Durham University

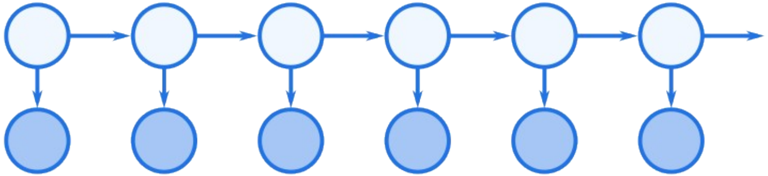
Introduction to Music Processing

Neural Models

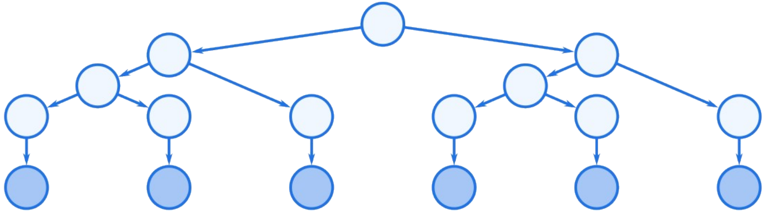
Dr Robert Lieck
robert.lieck@durham.ac.uk

Computational Models of Music

- **Sequential Models** (*n*-gram and (hidden) Markov)



- **Hierarchical Models** (context-free grammars)



- **Neural Networks** (RNNs, Transformers, WaveNet)

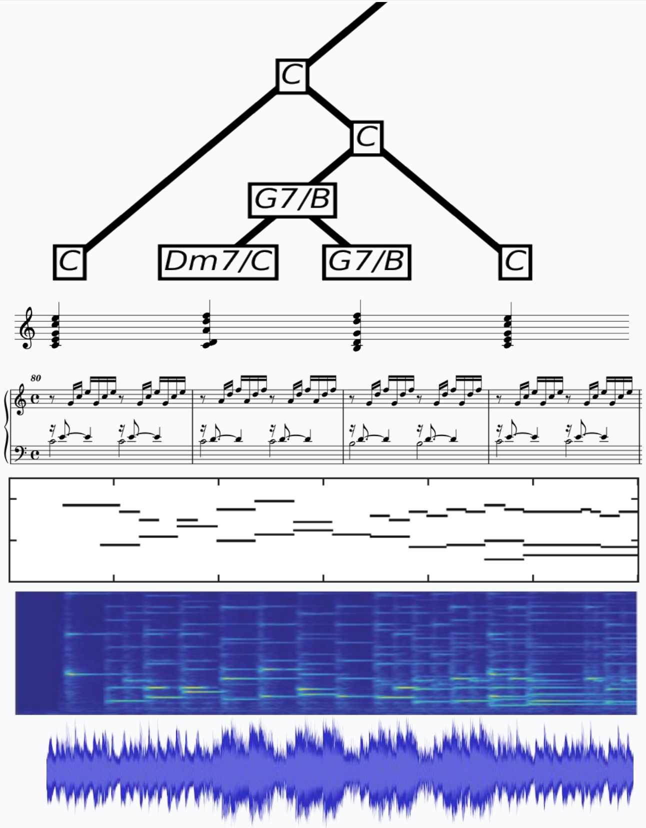
Durham University

Neural Networks

for Music Analysis and Generation

Music Representations

Analysis



Structure Analysis

Chords/Harmonies

Score

Pianoroll/MIDI

Spectrogram

Waveform/Audio

Benetos et al. (2019) / Lieck & Rohrmeier (2021)

Generation

Durham University

Combining Multiple Architectures & Representations

- **Source representation(s):** The representation(s) of the input data
- **Target representation:** The representation of the desired output
- **Intermediate/Auxiliary representation(s):** Any representation(s) used in intermediate or additional processing steps

Examples

- **Raw audio generation:** (audio) → (MIDI / tokens) → audio
- **Transcription:** audio → (spectrogram) → MIDI → score → (audio)
- **Chord analysis:** score → chords
- **Metrical analysis:** audio → onsets → beat → metre

Note: Encoder-decoder architectures combine analysis and generation!

Generative Audio Models ("best of big tech")

These models combine multiple architectures & representations:

- MusicGen (Meta): https://ai.honu.io/papers/musicgen/
  Copet J, Kreuk F, Gat I, et al (2024) Simple and controllable music generation. In: Advances in Neural Information Processing Systems
- MusicLM (Google): https://google-research.github.io/seanet/musiclm/examples/
  Agostinelli A, Denk TI, Borsos Z, et al (2023) MusicLM: Generating Music From Text http://arxiv.org/abs/2301.11325
- Jukebox (OpenAI): https://openai.com/research/jukebox | https://jukebox.openai.com/
  Dhariwal P, Jun H, Payne C, et al (2020) Jukebox: A Generative Model for Music http://arxiv.org/abs/2005.00341
- MuseNet (OpenAI): https://openai.com/research/musenet
  (2019)
- Music Transformer (Google): https://magenta.tensorflow.org/music-transformer
  Huang C-ZA, Vaswani A, Uszkoreit J, et al (2018) Music Transformer: Generating Music with Long-Term Structure

Generative Audio Models (“best of big tech”)

Common Strengths…

- Getting the “feel” right (high-level surface features of genres etc.)
- Basic musical structure (e.g. isochronous beat, “keyness”, surface texture)
- Decent short-term *structure* (range of seconds)
- Decent long-term *coherence* (e.g. staying in style/key)

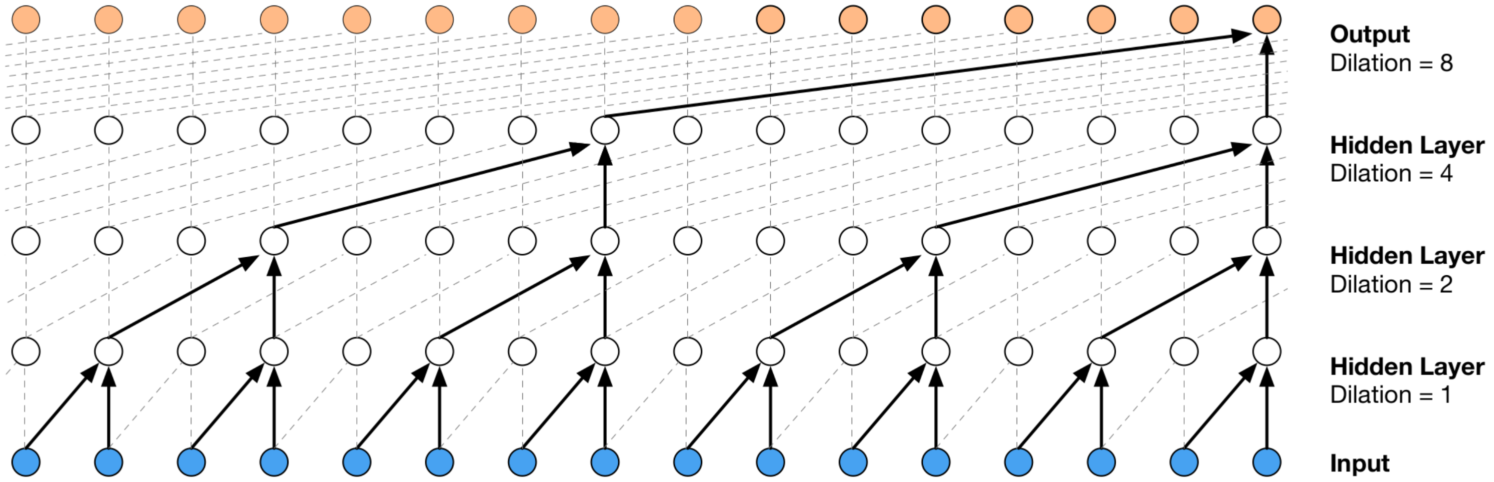
…and Shortcomings…

- Mediocre audio quality (esp. given the amount of training data)
  → This has changed in more recent models!
- Inaccurate details in all musical dimensions (timing, pitch, harmony, …)
- **No long-term structure** (range of minutes or more)

WaveNet: (conditional) → audio

Raw autoregressive audio generation (optionally conditional)

- Uses *dilated causal convolutional layers* (with residual and skip connections)
- μ-law companding algorithm to *discretise* to 256 prediction values (nowadays VQ approaches are commonly used)



Durham University

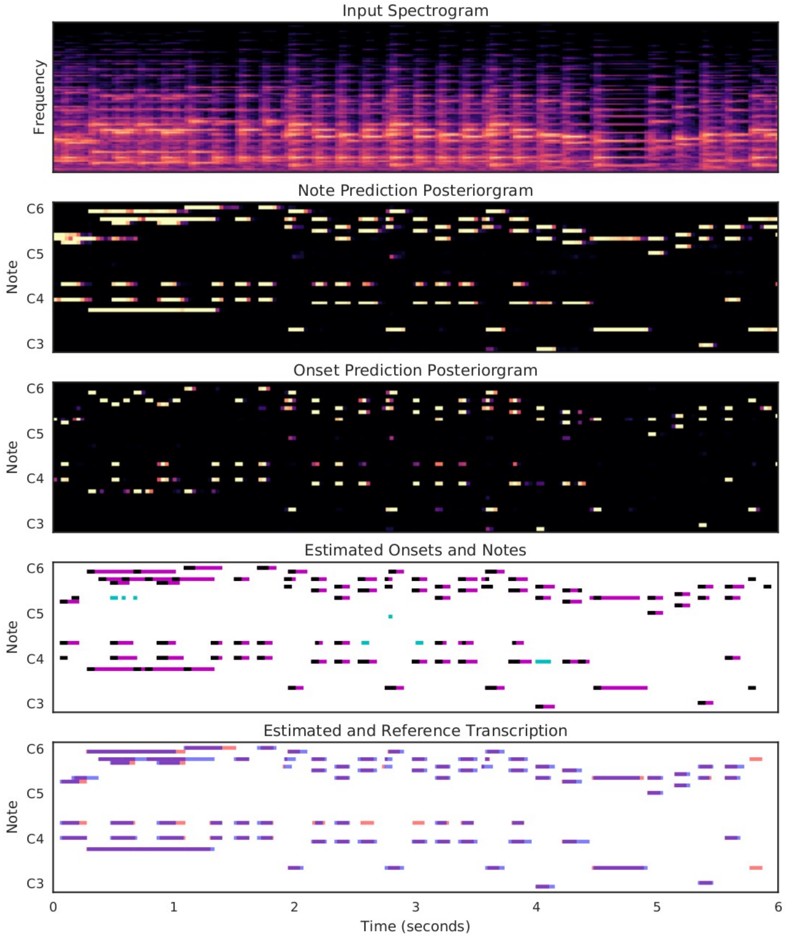
van den Oord A, Dieleman S, Zen H, et al (**2016**) WaveNet: A generative model for raw audio. arXiv preprint arXiv:160903499

Onset and Frames: audio → MIDI

Predict MIDI from raw audio

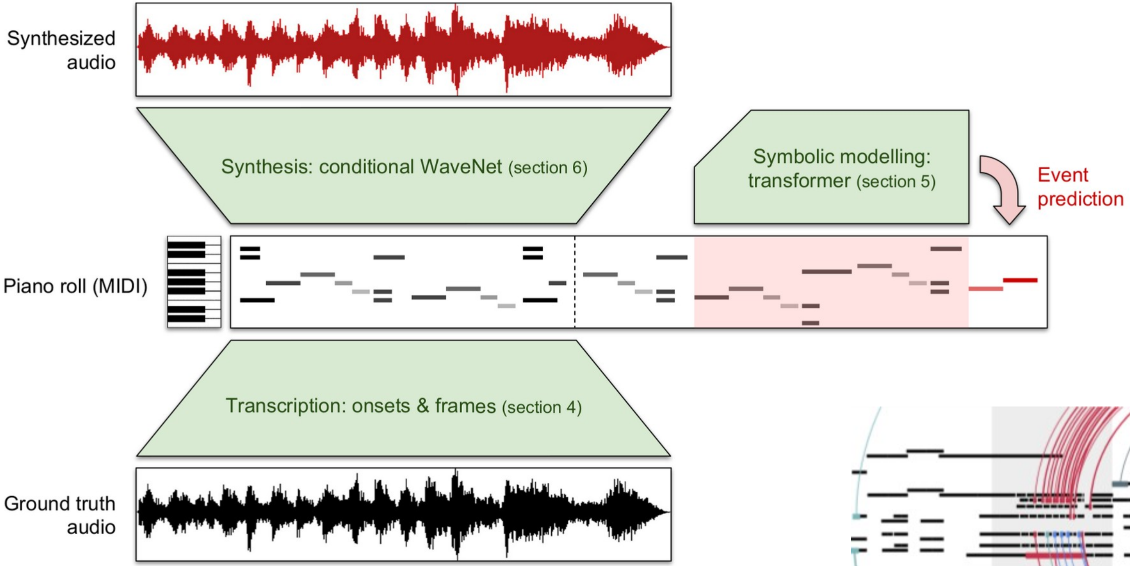
- Estimate notes (pitches)
- Estimate (note-wise) onsets
- Predict notes that also have an onset

Hawthorne C, Elsen E, Song J, et al (2018) Onsets and Frames: Dual-Objective Piano Transcription. In: Proceedings of the 19th International Society for Music Information Retrieval Conference

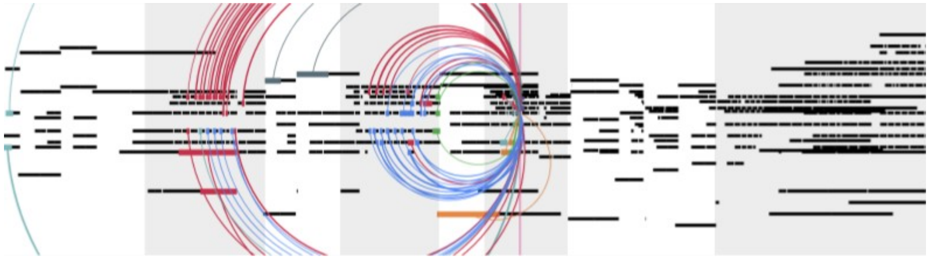


Music Transformer

- Autoregressive symbolic generation
- audio → symbolic → audio (Wave2Midi2Wave)



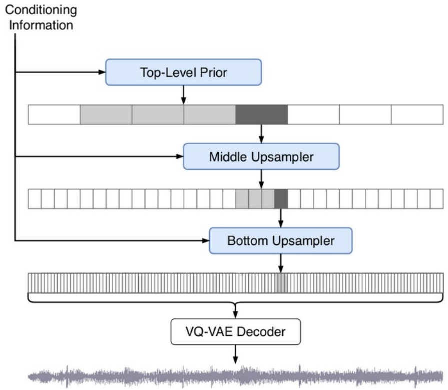
- Huang C-ZA, Vaswani A, Uszkoreit J, et al (2019) Music Transformer: Generating Music with Long-Term Structure. In: International Conference on Learning Representations
- Hawthorne C, Stasyuk A, Roberts A, et al (2019) Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: 7th international conference on learning representations, ICLR 2019, new orleans, LA, USA, may 6-9, 2019

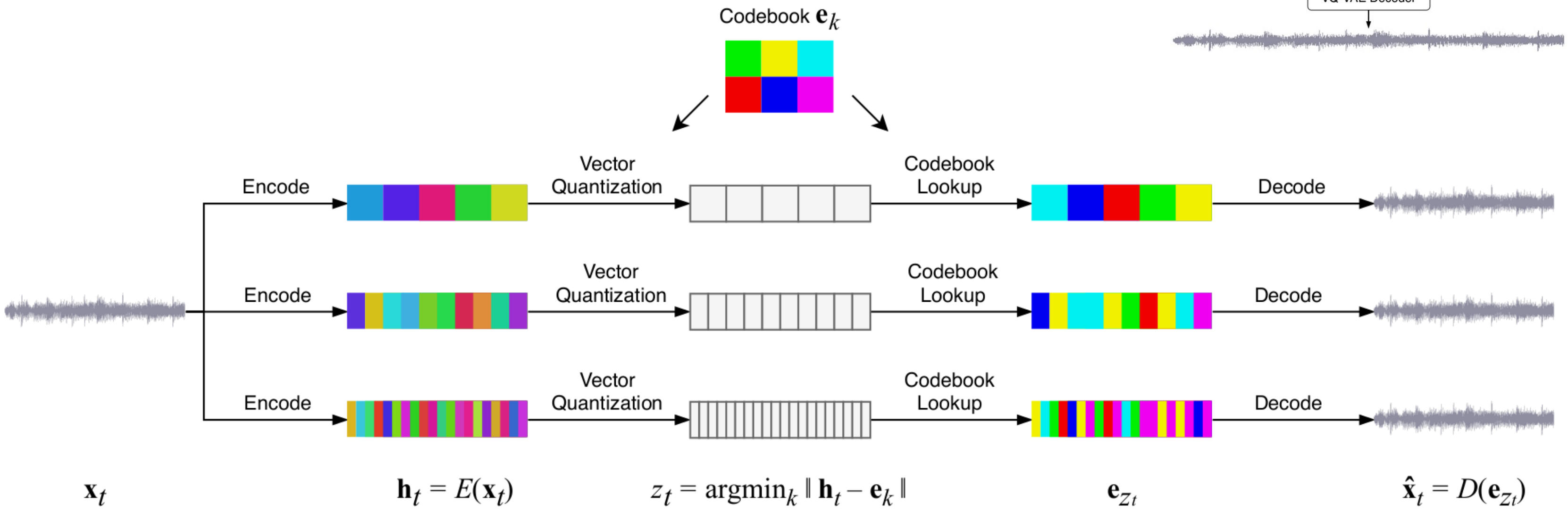


Durham University

Jukebox

- audio → discrete latent (VQ-VAE) → audio
- three separate time scales
- transformer in latent space





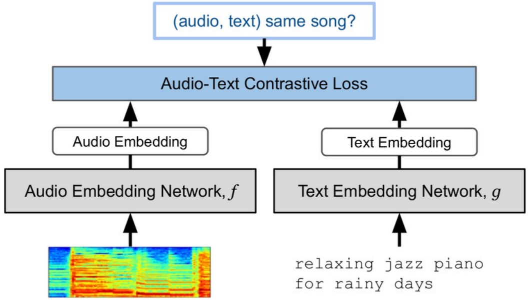
Durham University

Dhariwal P, Jun H, Payne C, et al (2020) Jukebox: A Generative Model for Music http://arxiv.org/abs/2005.00341

MusicLM

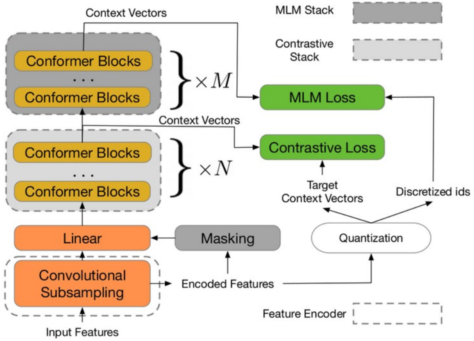
MuLan

- Joint audio-text embedding
- → relate text to audio



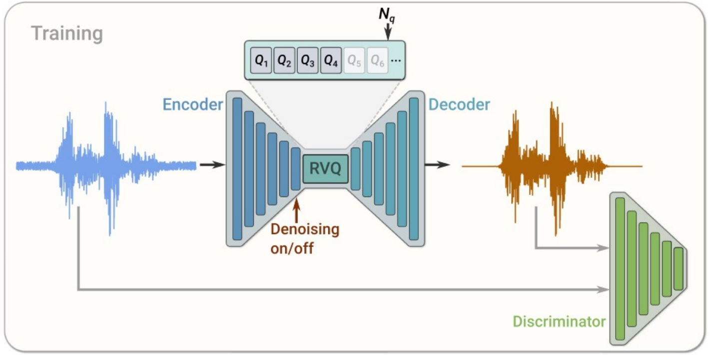
w2v-BERT

- Semantic audio tokens
- Masked contrastive loss
- → learn "meaning" of audio



SoundStream

- Discrete neural audio codec
- → capture waveform of audio

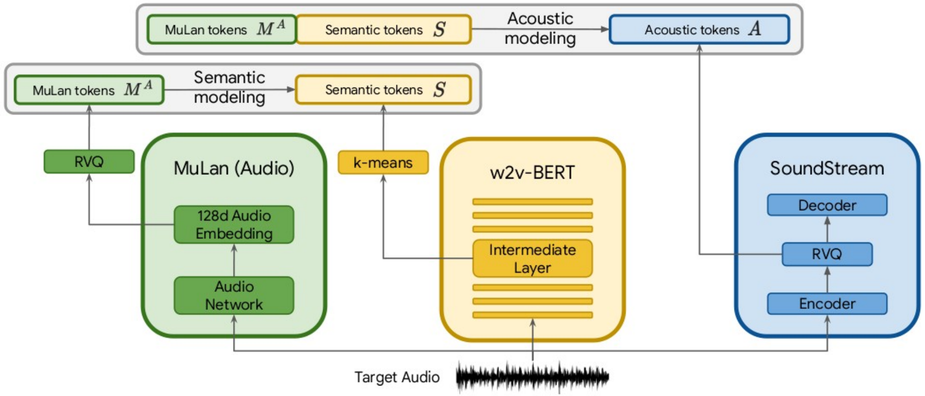


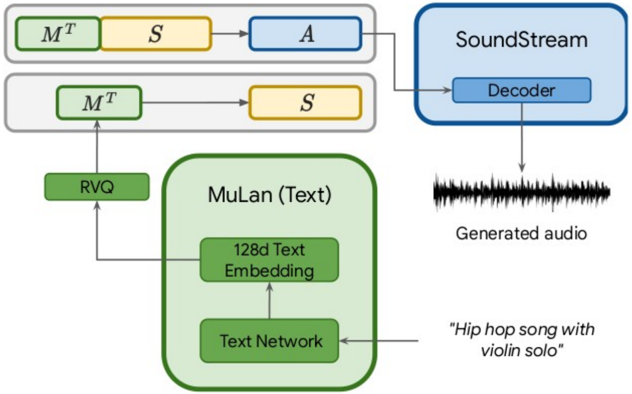
- Huang Q, Jansen A, Lee J, et al (2022) MuLan: A Joint Embedding of Music Audio and Natural Language. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference.
- Chung Y-A, Zhang Y, Han W, et al (2021) w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, Cartagena, Colombia, pp 244–250
- Zeghidour N, Luebs A, Omran A, et al (2022) SoundStream: An End-to-End Neural Audio Codec. IEEE/ACM Trans Audio Speech Lang Process 30:495–507. https://doi.org/10.1109/TASLP.2021.3129994

MusicLM

Combine three modules

1) Get embedding from audio (training) or text (inference)
2) Generate semantic audio tokens from embedding
3) Generate audio codes from embedding and tokens

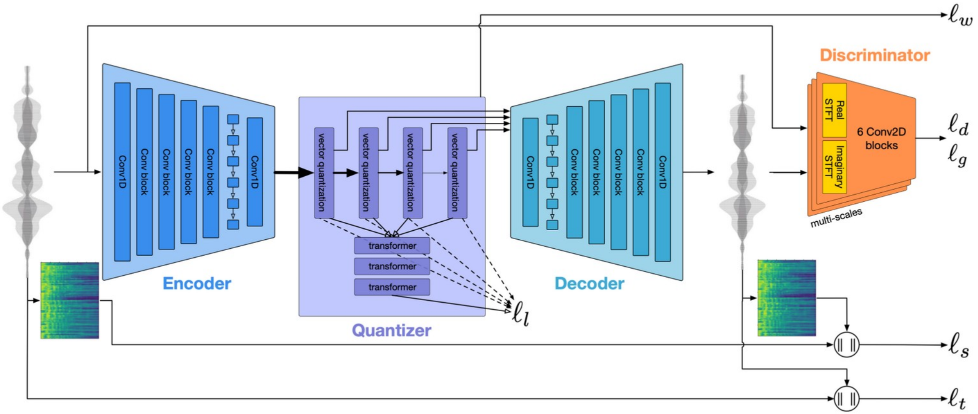


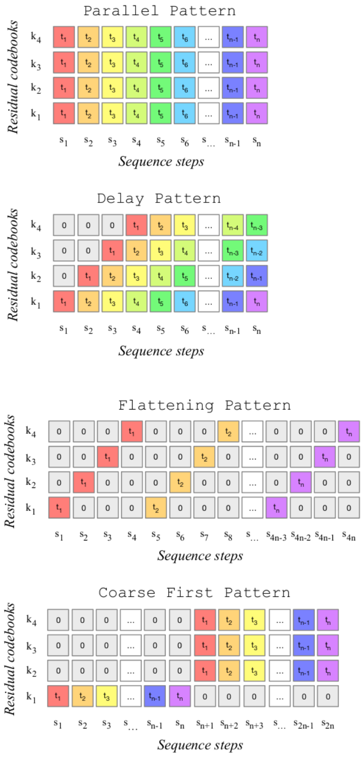


Agostinelli A, Denk TI, Borsos Z, et al (2023) MusicLM: Generating Music From Text http://arxiv.org/abs/2301.11325

MusicGen

- Encode to discrete tokens (multiple residual streams)
  - time and frequency (mel-spectrogram) reconstruction loss
  - discriminative STFT loss
- Model sequence of discrete tokens using transformer





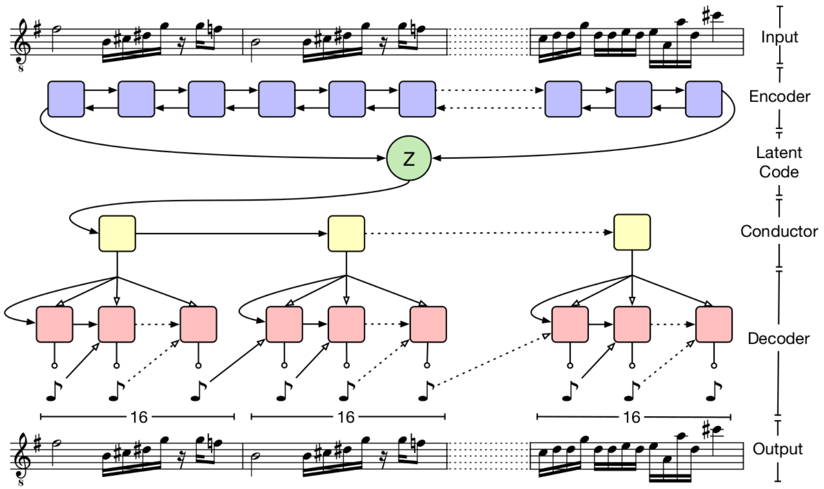
- Défossez A, Copet J, Synnaeve G, Adi Y (2023) High Fidelity Neural Audio Compression. Transactions on Machine Learning Research
- Copet J, Kreuk F, Gat I, et al (2024) Simple and controllable music generation. In: Advances in Neural Information Processing Systems

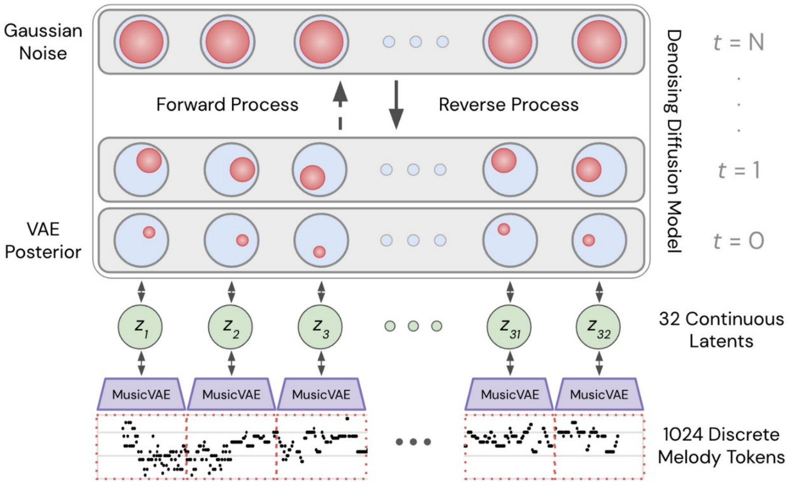
Durham University

Symbolic Generation with Diffusion Models

- MIDI → continuous latent (VAE) → MIDI
- continuous latent (diffusion) → MIDI

MusicVAE

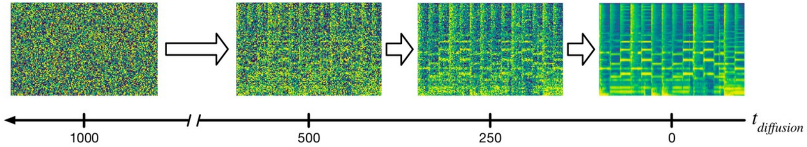


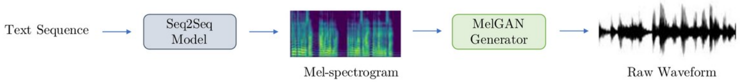


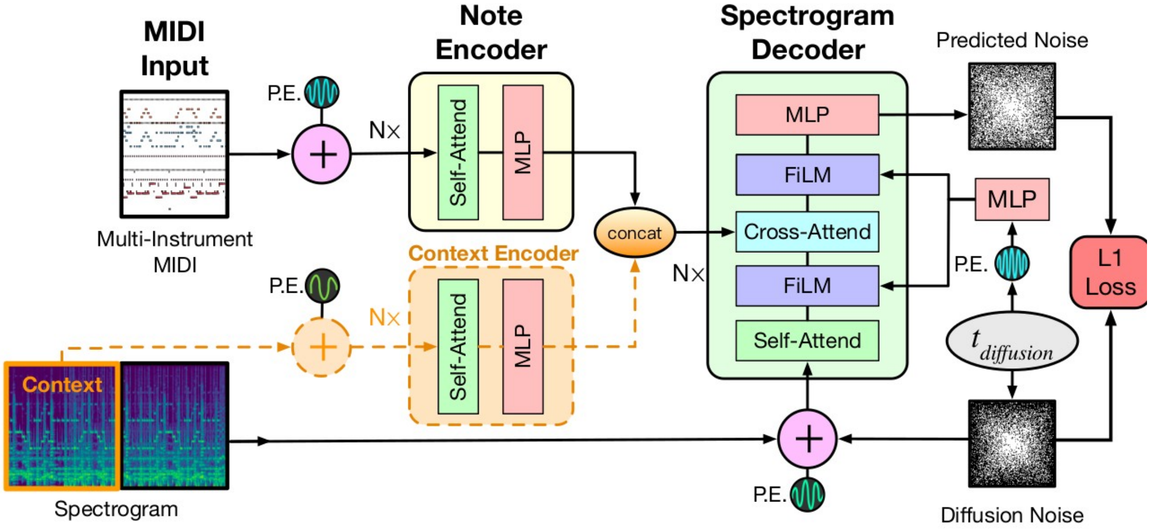
- Roberts A, Engel J, Raffel C, et al (2018) A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In: International Conference on Machine Learning. pp 4361–4370
- Mittal G, Engel J, Hawthorne C, Simon I (2021) Symbolic Music Generation with Diffusion Models. arXiv:210316091
- (Plasser M, Peter S, Widmer G (2023) Discrete diffusion probabilistic models for symbolic music generation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Macao, China, pp 5842–5850)

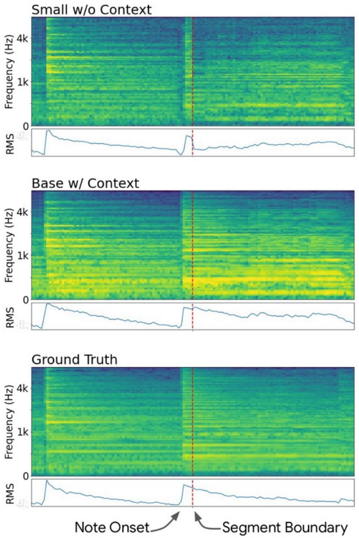
Durham
University

Spectrogram/Audio Generation with Diffusion Models









- Hawthorne C, Simon I, Roberts A, et al (2022) Multi-instrument Music Synthesis with Spectrogram Diffusion. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022
- Kumar K, Kumar R, de Boissiere T, et al (2019) Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems

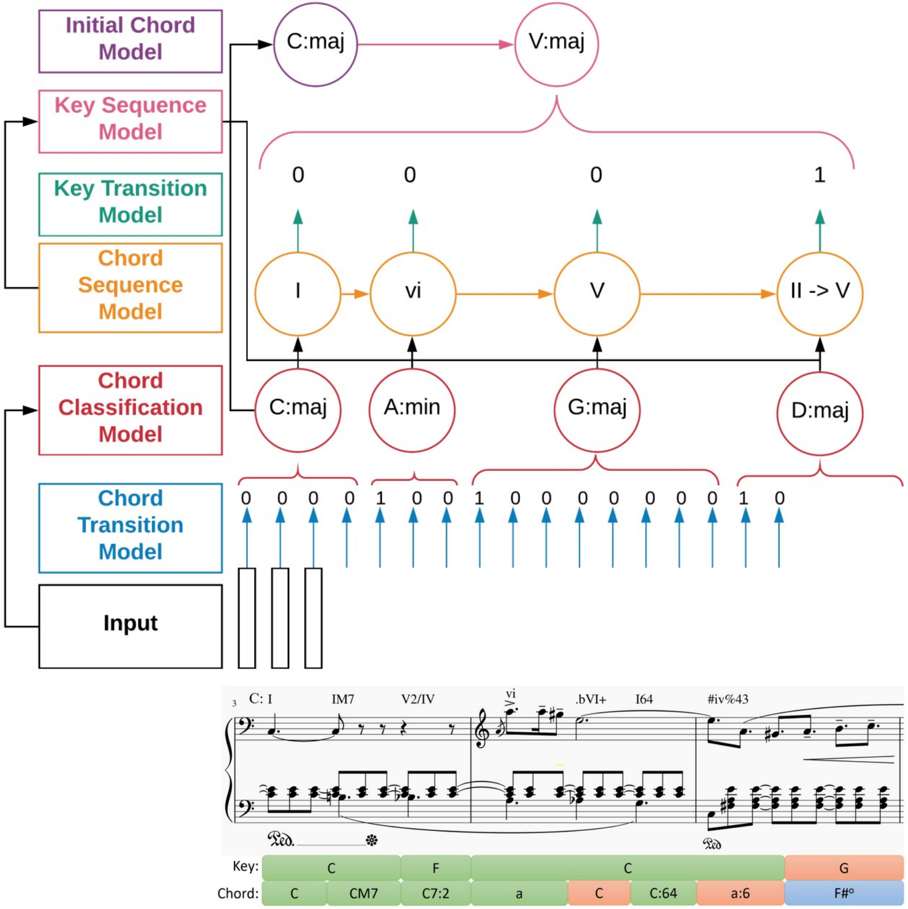
Harmonic Analysis

Vocabulary (1540 possible chords)

- 35 root pitches (A♭♭ – G♯♯), 70 keys (major/minor)
- 12 chord types
  - major, minor, augmented (triad, major 7th, minor 7th)
  - diminished (triad, minor 7th, diminished 7th)
- 3 or 4 inversions (triads/tetrads)

Architecture

- Modular with well-defined outputs
- Note features as one-hot vectors
- Bi-LSTM for sequence modules
- Using beam search for inference



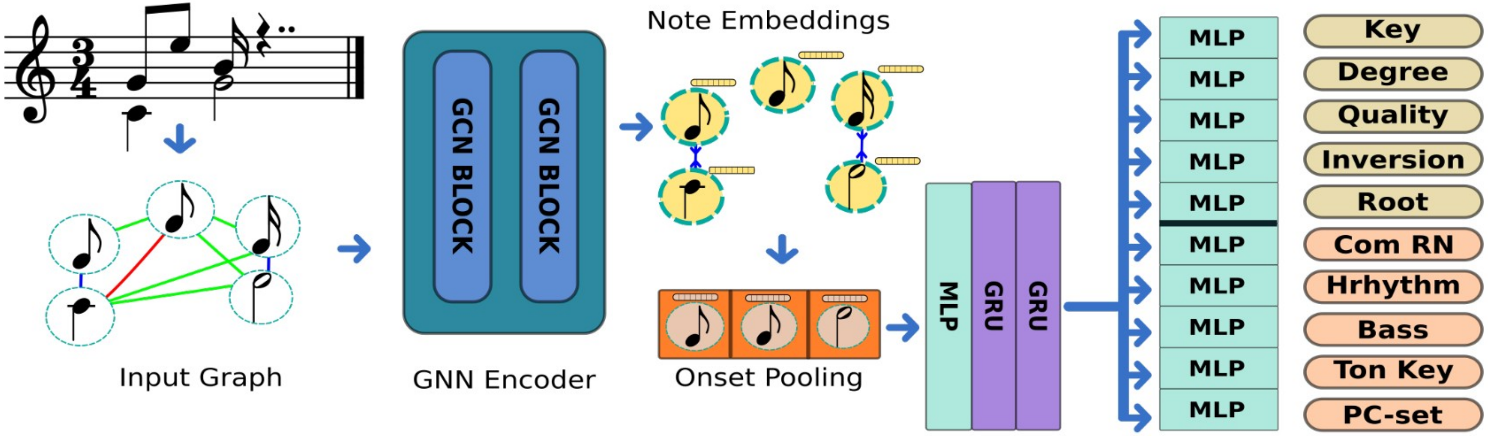
McLeod AP, Rohrmeier MA (2021) A modular system for the harmonic analysis of musical scores using a large vocabulary. In: International Society for Music Information Retrieval Conference. pp 435–442

Roman Numeral Analysis with Graph Neural Networks

- Explicitly represent structure in musical score
- Use graph convolutional neural network to compute embeddings
- Predict features of Roman numeral symbols independently

$$\boxed{\mathrm{I}^{(6)-5}_{(4)-3}/\mathrm{V}}$$

- composition: chord type, inversion, pitch changes
- root position: local key, tonicization



Karystinaios E, Widmer G (2023) Roman Numeral Analysis With Graph Neural Networks: Onset-Wise Predictions From Note-Wise Features. In: Proceedings of the 24th International Society for Music Information Retrieval Conference

References

1) van den Oord A, Dieleman S, Zen H, et al (2016) WaveNet: A generative model for raw audio. arXiv preprint arXiv:160903499

2) Roberts A, Engel J, Raffel C, et al (2018) A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In: International Conference on Machine Learning. pp 4361–4370

3) Hawthorne C, Elsen E, Song J, et al (2018) Onsets and Frames: Dual-Objective Piano Transcription. In: Proceedings of the 19th International Society for Music Information Retrieval Conference

4) Hawthorne C, Stasyuk A, Roberts A, et al (2019) Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: 7th international conference on learning representations, ICLR 2019, new orleans, LA, USA, may 6-9, 2019

5) Huang C-ZA, Vaswani A, Uszkoreit J, et al (2019) Music Transformer: Generating Music with Long-Term Structure. In: International Conference on Learning Representations

6) Kumar K, Kumar R, de Boissiere T, et al (2019) Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems

7) Dhariwal P, Jun H, Payne C, et al (2020) Jukebox: A Generative Model for Music http://arxiv.org/abs/2005.00341

8) Mittal G, Engel J, Hawthorne C, Simon I (2021) Symbolic Music Generation with Diffusion Models. arXiv:210316091

9) McLeod AP, Rohrmeier MA (2021) A modular system for the harmonic analysis of musical scores using a large vocabulary. In: International Society for Music Information Retrieval Conference. pp 435–442

10) Chung Y-A, Zhang Y, Han W, et al (2021) w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, Cartagena, Colombia, pp 244–250

References

11) Civit M, Civit-Masot J, Cuadrado F, Escalona MJ (2022) A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. Expert Systems with Applications 209:118190. https://doi.org/10.1016/j.eswa.2022.118190

12) Zeghidour N, Luebs A, Omran A, et al (2022) SoundStream: An End-to-End Neural Audio Codec. IEEE/ACM Trans Audio Speech Lang Process 30:495–507. https://doi.org/10.1109/TASLP.2021.3129994

13) Huang Q, Jansen A, Lee J, et al (2022) MuLan: A Joint Embedding of Music Audio and Natural Language. In: Rao P, Murthy HA, Srinivasamurthy A, et al (eds) Proceedings of the 23rd International Society for Music Information Retrieval Conference. Bengaluru, India, pp 559–566

14) Hawthorne C, Simon I, Roberts A, et al (2022) Multi-instrument Music Synthesis with Spectrogram Diffusion. In: Rao P, Murthy HA, Srinivasamurthy A, et al (eds) Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022. pp 598–607

15) Karystinaios E, Widmer G (2023) Roman Numeral Analysis With Graph Neural Networks: Onset-Wise Predictions From Note-Wise Features. In: Sarti A, Antonacci F, Sandler M, et al (eds) Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023. pp 597–604

16) Plasser M, Peter S, Widmer G (2023) Discrete diffusion probabilistic models for symbolic music generation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Macao, China, pp 5842–5850

17) Agostinelli A, Denk TI, Borsos Z, et al (2023) MusicLM: Generating Music From Text http://arxiv.org/abs/2301.11325

18) Défossez A, Copet J, Synnaeve G, Adi Y (2023) High Fidelity Neural Audio Compression. Transactions on Machine Learning Research

19) Copet J, Kreuk F, Gat I, et al (2024) Simple and controllable music generation. In: Advances in Neural Information Processing Systems