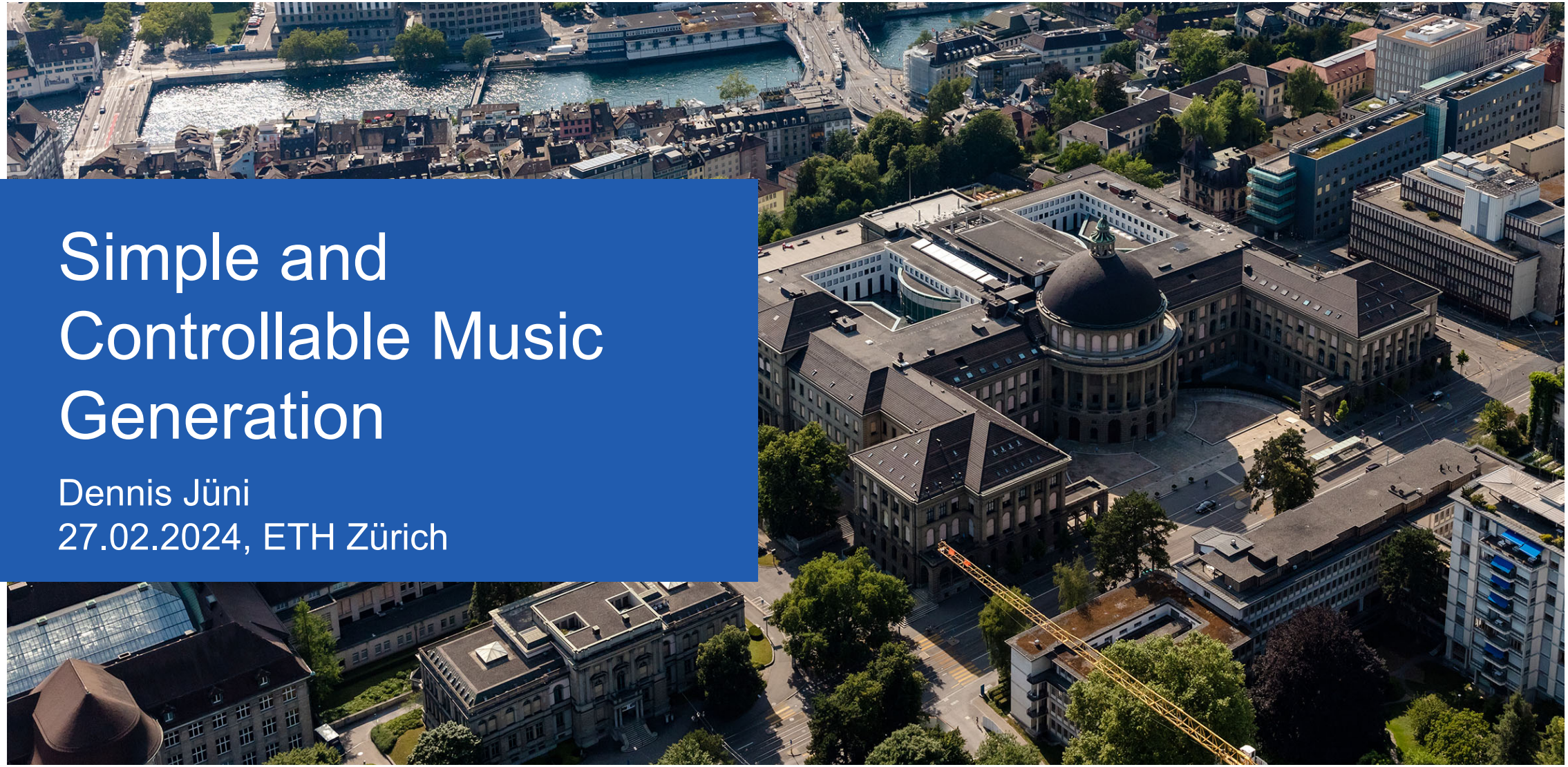
ETH zürich



Simple and Controllable Music Generation

Dennis Jüni

27.02.2024, ETH Zürich

Seminar in Deep Neural Networks

Motivation – Spot the Generated Tune!

1

Real

2

Generated

MusicGen GitHub [1]
MusicGen sample page [2]

Motivation – Challenges of Generating Music

$4 \cdot 60s \cdot 44kHz$
$> 10^6$ timesteps

Harmonies & melodies

Disharmony

Method set

Audio-text data

Previous Work

Riffusion

- Late 2022
- Diffusion model
- Not open source
- No paper

MusicLM

- Jan 2023
- By Google
- Hierarchical seq2seq
- Not open source
- 280’000 hours

Moûsai

- Oct 2023
- ETH
- 2-stage diffusion model
- Open source
- 2’500 hours

Agostinelli et al. [3]
Forsgren and Martiros [4]
Schneider et al. [5]

Previous Work

AudioLDM

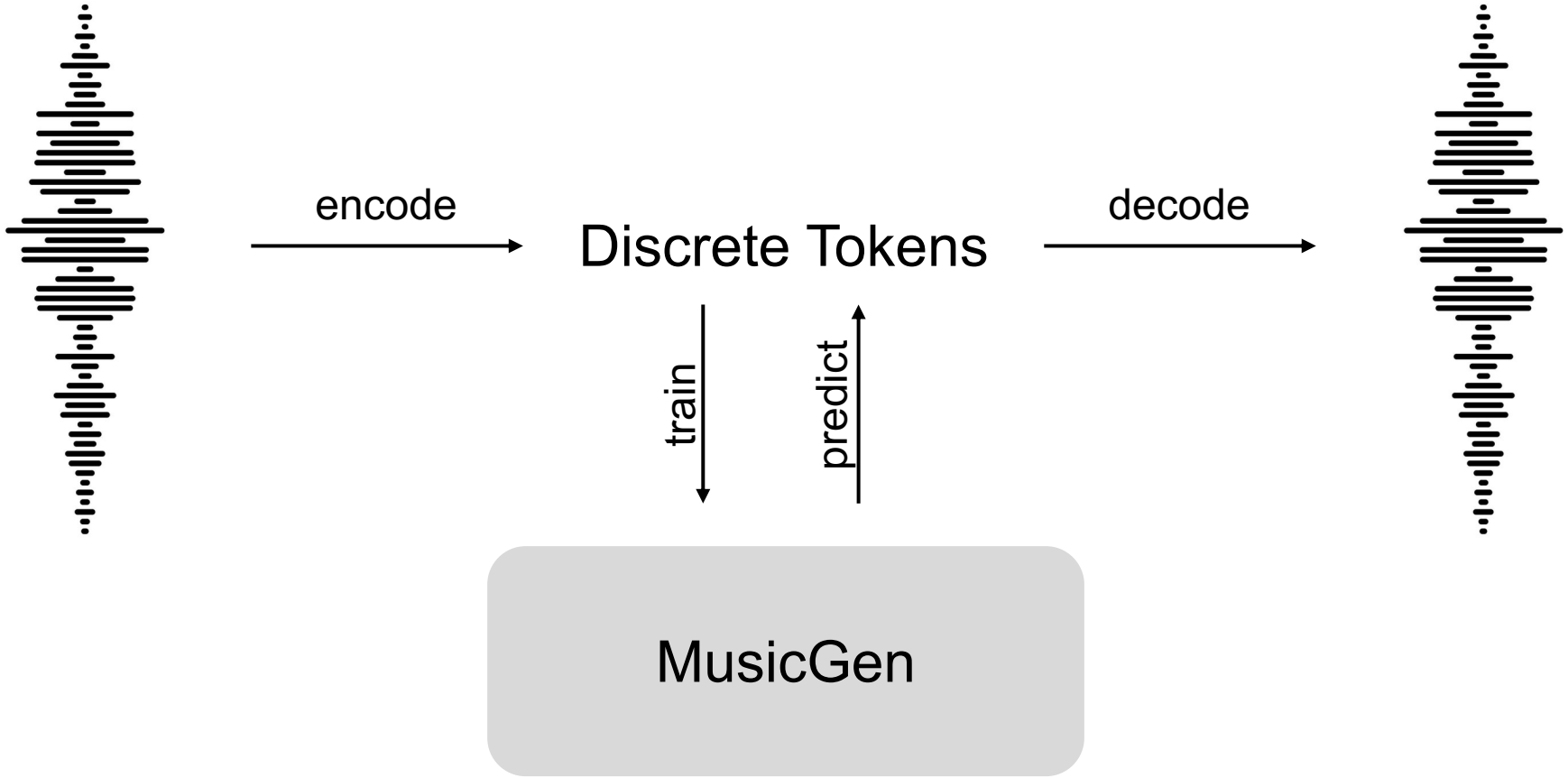
Stable Audio

Suno

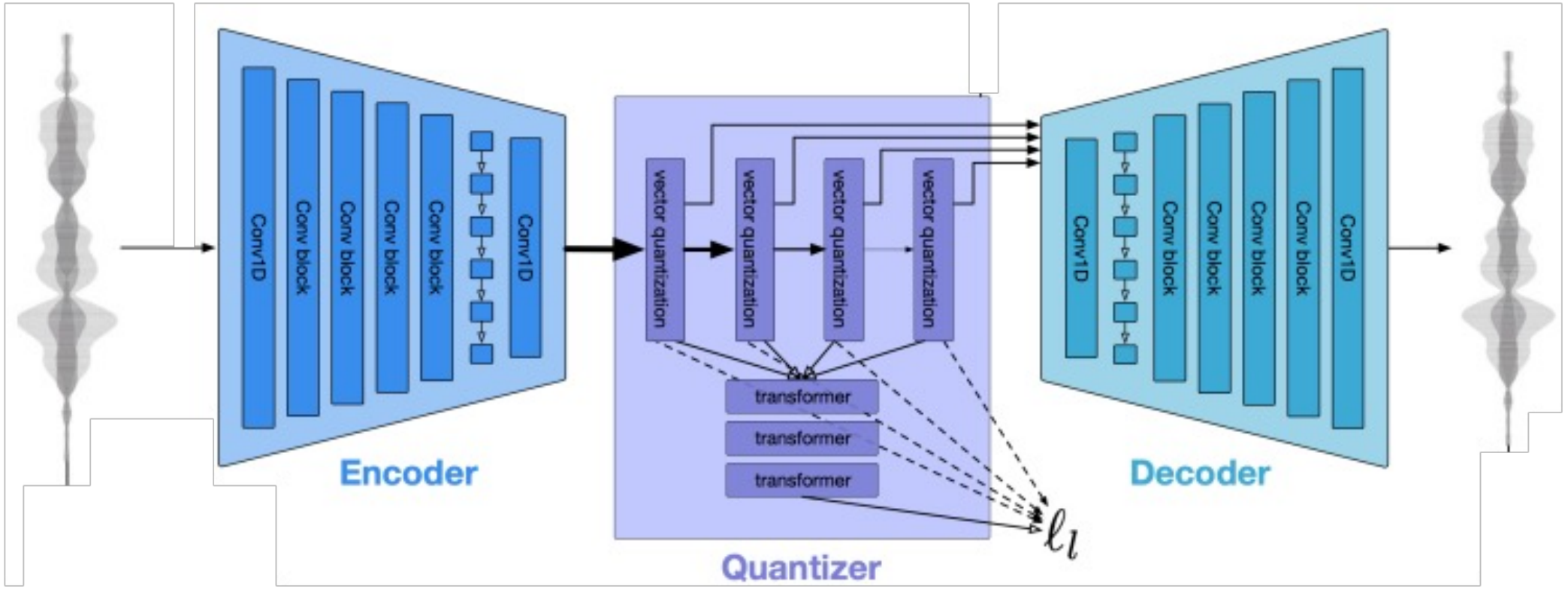
MeLoDy

Liu et al. [10]
Stable Audio Website [11]
Suno AI Website [12]
Lam et al. [13]

MusicGen – Overview



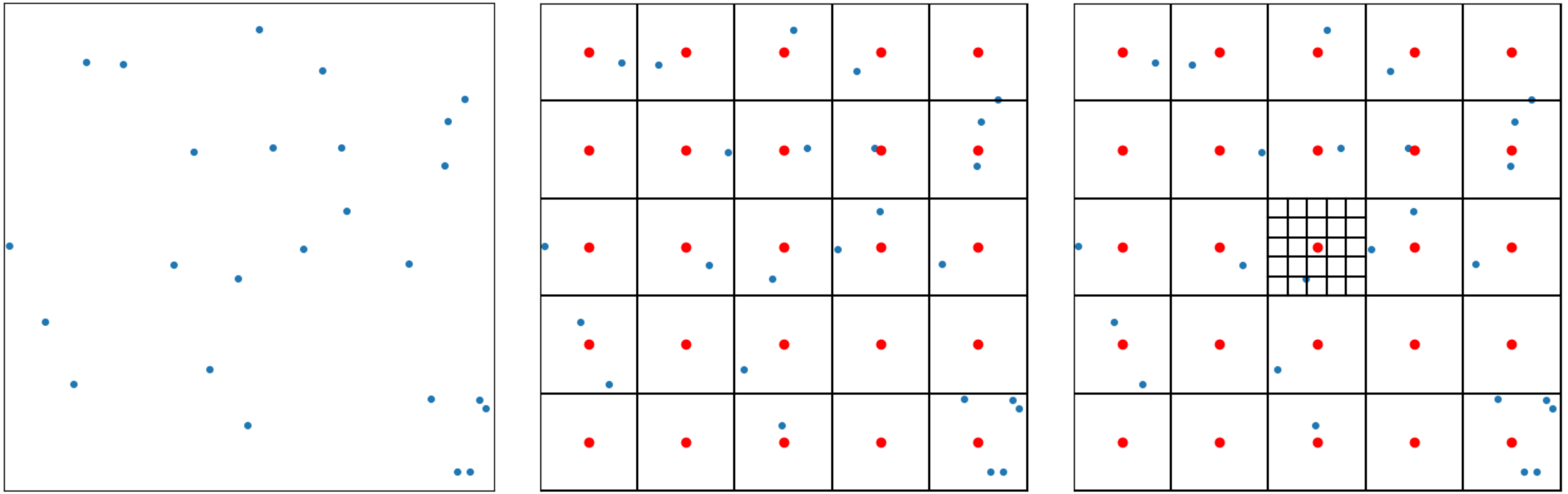
Transforming Music into Discrete Tokens – Encodec



Défossez et al. [6]

Transforming Music into Discrete Tokens – RVQ

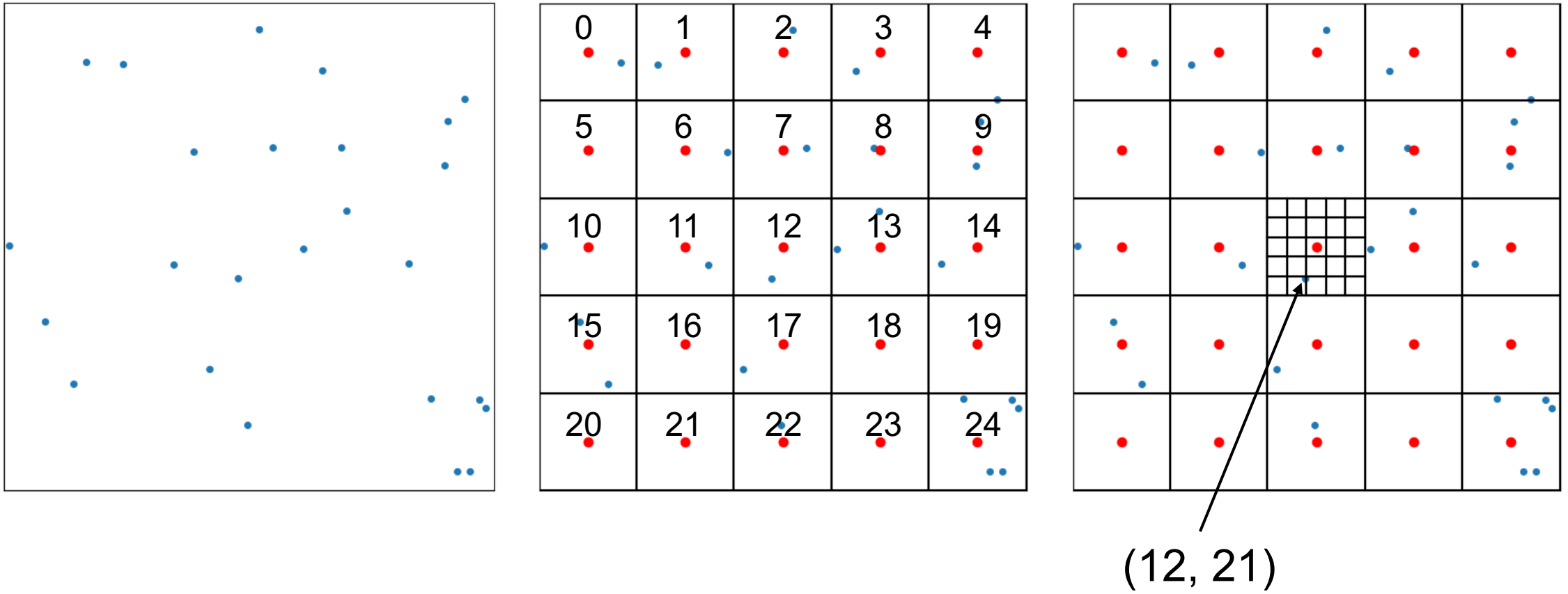
Simplified visualization of audio samples



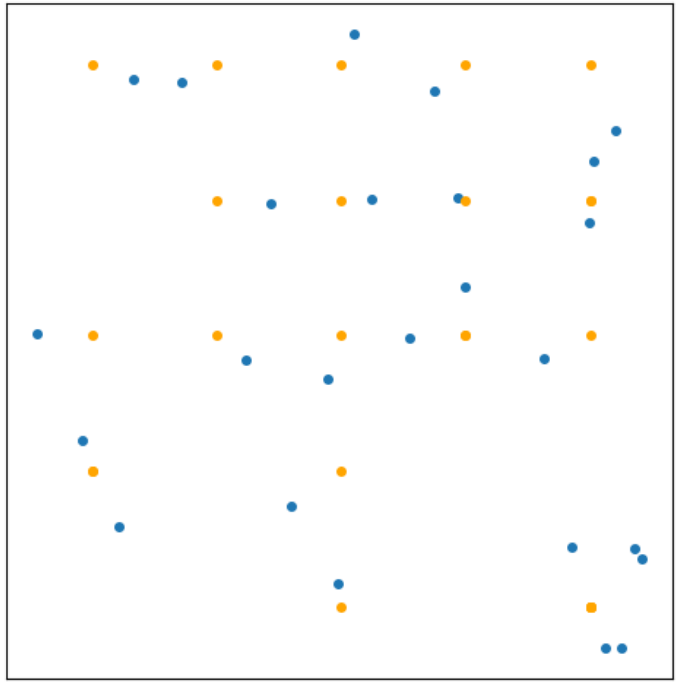
Inspired by Hawley [14]

Transforming Music into Discrete Tokens – RVQ

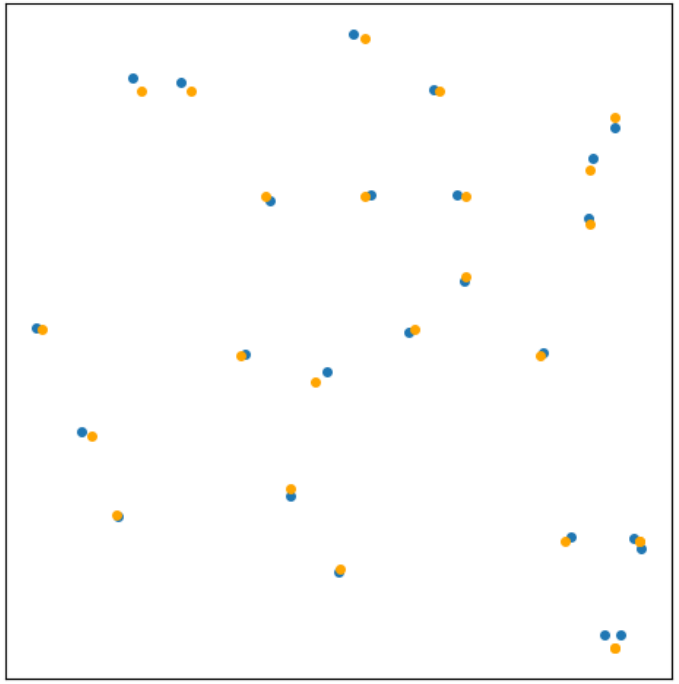
Simplified visualization of audio samples



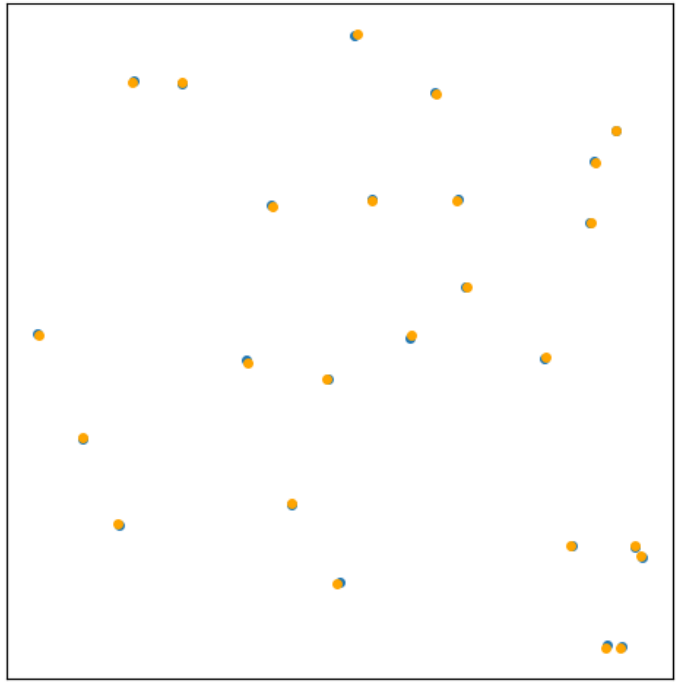
Transforming Music into Discrete Tokens – RVQ



1 codebook
MSE: $2.83 \cdot 10^{-3}$



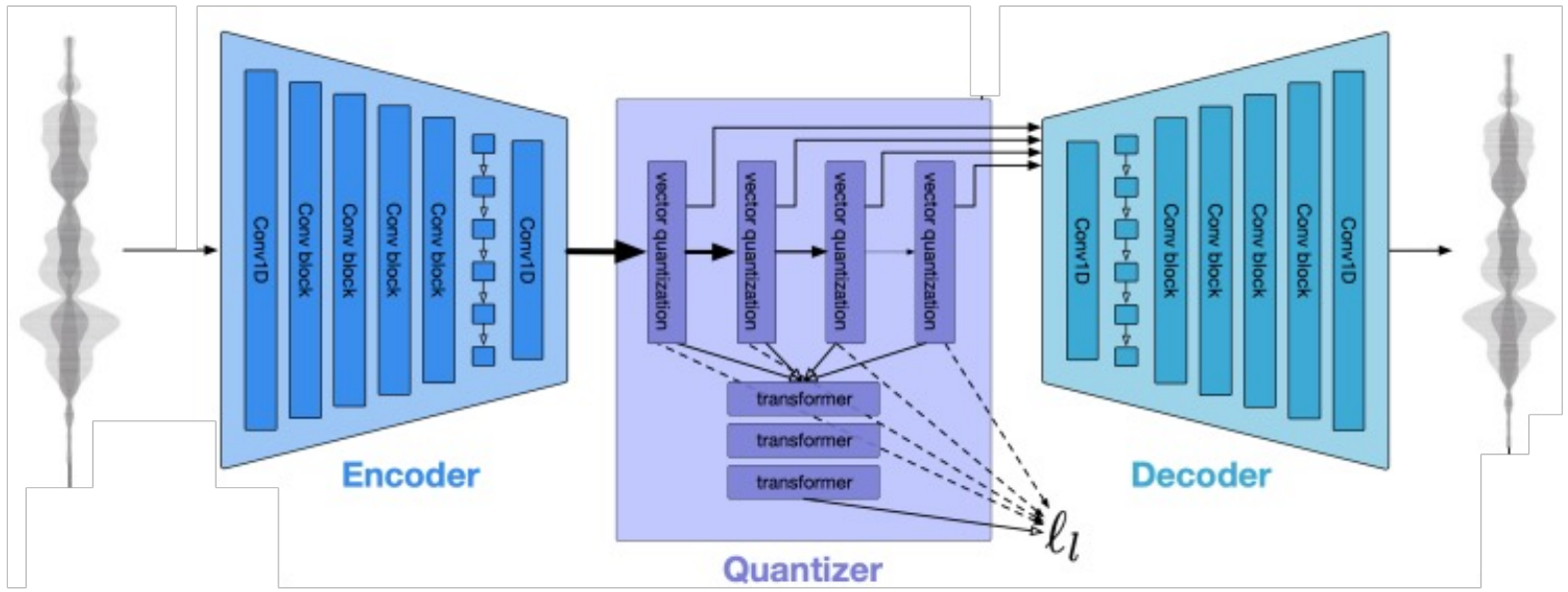
2 codebooks
MSE: $1.07 \cdot 10^{-4}$



3 codebooks
MSE: $5.23 \cdot 10^{-6}$

Transforming Music into Discrete Tokens – RVQ

Ground truth



Transforming Music into Discrete Tokens – RVQ

Ground truth

2 codebooks

MSE: 37.0

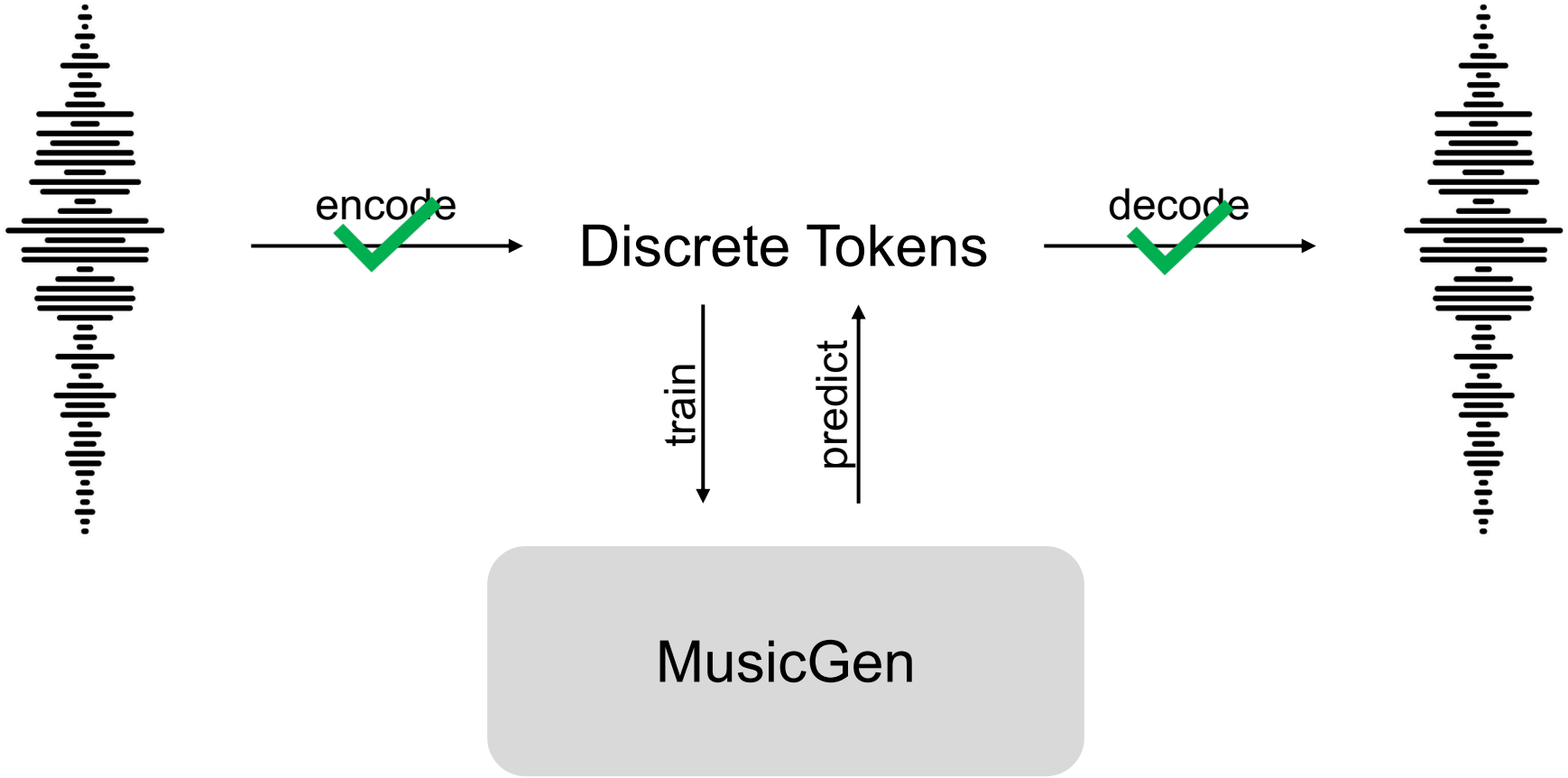
4 codebooks

MSE: 27.5

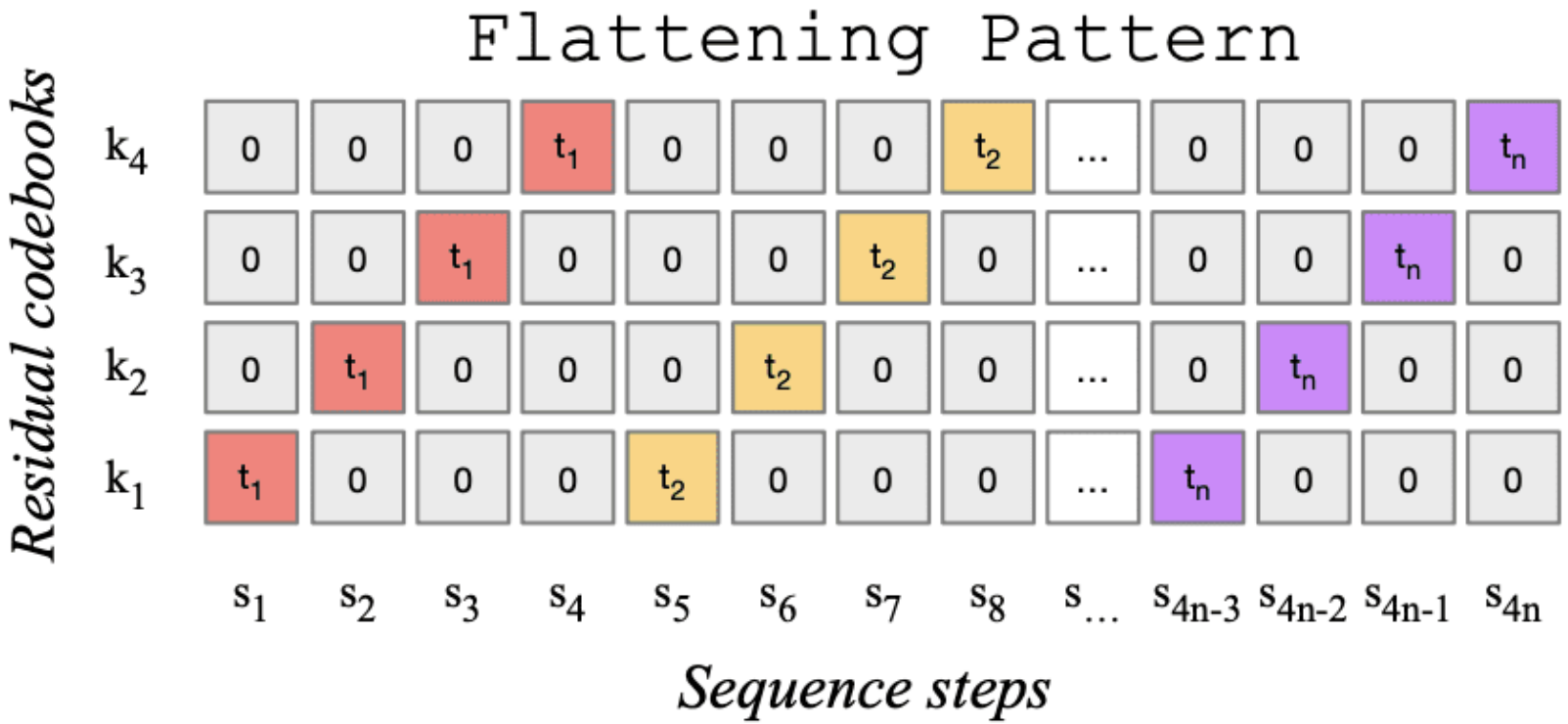
8 codebooks

MSE: 20.8

MusicGen – Overview

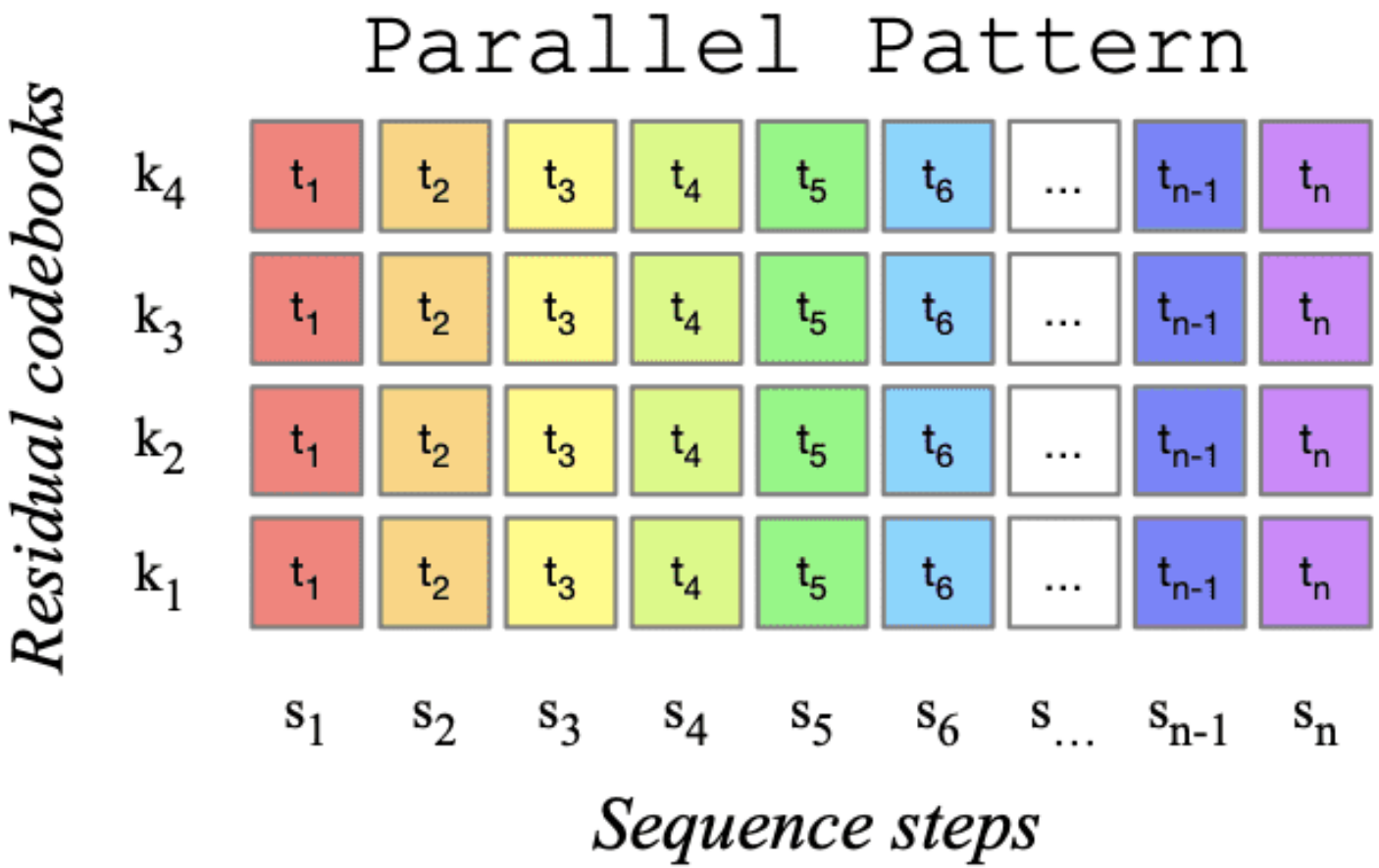


Codebook Patterns



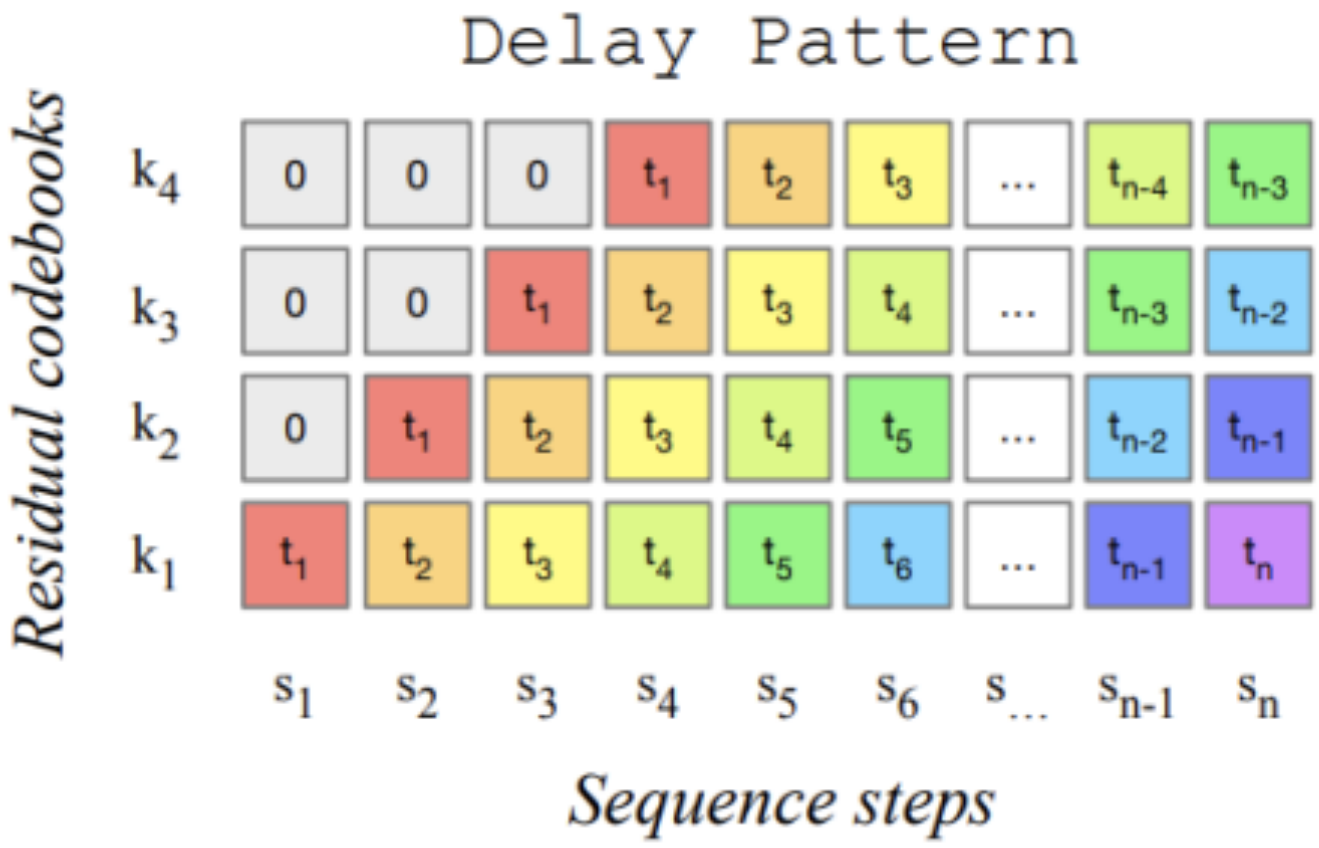
Copet et al. [7]

Codebook Patterns



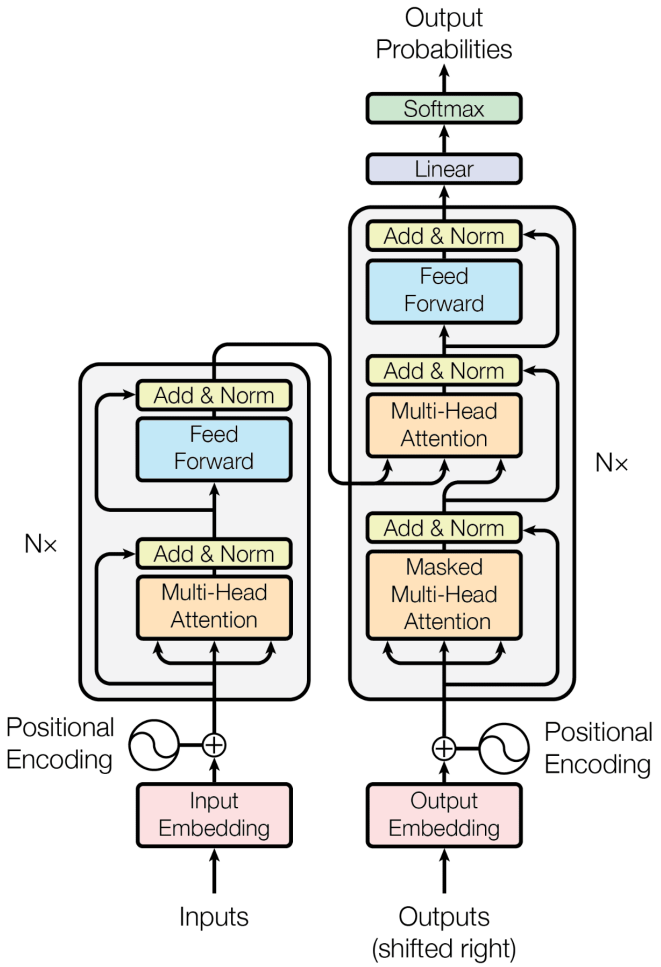
Copet et al. [7]

Codebook Patterns



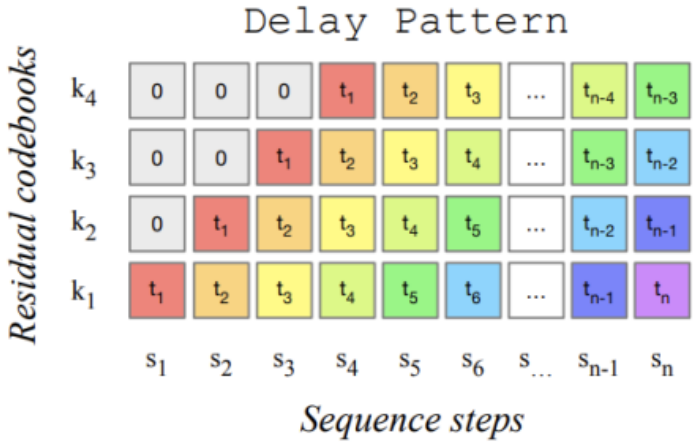
Copet et al. [7]

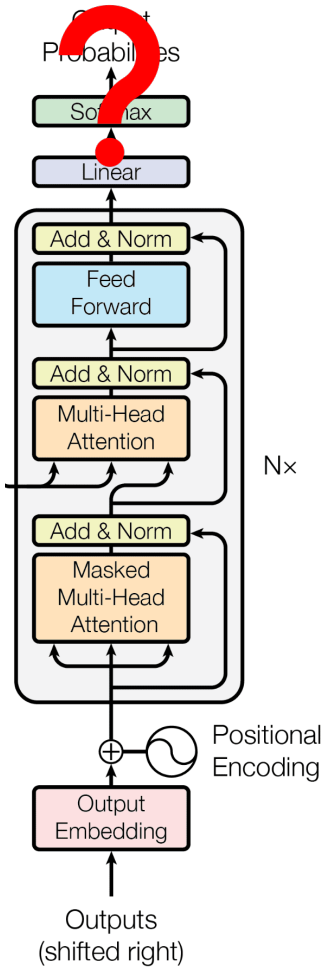
Model Architecture



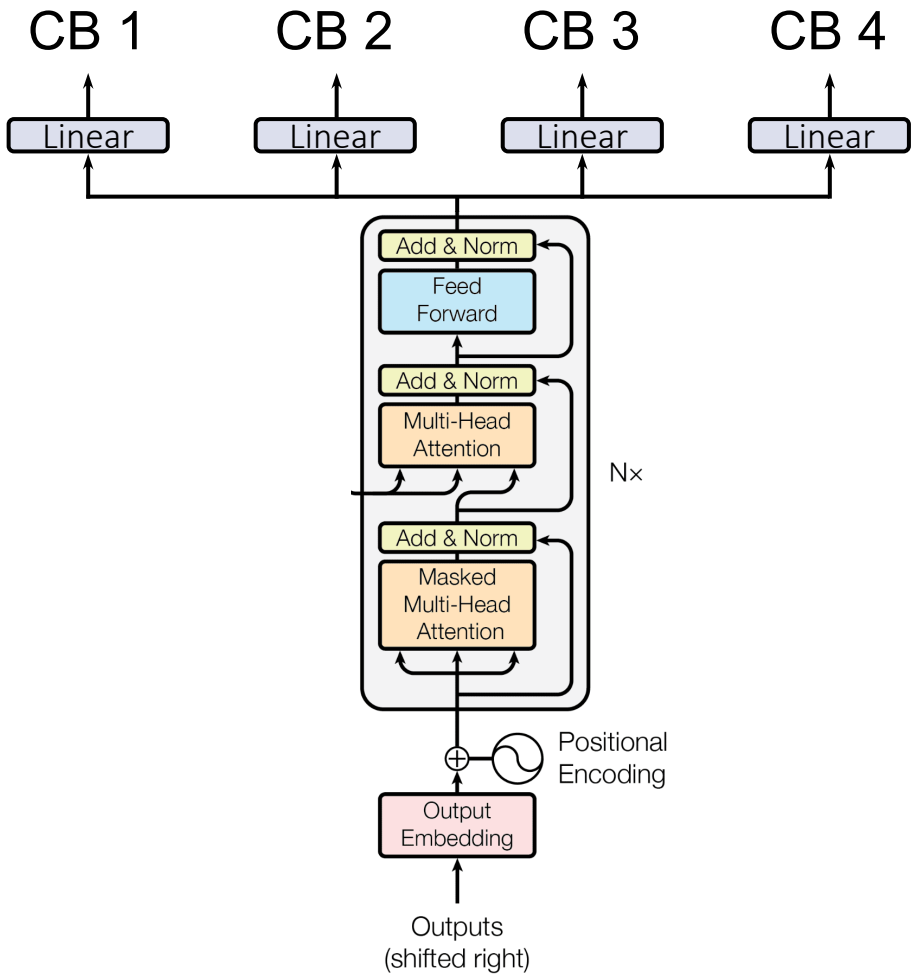
Vaswani et al. [8]

Model Architecture

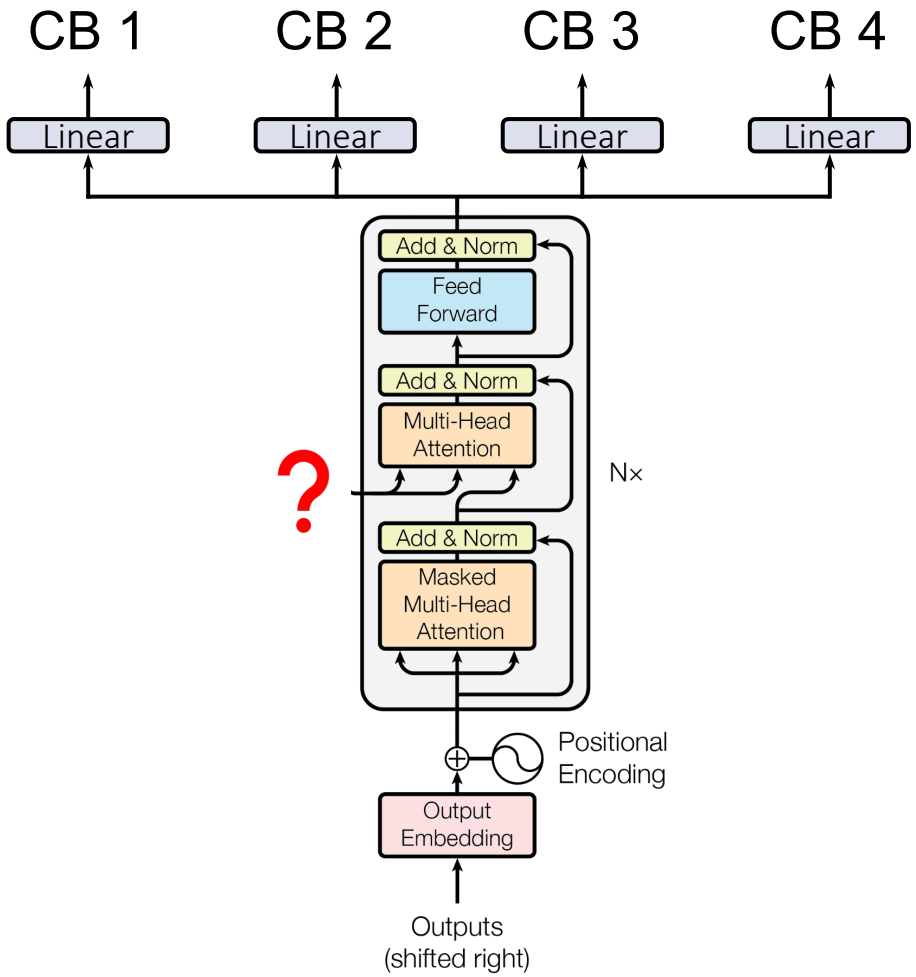




Model Architecture

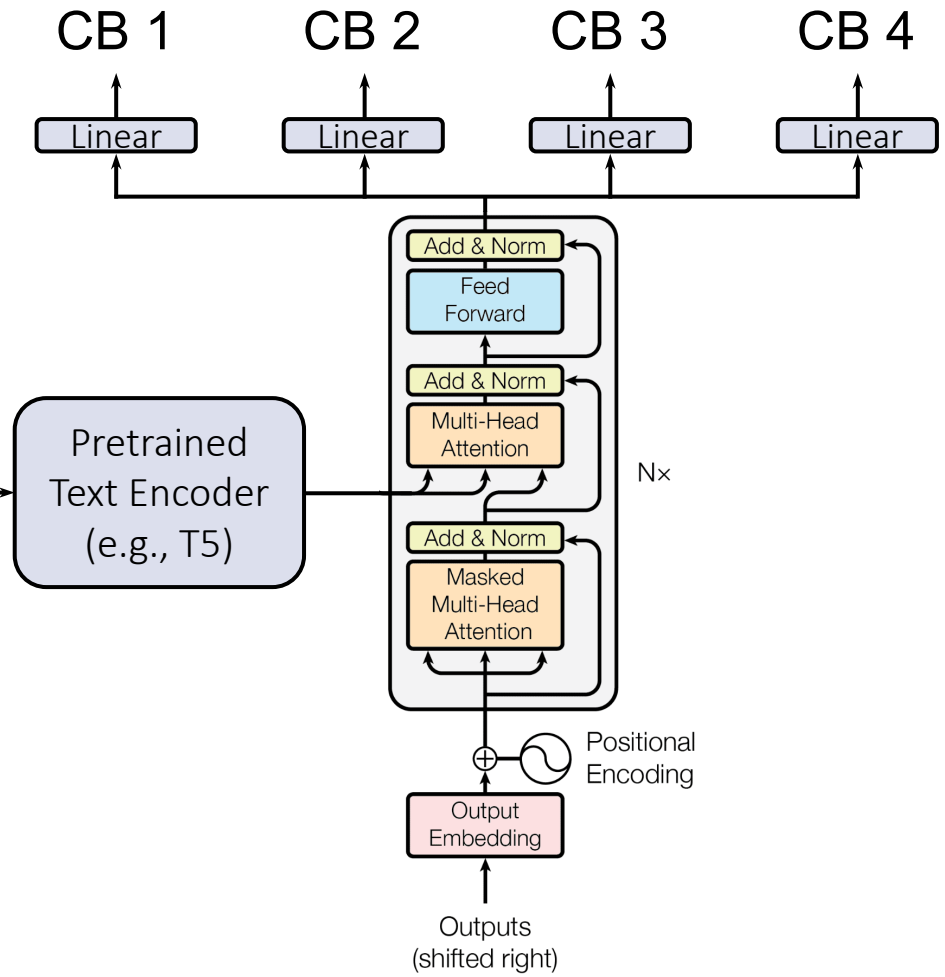


Model Architecture

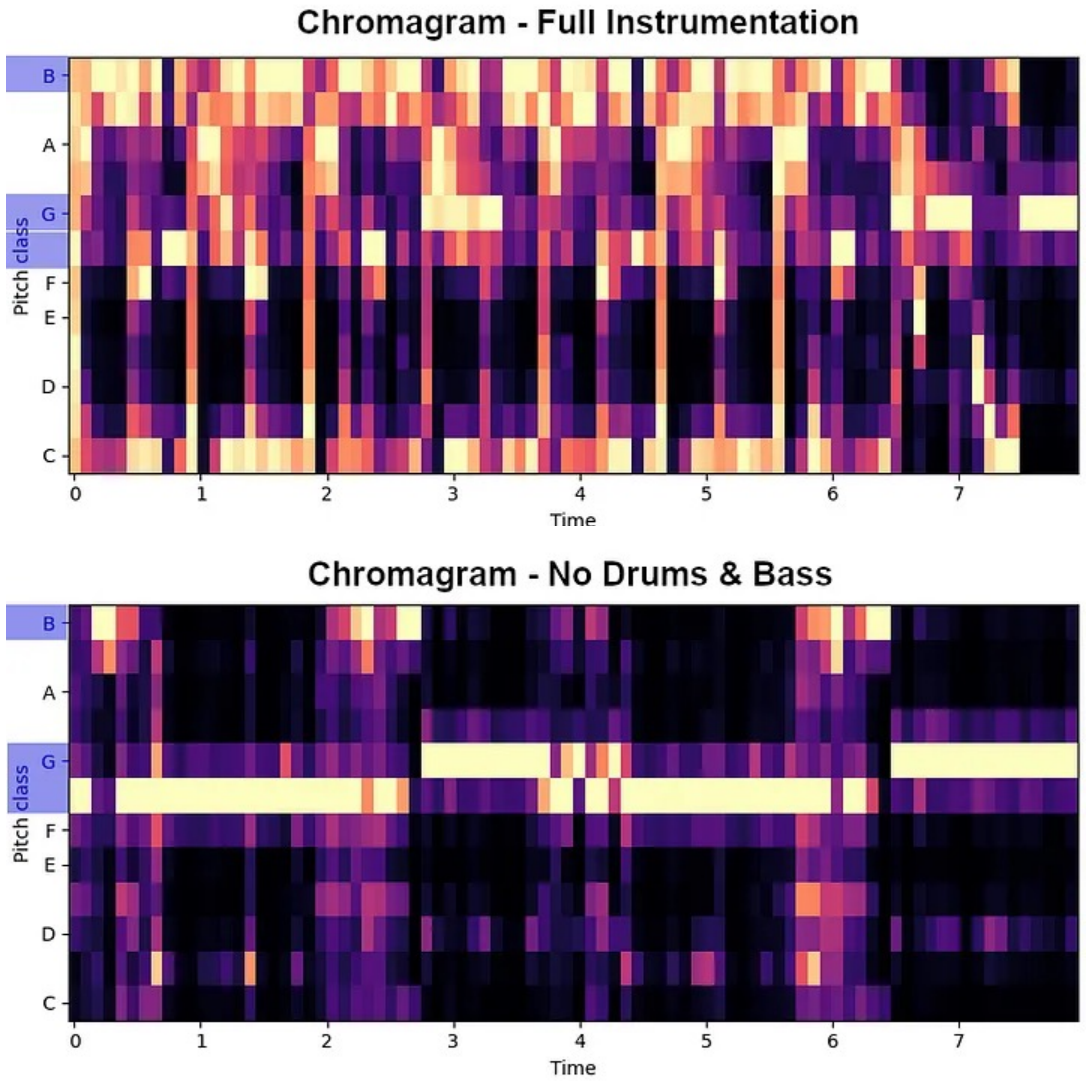


Text Conditioning

«Classic reggae track with an electronic guitar solo»

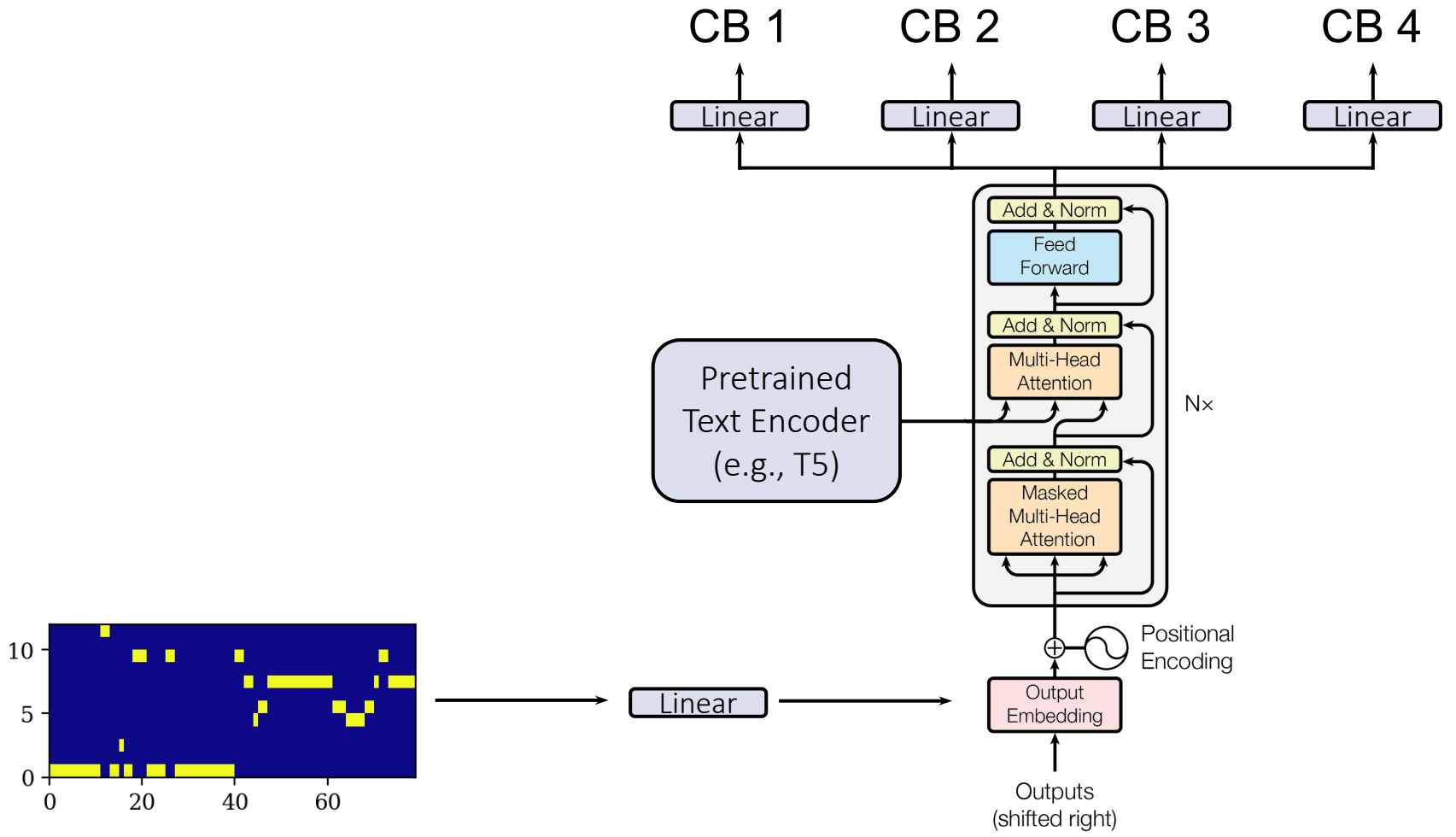


Melody Conditioning



TowardsDataScience Blogpost [9]

Melody Conditioning

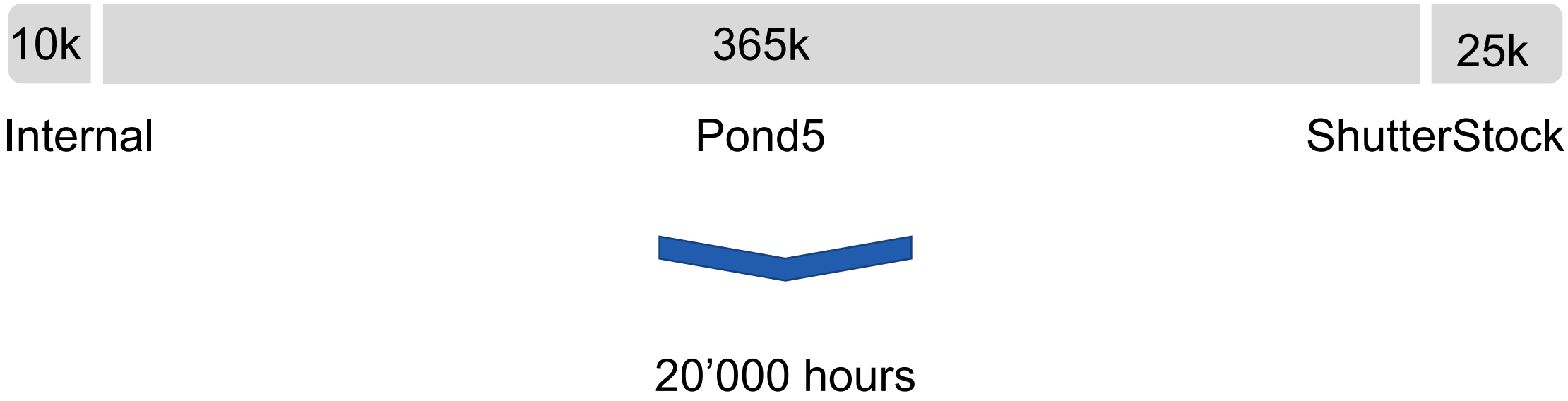


Copet et al. [7]

Melody & Text Conditioning

«An 80s driving pop song with heavy drums and synth pads in the background»

Training Data



→ MusicCaps for evaluation

Evaluation

Objective Metrics

- FAD
- KL divergence
- CLAP

Subjective Metrics

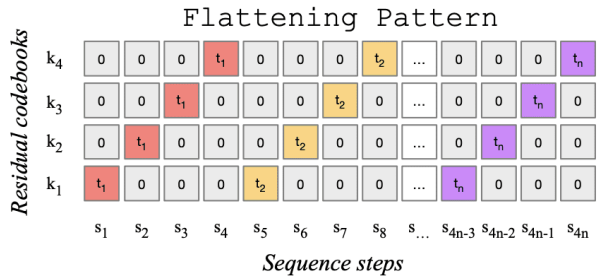
- Overall (1-100)
- Relevance (1-100)

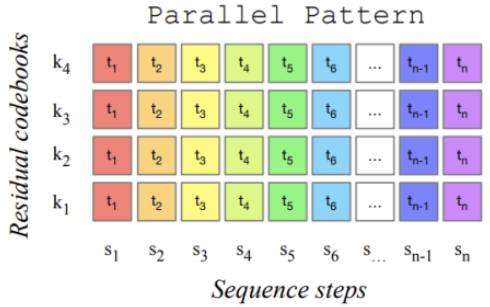
Results – Other models

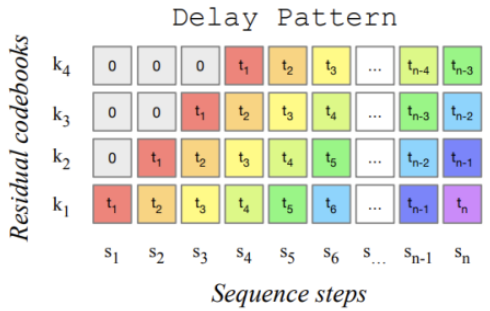
	FAD↓	KL↓	CLAP↑	OVL↑	REL↑
Riffusion	14.8	2.06	0.19	79.31 (±1.37)	74.20 (±2.17)
Moûsai	7.5	1.59	0.23	76.11 (±1.56)	77.35 (±1.72)
MusicLM	4.0	-	-	80.51 (±1.07)	82.35 (±1.36)
MusicGen (300M)	**3.1**	1.28	0.31	78.43 (±1.30)	81.11 (±1.31)
MusicGen (1.5B)	3.4	1.23	**0.32**	80.74 (±1.17)	**83.70** (±1.21)
MusicGen (3.3B)	3.8	**1.22**	0.31	**84.81** (±0.95)	82.47 (±1.25)
MusicGen w. random melody (1.5B)	5.0	1.31	0.28	81.30 (±1.29)	81.98 (±1.79)

Copet et al. [7]

Results – Patterns







	FAD↓	KL↓	CLAP↑	OVL↑	REL↑
Delay	0.96	0.52	0.35	**79.69** (±1.46)	79.67 (±1.41)
Parallel	2.58	0.62	0.27	72.21 (±2.49)	80.30 (±1.43)
Flattening	**0.86**	**0.51**	**0.37**	**79.71** (±1.58)	**82.03** (±1.10)

Copet et al. [7]

Results – Audio Samples

«A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a heroic battle.»

Riffusion

Moûsai

MusicLM

Stereo MusicGen

Weaknesses

No long-term structures

Melody representation limited

Lyrics

Copyright infringement risk

Questions & Discussion

Sources

[1]: https://github.com/facebookresearch/audiocraft/blob/main/dataset/example/electro_1.mp3

[2]: https://ai.honu.io/papers/musicgen/

[3]: Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[4]: S Forsgren and H Martiros. Riffusion-stable diffusion for real-time music generation. 2022. URL: https://riffusion.com/about.

[5]: Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion. arXiv preprint arXiv:2301.11757, 2023.

[6]: Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022.

[7]: J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[8]: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, edi- tors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[9]: https://towardsdatascience.com/how-metas-ai-generates-music-based-on-a-reference-melody-de34acd783

[10]: Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., ... & Plumbley, M. D. (2023). Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

[11]: https://www.stableaudio.com

[12]: https://www.suno.ai

[13]: Lam, M. W., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., ... & Wang, Y. (2024). Efficient neural music generation. *Advances in Neural Information Processing Systems*, *36*.

[14]: https://drscotthawley.github.io/blog/posts/2023-06-12-RVQ.html

ETH zürich

Dennis Jüni
djueni@student.ethz.ch

Seminar in Deep Neural Networks (FS 2023)