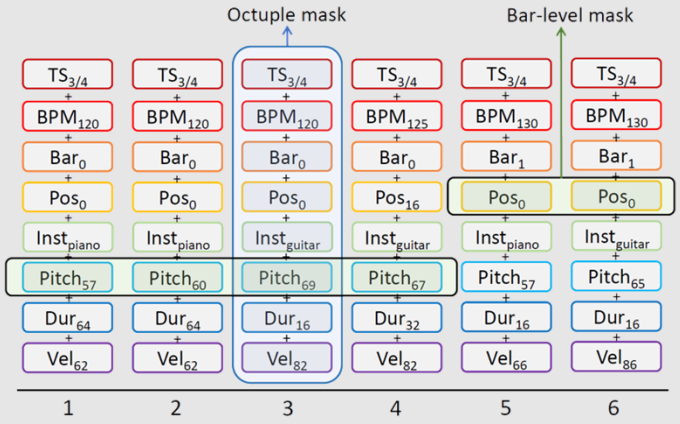
REMI Encoding



Compound Encoding (CP)

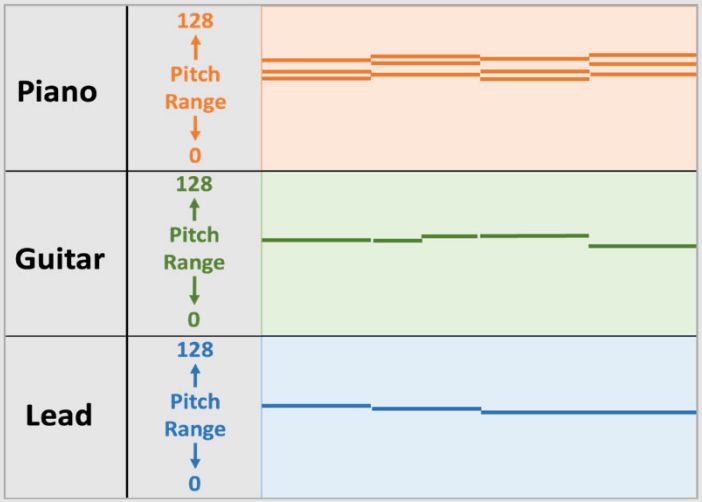


Octuple Encoding



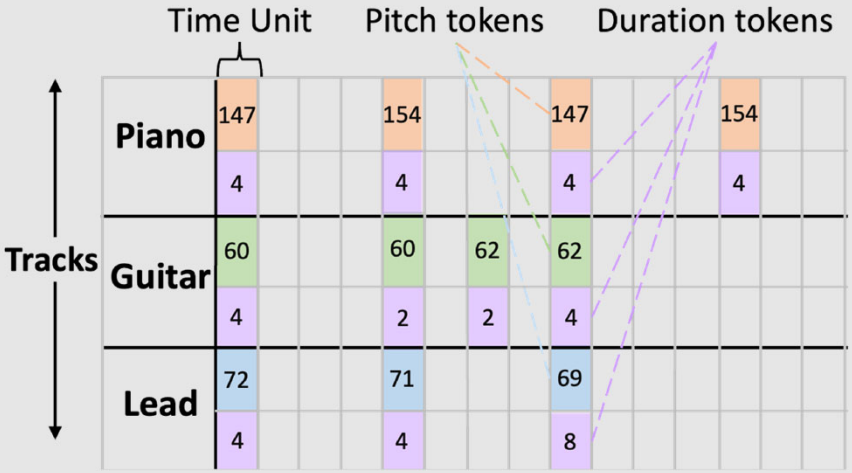
[1] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM international conference on multimedia, 1180–1188.
[2] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In Proceedings of the AAAI Conference on Artificial Intelligence, 178–186.
[3] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. Musicbert: Symbolic music understanding with large-scale pre-training. arXiv preprint arXiv:2106.05630 (2021).

Machine-readable



(d) Mixed/Combined (GETMusic [1])

- Tracks are separated, desirable for multi-track music modeling.
- Using integer value to replace 0/1 in Pianoroll, mode condensed.



[1] Ang Lv, Xu Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, and Rui Yan. 2023. GETMusic: Generating Any Music Tracks with a Unified Representation and Diffusion Framework. arXiv preprint arXiv:2305.10841 (2023).

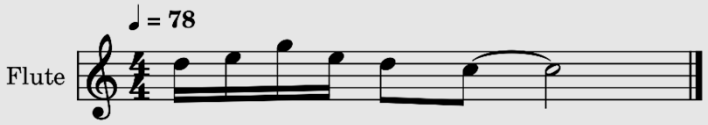
Machine-readable

(a) MusicXML: represent music as script

Problems:

- Not Playable
- Lengthy

(b) LilyPond: Shorter but
The notation is not unique in its identification




```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE score-partwise PUBLIC "-//Recordare//DTD MusicXML 3.1 Partwise//EN"
"http://www.musicxml.org/dtds/partwise.dtd">
<score-partwise version="3.1">
 <identification>
  <encoding>
   <software>MuseScore 3.6.2</software>
   <encoding-date>2023-06-16</encoding-date>
   <supports element="accidental" type="yes"/>
   <supports element="beam" type="yes"/>
   <supports element="print" attribute="new-page" type="yes" value="yes"/>
   <supports element="print" attribute="new-system" type="yes" value="yes"/>
   <supports element="stem" type="yes"/>
   </encoding>
  </identification>
 <defaults>
  <scaling>
   <millimeters>7.056</millimeters>
   <tenths>40</tenths>
   </scaling>
  <page-layout>
   <page-height>1683.27</page-height>
   <page-width>1190.82</page-width>
   <page-margins type="even">
    <left-margin>56.6894</left-margin>
    <right-margin>56.6894</right-margin>
    <top-margin>56.6894</top-margin>
    <bottom-margin>113.379</bottom-margin>
    </page-margins>
   <page-margins type="odd">
    <left-margin>56.6894</left-margin>
    <right-margin>56.6894</right-margin>
    <top-margin>56.6894</top-margin>
```

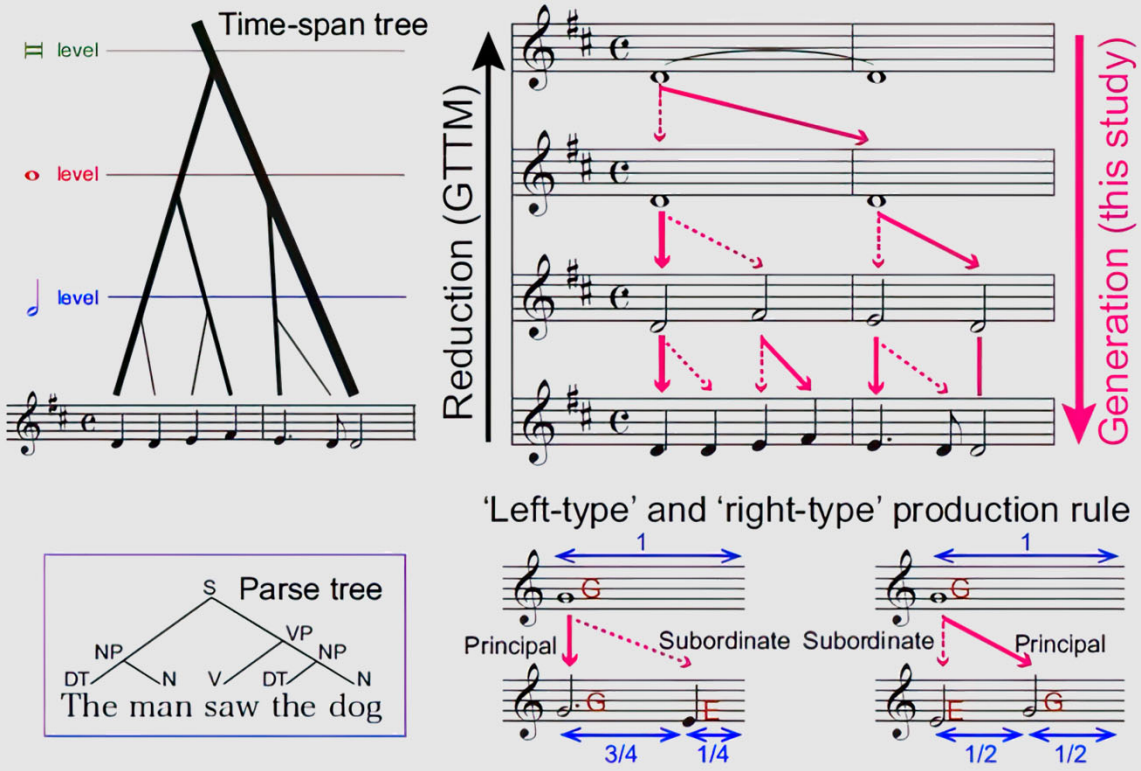
…… (340 lines in total)

Music generation with ChatGPT: https://www.holovaty.com/writing/chatgpt-music-generation/

Machine-readable

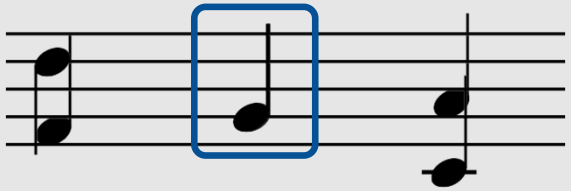
(b) Tree-like [1]: model music as a tree

- Potential in modeling music effectively because it contains the hierarchical musical structure.
- Difficult to obtain such representation and difficult to incorporate in music generation models.
- Increased complexity in polyphonic music



[1] Eita Nakamura, Masatoshi Hamanaka, Keiji Hirata, and Kazuyoshi Yoshii. 2016. Tree-structured probabilistic model of monophonic written music based on the generative theory of tonal music. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 276–280.

Computation Effectiveness: Embeddings



C:maj
D:min

“Relax Jazz fast tempo”

“Classical music composed by Russian composers of 1900s”

Why do you need contextual information?

Same music token contains different meanings in different surrounding contexts

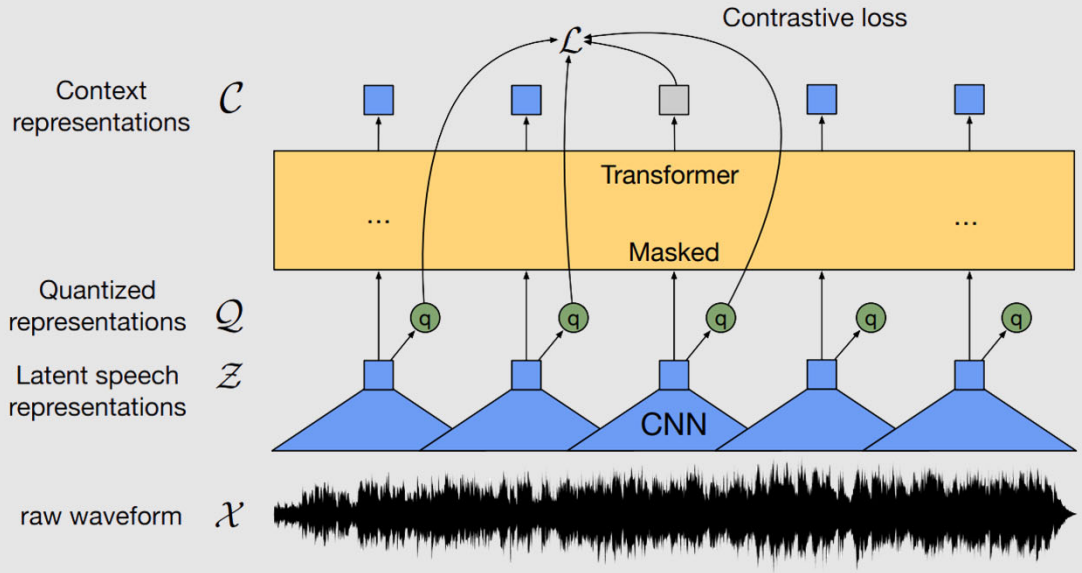
Encoding: same note is same token → **Embedding:** to obtain **contextual information**

Computation Effectiveness: Embeddings

Encoding → Embedding
to obtain **contextual information**

(a) Wav2vec2.0 [1]: Audio Embedding

- Self-supervised
- Predict some masked positions via its contexts

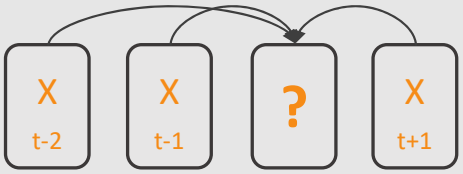


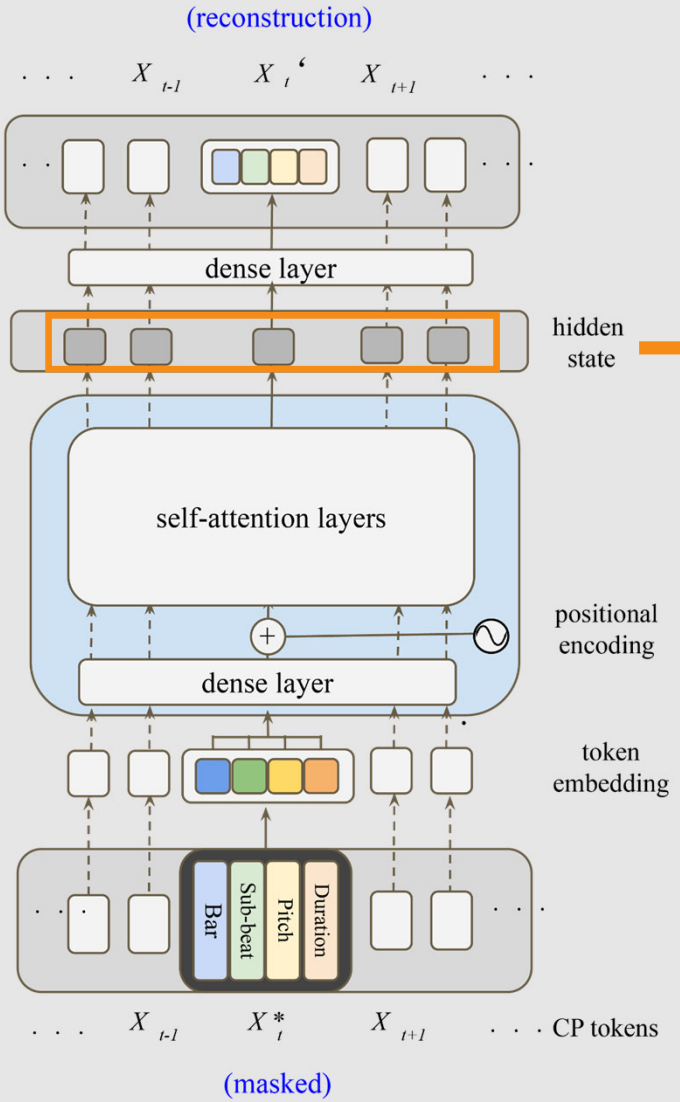
[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33, (2020), 12449–12460.

Computation Effectiveness

(b) MidiBERT [1]: Symbolic Embedding
(c) MusicBERT [2]: Symbolic Embedding

- Randomly mask some tokens
- Predict masked tokens from their contexts





Predicting from contexts

Masking

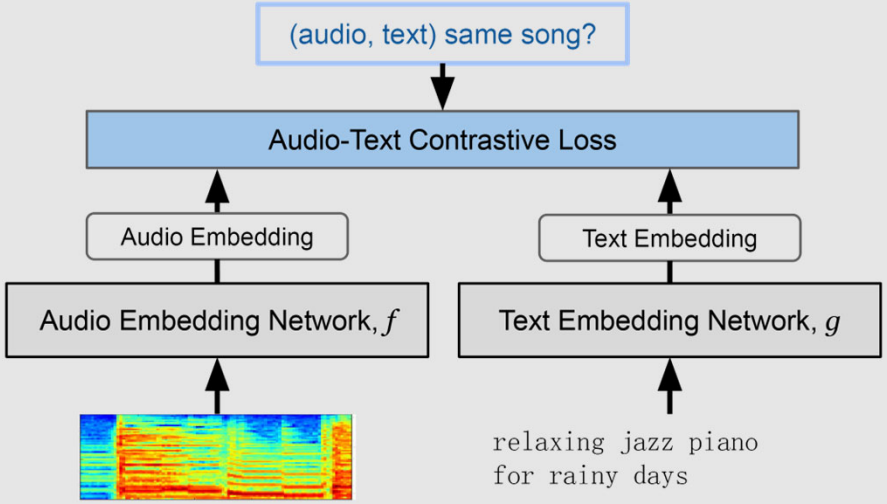
MidiBERT

[1] Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, Yi-Hsuan Yang, and others. 2021. MidiBERT-piano: large-scale pre-training for symbolic music understanding. arXiv preprint arXiv:2107.05223 (2021).
[2] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. Musicbert: Symbolic music understanding with large-scale pre-training. arXiv preprint arXiv:2106.05630 (2021).

Computation Effective

(d) MuLan [1]: Cross-modal Embedding (Text-Music)

- Trained on paired audio and its description texts
- Optimized by contrastive loss of audio and text embeddings



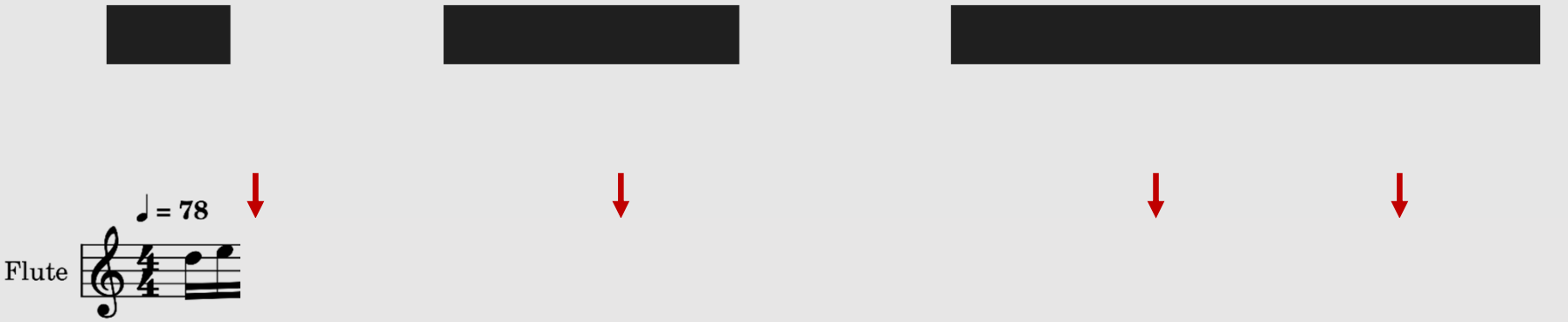
[1] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. arXiv preprint arXiv:2208.12415 (2022).

Encoder & Generation Model

From Looking One Step to the Whole Piece

Autoregressive

An Experiment



- The longer you have heard, the easier to predict the next notes.

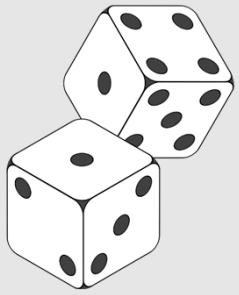
Music source: Tears of Suzume by Radwimps


Mozart Game
Mozart Game

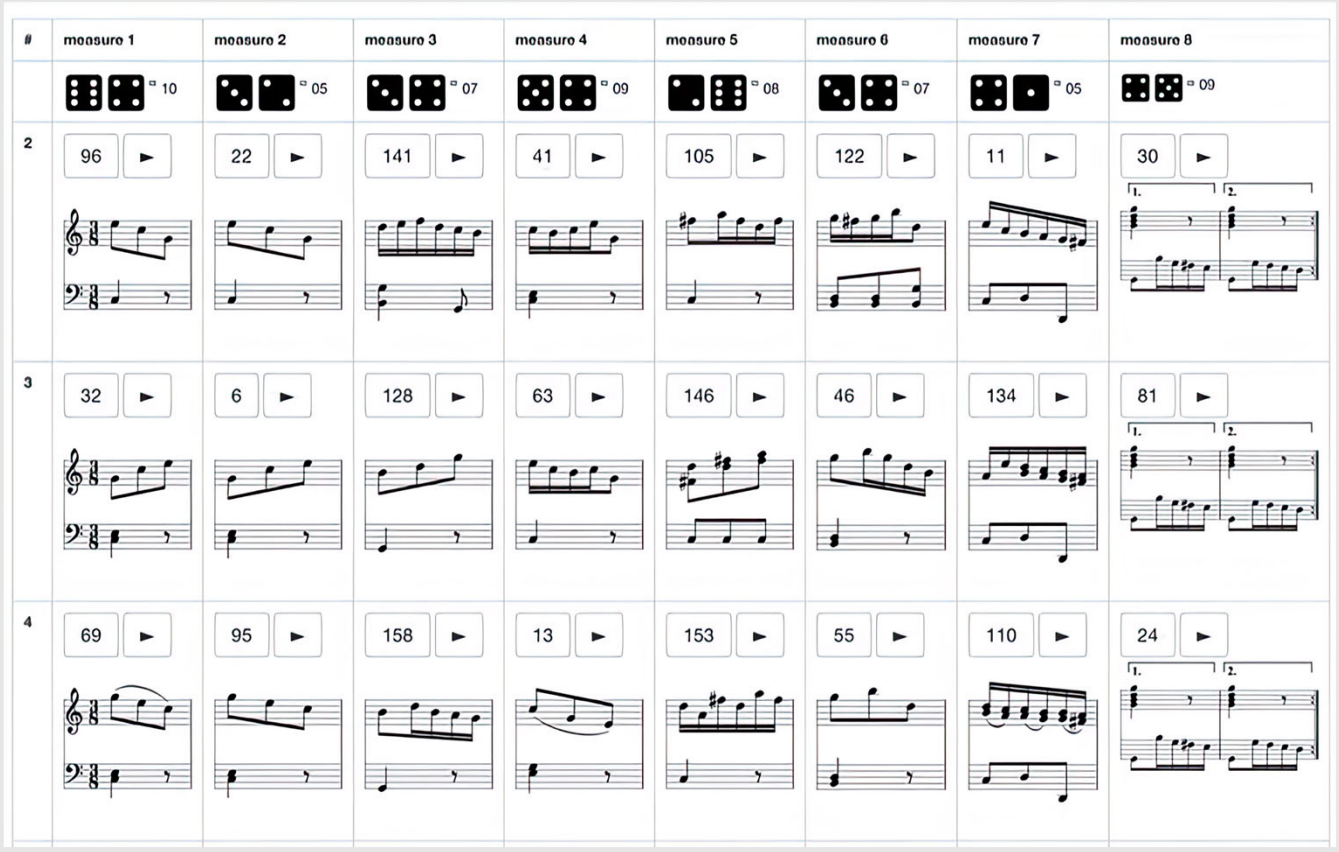
Autoregressive

Mozart Game Algorithm

1. Split music pieces into Fragments
2. Number the fragments
3. Each time toss two dices
4. Concatenate all samples
5. Go back to 2.

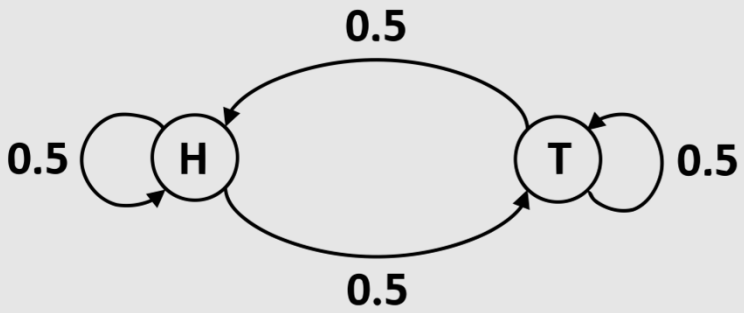


Demo



Mozart Game: https://en.wikipedia.org/wiki/Musikalisches_W%C3%BCrfelspiel

Markov Model



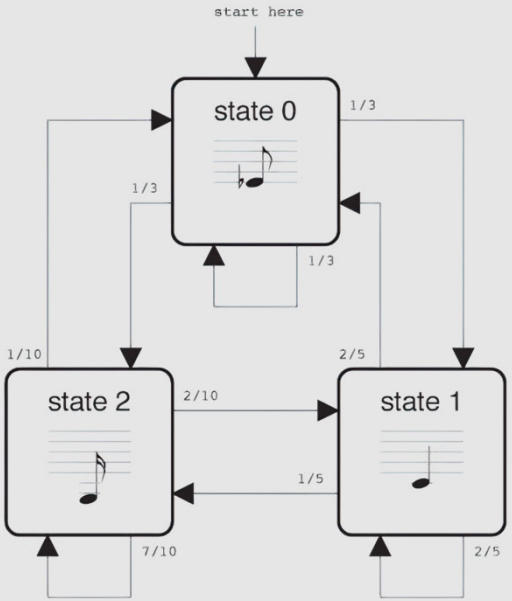
Markov Chain Example

Mozart Game: Transition of Fragments

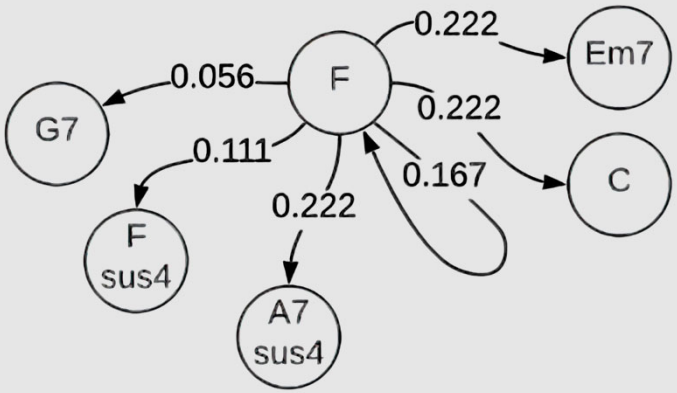
…	0.25	0.25	0.25	0.25
0.25	…	0.25	0.25	0.25
0.25	0.25	…	0.25	0.25
0.25	0.25	0.25	…	0.25
0.25	0.25	0.25	0.25	…

Music Generation with Markov Model

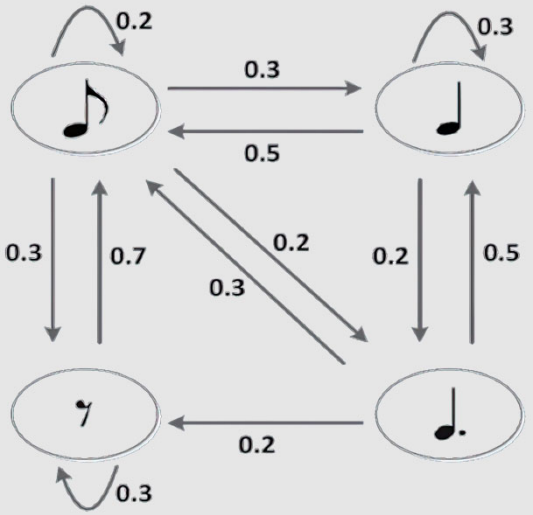
Melody Note [1]



Chord [2]



Rhythm/Duration [3]



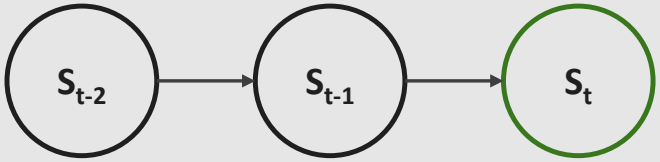
[1] http://codehop.com/three-node-markov-chain/
[2] https://towardsdatascience.com/markov-chain-for-music-generation-932ea8a88305
[3] Bongjun Kim and Woon Seung Yeo. 2013. Probabilistic prediction of rhythmic characteristics in Markov chain-based melodic sequences. In ICMC.

Music Generation with Markov Model

The Problem of increasing Markov order:

- Fixed Probabilities
- Losing connection and cohesion very soon
- **Exponential Increment** of the States

Demo



1-order MM

$$p(s_1|\emptyset), p(s_1|s_1), p(s_2|s_1), p(s_2|s_2),$$
$$p(s_3|s_2), p(s_3|s_3), p(\emptyset|s_3)$$

2-order MM

$$p(s_1|\emptyset), p(s_1|s_1,\emptyset), p(s_1|s_1,s_1),$$
$$p(s_2|s_1,\emptyset), p(s_2|s_2,s_1), p(s_2|s_2,s_2),$$
$$p(s_3|s_2,s_1), p(s_3|s_2,s_2), p(s_3|s_3,s_2),$$
$$p(s_3|s_3,s_3), p(\emptyset|s_3,s_2), p(\emptyset|s_3,s_3).$$

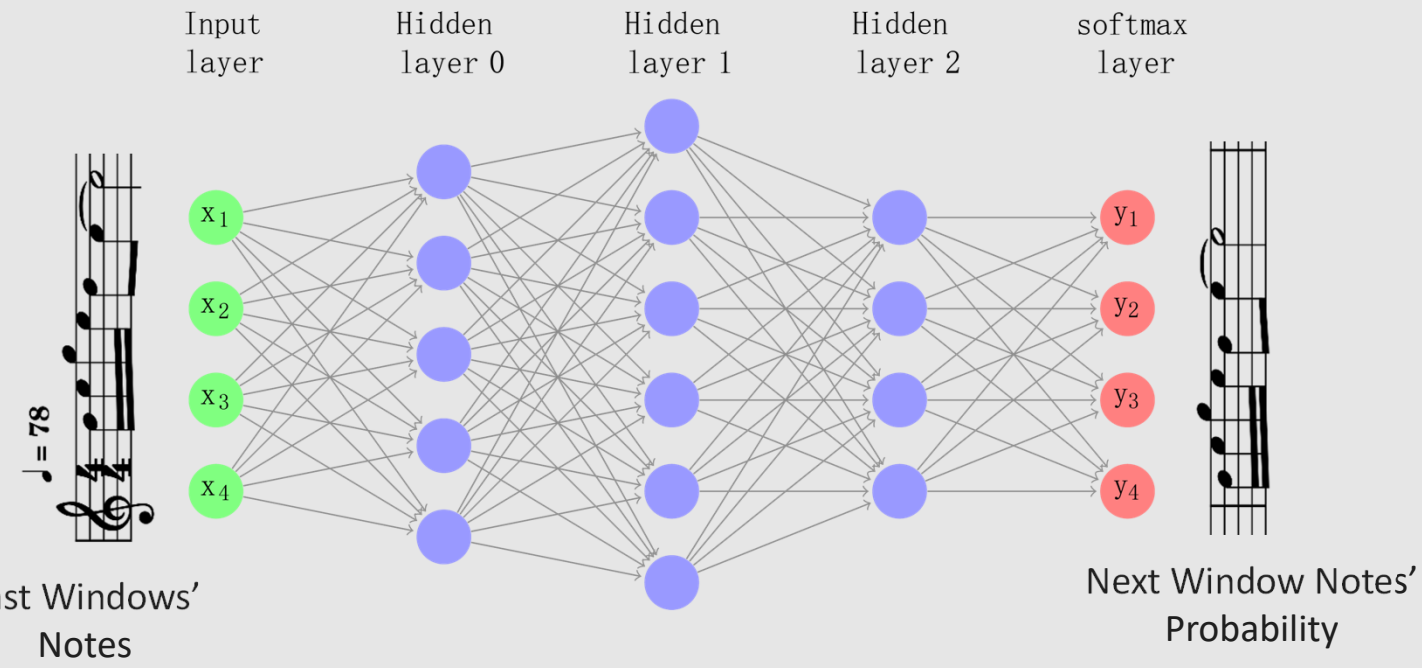
N-order MM

…

Music source: https://www.youtube.com/watch?v=Ghqchw9X6dw

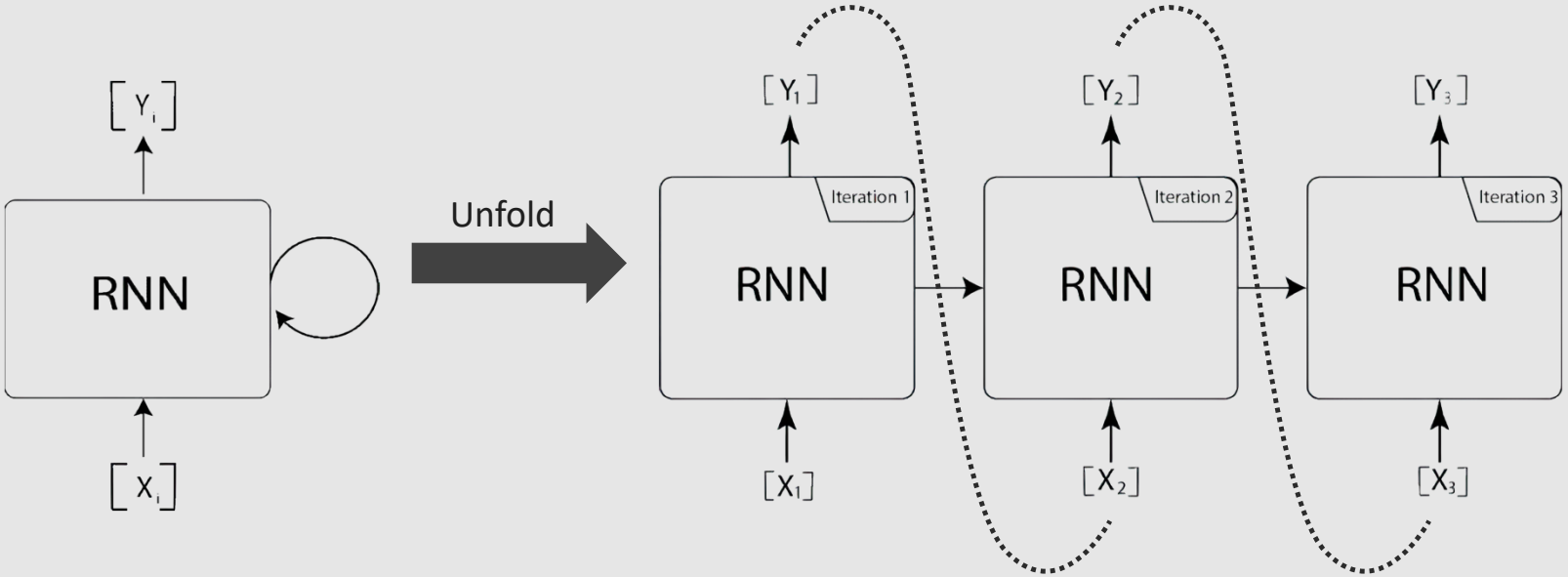
Compute Probability with Neural Networks

- Multilayer Perceptron (MLP)
- MLP cannot model sequential information
- Fixed length input
- Exponential increasing #parameters for longer inputs



[1] Peter M. Todd and Gareth Loy. 2003. A Connectionist Approach To Algorithmic Composition. In Music and Connectionism. 173–194.

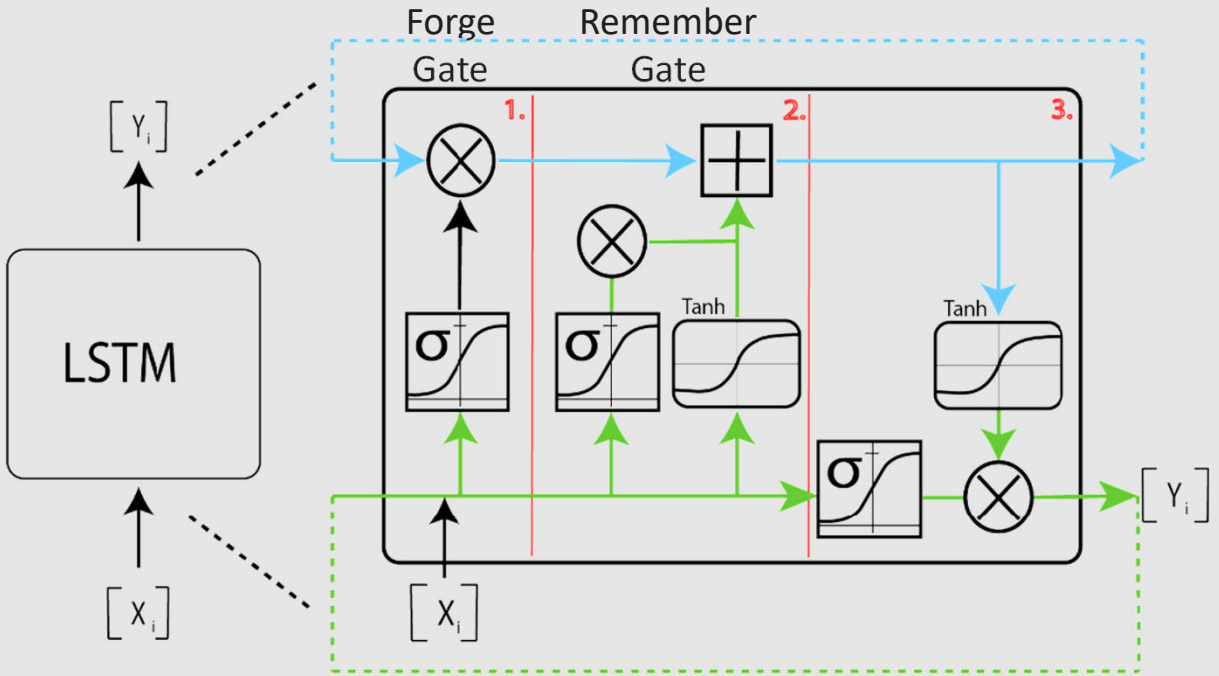
Sequential Model: RNN



- Parameter explosion problem: One token per step
- Fixed length problem: One token per step
- Ignorance of previous data: Hidden state passing and output feedback

Sequential Model: LSTM

Demo



- Based on RNN, add forget gate and remember gate to filter the insignificant/significant previous information.

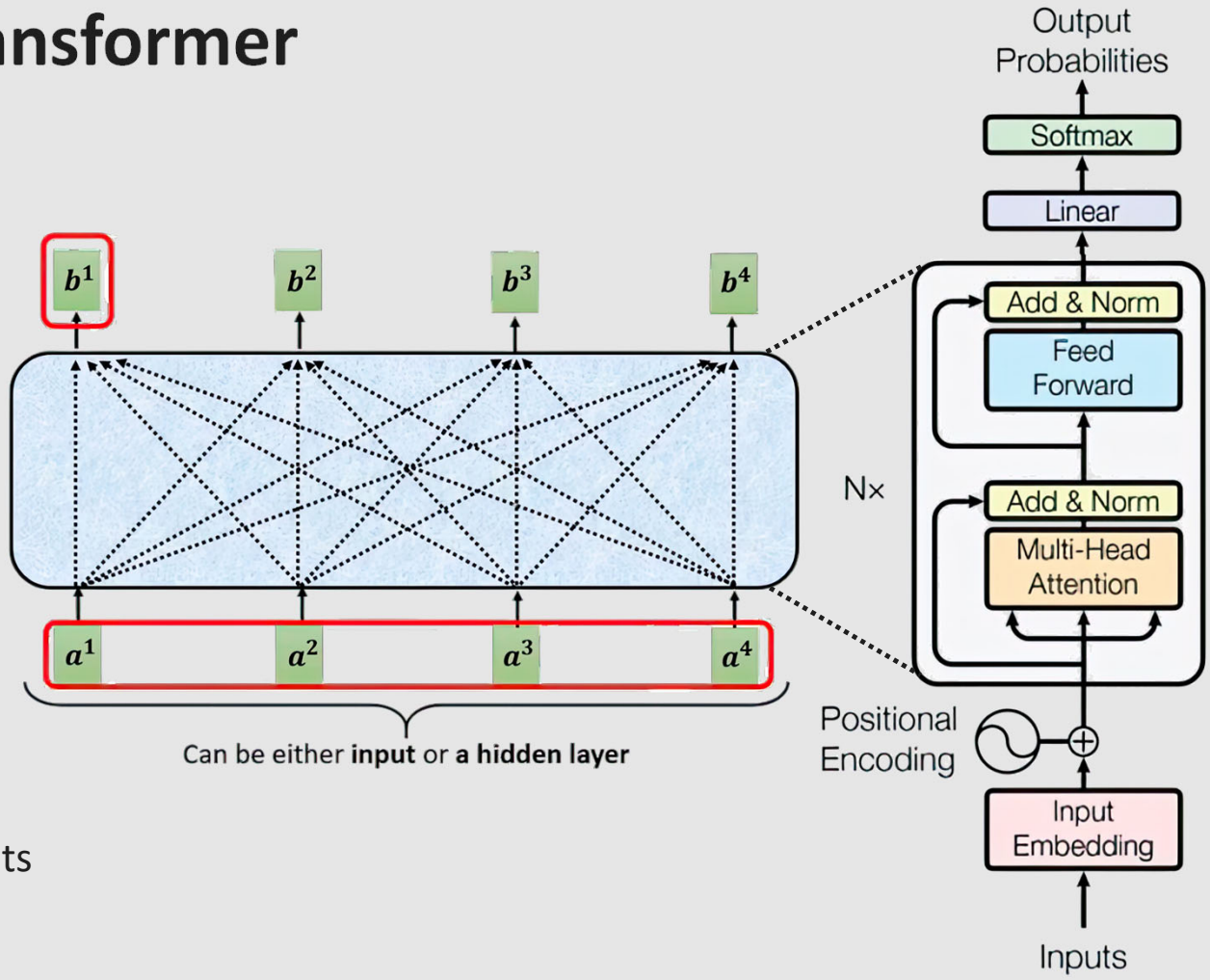
Music source: https://www.bilibili.com/video/BV17L411K7Ax/

Sequential Model: Transformer

- RNN/LSTM's Problems:
  - Slow computation for long sequences
  - Vanishing gradient issue
  - Local Perceptual field

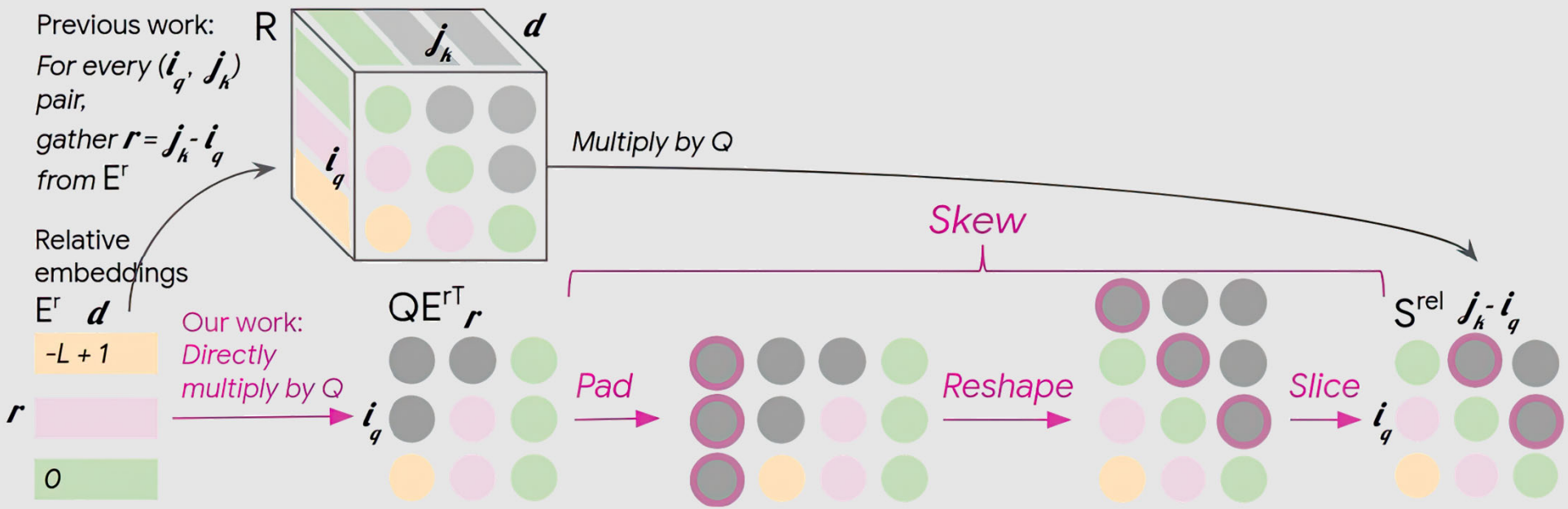
↓

- Music Transformer: Self-attention [1]
  - Like MLP but with dynamic weights
  - Like MLP but more efficient

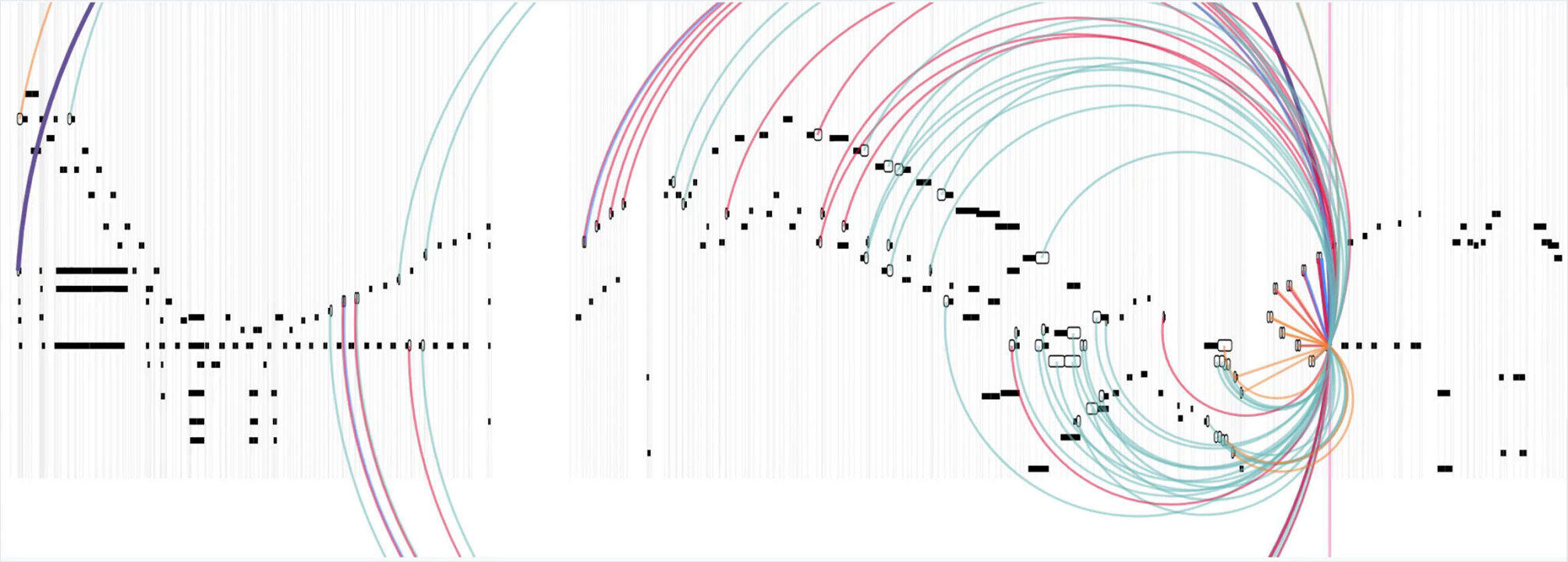


[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. arXiv preprint arXiv:1809.04281 (2018).

Sequential Model: Transformer



Sequential Model: Music Transformer Visualization [1]



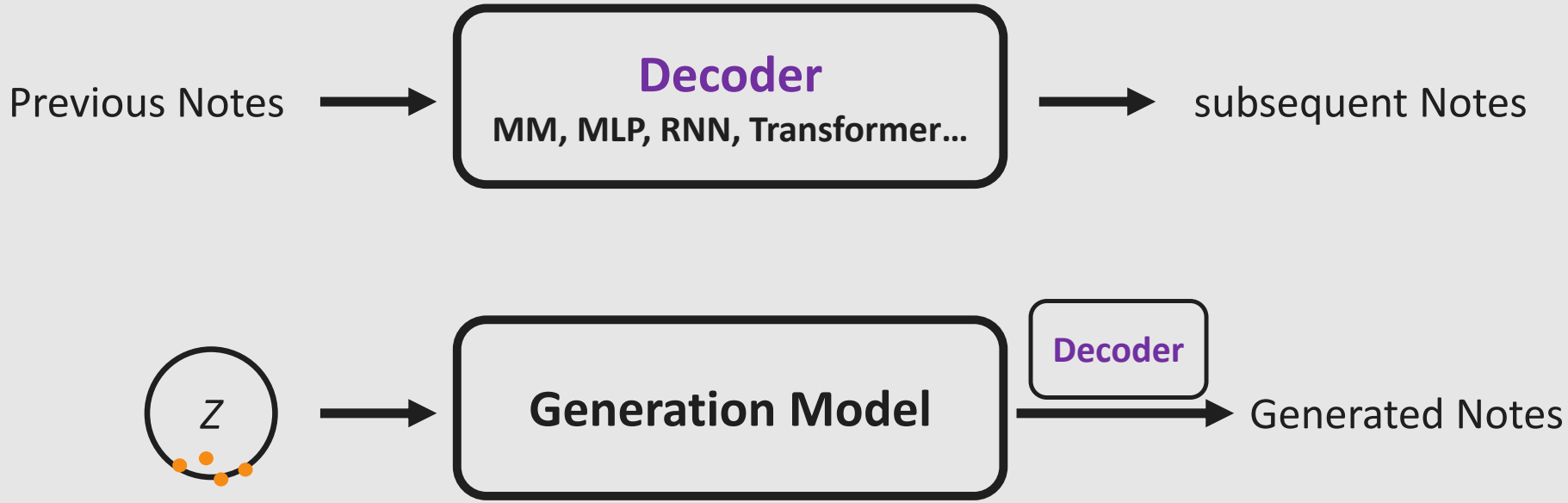
[1] Anna Huang, Monica Dinculescu, Ashish Vaswani, and Douglas Eck. 2018. Visualizing music selfattention. In Proc. NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language, 1.

Scan to vote

Which one sounds:

- the best overall?
- Musically structural?
- Human-like?
- Stable over time?
- Most “Relaxing Jazz”?

Generation Models: Sampling from a prior

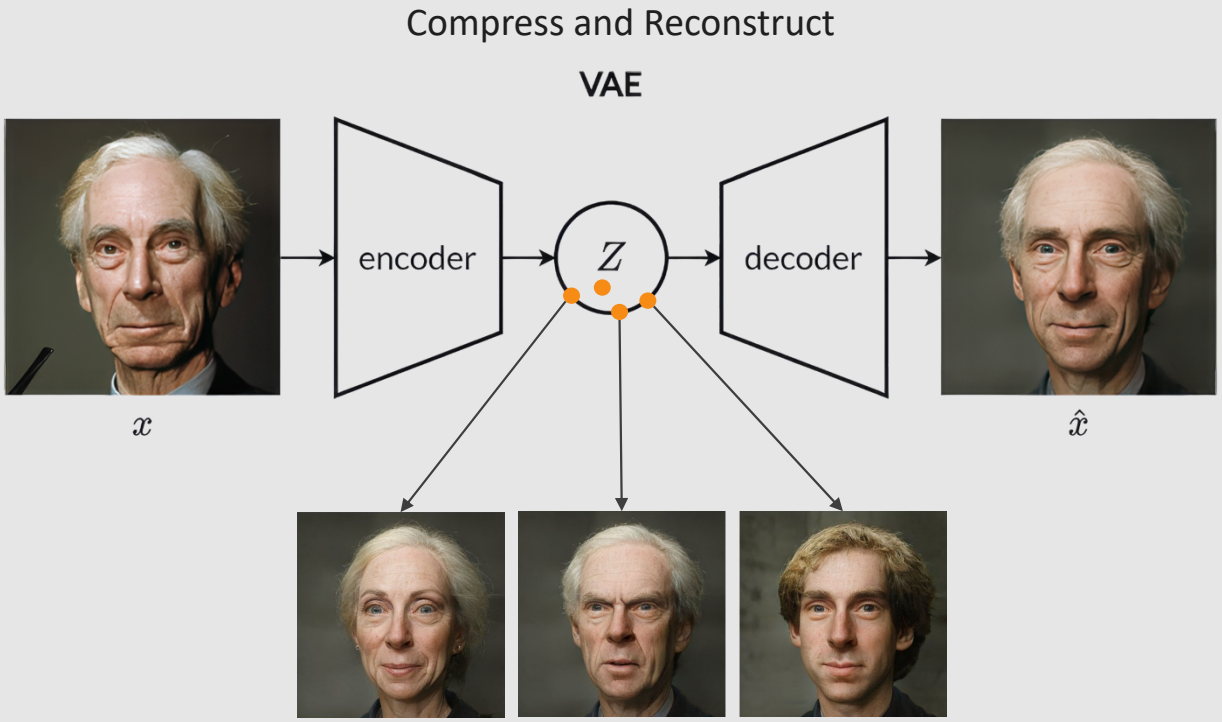


Generation Models: Sampling from a prior

(a) VAE [1]

- Encoder and Decoder can be MLP/RNN/LSTM/CNN/Transformer…
- VAE's idea:
  Training: Compress and Reconstruct

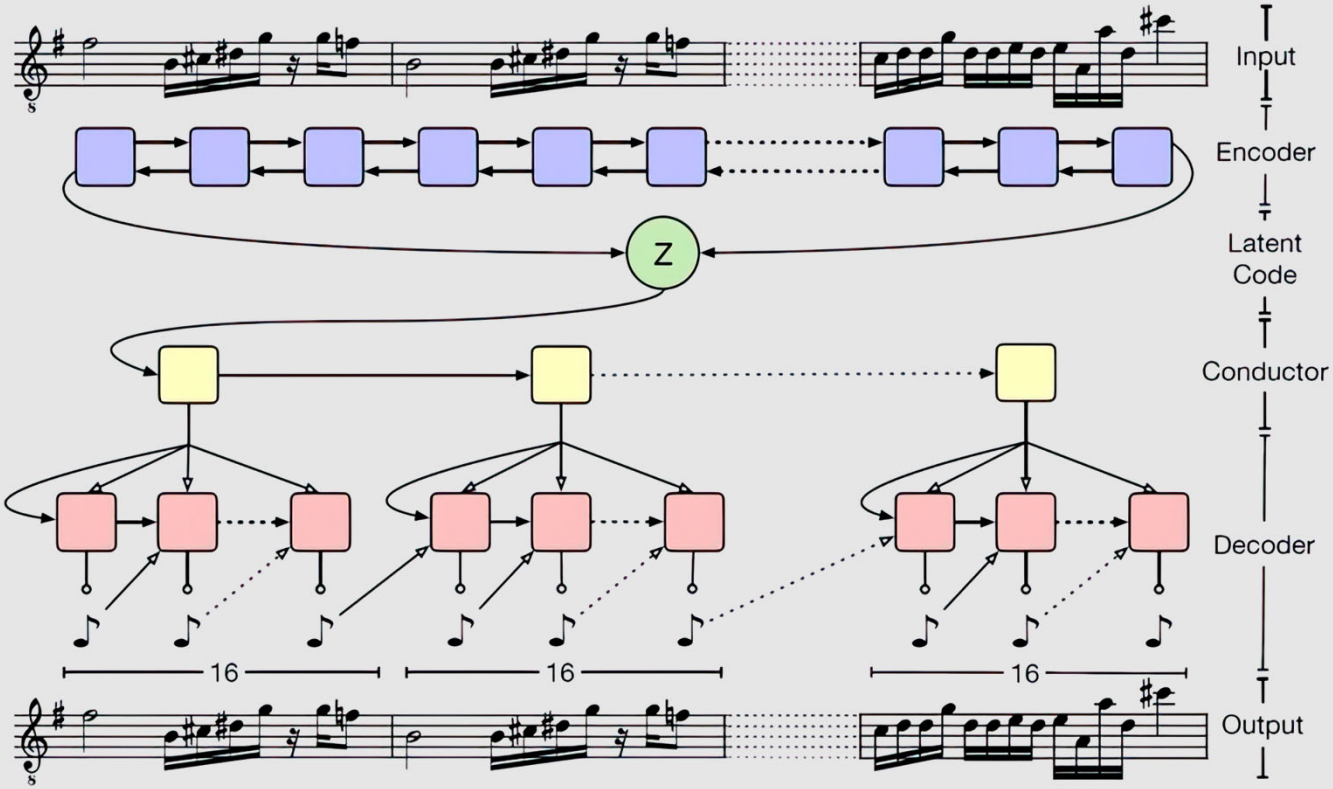
  Inference:
  - Random Sampling
  - Decoding the sample



[1] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. 2021. Variational Autoencoder. In Variational Methods for Machine Learning with Applications to Deep Networks. Springer International Publishing, Cham, 111–149.
Image source: Raphaël Millière. 2022. Deep learning and synthetic media. Synthese 200, 3 (2022), 231.

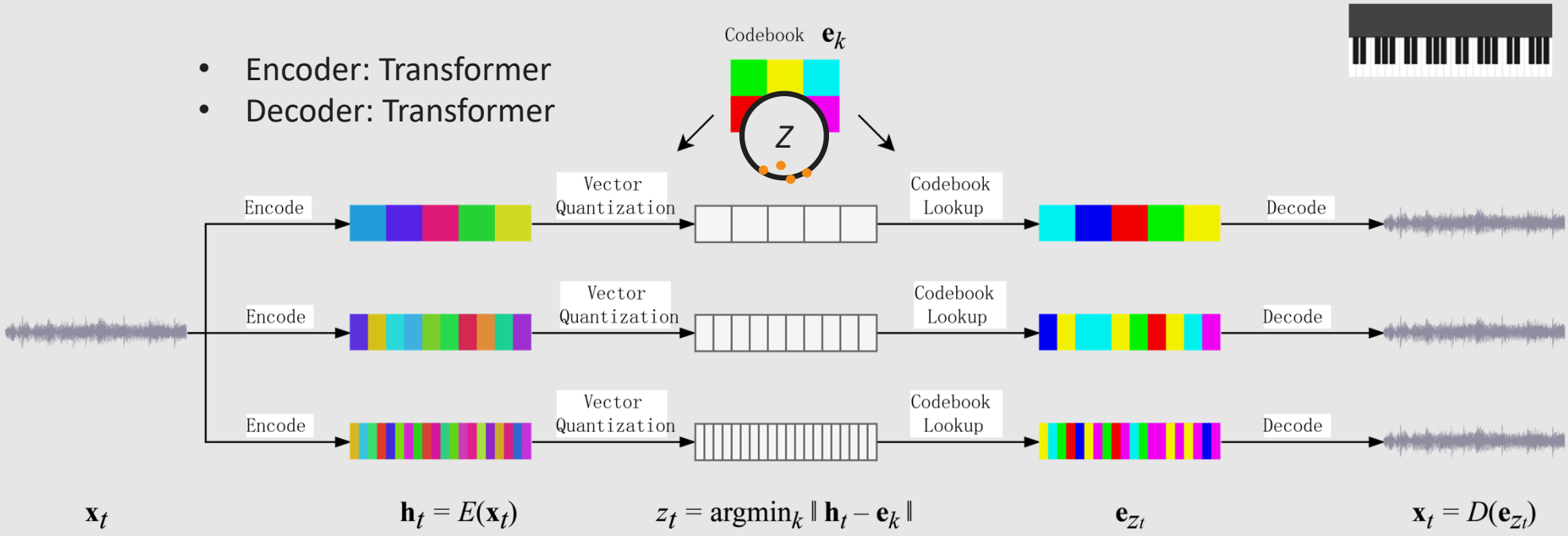
An example: MusicVAE

- Encoder: Hierarchical Bi-RNN
- Decoder: Hierarchical RNN



An example of VAE: Jukebox [1]

- Encoder: Transformer
- Decoder: Transformer



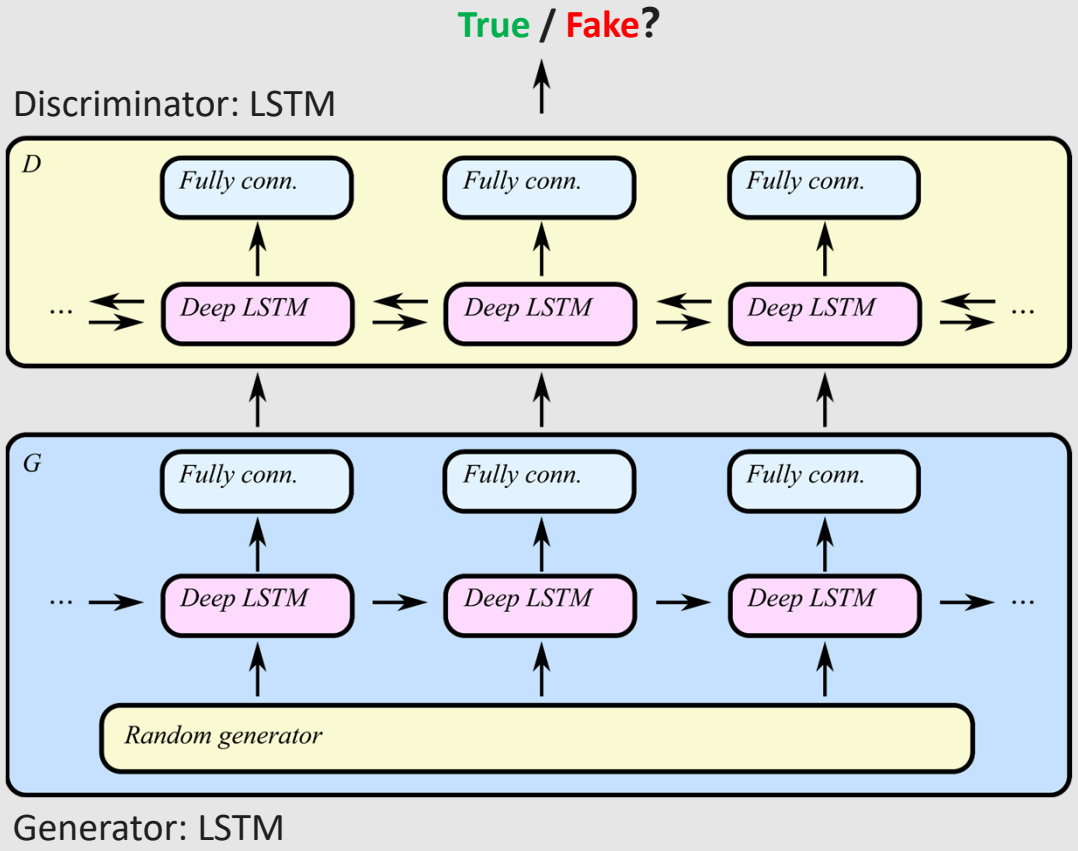
- Sample directly from song dataset, generated song contains music/lyrics/singing
- However the lyrics sometimes are meaningless words.

[1] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. CoRR abs/2005.00341, (2020). https://arxiv.org/abs/2005.00341

Generation Models: Sampling from a prior

(b) GAN

- Generator and Discriminator can be MLP/RNN/LSTM/CNN/Transformer…
- Training in Parallel
  - Generator become more powerful in generate human-like music
  - Discriminator become more powerful distinguish human compositions and generated music

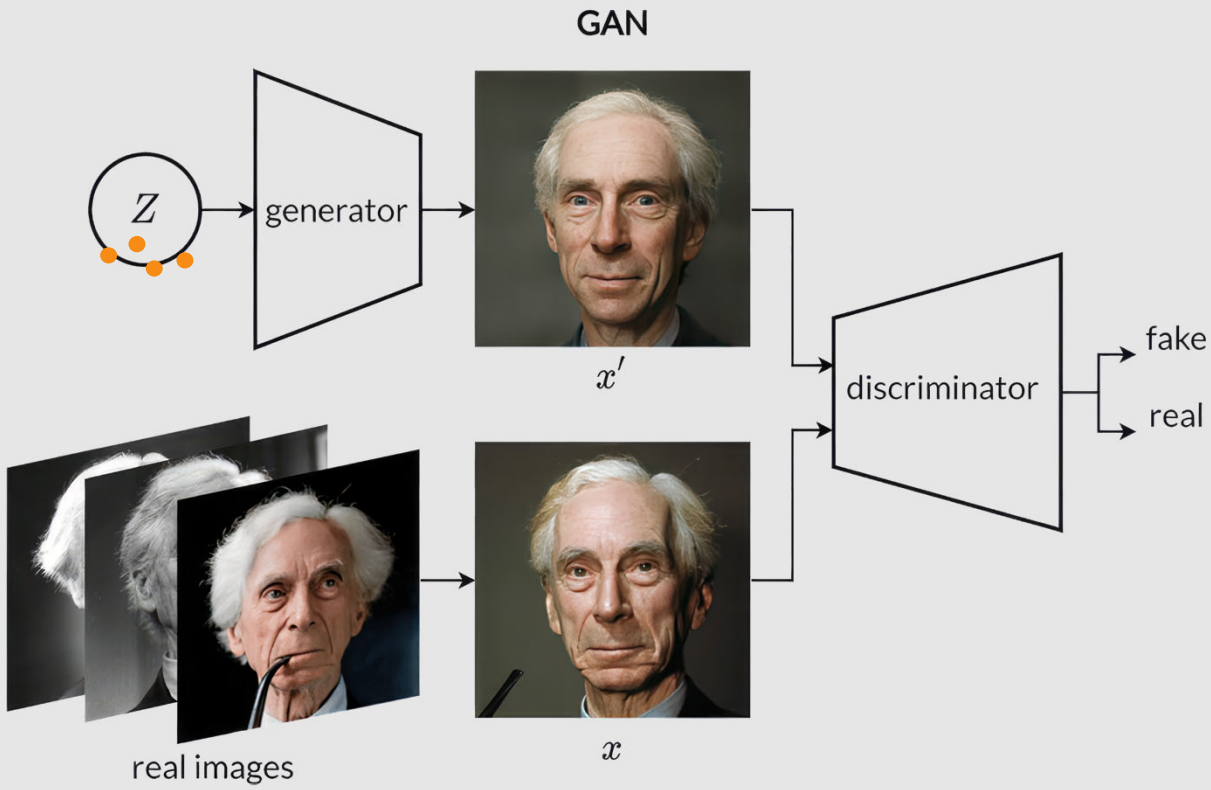


An example: C-RNN-GAN

Generation Models: Sampling from a prior

(b) GAN [1]

- Generator and Discriminator can be MLP/RNN/LSTM/CNN/Transformer…
- Training in Parallel
  - **Generator** becomes more powerful in generate human-like music
  - **Discriminator** becomes more powerful distinguish human compositions and generated music



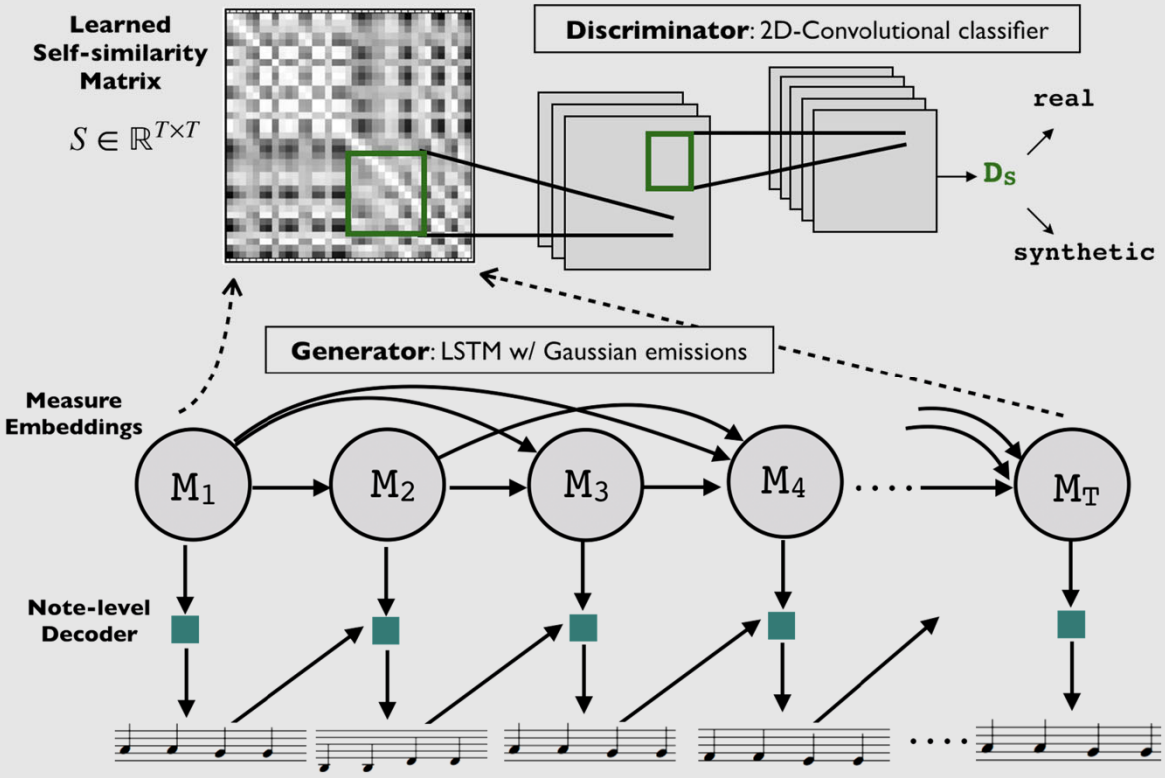
[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In Neural Information Processing Systems. https://arxiv.org/abs/1406.2661
Image source: Raphaël Millière. 2022. Deep learning and synthetic media. Synthese 200, 3 (2022), 231.

An example of GAN: self-repetition GAN

Discriminator: Using CNN to judge whether the self-similarity matrix looks like a real piece.

Some **repetitive note rhythmic patterns** can be observed in the generation.

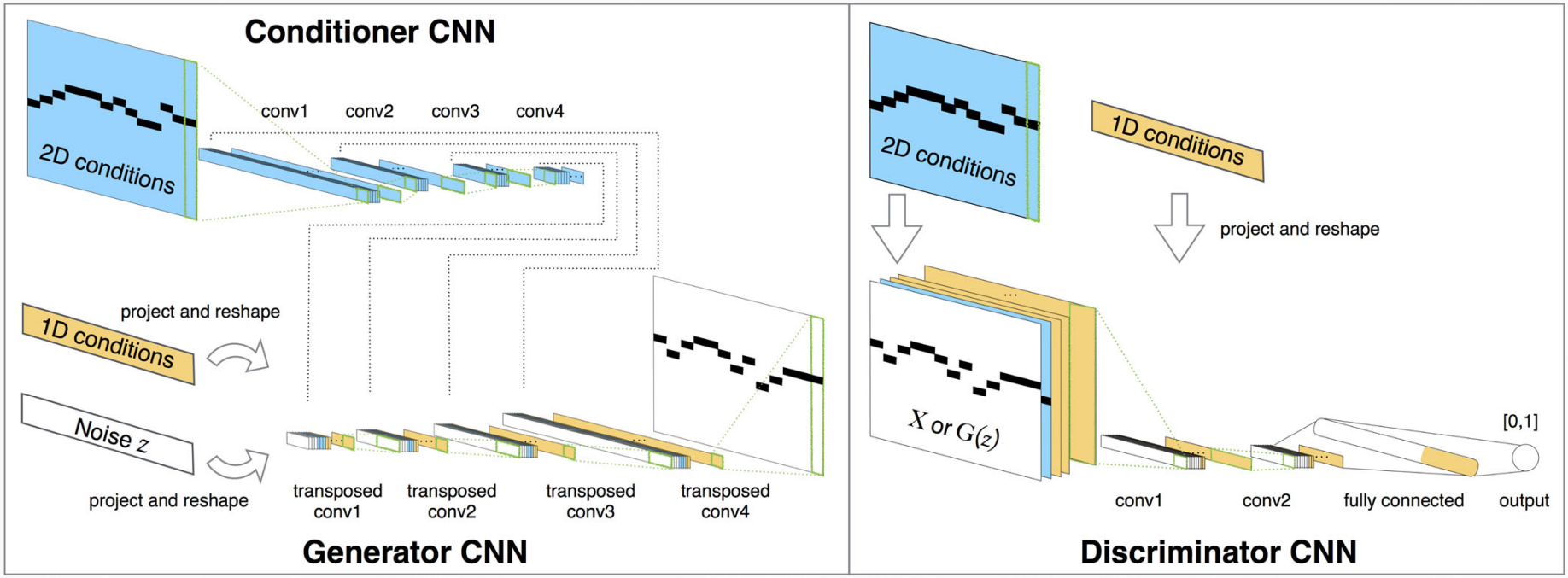
Generator: LSTM
Discriminator: CNN



[1] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2019. Modeling self-repetition in music generation using generative adversarial networks. In Machine Learning for Music Discovery Workshop, ICML.

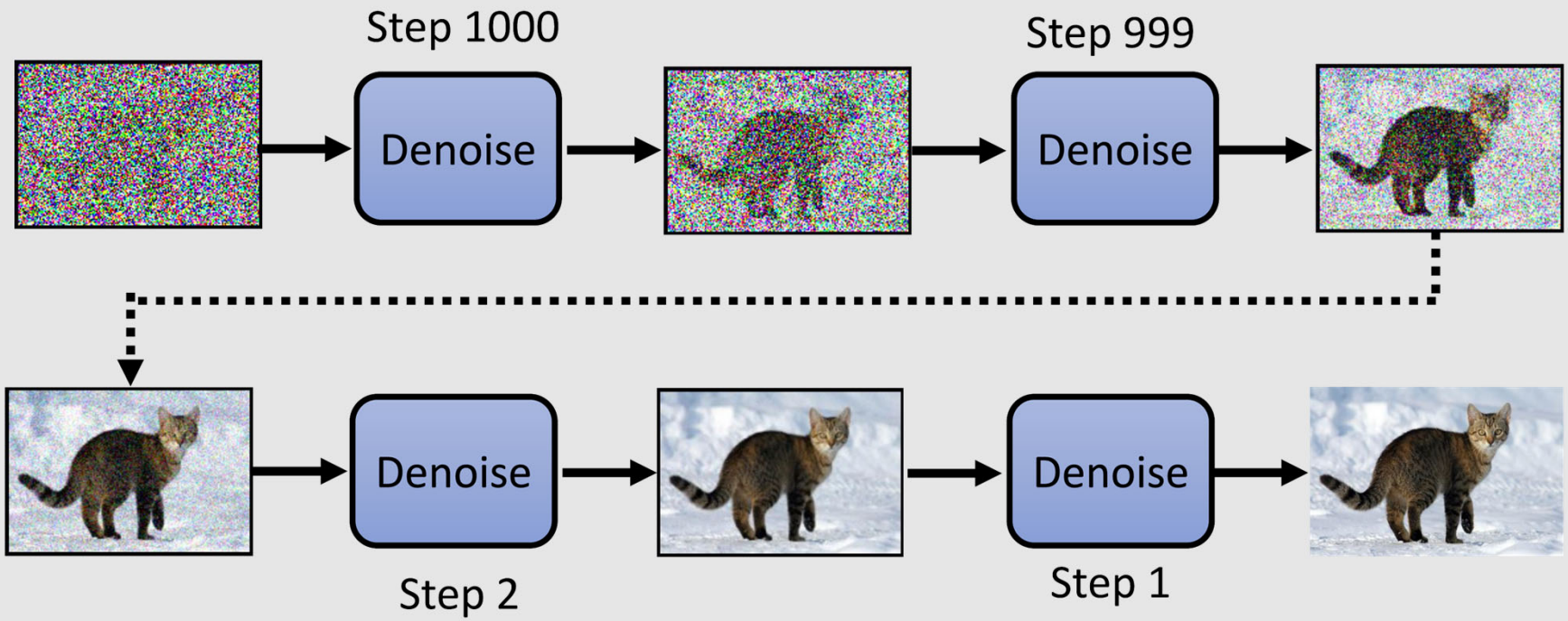
Previous bar

Chord



Generation Models: Sampling from a prior

(c) Diffusion Model [1] : combining autoregressive and non-autoregressive models



[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems 33, (2020), 6840–6851.
Image source: https://www.youtube.com/watch?v=azBugJzmz-o&t=281s

Generation Models: Sampling from a prior

(c) Diffusion Model: combining **autoregressive** and **non-autoregressive** models

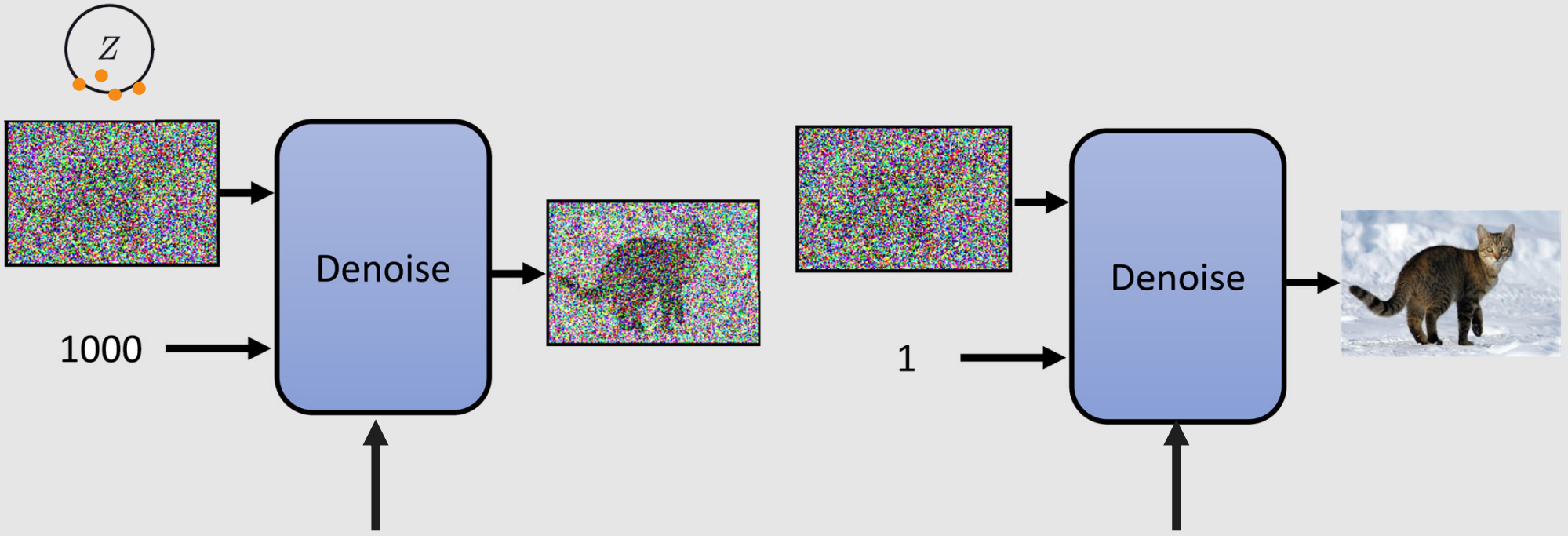


Image source: https://www.youtube.com/watch?v=azBugJzmz-o&t=281s

Generation Models: Sampling from a prior

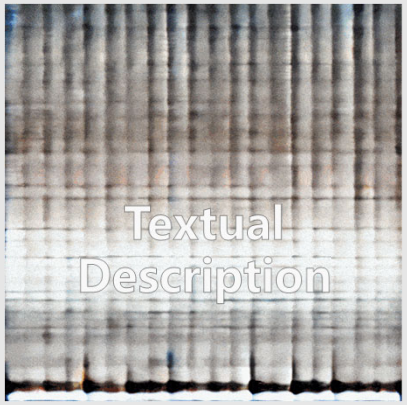
photograph of an astronaut riding a horse

Image source: https://riffusion.com/about

An example: Riffusion [1]

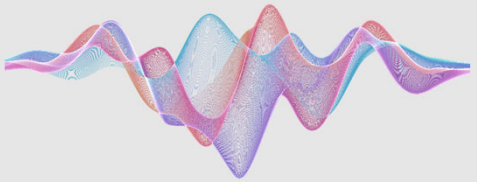
- Noise → Spectrum → music

Encoder: CNN for graph
Decoder: CNN for graph



Spectrogram generated by Stable Diffusion

Convert



Waveform music clips

Demo

Riffusion

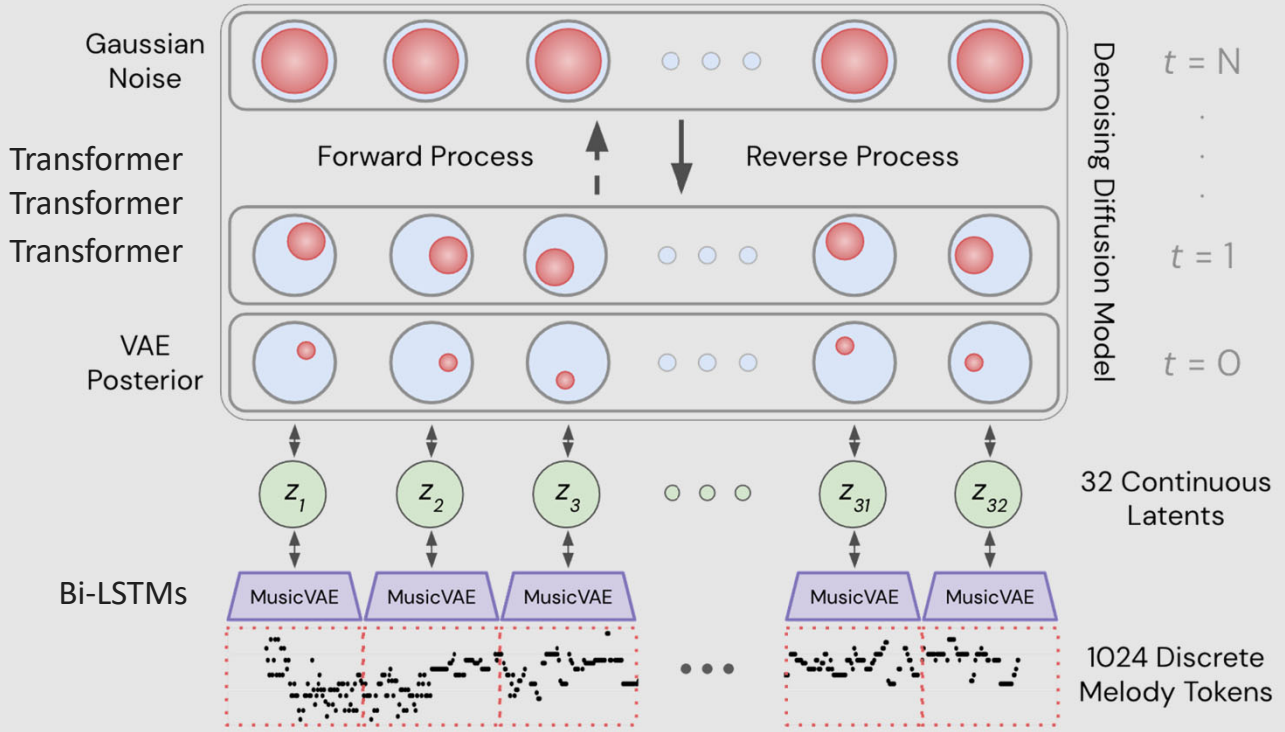
"A song in Singapore style representing the spirit of the course Sound and Music Computing"

[1]Seth* Forsgren and Hayk* Martiros. 2022. Riffusion - Stable diffusion for real-time music generation. (2022). Retrieved from https://riffusion.com/about

An example: Symbolic Diffusion Model [1]

- How to generate discrete data (MIDI) with diffusion model?

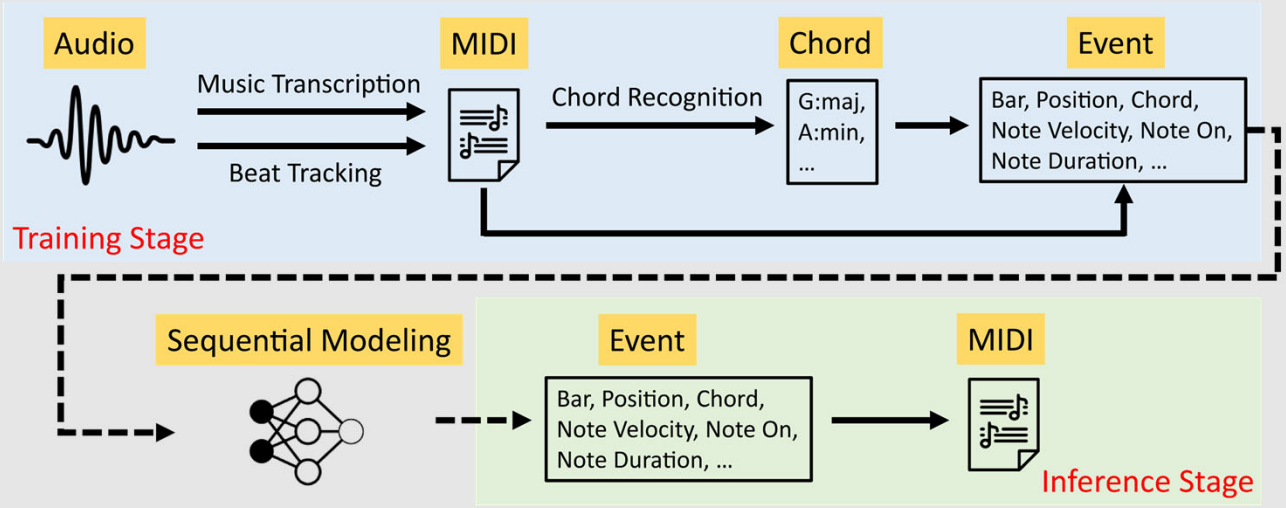
- Encoder: Bi-LSTM + MusicVAE
- Decoder: Transformer



[1] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. 2021. Symbolic music generation with diffusion models. arXiv preprint arXiv:2103.16091 (2021).

SOTA1: Pop Music Transformer [1]

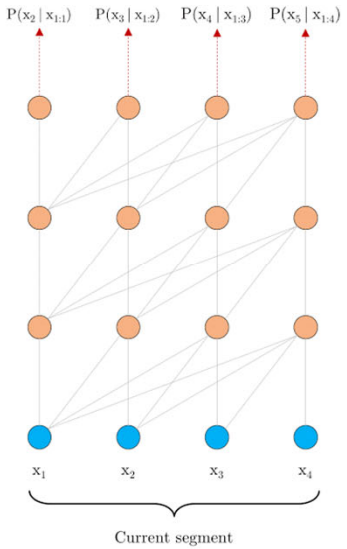
- Based on Music Transformer
  - 1. Using condensed encoding: **REMI**
  - 2. Propose Segment-level recurrence mechanism to connect sections of long inputs (**Transformer-XL**)
  - 3. Propose relative positional embedding among sections to replace absolute embedding (**Reusable**)
  - 4. Propose **beat and chord recognition** and incorporate them into the encoding.



[1] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM international conference on multimedia, 1180–1188.

SOTA1: Pop Music Transformer

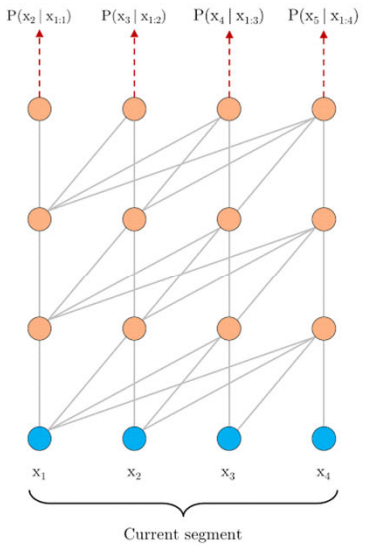
Traditional Transformer



- Segment-level recurrence mechanism

SOTA1: Pop Music Transformer

Transformer-XL

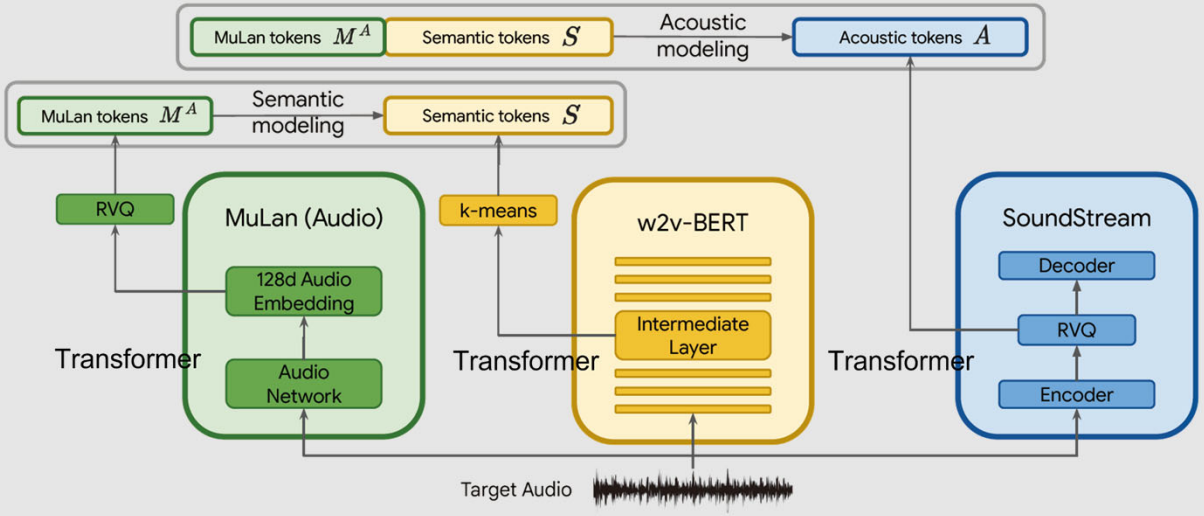


- Segment-level recurrence mechanism

SOTA2: MusicLM [1]

- Generating music from text description

- Three Transformers for:
  - Musical Feature Modeling
  - Semantic Modeling
  - Music Acoustic Modeling (Generator)



"The main soundtrack of an arcade game. It is fast-paced and upbeat, with a catchy electric guitar riff. The music is repetitive and easy to remember, but with unexpected sounds, like cymbal crashes or drum rolls."

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and others. 2023. Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023).

Controls

Making Music Generation

Controllable & Useful

Rule-based Control

Markov Transition Probability Matrix

	C	D	…	F	G
C	/	0.2	0.1	0.4	0.3
D	0.2	/	0.3	0.4	0.1
…	0.1	0.1	/	0.4	0.4
F	0.25	0.25	0.25	/	0.25
G	0.5	0.2	0.1	0.1	/

Composition Rules

- Position
- Tonal
- Beat
- …

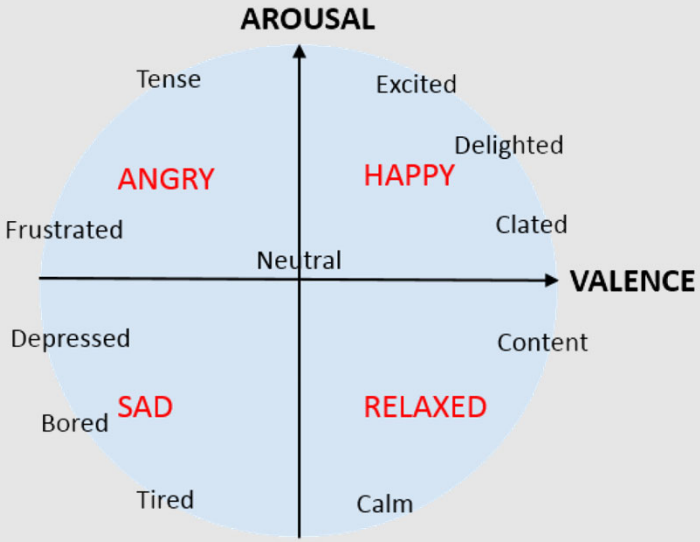
Modified Transition Matrix

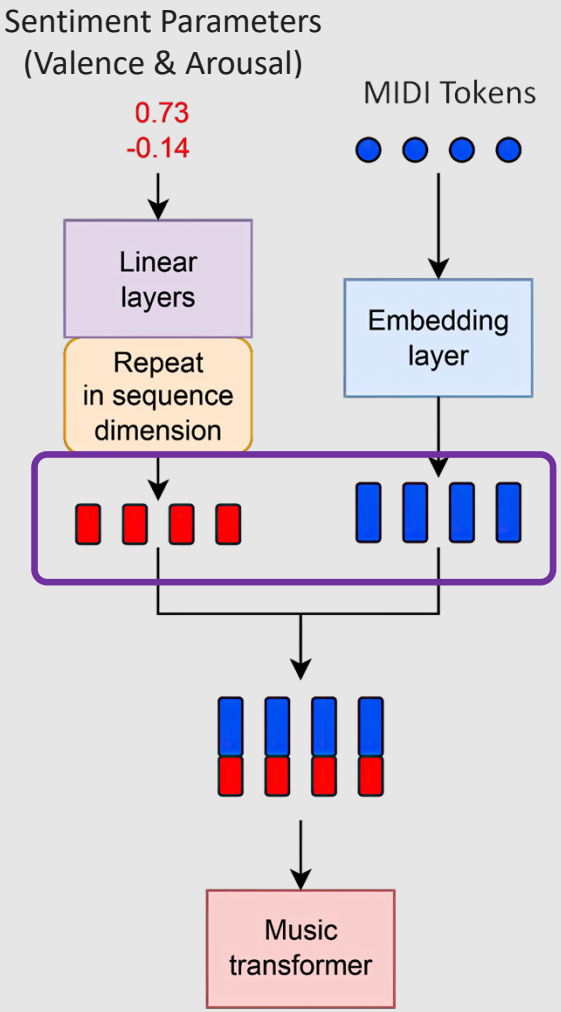
	C	D	…	F	G
C	**/**	**0.2**	**0.1**	**0.4**	**0.3**
D	**0.2**	**/**	**0.3**	**0.4**	**0.1**
…	**0.1**	**0.1**	**/**	**0.4**	**0.4**
F	**0.25**	**0.25**	**0.25**	**/**	**0.25**
G	**0.5**	**0.2**	**0.1**	**0.1**	**/**

Sentiment Control

Convert desirable sentiment to Valence & Arousal parameters

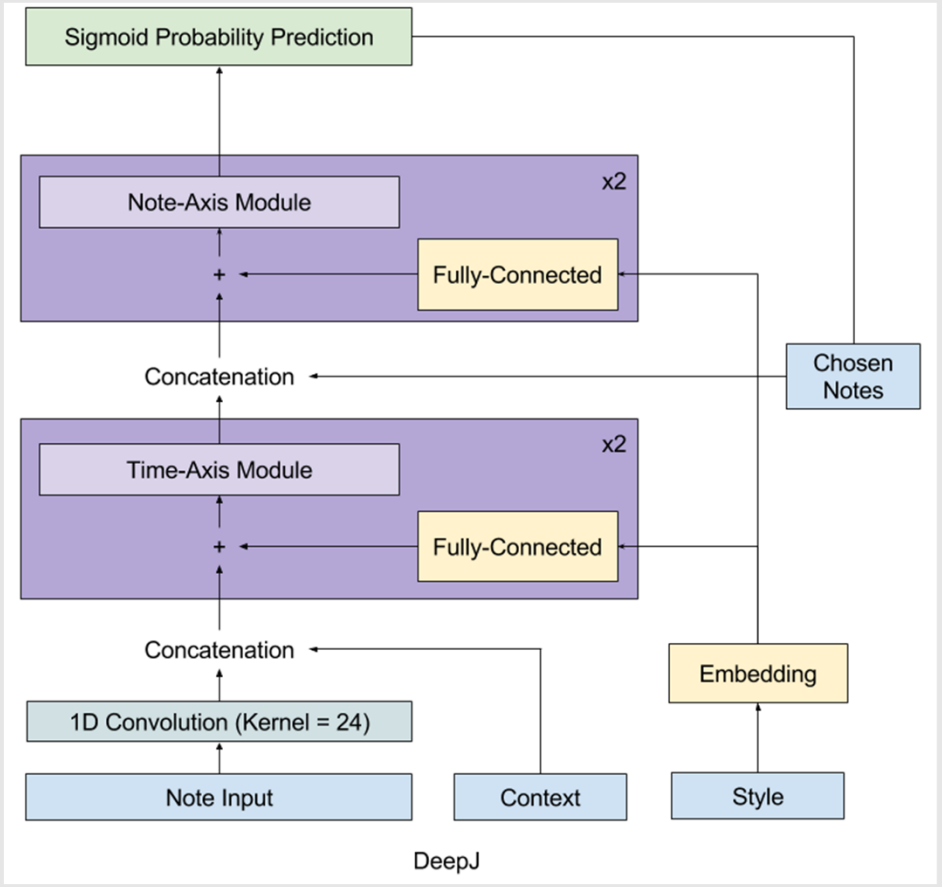
Expand the parameters and append to MIDI embeddings as controlling condition





[1] Serkan Sulun, Matthew EP Davies, and Paula Viana. 2022. Symbolic music generation conditioned on continuous-valued emotions. IEEE Access 10, (2022), 44617–44626.

X

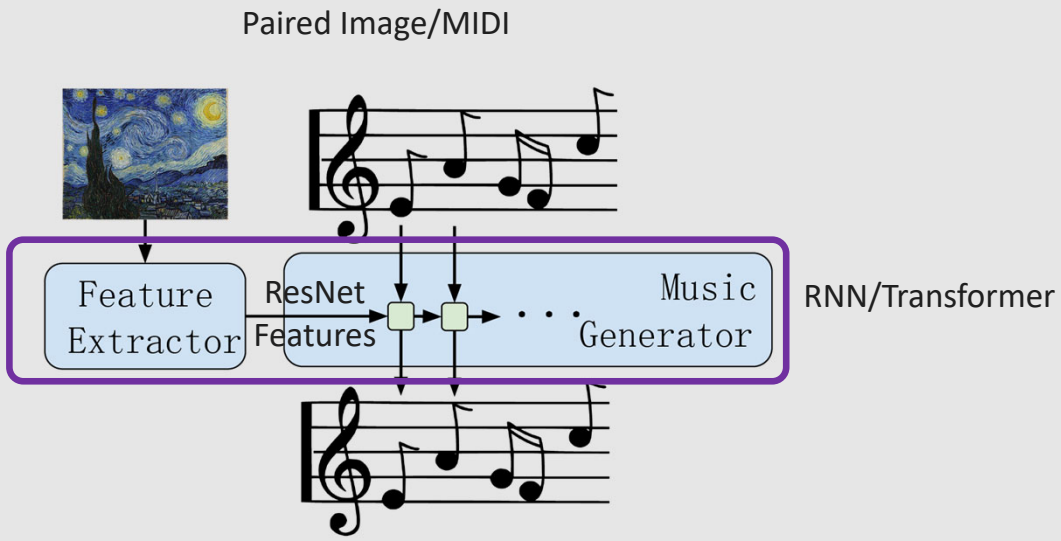

Sigmoid Probability Prediction
x2
Note-Axis Module
+
Fully-Connected
Concatenation
Chosen Notes
x2
Time-Axis Module
+
Fully-Connected
Concatenation
Embedding
1D Convolution (Kernel = 24)
Note Input
Context
Style
DeepJ

Music Generation from Image

Extract visual features of input image (ResNet)

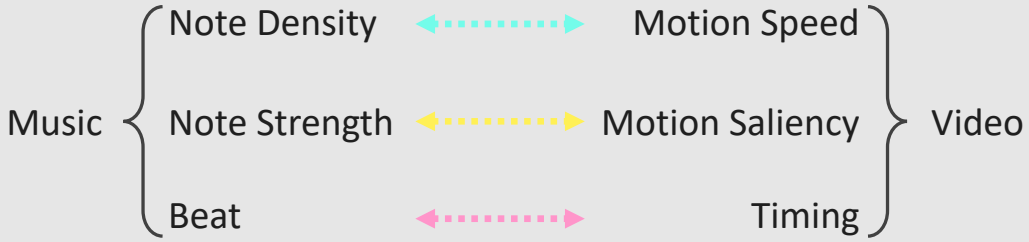
↓

Using extracted features as initial hidden state of the music generator (RNN/Transformer)

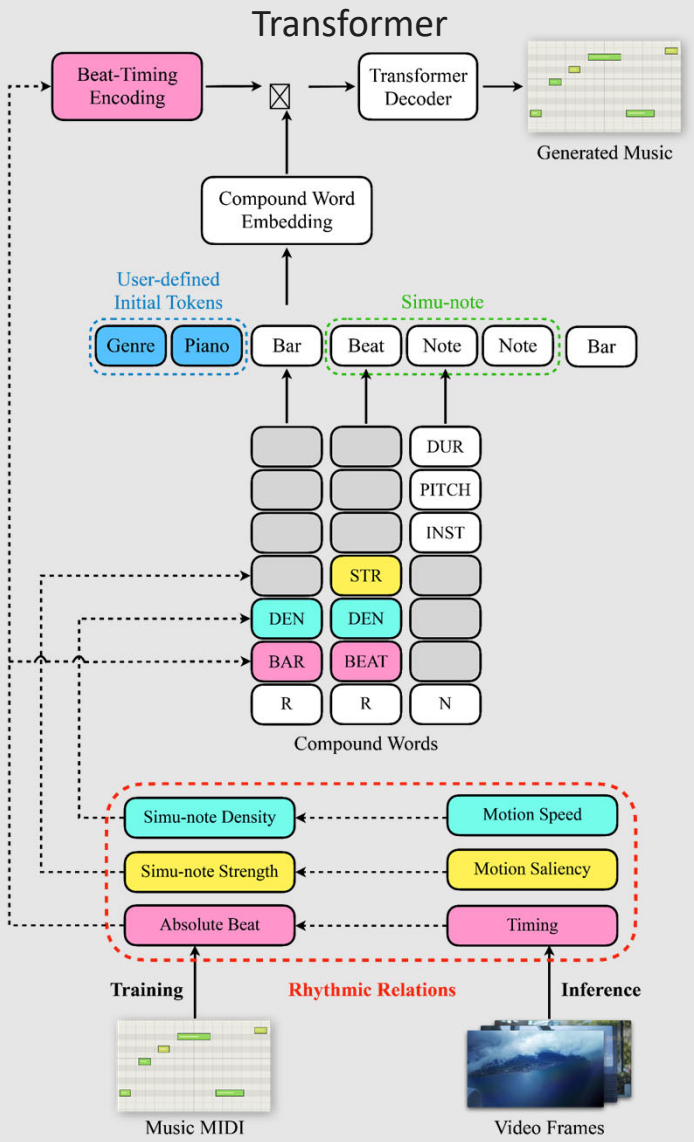


[1] Xiaodong Tan, Mathis Antony, and H Kong. 2020. Automated Music Generation for Visual Art through Emotion. In ICCC, 247–250.

Music Generation from Video [1]



Improve **interpretability**



[1] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video background music generation with controllable music transformer. In Proceedings of the 29th ACM International Conference on Multimedia, 2037–2045.

Music Generation from Video

Improve Interpretability of AMG

Demo

Video source: https://www.bilibili.com/video/BV1wh411n7d9/
Visualization source: credits to Liang Qihao

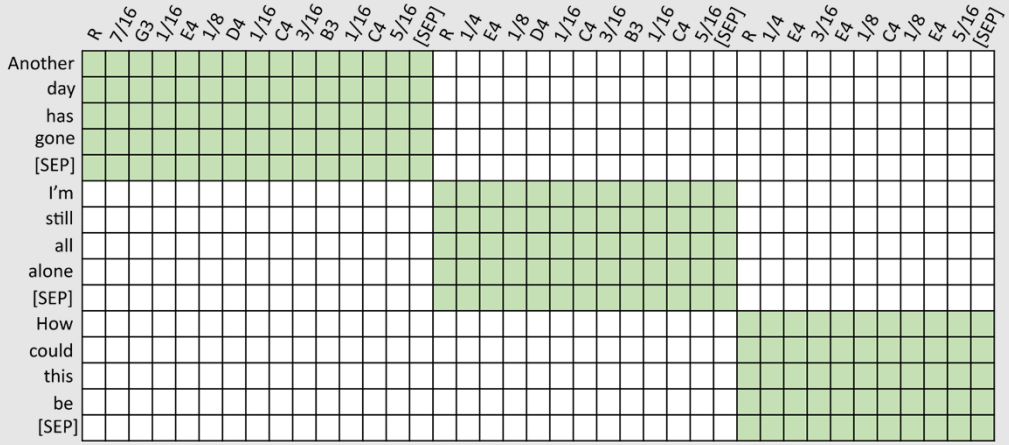
Music Generation from Lyrics

SongMASS

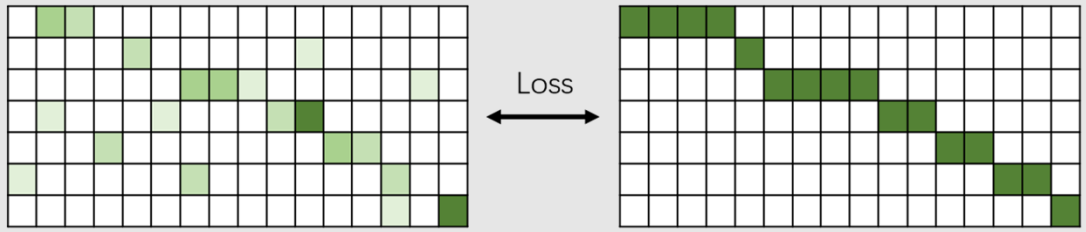
Singability: #Syllables matches #Notes

Using lyric-music align attention matrix to control the note number.

Generator: Transformer



Bar-level alignment attention

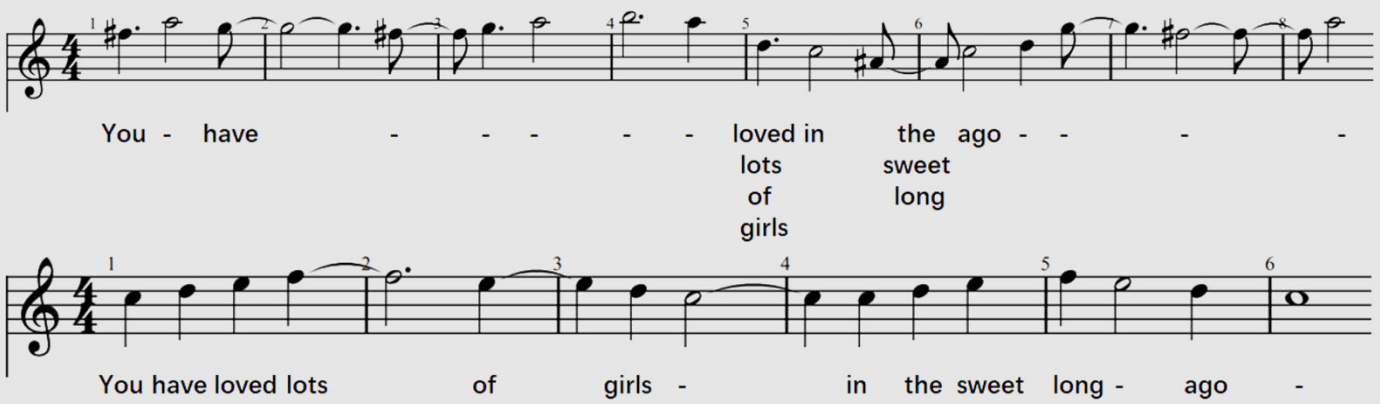


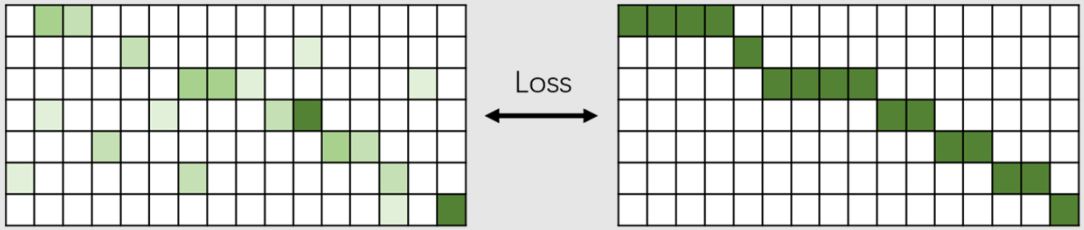
Word-level alignment attention

[1] Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In Proceedings of the AAAI Conference on Artificial Intelligence, 13798–13805.

Music Generation from Lyrics

SongMASS Demo



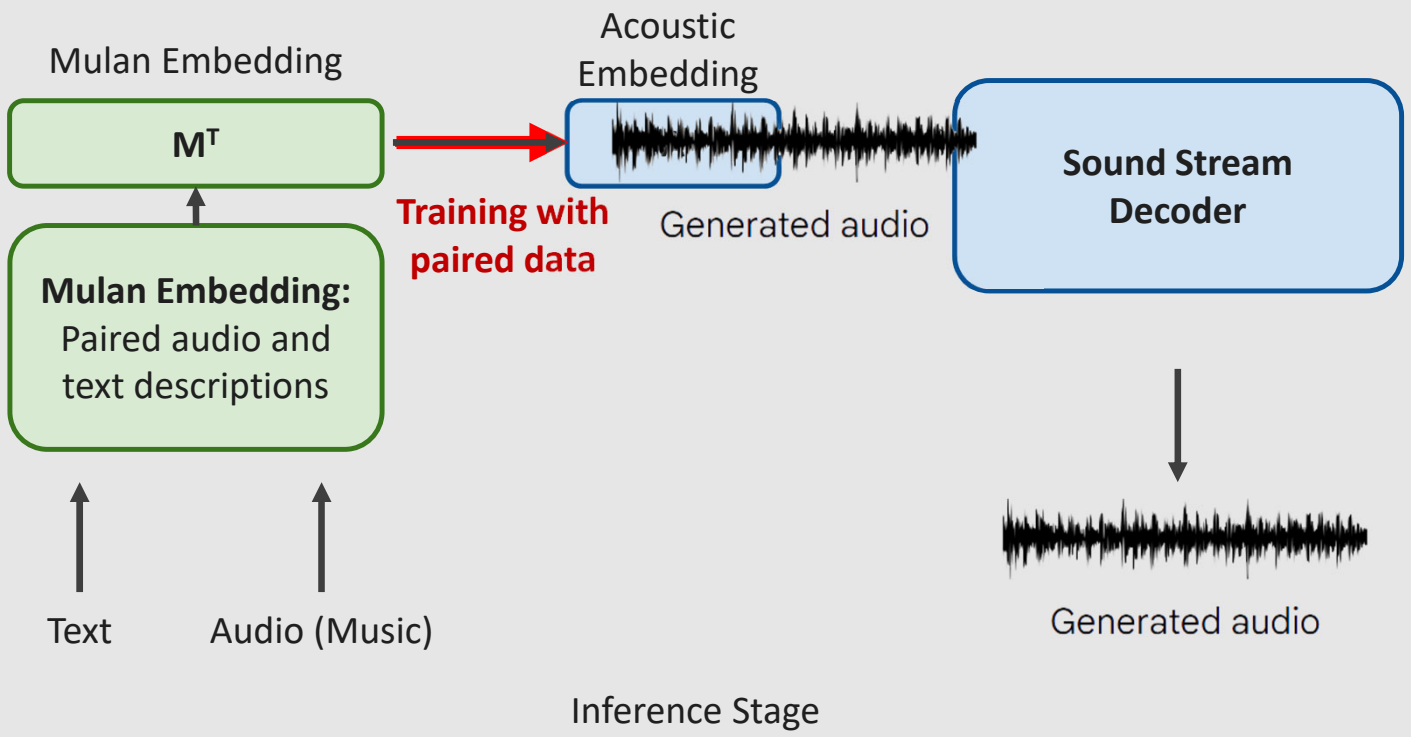


alignment
attention

Music source: https://ai-muzic.github.io/songmass/

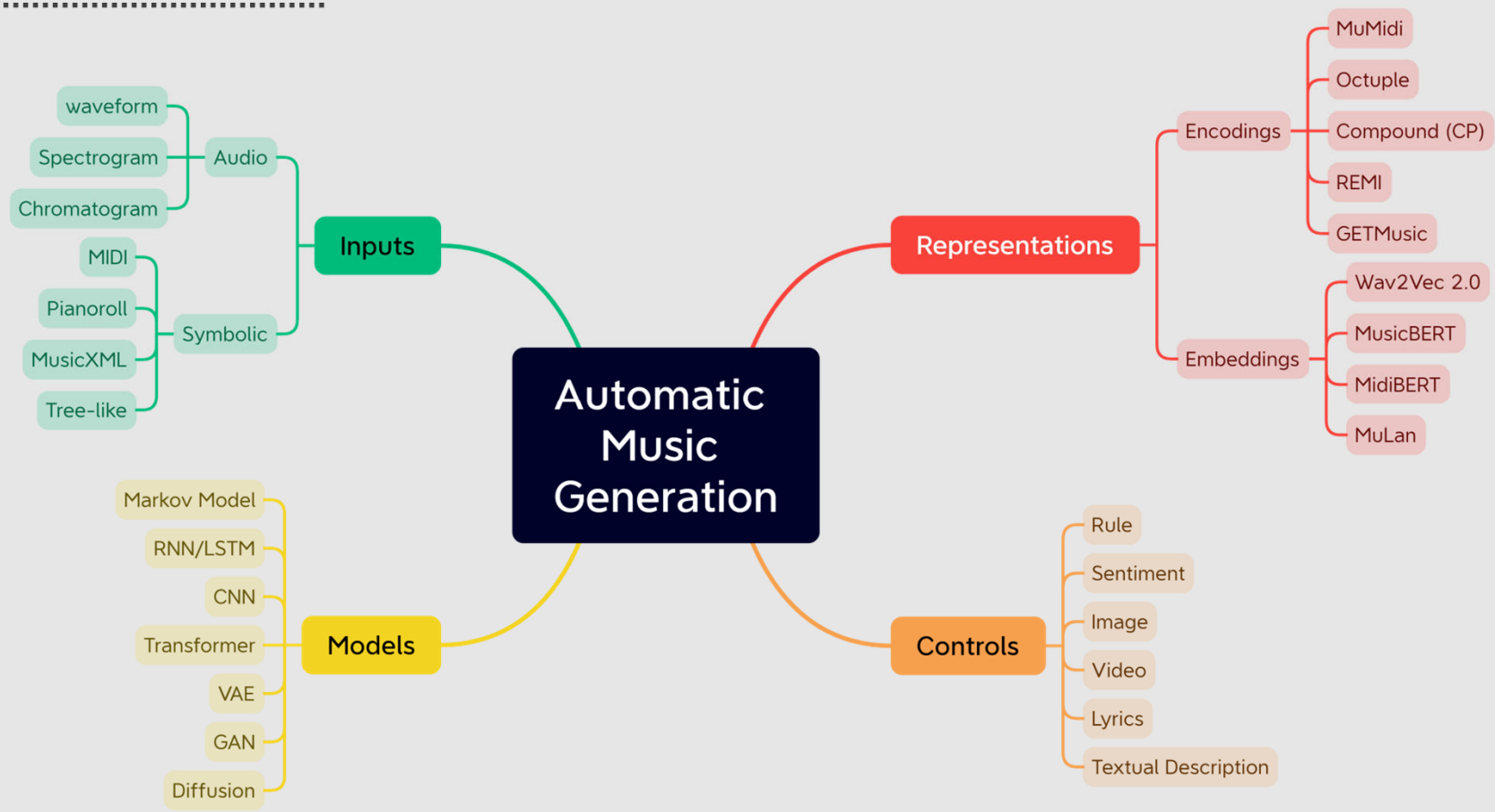
Music Generation from Description

MusicLM [1]:



[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and others. 2023. Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023).

Summary of AMG



Our Works

Personalized Music Generation

For Human Health and Potential

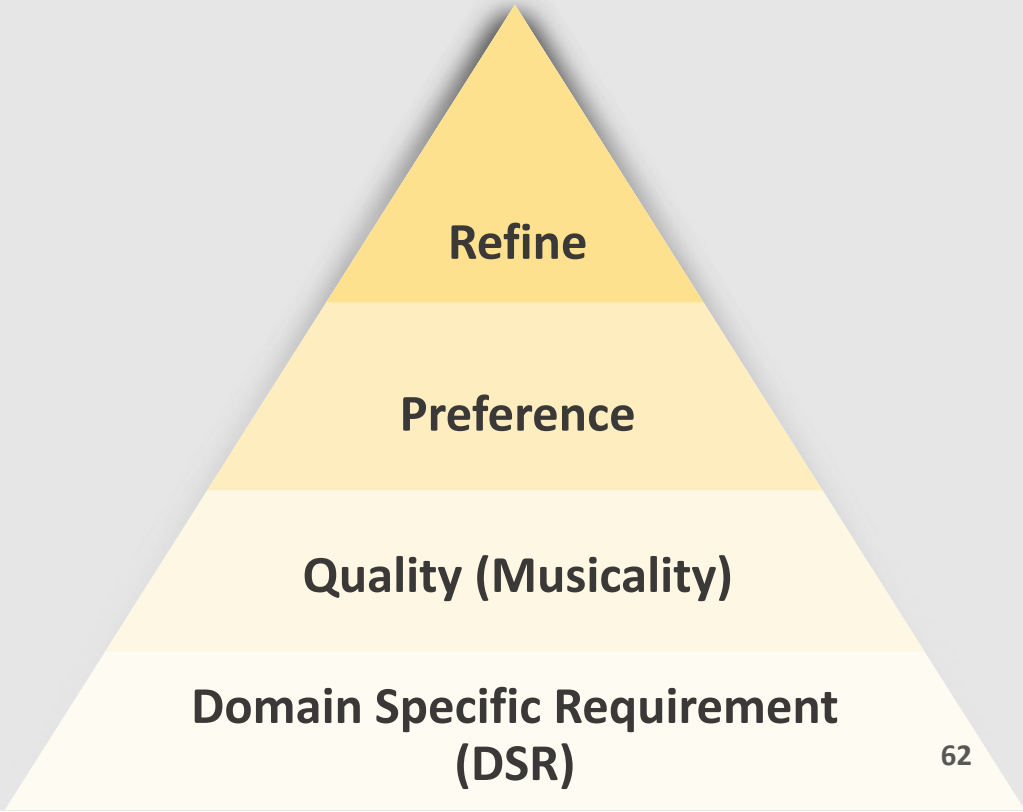
Motivation

Automatic Music Generation for Rhythmic Auditory Stimulation (RAS)

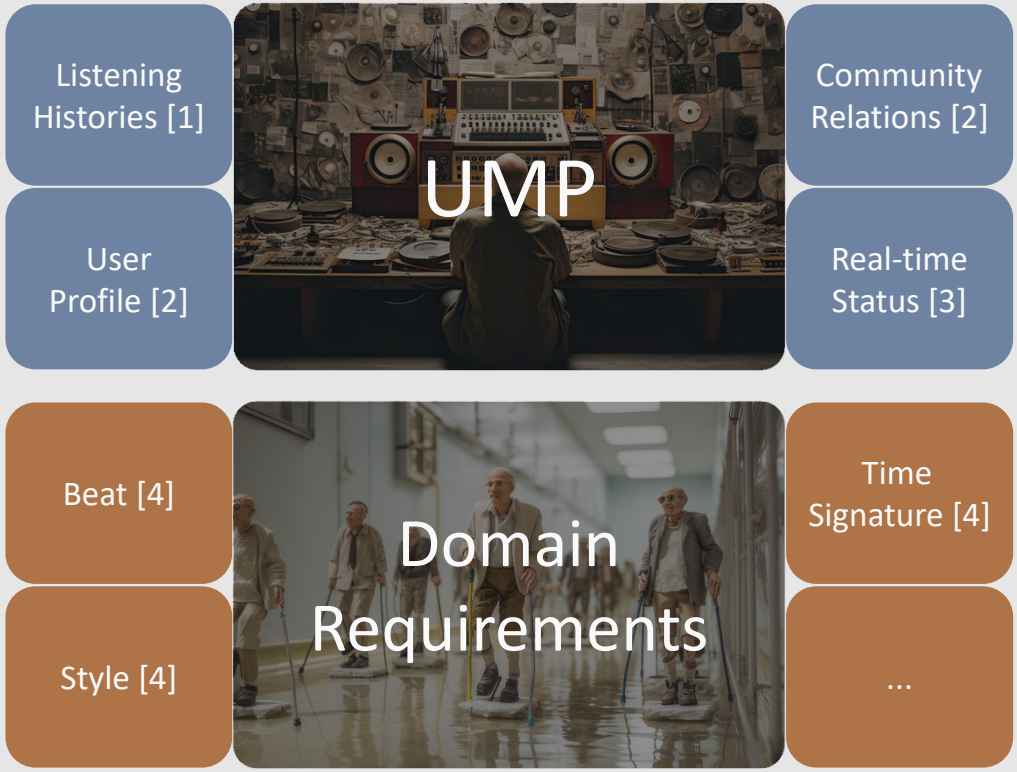
Stabilize gaits of Parkinson's Disease patients

Metronome (Tedious) → Music (Enjoyable)

4 Aspects of Requirement



User-Preference-Aware AMG

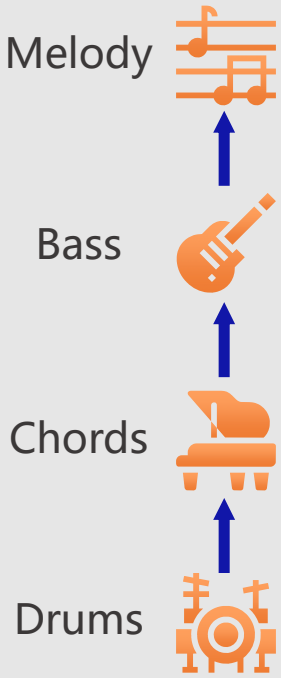


[1] Xichu Ma, Yuchen Wang, and Ye Wang. 2022. Content based User Preference Modeling in Music Generation. In Proceedings of the 30th ACM International Conference on Multimedia, 2473–2482.
[2] Paper under review.
[3] Project is underway.
[4] Shuqi Dai, Xichu Ma, Ye Wang, and Roger B Dannenberg. 2022. Personalised popular music generation using imitation and structure. Journal of New Music Research 51, 1 (2022), 69–85.

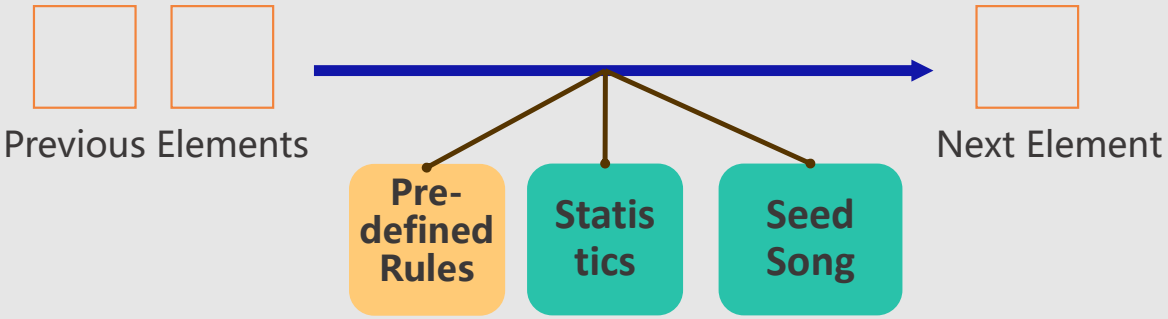
Music Generator: Markov Model + Rule Control

Rules | Statistics

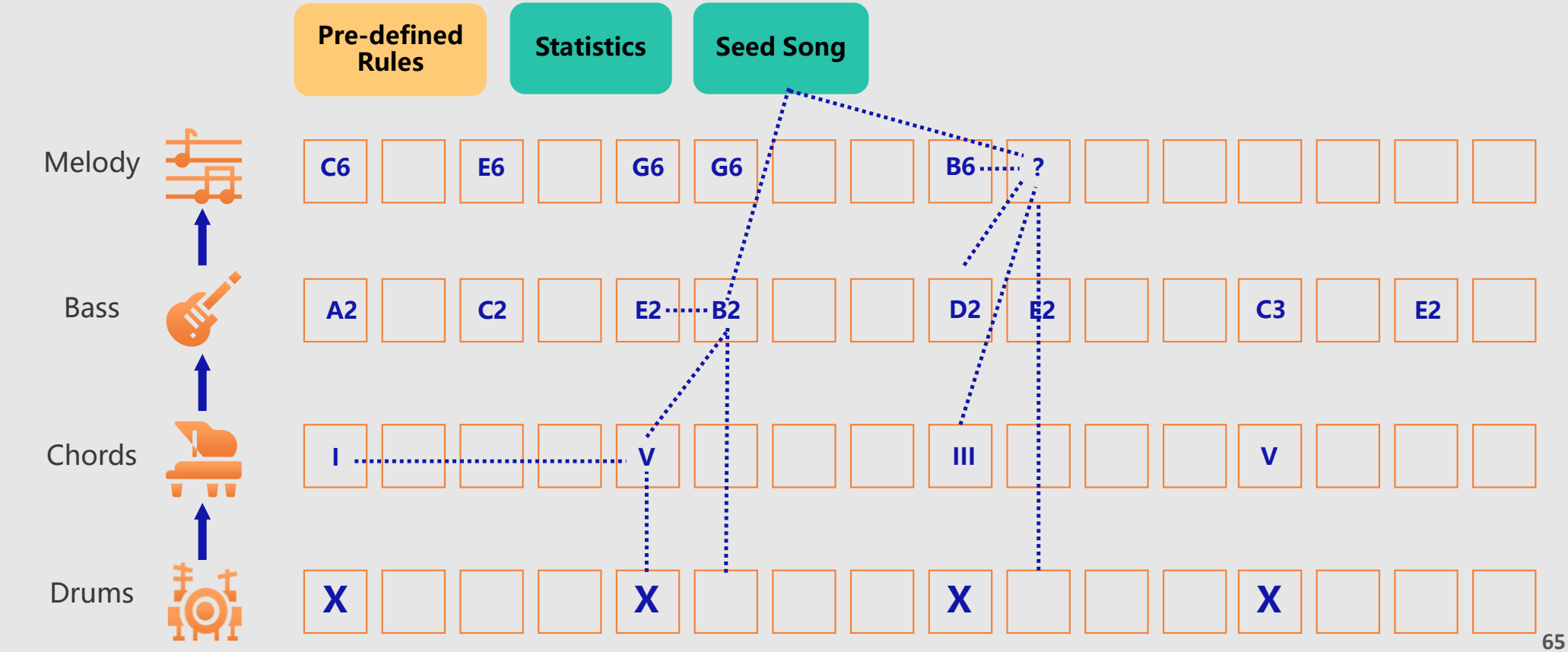
Stable, Controllable and Faster



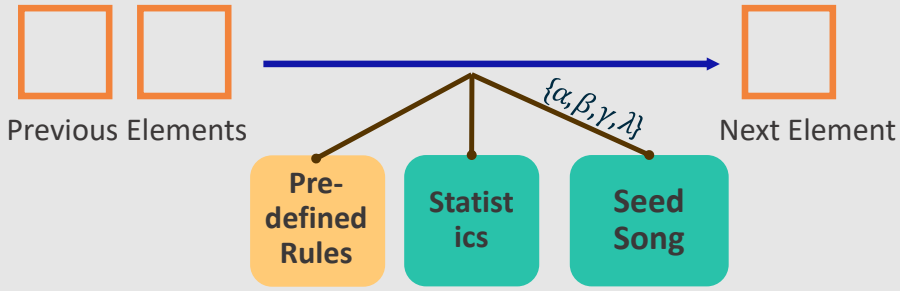
Generating each component Element by Element



Music Generator: Markov Model + Rule Control



Personalization 1.0: Imitating a Seed Song

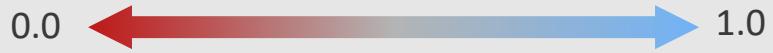


Seed Song (MIDI File)

A seed song is selected by patients,
Reflecting their music preference.

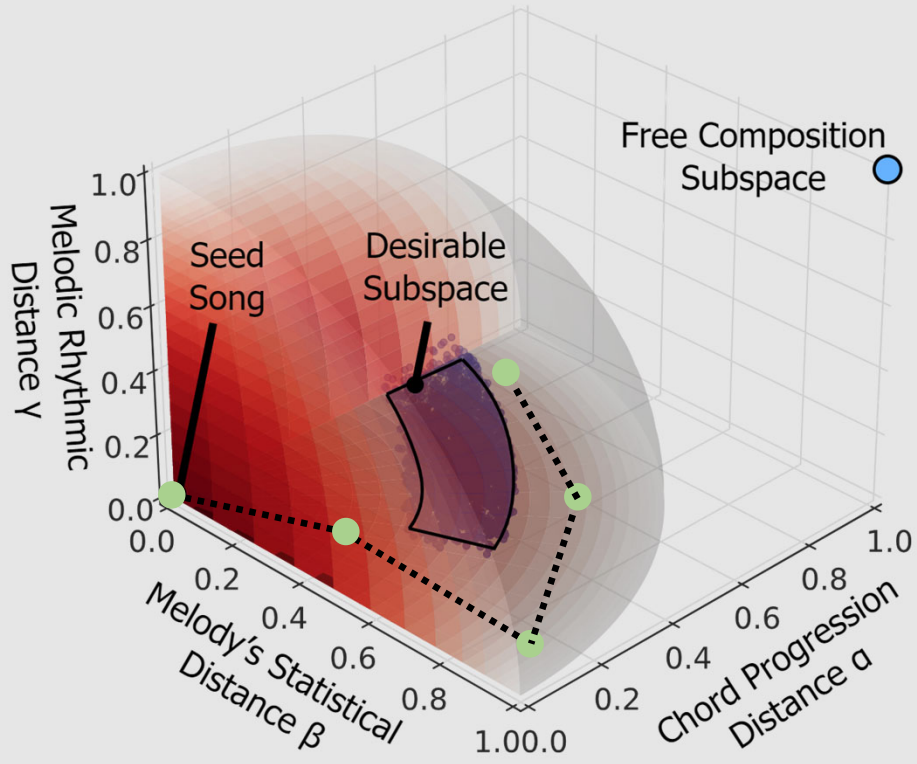
Chord Transition α | Melody Statistics β | Contour γ | Rhythm λ

4 Distance Parameters

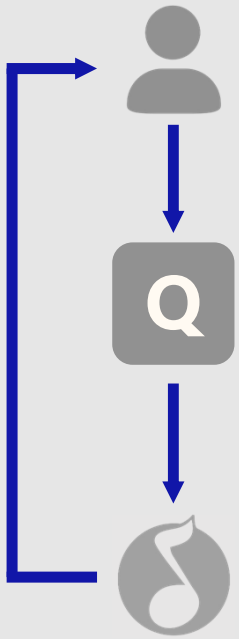


Identical to seed song — Free Composition

Personalization 1.0: Q-learning



Q-learning



User

Q-table
(Strategy)

A song
generated with
new parameters

Demos: Imitating Bach

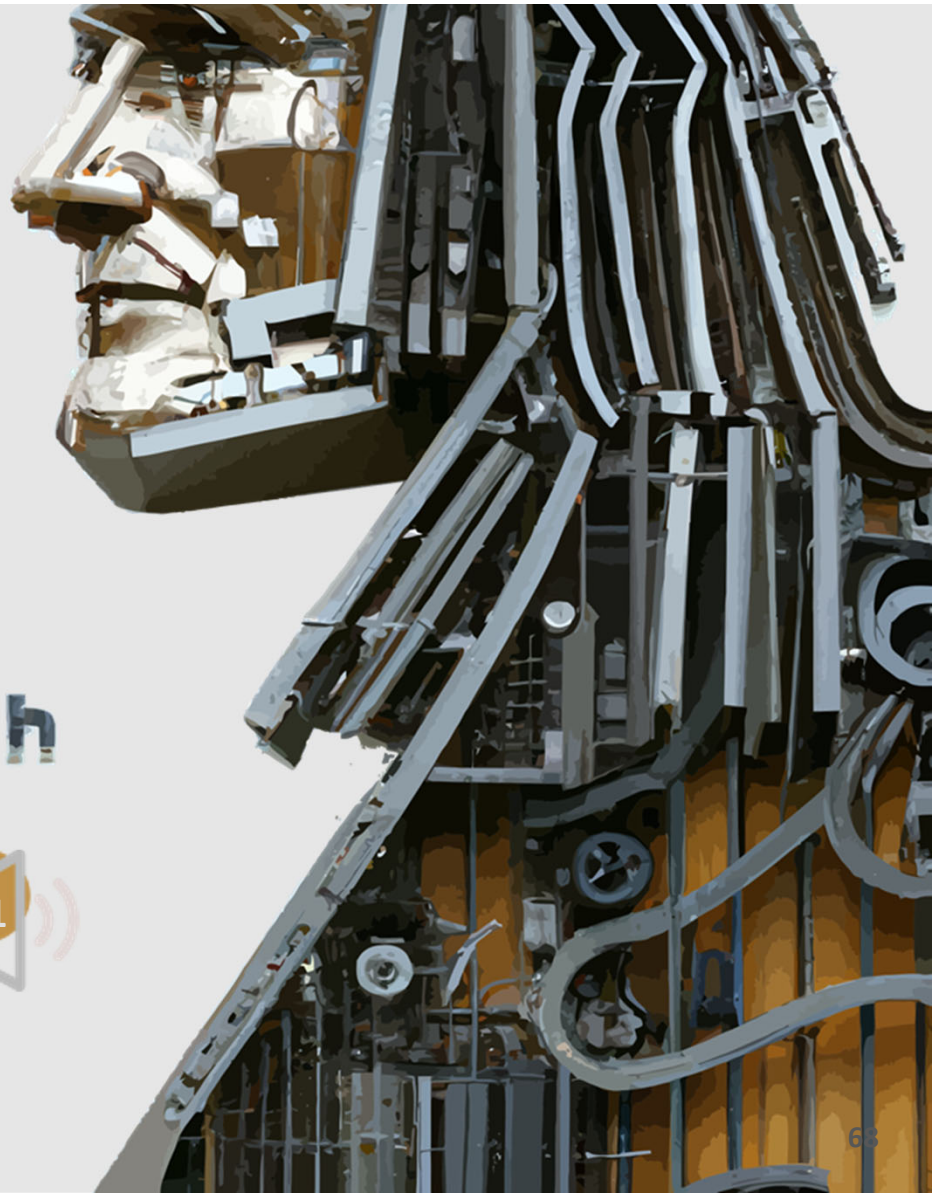
Seed Song: Badinerie

Chord Transition $\alpha = 0.4$

Melody Statistics $\beta = 0.4$

Contour $\gamma = 0.4$

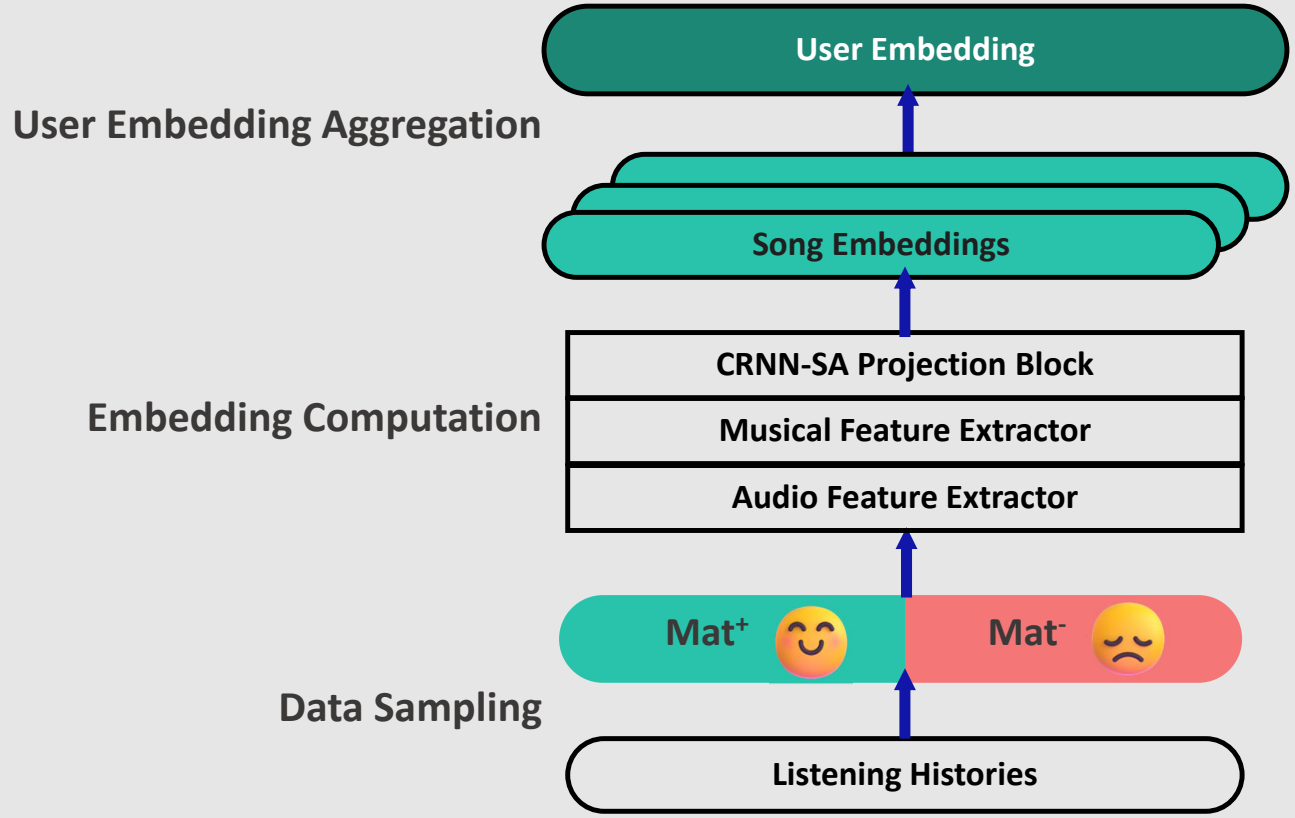
Rhythm $\lambda = 0.4$



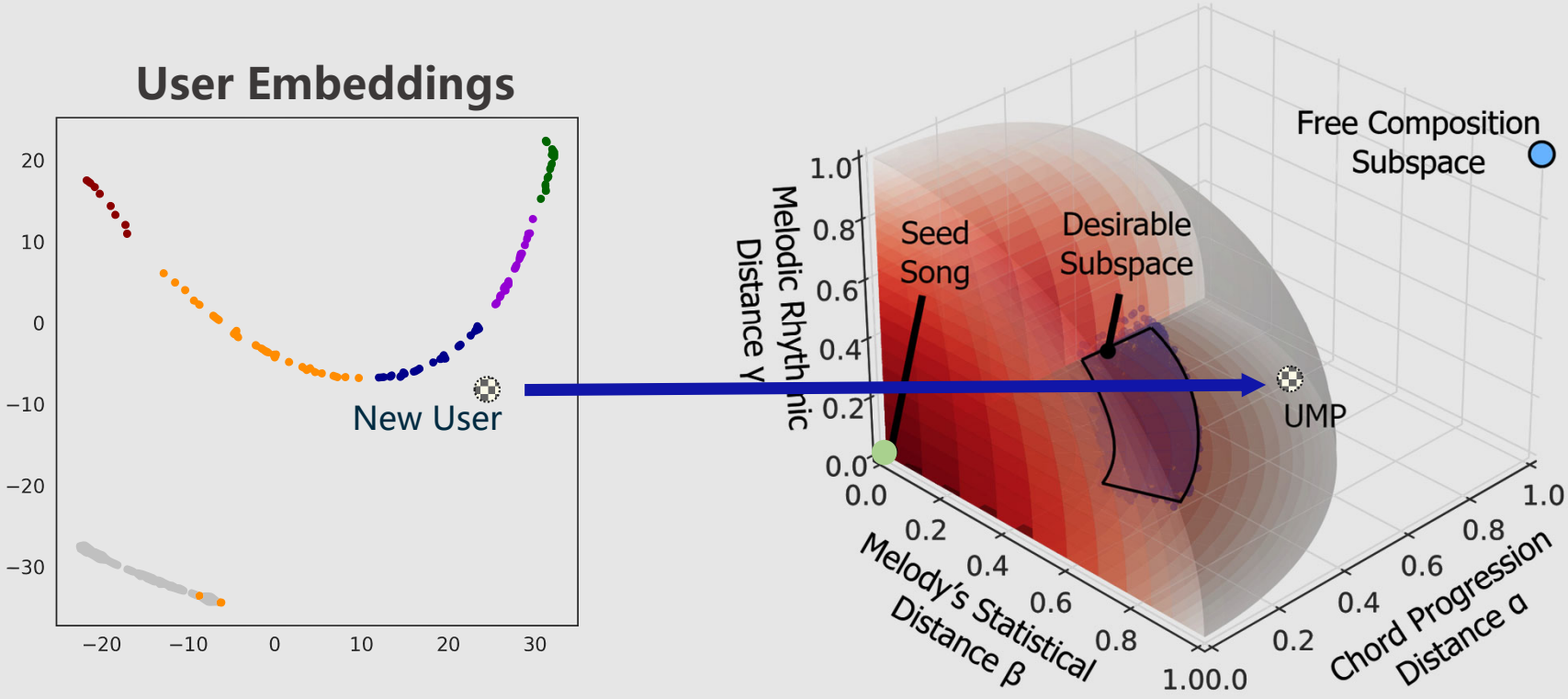
Personalization 2.0

Problem with UMP 1.0:
One user has only one parameter set for all songs.

Modeling User Music Preference from Listening Histories



Personalization 2.0



Projecting UMP to AMG parameters

Personalization 2.0

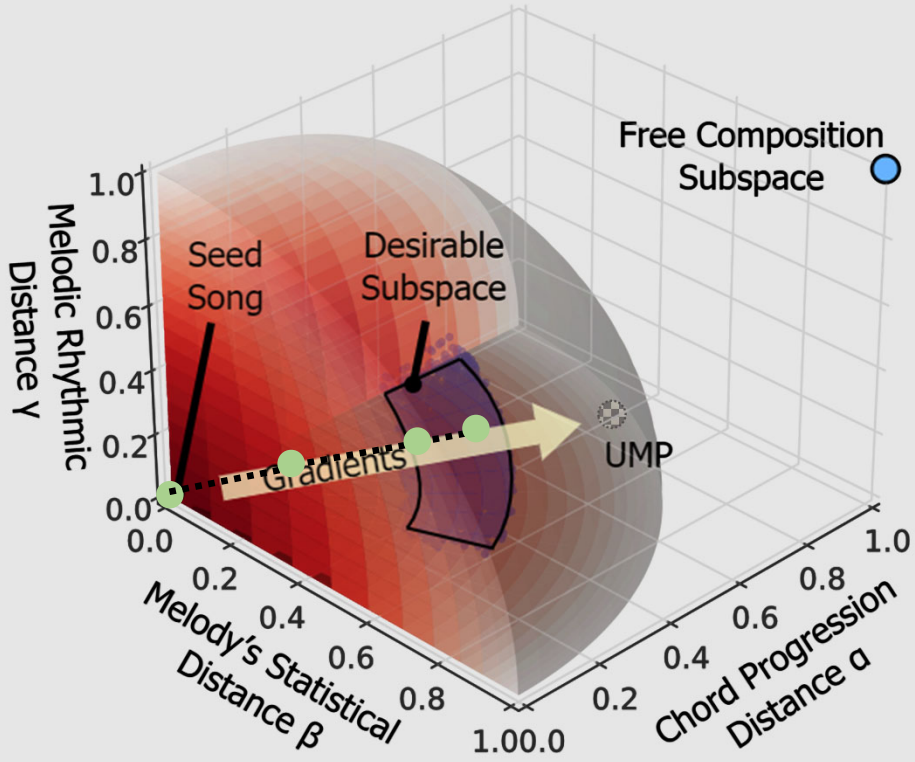
UMP's Direction √

4 Distance Parameters Update

Update the four distance parameters to minimize the loss (distance).

$$\arg\max_{\alpha,\beta,\gamma,\lambda} L^{bp} \Leftrightarrow \{\alpha,\beta,\gamma,\lambda\} -= \boxed{gd} * \boxed{sl}$$

Gradients Step length



Projecting UMP to AMG parameters:
One user has an independent parameter set for each seed song.

Personalization 3.0: Multi-facet UMP

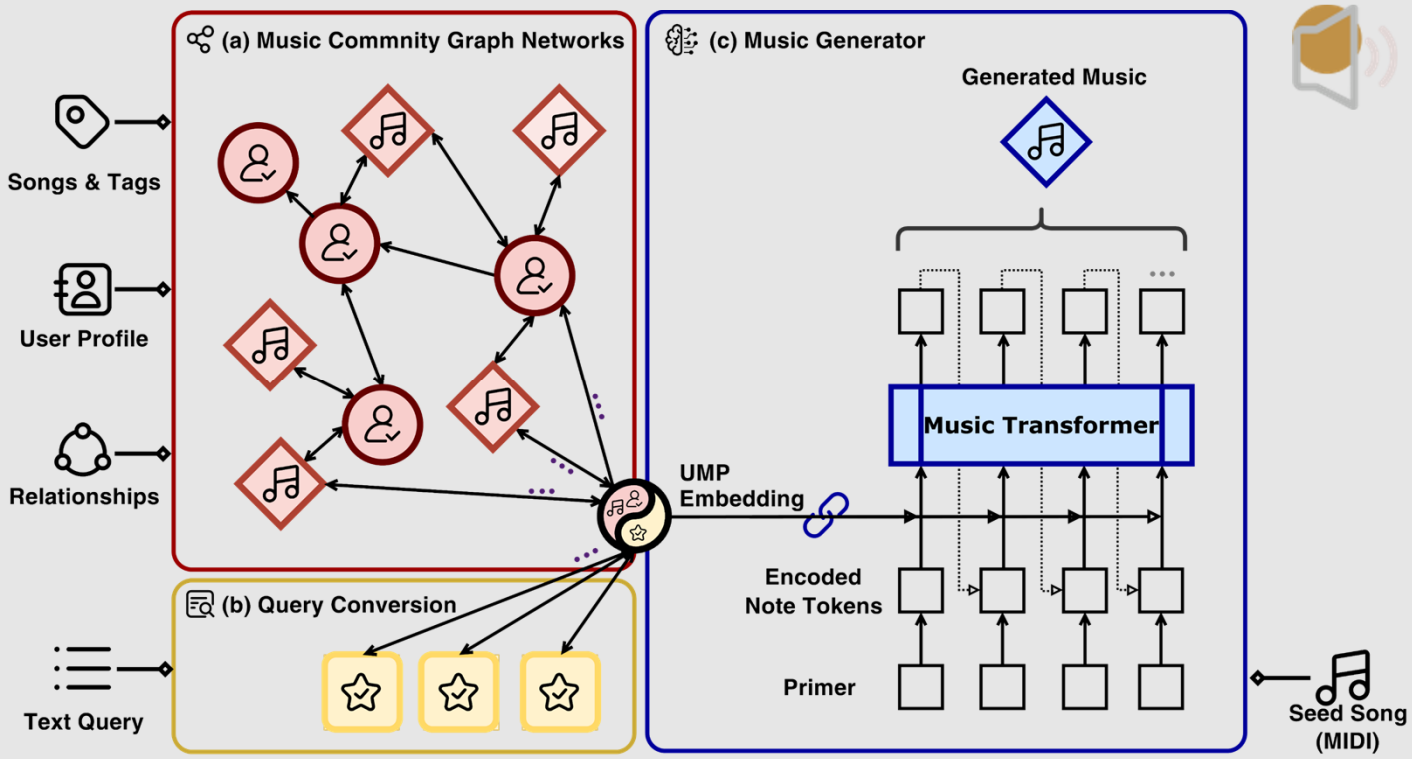
Model UMP from Music Community Graph

↓

Convert Text Query to Graph Nodes

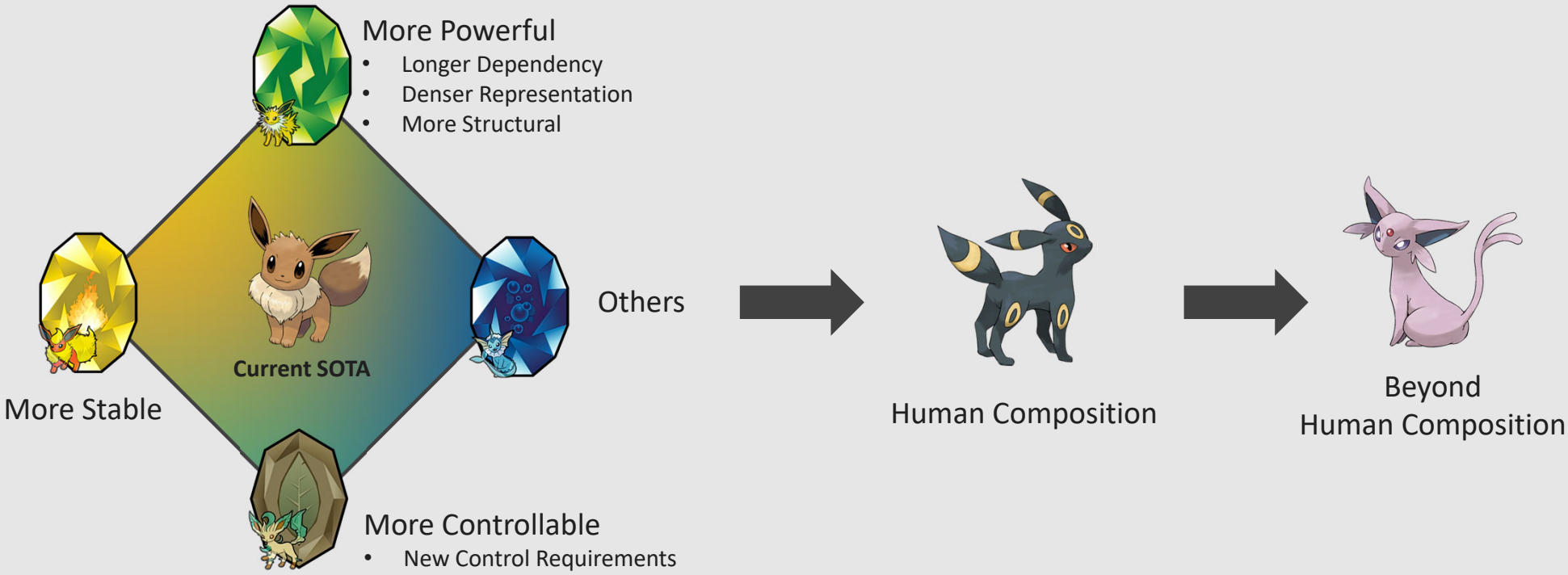
↓

Node Embedding conditioning AMG (MusicTransformer)



One user has multi-facet parameter sets for each seed song.

Challenges: Future Directions

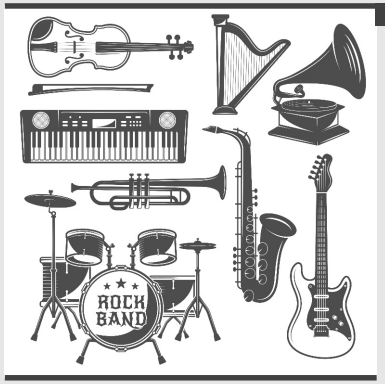


Related Topics

I

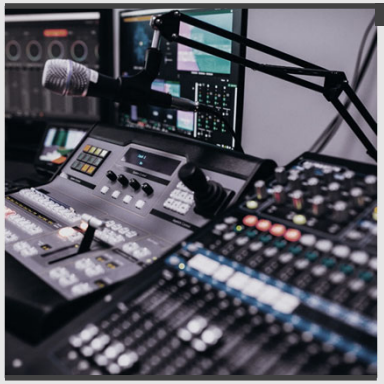
Song Writing: Lyrics Generation

II



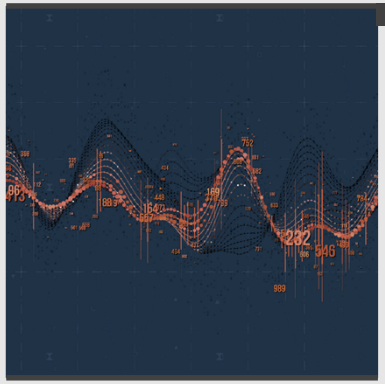
Accompaniment & Arrangement Generation

III



Performance Model

IV



Automatic Music Analysis/Understanding

Useful Datasets

- Music Information Retrieval (MIR) related datasets
  Audio, MIDIs, lyrics, drums, genres, …

  https://ismir.net/resources/datasets/
  https://gist.github.com/alexanderlerch/e3516bffc08ea77b429c419051ab793a
  https://paperswithcode.com/datasets?o=newest&task=music-generation&mod=audio
  https://github.com/pmlg/generative_music_playground (Data used in popular AMG models)

- Compared with LLMs, Image Generation and other areas:
  Music data is still scare and difficult to annotate.

THANKS FOR WATCHING

Question & Answer

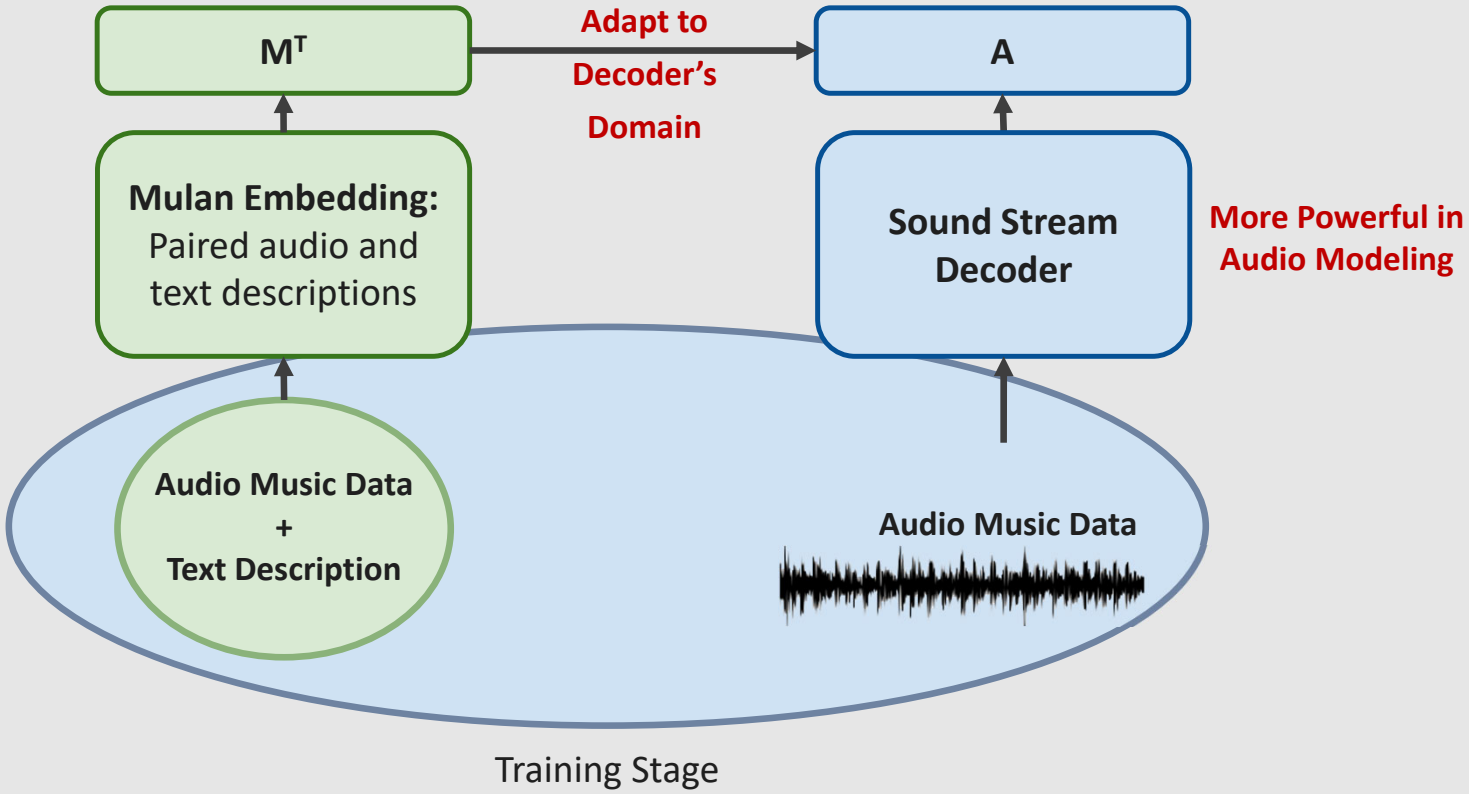
CS4347 · CS5647

Automatic Music Generation: More Effective, More Structural, More Controllable

Contact: Stan Ma (ma_Xichu@nus.edu.sg)

Music Generation from Description

MusicLM [1]:



[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and others. 2023. Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023).