Understanding Equivariant Self-Supervised Learning in Musical Pitch Class Space

Vincent Lostanlen, Yuexuan Kong, Gabriel Meseguer-Brocal, Mathieu Lagrange, Romain Hennequin

Understanding Equivariant Self-Supervised Learning in Musical Pitch Class Space

Vincent Lostanlen, Yuexuan Kong, Gabriel Meseguer-Brocal, Mathieu Lagrange, Romain Hennequin

***Abstract*—STONE, which stands for self-supervised tonality estimator, has recently demonstrated the practical feasibility of recognizing key signatures in music signals given little or no human annotation. In this article, we revisit STONE from a more theoretical standpoint. We show that cross-power spectral density (CPSD) defines a differentiable measure of harmonic discrepancy between key signature profiles (KSP). Having set the CPSD frequency to seven cycles per octave, we offer a geometric interpretation of this discrepancy via the circle of fifths and conduct an algebraic study to prove that all its local minima are global. We rely on the equivariance property of deep convolutional networks to prove that the STONE loss function is invariant to circular frequency shifts of the constant-Q transform. We conclude by identifying a phenomenon of spontaneous symmetry breaking in STONE: since modes of limited transposition (e.g., augmented, diminished, whole-tone) are multistable in the circle of fifths, the associated CPSD gradient is driven towards more "tonal" (i.e., asymmetric) scales, such as diatonic or pentatonic.**

***Index Terms*—Convolutional networks, cross-power spectral density, equivariance, music, self-supervised learning.**

I. Introduction

EQUIVARIANCE is an algebraic property which is useful in the design of deep convolutional networks (convnets) [1]. In the context of music signal processing, equivariance to pitch transposition may be accomplished by applying a constant-Q transform (CQT), i.e., a time–frequency representation in which the frequency boundaries between adjacent subbands follow a geometric progression: $\xi_n = 2^{n/Q}\xi_0$ where n and Q are integers and where ξ_0 is a positive constant [2].

Certain music information retrieval (MIR) tasks combine pitch equivariance with octave equivalence, i.e., *invariance* to pitch transposition for intervals which are multiples of the octave: $k = mQ$ for any $m \in \mathbb{Z}$. These considerations have inspired many techniques in feature engineering, e.g., chromagrams [3], spiral scattering [4], folded CQT [5]; and feature learning, e.g., deep chroma [6], spiral convnets [7], and Tonnetz-space transform [8]. Another line of research has proposed to discover pitch equivariance and octave equivalence from unlabeled CQT samples of musical sounds [9], [10]. Yet, until recently, these algebraic properties were regarded as insufficient for the modeling of high-level music information such as tonality or chords. For these tasks, the CQT–convnet pipeline and its octave-equivalent variants have historically been trained via supervised learning [11].

V. Lostanlen and M. Lagrange are with Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France. Y. Kong, G. Meseguer-Brocal, and R. Hennequin are with Deezer Research, Paris, 75009, France. This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project MuReNN (ANR-23-CE23-0007-01).

The situation changed in 2024 with the publication of the first self-supervised tonality estimator (STONE) [12]. Building upon prior work in self-supervised fundamental frequency estimation [13], STONE is a convnet which is trained to regress artificial pitch transpositions between any two unlabeled musical excerpts from the same audio track. It does so with two essential tools: a fully pitch-equivariant and octave-invariant convnet, named ChromaNet; and a differentiable loss function based on circular cross-power spectral density (CPSD). STONE learns a 12-dimensional representation of CQT, which we call *key signature profile* (KSP) because it correlates with key signature. A modified version of STONE, known as S-KEY, has recently surpassed supervised models in the automatic classification of major and minor keys [14]. But despite its practical success in MIR, STONE lacks a rigorous foundation.

In this article, we present a mathematical theory of equivariant self-supervised learning (SSL) in musical pitch space. The key idea is to harness the algebraic properties of the discrete Fourier transform (DFT), particularly the convolution theorem and modular arithmetics between integer exponents of the complex roots of unity. Our main result is that STONE learns to solve a well-posed inverse problem up to a global pitch class shift: the associated CPSD-based objective admits Q global minima, which coincide with paired vertices of a regular polygon in the complex plane. Optimizing convnet weights with this objective encourages equivariance to key signature versus robustness to other aspects of musical information, such as melody, rhythm, lyrics, and instrumentation.

Section II presents the STONE model, particularly the KSP operator for octave equivalence in the ChromaNet and the (ω, k)-discrepancy as a differentiable and nonconstrative objective for equivariant SSL. Section III shows that the CPSD-based objective is *sparsity-inducing*: at any of its global minima, the two ChromaNet predictions have a single nonzero entry, which may be interpreted as a key signature. Section IV shows that the objective is *invariant*: pitch-shifting both KSPs leaves the objective unchanged; in musical terms, the learned assignment of CQT to key signature is relative, not absolute. Section V shows that the objective is *spontaneously symmetry-breaking*: it is globally maximal at periodic KSPs; in particular those which describe modes of limited transposition, e.g., augmented, diminished, and whole-tone.

Together, these three properties prevent STONE from collapsing to a trivial solution, e.g., a constant KSP; explain how the lack of manual annotation may be obviated by prior knowledge about signal processing and musical harmony; and enable its interpretability in the practical context of tonality estimation.

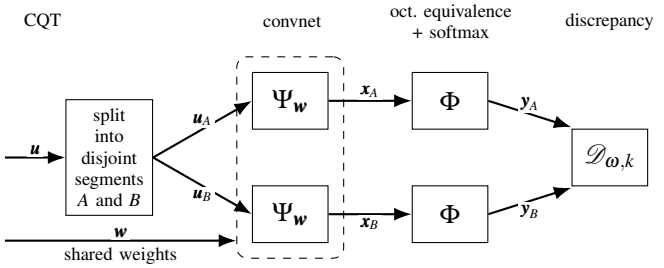

Fig. 1. Flowchart of a deep learning model for self-supervised tonality estimation (STONE). We split constant-Q transform (CQT) $\boldsymbol{u}$ of an unlabeled music signal into disjoint segments A and B, yielding a pair of CQTs $(\boldsymbol{u}_A, \boldsymbol{u}_B)$ which we assume to be in the same key. We feed $\boldsymbol{u}_A$ and $\boldsymbol{u}_B$ to a deep convolutional network (convnet) $\Psi_{\boldsymbol{w}}$, followed by a non-learnable operator Φ which performs octave equivalence and softmax. This results in two key signature profiles (KSP) $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$; i.e., nonnegative vectors summing to one in dimension Q. These KSPs serve as input to $\mathscr{D}_{\omega,k}$, a measure of harmonic discrepancy which is based on the cross-power spectral density between $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ and which is differentiable with respect to neural network weights $\boldsymbol{w}$.

Notation: Integers J and Q are the number of octaves and number of bins per octave in the CQT, respectively. The integer k is the pitch interval between the key signatures of A and B, measured in semitones if $\omega = 1$ or in perfect fifths if $\omega = 7$.

II. The STONE model for tonality estimation

Definition II.1. Given $Q \in \mathbb{N}^*$, the *standard Q-simplex* is the set of vectors in $\mathbb{R}^Q$ with nonnegative entries summing to one:

$$S_Q = \left\{ \boldsymbol{y} \in \mathbb{R}^Q \,\middle|\, \|\boldsymbol{y}\|_1 = 1;\ \forall q \in \mathbb{Z}/Q\mathbb{Z},\, \boldsymbol{y}[q] \geq 0 \right\}. \quad (1)$$

Definition II.2. Given $J, Q \in \mathbb{N}^*$, the *key signature profile (KSP)* operator is the function $\Phi : \mathbb{R}^{JQ} \to S_Q$ such that:

$$\forall \boldsymbol{x} \in \mathbb{R}^{JQ},\, \forall q \in \mathbb{Z}/Q\mathbb{Z},\, \Phi(\boldsymbol{x})[q] = \frac{\exp\left(\sum_{j=0}^{J-1} \boldsymbol{x}[jQ+q]\right)}{\sum_{q'=0}^{Q-1} \exp\left(\sum_{j=0}^{J-1} \boldsymbol{x}[jQ+q']\right)}. \quad (2)$$

Definition II.3. Given $Q \in \mathbb{N}^*$, $\omega \in \mathbb{Z}/Q\mathbb{Z}$ and $k \in \mathbb{Z}$, the (ω,k)*-discrepancy* between two KSP vectors $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ is:

$$\mathscr{D}_{\omega,k}(\boldsymbol{y}_A, \boldsymbol{y}_B) = \frac{1}{2} \left| e^{2\pi \mathrm{i} \omega k/Q} - \hat{\boldsymbol{y}}_A[\omega] \hat{\boldsymbol{y}}_B[\omega]^* \right|^2 \quad (3)$$

where $\hat{\boldsymbol{y}}_A[\omega] = \sum_{q=0}^{Q-1} \boldsymbol{y}_A[q] \exp(-2\pi \mathrm{i} \omega q/Q)$ is the discrete Fourier transform (DFT) of $\boldsymbol{y}_A$ at ω and idem for $\boldsymbol{y}_B$, and where the asterisk symbol denotes complex conjugation.

Note. The (ω,k)-discrepancy is a function of the circular cross-correlation $\boldsymbol{R}_{AB}$ between $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$, defined by: $\boldsymbol{R}_{AB}[\tau] = \sum_{q=0}^{Q-1} \boldsymbol{y}_A[q] \boldsymbol{y}_B[(q+\tau) \bmod Q]$. Indeed, applying a circular change of variable $\tau = (q'-q) \bmod Q$ to $\hat{\boldsymbol{y}}_A[\omega] \hat{\boldsymbol{y}}_B[\omega]^*$ leads to a convolution theorem of the form:

$$\begin{aligned} \hat{\boldsymbol{y}}_A[\omega] \hat{\boldsymbol{y}}_B[\omega]^* &= \sum_{q=0}^{Q-1} \sum_{q'=0}^{Q-1} \boldsymbol{y}_A[q] \boldsymbol{y}_B[q'] \exp\left(2\pi \mathrm{i} \frac{\omega(q'-q)}{Q}\right) \\ &= \sum_{q=0}^{Q-1} \sum_{\tau=0}^{Q-1} \boldsymbol{y}_A[q] \boldsymbol{y}_B[(q+\tau) \bmod Q] \exp\left(2\pi \mathrm{i} \frac{\omega \tau}{Q}\right) \\ &= \widehat{\boldsymbol{R}}_{AB}[\omega]^* = \widehat{\boldsymbol{R}}_{BA}[\omega]. \end{aligned} \quad (4)$$

$\widehat{\boldsymbol{R}}_{AB}$ is sometimes called *cross-power spectral density (CPSD)*.

Definition II.4. Given $Q \in \mathbb{N}^*$ and $\omega, k \in \mathbb{Z}/Q\mathbb{Z}$; two CQT spectrograms $\boldsymbol{u}_A$, $\boldsymbol{u}_B$; and a model $\Psi_{\boldsymbol{w}}$ with trainable weights $\boldsymbol{w}$, the *STONE loss* is the function $\mathscr{L}_{\omega,k} : W \to \mathbb{R}_{\geq 0}$ such that:

$$\mathscr{L}_{\omega,k}(\boldsymbol{w} \mid \boldsymbol{u}_A, \boldsymbol{u}_B) = \mathscr{D}_{\omega,k}\big((\Phi \circ \Psi_{\boldsymbol{w}})(\boldsymbol{u}_A), (\Phi \circ \Psi_{\boldsymbol{w}})(\boldsymbol{u}_B)\big), \quad (5)$$

where W is the domain of $\boldsymbol{w}$, Φ is the key signature profile operator (Definition II.2) and $\mathscr{D}_{\omega,k}$ is (ω,k)-discrepancy (Definition II.3). The vertical bar in $\mathscr{L}_{\omega,k}(\boldsymbol{w} \mid \boldsymbol{u}_A, \boldsymbol{u}_B)$ denotes a separation between trainable parameters and input data.

Note. In practice, we construct pairs $(\boldsymbol{u}_A, \boldsymbol{u}_B)$ by extracting two disjoint 15-second segments within the same unlabeled music signal $\boldsymbol{u}$. By setting $k = 0$ in $\mathscr{D}_{\omega,k}$, we assume that $\boldsymbol{u}_A$ and $\boldsymbol{u}_B$ are in the same (unknown) key. We concede that this assumption is occasionally disproven by the presence of key changes between segments A and B. However, key changes have become increasingly rare in mainstream pop music, at least at the 15-second time scale [1]. Thus, the founding hypothesis of STONE is that the rate of label noise (setting $k = 0$ despite occasional key changes) is outweighed by the sheer scale of available data for SSL, of the order of one million songs [14].

Proposition II.5. The STONE loss is differentiable.

Proof. The model $\Psi_{\boldsymbol{w}}$ is differentiable over W, by construction. The KSP operator Φ is differentiable over $\mathbb{R}^{JQ}$ as a composition of a linear transformation and a pointwise softmax (Equation 2). Lastly, $\mathscr{D}_{\omega,k}$ is differentiable with respect to the real and imaginary parts of both its input arguments as a polynomial of degree two (Equation 4). We conclude using Equation 5. ■

Note. We typically set $Q = 12$ to match twelve-tone equal temperament in music. For the sake of conciseness, we do not recall the dependency of Φ, $\mathscr{D}_{\omega,k}$, and $\mathscr{L}_{\omega,k}$ upon Q.

III. Sparsity properties

Definition III.1. Given $Q \in \mathbb{N}^*$ and $\omega \in \mathbb{Z}/Q\mathbb{Z}$, the *Kronecker delta symbol* at index $\alpha \in \mathbb{Z}/Q\mathbb{Z}$ is the vector such that, for every $q \in \mathbb{Z}/Q\mathbb{Z}$, $\boldsymbol{\delta}_\alpha[q] = 1$ if $q = \alpha \bmod Q$ and zero otherwise.

Theorem III.2. *Let $Q \in \mathbb{N}^*$ and $\omega, k \in \mathbb{Z}/Q\mathbb{Z}$. If $\gcd(\omega, Q) = 1$, the global minima of $\mathscr{D}_{\omega,k}$ over S_Q are the pairs $(\boldsymbol{y}_A, \boldsymbol{y}_B) \in S_Q$ of the form $(\boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_{\alpha+k})$ for every $\alpha \in \mathbb{Z}/Q\mathbb{Z}$.*

We will apply the lemma below, proven in the Appendix.

Lemma III.3. *Let P and Q be two coprime integers where $Q > 0$. For all $q, q' \in \mathbb{Z}/Q\mathbb{Z}$ where $q \neq q'$, $qP \bmod Q \neq q'P \bmod Q$.*

Proof of Theorem III.2. Since $\mathscr{D}_{\omega,k}$ is clearly nonnegative, it suffices to identify all antecedents of zero by $\mathscr{D}_{\omega,k}$ in S_Q. The identity $\mathscr{D}_{\omega,k}(\boldsymbol{y}_A, \boldsymbol{y}_B) = 0$ implies $\hat{\boldsymbol{y}}_A[\omega] \hat{\boldsymbol{y}}_B[\omega]^* = \exp(2\pi \mathrm{i} \omega k)$, which in turn implies $|\hat{\boldsymbol{y}}_A[\omega]||\hat{\boldsymbol{y}}_B[\omega]| = 1$ after applying the complex modulus. Yet, by the triangular inequality and by definition of the DFT, $\boldsymbol{y}_A \in S_Q$ implies $|\hat{\boldsymbol{y}}_A[\omega]| \leq \|\boldsymbol{y}_A\|_1 = 1$. Therefore, $\mathscr{D}_{\omega,k}(\boldsymbol{y}_A, \boldsymbol{y}_B) = 0$ corresponds to an equality case of the triangular inequality: all nonzero terms in the DFT of $\boldsymbol{y}_A$ at ω have the same phase. By contradiction, let us

[1] Visit: https://ethanhein.substack.com/p/identifying-modulations

assume that there are two such nonzero terms, indexed by q and q' respectively. Since ω is coprime with Q, we may apply Lemma III.3 and deduce that $\omega q \bmod Q \neq \omega q' \bmod Q$: in other words, $\exp(2\pi \mathrm{i}\omega q/Q)$ and $\exp(2\pi \mathrm{i}\omega q'/Q)$ have different phases modulo 2π, which leads to a contradiction. Furthermore, let us assume that there are no nonzero terms: this implies $\boldsymbol{y}_A \notin S_Q$ and leads to another contradiction. Having excluded both of these contradictions, we deduce that there exists a unique nonzero term in the DFT of $\boldsymbol{y}_A$ at ω: let us denote its index by $\alpha \in \mathbb{Z}/Q\mathbb{Z}$. Since $\boldsymbol{y}_A \in S_Q$, it follows that $\boldsymbol{y}_A[\alpha] = 1$ and $\boldsymbol{y}_A[q] = 0$ for $q \neq \alpha \bmod Q$. The same reasoning for $\boldsymbol{y}_B$ leads us to claim that there exists $\beta \in \mathbb{Z}/Q\mathbb{Z}$ such that $\boldsymbol{y}_B[\beta] = 1$ and $\boldsymbol{y}_B[q] = 0$ for $q \neq \beta \bmod Q$. Thus $\boldsymbol{y}_A = \boldsymbol{\delta}_\alpha$ and $\boldsymbol{y}_B = \boldsymbol{\delta}_\beta$. We compute the following half-angle factorization:

$$\begin{aligned}\mathscr{D}_{\omega,k}(\boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_\beta) &= \left| e^{2\pi \mathrm{i}\omega k/Q} - \hat{\boldsymbol{\delta}}_\alpha[\omega]\hat{\boldsymbol{\delta}}_\beta[\omega]^* \right|^2 \\ &= \left| e^{2\pi \mathrm{i}\omega k/Q} - e^{2\pi \mathrm{i}\omega(\beta-\alpha)/Q} \right|^2 \\ &= 4\left| \sin\left(2\pi \frac{\omega(k+\alpha-\beta)}{Q} \right) \right|^2. \end{aligned} \tag{6}$$

Solving $\mathscr{D}_{\omega,k}(\boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_\beta) = 0$ leads to $\omega(k+\alpha-\beta) = 0 \bmod Q$, which we rewrite as $\omega(\beta-\alpha) = \omega k \bmod Q$. Since ω is coprime with Q, we may apply Lemma III.3 again and deduce that $\beta = (\alpha + k) \bmod Q$, which completes the proof. ■

Note. The condition $\gcd(\omega, Q) = 1$ is necessary to guarantee that $\mathscr{D}_{\omega,k}$ has a countable (*a fortiori*, finite) number of global minima. Assuming $\gcd(\omega, Q) = m$ with $m > 1$, and letting a parameter ρ take any real value between zero and one, we may observe that $\mathscr{D}_{\omega,k}(\rho\boldsymbol{\delta}_\alpha + (1-\rho)\boldsymbol{\delta}_{\alpha+(Q/m)}, \boldsymbol{\delta}_{\alpha+k}) = 0$.

Theorem III.4. *Let $Q \in \mathbb{N}^*$ and $\omega, k \in \mathbb{Z}/Q\mathbb{Z}$. If $\gcd(\omega, Q) = 1$, all local minima of the (ω,k)-discrepancy operator are global.*

Proof. Let the bilinear map $\mathscr{B}_\omega$ denote the CPSD operator at frequency ω: for any $\boldsymbol{y}_A, \boldsymbol{y}_B$, $\mathscr{B}_\omega(\boldsymbol{y}_A, \boldsymbol{y}_B) = \hat{\boldsymbol{y}}_A[\omega]\hat{\boldsymbol{y}}_B[\omega]^*$ (see Equation 4). We define $\mathscr{E}_{\omega,k} : z \in \mathbb{C} \mapsto \frac{1}{2}|e^{2\pi \mathrm{i}\omega k/Q} - z|^2$ and decompose $\mathscr{D}_{\omega,k}$ as $(\mathscr{E}_{\omega,k} \circ \mathscr{B}_\omega)$. Up to a factor of one half, $\mathscr{E}_{\omega,k}$ is proportional to the squared Euclidean distance to the point $z_{\omega,k} = e^{2\pi \mathrm{i}\omega k/Q}$ in the complex plane. Thus, $\mathscr{E}_{\omega,k}$ is strongly convex and has a single local minimum over $\mathbb{C}$: $\mathscr{E}_{\omega,k}(z_{\omega,k}) = 0$. Analogously to the proof of Theorem III.2, we apply Lemma III.3 and an equality case of the triangular inequality to show that the antecedents of $z_{\omega,k}$ by the bilinear map $\mathscr{B}_\omega$ are the pairs $(\boldsymbol{y}_A, \boldsymbol{y}_B) \in S_Q$ of the form $(\boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_{\alpha+k})$ for every $\alpha \in \mathbb{Z}/Q\mathbb{Z}$. By Theorem III.2, all of these pairs are global minima for $\mathscr{D}_{\omega,k}$, which completes the proof. ■

Note. The result above may or may not translate to the loss function $\mathscr{L}_{\omega,k}$, depending on the choice of neural network architecture $\Psi_{\boldsymbol{w}}$ and initialization. If the Jacobian of $\boldsymbol{w} \mapsto (\Psi_{\boldsymbol{w}}(\boldsymbol{u}_A), \Psi_{\boldsymbol{w}}(\boldsymbol{u}_B))$ is singular at some $\boldsymbol{w} \in W$, the STONE gradient $\boldsymbol{\nabla}\mathscr{L}_{\omega,k}$ may vanish even though the partial derivatives of $\mathscr{D}_{\omega,k}$ with respect to $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ are nonzero. To reduce the risk of a singular Jacobian, [12] have designed $\Psi_{\boldsymbol{w}}$ as a residual network (ResNet); more specifically, a ConvNeXT [15].

IV. Invariance properties

Definition IV.1. Let G be a group with identity element 1_G and X be a set. A *(left) group action* of G on X is a function $\mathscr{A} : G \times X \to X$ such that, for all $\boldsymbol{x} \in X$, $\mathscr{A}(1_G, \boldsymbol{x}) = \boldsymbol{x}$; and such that, for all $\boldsymbol{x} \in X$, for all $g, g' \in G$, $\mathscr{A}(g, \mathscr{A}(g', \boldsymbol{x})) = \mathscr{A}(gg', \boldsymbol{x})$.

Definition IV.2. A function $\Phi : X \to Y$ is *equivariant* to G if there exist group actions $\mathscr{A}_X : G \times X \to X$ and $\mathscr{A}_Y : G \times Y \to Y$ such that, for all $\boldsymbol{x} \in X$ and $g \in G$, $\Phi(\mathscr{A}(g, \boldsymbol{x})) = \mathscr{A}_Y(g, \Phi(\boldsymbol{x}))$.

Proposition IV.3. The key signature profile (KSP) is equivariant to circular frequency shifts in the constant-Q transform.

Proof. Let $X = \mathbb{R}^{JQ}$ and $Y = \mathbb{R}^Q$. Consider the group $G = \mathbb{Z}$ and the group action $\mathscr{A}_X : G \times X \to X$ such that:

$$\forall g \in G, \forall q \in \mathbb{Z}/JQ\mathbb{Z}, \quad \mathscr{A}_X(g, \boldsymbol{x})[q] = \boldsymbol{x}[(q-g) \bmod JQ]. \tag{7}$$

Similarly, consider the group action $\mathscr{A}_Y : G \times Y \to Y$ such that:

$$\forall g \in G, \forall q \in \mathbb{Z}/Q\mathbb{Z}, \quad \mathscr{A}_Y(g, \boldsymbol{y})[q] = \boldsymbol{y}[(q-g) \bmod Q]. \tag{8}$$

Consider the sets $C_q = \{jQ + q \bmod JQ \mid 0 \leq j < J\}$ for each $q \in \mathbb{Z}/Q\mathbb{Z}$. Definition II.2 rewrites as:

$$\forall \boldsymbol{x} \in \mathbb{R}^{JQ}, \forall q \in \mathbb{Z}/Q\mathbb{Z}, \ \Phi(\boldsymbol{x})[q] = \frac{\prod_{n \in C_q} \exp \boldsymbol{x}[n]}{\prod_{q'=0}^{Q-1} \prod_{n' \in C_{q'}} \exp \boldsymbol{x}[n']}. \tag{9}$$

Given $g \in G$, consider $\mathscr{A}_G(g, C_q) = \{(n-g) \bmod JQ \mid n \in C_q\}$. By property of the Euclidean division: for every $n \in C_q$, there exists a unique $j \in \mathbb{Z}$ such that $n = jQ + q$. Thus:

$$\mathscr{A}_G(g, C_q) = \{(jQ + q - g) \bmod JQ \mid 0 \leq j < J\} = C_{(q-g)}. \tag{10}$$

Returning to Equation 9, a change of variables yields:

$$\begin{aligned}\mathscr{A}_Y(g, \Phi(\boldsymbol{x}))[q] &= \frac{\prod_{n \in C_{(q-g)}} \exp \boldsymbol{x}[n]}{\prod_{q'=0}^{Q-1} \prod_{n' \in C_{(q'-g)}} \exp \boldsymbol{x}[n']} \\ &= \frac{\prod_{n \in C_q} \exp \mathscr{A}_X(g, \boldsymbol{x}[n])}{\prod_{q'=0}^{Q-1} \prod_{n' \in C_{q'}} \exp \boldsymbol{x}[n']} \end{aligned} \tag{11}$$

Thus $\mathscr{A}_Y(g, \Phi(\boldsymbol{x})) = \Phi(\mathscr{A}_X(g, \boldsymbol{x}))$, concluding the proof. ■

Proposition IV.4. Given $Q \in \mathbb{N}^*$ and $\omega, k \in \mathbb{Z}/Q\mathbb{Z}$, consider the group $G = \mathbb{Z}$, the standard Q-simplex $Y = S_Q$ (Definition II.1), and the group action $\mathscr{A}_Y : G \times Y \to Y$ such that: $\forall g \in G, \forall q \in \mathbb{Z}/Q\mathbb{Z}, \mathscr{A}_Y(g, \boldsymbol{y})[q] = \boldsymbol{y}[(q-g) \bmod Q]$. For all $\boldsymbol{y}_A, \boldsymbol{y}_B \in S_Q$ and all $g_A, g_B \in G$, the (ω,k)-discrepancy (Definition II.3) satisfies:

$$\mathscr{D}_{\omega,k}(\mathscr{A}_Y(g_A, \boldsymbol{y}_A), \mathscr{A}_Y(g_B, \boldsymbol{y}_B)) = \mathscr{D}_{\omega,(k+g_A-g_B)}(\boldsymbol{y}_A, \boldsymbol{y}_B). \tag{12}$$

Proof. Let us denote the discrete Fourier transform (DFT) of $\mathscr{A}_Y(g_A, \boldsymbol{y}_A)$ by $\widehat{\mathscr{A}_Y(g_A, \boldsymbol{y}_A)}$. The DFT is equivariant to $\mathscr{A}_Y$: $\widehat{\mathscr{A}_Y(g_A, \boldsymbol{y}_A)}[\omega] = \exp(-2\pi \mathrm{i}\omega g_A/Q)\hat{\boldsymbol{y}}_A[\omega]$ and idem for $\widehat{\mathscr{A}_Y(g_B, \boldsymbol{y}_B)}$. Thus, by definition of (ω,k)-discrepancy:

$$\begin{aligned}&\mathscr{D}_{\omega,k}(\mathscr{A}_Y(g_A, \boldsymbol{y}_A), \mathscr{A}_Y(g_B, \boldsymbol{y}_B)) = \\ &\qquad \frac{1}{2}\left| e^{2\pi i\omega k/Q} - e^{2\pi i\omega(g_B-g_A)/Q}\hat{\boldsymbol{y}}_A[\omega]\hat{\boldsymbol{y}}_B[\omega]^* \right|^2. \end{aligned} \tag{13}$$

After multiplying the subtraction by $\exp(2\pi\mathrm{i}\omega(g_A - g_B)/Q)$, we recognize $\mathscr{D}_{\omega,(k+g_A-g_B)}(\boldsymbol{y}_A,\boldsymbol{y}_B)$ and conclude the proof. ■

Theorem IV.5. *The STONE loss function is invariant to circular frequency shifts in the constant-Q transform: given $Q \in \mathbb{N}^*$ and $\omega,k \in \mathbb{Z}/Q\mathbb{Z}$, for all $\boldsymbol{w} \in W$, $g \in \mathbb{Z}$, and $\boldsymbol{u}_A,\boldsymbol{u}_B \in \mathbb{R}^{JQ\times T}$,*

$$\mathscr{L}_{\omega,k}(\boldsymbol{w} \mid \mathscr{A}_U(g,\boldsymbol{u}_A),\mathscr{A}_U(g,\boldsymbol{u}_B)) = \mathscr{L}_{\omega,k}(\boldsymbol{w} \mid \boldsymbol{u}_A,\boldsymbol{u}_B), \quad (14)$$

where $\mathscr{A}_U(g,\boldsymbol{u}_A)[q,t] = \boldsymbol{u}_A[(q-g) \bmod JQ,t]$ is a cyclic group action of $\mathbb{Z}$ over $R^{JQ\times T}$ and idem for $\mathscr{A}_U(g,\boldsymbol{u}_B)$.

Proof. By Propositions IV.3 and IV.4 with $g_A = g_B = g$. ■

Note. The equivariance of $(\Phi \circ \Psi_{\boldsymbol{w}})$ offers the opportunity to include artificially transposed versions of $\boldsymbol{u}_A$ and $\boldsymbol{u}_A$ in the STONE loss, and derive a suitable value for the parameter k. For example, in [12], $\boldsymbol{u}_A$ and $\boldsymbol{u}_B$ are assumed to be in the same key and the modified STONE loss combines three terms:

$$\begin{aligned}\mathscr{L}^{\text{multi}}_{\omega,k}&(\boldsymbol{w} \mid \boldsymbol{u}_A,\boldsymbol{u}_B)\\ &= \mathscr{D}_{\omega,0}\big((\Phi\circ\Psi_{\boldsymbol{w}})(\boldsymbol{u}_A),(\Phi\circ\Psi_{\boldsymbol{w}})(\boldsymbol{u}_B)\big)\\ &+ \mathscr{D}_{\omega,k}\big((\Phi\circ\Psi_{\boldsymbol{w}})(\boldsymbol{u}_A),(\Phi\circ\Psi_{\boldsymbol{w}})(\mathscr{A}_U(k,\boldsymbol{u}_A))\big)\\ &+ \mathscr{D}_{\omega,k}\big((\Phi\circ\Psi_{\boldsymbol{w}})(\boldsymbol{u}_B),(\Phi\circ\Psi_{\boldsymbol{w}})(\mathscr{A}_U(k,\boldsymbol{u}_A))\big). \quad (15)\end{aligned}$$

Clearly, $\mathscr{L}^{\text{multi}}_{\omega,k}$ satisfies the same invariance property as in Theorem IV.5. At the same time, empirical evidence suggests that the pretext task $\mathscr{L}^{\text{multi}}_{\omega,k}$ is less susceptible to trivial collapse (i.e., predicting a constant KSP) than $\mathscr{D}_{\omega,0}$, while incurring a moderate (two-fold) increase in computational cost.

V. Symmetry-breaking properties

We have seen in Theorems III.2 and III.4 that all local minima of $\mathscr{D}_{\omega,k}$ are global minima. Moreover, we have characterized them as pairs of one-hot vectors, which are aperiodic patterns. Yet, by Proposition IV.4, $\mathscr{L}_{\omega,k}$ is invariant to circular pitch shifts. In physics, this phenomenon is known as spontaneous symmetry breaking (SSB) [16]: the stable states of a system—namely, $(\boldsymbol{\delta}_\alpha,\boldsymbol{\delta}_{\alpha+k})$ for $\alpha \in \mathbb{Z}/Q\mathbb{Z}$—are not symmetric (i.e., invariant) under a symmetry of the underlying "Hamiltonian" $\mathscr{D}_{\omega,k}$, understood as a function of $S_Q \times S_Q$.

We propose to elaborate on this observation by studying the value of $\mathscr{D}_{\omega,k}$ over pairs of periodic signals $(\boldsymbol{y}_A,\boldsymbol{y}_B)$, which are left unchanged by the action of proper subgroups of $G_Q = \mathbb{Z}/Q\mathbb{Z}$. Since G_Q is cyclic of order Q, these subgroups are of the form G_P with $P \mid Q$; i.e., P is a divisor of Q.

Note. In music theory, the binary vectors $\boldsymbol{y} \in \{0,1\}^Q$ which are fixed points of $\boldsymbol{y} \mapsto \mathscr{A}_Y(P,\boldsymbol{y})$ for some $P \mid Q$ are called "modes of limited transposition" [17]: e.g., tritone, augmented triad, diminished seventh chord, whole-tone scale, and more.

The theorem below shows that the STONE loss is at an unstable equilibrium if (but not only if) its inputs are periodic. For simplicity, we only discuss the case where both elements in the pair are equal ($\boldsymbol{y}_A = \boldsymbol{y}_B = \boldsymbol{y}$) and we have set $k = 0$.

Theorem V.1. *Given $Q \in \mathbb{N}^*$, $\omega \in \mathbb{Z}/Q\mathbb{Z}$ such that $\gcd(\omega,Q) = 1$, and $G = \mathbb{Z}$, consider the standard simplex $Y = S_Q$ (Definition II.1), the group action $\mathscr{A}_Y : G \times Y \to Y$ (Definition IV.1), and the $(\omega,0)$-discrepancy $\mathscr{D}_{\omega,0}$ (Definition II.3). For any $\boldsymbol{y} \in S_Q$, if $\boldsymbol{y}$ is periodic—i.e., if there exists $P \mid Q$, $1 \le P < Q$ such that $\boldsymbol{y} = \mathscr{A}_Y(P,\boldsymbol{y})$—then the gradient of "self-discrepancy" $\boldsymbol{y} \longmapsto \mathscr{D}_{\omega,0}(\boldsymbol{y},\boldsymbol{y})$ vanishes at $\boldsymbol{y}$: $\boldsymbol{\nabla}\mathscr{D}_{\omega,0}(\boldsymbol{y},\boldsymbol{y}) = 0$.*

We will apply the lemma below, proven in the Appendix.

Lemma V.2. *Given $Q \in \mathbb{N}^*$ and $\omega \in \mathbb{Z}/Q\mathbb{Z}$, for all $\boldsymbol{y} \in S_Q$,*

$$\mathscr{D}_{\omega,0}(\boldsymbol{y},\boldsymbol{y}) \le \frac{1}{2}. \quad (16)$$

Proof of Theorem V.1. Let us factorize Q as KP where P is the period of $\boldsymbol{y}$. For each $q \in \mathbb{Z}/Q\mathbb{Z}$, we perform the Euclidean division: $q = kP + r$ where $0 \le k \le K$ and $0 \le r < Q$. Denoting by $\mathscr{A}_Y^{(k)}$ the group action $\mathscr{A}_Y$ applied iteratively k times, the DFT of $\boldsymbol{y}$ at frequency ω rewrites as a double sum:

$$\begin{aligned}\hat{\boldsymbol{y}}[\omega] &= \sum_{k=0}^{K-1}\sum_{r=0}^{P-1} \mathscr{A}_Y^{(k)}(-P,\boldsymbol{y})[r]\exp\left(-2\pi\mathrm{i}\frac{\omega(kP+r)}{Q}\right)\\ &= \left(\sum_{k=0}^{K-1} e^{-2\pi\mathrm{i}\omega kP/Q}\right) \times \left(\sum_{r=0}^{P-1} \boldsymbol{y}[r]e^{-2\pi\mathrm{i}\omega r/Q}\right). \quad (17)\end{aligned}$$

We recognize the first factor as the sum of K^{th} complex roots of unity. Since K is integer, this sum equals zero. Equating $\hat{\boldsymbol{y}}[\omega]$ to zero in Definition II.3 yields $\mathscr{D}_{\omega,k}(\boldsymbol{y},\boldsymbol{y}) = \frac{1}{2}$. By Lemma V.2, this value is a global maximum for the application $\boldsymbol{y} \mapsto \mathscr{D}_{\omega,0}(\boldsymbol{y},\boldsymbol{y})$ over S_Q. Since this application is differentiable (see proof of Proposition II.5), its gradient vanishes at local extrema, and *a fortiori* at global maxima; which concludes the proof. ■

To recap, stochastic optimization with STONE is spontaneously symmetry-breaking because symmetric patterns have the highest self-discrepancy (Theorem V.1) while one-hot vectors have the lowest self-discrepancy (Theorem III.2).

VI. Conclusion

STONE, which stands for self-supervised tonality estimation, is a "hybrid deep learning" model [18]: what it lacks in human supervision, it makes up for in mathematical structure [12]. In this article, we have seen that the interpretability of learned STONE embeddings as key signature profiles (KSP) does not emerge spontaneously from large-scale training: it also depends upon a pretext task design which is informed by signal processing (CQT, CPSD), deep learning (ResNets, softmax), and basic music theory (octave equivalence and, optionally, the circle of fifths). Note that STONE is not strictly tied to the CQT: in principle, it generalizes to any filterbank with center frequencies in geometric progression: e.g., SWIPE kernels [19].

Beyond its usefulness in practice, STONE is the first proof of feasibility of a machine learning system which distinguishes key signatures from polyphonic audio without any prior concept of root, bass, triad, scale degree, or fundamental line. At the same time, its publication has led to further developments: for example, a recent paper has adapted the CPSD objective to a multi-class-token Transformer architecture for multitask SSL [20]. Lastly, from the perspective of music perception and cognition, the STONE model qualifies as "illiterate" (machine) listening [21], i.e., as remote from linguistic or symbolic determinations such as pitch class names. We leave the implications of this remark to future work.

REFERENCES

[1] R. Kondor and S. Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," in *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 2747–2755.

[2] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Proceedings of the Sound and Music Computing (SMC) Conference*, 2010.

[3] M. Müller, *Fundamentals of music processing: Audio, analysis algorithms, applications*. Springer, 2015.

[4] V. Lostanlen and S. Mallat, "Wavelet scattering on the pitch spiral," in *Proceedings of the Digital Audio Effects (DAFX) Conference*, 2015.

[5] J.-F. Ducher and P. Esling, "Folded CQT RCNN for real-time recognition of instrument playing techniques," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2019.

[6] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," 2016.

[7] V. Lostanlen and C.-E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2016.

[8] E. J. Humphrey, T. Cho, and J. P. Bello, "Learning a robust Tonnetz-space transform for automatic chord recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 453–456.

[9] V. Lostanlen, S. Sridhar, B. McFee, A. Farnsworth, and J. P. Bello, "Learning the helix topology of musical pitch," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[10] S. Sridhar and V. Lostanlen, "Helicality: An isomap-based measure of octave equivalence in audio data," in *Proceedings of the International Society for Music Information Retrieval Late-Breaking/Demo Session (ISMIR-LBD)*, 2020.

[11] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," *arXiv:1709.04396*, 2017.

[12] Y. Kong, V. Lostanlen, G. Meseguer-Brocal, S. Wong, M. Lagrange, and R. Hennequin, "STONE: Self-supervised tonality estimator," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2024.

[13] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "PESTO: Pitch estimation with self-supervised transposition-equivariant objective," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2023.

[14] Y. Kong, G. Meseguer-Brocal, V. Lostanlen, M. Lagrange, and R. Hennequin, "S-KEY: Self-supervised learning of major and minor keys from audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[16] A. Beekman, L. Rademaker, and J. Van Wezel, "An introduction to spontaneous symmetry breaking," *SciPost Physics Lecture Notes*, p. 011, 2019.

[17] A. Barate and L. A. Ludovico, "Generalizing Messiaen's modes of limited transposition to a n-tone equal temperament," in *Proceedings of the International Conference on Sound and Music Computing (SMC)*, 2015, pp. 287–293.

[18] G. Richard, V. Lostanlen, Y.-H. Yang, and M. Müller, "Model-based deep learning for music information research: Leveraging diverse knowledge sources to enhance explainability, controllability, and resource efficiency," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 51–59, 2025.

[19] D. Marttila and J. D. Reiss, "Improving neural pitch estimation with SWIPE kernels," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[20] Y. Kong, V. Lostanlen, R. Hennequin, M. Lagrange, and G. Meseguer-Brocal, "Multi-class-token transformer for multitask self-supervised music information retrieval," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025.

[21] H. Honing, *The illiterate listener: On music cognition, musicality and methodology*. Amsterdam University Press, 2011.

VII. APPENDIX

We prove Lemma III.3 in two fashions: first, with elementary modular arithmetic; second, with abstract algebra.

Proof (elementary). Given $q, q' \in \mathbb{Z}/Q\mathbb{Z}$, suppose that $qP \bmod Q = q'P \bmod Q$, i.e., that $(q-q')P = 0 \bmod Q$. Since P and Q are coprime, it follows that Q divides $(q-q')$. Yet, since $0 \leq q, q' < Q$, $-Q < (q-q') < Q$. The only multiple of Q in this range is zero; thus, $q = q'$. We conclude by contraposition. ■

Proof (advanced). Since $\gcd(P,Q) = 1$, the coset $P + Q\mathbb{Z}$ is a unit in the quotient ring $\mathbb{Z}/Q\mathbb{Z}$; i.e., $P + Q\mathbb{Z} \in (\mathbb{Z}/Q\mathbb{Z})^{\times}$. Therefore, $q + Q\mathbb{Z} \mapsto qP + Q\mathbb{Z}$ is a ring automorphism. ■

Proof of Lemma V.2. Definition II.3 with $\boldsymbol{y}_A = \boldsymbol{y}_B = \boldsymbol{y}$ yields:

$$\mathscr{D}_{\omega,0}(\boldsymbol{y},\boldsymbol{y}) = \frac{1}{2}\Big(1 - |\hat{\boldsymbol{y}}[\omega]|^2\Big)^2. \tag{18}$$

By the triangular inequality on the DFT: $|\hat{\boldsymbol{y}}[\omega]| \leq \|\boldsymbol{y}\|_1 \leq 1$. Thus: $0 \leq (1 - |\hat{\boldsymbol{y}}[\omega]|^2) \leq 1$, concluding the proof. ■