MUSIKA!
FAST INFINITE WAVEFORM MUSIC GENERATION

Marco Pasini and Jan Schlüter

Department of Computational Perception

JOHANNES KEPLER UNIVERSITY LINZ

Department of Computational Perception

A comprehensive collection of generated samples can be found on **marcoppasini.github.io/musika**

State-of-the-art waveform music generation systems are **slow**. We introduce **Musika**, which allows for much **faster than real-time generation of music of arbitrary length** on a consumer CPU.

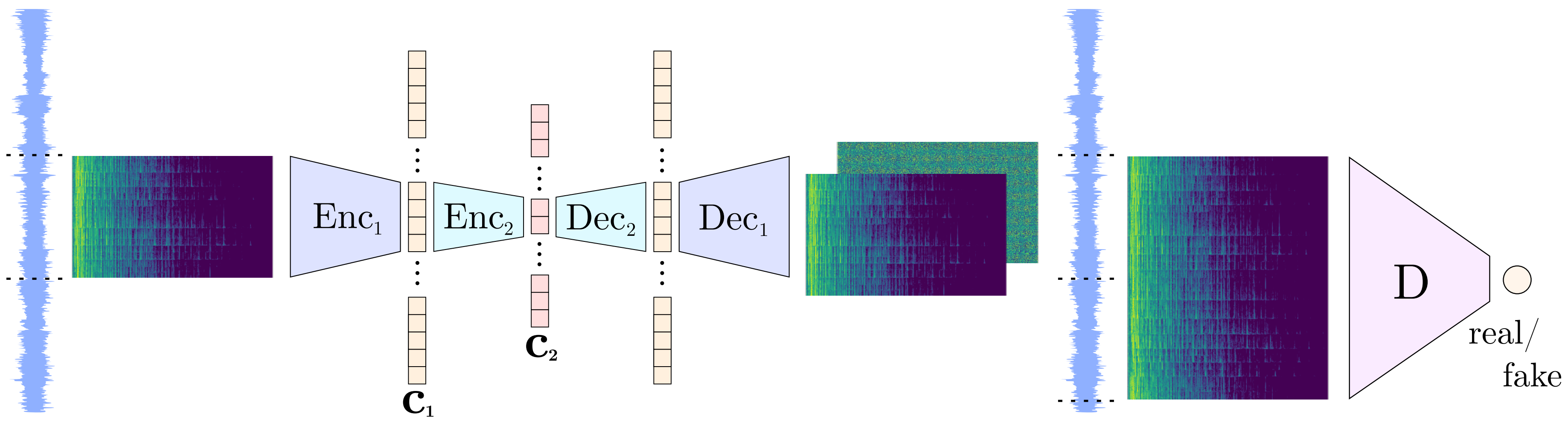

Fig. 1: **The proposed 2-level audio autoencoder**. A log-magnitude spectrogram is used as input, while the **decoder outputs magnitude and phase spectrograms** which are then inverted with iSTFT. A discriminator evaluates the magnitude spectrogram of two adjacent excerpts.

We train a **GAN** on compressed **invertible representations** from an efficient **Autoencoder**. **Coordinate** and **style vectors** allow parallel generation of arbitrarily **long** and **coherent** samples.

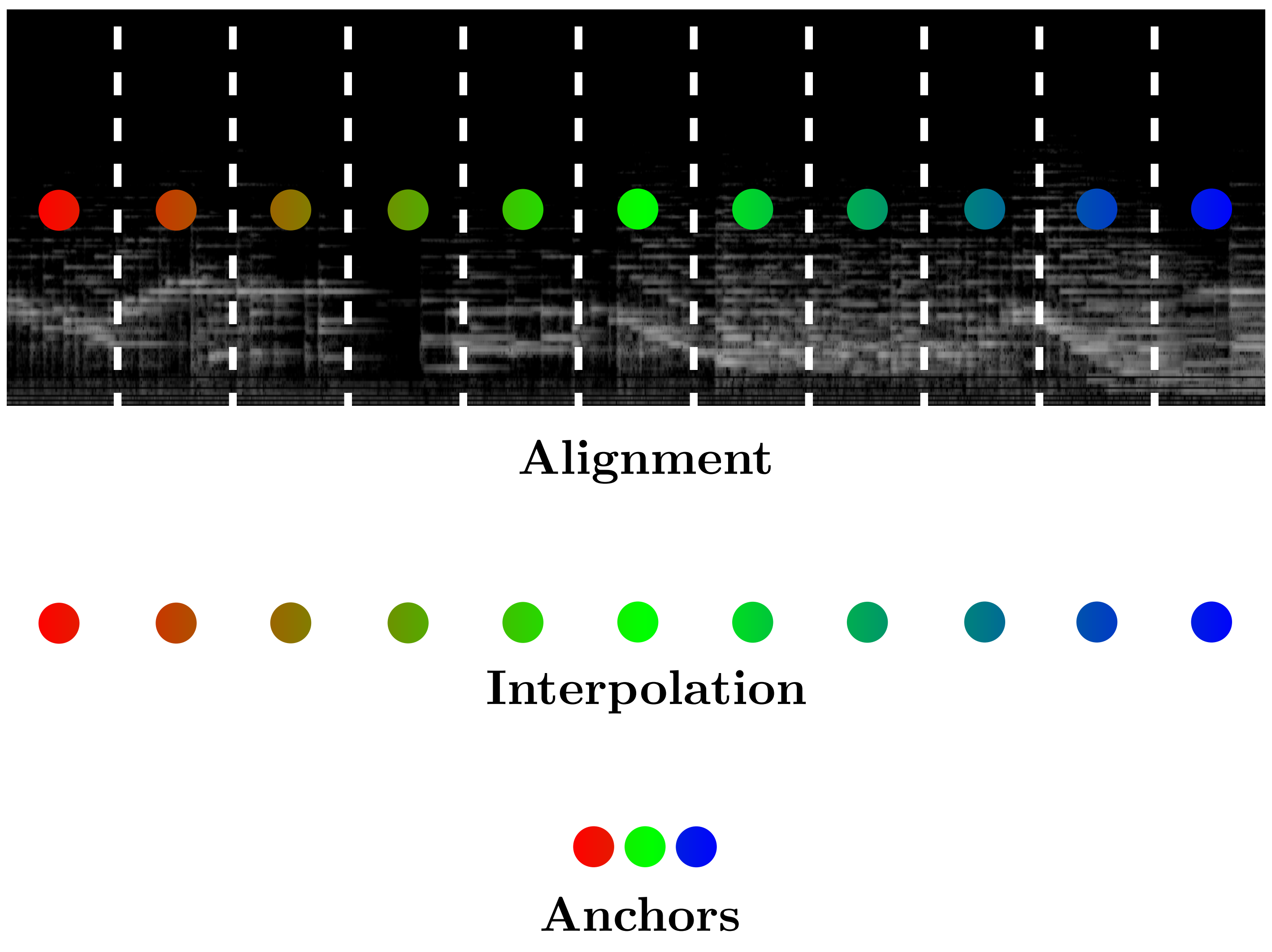

Fig. 2: **The alignment process**. We first **sample** random **anchor vectors** and linearly **interpolate between them**. The adversarial learning process **aligns the latent coordinate space with the generated latent vectors space**.

We first train the **Autoencoder** to produce **general representations**. We then train **GANs** to generate **Piano** and **Techno** music, with **note density** and **tempo** as conditioning.

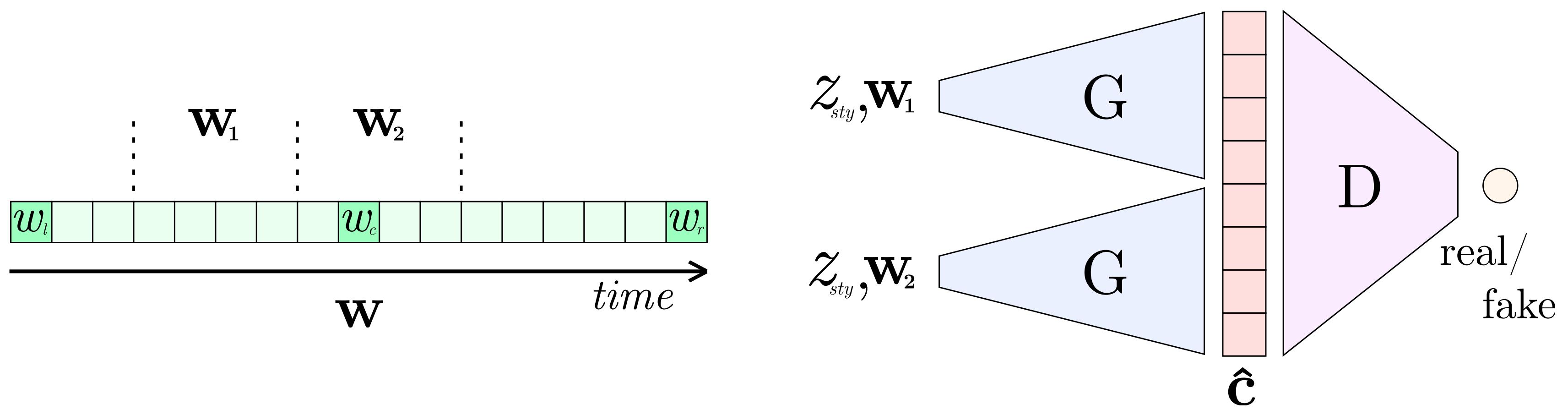

Fig. 3: **The proposed latent GAN training process**. Two adjacent latent coordinate sequences are randomly cropped from the linear interpolation between **three anchor vectors** and used as input to the generator, together with a **shared style vector**.

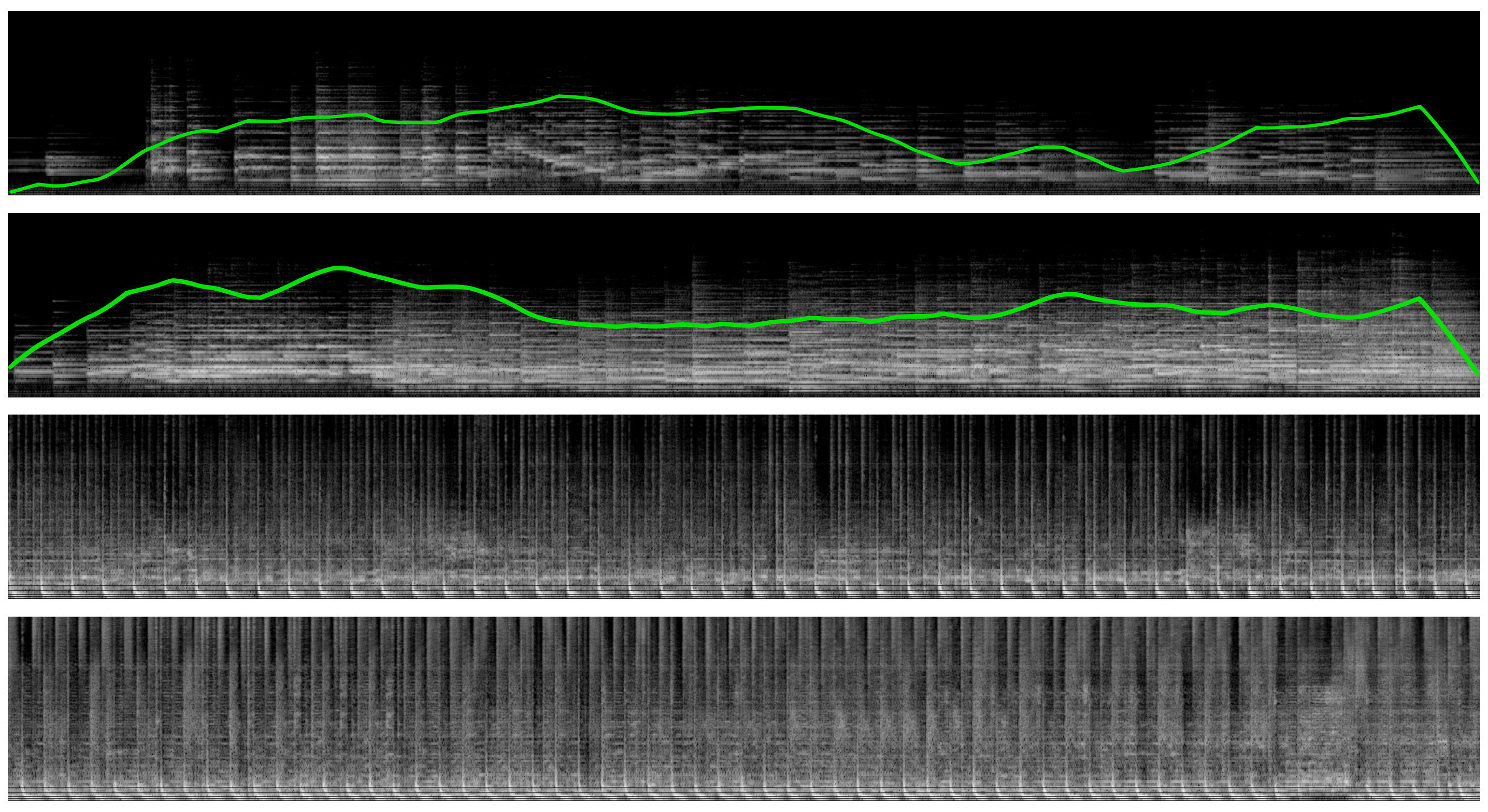

Fig. 4: Log-melspectrograms of **generated piano and techno samples from the conditional models**. For the piano samples (top row), we indicate the corresponding **note density conditioning** with a green line. The tempo used as conditioning for the techno samples (bottom row) is **120 bpm and 160 bpm**, respectively.

Model (Faster than real-time)	GPU	CPU
Musika Uncond. Piano	972x	**40x**
Musika Cond. Piano	921x	**40x**
UNAGAN[1] Piano	28x	11x
Musika Uncond. Techno	**994x**	39x
Musika Cond. Techno	917x	39x

Tab. 1: Comparison of **generation speed** between the different models.

Model	FAD
Musika Uncond. Piano	**1.641**
Musika Cond. Piano Rand.	2.150
UNAGAN[1] Piano	11.183

Tab. 2: **FAD evaluation** for generated piano music. We evaluate a conditional Musika model using random values of note density as conditioning.

We see our system as solving an important technical challenge -- **real-time music generation of sufficient quality, conditioned on user input** -- and hope it can advance the field of **human-AI co-creation**.

[1] Jen-Yu Liu et al. “Unconditional Audio Generation with Generative Adversarial Networks and Cycle Regularization”. In: INTERSPEECH. 2020.

Contact us at marco.pasini.98@gmail.com jan.schlueter@jku.at