A Geometrical Approach to Deep Generative Music Modeling

Supervisors. Pr. Carmine Cella (Sorbonne, IRCAM, UC Berkeley), Dr. Pierre Saint-Germier (IRCAM, CNRS), Dr. Edouard Oyallon (ISIR, CNRS, HdR)

Scientific Background. With the rising popularity of Deep Learning (DL), generative models are now capable of generating timbral and musical structures with very limited human control and intervention. However, the opacity of deep neural networks makes it difficult to characterize precisely the way these structures are represented in generative music models. This makes the question of both computational and musicological analysis of generative music models especially challenging.
In spite of this, geometric approaches to DL have been a topic of growing interest over the past decade [BBCV21] [Web25]. This area of research focuses on the analysis of underlying geometric invariance in data in order to explain the success and efficiency of various DL architectures. A notable example of this approach can be found in Mallat's geometric interpretation of image classifying deep convolutional networks [Mal16], via the formalism of Lie Groups. The geometric approach to machine learning has also found fruitful application in the audio domain. For instance, the use of Multiscale Spectrograms (MSS) with Differentiable Digital Signal Processing (DDSP), as proposed by recent studies, offers an affinized representation of the Weyl-Heisenberg group [Ard24] [EHGR20], establishing the notion of geometric invariance on the perceptual micro-timescale. Likewise, the Joint Time-Frequency Scattering (JTFS) transform formalizes notions of geometric invariance on the perceptual mesoscale [ALM19] [CHC+23]. While these approaches are successful under certain limitations, they pose issues in the context of larger musical timescales. Furthermore, geometric analysis of abstract musical features relevant to music theory such as rhythmic structure, musical form, phrase development, **do not exist** in the literature. The goal of this PhD is to develop a principled approach for bridging the gap between high-level musical structures and their low-level short-timescale representations, through learning.

Scientific objectives and justification of the approach. The scientific objective of the thesis is to leverage the unifying theoretical power of geometric deep learning in order to provide a better understanding of representations of sound and music in the field of generative musical AI. Our central research hypothesis is that group representation theory—which provides insight into the structural invariance preserved under different geometric transformations—is an appropriate framework to articulate the various types of representations involved in generative AI music. In particular, this involves *mathematical* representations of various generative models, *musicological* representations of the input and output of these models, and *cultural* representations of both DL as a technology, and generative music as a new type of computer music, making this an inherently interdisciplinary research endeavor.
First, this PhD project will focus on studying the computational properties of various deep learning (DL) models for music generation. Using synthetic data, experiments will be conducted to analyze geometric invariance within existing architectures such as DDSP [EHGR20], RAVE [CE21], and MusicGen [CKG+ 24]. This work will then be extended to diffusion models like REF, which have led to significant breakthroughs. Our approach involves incorporating Lie Group structures into key components such as linear attention mechanisms [Mamba]. The challenges fall into two main categories: **Model Design and Generation-** Developing an appropriate DL model capable of successfully generating meaningful musical signals. This requires expertise in deep learning, an area where E. Oyallon specializes, as well as proficiency in signal estimation, where C. Celle has expertise. **Latent Space Analysis-** Investigating whether the model naturally forms a structured latent space remains an open question.

Second, during this PhD, a wide ethnographic study will be conducted with computer scientists and musicians using these models. In particular, the candidate will take Ircam's ACIDs team, specialized in

music DL, as a case study. ACIDs is particularly relevant because it regularly collaborates with contemporary music composers both within and outside Ircam productions (Alexander Schubert, Holly Herndon), and contains as members engineers who are also musicians who use RAVE models in their own musical practices, involving experimental and improvisational music. This combination of methods will provide the desired integration of mathematical, musicological, and cultural representations.

The Collegium Musicae environment. This research aligns both with the "Improvisation, apprentissage, intelligence artificielle" axis of the Collegium Musicae, since improvisation is among the use of DL models covered by the project. It also aligns with the "Construction des savoirs musicaux" axis of the Collegium, since it proposes a new theoretical framework to articulate knowledge about the outputs of music generative models. This research plainly fits the approach of APM, geared towards the analysis of music and music technology in practice. Finally, the project further complements IRCAM's broader initiatives in computational creativity, human-machine co-creativity, and AI-assisted composition.

Profile Candidate. The candidate will have a double specialization in DL engineering and music. Ideally, they would have first-hand experience in the design, training, and use of audio and music generative models and formal training or demonstrable knowledge of the mathematical theory behind state-of-the art DL models. Conversely, formal training or demonstrable knowledge of music theory, music composition or musicology is expected.

References.

[ALM19] Joakim Anden, Vincent Lostanlen, and Stéphane Mallat. Joint Time-Frequency Scattering. IEEE Transactions on Signal Processing, 67(14):3704–3718, Jul 2019.
[Ard24] Max Ardito. An Acousmatic Approach to Neural Audio Synthesis. Master's thesis, McGill University, Montréal, Québec, Canada, 2024.
[BBCV21] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. https://arxiv.org/pdf/2104.13478, 2021.
[CE21] Antoine Caillon and Philippe Esling. RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis. arXiv:2111.05011, Dec 2021.
[CHC+ 23] Vahidi Cyrus, Han Han, Wang Changhong, Lagrange Mathieu, Fazekas György, and Lostanlen Vincent. Mesostructures: Beyond Spectrogram Loss in Differentiable Time–Frequency Analysis. Journal of The Audio Engineering Society, 71:577–585, September 2023.
[CKG+ 24] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation, 2024.
[EHGR20] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differential Digital Signal Processing. CoRR, abs/2001.04643, Jan 2020. https://arxiv.org/abs/2001.04643.
[Mal16] Stéphane Mallat. Understanding Deep Convolutional Networks. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150203, Apr 2016.
[Web25] Melanie Weber. Geometric Machine Learning. AI Magazine, 46(1):e12210, 2025.
[Mamba] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.