Artificial music theory: Interpreting the hidden spaces in neural-network models as a tool for music-theoretical investigation

Gabriele Cecchetti Roger T. Dean

The MARCS Institute for Brain, Behaviour and Development (Western Sydney University)

WESTERN SYDNEY UNIVERSITY

The MARCS Institute for Brain, Behaviour and Development

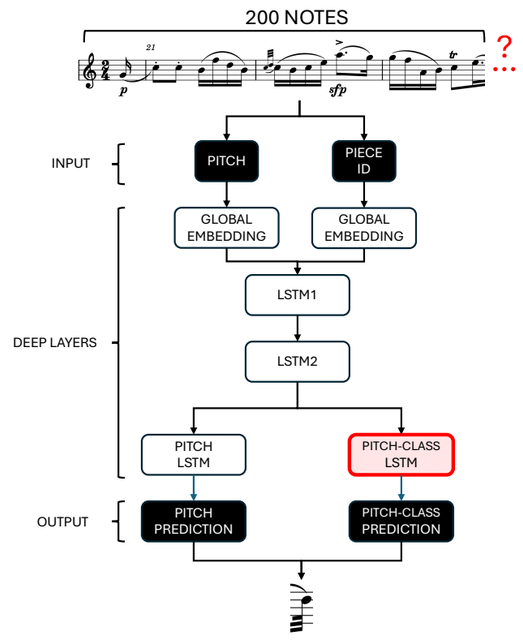
Emergent representations as artificial theories

What do machines hear? When exposed to a musical repertoire, humans learn structural abstractions such as chords and harmonic relationships between chords. Similarly, a Deep Neural Network (DNN) encodes representations that are useful towards performance in the training task. In principle, then, both humans and machines learn **a theory** of the musical repertoire they are exposed to. For example, Large Language Models are known to encode syntactic relations between words in terms of the **geometry** of the latent space where words are embedded. However, to what extent **musically and cognitively interpretable structural abstractions** emerge unsupervised in "black-box" DNNs when they are trained for music generation is unclear.

Tonal harmony. In tonal music, individual notes are understood as expressing underlying latent entities, the chords. The temporal succession of chords is ruled by syntactic principles. Some pairs of chords have a preferred goal-directed directionality that is temporally asymmetric: for example, dominant chords lead towards and resolve onto tonic chords, not vice versa. Discovering these structuring principles is a core goal of music theory: do DNNs learn a notion of **chord identity** and **harmonic directedness**?

Free improvisation. In non-idiomatic free improvisation, structure can emerge spontaneously during performance and remains largely unspoken and embodied. If DNNs can learn an **artificial theory** of a repertoire without enforcing prior theoretical commitments, can they be useful to uncover structure in improvised repertoires?

The model



Dataset

Distant Listening Corpus
1.283 pieces from tonal repertoires
(XVII-XX centuries)

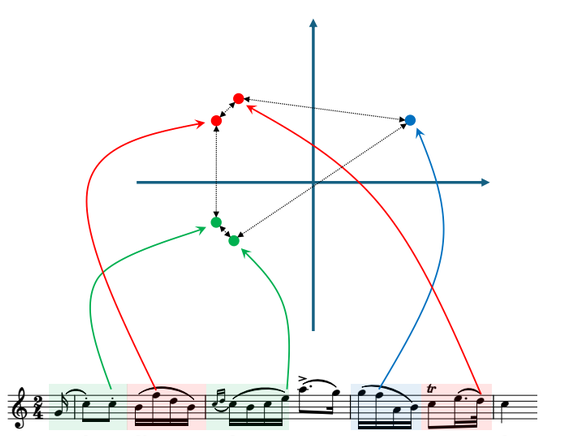
1.553.115 notes
242.867 expert-annotated chord labels
(used for analysis, not for training)

Training objective

	Pitch x	Pitch Class distr. $p(c)$
kind	numerical	categorical
loss	MSE $(x - \hat{x})^2$	Cross-Entropy $\sum \hat{p}(c) \ln p(c)$

The LSTM's hidden states assign each note to a location in a 512-dimensional embedding space. We examine the notes' embedding distances in Pitch Class LSTM.

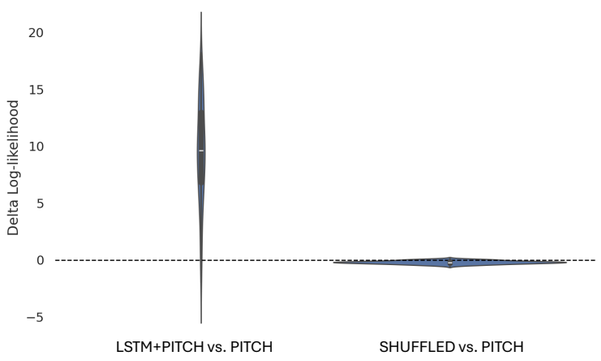
Chord identity



We used **three logistic regressions** to predict whether two notes express the same chord.

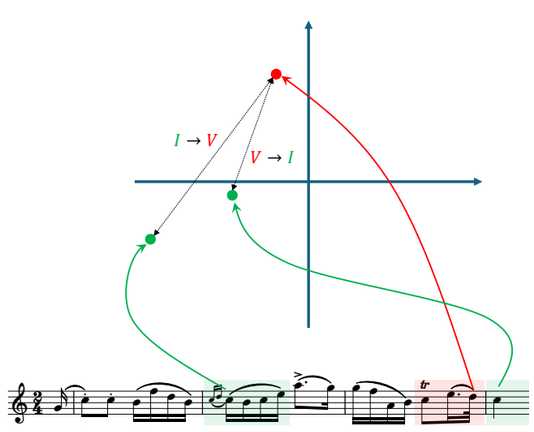
Predictors

- **LR1** LSTM embedding distances + pitch difference
- **LR2** pitch difference
- **LR3** LSTM embedding distances (randomly assigned) + pitch difference



Model comparisons show that LR1 outperforms the other models: the LSTM hidden states corresponding to individual notes encode information about their underlying chord abstraction.

Harmonic directedness

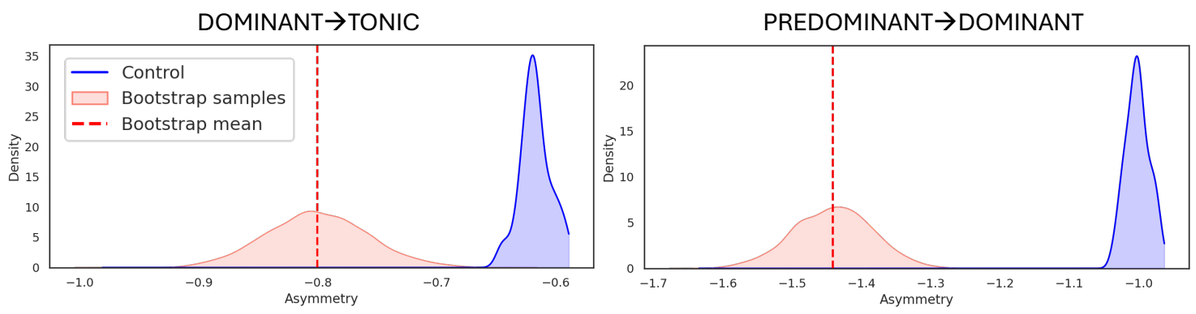


The **asymmetry score** is the difference in average embedding distance between notes expressing chords X and Y when Y follows X vs. when Y precedes X:

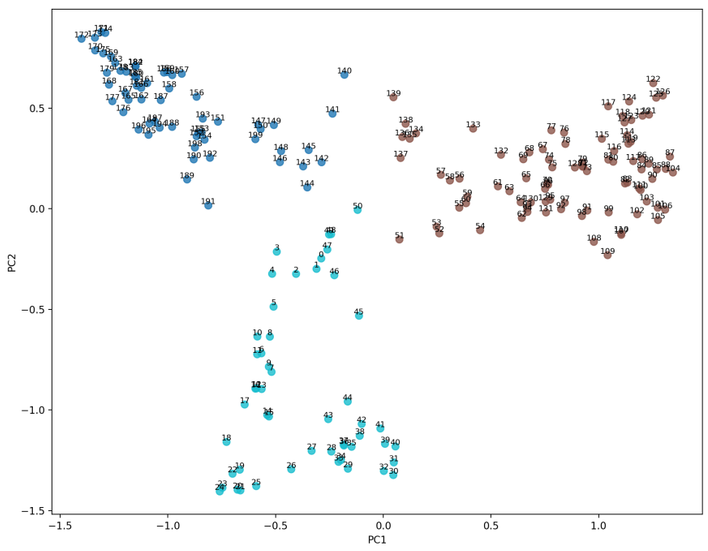
$$A(X \to Y) = \mathbb{E}_{X \to Y}\left[d(x, y)\right] - \mathbb{E}_{Y \to X}\left[d(y, x)\right]$$

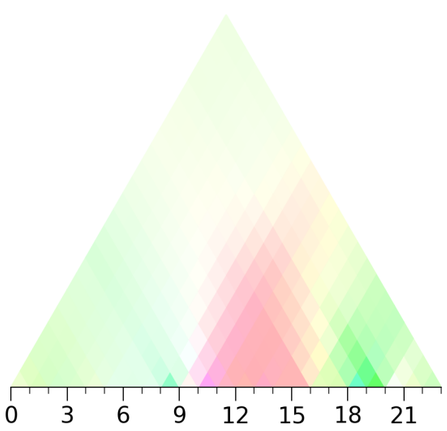
A **control distribution** was obtained by replacing the embedding vectors for every pair of notes with the embedding vectors of random notes at the same ordinal distance.

Negative asymmetry scores for goal-directed chord progressions indicate that the LSTM hidden states encode their syntactically preferred directionality.



Beyond tonality: freely improvised music

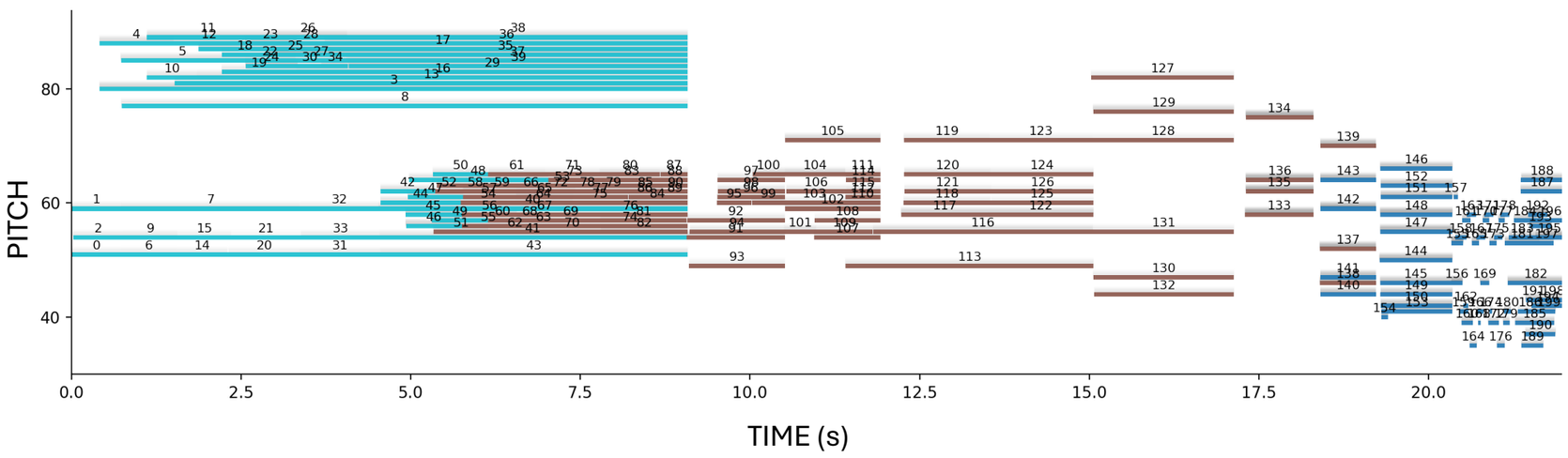




	Dataset	Genre	Frozen	Trained
Pretraining	MAESTRO	classical	—	full model
Finetuning	Dean & Forth	improvised	Pitch Embedding, LSTM 1 and 2	Piece Embedding, Pitch LSTM, Pitch-Class LSTM

We performed **unsupervised clustering** (K-means) on the LSTM latent space, selecting the number of clusters that optimizes the silhouette score. Results of the clustering are projected in 2D through Principal Component Analysis (left) and visualized on the piano-roll of the input (bottom).

The geometry of the LSTM hidden states reflects structural features of the improvised input sequence. In this example, the three clusters reveal a meaningful **temporal segmentation**. This loosely reflects the organisation of pitch materials in three regions, as highlighted in the wavescape (bottom-left) for the 2nd DFT component (diadicity).



Australian Research Council grant n. LP200301401

Gabriele.Cecchetti@westernsydney.edu.au